Empirical Mode Decomposition for adaptive AM-FM analysis of Speech

Rajib Sharma

EMPIRICAL MODE DECOMPOSITION FOR ADAPTIVE AM-FM ANALYSIS OF SPEECH

A

Thesis submitted

for the award of the degree of

Doctor of Philosophy

By

RAJIB SHARMA

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

OCTOBER 2017

Certificate

This is to certify that the thesis entitled "**EMPIRICAL MODE DECOMPOSITION FOR ADAPTIVE AM-FM ANALYSIS OF SPEECH**", submitted by **RAJIB SHARMA** (136102011), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati

Prof. S. R. Mahadeva Prasanna

Dept. of Electronics and Electrical Engineering (EEE)
Indian Institute of Technology Guwahati (IITG)
Guwahati - 781 039, Assam, India.

Dedicated To

My mother, **Arati Devee**, and my father **Jadu Nath Sarma**

&

My supervisor **Prof. S. R. M. Prasanna**

&

My two sisters, **Gayatri Sarma** and **Meenakshi Sarma**

&

The **Goddesses** and Gods, of Kamakhya Dham, over the Nilachal Hills

and

the **vibrant and colourful cultures and traditions** that exist because of their mythology.

Acknowledgements

Just like me, this *acknowledgements* write-up will be different than usual - *not just mere praise, but also cynical, and even funny* for some readers. So, please keep reading !

There come certain moments in your life which influence and affect the path of your life (I am not talking about women !). As I write this thesis, I soberly realize that joining the Ph.D. programme in IITG in July 2013, and writing an email to **Prof. S. R. Mahadeva Prasanna**, on July 27, 11:33 hrs IST were two significant events in my life. SRMP, as is his famous nickname, replied back on July 28, 07:32 hrs IST, accepting my request to supervise my Ph.D. work. I have felt that those moments were due to blessings of my ancestors, and the unknown and ambiguous powers that mankind worships. At that time, I had felt as if I had cracked a jackpot. Actually, I was both surprised and euphoric, since SRMP already had a full cabinet of researchers under him. I had felt obliged, at that time, to show him the respect and dedication that befits a *guru*. As I reflect back over the last four years, I would like to thank him for providing me the opportunities, and the inspiration, to develop as a researcher. I can only wish you and your family a great life, professor !

During this journey of Ph.D. life, I must make a special mention here about my visit to Argentina, between February - July 2016, where I was able to live the *western culture*. Also, drink *matte*. That period of my life could never have been possible without the support of Prof. SRMP. Thanks to my doctoral committee head, **Prof. Samarendra Dandapat**, for financing me on that trip - LOL ! Thanks also to Mawsumi Debanth (didi) for her help. The people I met, lived and communicated with during that period will forever make great memories in my life. I would specifically like to mention *the researchers at UNL, Santa Fe,* for all the great times - David Campo, Leandro Bugnon, Sebastián Vanrell, Leandro Vignolo, José Omar Chelotti, Cristian Yones, Nahuel Deniz, and the rest. And *the interesting people I met in Santa Fe* - Ruslan Genov/Rosen Pavlov (Bulgaria), Jan Möller (Alemania), Thompsom Tomatieli (Brazil), Rafael Nunez (Mexico), Prof. Silvio Cornú (Santa Fe) and Gabriela Picard (Santa Fe). Regarding my research work in Argentina, I must give a special mention about two gurus of mine - **Prof. Hugo Leonardo Rufiner**, and **Prof. Gastón Schlotthauer** - who have supervised two of my works. Also, M.A. Colominas at UNER, Parana. All of you - please visit India !

The official tag of a philosopher (Ph.D.) is truly humbling, and the biggest achievement of my life. But, one cannot become a philosopher - you are, or you are not. I am sure it also has to do with all that you observe and experience, and your environment since you were born. As such, my mother,

my first and principal guru, **Arati Devee**, deserves a special mention - she is one of the most selfless, hard-working and mentally tough persons I have met till now in my life. I will be indebted to her not only in this life but for the lives I am reborn (if that actually happens) ! I must also mention about my father, **Jadu Nath Sarma**, for providing me the education to survive in this world. Similarly my two sisters, **Gayatri Sarma** and **Meenakshi Sarma**, for all the memories, and their help and support, however little they may be.

I believe being born in a cultural place is a great experience. Being born in Kamakhya, around the many temples of the Hindu Goddesses and Gods, has been and will be something that I always proudly cherish. Maybe someday I can do something for that place ! There have been critical times in my life that I have received unsolicited help, from strange sources, which has guided me the right way. I believe those are somehow due to the blessings of my ancestors, who have left behind the wonderful cultures and traditions spread across my mesmerizing motherland, **India**. Part of my great childhood days was my experience of studying in a Catholic school (being a Hindu), **Don Bosco High School Guwahati**. I can never forget the teachers, like **Xavier Sir**, in my 12-year school life since kindergarten. Similarly, I will always remember my **friends and teachers** in my graduate school, **National Institute of Technology Silchar** - college life is a unique experience.

It is almost impossible that your personal life remains like a dc curve in a period of around 4 years. *Bull Shit* happens in life (pardon my toungue) ! I have had my ups and downs too, and I thank *the people around me* - Nagaraj Adiga, Banriskhem Khonglah, Subhasish Mandal, Suman Deb, Ramesh K. Bhukya, K. Balaji, Nagendra Kumar, and others - for being part of those experiences. During this period of life, I have met some inspirational people, and some despicable ones. The bad experiences in life are much more important - hope I do not forget them, but learn from them, and be better prepared for the rest of my life. In this context, I must make a special mention about Prof. Dandapat, who, I think (unlike other peoples opinions) has a very wise and intelligent viewpoint on the philosophy of life, and how he wants to live it. You would want to hear him, even though you, like me, may have very different opinions.

Finally, as I write this thesis, I realize that I could have done much better. I am easily attracted (distracted) to art, politics, history, music, and everything unique (diversity helps) - unique is beautiful. A lot of my time goes into *surfing the internet and youtube* - watching, absorbing, and discussing stuff. From defence videos, (particularly those beautiful military aircrafts), to world and international

politics, to people farming in their greenhouses, to ladies playing football (soccer) and softball (and baseball), to listening to songs in languages I do not know or understand, to reaction videos of people crying and laughing - the list is big and ever growing. But, that is part of me - if I had been more focussed, I would have had better results - may be, who knows (definitely, I would say). In my leisure, I also watch a lot of **Japanese anime and manga** (Naruto, Bleach, One Piece as examples) and series (The Big Bang Theory, Game Of Thrones as examples). Thanks to the people who make those. Lastly, I must mention that I am a very patriotic person. As such, I must thank the **selfless warriors of India** because of whom I am alive, and sleep peacefully, in this troubled world.

With this, I seek the blessings of my well-wishers and gurus so that I may win every battle in life, defeat my foes to submission (I mean it !), and make you proud that I have been part of your life at some point in time ! Bless me for success ! Bless me so that I become focussed like Arjuna and Laxmana (characters in Indian mythology) ! For as Lord Baelish (a.k.a. Littlefinger) in Game of Thrones says in his famous quote : **"Chaos isn't a pit. Chaos is a ladder. Many who try to climb it fail and never get to try again. The fall breaks them. And some, are given a chance to climb. They refuse, they cling to the realm or the gods or love. Illusions. Only the ladder is real. The climb is all there is."**

Finally, I would like to share the poem **"Invictus"**, that I also shared in my last farewell mail, the day I left my job in Ericsson Kolkata on 28 October, 2012.

Invictus

Out of the night that covers me,
Black as the pit from pole to pole,
I thank whatever gods may be
For my unconquerable soul.

In the fell clutch of circumstance
I have not winced nor cried aloud.
Under the bludgeonings of chance
My head is bloody, but unbowed.

Beyond this place of wrath and tears

Looms but the Horror of the shade,

And yet the menace of the years

Finds, and shall find me, unafraid.

It matters not how strait the gate,

How charged with punishments the scroll.

I am the master of my fate.

I am the captain of my soul.

(William Ernest Henley)

God bless you all (meaning my well-wishers only) !

Rajib Sharma

Abstract

This thesis investigates and endorses a non-linear and non-stationary data analysis technique, called *Empirical Mode Decomposition* (EMD), as an alternative to the conventional methods of speech analysis. The popular methods of speech analysis rely on the assumptions of short-time stationarity and linearity of the processes producing the speech signal. These assumptions are arguable, and incorporated more from the viewpoint of convenience and simplicity of analysis. The *source-filter theory* of speech production, and the *Mel filterbank*, are inherently dependent on the *Short Time Fourier Transform* (STFT). They are unable to capture information embedded in the *non-linear characteristics* of the speech signal, and are *non-adaptive*.

Amplitude Modulation - Frequency Modulation (AM-FM) analysis aims to represent the speech signal in terms of a finite number of AM-FM signals, representing its time-varying *vocal tract resonances*. The simple method of *Multiband Demodulation Analysis* (MDA) tries to achieve this representation by using a filterbank comprising of a large number of overlapping band-pass filters. This, however, makes MDA non-adaptive. As such, the ability of EMD to decompose any real world signal into *oscillatory* or AM-FM components, called its *Intrinsic Mode Functions* (IMFs), without using any *a priori* basis, has caught the attention of the research community. However, along with this wonderful capability also comes some undesired characteristics, more specifically *mode-mixing*.

This thesis commences with a study of EMD as an *adaptive* AM-FM analysis technique, its characteristics, and its advanced noise-assisted variants used to curtail mode-mixing. Based on this study, this thesis first proposes a *modified* EMD algorithm, captioned MEMD, which curtails mode-mixing, but at a much faster rate than the existing advanced noise-assisted variants. Thereafter, a study is done on the ability of EMD, and its variants, in decomposing the speech signal into IMFs which represent its latent *source* and *system* characteristics. Following this study, this thesis proposes the application of EMD, and its variants, in two speech processing applications. Firstly, this thesis proposes a principle or framework to detect the *Glottal Closure Instants* (GCIs) of the speech signal, using its IMFs. The objective here is to provide an alternative to the state-of-the-art methods based on short-time *Linear Prediction* (LP) analysis, which is based on the source-filter theory. Secondly, this thesis investigates the capability of the IMFs of the speech signal in capturing speaker-specific information which can be complementary to the Mel filterbank.

The major contributions of this thesis are as follows :

- Development of a *modified* EMD algorithm, captioned MEMD, which reduces mode-mixing, and is faster than the current noise-assisted EMD methods like the *Ensemble Empirical Mode Decomposition* (EEMD), and its newer versions.
  - Show that the MEMD algorithm reduces mode-mixing manifested in the EMD algorithm.
  - Show that the MEMD algorithm provides a *better distribution of the formants structure of voiced speech*, in its IMFs, than the EMD algorithm.
  - Show that the MEMD algorithm is faster than the noise-assisted EMD methods.

- Study the ability of EMD, MEMD, and a recently developed noise-assisted EMD method, called the *Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (ICEEMDAN), in extracting the source and system characteristics of the speech signal.
  - Investigate, using synthetic speech signals, the effect of change of *pitch* or *fundamental frequency*, *center* or *resonant frequencies* of the vocal tract resonators, and their *bandwidths*, on the decomposition of the speech signal.
  - Investigate, using *cepstral* or *homomorphic* analysis, the source and system characteristics represented in the IMFs of natural speech signals corresponding to different *phones* or *speech sounds*.
  - Investigate the capability of the IMFs of speech signals in representing their latent source and system characteristics, when the speech signals are subjected to telephone channel codecs.

- Propose a principle/framework for detecting the GCIs of voiced speech, without using short-time LP analysis, and which can provide reliable GCIs estimates under varied conditions.
  - Develop a method for detecting the GCIs of the speech signal, from its IMFs obtained using ICEEMDAN.
  - Develop a method for detecting the GCIs of the speech signal, from its IMFs obtained using MEMD.

- Show that the performances are consistent under clean, noisy, and telephone channel conditions, and comparable with the state-of-the-art methods.

- Show that the adaptive filterbank nature of EMD/MEMD captures speaker-specific information that can complement the Mel filterbank.
  - Show that only a small subset of the IMFs are useful in augmenting the performance of text-independent *Speaker Verification* (SV) system.
  - Show that *cepstral* or *energy-like* features, obtained from a small subset of the IMFs, are more useful than the higher dimensions of the *Mel filterbank Cepstral Coefficients* (MFCCs), in capturing speaker characteristics.
  - Show that a small subset of the IMFs is more useful than a large but fixed *Gabor filterbank*.
  - Show that the IMFs are useful not only for *normal* speech, but also for *fast* and *whispered* speaking styles, and under *insufficient test data* conditions.

Keywords: Non-linear and non-stationary, Mel filterbank, MFCCs, AM-FM , EMD, EEMD, ICEEMDAN, MEMD, IMFs, Vocal tract resonators, Formants, Source, System, Pitch, Cepstrum, GCIs, SV, Adaptive filterbank

Contents

List of Figures

List of Tables

List of Acronyms

AM-FM	Amplitude Modulation – Frequency Modulation
BPF	Band Pass Filter
BPF-IGE	Band Pass Filtering based mimicking of IGE
BPF-MGE	Band Pass Filtering based mimicking of MGE
CEEMD	Complementary Ensemble Empirical Mode Decomposition
CEEMDAN	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
DCF	Detection Cost Function
DCT	Discrete Cosine Transform
dEGG	difference ElectroGlottoGraph
DESAs	Discrete-time Energy Separation Algorithms
DFT	Discrete Fourier Transform
Di	Method of obtaining the Interpolation Points, where i=1,2,3,4
DWT	Discrete Wavelet Transform
DYPSA	Dynamic Programming and Phase Slope Algorithm
EAs	Evolutionary Algorithms
EEMD	Ensemble Empirical Mode Decomposition
EER	Equal Error Rate in %
EF	Epoch Filter
EGG	ElectroGlottoGraph
EMD	Empirical Mode Decomposition
Ext. MFCCs	Extended Mel Filterbank Cepstral Coefficients
FAR	False Alarm Rate in %
fGn	fractional Gaussian noise
Fi	The i^{th} principal formant, where i=1,2,3,4

GCIs	Glottal Closure Instants
GD	Group Delay
GMM	Gaussian Mixture Model
HEGD	Hilbert Envelope and Group Delay
HHT	Hilbert Huang Transform
HTL	High-Time Liftered
IA	Identification Accuracy in ms
IA'	Accuracy to $\pm$ 0.25 ms, in %
IAE	Instantaneous Amplitude Envelope
ICEEMDAN	Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
IDFT	Inverse Discrete Fourier Transform
IE	Identification Error in Hz
IF	Instantaneous frequency
IGE	ICEEMDAN based GCIs Estimation
ILPR-PI	Integrated Linear Prediction Residual using Plosion Index
IMF	Intrinsic Mode Function
IPs	Interpolation Points
IR	Identification Rate in %
ITU	International Telecommunication Union
LDA	Linear Discriminant Analysis
LP	Linear Prediction
LTI	Linear Time Invariant
LTL	Low-Time Liftered
M1-EMD	Modified Empirical Mode Decomposition method 1
M2-EMD	Modified Empirical Mode Decomposition method 2
M3-EMD	Modified Empirical Mode Decomposition method 3
MDA	Multiband Demodulation Analysis
MEMD	Modified Empirical Mode Decomposition
MFCCs	Mel Filterbank Cepstral Coefficients

MGE	MEMD based GCIs Estimation
MR	Miss Rate in %
OLA	OverLap and Add
PS	Power Spectrum
PSD	Power Spectrum Density (dB)
RAPT	Robust Algorithm for Pitch Tracking
RMS	Root Mean Square
SEDREAMS	Speech Event Detection using the Residual Excitation And a Mean-based Signal
SEDS	Source Enhanced and Detrended Speech
SF	Number of Spurious Formants
STFT	Short-Time Fourier Transform
SV	Speaker Verification
T-1	Type-1 telephone quality speech obtained using the VOICEBOX toolbox
T-2	Type-2 telephone quality speech obtained in accordance with the ITU standards
TEO	Teager Energy Operator
UBM	Universal Background Model
VAD	Voice Activity Detection
WCCN	Within Class Covariance Normalization
WPT	Wavelet Packets Transform
WT	Wavelet Transform
YAGA	Yet Another GCI detection Algorithm
ZFR	Zero Frequency Resonator

List of Symbols

Chapter 1

$s(n)$	Discrete-time digital speech signal/utterance, or a segment of it
$e(n)$	Impulse excitation signal at the glottis, according to the source-filter theory
$g(n)$	Impulse response of the low pass filter at the glottis, according to the source-filter theory
$v(n)$	Impulse response of the vocal tract resonator system, according to the source-filter theory
$r(n)$	Impulse response of the high pass filter at the lips, according to the source-filter theory
$S(z)$	Z-Transform of $s(n)$
$E(z)$	Z-Transform of $e(n)$
$G(z)$	Z-Transform of $g(n)$
$V(z)$	Z-Transform of $v(n)$
$R(z)$	Z-Transform of $r(n)$
F_s	Sampling frequency of a discrete-time digital signal
F_0	Pitch or fundamental frequency of a voiced speech signal
K_i	Number of excitation impulses in $e(n)$, for a voiced speech signal
p_k , $p_k^\star$	Locations of the k^{th} conjugate pole pair in $V(z)$
P	Number of conjugate pairs of poles in $V(z)$
q_k	The k^{th} coefficient of the polynomial representing the denominator of $V(z)$
b	Location of the double pole of the transfer function of the glottis, $G(z)$
c	Location of the zero of the transfer function of the lips, $R(z)$

$\delta(.)$	Unit impulse function, both continuous-time and dicrete-time, depending on the variable inside the brackets
$s(t)$	Continuous-time speech signal/utterance, or a segment of it
$S(\tau, f)$	STFT of $s(t)$, τ being the time variable, and f the analog frequency
$w(t)$	Continuous-time symmetric window of finite time-width
$s'(n)$, $S'(z)$	Pre-emphasized discrete-time digital speech signal, and its Z-Transform
$\hat{V}(z)$	Transfer function of a hypothetical single resonator vocal tract system, estimated using LP analysis
$\hat{q}_1$, $\hat{q}_2$	The two coefficients of the polynomial representing the denominator of $\hat{V}(z)$
$\tilde{q}_1$, $\tilde{q}_2$	Errors in the two estimated coefficients ($\hat{q}_1$, $\hat{q}_2$) with respect to their actual values (q_1 , q_2)
b_1 , b_2	The two coefficients of the polynomial representing the numerator hypothetical vocal tract system with a single antiresonance
$x^i_{k,j}$	Subband energy of the k^{th} frequency band of the j^{th} speech frame belonging to the i^{th} speaker
u^i_k	Average sub-band energy of the k^{th} frequency band for the i^{th} speaker
N^i	Number of speech frames for all the utterances belonging to the i^{th} speaker
u_k	Average sub-band energy of the k^{th} frequency band for all the speakers
M^s	Number of speakers
F^{ratio}_k	F-ratio for the k^{th} frequency band
$R_k(t)$, N_f	The k^{th} AM-FM component (representing a vocal tract resonance) of $s(t)$, and the number of such resonances
$a^R_k(t)$	Amplitude modulating signal of $R_k(t)$
f^R_k	Resononant or center frequency of $R_k(t)$
θ	Constant Phase of $R_k(t)$
$x(t)$, $X(f)$	An arbitrary continuous-time signal, and its Fourier Transform
$H[.]$	Hilbert Transform operator
$x_a(t)$, $X_a(f)$	Analytical signal corresponding to $x(t)$, and its Fourier Transform

$a(t)$, $\phi(t)$, $f(t)$	Instantaneous amplitude envelope, phase function, and frequency function of $x(t)$
$x(n)$, $y(n)$	An arbitrary discrete-time digital signal, and its first difference, used in TEO
$\Psi[.]$	TEO
$\omega(n)$, $a(n)$	Instantaneous angular frequency function, and envelope function, estimated from $x(n)$.
f_{res}	An arbitrary formant frequency of a voiced speech signal, centered at which a Gabor BPF is applied
$s_{bpf}(n)$	Signal obtained after applying a Gabor BPF on a voiced speech signal
$s_{syn}(n)$	Synthetic voiced speech signal generated with a synthetic single resonator vocal tract system
$v_{syn}(n)$, $V_{syn}(z)$	Impulse response of a synthetic single resonator vocal tract system, and its Z-Transform
$p_{syn,1}$, $p^{\star}_{syn,1}$	Locations of the conjugate pole pair in $V_{syn}(z)$
$s_{mod}(n)$	Synthetic AM-FM signal
$h_g(t)$, $H_g(f)$	Continuous-time Gabor filter, and its Fourier Transform
f_g , α_g	Analog center frequency, and bandwidth controlling parameter, of $H_g(f)$
F_w , B_w	Mean amplitude weighted instantaneous frequency, and bandwidth, of an analysis frame of an arbitrary AM-FM signal
t_0 , T	Starting instant, and time-duration, of an analysis frame of an arbitrary AM-FM signal
t_{fr} , f^g_k	Starting instant of a analysis frame, and analog center frequency of the k^{th} Gabor filter of a Gabor filterbank, used in creating a pyknogram
$F_w(t_{fr}, f^g_k)$	Time-frequency distribution of an analysis frame using the k^{th} Gabor filter, in a Pyknogram
$F_w(t, f)$	Complete Time-frequency distribution, for all analysis frames and all the Gabor filters, in a Pyknogram

Chapter 2

$s(t)$	Continuous-time speech signal/utterance, or a segment of it
$r_k(t)$, $h_k(t)$	The k^{th} outer residue, and IMF, of a continuous-time signal, generally speech
η , N	Sifting iteration index, and the number of sifting iterations, in a sifting process
t_{min} , t_{max}	X-coordinates of the IPs, corresponding to the minima and maxima envelopes
y_{min} , y_{max}	Y-coordinates of the IPs, corresponding to the minima and maxima envelopes
$e_{min}(t)$, $e_{max}(t)$, $e_m(t)$	Minima, maxima, and mean envelopes
M	Maximum number of IMFs to be extracted after which the EMD (or any variant of EMD) algorithm is terminated
$r_M(t)$	Final residue of a continuous-time signal, generally speech
$s(n)$	Discrete-time digital speech signal/utterance, or a segment of it
$r_M(n)$, $h_k(n)$	Final residue, and the k^{th} IMF, of $s(n)$
IMF_k	The k^{th} IMF
$x_g(t)$, $x_s(t)$, $x_n(t)$	Noisy sinusoidal signal, its sinusoidal component, and its white noise component
$a_k(t)$, $\theta_k(t)$, $f_k(t)$	Instantaneous amplitude envelope, phase, and frequency, of $h_k(t)$
$H(f,t)$	Hilbert spectrum
$h(f)$, T_{dur}	Marginal Hilbert spectrum, and the total duration of the signal for which the Hilbert spectrum is computed
$IE(t)$	Instantaneous energy density computed from the Hilbert spectrum
$x_H(n)$, σ_H , H	A sequence of fGn, its standard deviation, and its Hurst component
$r_H(m)$	Autocorrelation function of $x_H(n)$ at lag m
$\mathbb{E}$	Expectation operator
$z_H(k)$	Number of zero crossings of the k^{th} IMF of $x_H(n)$
ρ_H	Average decrease rate of the number of zero crossings of the IMFs of $x_H(n)$

$S_{k,H}(f)$	PSD of the k^{th} IMF of $x_H(n)$
α_H	Parameter related to Hurst component
$V_H(k)$	Variance of the k^{th} IMF of $x_H(n)$
κ_H	Slope of log-linearized variance curve of the IMFs of $x_H(n)$
H_{est}	Estimated Hurst component value
$\gamma(t)$	Minimizing parameter of the local-global stopping criterion
Θ_1 , Θ_2	Thresholds for the local-global stopping criterion
α_F	Parameter representing the fraction of the duration of the signal, in the local-global stopping criterion
$x_{mix}(t)$, $x_l(t)$, $x_h(t)$	Binary mixture of sinusoids, its low-frequency component, and its high-frequency component
F_s	Sampling frequency of a discrete-time digital signal
a_l , a_h , a_{rat}	Amplitude of the low-frequency sinusoid, that of the high-frequency sinusoid, ratio of the two amplitudes
f_l , f_h , f_{rat}	Frequency of the low-frequency sinusoid, that of the high-frequency sinusoid, ratio of the two frequencies
ϕ_l , ϕ_h , ϕ_d	Phase of the low-frequency sinusoid, that of the high-frequency sinusoid, difference of the two phases
$d_1^{(N)}(a_{rat}, f_{rat}, \phi)$	First IMF obtained from EMD of $x_{mix}(t)$
$c_1^{(N)}(a_{rat}, f_{rat}, \phi_d)$	Parameter evaluating the ability of EMD to decompose $x_{mix}(t)$
$s^l(t)$	The l^{th} noisy copy of the speech signal, $s(t)$, in EEMD
$w^l(t)$, L	The l^{th} white noise signal, and the number of white noise signals, in EEMD/ICEEMDAN
β	Parameter which controls the variance of noise with respect to the signal, in EEMD
$h_k^l(t)$, $r_M^l(t)$	The k^{th} IMF, and the final residue, of $s^l(t)$, in EEMD
$\bar{s}(t)$	Sum of the IMFs of $s(t)$, obtained after EEMD
$var[.]$	Operation of calculating variance
$\bar{s}(n)$	Sum of the IMFs of $s(n)$, obtained after EEMD
$S_k(f)$, F_k^m	Power spectrum of the k^{th} IMF, and its mean frequency

$e_{\mathrm{EGG}}(n)$	Discrete-time digital EGG signal
R_k^e	Maximum correlation of the k^{th} IMF of $s(n)$ with $e_{\mathrm{EGG}}(n)$
$H_{pre}(z)$	High-pass filter used for pre-emphasis
$E_k[.]$	Operation of extracting the k^{th} IMF using EMD
$\Upsilon[.]$	Operation of extracting the local mean of the signal using EMD
$r_{k-1}(t)$, $r_{k-1}^l(t)$	The $(k-1)^{th}$ outer residue, and its noisy version mixed with the l^{th} noise signal, in ICEEMDAN
ϵ_0 , β_0 , β_k	Parameters which control the SNR in ICEEMDAN
$std(.)$	Operation of calculating standard deviation
r_e	Reconstruction error of a speech signal
$\mathcal{W}_s^{\psi}(\tau, r)$	Continuous-time WT of $s(t)$, using mother wavelet $\psi(t)$, at time τ and scale r
$\mathbb{W}_s^{\psi}[\tau, 2^j]$	DWT of $s(n)$, using mother wavelet $\psi(n)$, at time τ and level j

Chapter 3

$x(t)$, $X(f)$	An arbitrary continuous-time signal, and its Fourier Transform
$s(t)$	Continuous-time speech signal/utterance, or a segment of it
$r_k(t)$, $h_k(t)$	The k^{th} outer residue, and IMF, of s(t)
η , N	Sifting iteration index, and the number of sifting iterations, in a sifting process
t_{min} , t_{max}	X-coordinates of the IPs, corresponding to the minima and maxima envelopes
y_{min} , y_{max}	Y-coordinates of the IPs, corresponding to the minima and maxima envelopes
$e_{min}(t)$, $e_{max}(t)$, $e_m(t)$	Minima, maxima, and mean envelopes
M	Maximum number of IMFs to be extracted after which the EMD (or any variant of EMD) algorithm is terminated
$r_M(t)$	Final residue of a continuous-time signal
$s(n)$	Discrete-time digital speech signal/utterance, or a segment of it

$r_M(n)$, $h_k(n)$	Final residue, and the k^{th} IMF, of $s(n)$
$z(t)$	Inner residue signal
F_s	Sampling frequency of a discrete-time digital signal
F_k^m , nF_k^m , $S_k(f)$	Mean frequency, log-normalized mean frequency, and power spectrum, of the k^{th} IMF
IMF_k	The k^{th} IMF
M^{uc}	Number of components obtained from an unconstrained decomposition by EMD (or its variant)
r_e^{uc}	Reconstruction error of a speech signal, for an unconstrained decomposition
M_{trend}^{uc}	Number of very low-frequency trend-like IMFs of a speech signal, obtained from an unconstrained decomposition
$s^x(n)$, $w^x(n)$	Noisy discrete-time digital speech signal, and its white noise component, the SNR being x dB
$h_k^x(n)$, e_k^x	The k^{th} IMF of $s^x(t)$, and its similarity error with $w^x(n)$
F_i^r , F_i^{est}	Reference formant frequency, and estimated formant frequency, of Fi
$H_{pre}(z)$	High-pass filter used for pre-emphasis
N_v , IE_i	Total number of voiced frames, and the Identification Error, for Fi of a speech signal/utterance
$F^{det} = \{F_1^{det}, F_2^{det}, ..., F_D^{det}\}$	Set of peaks detected using LP analysis on the IMFs, D being the number of peaks

Chapter 4

$s(n)$	Discrete-time digital speech signal/utterance, or a segment of it
F_s	Sampling frequency of a discrete-time digital signal
$e(n)$	Impulse excitation signal at the glottis, according to the source-filter theory
$g(n)$	Impulse response of the low pass filter at the glottis, according to the source-filter theory

$v_r(n)$	Impulse response of the r^{th} resonator of a vocal tract resonator system, according to the source-filter theory
$r(n)$	Impulse response of the high pass filter at the lips, according to the source-filter theory
$S(z)$, $S(f)$	Z-Transform, and DFT, of $s(n)$
$E(z)$	Z-Transform of $e(n)$
$G(z)$	Z-Transform of $g(n)$
$V_r(z)$	Z-Transform of $v_r(n)$
$R(z)$	Z-Transform of $r(n)$
K_i	Number of excitation impulses in $e(n)$, for a voiced speech signal
F_0	Pitch or fundamental frequency of a voiced speech signal
p_r , $p_r^\star$	Locations of the r^{th} conjugate pole pair in $V_r(z)$
f_r , B_r	Analog center frequency, and bandwidth, of $V_r(z)$
M	Maximum number of IMFs to be extracted after which the EMD (or any variant of EMD) algorithm is terminated
$g^e(n)$, $G^e(z)$, $G^e(f)$	Glottal constituent or source component of $s(n)$, its Z-Transform, and its DFT
$v_r^e(n)$, $V_r^e(z)$	The r^{th} vocal tract resonator constituent of $s(n)$, and its Z-Transform
$r^e(n)$, $R^e(z)$	Lip-radiation constituent of $s(n)$, and its Z-Transform
$h_k(n)$	The k^{th} IMF, of $s(n)$
IMF_k	The k^{th} IMF
$C[.\ ,\ .]$, $\sigma[.\ ,\ .]$	Operators for evaluating maximum correlation coefficient, and cross-covariance, between two signals
$\mathbb{E}$	Expectation operator
K_B	Parameter for changing the bandwidth of the vocal tract resonators
K_f	Parameter for changing the center or resonant frequencies of the vocal tract resonators
$h^v(n)$, $H^v(z)$, $H^v(f)$	System constituent of a speech signal, its Z-Transform, and its DFT
$v(n)$, $V(z)$	Combined impulse response of the cascade of vocal tract resonators, and its Z-Transform

$\hat{s}(n)$, $\hat{g}^e(n)$, $\hat{h}^v(n)$	Cepstrum of $s(n)$, its source component, and its system component
$l(n)$	Liftering window
$\sqcap[.]$, $\sqcup[.]$	Operation of obtaining LTL cepstrum, and operation of obtaining HTL cepstrum
$\mathcal{F}\{.\}$	Operation of DFT
$\bigcap_K$	Euclidian distance between the LTL cepstra of a speech signal and the partial sum of its IMFs, K being the IMF number till which the summation is considered
$\bigcup_K$	Euclidian distance between the HTL cepstra of a speech signal and the partial sum of its IMFs, K being the IMF number from which the summation is considered
$s^T(n)$, $h_k^T(n)$	ITU-coded version of $s(n)$, and its k^{th} IMF
$\bigwedge^T$, $\bigvee^T$	Euclidian distance between the LTL cepstra, and the HTL cepstra, of $s(n)$ and $s^T(n)$
$\bigwedge_k^T$, $\bigvee_k^T$	Euclidian distance between the LTL cepstra, and the HTL cepstra, of $h_k(n)$ and $h_k^T(n)$

Chapter 5

$s(n)$	Discrete-time digital speech signal/utterance, or a segment of it
$e(n)$	Impulse excitation signal at the glottis, according to the source-filter theory
F_s	Sampling frequency of a discrete-time digital signal
F_0	Pitch or fundamental frequency of a voiced speech signal
f_r , B_r	Analog center frequency, and bandwidth, of the r^{th} vocal tract resonator
F^m	Mean frequency of any arbitrary signal
F_k^m , $S_k(f)$	Mean frequency, and power spectrum, of the k^{th} IMF of $s(n)$
l_m^{ref}	Location of the m^{th} impulse of the excitation signal of a synthetic speech signal

$a_3(n)$, $f_3^m(n)$	Time-varying amplitude envelope, and the average frequency, of the third IMF of a synthetic speech signal
F_0^{ref}	Reference average pitch frequency, estimated from a dEGG signal
$h_k(n)$	The k^{th} IMF, of $s(n)$
IMF_k	The k^{th} IMF
$s_d(n)$	Detrended speech signal, constructed from the IMFs of $s(n)$
F_0^{est}	Estimated average pitch frequency of a speech signal, using the RAPT algorithm
$s_R(n)$	Signal from which the regions of search are obtained
R_r	The r^{th} region of search of the GCIs
l_r^i , I_{ige}	The r^{th} initial GCI estimate, and the number of initial GCIs estimates, obtained from the IGE/BPF-IGE algorithm
$s_e(n)$	SEDS signal obtained from $s(n)$
L_r	The r^{th} region, within which false GCIs estimates are eliminated
W , ν	Window size, and its controlling parameter, for eliminating spurious GCIs estimates
v_r^i	Amplitude of $s_e(n)$ at l_r^i. Also, amplitude of $e_d(n)$ at n_r^i
l_r^{est}	The r^{th} final GCI estimate
F_{ige}	Number of final GCIs estimates, obtained from the IGE/BPF-IGE algorithm
$s_h(n)$, n_d	Second difference of $s_d(n)$, and its minima locations/time-instants
Δ	Difference operation for a discrete-time digital signal
R_i^r , I_{mge}	The r^{th} initial region of search of the GCIs, and the number of such regions, obtained in the MGE/BPF-MGE algorithm
n_r^i	The time-instant of minimum amplitude of $e_d(n)$, for the region R_r^i
R_r^f , F_{ige}	The r^{th} final region of search, and the number of final GCIs estimates, obtained in the MGE/BPF-MGE algorithm
$s^i(n)$, $S^i(f)$	The i^{th} segment/frame of $s(n)$, and its DFT
$s_d^i(n)$, $S_d^i(f)$	Detrended speech segment obtained by high-pass filtering $s^i(n)$, and its DFT

$s_R^i(n)$, $S_R^i(f)$	Signal obtained by band-pass filtering $s^i(n)$ around its first two pitch harmonics, and its DFT
$s_e^i(n)$, $S_e^i(f)$	Signal obtained by band-pass filtering $s^i(n)$ around its first five pitch harmonics, and its DFT

Chapter 6

F_s	Sampling frequency of a discrete-time digital signal
$s(n)$	Discrete-time digital speech signal/utterance, or a segment of it
IMF_k	The k^{th} IMF
F_k^m , $S_k(f)$	Mean frequency, and power spectrum, of the k^{th} IMF or its segment/frame
$h_k(n)$, $h_k^i(n)$	The k^{th} IMF of $s(n)$, and its i^{th} segment/frame
N_f	Number of samples in a segment/frame of a discrete-time digital speech signal
K	Number of IMFs from which features are extracted
l_k^i	Log Sum Squared Amplitude feature obtained from the k^{th} IMF or AM-FM component of the i^{th} voice/speech frame of a speech signal
L_K^i	Log Sum Squared Amplitude feature vector obtained from K number of IMFs or AM-FM components of the i^{th} voice/speech frame of a speech signal
L_K	Log Sum Squared Amplitude feature set obtained from K number of IMFs or AM-FM components of all voice/speech frames of a speech signal
$\bar{V}_K^i = [\bar{v}_1^i, \bar{v}_2^i, ..., \bar{v}_K^i]^T$	An arbitrary K-dimensional feature vector extracted from the i^{th} speech frame
$\text{c}\bar{V}_K^i = [\text{c}\bar{v}_1^i, \text{c}\bar{v}_2^i, ..., \text{c}\bar{v}_K^i]^T$	Refined K-dimensional feature vector after applying DCT on $\bar{V}_K^i$
$\text{c}\bar{V}_K$	Refined K-dimensional feature set for all voice/speech frames
cL_K^i	Refined Log Sum Squared Amplitude feature vector
cL_K	Refined Log Sum Squared Amplitude feature set

List of Symbols

g_k^i	Sum Log Squared Amplitude feature obtained from the k^{th} IMF or AM-FM component of the i^{th} voice/speech frame of a speech signal
G_K^i	Sum Log Squared Amplitude feature vector obtained from K number of IMFs or AM-FM components of the i^{th} voice/speech frame of a speech signal
G_K	Sum Log Squared Amplitude feature set obtained from K number of IMFs or AM-FM components of all voice/speech frames of a speech signal
cG_K^i	Refined Sum Log Squared Amplitude feature vector
cG_K	Refined Sum Log Squared Amplitude feature set
e_k^i	Entropy feature obtained from the k^{th} IMF or AM-FM component of the i^{th} voice/speech frame of a speech signal
$p_k^i(x)$	Probability distribution of the amplitudes of $h_k^i(n)$ derived by 16-level histogram quantization
$X = \{x_1, x_2, ..., x_{16}\}$	Alphabet of 16 symbols obtained by histogram quantization of $h_k^i(n)$
E_K^i	Entropy feature vector obtained from K number of IMFs or AM-FM components of the i^{th} voice/speech frame of a speech signal
E_K	Entropy feature set obtained from K number of IMFs or AM-FM components of all voice/speech frames of a speech signal
cE_K^i	Refined Entropy feature vector
cE_K	Refined Entropy feature set
$H_{pre}(z)$, c_{pre}	High-pass filter used for pre-emphasis, its zero coefficient
F_0	Pitch or fundamental frequency of a voiced speech signal
f_r , B_r	Analog center frequency, and bandwidth, of the r^{th} vocal tract resonator
S_u	Mean supervector of a speech utterance
S_{UBM}	Mean supervector of the UBM dataset
T_v	Total variability matrix
i_u	i-vector of a speech utterance
$\hat{i}_{tr}$	i-vector of a training speech utterance, after being processed by LDA and WCCN

$\hat{i}_{ts}$	i-vector of a testing speech utterance, after being processed by LDA and WCCN
i_s	Cosine kernel scrore between $\hat{i}_{tr}$ and $\hat{i}_{ts}$
λ^s , D	Model (GMM) of the feature space of a speaker, and dimension of the feature space
w_m^s , μ_m^s , Σ_m^s	Mixture weight, mixture mean, and mixture covariance (diagonal), of the m^{th} Gaussian mixture of a speaker GMM
$\bar{X} = \{\bar{X}_1, \bar{X}_2, ..., \bar{X}_F\}$	An arbitrary feature set of F number of feature vectors
$\mathcal{N}(\bar{X}_i/\mu_m^s, \Sigma_m^s)$	Gaussian distribution function for the m^{th} Gaussian mixture
$\mathrm{S}c = \{\mathrm{sc}_1, \mathrm{sc}_2, ..., \mathrm{sc}_S\}$	Set of scores obtained after testing against the speaker models
ξ^s , ξ_0^s	Decision threshold for verifying a claim in SV, and a particular value of it
S_G	Number of genuine/true claims that is accepted in SV
S_I	Number of imposter/false claims that is rejected in SV
S_M	Number of genuine/true claims that is rejected in SV
S_F	Number of imposter/false claims that is accepted in SV
P_M , P_F	Probability of Miss, and Probability of False Alarm
C_M , C_F	Cost of a miss, and cost of a false alarm
P_T	Probability of genuine/true claims
C_{ξ^s}	Cost parameter of the DCF at any given value of the decision threshold

1

Introduction

Contents

1. Introduction

Speech is the principal method of communication amongst human beings. It is a signal generated by a complex psycho-acoustic process developed as a result of thousands of years of human evolution. However, it is not just a tool for communication. It is a signal which contains a multitude of information like the speaker's age, height, emotion, accent, health and physiological disorders, identity, etc., which give rise to the various fields of Speech Processing today [2, 13, 14]. However, speech is a highly non-linear and non-stationary signal, and hence unearthing such information is not a trivial task [2, 15]. Even though methods for analyzing non-stationary signals like the *Wavelet Transform* (WT) and the Wigner-Ville Transform have been developed, they have not been popular in the speech processing community mainly because they decompose the signal in an alternate domain and introduce additional complexity to the analysis [16–18]. Thus, the *source-filter theory* of speech production has remained the backbone of speech processing [2, 13, 14]. The treatment of the speech signal as being linear and stationary for short intervals of time (10-50 ms) gives a simplistic and time-affordable analysis. Such an analysis, though, is arguable and provides an oversimplified view of the phenomena related to speech production [1, 15, 19, 20]. Thus, the *Linear Prediction* (LP) analysis of speech provides us with a noisy excitation signal as a representation of the glottal source, and a vocal tract filter which represents the resonant cavities of the vocal tract (the high-pass filter characteristics of the lips is included in the filter) [2, 13, 14]. Further, the oversimplification of the speech production process makes LP analysis vulnerable to errors [21].

From the speech perception point of view, the *Mel filterbank*, which is based on the characteristics of the human ear, has been widely appreciated in speech processing [2, 13, 14]. The *Mel Filterbank Cepstral Coefficients* (MFCCs) are derived solely from the magnitude spectrum (or power spectrum) of the speech signal, while neglecting its phase spectrum. However, the phase spectrum of speech is equally critical to speech perception as the magnitude spectrum, and has found important use in many speech processing applications [12, 22–26]. Again, it has been observed that while the Mel filterbank is used for a variety of speech applications, it does not always provide the optimal features, and filterbanks tuned for different applications might be more suitable for better results [3, 27–30].

To overcome some of these limitations of conventional speech analysis the sinusoidal representation of speech was proposed, which models the speech signal as being constituted of a finite number of time-domain sinusoidal signals, in terms of their frequencies, amplitudes, and phases [4]. The sinusoidal model and the Teager Energy Operator (TEO) provided the impetus for the *Amplitude Modulation*

- *Frequency Modulation* (AM-FM) representation of the speech signal [15, 31–35]. The concept of Multiband Demodulation Analysis (MDA) was introduced, wherein the speech signal is passed through a *parallel bank of fixed band-pass filters*, generating different time-domain signals from the speech signal. The generated signals are then represented in terms of their instantaneous frequencies and amplitudes, as estimated from the Hilbert Transform or TEO [15,34,35]. In the recent years, such a representation has been found to be useful in many areas of speech processing [15].

The rest of this chapter is organized as follows: Section 1.1 discusses the non-linearity and non-stationarity of speech. The limitations of conventional short-time analysis of speech, with emphasis on the source-filter model and the Mel filterbank, are discussed. Section 1.2 presents AM-FM analysis of speech as a means for processing the speech signal in both a non-stationary and a non-linear framework. The principal philosophy and methodology behind AM-FM analysis are reviewed. Section 1.3 discusses the motivation of this thesis. Section 1.4 presents the layout of this thesis.

1.1 Limitations of conventional short-time analysis of speech

"*Much of what speech scientists believe about the mechanisms of speech production and hearing rests less on an experimental base than on a centuries-old faith in linear mathematics.*" - Teager & Teager, 1990.

To verify the validity of the source-filter theory of speech production, Teager measured the air flow at different positions inside the oral cavity. To his surprise, he found that most of the air flow was concentrated along the roof of the mouth and along the surface of the tongue. There was very little airflow at the center of the oral cavity, as depicted in Figure 1.1. Later, Teager also observed the presence of radial and axial airflows in the vocal tract. Thus, the air flow in the speech production system is not laminar, and hence the planar wave assumptions upon which the linear source-filter theory is based may be deemed arguable [1, 15, 36, 37].

To analyze the simplification achieved in the linear source-filter theory, we may look at Figures 1.2 and 1.3. Figure 1.2 shows the detailed speech production apparatus, which includes, apart from the main vocal tract, the nasal cavity, and the cavities of the *hypopharynx* and the *piriform fossa*. This complex structure is modeled by the source-filter theory into a far simpler structure - a concatenation of multiple tubes with different cross-sectional areas, as shown in Figure 1.3. This drastic simplification allows the vocal tract to be modeled as a linear filter, comprising of a cascade of resonators only. The

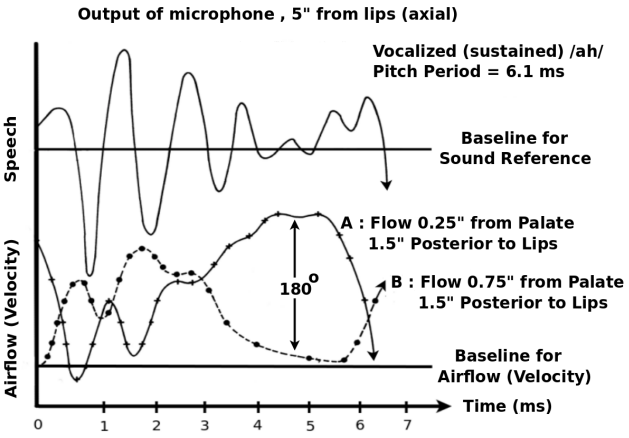


Figure 1.1: Figure redrawn from [1]. Three time traces for a vocalized vowel 'ah' produced by a male speaker. The traces provide experimental verification of separated flow. The topmost waveform is that of the speech signal recorded by a microphone placed 5" from the lips. The two waveforms at the bottom, A (solid) and B (dashed), represent airflows at two different positions inside the mouth, measured simultaneously with the recorded speech. The airflow inside the mouth is measured by a 0.7mm x 0.0005cm hot wire sensor, at a temperature of 200° C, with the wire kept normal to the flow.
< 1 > A represents air flow at a distance of 0.25" from the palate, and B represents air flow at a distance of 0.75" from the palate.
< 2 > A and B are 180° out of phase.
The waveforms show that most of the air flow occurs close to the palate, as represented by A.

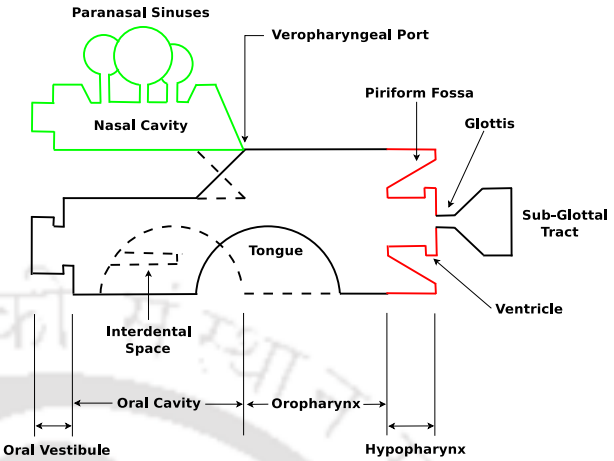


Figure 1.2: Figure redrawn from [2]. Acoustic design of the vocal tract.

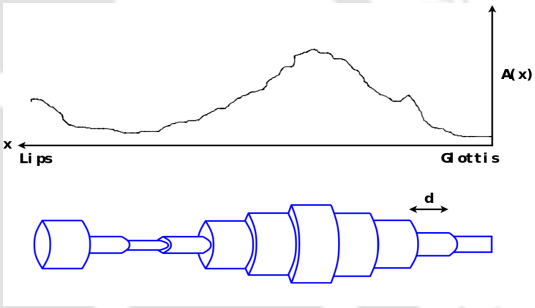


Figure 1.3: Area function of the vocal tract from the glottis to the lips (above). The simplified model of the vocal tract as a concatenation of multiple tubes with different cross-sectional areas.

vocal tract resonances are also referred to as the *speech resonances*. During the production of *voiced speech* signals (vowel-like sounds), they are popularly referred to as the *formants* [2, 13, 14].

A further simplification in speech analysis is obtained by considering this simplified model as being invariant for short segments of time, considering that the movements in the human vocal tract system are limited to such rates. This allows speech to be considered a quasi-stationary signal, whose short-time segments are the output of a Linear Time Invariant (LTI) system [2, 13, 14]. Thus, a *short segment* of a discrete-time digital *voiced speech* signal, $s(n)$, may be considered to be the output of a periodic excitation pulse, $e(n)$, which is low-pass filtered by the glottis, $g(n)$, processed by a cascade

of resonators in the vocal tract, $v(n)$, and finally radiated out of the lips by high-pass filtering, $r(n)$.

$$s(n) = e(n) * g(n) * v(n) * r(n) \;\longleftrightarrow\; S(z) = E(z)G(z)V(z)R(z)\ , \tag{1.1}$$

$$e(n) = \sum_{k=0}^{K_i-1} \delta(n - k\frac{F_s}{F_0})\ , \tag{1.2}$$

$$V(z) = \frac{1}{\prod_{k=1}^{P}(1 - p_k z^{-1})(1 - p_k^{\star} z^{-1})} = \frac{1}{\sum_{k=1}^{2P} 1 + q_k z^{-k}}\ , \tag{1.3}$$

$$G(z) = \frac{1}{(1 - bz^{-1})^2}\ ,\ b \in \mathbb{R},\ b \lesssim 1\ ,\ R(z) = 1 - cz^{-1}\ ,\ c \in \mathbb{R},\ c \approx 1 \tag{1.4}$$

In equation (1.2), K_i represents the total number of impulses corresponding to the speech segment $s(n)$. F_s represents the *sampling frequency* of $s(n)$, and F_0 the *pitch* or *fundamental frequency* of $s(n)$, also that of the excitation signal, $e(n)$. In equation (1.3), p_k and $p_k^{\star}$ represent a pair of complex conjugate poles, the number of such pairs being denoted by P. q_k represents the k^{th} coefficient of the simplified transfer function in standard polynomial form. For *unvoiced speech* (consonant-like sounds), the glottis does not play any part (all-pass filter), whereas the source excitation is modelled as a random noise sequence instead of a train of impulses [2, 13, 14]. While these simplifications make analysis easier, it almost certainly limits the capability of capturing the information embedded in the *dynamics* or the *fine structure* of speech.

From the speech perception point of view, the Mel filterbank is used to imitate the characteristics of the human ear. The MFCCs, which are widely used in speech processing applications, however, do not incorporate the phase spectrum of speech. The inability to accommodate the phase spectrum of speech in the MFCCs has led to limitations in the performance of many speech processing applications [15]. Further, one may argue that while the human ear does multi-tasking, it also has an unparalleled computer with mysterious capabilities at its disposal - the brain. Henceforth, for machines to replicate the performance of the human cognitive system it may be more beneficial to construct application-oriented filterbanks, instead of using the Mel filterbank for all purposes. These limitations of conventional speech production and perception modeling and analysis are discussed below.

1.1.1 Limitations of Fourier Analysis

There are two basic requirements of the data for Fourier-analysis to make sense [18].

(a) The data must be stationary.

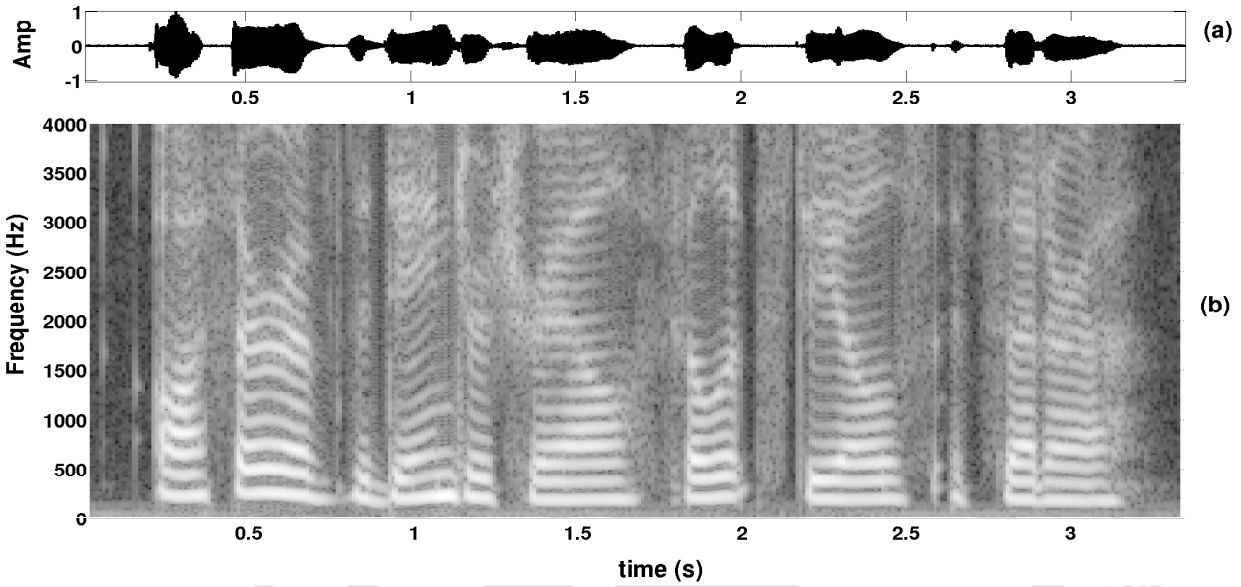


Figure 1.4: (a) Speech ; (b) Spectrogram of (a). A framesize of 25 ms, with a frameshift of 10 ms, is used.

(b) The data must be generated by a linear system.

However, real-world data, like speech, seldom meet these requirements. As such, Fourier analysis requires many additional harmonics to simulate non-uniformity (abrupt changes) in the data. *It spreads the energy over a wide frequency range.* A simple example is the delta function, which produces a flat Fourier spectrum.

$$\delta(t) \longleftrightarrow 1, -\infty < f < \infty$$

As such, for non-stationary signals, Fourier analysis produces a multitude of components which combine perfectly mathematically but may not represent meaningful characteristics of the signal [18]. From a different perspective, Fourier analysis cannot track the change in the frequency content of the signal with time, as all its components are spread over the entire time-scale. As a way of countering this particular limitation, the *Short-Time Fourier Transform* (STFT) has been the utilized, particularly for speech signal processing, wherein Fourier analysis is done for *short, fixed segments* of the speech signal [2, 13, 14]. Given a continuous-time speech signal, $s(t)$, its continuous-time STFT is given by,

$$S(\tau, f) = \int_{-\infty}^{\infty} s(t)w(t-\tau)\ e^{-j2\pi ft}dt\ , \tag{1.5}$$

where $w(t)$ represents a symmetric window of finite time-width, and τ the time-instant at which the window is placed. Thus, depending on the width of $w(t)$, STFT provides a fixed time and frequency resolution of the signal. Figure 1.4 represents the time-varying STFT magnitude spectrum of a speech utterance (signal) in the form of an image, popularly known as the *Spectrogram*, where the STFT spectrum is evaluated at every 10 ms, considering a time-window of 25 ms. Clearly, STFT is not an adaptive signal analysis method [38]. There is no "correct" window size, and it varies not only with the task at hand but even within a particular signal. Also, even if a signal-segment is truly stationary, and is constituted of a finite number of sinusoids, Fourier analysis would reveal the true spectrum (only the constituents) only if the signal-segment was of infinite duration. Lesser the duration of the segment, greater is its perceived non-uniform (abruptly changing) and non-stationary characteristics, and wider the STFT spectrum.

1.1.2 Limitations of LP analysis

LP analysis has been the cornerstone for speech analysis based on the source-filter theory. LP analysis is used to estimate the vocal tract resonances (popularly called *formants* in the case of voiced speech) and the excitation source of the speech signal [2, 13, 14]. However, in accordance with the source-filter theory, LP analysis does not model the anti-resonances of the speech production system. Also, it does not model the resonances in the cavities of the hypopharynx and the piriform fossa, which influence the overall spectrum of speech. As a result, LP analysis is prone to inaccurate estimation of the speech resonances, and the excitation signal represented by the LP residual [21]. These inaccuracies are best visualized in the LP residual, which turns out to be a highly fluctuating noise-like signal, instead of the ideal train of impulses, as shown in Figure 1.5.

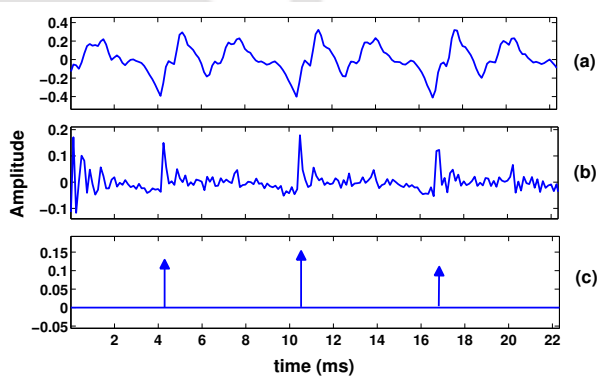


Figure 1.5: **(a) A voiced speech signal ; (b) its LP residual ; (c) the ideal excitation signal.**

Let us consider a segment of a discrete-time digital voiced speech signal, $s(n)$, having transfer

function $S(z)$. Using equations (1.1)-(1.4), and considering $b = c = 1$, we get,

$$s(n) \longleftrightarrow S(z) = E(z)G(z)V(z)R(z) = E(z)\frac{1}{1-z^{-1}}\frac{1}{\sum_{k=1}^{2P} 1 + q_k z^{-k}} ,$$

$$s'(n) = s(n) - s(n-1) \longleftrightarrow S'(z) = S(z)(1-z^{-1}) = E(z)\frac{1}{\sum_{k=1}^{2P} 1 + q_k z^{-k}} = E(z)V(z) ,$$

where $S'(z)$ represents the transfer function of the pre-emphasized (high-pass filtered) speech signal, $s'(n)$. It is evident that any inaccuracies in the estimation of the LP coefficients, q_ks, would result in inaccurate estimates of the formants, and an overall inaccurate representation of the spectrum of speech. Let us assume, hypothetically, that the vocal tract response, $V(z)$, consists of a single resonator only, and let a_1 and a_2 be its corresponding coefficients.

$$V(z) = \frac{1}{1 + q_1 z^{-1} + q_2 z^{-2}}$$

Let $\hat{q}_1$ and $\hat{q}_2$ be the corresponding coefficients of $V(z)$ as estimated from LP analysis. Let $\hat{q}_1 = q_1 + \tilde{q}_1$ and $\hat{q}_2 = q_2 + \tilde{q}_2$, where $\tilde{q}_1$ and $\tilde{q}_2$ represent the inaccuracies of estimation in LP analysis.

$$\hat{V}(z) = \frac{1}{1 + \hat{q}_1 z^{-1} + \hat{q}_2 z^{-2}}$$

The LP residual is then estimated as

$$\hat{E}(z) = S'(z)\hat{V}(z)^{-1} = S'(z)\Big\{V(z)^{-1} + \tilde{q}_1 z^{-1} + \tilde{q}_2 z^{-2}\Big\} ,$$

$$\hat{E}(z) = E(z) + S'(z)\tilde{q}_1 z^{-1} + S'(z)\tilde{q}_2 z^{-2} ,$$

$$\hat{e}(n) = e(n) + \tilde{q}_1 s'(n-1) + \tilde{q}_2 s'(n-2)$$

The last two terms of $\hat{e}(n)$ causes the LP residual to represent a noisy estimate of the actual excitation signal. Let us, now, consider the true vocal tract system to be composed of a pair of conjugate zeros (representing an anti-resonance) along with the pair of poles. Let us also assume that the inaccuracies in estimating the poles is negligible, i.e., $\tilde{q}_1 \approx \tilde{q}_2 \approx 0$.

$$V(z) = \frac{1 + b_1 z^{-1} + b_2 z^{-2}}{1 + q_1 z^{-1} + q_2 z^{-2}}$$

In this case, the estimated LP residual takes the form

$$\hat{E}(z) = S'(z)\hat{V}(z)^{-1} = E(z)\Big\{1 + b_1 z^{-1} + b_2 z^{-2}\Big\} ,$$

$$\hat{e}(n) = e(n) + b_1 e(n-1) + b_2 e(n-2)$$

If $b_2 > 1$, i.e, the zeros are outside the unit circle, the large values in the LP residual may not occur at the actual excitation instants, but after two samples. Thus, the unmodeled anti-resonances could severly affect the accuracy of the LP residual in estimating the excitation source. Apart from this, the inaccurate estimation of the phase angles of the formants is found to cause bipolar swings near the excitation instants, which make their identification difficult [21].

1.1.3 Importance of phase in speech perception

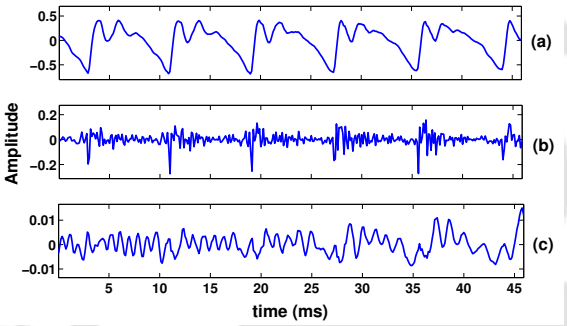


Figure 1.6: (a) Speech ; (b) Phase-only reconstruction of speech ; (c) Magnitude-only reconstruction of speech. Rectangular window (framesize) of 1024 ms duration, with 75% overlap between frames is used.

Table 1.1: Consonant intelligibility of Phase-only and Magnitude-only reconstructed speech for different analysis window sizes, and different window types. Values in the table are quoted from [12].

Window	Intelligibility (%)			
	Magnitude - only		Phase - only	
	32 ms	1024 ms	32 ms	1024 ms
Hamming	84.2	14.1	59.8	88.0
Rectangular	78.1	13.2	80.0	89.3

The understanding of phase does not come as intuitively as that of energy or amplitude. This is probably the reason why the magnitude spectrum is mostly involved in analysis, whereas the phase spectrum remains neglected. Even the MFCCs do not incorporate the phase spectrum of speech. However, the phase spectrum obtained from the STFT of speech has been found to be particularly important in speech perception [12, 25, 26]. Phase-only reconstructed speech, i.e., speech signal reconstructed using only its phase spectrum while keeping the magnitude spectrum fixed to unity, is found to be highly intelligible, particularly when rectangular windows are used for analysis. Compared to this, magnitude-only reconstructed speech, i.e., speech signal reconstructed using only its magnitude spectrum while keeping the phase spectrum fixed to zero, is found to be less intelligible. The intelligibility of magnitude-only reconstructed speech is also limited to shorter analysis windows. Table 1.1 lists the consonant intelligibility averaged over 12 listeners, for 16 consonants spoken by Australian English speakers in *vowel-consonant-vowel* context [12, 25]. The aforementioned observations are evidenced in the table. Again, the phase-only reconstructed speech, as shown in Figure 1.6, is also observed to carry information about the *epochal events* or the *glottal closure instants* [5–11, 21, 39] in voiced speech [12, 25, 26]. This is particularly evidenced for large

analysis windows.

1.1.4 Inadequacies of the Mel filterbank and the source-filter theory

To validate the utility of the MFCCs in characterizing speaker information vs. speech information, over the broader speech spectrum (0-8 kHz), the *Fisher's F-ratio* [3], which is a ratio of the inter-speaker variance to the intra-speaker variance, is computed for utterances of different speakers. For this experiment, 60 triangular filters, which are uniformly spaced in the linear frequency scale, are used [3]. Every speech frame, of a given utterance, is subjected to the filterbank, to obtain 60 sub-band energies for the frame. Let $x^i_{k,j}$ be the sub-band energy of the k^{th} frequency band of the j^{th} speech frame belonging to the i^{th} speaker. The average sub-band energy of the k^{th} frequency band for the i^{th} speaker is given by,

$$u^i_k = \frac{1}{N^i} \sum_{j=1}^{N^i} x^i_{k,j} \ , \ k = 1, ..., 60 \ ,$$

where N^i is the total number of speech frames for all the utterances belonging to the i^{th} speaker. Then, the average sub-band energy of the k^{th} frequency band for all the speakers is given by,

$$u_k = \frac{1}{M^s} \sum_{i=1}^{M} u^i_k \ ,$$

where M^s is the total number of speakers. The F-ratio for the k^{th} frequency band is then given by,

$$F^{ratio}_k = \frac{\frac{1}{M^s} \sum_{i=1}^{M^s} (u^i_k - u_k)^2}{\frac{1}{M^s N^i} \sum_{i=1}^{M^s} \sum_{j=1}^{N^i} (x^i_{k,j} - u^i_k)^2} \ , \ k = 1, ..., 60 \tag{1.6}$$

As can be seen from equation (1.6), the numerator of F^{ratio}_k gives the variation of energy in the k^{th} frequency band for different speakers. The denominator gives the variation of energy in the k^{th} frequency band for the same speaker. Thus, a high value of F-ratio for a given frequency band indicates the presence of high speaker-specific information in the band. Contrarily, a low value of F-ratio indicates the presence of speech-specific information in the band. Figure 1.7 shows the F-ratio (in dB) computed for different sessions of the NTT-VR corpus [40], and for different types of speech of the CHAINS corpus [41]. As can be seen from the figure, for the NTT-VR corpus the F-ratio is high roughly in three regions - below 500 Hz, around 5000 Hz, and around 7000 Hz. These are the regions which carry most of the speaker-specific information. The rest of the regions, particularly between 1 - 4 kHz, have low F-ratio values and carry the message information of speech. In the case of the

CHAINS corpus, again, it may be observed that the F-ratio values starts to increase after their lowest point (between 3-4 kHz), as the frequency increases. This is found to be true irrespective of the type of articulation or speaking style.

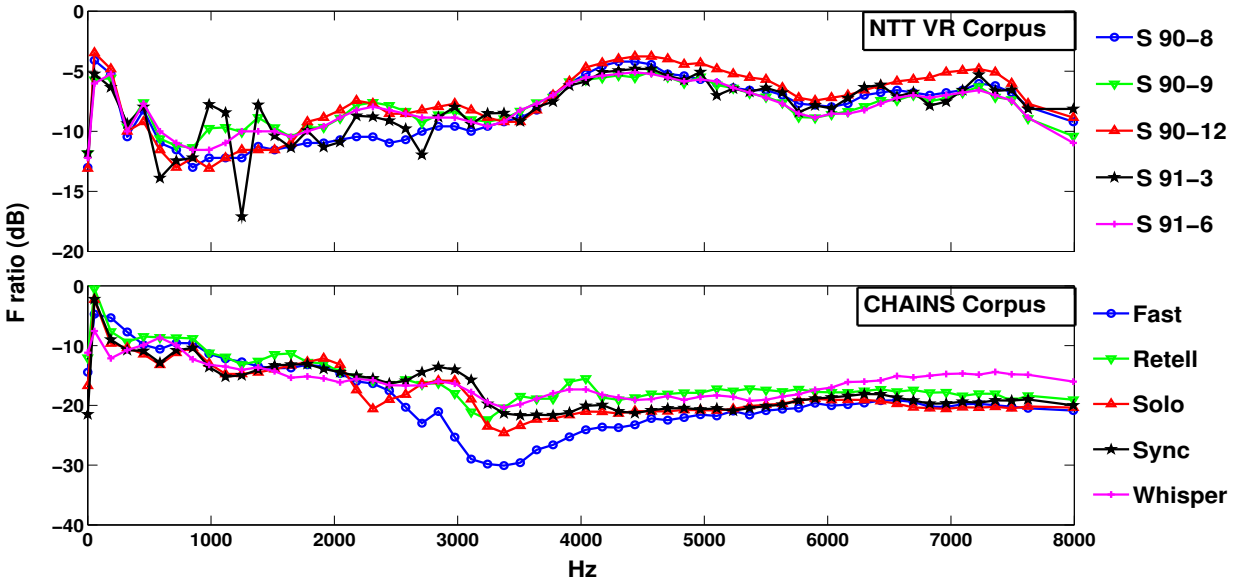


Figure 1.7: F-ratio for different frequency bands of speech, evaluated on the NTT (redrawn from [3]) and CHAINS databases.

The three frequency bands of the F-ratio curves of the NTT-VR corpus, carrying speaker-specific information, may be attributed to three different aspects of the speech production system. The high values of F-ratio below 500 Hz signify the fundamental frequency variation or the variability of the glottal source amongst speakers. Similarly, the high values of F-ratio around 7 kHz might be attributed to the vocal tract constrictions in the production of unvoiced speech. The high speaker discrimination information between 4 - 6 kHz, however, is believed to be contributed by the resonances and anti-resonances of the hypopharynx and the piriform fossa - the structures which are not included in the source-filter theory [3, 42–44]. Henceforth, attempts have been made to incorporate the velocity-to-velocity transfer functions of the hypopharynx and the piriform fossa in the source-filter theory based speech production model [42–44]. These attempts have revealed that these structures in the lower vocal tract significantly change the spectrum of voiced speech above 3.5 kHz, producing stable formants in the higher frequency spectrum of speech [42–44].

The above experiments suggest that the conventional Mel filterbank (Figure 1.8), which progressively de-emphasizes higher frequencies, may not be optimal for speaker recognition purposes.

Based on these observations, alternate avenues for speaker recognition are being explored, particularly those emphasizing the higher frequency spectrum of speech [3, 15, 27, 28, 45–49]. Some of these experiments use different filterbanks, as opposed to the conventional Mel filterbank [3, 27, 28], while others use an AM-FM representation of speech [45, 47–49], which is discussed in the following section. Even for speech recognition, the high resolution of the Mel filterbank at low frequencies might affect the machine recognition of speech. Also, the peaks and valleys within 1-4 kHz indicate that the Mel Filterbank may not represent the optimal filterbank for speech recognition.

Stretching the above discussion, one may consider that while the MFCCs remain the most widely used features for most speech processing applications (e.g. speech recognition [50], speaker verification [51], emotion recognition [52–55], language recognition [56]), and even for non-speech acoustic signal processing tasks (e.g. music information retrieval [57]), it is quite ambitious to assume that they would provide the best possible performance for all applications. In fact, with this viewpoint, many alternatives to the Mel filterbank have been recently introduced, allowing to improve the performances in different tasks. Most of these alternatives consist of modifications to the classical filterbank [58–64]. To improve the feature extraction process, a common strategy consists of designing filterbanks using data-driven optimization procedures [65–68]. In this direction, different methodologies based on non-stationary data analysis techniques like *Evolutionary Algorithms* (EAs) [29, 30, 69–71] and *Wavelet Transform* (WT) [16, 17, 38, 72–76] have been utilized in different speech processing applications.

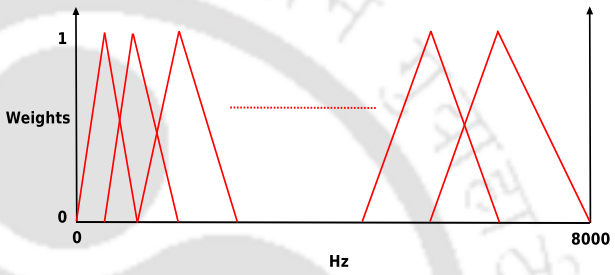


Figure 1.8: **Mel filterbank in the linear frequency scale.**

However, while these techniques optimize the features for a particular application, the optimization is not adaptive to individual speech utterances. The idea is to find out the frequency bands which are important for a particular application. From this viewpoint, they try to tackle the problems posed by the non-stationarity of the speech signal, in an average sense. Further, they do not take into consideration the inherent *non-linearity* of the speech production system or try to capture information embedded in the non-linear characteristics of speech. These techniques process the speech signal in a "linear" framework [16, 17, 38]. Henceforth, they are bound to exhibit limitations in analyzing non-linear signals [18]. Henceforth, a representation of the speech signal is desired which takes into

consideration not only its non-stationarity but also its non-linear dynamics. In this direction, we must head towards AM-FM analysis of speech.

1.2 AM-FM analysis of speech

As an attempt to overcome some of the limitations of traditional STFT analysis of speech, the sinusoidal representation of speech was proposed [4]. This model represented the glottal excitation source, and hence the speech signal, as being constituted of a finite number of sinusoids, as shown in the block diagram of Figure 1.9. The frequencies, amplitudes and phase angles of the sinusoids, for each speech frame, are derived from the peaks of its STFT spectrum. As such, this representation tries to reduce the redundancies of the STFT spectrum but does not really tackle the problems of non-linearity and non-stationarity. Its principal demerit, however, is that it requires the evaluation of a number of parameters. Also, as the process involves *peak-picking* of the STFT spectrum, it is, in essence, a miniature version of the STFT representation of speech [4]. Because of this reason, the sinusoidal representation of speech is not a complete decomposition, i.e., the components derived from it cannot synthesize exactly the same speech signal from which they have been derived.

Even though the sinusoidal model did not address the inherent non-linearity of the speech production mechanism, it aroused the possibility that the representation of the speech signal by *a small finite number of meaningful sinusoids* could be an effective mechanism for speech analysis. The next question was how to extract sinusoid-like waveforms from the speech signal without using linear and stationary analysis, i.e., the Fourier spectrum. This leads us to the next development in speech analysis - the *Amplitude Modulation - Frequency Modulation* (AM-FM) representation of speech. The AM-FM representation aims to represent the speech signal as the sum of a finite number of narrowband signals, with slowly varying amplitudes and frequencies. Thus, each component of the speech signal, under this representation, is an AM-FM signal, and not a sinusoid, with limited degrees of amplitude and frequency modulation. Ideally, one would want such AM-FM components to be centered around the resonances or the centers

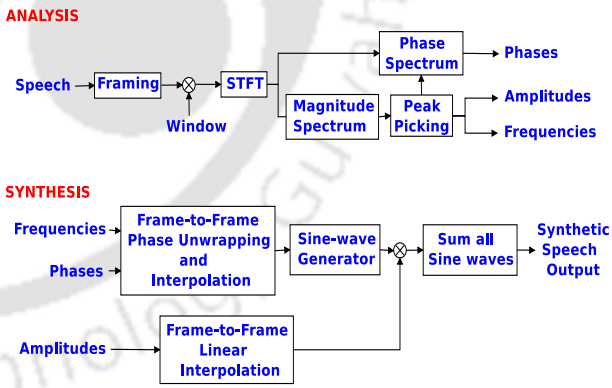


Figure 1.9: **Figure redrawn from [4]. Analysis and Synthesis process of the sinusoidal model of speech.**

of energy of the speech signal [15, 32, 34, 35]. Thus, under AM-FM analysis, a continuous-time speech signal, $s(t)$, may be ideally represented as,

$$s(t) = \sum_{k=1}^{N_f} R_k(t) \ , \ R_k(t) = a_k^R(t) \cos\left[2\pi\{f_k^R t + \int_0^t q_k^R(\tau)d\tau\} + \theta\right] \ , \tag{1.7}$$

where $R_k(t)$ represents an AM-FM signal having a center frequency, f_k^R, corresponding to a vocal tract resonator, and N_f represents the number of resonators of the speech signal. The amplitude and frequency modulating signals of $R_k(t)$ are given by $a_k^R(t)$, and $q_k^R(t)$, respectively, and θ is a constant phase. Henceforth, in order to realize the AM-FM representation, a demodulation technique is required which could estimate the instantaneous amplitude envelope and instantaneous frequency of each AM-FM component of the speech signal. One of the popular techniques for this purpose is the Hilbert Transform [15–18]. The Hilbert Transform is a reliable estimator of the frequency and amplitude envelope functions of a *monocomponent* signal, provided certain conditions are met [15–18]. These conditions are :

(i) : The frequency variation should not be large, i.e., the signal should be narrowband.

(ii) : The amplitude variation should not be large.

(iii) : The rate of frequency and amplitude variation should not be large.

Assuming these conditions are satisfied, the Hilbert Transform, $H[x(t)]$, of a signal, $x(t)$, is derived from its Fourier Transform as ,

$$x(t) \ \longleftrightarrow \ X(f) \ , \ \frac{1}{\pi t} \ \longleftrightarrow \ -j \ sgn(f) = \begin{cases} -j \ , \ f > 0 \\ j \ , \ f < 0 \end{cases} ,$$

$$H[x(t)] = x(t) * \frac{1}{\pi t} \ \longleftrightarrow \ -j \ sgn(f)X(f) = \begin{cases} -jX(f) \ , \ f > 0 \\ jX(f) \ , \ f < 0 \end{cases}$$

The instantaneous frequency function, $f(t)$, and the amplitude envelope function, $a(t)$, is derived from the analytical signal, $x_a(t)$, which is devoid of any negative-frequency Fourier components.

$$x_a(t) = x(t) + jH[x(t)] = a(t)e^{j\phi(t)} \ \longleftrightarrow \ X_a(f) = \begin{cases} 2X(f) \ , \ f > 0 \\ 0 \ , \ f < 0 \end{cases} , \tag{1.8}$$

$$a(t) = |x_a(t)| \ , \ \phi(t) = \arctan \frac{\Im\{x_a(t)\}}{\Re\{x_a(t)\}} \ , \ f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{1.9}$$

Correspondingly, the *Discrete Fourier Transform* (DFT) [2, 13, 14] is utilized for estimating the

instantaneous frequency and amplitude envelope of any discrete-time digital signal, $x(n)$.

1.2.1 The Teager Energy Operator and the proof of non-linearity in speech

Even though the Hilbert Transform can track frequency and amplitude variations in a signal, it is based on the Fourier Transform, and hence some of the limitations of Fourier analysis are also associated with it. This led to the development of the *Teager Energy Operator* (TEO), for tracking the instantaneous frequencies and amplitude envelopes of AM-FM signals [15, 31, 33, 34]. The TEO, $\Psi[x(n)]$, of a discrete-time digital signal, $x(n)$, is an estimate of the total instantaneous energy of the process generating the signal, and is given by,

$$\Psi[x(n)] = x^2(n) + x(n-1)x(n+1) \tag{1.10}$$

The *Discrete-time Energy Separation Algorithms* (DESAs), or the *Teager-Kaiser algorithms*, are used to estimate the envelope and frequency functions of discrete-time AM-FM signals. Out of the many DESAs, the more popular DESA-1 algorithm is given by,

$$\omega(n) = \arccos\left\{1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right\}, \tag{1.11}$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{1 - \left\{1 - \frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]}\right\}^2}}, \quad where\ y(n) = x(n) - x(n-1) \tag{1.12}$$

where $\omega(n)$ and $a(n)$ are the instantaneous angular frequency and envelope functions, respectively, estimated from $x(n)$. As with the Hilbert Transform, the same requirements are equally applicable to DESAs for tracking AM-FM signals [34]. However, DESAs are much simpler and efficient algorithms than the Hilbert Transform, and are free from the limitations of Fourier analysis. DESAs laid the foundation for the independent investigation of speech signals in terms of their constituent AM-FM signals, without the assumptions of the source-filter theory, and the involvement of Fourier analysis.

To evaluate whether speech is indeed the output of a linear resonator system, a simple experiment is performed [33, 34]. An arbitrary voiced speech signal, $s(n)$, with $F_s = 8$ kHz, is considered, and its formant frequencies are evaluated by a 12-order LP analysis. The signal, $s(n)$, is then band-pass filtered by a Gabor filter around one of its formant frequencies, f_{res}, to obtain a filtered output $s_{bpf}(n)$. In our case, the second formant (estimated using LP analysis), with f_{res} =1200 Hz, is considered, and the Gabor filter bandwidth is taken as 400 Hz. Again, a synthetic voiced speech signal, $s_{syn}(n)$, is generated by exciting a single resonator vocal tract system, $v_{syn}(n)$, having resonant frequency f_{res} =

1200 Hz, with a train of impulses having a frequency (F_0) of 100 Hz.

$$s_{syn}(n) = [-\sum_{-\infty}^{\infty} \delta(n - k\frac{F_s}{F_0})] * v_{syn}(n) \ ,$$

$$V_{syn}(z) = \frac{1}{(1 - p_{syn,1}z^{-1})(1 - p^{\star}_{syn,1}z^{-1})} \ , \ p_{syn,1} = 0.98 \times e^{j2\pi f_{res}/F_s}$$

Also, an AM-FM signal, $s_{mod}(n)$, is generated with gradually varying amplitude envelope and frequency.

$$s_{mod}(n) = \begin{bmatrix} a(n)\cos\{0.2\pi(n-100) + \pi(n-100)^2/4000\} \ , \ n = 0, ..., 200 \\ a(n)\cos\{0.25\pi(n-200) + \pi(n-200)^2/4000 + \pi\} \ , \ n = 201, ..., 400 \end{bmatrix} ,$$

$$\text{where } a(n) = 1 + 0.25\cos(\pi n/100)$$

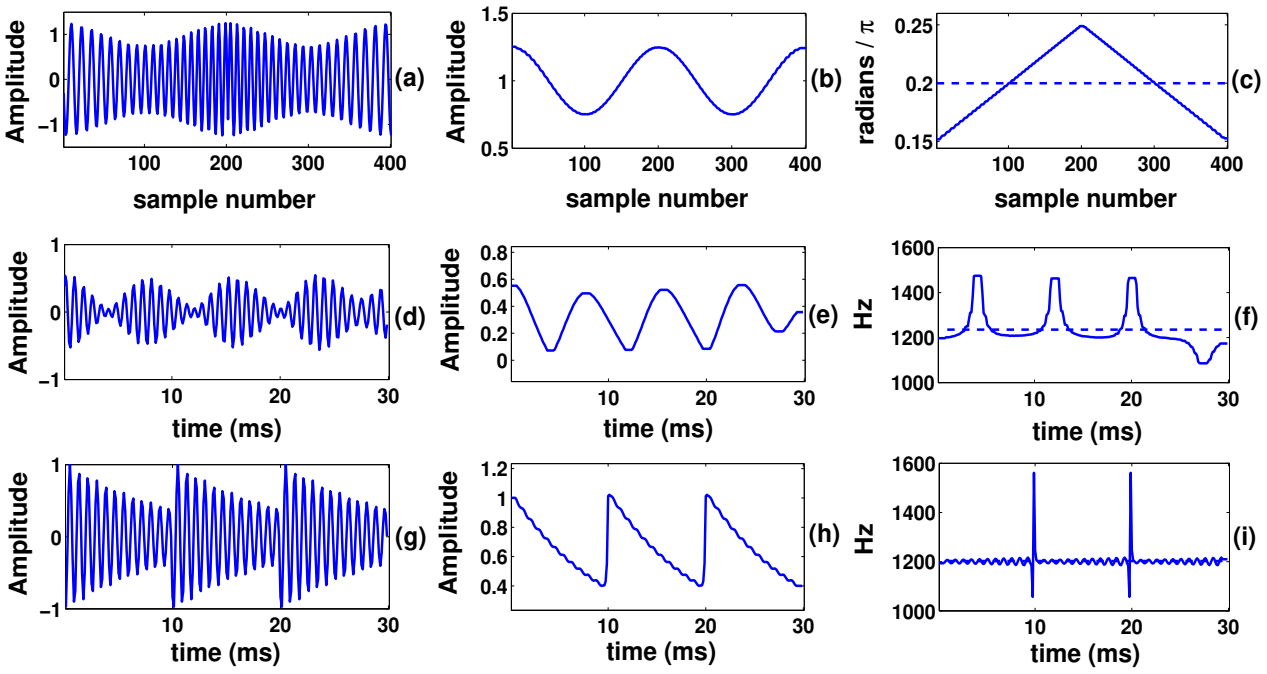


Figure 1.10: (a) $s_{mod}(n)$ (b) Estimated amplitude envelope of $s_{mod}(n)$ using DESA-1 (c) Estimated instantaneous frequency of $s_{mod}(n)$ using DESA-1. Dashed line shows average instantaneous frequency. 11-point median filter is used to smooth the estimates ; (d) $s_{bpf}(n)$ (e) Estimated amplitude envelope of $s_{bpf}(n)$ using DESA-1 (f) Estimated instantaneous frequency of $s_{bpf}(n)$ using DESA-1. Dashed line shows average instantaneous frequency. 11-point median filter is used to smooth the estimates ; (g) $s_{syn}(n)$ (h) Estimated amplitude envelope of $s_{syn}(n)$ using DESA-1 (i) Estimated instantaneous frequency of $s_{syn}(n)$ using DESA-1. Dashed line shows average instantaneous frequency.

The amplitude envelope and frequency functions of $s_{mod}(n)$, $s_{bpf}(n)$, and $s_{syn}(n)$ are then estimated using DESA-1. Figure 1.10 shows the plots of the estimates. As $s_{syn}(n)$ is generated from an LTI

system, the frequency estimate of $s_{syn}(n)$ is almost constant within an excitation period. Jumps in the frequency function indicate the impulse locations. Similarly, the amplitude envelope is an exponentially decaying function within a pitch period, and jumps occur at excitation instants. In contrast to this, the amplitude envelope of $s_{bpf}(n)$ is a more like a sinusoid, even within a glottal cycle. The frequency function also increases and decreases with time within a pitch period. These characteristics are similar to the estimates obtained from the AM-FM signal, $s_{mod}(n)$. These observations suggest that even within a glottal cycle, the speech signal may not be considered as the output of an LTI system, but rather as a combination of AM-FM signals. Such AM-FM components are mainly contributed by the resonances of the vocal tract system, and analyzing them individually might be useful for various speech processing tasks.

1.2.2 Multiband Demodulation Analysis and the Pyknogram

Though the objective of AM-FM analysis is to obtain a representation of the speech signal as a sum of its resonances, which represent narrowband AM-FM signals, there are two obstacles in this formulation. Firstly, how to obtain the formant frequencies without short-time Fourier and LP analysis ? Secondly, how to ensure that the sum of the components adds up to be exactly the same speech signal ? To circumvent the first problem, the framework of *Multiband Demodulation Analysis* (MDA) was proposed [15,32,35]. In this framework, the speech signal is passed through a *fixed parallel bank of overlapping band-pass filters*, as shown in Figure 1.11. This ensures that even if the formant frequencies vary with time, one of the filters will pick up the speech resonances at any given instant. However, as the filters are overlapping, and not disjoint, the output components may approximate, but will never exactly add up to be the same speech signal. So, just like in sinusoidal analysis, the synthesis is approximately true, if a large number of filters with less overlap is used, but not complete.

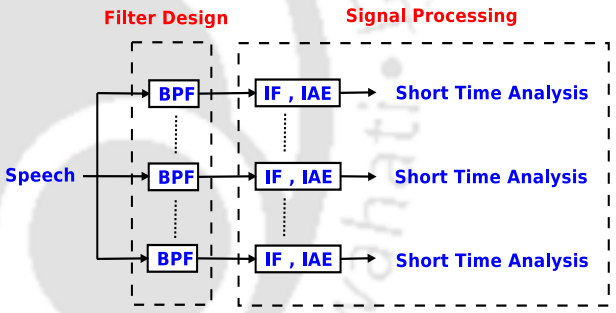


Figure 1.11: **Multiband Demodulation Analysis (MDA). BPF : Band-pass Filter , IF : Instantaneous frequency , IAE : Instantaneous Amplitude Envelope.**

As seen in Figure 1.11, there are two stages in MDA. The first stage involves the design of the filters, which may vary for different speech processing tasks. Three main questions are to be answered.

(i) : What filter to use, and the number of filters ?

(ii) : The center frequencies of the filters ?

(iii) : The bandwidths of the filters ?

Generally, a Gabor filter, $h_g(t)$, is used for filtering, as it has a low value of time-bandwidth product, and it does not produce sidelobes [32, 34, 35].

$$h_g(t) = e^{-\alpha_g^2 t^2}\cos(2\pi f_g t) \;\longleftrightarrow\; H_g(f) = \frac{\sqrt{\pi}}{2\alpha_g}\left[e^{-\frac{\pi^2(f-f_g)^2}{\alpha_g^2}} + e^{-\frac{\pi^2(f+f_g)^2}{\alpha_g^2}}\right],$$

where f_g is the center frequency of the filter, and α_g controls the bandwidth of the filter.

The second stage involves the processing of the time-domain AM-FM signals, obtained from the band-pass filterbank. The instantaneous amplitude envelope and frequency function of each AM-FM signal is estimated using DESAs or the Hilbert Transform. They are then utilized to obtain short-time estimates of mean instantaneous frequency and bandwidth. Two most used formulations are the mean amplitude weighted instantaneous frequency, F_w, and the mean amplitude weighted instantaneous bandwidth, B_w, given by,

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)a^2(t)dt}{\int_{t_0}^{t_0+T} a^2(t)dt}\;,\; B_w^2 = \frac{\int_{t_0}^{t_0+T}[\{\dot{a}(t)/2\pi\}^2 + \{f(t)-F_w\}^2 a^2(t)]dt}{\int_{t_0}^{t_0+T} a^2(t)dt}\;, \tag{1.13}$$

where t_0 and T are the starting instant and time-duration of the analysis frame, respectively, and $a(t)$ and $f(t)$ are the instantaneous amplitude envelope and frequency, respectively, of the AM-FM signal under consideration.

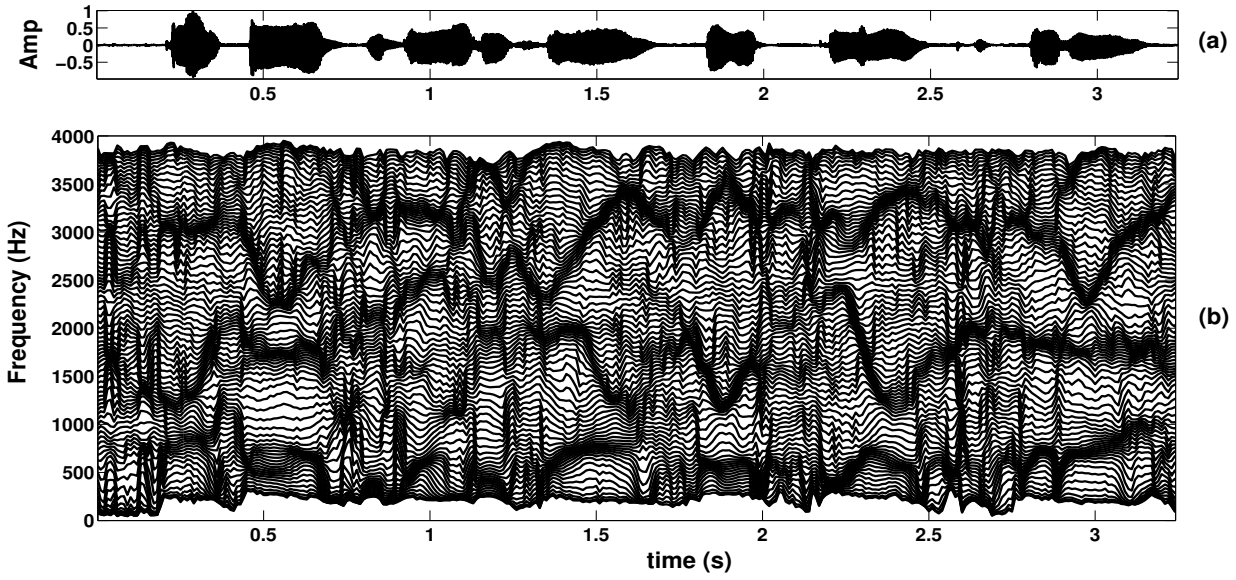


Figure 1.12: (a) Speech ; (b) Pyknogram of (a) using 80 Gabor band-pass filters of 1000 Hz effective RMS bandwidth. A framesize of 25 ms with a frameshift of 10 ms is used.

The usefulness of AM-FM analysis may be appreciated in an important time-frequency representation of the speech signal - the *speech pyknogram* [35]. Figure 1.12 shows a typical pyknogram constructed from a speech signal of $F_s = 8$ kHz. It is formed by MDA of the speech signal, using 80 Gabor filters uniformly spaced in the linear frequency scale. The Gabor filters have an effective *Root Mean Square* (RMS) bandwidth of 1000 Hz. For every speech component obtained from MDA, corresponding to a Gabor filter with analog center frequency $\{f_k^g \mid k = 1, ..., 80\}$, a short time-frequency estimate $F_w(t_{fr}, f_k^g)$ is obtained at every 10 ms time interval, t_{fr}, using equation (1.13). The time-duration of the analysis frame is taken as $T = 25$ ms. This results in the time-frequency distribution, $F_w(t, f)$, called the pyknogram. The Greek word "pykno" means dense. As can be seen from Figure 1.12, the dense clusters of curves in the pyknogram indicate the trajectories of different formants. It is a much more vivid representation than the spectrogram of Figure 1.4, for the same speech file, created using the same time resolution and analysis window size. Henceforth, the pyknogram is processed to identify regions with dense clusters for the purpose of tracking formant frequencies and their bandwidths [35].

Based on the philosophy and methodology discussed above, AM-FM analysis has been applied successfully in different speech processing tasks, particularly in the fields of speech and speaker recognition [15, 45, 47–49, 77–79].

1.3 Motivation and scope of work

The above discussion suggests how AM-FM analysis could be useful for various speech processing applications. Its objective is to capture the non-linear dynamics of the speech signal, which is generally neglected in the conventional short-time processing of the speech signal. It is a way of processing the speech signal without being subjected to the assumptions of short-time stationarity and linearity. However, no method is perfect, and hence traditional AM-FM analysis has its limitations and demerits as well. MDA provides a *fixed* analysis, determined by the design of the filterbank. As such, apart from the useful components which carry the vocal tract resonances and the glottal source information, a multitude of other components are also generated, as seen in the pyknogram. The ideal objective of AM-FM analysis, to represent the speech signal in terms of its time-varying resonances only, as encapsulated in equation (1.7), is still left desired. Again, as the filters are overlapping, the sum of the components can never synthesize exactly the speech signal - it is not a *complete decomposition*.

As such, there is a definite requirement to develop AM-FM analysis, to curb its limitations, and make it a *complete, adaptive,* and *compact* analysis.

If there were a method of performing AM-FM analysis such that it could *completely decompose* the speech signal into a finite number of time-domain components, without involving any computation of parameters, and without using short-time processing of the data, it would be more appealing to the speech community. It is desired that such a method be able to decompose the speech signal into components, whose frequency spectra are dominated by the resonant frequencies (and the fundamental frequency for voiced speech) alone. Thus, the components must be extracted adaptively and must reflect the changing nature of the resonances within the speech signal. Such a decomposition would produce less, but *meaningful*, speech components. Ideally, the frequency spectra of the components so generated should not overlap, and each component should carry information about a single vocal tract resonator or the glottal source only. *Such components then may be considered narrowband with respect to the speech signal, and therefore the piecewise stationarity and linearity assumptions might be more applicable to them. Thus, even conventional short-time analysis, based on the source-filter theory, might be more effective, provided such speech components are available.* In the pursuit of such time-domain speech components, we explore, in this thesis, the technique of *Empirical Mode Decomposition* (EMD), for decomposing the speech signal [18]. This thesis is dedicated to using EMD, and improving it, as an AM-FM analysis tool for speech processing applications. Henceforth, the next chapter of this thesis is dedicated to the study of EMD, which will provide the backdrop for the experimental work presented in the subsequent chapters.

1.4 Organization of the Thesis

In light of the issues discussed in this chapter, the rest of this thesis is organized into seven chapters. The content of each chapter is summarized below :

- Chapter 2 reviews the technique of EMD as an adaptive method of AM-FM analysis of speech. The various aspects of EMD are discussed, including its benefits and limitations. The advancements of EMD (using noise-assisted data analysis) to counter its demerits are presented. The chapter concludes by pointing out the areas in which EMD will be explored in this thesis.

- In Chapter 3, a *modified* EMD algorithm, denoted as MEMD, is proposed, which can effectively

inhibit the phenomenon of *mode mixing* manifested in EMD, but at a fraction of the time-cost incurred by the popular advanced noise-assisted versions of EMD. Detailed analysis of the AM-FM components of the speech signal, obtained from MEMD, is done, especially from the perspective of the *distribution of the formants* of voiced speech. The benefits of MEMD are illustrated not only with respect to EMD, but also in comparison with standard AM-FM analysis and WT.

- In Chapter 4, a detailed analysis is presented on the ability of the *Intrinsic Mode Functions* (IMFs), or the AM-FM components, of the speech signal, to represent its source and system characteristics. The IMFs of EMD, MEMD, and a noise-assisted EMD variant called the *Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (ICEEMDAN), are compared in their abilities to manifest the latent information in the speech signal. Experiments are firstly done on synthetic speech signals, and then on natural speech signals. Different phones or speech sounds are considered to evaluate if the characteristics of the IMFs differ with respect to the speech sounds. Experiments are also conducted to evaluate how the nature of the IMFs change if the speech waveform is modified by telephone channel codecs.

- In Chapter 5, two algorithms are proposed, which uses the IMFs of the speech signal, obtained from two advanced variants of EMD - ICEEMDAN and MEMD, for detecting the *Glottal Closure Instants* (GCIs) in the voiced regions of the speech signal/utterance. The objective of this work is to investigate if non-linear and non-stationary analysis could be used to detect the GCIs of the speech signal, as opposed to most of the state-of-the-art methods which rely on short-time LP analysis. Simultaneously, the goal is to estimate the GCIs *reliably under different conditions* - clean, noisy, and telephone channel conditions - unlike most of the state-of-the-art methods which fail to do so.

- In Chapter 6, the IMFs of the speech signal are investigated for carrying speaker-specific information. Three different features are extracted from the IMFs of the speech signal and used in the task of text-independent *Speaker Verification* (SV). The IMFs are obtained from EMD and MEMD. The objective of the work is to study whether the IMFs can complement the MFCCs in enhancing the performance of the SV system. The experiments are performed not

only for *normal* speech articulation, but also for *fast* and *whispered* speaking styles, and also for *short-length* test utterances. These broad set of experiments portrays the utility of the features in practical deployment of the SV system.

- Chapter 7 summarizes the works presented in this thesis, highlights the main contributions of the work, provides some self-criticism, and gives some directions for future research.

2

Empirical Mode Decomposition - A Review

Contents

One of the limitations of the traditional AM-FM analysis, based on MDA, is that it acts as a fixed filterbank. As such, there have been efforts to break this fixed filterbank structure, using alternative methods than MDA, to make AM-FM analysis adaptive to the speech signal [80, 81]. But, more is desired. The limitations of short-time processing of speech, apart from the fixed time-frequency resolution, lies in the inability to capture the dynamics of the speech signal [82–85]. As such, the first difference (Δ or *velocity* coefficients), and the second difference ($\Delta\Delta$ or *acceleration* coefficients), are often utilized on top of the normal speech processing features, as in the case of MFCCs [82]. To minimize this limitation, efforts have been also made to utilize dynamic frame rate and length based on the time-varying properties of the speech signal [83–85]. But, neither of these solutions which try to capture the dynamics of the speech signal, just like WT, cater to the problem of non-linearity of the speech signal. Again, the analysis becomes more complicated. Thus, there is a definite requirement of a technique for speech processing, which is more effective than the currently available tools, yet which does not complicate the analysis. The various features that such a technique is required to have may be summarized as follows :

- **(i) Complete and compact decomposition :** Fourier analysis gives a complete decomposition, i.e, the sum of the components add up to be the exact same speech signal. But, it is limited in analyzing signals which are the output of non-linear and non-stationary processes. On the other hand, AM-FM analysis relies on a large overlapping bank of filters to extract signal components. Thus, a non-linear and non-stationary signal decomposition technique is required, which is data adaptive, has little complexity, and produces time domain components which add up to be the exact same speech signal. The number of such components needs to be countably finite.

- **(ii) Unique and Meaningful Components :** The information in speech resides in the various resonant structures that shape and modulate the air flow passing through the voice production apparatus, and the glottis that controls the amount of air flowing out through the apparatus. Thus, the desired decomposition technique should be able to extract time-domain components of the speech signal, which manifests the characteristics of its resonators and glottal source. The information carried by them, ideally, should not overlap. In short, the components should carry unique and meaningful information of the speech signal.

- **(iii) No short-time processing and parameter computation :** The components should be obtained from the desired decomposition without any short-time processing of the speech signal, and

without any short-time computation of parameters, unlike in the case of sinusoidal analysis.

- **(iv) Reliable instantaneous frequency and amplitude estimates :** The components derived from the desired decomposition should be narrowband, and have limited fluctuations in amplitude and frequency, so that reliable estimates of instantaneous frequency and amplitude envelope could be obtained from either DESAs or the Hilbert Transform.

In the search of such a technique, we explore the method of *Empirical Mode Decomposition* (EMD), which is the topic of discussion of this chapter. The rest of the chapter is organized as follows : Section 2.1 discusses in detail the working mechanism of EMD, and the various characteristics of the decomposition, its advantages, and limitations. Section 2.2 presents the advancements in EMD, which cater to its limitations. Section 2.3 compares the speech components obtained from EMD with that obtained from traditional AM-FM analysis and WT. Section 2.4 presents the scope of work of this thesis.

2.1 Empirical Mode Decomposition

As discussed above, a data-adaptive and complete analysis technique is required, which can decompose the speech signal into a finite number of meaningful time domain components without the requirement of the assumptions of short-time linearity and stationarity, such that reliable instantaneous amplitude envelope and frequency estimates could be obtained from them. With this objective, we look towards the technique of *Empirical Mode Decomposition* (EMD) [18, 86, 87] for processing speech signals. EMD is a method that decomposes a signal into *oscillatory* or AM-FM components, called *Intrinsic Mode Functions* (IMFs), in a completely data-driven manner, without the requirement of any *a priori* basis. Due to this capability, EMD

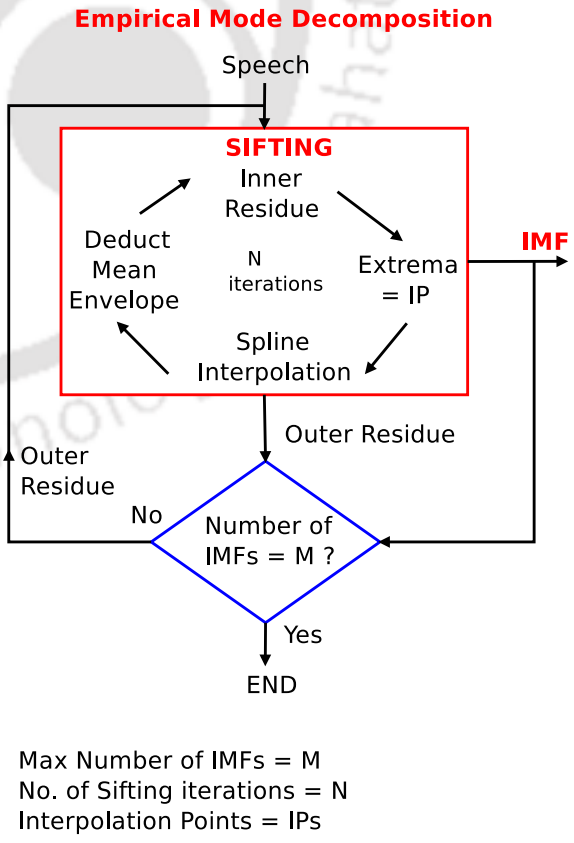


Figure 2.1: Flowchart of EMD.

has gained widespread recognition for processing real-world signals for different real-world applications [87–93, 93–103]. Figure 2.1 shows the flowchart of the EMD process. The pseudocode for the same is given below :

Pseudocode for EMD : Let $s(t)$ be a continuous-time speech signal.

•**(i)** Let $r_0(t) = s(t)$. We subject an *outer residue*, $r_{k-1}(t)$, to a *sifting process* to obtain an IMF, $h_k(t)$, and another *outer residue*, $r_k(t)$, from it. In other words, the k^{th} sifting process decomposes the $(k-1)^{th}$ outer residue, $r_{k-1}(t)$, into the k^{th} IMF, $h_k(t)$, and the k^{th} *outer residue*, $r_k(t)$.

The *sifting process* for EMD is given as :

Let $h_{k-1}^0(t) = r_{k-1}(t)$. Repeat the following steps for each *sifting iteration*. Let η represent the sifting iteration index, where $\eta = 1, 2, ..., N$.

⋆**(a)** Given the *inner residue* signal, $h_{k-1}^{\eta-1}(t)$, find the maxima and minima locations of $h_{k-1}^{\eta-1}(t)$. These locations are to be used as x-coordinates of the *Interpolation Points* (IPs), to be used for cubic spline (third order polynomial) interpolation.

$$t_{max} = \{t : \frac{d}{dt}h_{k-1}^{\eta-1}(t) = 0 \ , \ \frac{d^2}{dt^2}h_{k-1}^{\eta-1}(t) < 0\} \ ,$$
$$t_{min} = \{t : \frac{d}{dt}h_{k-1}^{\eta-1}(t) = 0 \ , \ \frac{d^2}{dt^2}h_{k-1}^{\eta-1}(t) > 0\}$$

⋆**(b)** Obtain the y-coordinates of the IPs from $h_{k-1}^{\eta-1}(t)$.

$$y_{max} = h_{k-1}^{\eta-1}(t_{max}) \ , \ y_{min} = h_{k-1}^{\eta-1}(t_{min})$$

⋆**(c)** Create the maxima envelope, $e_{max}(t)$, using cubic spline interpolation, with the IPs as $\{t_{max} \ , \ y_{max}\}$. Create the minima envelope, $e_{min}(t)$, using cubic spline interpolation, with the IPs as $\{t_{min} \ , \ y_{min}\}$. Deduce the mean envelope, $e_m(t)$, as,

$$e_m(t) = \frac{e_{max}(t) + e_{min}(t)}{2}$$

⋆**(d)** $h_{k-1}^{\eta}(t) = h_{k-1}^{\eta-1}(t) - e_m(t)$. Go to step (a). Stop when $\eta = N$.

•**(ii)** Set $h_k(t) = h_{k-1}^N(t)$. Obtain $r_k(t) = r_{k-1}(t) - h_k(t)$.

•**(iii)** Go to step (i). Ideally, the decomposition is to be stopped when the *outer residue* takes the form of a trend, i.e., the number of extrema in $r_k(t)$ is 2 or less [18, 86, 87]. Practically, however, the decomposition may be stopped when a user-defined maximum number (M) of AM-FM components,

or IMFs, has been extracted, as shown in Figure 2.1.

$$s(t) = r_M(t) + \sum_{k=1}^{M} h_k(t) \tag{2.1}$$

For a discrete-time digital speech signal, $s(n)$, appropriate discrete-time operations replace the continuous-time operations. The decomposition may be represented as,

$$s(n) = r_M(n) + \sum_{k=1}^{M} h_k(n) \tag{2.2}$$

Equations (2.1) and (2.2) represent the decomposition of the signal in terms of its IMFs and its final residue, which is a trend-like signal. One of the biggest advantages of this process is that it requires no *a priori* basis.

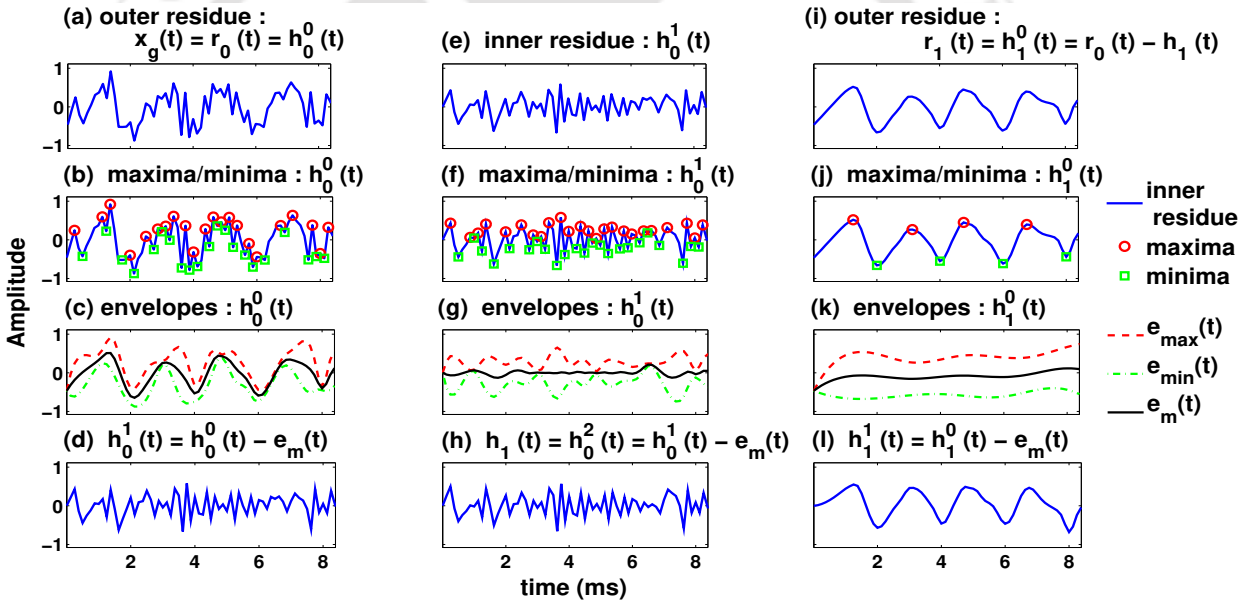


Figure 2.2: **Simulation of the EMD algorithm using a noisy sinusoidal signal,** $x_g(t)$.

To illustrate the mechanism involved, we use a signal, $x_g(t) = x_s(t) + x_n(t)$, where $x_s(t) = \cos(2\pi \times 500t)$, and $x_n(t)$ is a zero mean Gaussian White noise signal such that the *Signal to Noise Ratio* (SNR) is 0 dB. The signal, $x_g(t)$, may loosely resemble a *glottal air-volume velocity* signal that excites the vocal tract system during the production of voiced speech [2, 13, 14]. Figure 2.2 shows the working mechanism of EMD on $x_g(t)$ (instead of a speech signal). In this example, we consider $N = 2$ sifting iterations per sifting process. The first sifting process is completely illustrated, resulting in the first IMF, $h_1(t)$, after $N = 2$ sifting iterations. A new outer residue, $r_1(t)$, is, also, thereby obtained. The first iteration of the sifting process applied on the new outer residue is also shown.

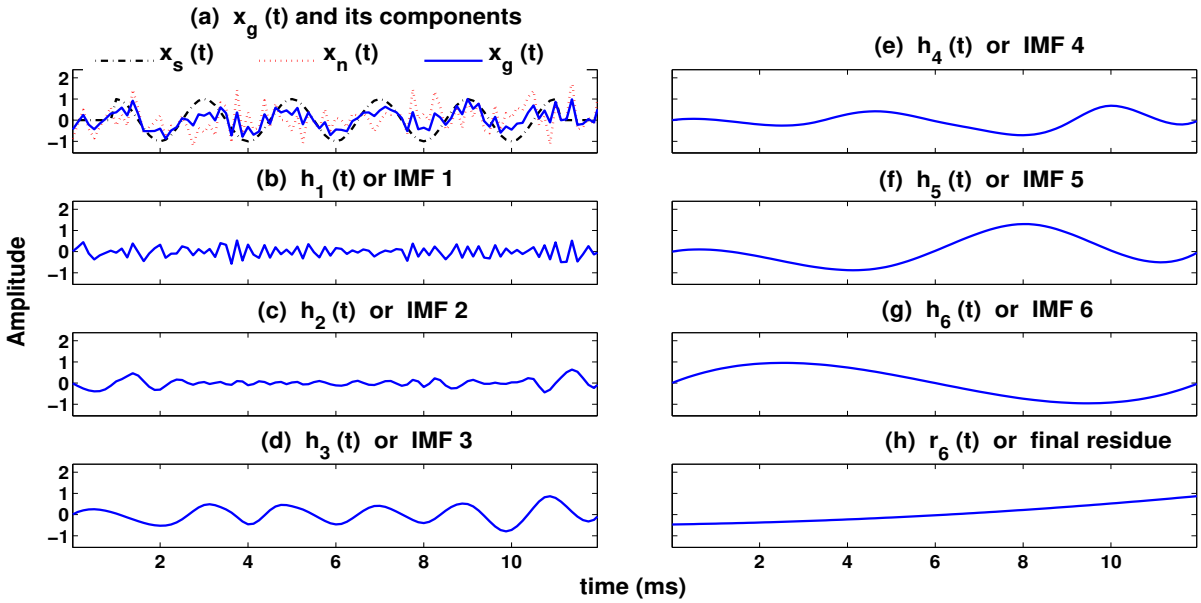


Figure 2.3: **IMFs obtained from EMD of $x_g(t)$. $N = 10$ and $M = \infty$ are considered. The decomposition naturally stops at $M = 6$.**

Figure 2.3 shows the IMFs obtained from EMD of $x_g(t)$, where $N = 10$ sifting iterations are used per sifting process. The decomposition is allowed to stop naturally by keeping no maximum limit ($M = \infty$) on the number of IMFs. Under this condition, the decomposition stops automatically when the final residue has insufficient extrema to construct the maxima and minima envelopes. For the noisy sinusoid, $x_g(t)$, the final residue is obtained after six IMFs ($M = 6$) have been extracted, as shown in the figure. One can easily observe the similarity between IMF_3 and the pure sinusoid, $x_s(t)$, and that between IMF_1 and the White noise signal, $x_n(t)$, which portrays the ability of EMD to segregate the components of a signal.

It may be noted that even though we have used continuous-time notations in the above discussion, we have implemented discrete-time digital versions of the signals. This is also true for various other examples that we would be using in this thesis. The continuous-time notations are used from time-to-time for notational convenience, and easy understanding.

2.1.1 The importance of the sifting process

As explained above, the EMD process results in a finite number of time-domain components, $h_k(t)$, $k = 1, 2, ..., M$, called the IMFs, and a final residue signal, $r_M(t)$, which is the low-frequency trend of the signal [18, 86, 87]. An IMF is defined as a signal having the following properties.

(i) The number of extrema and the number of zero-crossings in an IMF must either be equal or differ at most by one.

(ii) At any time-instant, the mean value of its two envelopes, defined by its local maxima and minima, respectively, is zero.

Thus, the aim of EMD is to obtain *oscillatory functions* from the signal. The process of *sifting* is designed to achieve this purpose. As discussed in Section 1.2, both the Hilbert Transform and TEO require the signal to be narrowband (ideally monocomponent), with limited degrees of frequency and amplitude modulation, for accurate demodulation. The above-mentioned properties of an IMF make it locally narrowband and symmetric, which enables more reliable estimation of its instantaneous frequency and amplitude envelope.

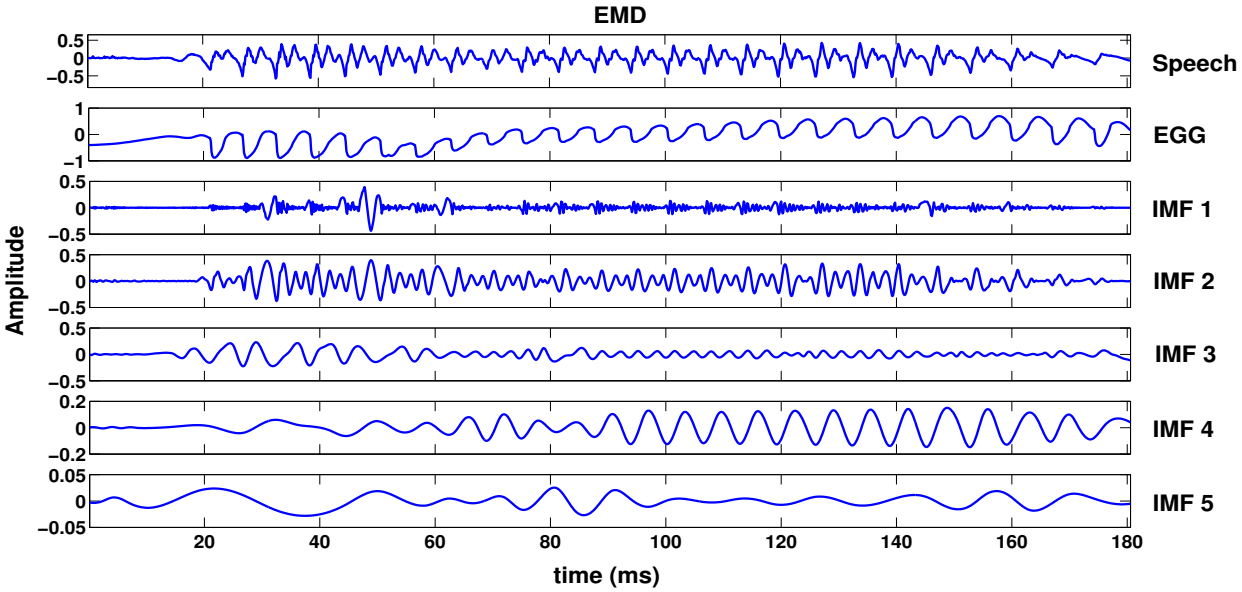


Figure 2.4: A natural speech signal, its EGG, and its first five IMFs (obtained by EMD).

Having applied EMD on a synthetic signal, we now apply it to a natural speech signal. Figure 2.4 shows the first five IMFs obtained from a natural speech signal/utterance, where $N = 10$ sifting iterations have been used per sifting process, and the decomposition is curtailed at $M = 9$. The second plot of the figure shows the *ElectroGlottoGraph* (EGG) signal [104, 105] corresponding to the speech signal. The EGG signal represents a measurement of the movements of the vocal folds during the production of voiced speech. As can be seen from the figure, there is a strong similarity between IMF_4 and the EGG signal, which reflects the ability of EMD to extract information about the glottal source producing the speech signal [106]. This shows the ability of EMD to extract latent information

from the signal, in its IMFs. Again, as mentioned earlier, no *a priori* basis function is used during the decomposition. Further, the process is carried out on the entire data stream (complete speech utterance), without segregating it into blocks/segments/frames (unlike in the case of STFT), and no parameter computations are involved.

2.1.2 Hilbert Huang Transform as a generalized Fourier Transform

Having derived the IMFs from the signal, they are represented in terms of their instantaneous amplitude envelopes and frequencies using the Hilbert Transform. This entire process of extracting IMFs from the data, and representing them in terms of their instantaneous amplitude envelopes and frequencies, is termed as *Hilbert Huang Transform* (HHT) [18, 86, 87, 107]. We have, from equation (2.1),

$$s(t) = \sum_{k=1}^{M} h_k(t) \; + \; r_M(t) = \sum_{k=1}^{M+1} h_k(t) \tag{2.3}$$

Each component, $h_k(t)$, derived from the signal, can be represented using the Hilbert Transform, using equations (1.8) and (1.9), as,

$$\begin{aligned} h_k(t) &\xrightarrow{\text{Hilbert Transform}} a_k(t)e^{j\theta_k(t)} \; , \\ h_k(t) = \Re\Big\{a_k(t)e^{j2\pi\int f_k(t)dt}\Big\} \; , \; f_k(t) &= \frac{1}{2\pi}\frac{d}{dt}\theta_k(t) \; , \end{aligned} \tag{2.4}$$

where $a_k(t)$, $\theta_k(t)$, and $f_k(t)$ represent the instantaneous amplitude envelope, phase, and frequency, respectively, of $h_k(t)$. The signal can then be represented as,

$$s(t) = \Re\Big\{ \sum_{k=1}^{M+1} a_k(t)e^{j2\pi\int f_k(t)dt} \Big\} \tag{2.5}$$

The standard Fourier representation of the same signal is given by,

$$s(t) = 2\pi \int_{-\infty}^{\infty} S(f)e^{j2\pi ft}df \tag{2.6}$$

Comparison of equations (2.5) and (2.6) shows that HHT is a generalization of the Fourier Transform, without the limitations of it. HHT represents the signal in terms of a finite number of components, unlike the Fourier Transform. While the amplitude envelope and frequency of each component in Fourier representation is constant for an infinite time duration, it is time varying in the case of HHT. HHT is a complete, compact, and adaptive Fourier representation of the signal [18, 86, 87, 107]. This formulation, when presented in terms of an image, is called the Hilbert

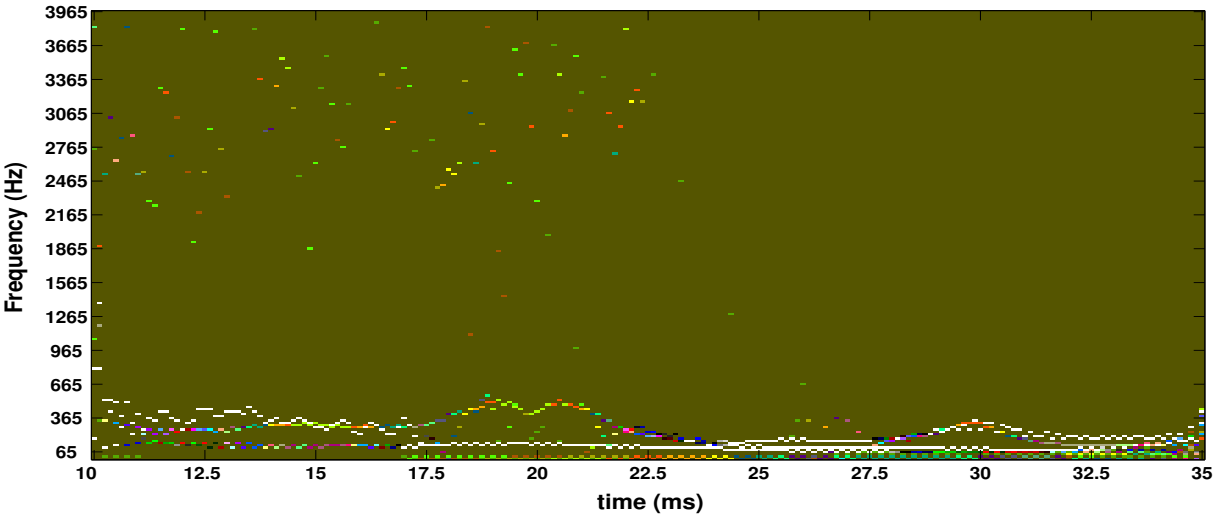


Figure 2.5: **Hilbert spectrum of the speech signal used in Fig2.4, using EMD.**

spectrum [18, 86, 87, 107]. The Hilbert spectrum can be defined as the time-frequency distribution of the instantaneous energy envelope, which is the squared magnitude of the amplitude envelope.

$$H(f,t) = \{a_k^2(t) \mid f_k(t), t\} \ , \ k = 1, ..., K \leq M \tag{2.7}$$

In general, the last few components, which are low-frequency trend-like waveforms, are excluded from the spectrum, as they have high energy and obscure the image [18, 86, 87, 107]. Figure 2.5 shows the Hilbert spectrum for a section of the speech signal used in Figure 2.4. As is evident from the spectrum, most of the energy in the spectrum lies within 60-500 Hz, which is the pitch frequency range, i.e., the frequency range of vibration of the vocal folds in the glottis (during the production of voiced speech). As such, this spectrum can be post-processed to obtain the instantaneous pitch frequency [88]. From the Hilbert spectrum, the marginal Hilbert spectrum may be derived as,

$$h(f) = \int_{t=0}^{T_{dur}} H(f,t)dt \tag{2.8}$$

The marginal Hilbert spectrum gives the probability that an oscillation of frequency f could have occurred locally at some time during the entire duration (T_{dur}) of the signal. Similarly, the instantaneous energy density can be computed from the Hilbert spectrum, which reflects the energy fluctuations in the signal with respect to time [18, 86, 87, 107].

$$IE(t) = \int_f H(f,t)df \tag{2.9}$$

2.1.3 The dyadic filterbank nature of EMD

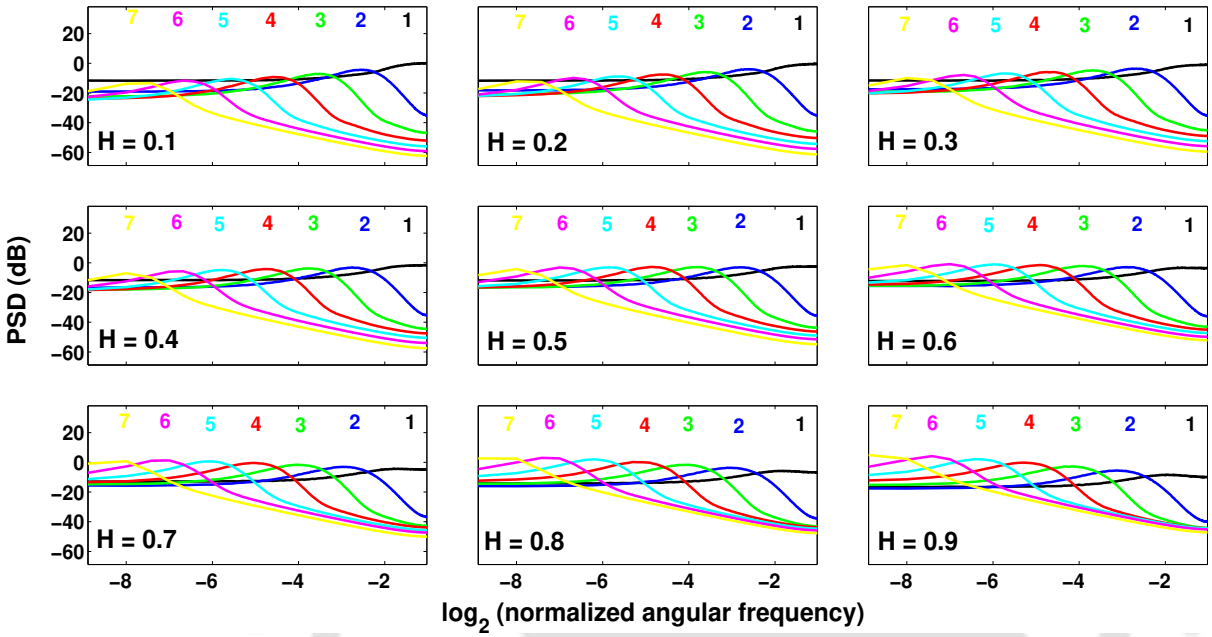


Figure 2.6: **IMF power spectra in the case of fractional Gaussian noise, for Hurst exponent H = { 0.1 , 0.2 , ... , 0.9 }. The estimated PSDs (in dB) are plotted as a function of the logarithm of the normalized frequency for the first seven IMFs. The IMF number is mentioned above the peak of the corresponding power spectrum. For each of the nine H values, the spectral estimates have been computed on the basis of 5000 independent sample paths of 512 data points.**

While EMD is an effective decomposition process, it is an algorithm without a solid mathematical framework. Even though efforts have been made to provide some mathematical representation [108], deducing conclusions about its behavior is not straightforward. To have a better understanding of the behavior of the process, studies were carried out on the decomposition of noise by EMD, by Wu and Huang, and Flandrin et al., separately [87, 107, 109–113]. For the experiments, fractional Gaussian noise (fGn) was used. The autocorrelation function of a fractional Gaussian noise sequence, $x_H(n)$, of variance σ_H^2, is given by,

$$r_H(m) = \mathbb{E}\Big[x_H(n)x_H(n+m)\Big] = \frac{\sigma_H^2}{2}\Big\{|m-1|^{2H} - 2|m|^{2H} + |m+1|^{2H}\Big\}\,, \tag{2.10}$$

where $\mathbb{E}$ represents the expectation operator, and m the lag. As is evident from equation (2.10), the parameter H, $0 < H < 1$, called the *Hurst component* [110–112], controls the nature of the signal. For $H = 0.5$, $x_H(n)$ becomes a White Gaussian noise sequence. For $0 < H < 0.5$, the power spectrum of $x_H(n)$ is of high-pass nature, whereas $0.5 < H < 1$ produces $x_H(n)$ of low-pass nature.

For the experiments, 5000 realizations of fGn were generated for each of the nine H values given by $H = 0.1, 0.2, ..., 0.9$. The fGn sequences were of 512 samples length. Each fGn sequence was then decomposed by EMD, and the properties of the first seven IMFs were then examined.

Figure 2.6 shows the plots of the *Power Spectrum Densities* or PSDs (averaged for 5000 fGn sequences) of the IMFs, for $H = \{0.1,\ 0.2,\ ...,\ 0.9\}$. The PSD of a discrete-time digital signal is given by its squared magnitude DFT spectrum, normalized with respect to its total energy. The plots show that barring the first IMF, the rest of the IMFs (IMFs 2-7), for all the H values, have power spectra having band-pass nature. The frequencies corresponding to the peaks of these band-pass spectra, approximately decrease by a factor of 2 as the IMF order (number) increases. In other words, starting from the second IMF, EMD acts as a *dyadic filterbank* on the signal [87, 107, 109–113]. The first IMF exhibits a weak high-pass spectrum for all the H values.

The IMFs being locally zero mean signals, the number of zero crossings of an IMF could be used to estimate the dominant frequency that it represents. Figure 2.7 shows the plots of the number of zero crossings of the IMF vs. the IMF number. As can be seen, the curves in Figure 2.7(a) have an approximate slope of -1, which means that the dominant frequency reflected in an IMF is half of that of its preceding IMF. The average decrease rate of the number of zero-crossings, ρ_H, for each of the H values, is plotted in Figure 2.7(b), which ascertains this observation. Thus, if $z_H(k)$ represents the number of zero-crossings of the k^{th} IMF, we have,

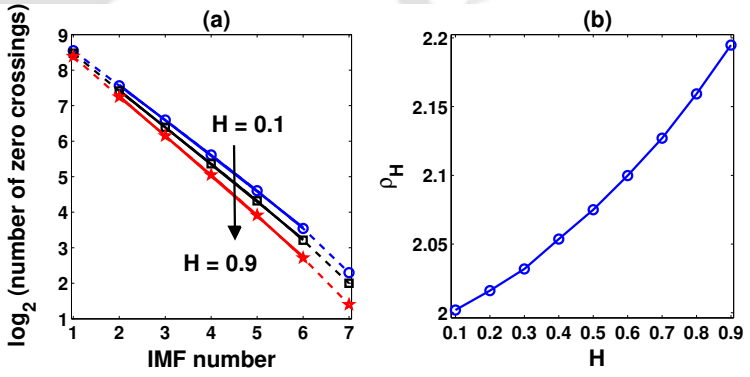


Figure 2.7: **(a) Average number of zero-crossings of the first seven IMFs of fractional Gaussian noise. For clarity, only those curves corresponding to H = 0.1 (bubbles), H = 0.5(squares) and H = 0.9 (stars) have been plotted in the diagram. The remaining cases lead to regularly intertwined similar curves. The superimposed solid lines correspond to linear fits within the IMF range k = 2 to 6 ; (b) Corresponding decrease rate of zero-crossings.**

$$z_H(k') = \rho_H^{(k'-k)} z_H(k),\ k' > k \geq 1\ , \tag{2.11}$$

$$\rho_H = 2.01 + 0.2(H - 0.5) + 0.12(H - 0.5)^2 \approx 2 \tag{2.12}$$

Given that the dominant frequency of an IMF is approximately half of its immediately preceding

IMF, the PSDs of the band-pass IMFs can then be approximately related to one another as,

$$S_{k',H}(f) = \rho_H^{\alpha_H(k'-k)} S_{k,H}(\rho_H^{[k'-k]} f) \ , \ k' > k \geq 2 \ ; \ \rho_H \approx 2 \ , \ \alpha_H = 2H - 1 \tag{2.13}$$

Figure 2.8 plots the variances of the IMFs vs. their IMF numbers, for $H = \{0.1,\ 0.5,\ 0.9\}$. As can be seen from the figure, the variances of the IMFs decrease with respect to their IMF numbers, at a rate dependent on the H value. For the case of White noise ($H = 0.5$), the slope of the curve is approximately -1, i.e., the IMF energy decreases by a factor of 2 as the IMF order increases.

$$V_H(k') = \rho_H^{(\alpha_H - 1)(k'-k)} V_H(k),\ k' > k \geq 2 \ ; \ \rho_H \approx 2 \ , \ \alpha_H = 2H - 1$$
$$V_H(k') = \rho_H^{2(H-1)(k'-k)} V_H(k),\ k' > k \geq 2 \tag{2.14}$$

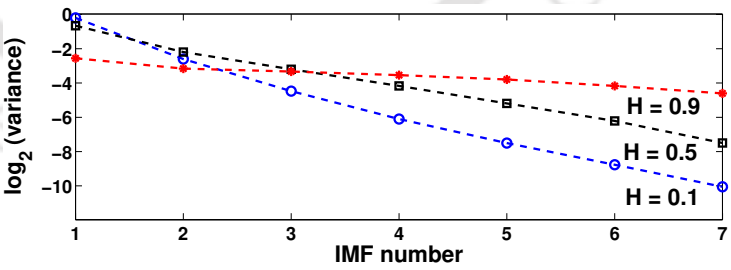
The slope (κ_H) of the log-linearized version of equation (2.14) can be used to estimate the Hurst component as,

$$V_H(k') = C \rho_H^{2(H-1)k'},\ k' \geq 3 \ ,$$
$$\log_2 V_H(k') \approx \log_2 C + 2(H-1)k' = C' + \kappa_H k',\ k' \geq 3 \ , \tag{2.15}$$
$$H_{est} = 1 + \frac{\kappa_H}{2} \tag{2.16}$$

The estimated Hurst component values, H_{est}s, obtained by this process, are shown in Figure 2.8.

2.1.4 Some aspects of EMD

Apart from the fact that EMD does not have a robust mathematical framework, there are some aspects of the decomposition that are to be catered to. One of them is the sampling rate of the signal. As is evident from the flowchart of EMD (Figure 2.1), and its pseudocode, the detection of extrema is crucial to the process. Again, the instantaneous frequency derived using the Hilbert Transform depends on differentiation with respect to time, as



H	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
H_{est}	0.23	0.28	0.34	0.41	0.5	0.58	0.68	0.77	0.87

Figure 2.8: **Estimated $\log_2(variance)$ of each of the first seven IMFs, in the case of fractional Gaussian noise, for $H = \{0.1, 0.5, 0.9\}$. The values of the empirical (energy-based) variance estimates are mentioned at the bottom for all the nine H values.**

is evident from equation (1.9). For these reasons, it is beneficial for the decomposition if the signal is sampled at much above the Nyquist rate [18, 86, 87, 114]. Another issue that needs to be catered to is *"end-effects"*, which are large swings that occur at the ends of the IMFs due to cubic spline fitting. By zero-padding the ends of the signal, however, *"end-effects"* could be curtailed to a certain extent [18, 86, 87]. More details on *"end-effects"*, and how to curtail it, could be found in [115]. As a speech signal (utterance) has silence regions (pauses) at the beginning and end of the signal, *"end-effects"* is not too concerning. Apart from this, the *sifting criterion* is another aspect that needs attention. If the number of sifting iterations is less, the decomposed signals would not be eligible to be IMFs, which would cause erroneous estimates of their instantaneous frequencies and amplitude envelopes. On the other hand, over-sifting would result in the smoothing of the IMFs, and thus they would become more like sinusoids, and the decomposition may converge towards Fourier analysis. A number of *sifting criteria* have been proposed to ascertain that the IMFs adhere to their defined properties [18, 86, 87, 116–118]. All of them need some parametric tuning, and none of them may be deemed significantly better than the others. Amongst them, one recent and popular criterion is the *local-global stopping criterion*, proposed in [116]. This criterion is based on minimizing the parameter $\gamma(t) = \left|\frac{[e_{max}(t)+e_{min}(t)]/2}{[e_{max}(t)-e_{min}(t)]/2}\right|$, for ascertaining globally small fluctuations of the mean envelope signal even for locally large fluctuations of the signal. Two thresholds, Θ_1 and Θ_2, are used to control the number of iterations, N, in every sifting process. When $\gamma(t) < \Theta_2$ for a fraction α_F of the duration of the signal, and $\gamma(t) < \Theta_1$ for the remaining fraction of the duration of the signal, the sifting process is stopped. Default values of the parameters are : $\alpha_F \approx 0.05, \Theta_1 \approx 0.05, \Theta_1 = 10\Theta_2$. In general, it has been found that the *dyadic filterbank* nature of EMD is well maintained, for fGn with both flatband and skewed spectra, if the number of sifting iterations is around ten ($N = 10$) [118, 119].

Besides the above-mentioned marginal issues, there are two major aspects of EMD that need to be discussed, particularly for decomposing speech signals :

(i) Ability to separate frequency components.

(ii) Mode-mixing.

To examine the ability of EMD to separate different frequency components of the signal, a signal, $x_{mix}(t)$, composed of a lower-frequency sinusoid, $x_l(t)$, and a higher-frequency sinusoid, $x_h(t)$, is considered [120].

$$x_{mix}(t) = x_l(t) + x_h(t) = a_l \cos(2\pi f_l t + \phi_l) + a_h \cos(2\pi f_h t + \phi_h) \ , \ f_l, f_h \ll F_s \ ,$$

where F_s is the sampling frequency of the discrete-time digital version of the signal. To simplify the experiment, and without any loss of generality, $x_{mix}(t)$ is considered as,

$$x_{mix}(t) = a_{rat}\cos(2\pi f_{rat}t + \phi_d) + \cos(2\pi t)\ ,\ f_{rat} = \frac{f_l}{f_h}\ ,\ a_{rat} = \frac{a_l}{a_h}\ ,\ \phi_d = \phi_l - \phi_h \tag{2.17}$$

The signal $x_{mix}(t)$ is decomposed by EMD, and the following parameter is computed.

$$c_1^{(N)}(a_{rat}, f_{rat}, \phi_d) = \frac{||d_1^{(N)}(a_{rat}, f_{rat}, \phi_d) - \cos(2\pi t)||_2}{||a_{rat}\cos(2\pi f_{rat}t + \phi_d)||_2}\ , \tag{2.18}$$

where $d_1^{(N)}(a_{rat}, f_{rat}, \phi_d)$ is the first IMF obtained from the decomposition of $x_{mix}(t)$, where N sifting iterations have been used in the sifting process. In the experiment, $N = 10$ is used, whereas ϕ_d is kept constant. The parameter $c_1^{(10)}(a_{rat}, f_{rat}, \phi_d)$ is averaged over different values of $\phi_d \in [0, 2\pi)$. Thus, $c_1^{(10)}(a_{rat}, f_{rat}, \phi_d)$, represents a function of the frequency and amplitude ratios, f_{rat} and a_{rat}, respectively, where $||.||_2$ denotes the Euclidian norm. $c_1^{(10)}(a_{rat}, f_{rat}, \phi_d)$ gives a measure of whether EMD could successfully extract the components of the signal, $x_{mix}(t)$, or not [120].

Figure 2.9 plots the distribution of $c_1^{(10)}(a_{rat}, f_{rat}, \phi_d)$ as an image, with $f_{rat} \in\ ]0, 1[$ and $a_{rat} \in\ [0.01, 100]$ being the independent variables. The whiter regions of Figure 2.9 indicate that the components have been properly extracted, whereas the darker shades indicate that proper decomposition could not be achieved by EMD. As can be seen from the figure, for EMD to successfully decompose the signal into its actual constituents, there is a dependency on both f_{rat} and a_{rat}. There is a hard cut-off, $f_{rat} \lesssim 0.67$, irrespective of a_{rat}, only below which the constituents could be satisfactorily segregated. Also, even within this limit, the performance of segregation decreases as the relative strength of the lower-frequency component increases with respect to the higher-frequency component. Ideally, for proper segregation of the components, $a_{rat} \lesssim 1$ is required [120].

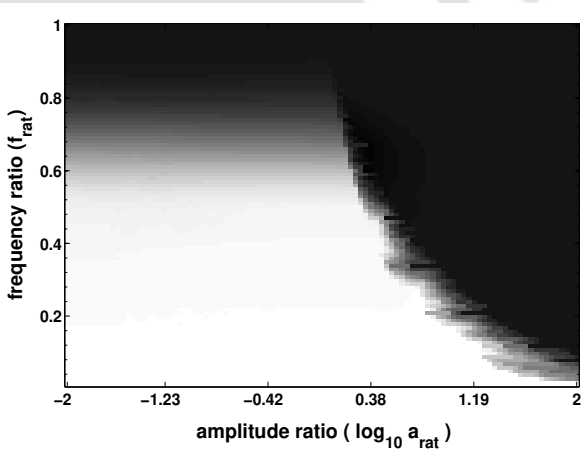


Figure 2.9: 2-D projection of $c_1^{(10)}(a_{rat}, f_{rat}, \phi_d)$ onto the (a_{rat}, f_{rat}) plane of amplitude and frequency ratios.

This simple experiment of segregating the sinusoidal constituents of $x_{mix}(t)$ gives us an idea about the difficulties involved in extracting the *true* components of a non-linear and non-stationary signal

like speech. To add to the problem, most of the energy of the speech signal is present in its voiced regions, which have a high spectral slope of - 6 dB/octave [2,13,14]. This causes the higher-frequency spectrum of speech to be overshadowed by its lower-frequency spectrum. Thus, the characteristics of the speech spectrum are not in tune with the requirements of the amplitude ratio, needed for successful segregation of its components by EMD. Due to this fact, EMD is limited in extracting meaningful IMFs which characterize the higher-frequency content of the speech signal. Henceforth, as is discussed later, most of the *vocal tract resonances* or *formants* of voiced speech are captured by the first IMF alone [121, 122]. The second IMF captures the first formant only, and the rest of the IMFs are of a lower-frequency and may represent the glottal source information [106, 121, 122].

Thus, like any other technique, EMD also has its due share of limitations. However, the most important phenomenon in the EMD process is the phenomenon of *mode-mixing*. Mode-mixing may be defined as the presence of *disparate frequency scales* within an IMF, and/or the presence of the same frequency scale in multiple IMFs [18, 86, 87, 107, 116, 119]. It is vividly observed in the case of non-stationary signals in which oscillations of *disparate frequency scales* occur intermittently. In reality, mode-mixing is not unexpected, as EMD is designed to locally separate a signal into low-frequency and high-frequency components. However, for many applications, this phenomenon may hinder the utility of the IMFs, and one may instead want IMFs which have a narrower frequency spectrum, as is desired ideally in AM-FM analysis. As an example of this phenomenon, we may consider the IMFs extracted from the speech signal in Figure 2.4. As is seen in the figure, IMF_1, which mostly consists of high-frequency oscillations, is corrupted in between by lower-frequency signals. Similarly, the primary frequency scales reflected in IMFs 2-4 seem to be distributed amongst them. There are parts of IMF_3 that appear to have the same frequency scale as that of major parts of IMF_4. Similarly, parts of IMF_2 seem to carry a low amplitude oscillation, which is mainly present in IMF_3. If such frequency variations are too large within an IMF, then the instantaneous frequency and amplitude envelope functions, estimated from it, would not be reliable, as discussed earlier.

2.2 Developments of EMD

To reduce the effects of mode-mixing in extracting IMFs from real physical signals, many modifications have been proposed to the EMD algorithm [105, 119, 123–127]. However, the best results have come by the infusion of noise to the signal. It was observed that by combining the signal

with finite amplitude White noise, before feeding it to the EMD process, mode-mixing could be curtailed satisfactorily. This development was termed the *Ensemble Empirical Mode Decomposition* (EEMD) [119].

The idea of infusing finite amplitude White noise into the signal serves an important purpose - the flat spectrum of White noise balances the skewed spectrum of speech to a certain extent. It lends energy to the subdued higher-frequency spectrum of speech, which makes the extraction of the higher-frequency content of the speech signal much more feasible for the EMD process. Simultaneously, another effect occurs - the addition of finite amplitude White noise increases the number of extrema present in the *inner residue* signal, which when used as IPs in a sifting iteration, leads to better estimates of the maxima and minima envelopes of the signal [119]. Figure 2.10 shows the flowchart of EEMD. The pseudocode for it is given below.

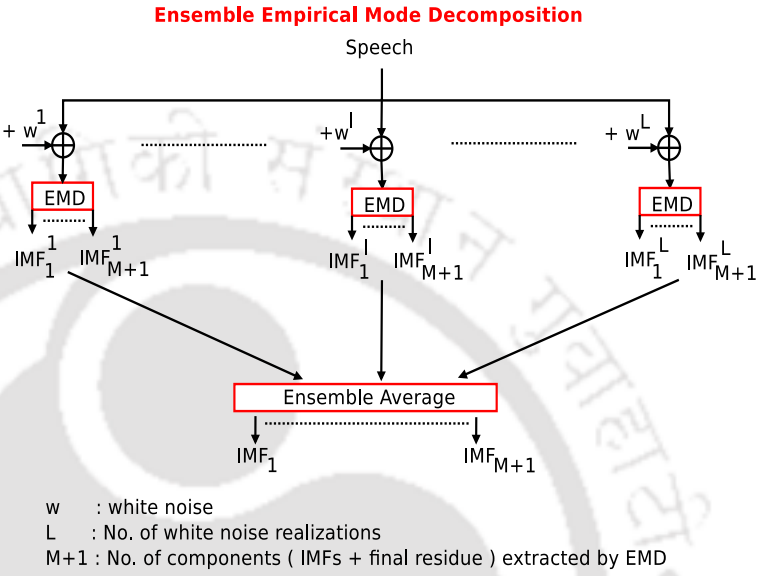


Figure 2.10: Flowchart of Ensemble Empirical Mode Decomposition.

Pseudocode for EEMD : Let $s(t)$ be a continuous-time speech signal.

•(**i**) Create L noisy copies of the signal, $s(t)$, using L independent realizations of finite amplitude normally distributed White noise (having zero mean and unit variance), $\mathcal{N}(0,1)$.

$$s^l(t) = s(t) + \beta w^l(t),\ l = 1, 2, ..., L\ , \tag{2.19}$$

where $s^l(t)$ is the l^{th} noisy copy of $s(t)$, obtained by adding the l^{th} White noise signal, $w^l(t)$, to it. The parameter, $\beta > 0$, controls the variance of the noise with respect to the signal. Generally, the variance of the noise is taken as 10-40 % of the variance of the signal [119].

•(**ii**) Decompose each $s^l(t)$ using EMD, to obtain M number of IMFs from the noisy signal. Generally, $N = 10$ sifting iterations are used in the sifting process, but may be adjusted for the task at hand and

with respect to the level of added noise [119].

$$s^l(t) = \sum_{k=1}^{M} h_k^l(t) + r_M^l(t) = \sum_{k=1}^{M+1} h_k^l(t) \ , \ l = 1, 2, ..., L \tag{2.20}$$

•(**iii**) The final components are obtained as the ensemble average of the components obtained from each noisy copy of the signal.

$$h_k(t) = \frac{1}{L} \sum_{l=1}^{L} h_k^l(t) \ , \ k = 1, 2, ..., M+1 \ , \tag{2.21}$$

$$\bar{s}(t) = \sum_{k=1}^{M+1} h_k(t) = \frac{1}{L} \sum_{k=1}^{M+1} \sum_{l=1}^{L} h_k^l(t) \tag{2.22}$$

It is expected that as the number of White noise realizations, L, is increased, the effect of noise would cancel out. If $var[.]$ denotes the operation of calculating the variance, we have,

$$var[\bar{s}(t)] = var\Big[\frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{M+1} h_k^l(t)\Big] \ ,$$

$$var[\bar{s}(t)] = var\Big[\frac{1}{L} \sum_{l=1}^{L} s^l(t)\Big] = var\Big[\frac{1}{L} \sum_{l=1}^{L} \{s(t) + \beta w^l(t)\}\Big] \ ,$$

$$var[\bar{s}(t)] = var[s(t)] + \frac{1}{L}\beta^2 \ , \tag{2.23}$$

$$var[\bar{s}(t)] = var[s(t)] \ , \ L \to \infty \ , \tag{2.24}$$

$$\bar{s}(t) = s(t) = \sum_{k=1}^{M+1} h_k(t) \ , \ L \to \infty \tag{2.25}$$

For a discrete-time digital speech signal, $s(n)$, the above-mentioned continuous-time operations would be replaced by discrete-time operations, and equations (2.22) and (2.25) would converge to,

$$\bar{s}(n) = s(n) = \sum_{k=1}^{M+1} h_k(n) = \frac{1}{L} \sum_{k=1}^{M+1} \sum_{l=1}^{L} h_k^l(n) \ , \ L \to \infty \tag{2.26}$$

Figure 2.11 shows the IMFs obtained by EEMD of the same speech signal, which was decomposed by EMD in Figure 2.4. $L = 10$ White noise realizations have been used in the process, and the variance of noise has been kept at 20 %. $N = 10$ and $M = 9$ are used in the decomposition. It is evident from the figures that EEMD produces components with much lesser mode-mixing than EMD. Also, the IMFs of EEMD have a much better representation of the higher-frequency spectrum of speech, as is reflected in the Hilbert spectrum presented in Figure 2.12. To quantify this observation, we calculate the *mean frequency* of the IMFs generated by EMD and EEMD [106].

The mean frequency of IMF_k, denoted as F_k^m, gives an indication of the dominant frequency

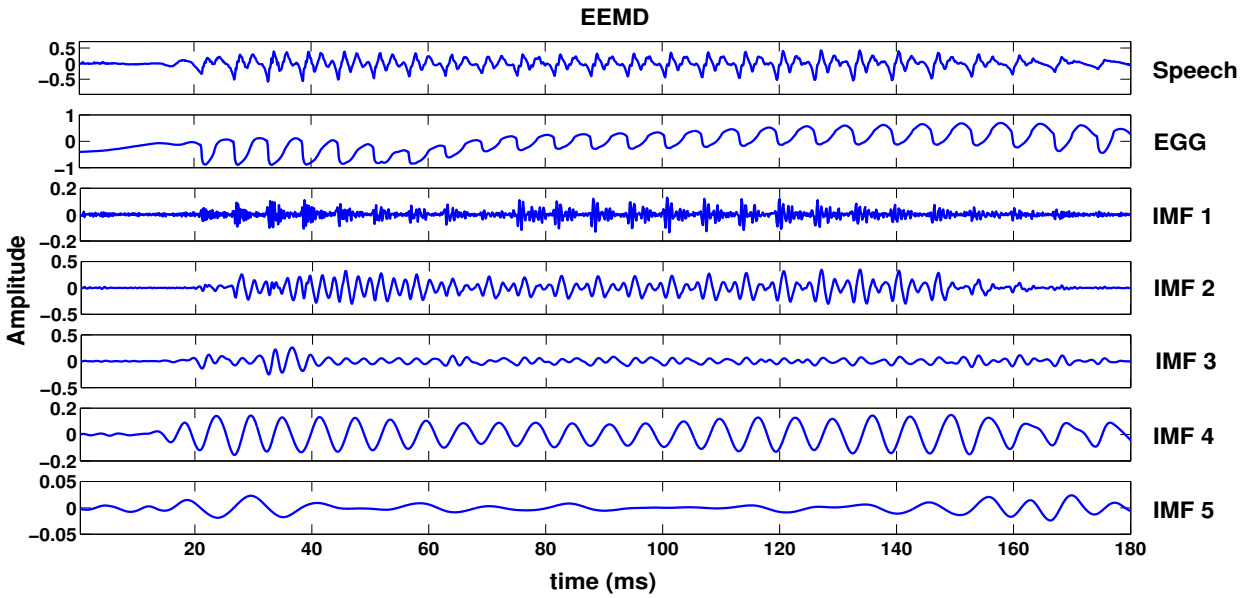


Figure 2.11: (a) Speech ; (b) EGG ; (c)-(g) are IMFs 1-5 of the speech signal obtained by EEMD.

manifested in the IMF. Mathematically, it gives the *central tendency of the power spectrum* of the IMF. For a continuous-time speech signal, $s(t)$, the mean frequency of IMF$_k$, or $h_k(t)$, may be obtained as,

$$F_k^m = \int_{f=0}^{\infty} \frac{f \times S_k(f)df}{\int_{f=0}^{\infty} S_k(f)df} \ , \ k = 1, 2, .., M+1 \ , \tag{2.27}$$

where f and $S_k(f)$ represent the analog frequency and the power spectrum (squared magnitude spectrum) of IMF$_k$, respectively. In our practical implementation, of course, we utilize a discrete-time digital version of $s(t)$. For such a digital speech signal, $s(n)$, obtained by sampling $s(t)$ at a rate F_s, we may obtain the mean frequency of IMF$_k$, $h_k(n)$, as,

$$F_k^m = \sum_{f=0}^{F_s/2} \frac{f \times S_k(f)}{\sum_{f=0}^{F_s/2} S_k(f)} \ , \ k = 1, 2, .., M+1 \tag{2.28}$$

In the above equation, f represents the analog frequencies corresponding to the discrete frequencies of the DFT spectrum of $h_k(n)$. $S_k(f)$ represents the power spectrum (squared magnitude spectrum) evaluated at these frequencies.

Figure 2.13(a) shows how the lower order IMFs of EEMD have a much higher mean frequency than that of EMD, thus giving a better representation of the higher-frequency spectrum of the speech signal. To evaluate how the lower-frequency information of speech is represented by EMD and EEMD, the *maximum correlation* of the discrete-time digital EGG signal, $e_{\text{EGG}}(n)$, with respect to the IMFs,

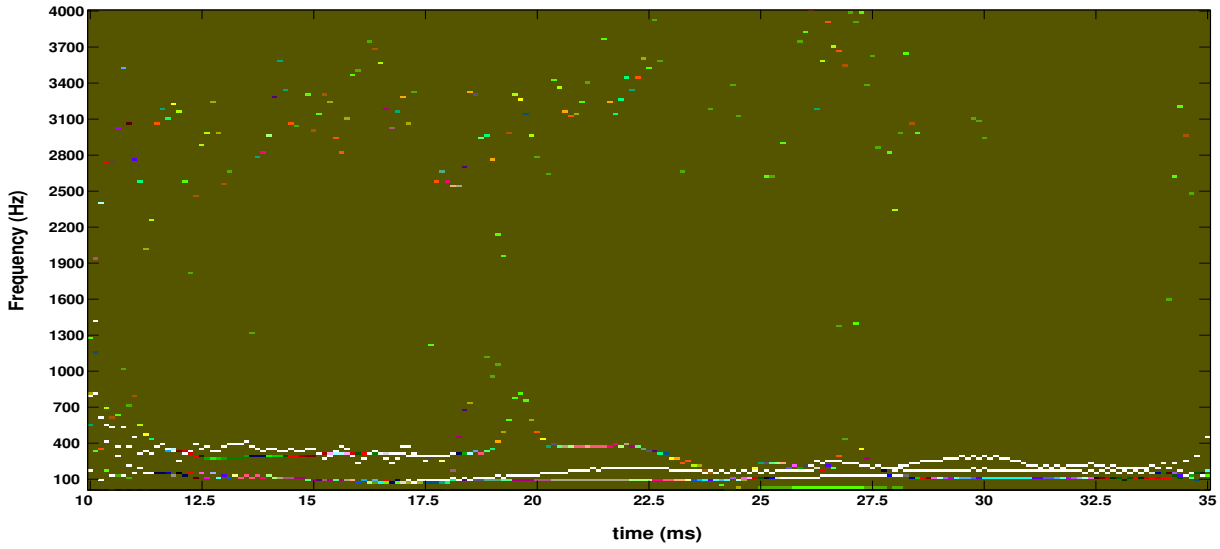


Figure 2.12: Hilbert Spectrum of the speech signal, used in Figure 2.11, using EEMD.

obtained from both EMD and EEMD, is evaluated.

$$R_k^e = \max_m \Big\{ \sum_n h_k(n) e_{\mathrm{EGG}}(n+m) \Big\} \ , \ k = 1, 2, ..., M+1 = 10 \ , \tag{2.29}$$

where R_k^e represents the maximum correlation of IMF_k with the EGG signal. Figure 2.13(b) plots the values of R_k^e for both EMD and EEMD [106]. As is evident from the figure, EEMD reflects the glottal source information in a better way than EMD. Also, the source information is less distributed amongst the components of EEMD than that of EMD, a consequence of reduced mode-mixing. In general, the source information is found to be distributed almost entirely amongst two consecutive IMFs in the case of EEMD [106].

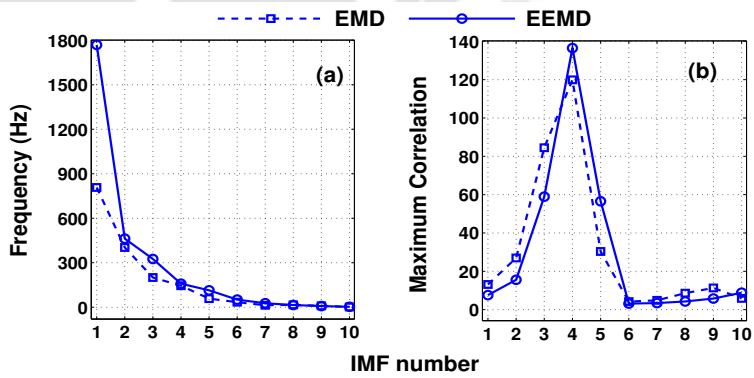


Figure 2.13: (a) Mean Frequency of the IMFs - EMD vs. EEMD ; (b) Maximum correlation of the IMFs with the EGG signal - EMD vs. EEMD.

Finally, we may look at the distribution of the speech resonances (vocal tract resonances) in the IMFs. Figure 2.14 shows the magnitude spectra of the LP filters of the first four IMFs of a voiced speech signal of the TIMIT corpus [128]. A 24-order LP analysis is used on a 20 ms voiced segment of the 16 kHz speech signal, and its corresponding IMFs. The speech segment (not its IMFs) is pre-

emphasized by a high-pass filter, $H_{pre}(z) = 1 - 0.98z^{-1}$, prior to LP analysis. The reference formant frequencies are obtained from the VTR Formants database [129]. For better visualization, the spectra are plotted only up to 4 kHz, within which the first four principal formants of the speech signal are generally confined. As can be seen from the figure, the first IMF of EMD carries all the formants, except the first formant, which is carried by the second IMF [121, 122]. In the case of EEMD, the formants structure is more evenly distributed amongst the first four IMFs. Thus, a better spectral segregation is achieved in the case of EEMD, compared to that of EMD. The IMFs of EEMD, hence, may be considered to be better suited for AM-FM analysis.

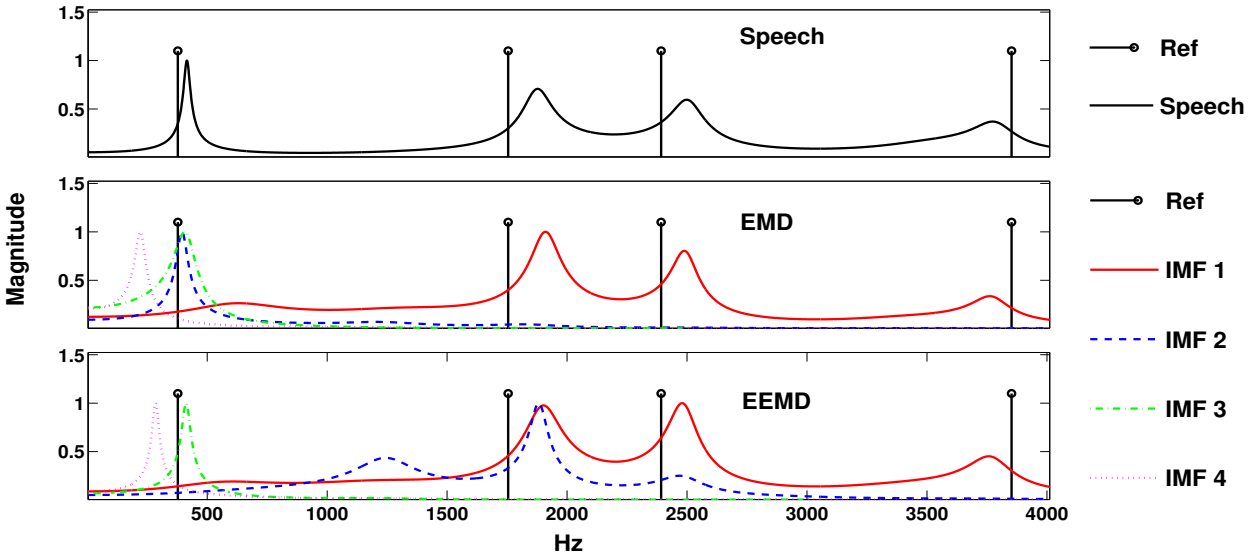


Figure 2.14: **Formants distribution in the IMFs - EMD vs. EEMD. (a) Normalized magnitude spectrum of the LP filter of pre-emphasized voiced speech ; Normalized magnitude spectra of the LP filters of the first four IMFs, derived from (b) EMD of voiced speech, and (c) EEMD of voiced speech.**

2.2.1 Improvements in EEMD

While there are many merits of EEMD, there are also certain limitations to the process. One of them is its efficiency. As is reflected in equation (2.24), a large number of White noise realizations is required to average out the effect of noise [119]. Again, there is no guarantee that each of the noisy signal copies would produce the same number of IMFs, which creates problems in averaging the IMFs. One way to circumvent this problem is to restrict the EMD decomposition to a fixed smaller number (M) of IMFs (as shown in Figure 2.1) than that would be obtained if the decomposition is allowed to continue till a trend with only two extrema remains. In the recent years, efforts to efficiently cancel

out the noise infused with the signal has led to many EEMD variants [105, 126, 127].

It was observed that using White noise in pairs of opposite polarities substantially reduces the effect of the added noise in the IMFs finally derived from EEMD. This development was termed the *Complementary Ensemble Empirical Mode Decomposition* (CEEMD) [126]. However, the problem that the number of IMFs produced could still be different for the different EMD processes of an EEMD or CEEMD decomposition, still existed. To circumvent this problem, an algorithm was designed, which not only decomposes the signal, but parallely also the White noise realizations. The IMFs obtained from the White noise realizations, which could be interpreted as correlated noise signals, are then fused with the outer residue of the signal, at the beginning of each sifting process. The signal IMFs are obtained progressively after averaging the results at each stage. This algorithm was termed the *Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (CEEMDAN) [127]. However, it was observed that CEEMDAN sometimes produced some high-frequency low-amplitude spurious IMFs, in the decomposition. To overcome this problem, the *Improved Complete Ensemble Empirical Mode Decomposition with Adaptive Noise* (ICEEMDAN) [105] was developed, which makes some subtle and effective modifications to the CEEMDAN algorithm. The pseudocode of ICEEMDAN is given below :

Algorithm for ICEEMDAN : Let $s(t)$ be a continuous-time speech signal. Let $E_k[.]$ be the operator which denotes the operation of extracting the k^{th} IMF from any signal, $x(t)$, using EMD. Then, if $\Upsilon[x(t)]$ denotes the local mean of the signal, we have, $E_1[x(t)] = x(t) - \Upsilon[x(t)]$. Let $w^l(t)$ denote the l^{th} realization of zero mean unit variance White noise. $w^l(t) \sim \mathcal{N}(0,1)$.

For $k = 1, 2, ..., M$, repeat the following steps.

•(**i**) Let $r_0(t) = s(t)$. The $(k-1)^{th}$ residue, $r_{k-1}(t)$, is mixed with noise as,

$$r^l_{k-1}(t) = r_{k-1}(t) + \beta_{k-1} E_k[w^l(t)] \ , \ l = 1, 2, ..., L \ ,$$

where β_{k-1} is used to control the SNR at each stage of the decomposition.

•(**ii**) The k^{th} IMF, $h_k(t)$, is derived as,

$$E_1[r^l_{k-1}(t)] = r^l_{k-1}(t) - \Upsilon[r^l_{k-1}(t)] \ , \ l = 1, 2, ..., L \ ,$$

$$r_k(t) = \frac{1}{L} \sum_{l=1}^{L} \Upsilon[r^l_{k-1}(t)] \ ,$$

$$h_k(t) = r_{k-1}(t) - r_k(t)$$

•(**iii**) Go to step (i). Stop when a maximum number of IMFs, M, is extracted, i.e., when $k = M$.

$$s(t) = r_M(t) + \sum_{k=1}^{M} h_k(t) = \sum_{k=1}^{M+1} h_k(t) \tag{2.30}$$

Generally, and in this thesis, $\beta_0 = \epsilon_0\ std(s(t))/std(E_1[w^l(t)])$, and $\beta_{k-1} = \epsilon_0\ std(r_{k-1}(t)),\ k \geq 2$. In the above-mentioned parameters, $std(.)$ denotes the operation of calculating *standard deviation*. In this thesis, $\epsilon_0 = 0.2$. The number of iterations in a complete sifting process is determined by the local-global stopping criterion. The maximum number of iterations per sifting process is not allowed to exceed 15, i.e., $N \leq 15$ [105, 116, 130, 131]. Finally, we may note that for a discrete-time digital speech signal, $s(n)$, the ICEEMDAN algorithm and the various parameters would undergo suitable modifications/adjustments, and equation (2.30) may be represented as,

$$s(n) = r_M(n) + \sum_{k=1}^{M} h_k(n) = \sum_{k=1}^{M+1} h_k(n) \tag{2.31}$$

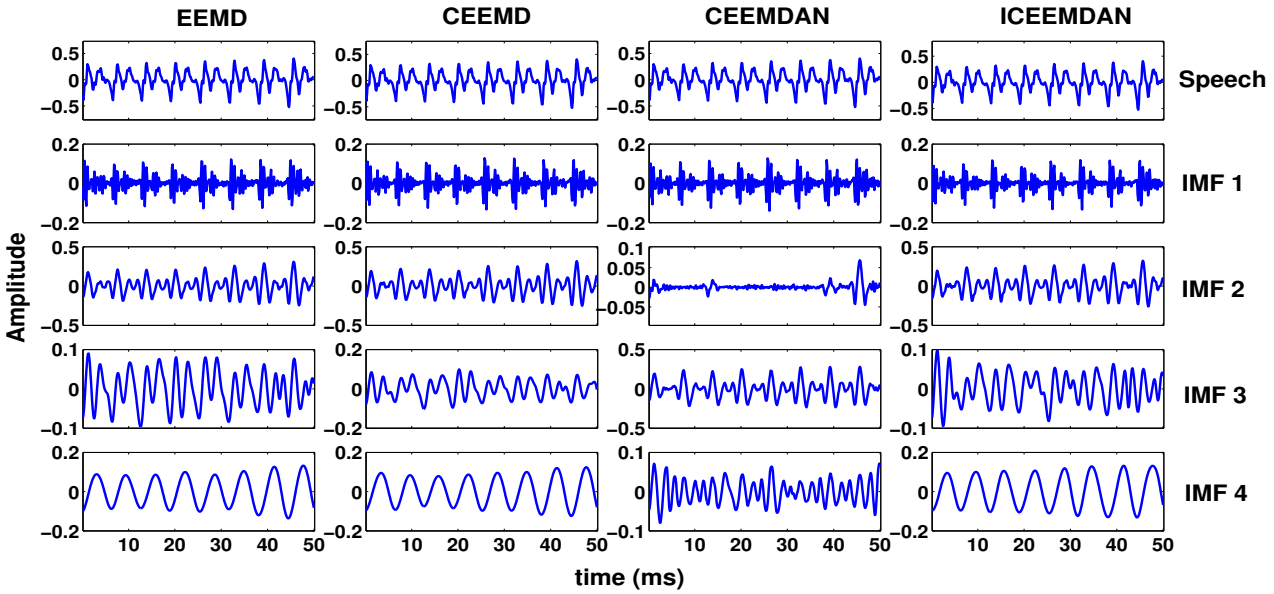


Figure 2.15: **The first four IMFs obtained from a speech signal using (from left to right column) EEMD, CEEMD, CEEMDAN, and ICEEMDAN. $L = 20$ White noise realizations (10 White noise pairs for CEEMD) are used in the processes. The number of sifting iterations are kept fixed to $N = 10$.**

Figure 2.15 shows the first four IMFs derived from a speech signal using EEMD, CEEMD, CEEMDAN, and ICEEMDAN. To maintain uniformity, $N = 10$ sifting iterations are used for all the four algorithms, and the local-global stopping criterion is not used for ICEEMDAN in this case. It may be observed that, in this example, there is no significant advantage of any one variant over the other [106]. In the case of CEEMDAN, a spurious mode (IMF_2) is exhibited. Figure 2.16 shows the

reconstruction error of the four algorithms, for the speech signal decomposed in Figure 2.15. Given a speech signal, $s(n)$, and its IMFs (including the final residue), $\{h_k(n), k = 1, 2, ..., M+1\}$, the *reconstruction error*, r_e, is given by,

$$r_e = 10 \log_{10} ||s(n) - \sum_k h_k(n)||_2 \tag{2.32}$$

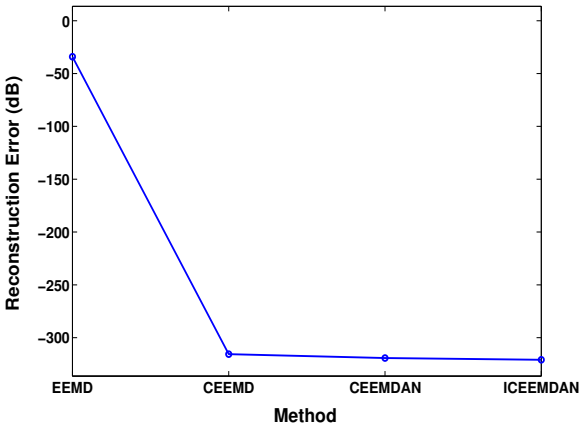


Figure 2.16: **The reconstruction error (in dB) of the speech signal shown in Figure 2.15, for the methods - EEMD, CEEMD, CEEMDAN and ICEEMDAN.**

Table 2.1: **Computational time of EMD, EEMD, CEEMD, CEEMDAN, and ICEEMDAN, in decomposing a speech signal of around 3.5 seconds duration. Ten IMFs are extracted from the EMD variants, where $N = 10$ sifting iterations are used per sifting process. Only $L = 10$ White noise realizations are used. The algorithms are implemented in the GUI mode of MATLAB, on a machine having an Intel i5 quad-core processor of 3.2 GHz clock frequency, and 4 GB RAM.**

Method	EMD	EEMD	CEEMD	CEEMDAN	ICEEMDAN
Time (s)	0.83	15.62	14.94	32.88	30.11

As Figure 2.16 shows, the reconstruction errors for the EEMD variants are lower than that of EEMD. However, it may be noted that even for EEMD, the reconstruction error is quite low. However, the processing time, for both EEMD and its variants remains quite large. Table 2.1 lists the time taken to extract ten IMFs ($M = 9$) from a speech signal ($F_s = 8$ kHz) of around 3.5 s duration, by EMD, EEMD, and the EEMD variants. A fixed number of sifting iterations, $N = 10$, is considered for all the methods, for a fair comparison, and the local-global stopping criterion is not used in this case for ICEEMDAN. Only $L = 10$ White noise realizations are used for the noise-assisted methods. As is clear from the table, EEMD and its variants are time costly, and hence, they currently are limited in use in real-time applications, despite their obvious merits. Efficient coding, of course, may alleviate this drawback substantially. Regardless, EEMD and its variants are quite useful in applications which are not real-time.

2.3 Comparison with other speech processing approaches

As discussed in Chapter 1, non-stationary signal analysis techniques, like Wavelet Transform (WT) and AM-FM analysis, provide an alternative to conventional speech analysis. Henceforth, we need to weigh the effectiveness of EMD with respect to such techniques. While AM-FM analysis has been discussed in detail in Section 1.2, a small discussion on WT is worthwhile at this juncture.

As a way of overcoming the time and frequency resolution limitations of STFT, WT was introduced [16, 17, 38]. The continuous-time WT of a signal, $s(t)$, is given by,

$$\mathcal{W}_s^{\psi}(\tau, r) = \frac{1}{\sqrt{|r|}} \int_{-\infty}^{\infty} s(t)\psi^* \left(\frac{t-\tau}{r}\right) dt \tag{2.33}$$

where $\psi(t)$, called the *mother wavelet*, represents an oscillatory signal of finite time-width. Here, r represents the scale, and τ the time around which the signal is analyzed. Thus, WT allows the visualization of the signal at different scales, depending on the value of r. A comparison of equation (1.5) with equation (2.33) reveals that WT, basically, works as an adjustable window STFT. The discrete version of the continuous-time WT, called the *Discrete Wavelet Transform* (DWT), is popularly used for analyzing non-stationary discrete-time digital signals. For a discrete-time digital speech signal, $s(n)$, the *dyadic* DWT, at j^{th} level of decomposition, is obtained as,

$$\mathbb{W}_s^{\psi}\left[\tau, 2^j\right] = \sum_n s(n)\psi^*_{\tau,2^j}(n) \tag{2.34}$$

A comparison between the time-frequency resolutions of STFT and DWT may be visualized in Figure 2.17. The decomposition obtained by DWT is further extended by the *Wavelet Packets Transform* (WPT), which applies the filtering process of the *binary decomposition tree* to both the low-frequency and high-frequency component of the signal, at each stage of the decomposition. Thus, an over-complete dictionary is obtained by WPT, providing more flexibility for the analysis of specific frequency bands. From the decomposition provided by WPT, different sub-trees can be selected in order to extract the desired

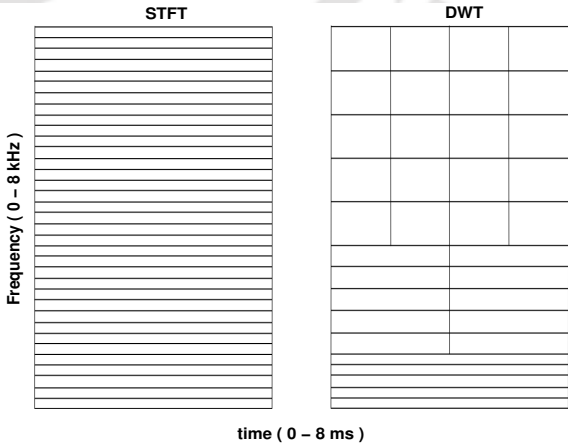


Figure 2.17: **Comparison between time-frequency resolution of STFT (left) and 3-level DWT (right), calculated for a 8 ms signal of sampling frequency 16 kHz.**

information. The flexibility provided by WPT, however, comes along with the challenging problem of choosing the optimum set of coefficients (for a particular application) among all the possible combinations [132–135]. Another concern in the design of useful wavelet-based decomposition is the choice of an adequate wavelet family, and associated parameters, that suit the characteristics of the signal of interest and the problem at hand [136–138]. Thus, the selection of the *mother wavelet*, $\psi(t)$, is critical to Wavelet analysis. As an illustration of this point, we may consider the case of a speech signal decomposed by 10-level DWT, using the 'Daubechies-4' and 'Biorthogonal-2.4' wavelets. The first five time-domain detail components of a digital speech signal, reconstructed by Inverse DWT of the first five detail coefficients, are shown in Figure 2.18. It is evident that changing the *mother wavelet* changes the decomposition. Apart from this, the Wavelet analysis does not tackle the problem of *non-linearity* of the speech signal [1, 15, 19, 20]. As mentioned earlier, WT (hence DWT and its subsidiaries) is essentially an *adjustable window* STFT, and hence not applicable for analyzing non-linear systems [18, 87]. This is where EMD, and hence its variants, scores over STFT and WT. EMD and its variants are able to extract the components of the signal, without requiring any *a priori* basis. Further, the *sifting process* is a non-linear process, and hence EMD may be deemed applicable for analyzing signals produced by non-linear systems [18, 87].

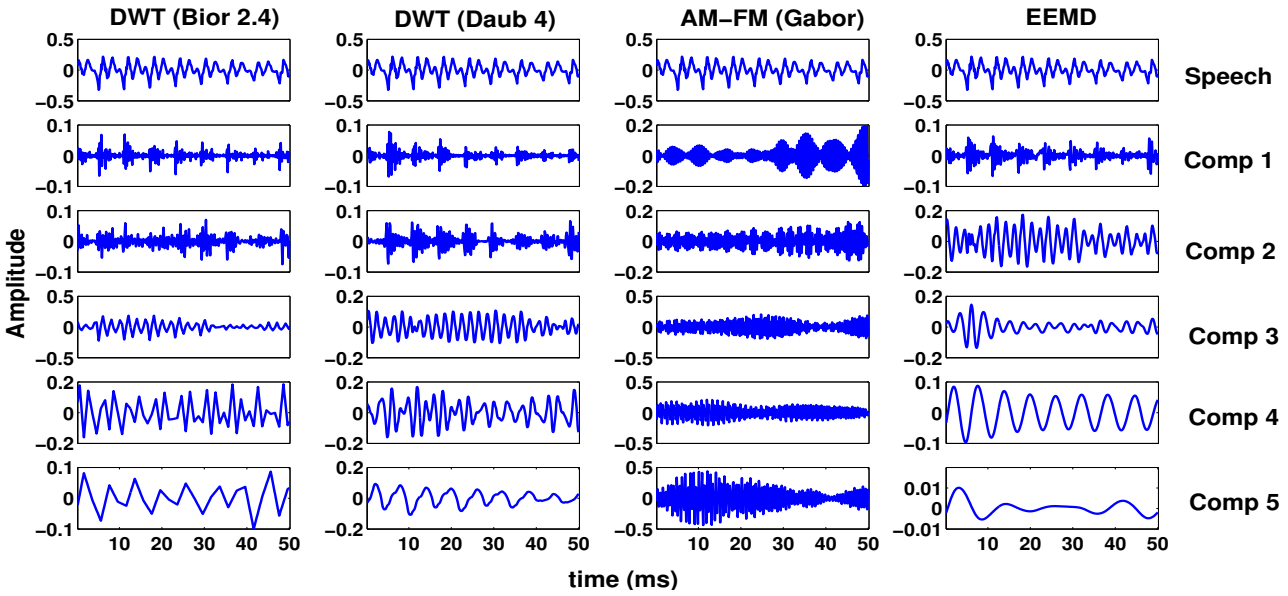


Figure 2.18: **The first five components (in decreasing order of frequency content) of a speech signal, as obtained from DWT (Biorthogonal 2.4 wavelet), DWT (Daubechies 4 wavelet), AM-FM analysis (20 filter linear Gabor filterbank), and EEMD.**

Unlike STFT and WT, AM-FM analysis (MDA) may be used for dealing with both the non-

stationary and the non-linear characteristics of the speech signal, as discussed in Section 1.2. The basic aim of AM-FM analysis is to represent the speech signal in terms of AM-FM components, which are dominated by its resonant frequencies, as reflected in equation (1.7). But, as the speech resonances are not known *a priori*, traditional AM-FM analysis uses a large bank of overlapping band-pass filters, to obtain AM-FM signals from the signal, which are used for further analysis [4, 15, 34, 35]. As such, the design of the filterbank remains an open issue. Figure 2.18 (third column from the left) shows the first five high-frequency components obtained using a Gabor filterbank of 20 uniformly spaced filters in the Hz scale, each having an effective bandwidth of 400 Hz. It is evident that AM-FM analysis would produce a significant number of redundant components, which may not be useful for analysis. Figure 2.18 (fourth column from the left) also presents the first five IMFs of the speech signal, derived using EEMD, for comparison, which reflects its superiority over DWT and traditional AM-FM analysis. As the variants of EEMD perform similarly, they are not shown in the figure.

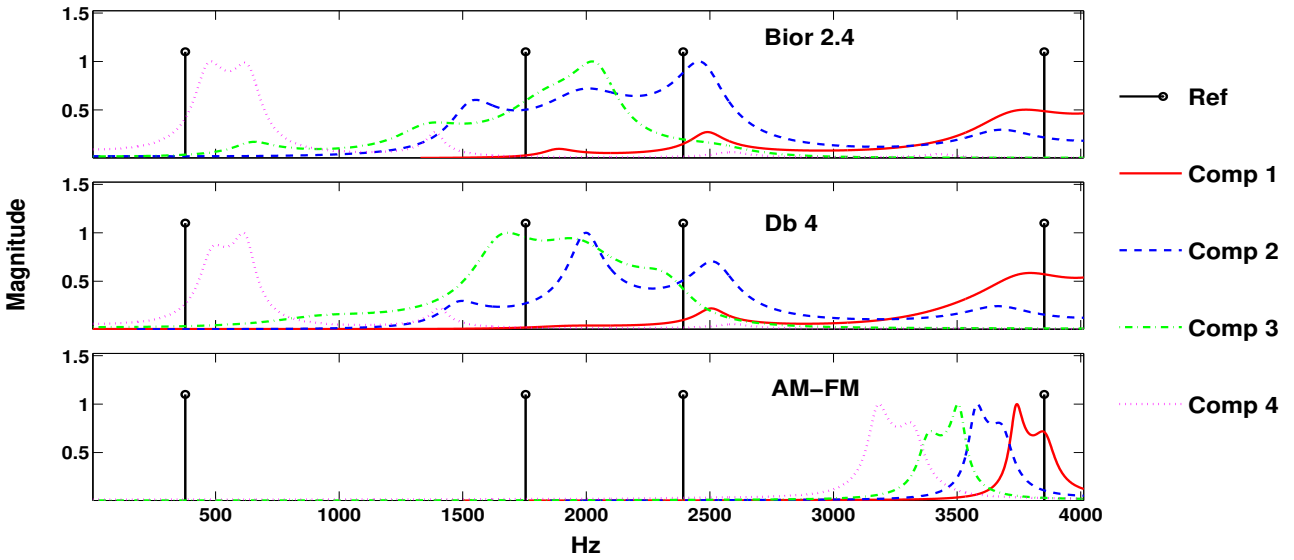


Figure 2.19: **Normalized magnitude spectra of LP filters of the first four high-frequency components of the speech signal used in Figure 2.14. The components are obtained from (a) DWT (Biorthogonal 2.4 wavelet) ; (b) DWT (Daubechies 4 wavelet) ; (c) AM-FM analysis, using a 20 filter linear Gabor filterbank.**

Figure 2.19 shows the magnitude spectra of the LP filters of the first four high-frequency components, obtained from DWT and AM-FM analysis, of the same voiced speech segment ($F_s = 16$ kHz) that has been used in Figure 2.14. Irrespective of the type of mother wavelet used, DWT operates as a *fixed dyadic filterbank* [16, 17, 38]. The components derived clearly show the overlap of formants information, and also seem to lack precision in capturing the formants. In the case of

AM-FM analysis, the components are obtained by Gabor filtering the speech segment, with a 40-filter Gabor filterbank, each filter having a bandwidth of 400 Hz. The LP magnitude spectra for only the four highest frequency components, corresponding to the Gabor filters having center frequencies below 4 kHz, are plotted. As reflected in the figure, AM-FM analysis is much more precise, but it produces a multitude of components which do not capture the vocal tract resonances, which may not be useful for speech analysis. A comparison of Figures 2.14 and 2.19 shows that, out of all the techniques, EEMD provides the best solution to achieving the ideal goal of AM-FM analysis of speech - a limited number of components which encapsulates the vocal tract resonances precisely. More importantly, it does so without using any *a priori* basis or pre-designed filterbank, and hence is adaptive to the signal at hand. In simple words, EEMD (hence its advanced versions) paves the way for *adaptive* AM-FM analysis of the speech signal.

Finally, we may look at the time-cost of the various speech analysis techniques discussed till now. Table 2.2 enlists the time taken by DWT, AM-FM analysis, EMD, and EEMD in decomposing a digital speech signal (F_s = 8 kHz). The same speech signal for which Table 2.1 is generated, is used in this case. Ten time-domain components are generated from the DWT decomposition of the speech signal using Biorthogonal 2.4 wavelet. For AM-FM analysis, a 20-filter overlapping Gabor filterbank is used. As can be observed from the table, DWT and AM-FM analysis are faster algorithms. EEMD is extremely time costly. EMD though not the fastest, it has an acceptable time-cost, which may be improved by efficient coding.

Table 2.2: Computational time of EMD, EEMD, DWT and AM-FM analysis, in decomposing a speech signal of around 3.5 seconds duration. Ten IMFs are extracted from EMD and EEMD, where 10 sifting iterations are used per sifting process. Only $L = 10$ White noise realizations are used for EEMD. 10 components are also extracted from DWT using Biorthogonal 2.4 wavelet. 20 components are extracted from AM-FM analysis using a Gabor filterbank, each filter having a bandwidth of 400 Hz. The algorithms are implemented in the GUI mode of MATLAB, on a machine having an Intel i5 quad-core processor of 3.2 GHz clock frequency, and 4 GB RAM.

Method	EMD	EEMD	DWT	AM-FM
Time (s)	0.83	15.62	0.42	0.40

2.4 Scope for present work

This chapter discusses the technique of EMD from the perspective of an adaptive AM-FM analysis technique for speech signal processing. The chapter is dedicated entirely to discussing the characteristics of EMD, its advantages and disadvantages, and its noise-assisted advanced versions which try to diminish its demerits. The decomposition of the speech signal by EMD, and its advanced versions, is analyzed. The segregation of the *vocal tract resonances* or the *formants* of

voiced speech, in its IMFs, is depicted, to project EMD as a tool for performing AM-FM analysis.

As a result of all the attractive and positive aspects of EMD, it has already been applied in various speech processing applications [87–103, 106, 112, 122, 139–157]. We may categorize some of these explorations as follows :

(i) Speech Analysis - Enhancement/Denoising [90, 112, 139–144], Pitch Tracking [88, 89, 145, 146], Formant Tracking [122, 147], Pathological Voice Analysis [91, 148, 149], etc.

(ii) Feature Extraction - Noise and audio Classification [150, 151], Emotion Classificatio [152, 153], Speaker Recognition [92, 154], Voiced/Unvoiced speech classification [106, 155–157] etc.

The objective of this thesis, however, is not a mere application of the IMFs to speech processing applications. The **first objective** of this thesis, which is addressed in the next chapter, **Chapter 3**, is to customize EMD as a tool for speech processing, in a time-efficient manner. As has been discussed, the utility of EMD, at its current state, is limited, owing to *mode-mixing*. EEMD and its advanced versions provide a much better decomposition of the speech signal, but at an almost unbearable time-cost. Hence, they have had limited use. If EMD could be customized to produce meaningful IMFs of speech, which are similar to that of EEMD, but at a much lesser time cost, it would become a much more attractive and useful tool for the speech community. For this purpose, experiments are performed so as to mimic the process of EEMD in the EMD algorithm. This results in a *modified version of* EMD (MEMD), which decomposes the speech signal with much lesser mode-mixing than EMD, and only at a fraction of the time endured by EEMD or its variants.

The **second objective** of this thesis is to investigate the ability of EMD, and its variants - MEMD and ICEEMDAN - in decomposing the speech signal into meaningful IMFs. The information contained in the speech signal may not be solely attributed to its waveform shape, whereas the decomposition provided by EMD and its variants is entirely based on the waveform characteristics of the speech signal. As such, it becomes necessary to evaluate the characteristics of the IMFs for varying nature of the speech signal. **Chapter 4** of this thesis is dedicated to this objective. Firstly, controlled experiments are done on synthetic speech signals to find out how the vocal tract resonators and the glottal source is manifested in their IMFs. The effects of variations of the fundamental frequency of the excitation source, and center frequencies and bandwidths of the resonators, are studied. Thereafter, using standard *cepstral* or *homomorphic* analysis [2, 13, 14], the distribution of the *source* and *system* characteristics of natural speech signals, in their IMFs, is studied. Different *phones* or *speech sounds*

uttered by male and female speakers are analyzed separately. The natural speech signals are further converted to telephone quality speech, using *International Telecommunication Union* (ITU) codecs, and their IMFs are analyzed.

The **third objective** of this thesis is to demonstrate the utility of the IMFs, derived from some advanced version of EMD, in performing a task that is generally confined to short-time LP analysis. For this purpose, we explore the task of detecting the *epochs* or *Glottal Closure Instants* (GCIs) of voiced speech [5–11, 21, 39]. Almost all the state-of-the-art methods for extracting the GCIs are based on processing the LP-residual signal obtained after LP analysis of the speech signal. While most of these techniques perform excellently under clean and controlled data conditions, when the speech signal is subjected to external noise or telephone channel conditions, they are not as effective [10, 11]. Such reduced performances may be correlated to the fact that the assumptions of short-time stationarity and linearity are further weakened when the speech signal is subjected to external influences. Therefore, if the IMFs could be utilized for estimating the GCIs, it would circumvent the assumptions of short-time stationarity and linearity, and might provide credible performances across varied conditions. With this viewpoint, in **Chapter 5**, the IMFs obtained from ICEEMDAN, and MEMD, are utilized for detecting the GCIs in the voiced regions of the speech signal, under clean, noisy, and telephone-channel conditions. Needless to say, the insights gained in Chapter 4, regarding the segregation of source and system information in the IMFs, would be useful in this task.

The **fourth objective** of this thesis is to show that the adaptive filterbank nature, manifested in the IMFs, captures speaker specific information which can complement the Mel filterbank. This would further showcase the utility of EMD (and its variants) as a speech signal analysis method. As discussed earlier, the Mel filterbank is a fixed structure, while the IMFs are unique to every speech signal. In fact, even for the same speech signal, the nature of the IMFs vary as the speech sound (hence the speech waveform) changes. In this context, the IMFs represent a concise but adaptive filterbank. As such, *cepstral* or *energy-like* features extracted from the IMFs could carry speaker-specific information that is different from that manifested in the MFCCs. With this viewpoint, in **Chapter 6**, the IMFs obtained from EMD, and MEMD, are utilized to extract three different feature-sets, which in combination with the MFCCs, are utilized for the task of text-independent *Speaker Verification* (SV). The experiments are performed on two different databases, using two different modeling techniques. The experiments are performed for three different speaking styles - *normal*, *fast*, and *whispered* - and also for *limited*

test data conditions. These large sets of experiments portray the practicality and generalization of the observations.

The final chapter, **Chapter 7**, of the thesis, summarizes the works, the contributions, and discusses future directions of work.

3

Modified Empirical Mode Decomposition

Contents

Outline

The objective of this work is to obtain the IMFs of the speech signal with reduced mode mixing, and in a time-efficient manner. The idea is to try and mimic the EEMD process, while ensuring time-efficiency. To this effect, modifications are proposed to the EMD algorithm, based on the critical nature of the *Interpolation Points* (IPs) used for cubic spline interpolation in EMD. The effect of using different sets of IPs, other than the extrema of the inner residue (as used in conventional EMD), is analyzed. It is found that having more IPs is beneficial to the decomposition, but only up to a certain limit, beyond which the characteristic *dyadic filterbank* nature of EMD breaks down. For certain sets of IPs, these modified EMD processes perform better than EMD, giving better frequency separability between the IMFs, and an enhanced representation of the higher-frequency content of the speech signal. A study of the *distribution of the formants*, in the IMFs of the speech signal, is done using LP analysis of the IMFs. It is found that the IMFs of the proposed EMD variants have a much better distribution of the formants structure within them, with reduced overlapping amongst their LP filter spectra, compared to that of conventional EMD. Henceforth, when subjected to the task of formants estimation of voiced speech, using LP analysis, the IMFs of the modified EMD processes cumulatively exhibit a superior performance than that of standard EMD, or the speech signal itself, under both clean and noisy conditions.

3.1 Introduction

The mechanism of EMD is critically dependent on generating the *mean envelope* of the *inner residue* signal. In the case of EMD, this is achieved by using the extrema (maxima and minima) of the inner residue as the anchor points, or the *Interpolation Points* (IPs), over which cubic splines are fitted. In the case of EEMD, the same process is employed, yet the IMFs obtained manifest significantly reduced *mode-mixing* - why is it so ? The difference between the EMD and the EEMD algorithm is not in the mechanism employed, but in the signal decomposed - the signal in the case EEMD is a noisy version of the actual signal which is required to be decomposed. *How could the addition of finite amplitude White noise provide a better decomposition, rather than inferior (noisy) IMFs ?* The answer to this question is two-fold. One aspect of this answer lies in the effect of White noise on the frequency spectrum of the signal, which in our case is the speech signal. The other aspect lies in the effect of White noise on the time domain speech waveform itself. To appreciate the first aspect, we

need to consider the observations regarding the ability of EMD to segregate frequency components, as discussed in Section 2.1.4, and encapsulated in Figure 2.9. It was observed that for EMD to satisfactorily decompose the binary mixture of sinusoids, apart from a considerable frequency gap between the components, the higher-frequency component was required to have comparable amplitude with respect to the lower-frequency component. This is not a favourable condition for decomposing a speech signal, as most of the energy in the speech signal is concentrated in its *voiced* regions, which exhibit a spectral slope of - 6 dB/octave [2,13,14]. In other words, the higher frequency spectrum of the speech signal is subdued by its lower-frequency spectrum, which is undesirable for EMD. The addition of White noise serves an important purpose here. White noise has a flat spectrum ; the mixture of the speech signal with a White noise realization reduces the sharp imbalance of energy that is prevalent in the original speech signal. This is the frequency domain interpretation of why EEMD, and so its variants, provide a better decomposition of the speech signal, compared to that of EMD. From the time domain perspective, there is one important visual observation - the increase in the number of extrema in the speech signal upon combining it with a White noise realization. Hence a question naturally arises - *"Did the increase in the number of IPs have any effect in the EEMD decomposition ? If so, could a similar decomposition be achieved without using noise, but by increasing the number of IPs in a certain way ?"*

It is the above-mentioned time-domain aspect of the EEMD decomposition that forms the basis for the experiments performed as part of this chapter. The objective of the experiments lies in the hypothesis that - *"If the number of IPs in the inner residue signal could be increased, effectively, by some simple signal processing method, rather than the infusion of noise, it might be possible to obtain a better signal decomposition, compared to that of conventional EMD"*. Such a decomposition is expected to be much faster, and computationally cheaper than EEMD, as the signal would be decomposed once, unlike in the case of EEMD where multiple noisy copies of the signal need to be decomposed. Even though such a decomposition might still be inferior to EEMD (hence its variants), it might be more applicable for speech processing applications, and other real world data analysis, where time is of precious value. With this viewpoint, three different ways are utilized to change the IPs from the conventional EMD process. The modified EMD processes are named M1-EMD, M2-EMD, and M3-EMD. The EMD variants are compared with conventional EMD, AM-FM analysis, and WT as well. To measure the ability of these EMD variants in segregating the frequency spectrum of speech,

a qualitative study of the distribution of the formants in the IMFs is performed. To quantify the advantage of the better splitting of the speech spectrum, the IMFs of EMD, and its proposed modified variants, are subjected to the task of detecting the first four principal formants of voiced speech.

The rest of the chapter is organized as follows: Section 3.2 describes the experimental setup of modifying the IPs of standard EMD. Section 3.3 analyzes the effect of changing the IPs on the filterbank nature of EMD. Section 3.4 discusses the convergence and robustness of the modified EMD processes. Section 3.5 discusses the utility of the proposed EMD variants with respect to other speech analysis methods. In Section 3.6, the distribution of the frequency spectrum of the speech signal, in its IMFs, in terms of their LP filter spectra, is studied. Section 3.7 encapsulates the results of formants estimation, from the IMFs of speech. Section 3.8 concludes this work.

3.2 Modifying EMD by changing the IPs

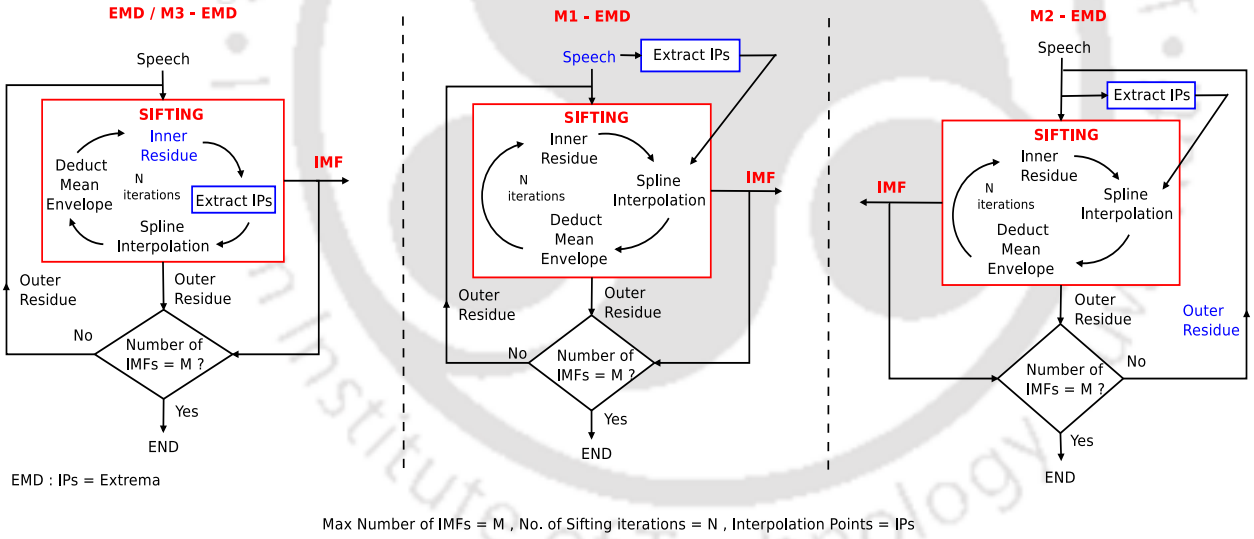


Figure 3.1: Left to right : EMD and M3-EMD , M1-EMD , M2-EMD

Figure 3.1 shows the different EMD variants that we obtain by extracting different sets of IPs, for a given sifting iteration. As illustrated in the leftmost diagram of Figure 3.1, in standard EMD, the extrema locations (x-coordinates / time-instants), and the amplitude (y-coordinates) of the *inner residue* at those locations, serve as the IPs (x-coordinates,y-coordinates) for every sifting iteration. In this work, we experiment with three different signals - the *speech* signal (M1-EMD) , the *outer residue* signal (M2-EMD), and the *inner residue* signal (EMD and M3-EMD) - for obtaining the IPs. To increase the number of IPs, effectively, with respect to the standard EMD process, an i^{th}, $i = 1, 2, 3, 4,$

order differentiation is done on the signal from which the IPs locations (x-coordinates) are being extracted. The differentiation of any signal, $x(t)$, is a high-pass filtering operation.

$$\frac{d^i}{dt^i}x(t) \leftrightarrow (j2\pi f)^i X(f)$$

The component $(j2\pi f)^i$ causes the differentiated signal to have a much enhanced higher-frequency spectrum, creating more fluctuations, and hence extrema. As such, the extrema locations of the differentiated signal are taken as x-coordinates of the IPs. This process is denoted as **Di, i=1,2,3,4**. Thus, the notation **Method (Di), i=1,2,3,4**, is used in this work to cumulatively indicate the method used for decomposition and that for extracting the IPs. For example, M1-EMD (D2) indicates that M1-EMD has been used to extract the IMFs from the speech signal, which has been twice differentiated for obtaining the IPs. To clarify the processes, we first re-represent the pseudocode for the EMD algorithm (presented in Section 2.1) in a manner which would explain the objective of our experiments. The modifications done to the EMD algorithm for the modified EMD processes - M1-EMD (Di), M2-EMD (Di), and M3-EMD (Di) - are then presented.

Pseudocode for EMD : Let $s(t)$ be a continuous-time speech signal.

•(**i**) Let $r_0(t) = s(t)$. We subject an *outer residue*, $r_{k-1}(t)$, to a *sifting process* to obtain an IMF, $h_k(t)$, and another *outer residue*, $r_k(t)$, from it. In other words, the k^{th} sifting process decomposes the $(k-1)^{th}$ outer residue, $r_{k-1}(t)$, into the k^{th} IMF, $h_k(t)$, and the k^{th} *outer residue*, $r_k(t)$.

The *sifting process* for EMD is given as :

Let $h^0_{k-1}(t) = r_{k-1}(t)$. Repeat the following steps for each *sifting iteration*. Let η represent the sifting iteration index, where $\eta = 1, 2, ..., N$.

⋆(**a**) **Obtain the x-coordinates of the IPs to construct the maxima and the minima envelope, i.e., obtain t_{max} and t_{min}, respectively.**

⋆(**b**) Obtain the y-coordinates of the IPs from the *inner residue* signal, $h^{\eta-1}_{k-1}(t)$.

$$y_{max} = h^{\eta-1}_{k-1}(t_{max}) \ , \ y_{min} = h^{\eta-1}_{k-1}(t_{min})$$

⋆(**c**) Create the maxima envelope $e_{max}(t)$ using cubic spline interpolation, with the IPs as $\{t_{max} \ , \ y_{max}\}$. Create the minima envelope $e_{min}(t)$ using cubic spline interpolation, with the

IPs as $\{t_{min}\ ,\ y_{min}\}$. Deduce the mean envelope $e_m(t)$ as,

$$e_m(t) = \frac{e_{max}(t) + e_{min}(t)}{2}$$

$\star$**(d)** $h_{k-1}^{\eta}(t) = h_{k-1}^{\eta-1}(t) - e_m(t)$. Go to step (a). Stop when $\eta = N$.

•**(ii)** Set $h_k(t) = h_{k-1}^{N}(t)$. Obtain $r_k(t) = r_{k-1}(t) - h_k(t)$.

•**(iii)** Go to step (i). Ideally, the decomposition is to be stopped when the *outer residue* takes the form of a trend, i.e., the number of extrema in $r_k(t)$ is 2 or less [18, 86, 87]. Practically, however, the decomposition may be stopped when a user-defined maximum number (M) of IMFs has been extracted, as shown in Figure 3.1.

$$s(t) = r_M(t) + \sum_{k=1}^{M} h_k(t) = \sum_{k=1}^{M+1} h_k(t) \tag{3.1}$$

For a discrete-time digital speech signal, $s(n)$, appropriate discrete-time operations would replace the continuous-time operations, and the decomposition may be represented as,

$$s(n) = r_M(n) + \sum_{k=1}^{M} h_k(n) = \sum_{k=1}^{M+1} h_k(n) \tag{3.2}$$

By changing the way the IPs are being obtained, i.e., by modifying step (a) of the sifting process, we obtain the following EMD variants.

EMD : The x-coordinates of the IPs are obtained from the *inner residue* signal in every sifting iteration.

$$z(t) = h_{k-1}^{n-1}(t)\ ,$$
$$t_{max} = \{t:\ \frac{d}{dt}z(t) = 0,\ \frac{d^2}{dt^2}z(t) < 0\}\ ,\ t_{min} = \{t:\ \frac{d}{dt}z(t) = 0,\ \frac{d^2}{dt^2}z(t) > 0\}$$

M1-EMD (Di) : The x-coordinates of the IPs are obtained from the *differentiated speech* signal. Hence, they may be estimated only once for the entire EMD process, and stored in memory. For a given Di, i = 1,2,3,4, we have,

$$z(t) = \frac{d^i}{dt^i}s(t)\ ,$$
$$t_{max} = \{t:\ \frac{d}{dt}z(t) = 0,\ \frac{d^2}{dt^2}z(t) < 0\}\ ,\ t_{min} = \{t:\ \frac{d}{dt}z(t) = 0,\ \frac{d^2}{dt^2}z(t) > 0\}$$

M2-EMD (Di) : The x-coordinates of the IPs are obtained from the *differentiated outer residue*

signal. Hence, they may be estimated only once for one entire sifting process, and stored in memory. For a given Di, i = 1,2,3,4, we have,

$$z(t) = \frac{d^i}{dt^i} r_{k-1}(t) \ ,$$
$$t_{max} = \{t : \ \frac{d}{dt}z(t) = 0, \ \frac{d^2}{dt^2}z(t) < 0\} \ , \ t_{min} = \{t : \ \frac{d}{dt}z(t) = 0, \ \frac{d^2}{dt^2}z(t) > 0\}$$

M3-EMD (Di) : The x-coordinates of the IPs are obtained from the *differentiated inner residue* signal in every sifting iteration. For a given Di, i = 1,2,3,4, we have,

$$z(t) = \frac{d^i}{dt^i} h_{k-1}^{n-1}(t) \ ,$$
$$t_{max} = \{t : \ \frac{d}{dt}z(t) = 0, \ \frac{d^2}{dt^2}z(t) < 0\} \ , \ t_{min} = \{t : \ \frac{d}{dt}z(t) = 0, \ \frac{d^2}{dt^2}z(t) > 0\}$$

For discrete-time digital signals, appropriate discret-time operations would substitute the continuous-time operations for the above-mentioned processes. It can be seen from the pseudocode of M3-EMD (Di) that if the order of differentiation, i, is kept to 0, i.e, the *inner residue* signal, $h_{k-1}^{n-1}(t)$, is not differentiated at all, M3-EMD actuates to the standard EMD process. To avoid the generation of trend-like low-frequency signals, we limit the decomposition to ten levels ($M = 9$) only, which results in nine IMFs and the final residue signal as the components of the speech signal.

3.3 Effect of changing the IPs on EMD

To evaluate the effects of the proposed EMD variants, we examine the filterbank nature exhibited by them. For this purpose, we calculate the *mean frequency* and the *log-normalized mean frequency* of the IMFs generated by EMD and its variants. The mean frequency of IMF$_k$ of a speech signal, $s(n)$, gives an indication of the dominant frequency reflected in the IMF, and is given by equation (2.28).

$$F_k^m = \int_{f=0}^{F_s/2} \frac{f \times S_k(f)df}{\int_{f=0}^{F_s/2} S_k(f)df} \ , \ k = 1, 2, .., M + 1 = 10 \ , \tag{3.3}$$

where $S_k(f)$ represents the power spectrum (squared magnitude spectrum) of IMF$_k$, and F_s is the *sampling frequency* of the speech signal. The log-normalized mean frequency (nF_k^m) of IMF$_k$ is derived as,

$$nF_k^m = \log_2(F_k^m/F_1^m) \ , \tag{3.4}$$

where F_1^m is the mean frequency of IMF$_1$.

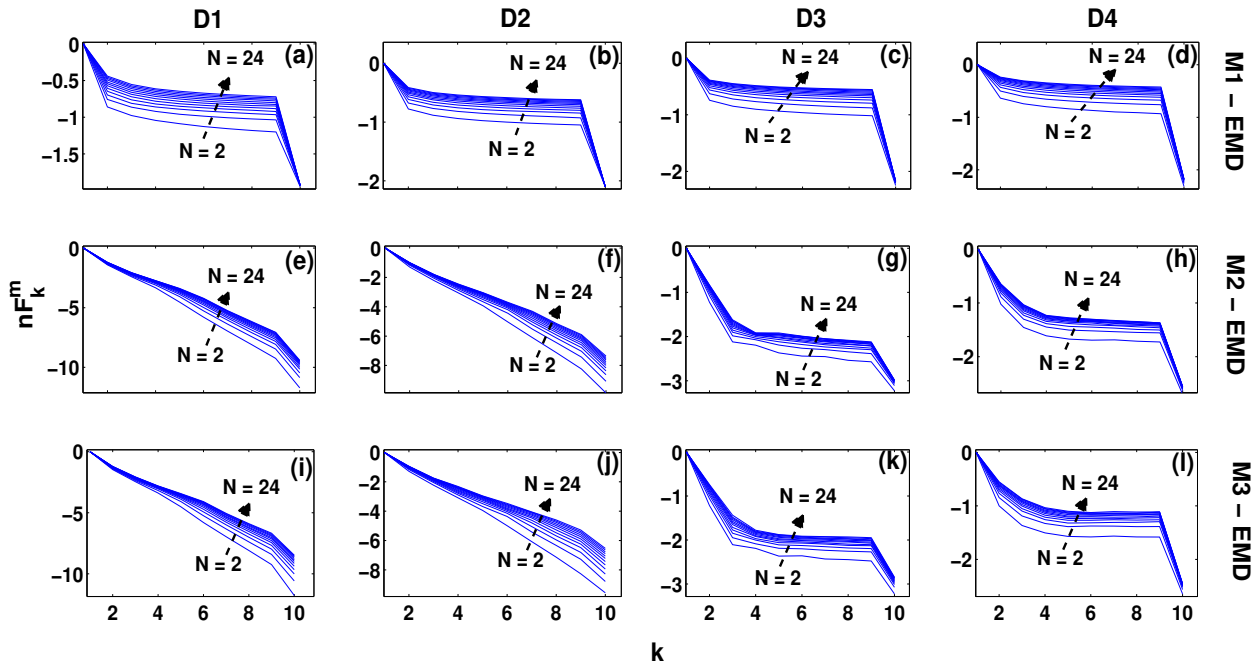


Figure 3.2: **Plots of the medians of the log-normalized mean frequencies for 1000 files of the CMU-Arctic database. Each row signifies an EMD-variant , and each column a method of extracting IPs. The number of sifting iterations, N, is varied in steps of 2, from 2 to 24. (a) M1-EMD (D1) ; (b) M1-EMD (D2) ; (c) M1-EMD (D3) ; (d) M1-EMD (D4) ; (e) M2-EMD (D1) ; (f) M2-EMD (D2) ; (g) M2-EMD (D3) ; (h) M2-EMD (D4) ; (i) M3-EMD (D1) ; (j) M3-EMD (D2) ; (k) M3-EMD (D3) ; (l) M3-EMD (D4).**

Figure 3.2 plots the median values of the log-normalized mean frequencies of the ten IMFs (the final residue is taken as IMF_{10}) generated by using the different EMD variants for different Di, i = 1,2,3,4. The plots are generated from 1000 files, i.e, 200 random speech signals (utterances) from each of the five speakers of the CMU-Arctic database [158]. The signals are downsampled to $F_s = 8$ kHz prior to their decomposition. Each plot in Figure 3.2 consists of 12 curves, corresponding to the *number of sifting iterations* (N) used, where N is varied from 2 to 24, in steps of 2.

For an *ideal dyadic filterbank*, the plot of nF_k^m vs. k, where $k = 1, 2, .., 10$, should be a straight line with a slope of -1. The plot of nF_k^m vs. k, for the ten IMFs derived from EMD, is shown in Figure 3.3(a), which reflects a close approximation to the ideal filterbank. An examination of the plots of Figure 3.2 shows that only the plots of Figure 3.2(e),(f),(i), and (j), corresponding to M2-EMD (D1), M2-EMD (D2), M3-EMD (D1), and M3-EMD (D2), respectively, exhibit a similar nature. This shows that out of the three EMD variants, only M2-EMD and M3-EMD are capable of maintaining the dyadic filterbank nature, like that of EMD. However, they can do so only if certain orders of differentiation are used for extracting the IPs. In essence, M2-EMD and M3-EMD work only if Di,

i ≤ 2 is used. The rest of the plots of Figure 3.2 show flat curves, contrary to the dyadic filterbank nature. M1-EMD completely fails to maintain the filterbank nature, irrespective of the process of extracting the IPs, as exhibited in Figure 3.2(a)-(d). The flatness of the curves between IMFs 2-9 of M1-EMD indicate that they represent similar signals, exhibiting similar dominant frequencies within them. The log-normalized mean frequency of the final residue or IMF_{10}, nF_{10}^{m}, has a sudden dip from the flat curves. This signifies that IMF_{10} is of significantly lower-frequency than the other IMFs - there is a discontinuity in the process. In a nutshell, M1-EMD (D1-D4) represents a complete breakdown of the EMD process. Similar to M1-EMD (D1-D4), the curves corresponding to M2-EMD (D3,D4) and M3-EMD (D3,D4) also exhibit flatness, with a sudden dip at IMF_{10}. They too represent the breakdown of the EMD process.

As has been discussed in Chapter 2, in the case of fGn, EMD exhibits a dyadic filterbank nature [109–111]. And this nature has been found to be best maintained when $N \approx 10$ [118, 119]. This characteristic also seems to hold true in the case of speech signals, as reflected in Figure 3.3(a). Further, the value of N - varied within reasonable limits, from 2 to 24 - doesn't seem to have a drastic effect on the decomposition, as seen in Figure 3.3(a). The curves in Figure 3.2 also reflect the same observation - whether the EMD variants succeeded or failed in maintaining the filterbank nature, the value of N did not have any significant effect on the same.

In Figure 3.3, we compare the mean frequencies of the IMFs generated by M2-EMD and M3-EMD (using D1 and D2 for extracting the IPs) with respect to that generated by EMD. As in Figure 3.2, each plot in Figure 3.3 consists of 12 curves, corresponding to different values of N (varied in steps of 2, from 2 to 24). Each curve in Figure 3.3(b)-(h) represents the median of the mean frequencies of the ten IMFs generated from the same 1000 speech files, from which the plots in Figure 3.2 were generated. The curves of Figure 3.3(b),(f) reflect the same story as that of Figure 3.2(a),(b) - M1-EMD fails in decomposing the speech signal. The curves in Figure 3.3(c) represent the mean frequencies corresponding to the curves in Figure 3.2(e). Similarly, Figure 3.3(g) corresponds to Figure 3.2(f). It can be observed from Figure 3.3(c),(e), and (g) that the IMFs of M2-EMD (D1) and M2-EMD (D2) have higher mean frequencies than that of EMD. Further, the IMFs of M2-EMD (D2) have a significantly higher mean frequency than that of M2-EMD (D1). This means that M2-EMD (D2) produces IMFs that better represent the higher-frequency spectrum of speech, compared to that of M2-EMD (D1) and EMD. As the mean frequencies are higher, the frequency separation between the

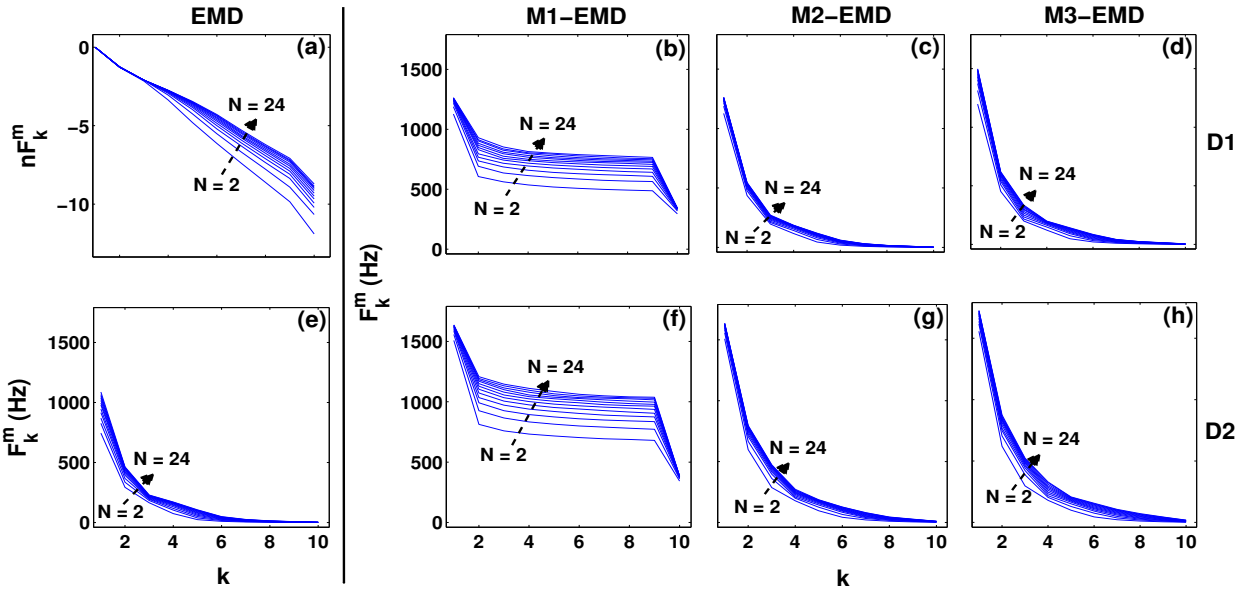


Figure 3.3: **Plots of the medians of the log-normalized mean frequencies and mean frequencies for 1000 files of the CMU-Arctic database. The number of sifting iterations, N, is varied in steps of 2, from 2 to 24. (a) log-normalized mean frequencies of the IMFs generated by EMD ; mean frequencies of the IMFs generated by (b) M1-EMD (D1) ; (c) M2-EMD (D1) ; (d) M3-EMD (D1) ; (e) EMD ; (f) M1-EMD (D2) ; (g) M2-EMD (D2); (h) M3-EMD (D2).**

IMFs of M2-EMD (D2) are also larger, and the possibility of mode-mixing lesser. The curves in Figure 3.3(d),(h) represent the mean frequencies corresponding to the curves in Figures 3.2(i),(j). A comparison of the curves of Figure 3.3(d),(e) and (h) shows that the IMFs of M3-EMD (D2) have much higher mean frequencies than that of M3-EMD (D1) and EMD. Similar to M2-EMD (D2), M3-EMD (D2) also gives a better representation of the higher-frequency content of the speech signal, producing IMFs that have larger frequency separation amongst them, with a lesser possibility of mode-mixing. Between the two, however, there is little to choose - M2-EMD (D2) and M3-EMD (D2) behave very similarly, as can be seen from Figure 3.3(g),(h).

Figure 3.4 shows the first five IMFs generated by EMD and its variants, using D2 for extracting the IPs, from a speech signal/utterance of the CMU-Arctic database, where $N = 10$ sifting iterations have been used for the decomposition. As is evident from the figure, in the case of M1-EMD (D2), the spline interpolation causes similar IMFs (IMF_2 and beyond) of low amplitude to be produced, thus supporting the flat curves of Figures 3.2 and 3.3. The lower-frequency content of the speech signal is almost entirely represented by the final residue (IMF_{10}), not shown in the figure. In fact, M1-EMD goes into an endless loop, if a maximum number of IMFs (M=9) is not specified. Also is evident from Figure

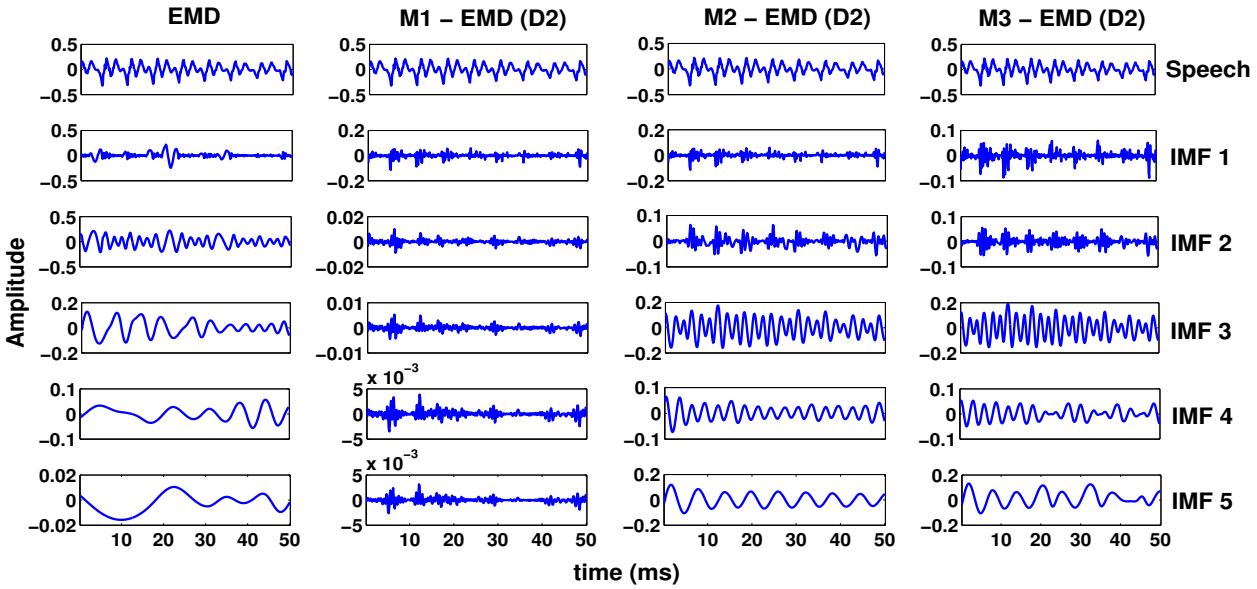


Figure 3.4: **The first five IMFs generated from an arbitrary speech file from the CMU-Arctic database, using EMD, M1-EMD (D2), M2-EMD (D2), and M3-EMD (D2) respectively, from left to right. Number of sifting iterations, $N = 10$, is used.**

3.4 that M2-EMD (D2) and M3-EMD (D2) produce IMFs with lesser *mode-mixing* than EMD. In the case of EMD, IMF_1 contains mainly higher-frequency oscillations, which are interrupted in between by sinusoid-like signals of much lower-frequency. The other IMFs represent sinusoid-like signals, whose frequencies differ at different times. None of them seem to represent a particular dominant frequency scale. In the case of M2-EMD (D2) and M3-EMD (D2), IMF_1 is much less corrupted. All the IMFs of M2-EMD (D2) and M3-EMD (D2) seem to represent frequency scales, which are particular to them only. There is very little intermingling of the different oscillations represented by them. Again, IMFs 1,2 and 3 of M2-EMD (D2) and M3-EMD (D2) exhibit oscillations that are not represented at all by EMD. Further, as a benefit of reduced mode-mixing, the lower-frequency IMFs (IMFs 3-5) of M2-EMD (D2) and M3-EMD (D2), are much more sinusoidal. These observations ascertain that M2-EMD (D2) and M3-EMD (D2) have the ability to extract better time-domain components of the speech signal, compared to standard EMD.

3.3.1 Simulation of the sifting process

Figure 3.5 gives an illustration of the sifting process for EMD, M2-EMD (D2), and M3-EMD (D2) using a small voiced speech segment of an arbitrary speech signal/utterance from the CMU Arctic database. The first row of each column shows the speech signal, $s(t)$, to be subjected to a

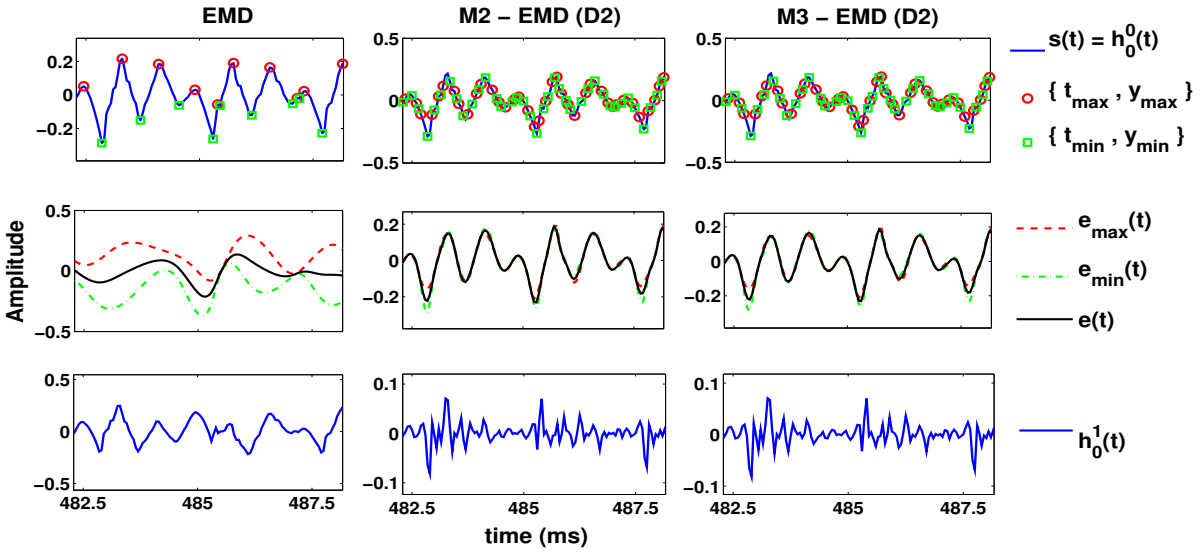


Figure 3.5: **Illustration of the sifting process of EMD, M2-EMD (D2), and M3-EMD (D2), using a portion of an arbitrary speech signal (utterance) from the CMU-Arctic database.**

sifting process, to obtain the first IMF. Hence, it is the *first outer residue*, and also the *first inner residue* of the *first sifting process*. $s(t) = r_0(t) = h_0^0(t)$.

The IPs for constructing the maxima envelope, $e_{max}(t)$, and the minima envelope, $e_{min}(t)$, are indicated by bubbles and square boxes, respectively. The maxima envelope, minima envelope, and the mean envelope, $e(t)$, are shown in the second row. It can be observed that the mean envelope almost follows the shape of the first *inner residue*, $h_0^0(t)$, in the case of the EMD variants, unlike in the case of standard EMD. This is due to the large number of IPs that is employed in the EMD variants. Hence, when this mean envelope is subtracted from $h_0^0(t)$, the *fine fluctuations* present in $h_0^0(t)$ are revealed, as is represented by $h_0^1(t)$. $h_0^1(t) = h_0^0(t) - e(t)$

The same processes are again applied on $h_0^1(t)$, in the second *sifting iteration*, to obtain the second *inner residue*, $h_0^2(t)$, and so on. It is evident from the figure that increasing the number of IPs enables the mean envelope to follow the shape of the *inner residue* signals, thereby revealing the fine fluctuations embedded in them.

As mentioned before, we have used notations for continuous-time signals for simplicity of explanation and representation, even though, we are, in fact, processing discrete-time digital signals.

3.4 Convergence and Robustness of the EMD variants

One of the properties of EMD is that it provides a *complete* decomposition [18, 87] of the signal, i.e., the sum of the components derived from it gives us back the exact same signal, as reflected in equations (3.1) and (3.2). We may verify the same practically using the speech signal shown in Figure 3.4. The speech signal, $s(n)$, in this case, is allowed to be decomposed freely without putting any constraint on the number of IMFs ($M = \infty$), i.e, the signal is decomposed till the trend is obtained where the number of extrema is only 2 or less. N is kept fixed to 10. Let M^{uc} denote the number of components obtained from the decomposition, including the final residue.

$$s(n) = r_{M^{uc}-1}(n) + \sum_{k=1}^{M^{uc}-1} h_k(n) = \sum_{k=1}^{M^{uc}} h_k(n) \tag{3.5}$$

The *reconstruction error* of the speech signal is then given by,

$$r_e^{uc} = ||s(n) - \sum_{k=1}^{M^{uc}} h_k(n)||_2 \tag{3.6}$$

Let M_{trend}^{uc} denote the number of very low-frequency trend-like IMFs obtained from the unconstrained decomposition. Considering that the *pitch frequency* of normal human speech doesn't fall below ≈ 50 Hz [2, 13, 14], we define, in the case of speech, the trend-like components as the components whose mean frequencies are less than 50 Hz.

$$\{h_k(n) \; is \; a \; trend \mid F_k^m < 50 \text{ Hz} \; , \; k = 1, 2, ..., M^{uc}\} \tag{3.7}$$

Table 3.1 shows the r_e^{uc}, M^{uc} and M_{trend}^{uc} values for the speech signal shown in Figure 3.4, as decomposed by EMD, M2-EMD (D2), and M3-EMD (D2). Though, theoretically, the r_e^{uc} value should be 0, practically a small error is observed, owing to machine precision. Also, it is observed that a few more components are extracted by the EMD variants than EMD. Correspondingly, the number of trend-like IMFs are more, as well as the number of *possibly useful components*, given by the $\{M^{uc} - M_{trend}^{uc}\}$ values. Thus, the EMD variants doesn't produce an excessive number of components, while providing a better decomposition - they converge equally well as EMD.

Table 3.1: A comparison of the reconstruction error (r_e^{uc}), number of components derived M^{uc}, and the number of trend-like components (M_{trend}^{uc}) extracted from the speech signal used in Figure 3.4. EMD, M2-EMD (D2) and M2-EMD (D2) are used for decomposing the speech signal, with $M = \infty$ and $N = 10$.

EMD Variant →	EMD	M2-EMD (D2)	M3-EMD (D2)
r_e^{uc}	1.4974×10^{-13}	6.1558×10^{-10}	4.3268×10^{-10}
M^{uc}	17	21	23
M_{trend}^{uc}	10	12	14
$M^{uc} - M_{trend}^{uc}$	7	9	9

To analyze the performance of the EMD variants with respect to EMD, under noisy conditions, we consider the speech signal depicted in Figure 3.4, corrupted by varying levels of *White noise* [159]. The corrupted speech signal is then decomposed by EMD, M2-EMD (D2), and M3-EMD (D2), with $M = 9$ and $N = 10$. Let $s^x(n)$ represent the speech signal after being corrupted by a White noise sequence, $w^x(n)$, such that the SNR is x dB. Thus,

$$s^x(n) = s(n) + w^x(n) \tag{3.8}$$

$$s^x(n) = r_9^x(n) + \sum_{k=1}^{9} h_k^x(n) = \sum_{k=1}^{10} h_k^x(n) \tag{3.9}$$

where $h_k^x(n)$ denotes the k^{th} IMF generated from the noisy speech signal. The *similarity error* of IMF$_k$ with respect to the White noise signal is then calculated as,

$$e_k^s = \frac{||h_k^x(n) - w^x(n)||_2}{||w^x(n)||_2} \tag{3.10}$$

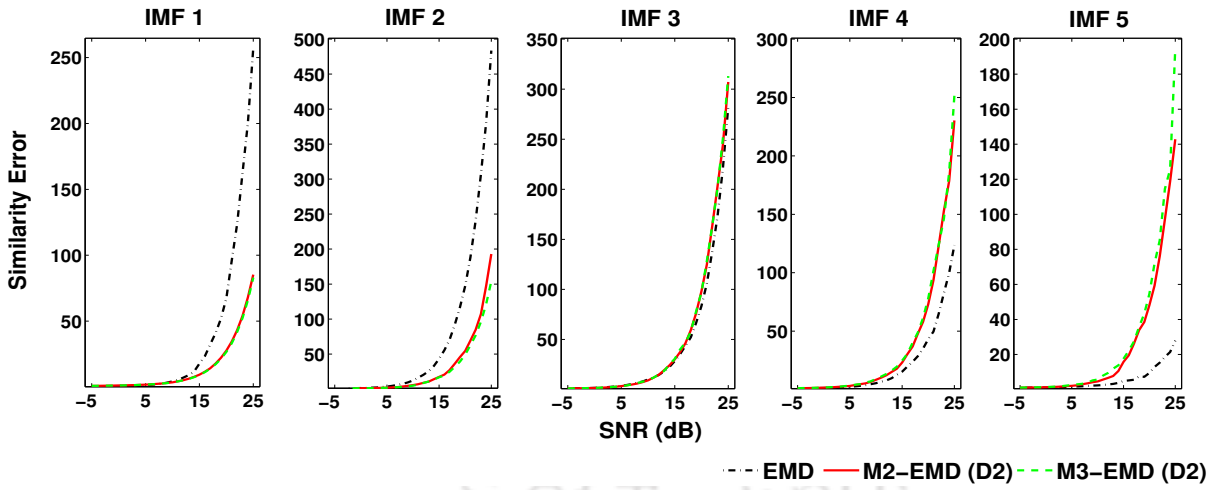


Figure 3.6: **The similarity error between the White noise signal and the first five IMFs derived from the corrupted speech signal, at varying SNRs. The IMFs are derived using EMD, M2-EMD (D2), and M3-EMD (D2) on the corrupted speech signal.**

Figure 3.6 plots the *similarity errors* of the first five IMFs generated by EMD, M2-EMD (D2) and M3-EMD (D2), for the speech signal corrupted by different levels of White noise ($x = -5, 0, ..., 25$). It is evident from the figure that IMFs 1 and 2, generated by the EMD variants, have a much stronger similarity with the White noise signal, compared to that of EMD. Subsequently, the similarity decreases as the IMF order increases. Thus, in the case of the EMD variants, the first few IMFs are able to extract the noise embedded in the signal, whereas the latter IMFs represent the characteristics of the speech signal. In the case of EMD, however, the effect of White noise percolates across even the latter

IMFs, as is reflected in the much lesser *similarity errors* of IMF_4 and beyond. This shows that the EMD variants provide a much better decomposition of the speech signal, even under the effect of noise.

3.5 Comparison of the EMD variants with other time-frequency analysis methods

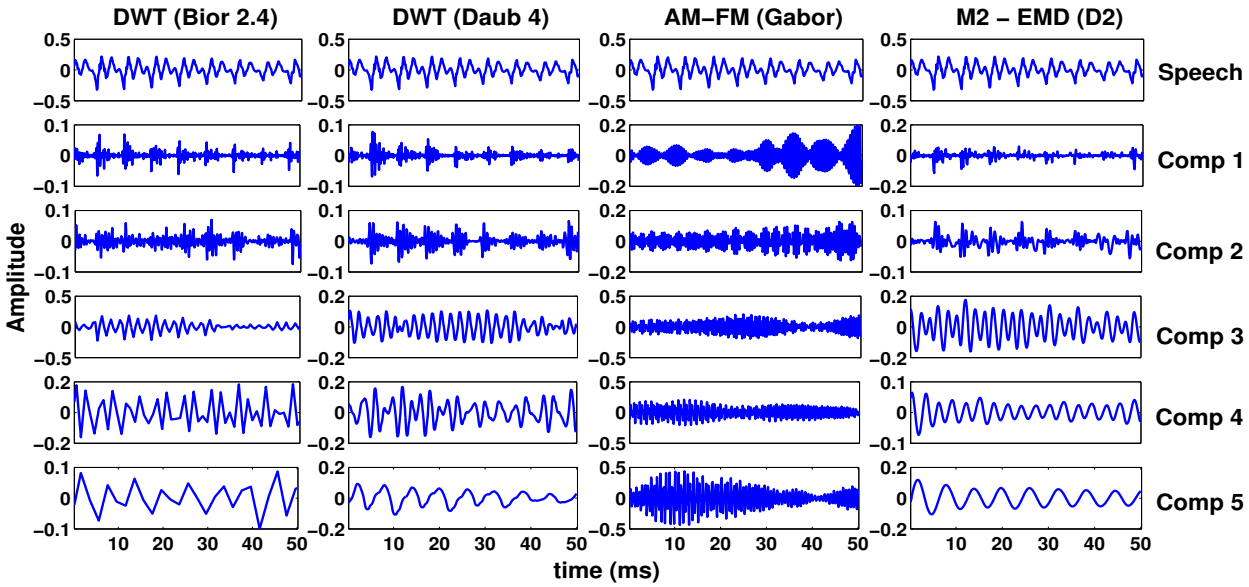


Figure 3.7: **The first column shows the first five time-domain components obtained by Inverse DWT of the first five detail coefficients, obtained by DWT of a speech signal, using 'Biorthogonal 2.4' wavelet. The second column shows the same, but for 'Daubechies 4' wavelet. The third column shows the first five components (in decreasing order of frequency) derived by AM-FM analysis of the speech signal, using a 20-filter Gabor filterbank, each filter having an effective bandwidth of 400 Hz. The fourth column shows the first five IMFs obtained from M2-EMD (D2) of the speech signal.**

For the purpose of comparing the EMD variants with DWT and conventional AM-FM analysis, we consider the case of a discrete-time digital speech signal (F_s = 8 kHz) taken from the CMU-Arctic database. The speech signal is decomposed by 10-level DWT, using the 'Daubechies-4' and 'Biorthogonal-2.4' wavelets. The first five time-domain detail components, reconstructed by Inverse DWT of the first five detail coefficients, for the two wavelet types, are shown in the first two columns of Figure 3.7, respectively. The third column of Figure 3.7 shows the first five high-frequency components obtained using a small Gabor filterbank of 20 uniformly spaced filters (generally more than 40 filters are used) in the Hz scale, each having an effective bandwidth of 400 Hz. The last column of Figure 3.7 shows the first five IMFs of the speech signal, derived using M2-EMD (D2), which reflects its superiority over DWT and traditional AM-FM analysis. The IMFs of M3-EMD (D2) are similar to that of M2-EMD (D2), and hence not shown here. If these IMFs are able to reflect the formants

structure of speech effectively, then they could be extremely useful for speech processing applications, which is the subject of investigation in the next section.

3.6 Distribution of the formants amongst the IMFs

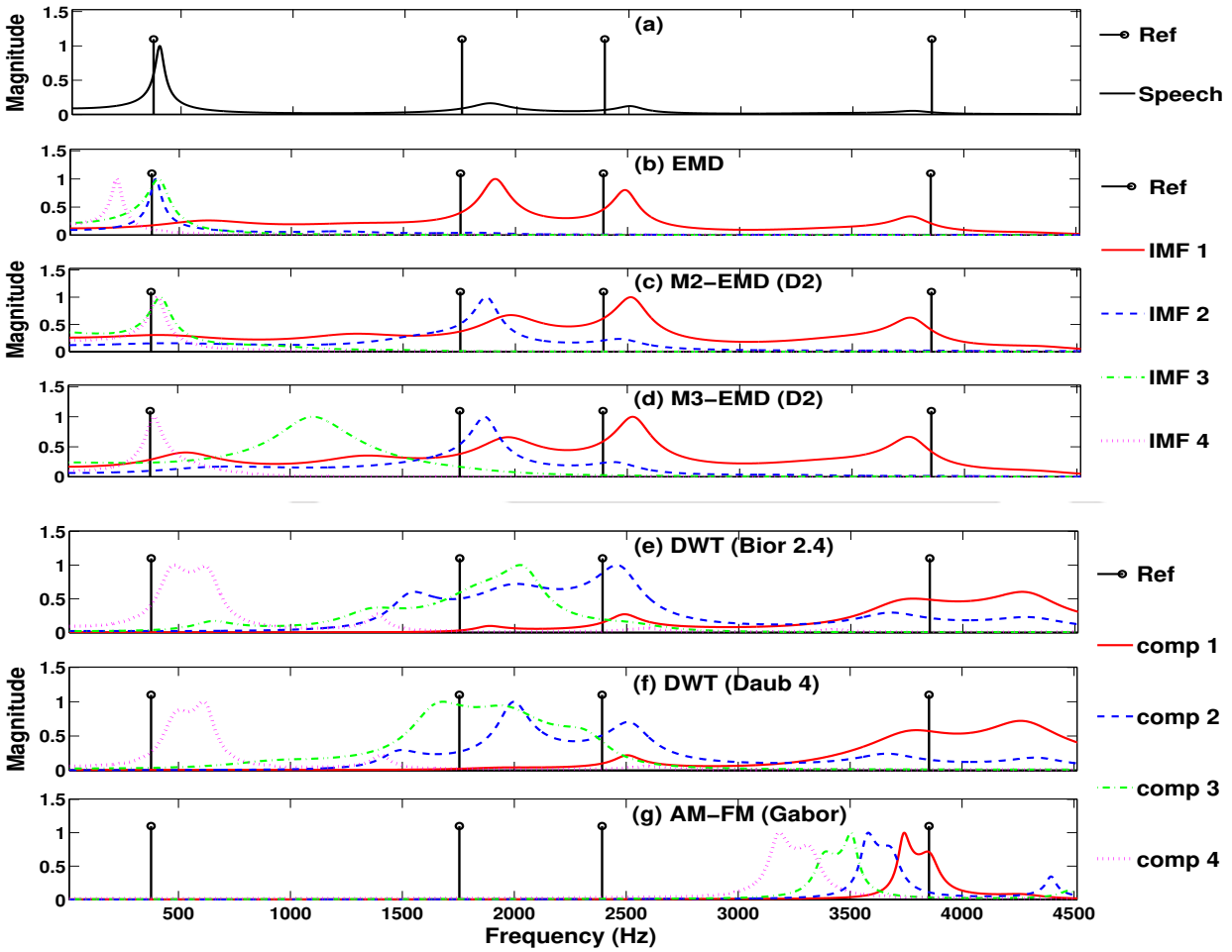


Figure 3.8: **The normalized magnitude spectra of the vocal-tract filters estimated by LP analysis of (a) speech ; IMFs 1-4 obtained from (b) EMD, (c) M2-EMD (D2), (d) M3-EMD (D2) ; the first four high-frequency components obtained from (e) DWT using Biorthogonal 2.4 wavelet, (f) DWT using Daubechies 4 wavelet ; (g) AM-FM analysis using a 40 filter Gabor filterbank with effective bandwidth of 400 Hz. Vertical stem lines indicate reference formant frequencies obtained from the VTR Formants database.**

To understand the distribution of the spectral content of speech in its IMFs, we study the distribution of the vocal tract resonances in them. For this purpose, we conduct an LP analysis of the voiced segments of the speech signal, and the first four IMFs derived from it using EMD and its variants. Figure 3.8(a)-(d) shows the normalized filter magnitude spectra of a 16 kHz voiced speech segment (20 ms), taken from the TIMIT database [128], and its corresponding IMFs. The spectra

are shown only up to around 4.5 kHz (within which the first four principal formants generally exist) for better visualization. The filter spectra are derived using 24-order LP analysis on the signals. The IMFs are obtained from EMD, M2-EMD (D2) and M3-EMD (D2). As seen from Figure 3.8(a), the speech signal exhibits resonance peaks in the vicinity of the formants, where the reference formant frequencies of the first four formants, as indicated by vertical lines, have been taken from the VTR Formants database [129]. In the case of EMD, as seen in Figure 3.8(b), IMF_1 exhibits strong peaks (much stronger than those exhibited by the speech signal) around the second, third and fourth formants. However, the first formant is carried mainly by IMF_2 and IMF_3. This shows that even IMF_3 can carry the formant information of voiced speech, resulting from an overlap of information between the two IMFs. In the case of M2-EMD (D2), Figure 3.8(c), the filter spectrum of IMF_1 has strong peaks around the third and fourth formants. The filter spectrum of IMF_2 has a strong peak around the second formant. The filter spectra of IMFs 3 and 4 both have strong peaks around the first formant, which reflects the sharing of information between them. In the case of M3-EMD (D2), Figure 3.8(d), IMFs 1 and 2 exhibit filter spectra similar to that of IMFs 1 and 2 of M2-EMD (D2), respectively. However, IMF_3 doesn't exhibit any strong peaks around any of the reference formants. IMF_3, thus, represents a component of the voiced speech signal which may not be very useful for analysis. Again, just as in the case of M2-EMD (D2), IMF_4 of M3-EMD (D2) manifests the first formant.

The plots in Figure 3.8(a)-(d) depict how evenly the spectral content of the voiced speech signal is distributed amongst the IMFs of M2-EMD (D2) and M3-EMD (D2), compared to the case of EMD. The stronger peaks also provide a possibility of more accurate formants estimation from the filter spectra of the IMFs, compared to the spectrum of the speech signal alone. At this point, we may, for the sake of comparison, consider the filter spectra of the first four high-frequency components obtained from DWT, and AM-FM analysis, of the voiced speech segment. Figure 3.8(e) shows the filter spectra corresponding to the first four time-domain detail components of the same voiced speech segment, obtained by DWT using Biorthogonal 2.4 wavelet. Figure 3.8(f) shows the same for Daubechies 4 wavelet. Figure 3.8(g) shows the filter spectra of first four high-frequency components, obtained by Gabor filtering the speech segment, with a 40-filter Gabor filterbank, each filter having a bandwidth of 400 Hz. The high-frequency components, in this case, are limited to the Gabor filters with center frequencies below 4 kHz. A comparison of the plots of Figure 3.8(c),(d) with that of Figure 3.8(e)-(g)

shows how inefficient DWT and AM-FM analysis are, compared to the EMD variants, in capturing the formants information in its components.

In line with the above observations, we now need to quantify the contribution of each of the first four IMFs to the actual filter spectrum of the speech signal. For this purpose, we calculate the *Identification Rate* (IR) of the four principal formants, as estimated from the IMFs. The formant frequencies are identified as the peaks of the filter magnitude spectra of the signals. A 24-order LP analysis is used to generate the filter spectra of the signals. The IR, depicted in Figure 3.10, is defined as,

Identification Rate (IR) : Let Fi, i=1,2,3,4, be a formant, with its reference formant frequency F_i^r, $i = 1, 2, 3, 4$, corresponding to a voiced frame of a given speech signal. The formant is considered identified in that voiced frame, if there is atleast one estimated formant in the range of $F_i^r \pm \min\{20\ \%\ of\ F_i^r\ ,\ 250\ \text{Hz}\}$. The percentage (%) of voiced frames in which Fi is identified, gives the IR of Fi for the speech signal.

Database : The experiments are performed on 512 speech files or utterances (F_s = 16 kHz), of the TIMIT corpus, for which the reference formant frequencies are available in the VTR Formants database [129].

Table 3.2: Mean of the identification rate IR (%) of the four principal formants, denoted as F1, F2, F3, and F4, estimated from 512 files of TIMIT database. The formants are estimated from the speech signal, and its first four IMFs, derived using EMD, M2-EMD (D2) and M3-EMD (D2). The speech signal is pre-emphasized prior to LP analysis.

Method →	EMD				M2-EMD (D2)				M3-EMD (D2)			
Formant →	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
Speech	80	90	88	82	–	–	–	–	–	–	–	–
IMF 1	37	86	86	77	26	82	87	79	24	65	85	80
IMFs 1-2	80	92	88	83	50	96	91	85	40	95	92	85
IMFs 1-3	90	92	89	85	83	97	92	88	75	97	93	88
IMFs 1-4	91	92	89	86	93	97	93	90	91	97	93	90

Table 3.2 shows the IRs of the formants, as estimated from the LP filter spectrum of the speech signal, and that of each of its first four IMFs derived from EMD, M2-EMD (D2), and M3-EMD (D2). The speech signal is pre-emphasized by a high-pass filter, $H_{pre}(z) = 1 - 0.98z^{-1}$, prior to LP analysis. The IMFs are not pre-emphasized. Frames of 20 ms, with a frameshift of 10 ms, are used in the analysis. Only voiced frames, which are considered as the frames of the speech signal having energy greater than 10 % of the maximum frame energy, are used for analysis. To visualize the contribution

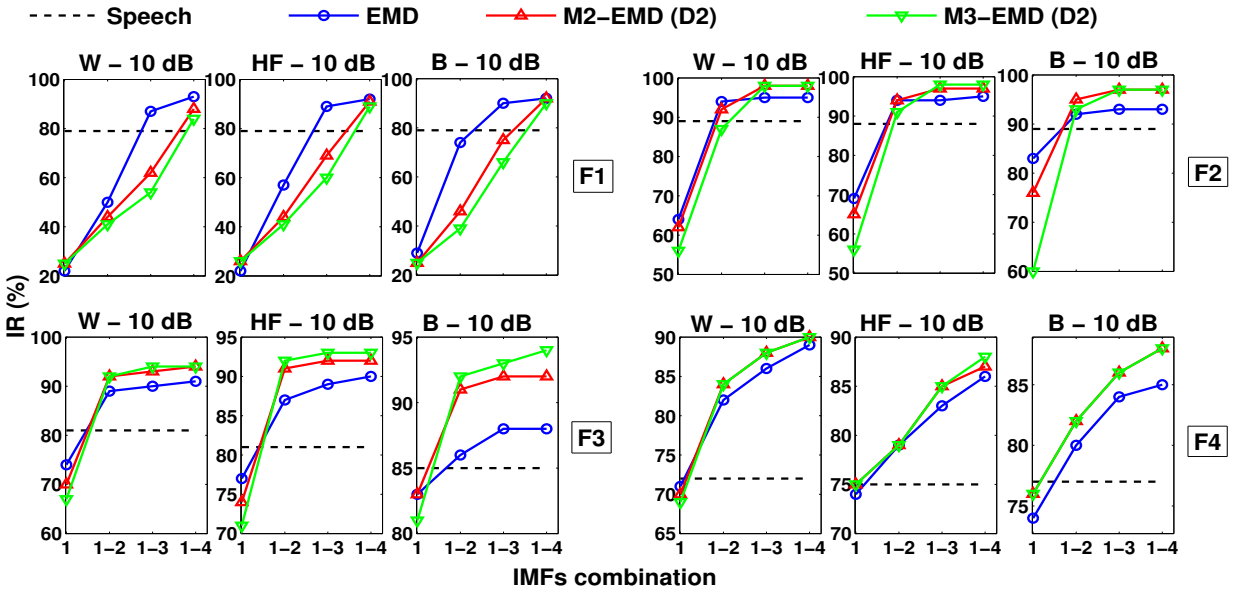


Figure 3.9: Mean of the identification rate IR (%) of the four principal formants, estimated from 512 files of TIMIT database. The first four formants, denoted as F1, F2, F3 and F4, are estimated from the speech signal, and its first four IMFs, derived using EMD, M2-EMD (D2) and M3-EMD (D2). The speech signal is pre-emphasized prior to LP analysis. The speech signals are corrupted by White (W), HFchannel (HF) and Babble (B) noise, at an SNR of 10 dB.

of each IMF towards a particular formant, the number of IMFs is increased progressively from 1 to 4. It can be seen from Table 3.2 that the IR of the fourth formant (F4), as estimated from IMF_1, is comparable to that estimated from the speech signal, for all the methods. In the case of EMD, the IR of the first formant (F1), for clean conditions, increases sharply from 37 % to 80 %, when IMF_2 is added to the analysis. The inclusion of IMF_3 further increases the IR to 90 %. IMF_4, however, doesn't contribute significantly to the estimation of any of the formants. In the case of M2-EMD (D2), IMFs 3 and 4, both significantly enhance the IR for F1. IMF_2 provides significant gains in the estimation of both the first and the second formants, F1 and F2, respectively. The higher formants are mainly carried by IMF_1. Similar observations can be made in the case of M3-EMD (D2). Figure 3.9 shows the distribution of the formants, as described above, but for the speech signals corrupted by moderate level (SNR = 10 dB) of noise, taken from the NOISEX-92 database [159]. It can be seen from the figure that under the influence of high-frequency noise (HFchannel and White), the third and fourth formants, F3 and F4, are almost completely manifested in IMFs 1 and 2, whereas F1 and F2 are strongly manifested in IMFs 3 and 4. This is particularly seen in the case of the EMD variants, which suggests that the reduction in the spectral slope of voiced speech, caused by the addition of high-frequency noise, helps in better spectral segregation. The addition of low-frequency

noise (Babble), has an opposite effect, and inhibits a better spectral segregation, particularly in the case of EMD.

Table 3.2 and Figure 3.9 show how a better distribution of the formants is given by M2-EMD (D2) and M3-EMD (D2) compared to EMD. The enhancement in the formants estimation performance of the four principal formants, with the addition of more IMFs, shows that the IMFs of the EMD variants have less overlapping filter spectra - if one IMF doesn't carry the information of a formant, another does. Further, all the methods show enhancement, though of differing degrees, in the identification of the formants, when more IMFs are included in the analysis - the IR is significantly better when all the four IMFs are cumulatively considered, compared to the speech signal alone. This proves that the formants structure is not just distributed between the first two IMFs, even for EMD, particularly under noisy conditions.

3.7 Results and Discussion

Having observed how the vocal tract spectrum of the speech signal is evenly distributed amongst its first four IMFs derived from the EMD variants, compared to that of EMD, we now demonstrate the merit of the improved decompositions obtained from them, by quantifying their performance in the task of formants estimation of voiced speech. For this purpose, two metrics - *Identification Rate* (IR) and *Identification Error* (IE) - are used. IR has already been defined in the previous section. IE, illustrated in Figure 3.10, is defined as -

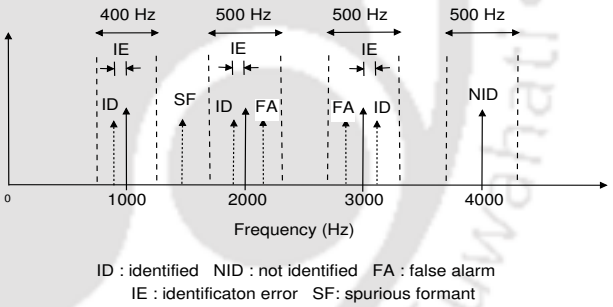


Figure 3.10: **Illustration of the metrics used for evaluation. The vertical solid arrows indicate the reference formant frequencies at 1000 Hz, 2000 Hz, 3000 Hz and 4000 Hz. Shorter dotted arrows indicate the estimated formants. The allowed range of search for the estimated formants is indicated by dashed vertical lines.**

Identification Error (IE) : Let Fi, i=1,2,3,4, be a formant, with its reference formant frequency F_i^r, $i = 1, 2, 3, 4$, corresponding to a voiced frame of a given speech signal. Let the formant estimate, in the range of $F_i^r \pm \min\{20\ \%\ of\ F_i^r\ ,\ 250\ \text{Hz}\}$, and closest to F_i^r, be denoted as F_i^{est}. The absolute difference in frequency between the reference and the closest estimate gives the identification error for that formant, in that voiced frame. The average of the identification error for that formant, for all the voiced frames of the speech signal, gives the identification error of Fi, for that speech signal.

$\text{IE}_i = \frac{1}{N_v} \sum_{frame=1}^{N_v} |F_i^r - F_i^{est}|$, $i = 1, 2, 3, 4,$

where N_v represents the total number of voiced frames in the speech signal.

Database and Method : The performance metrics, IR (%) and IE (Hz), are evaluated on 512 speech files/utterances (F_s = 16 kHz), of the TIMIT corpus, for which the reference formant frequencies are available in the VTR Formants database [129]. As has been illustrated earlier, a 24-order LP analysis is used on the signal (whether the speech signal or its IMF) to obtain the magnitude spectrum of its LP filter, whose peaks provide the estimates of the formants. The LP analysis is done on short segments or frames of 20 ms, with the frames being shifted by 10 ms (50 % overalp between the frames). Out of all the frames, only the voiced frames are considered for performance evaluation. The voiced frames are considered as those frames of the speech signal whose energy is atleast 10 % of the maximum frame energy of the speech signal. If the signal is the speech signal, it is pre-emphasized prior to LP analysis. However, if the signal is an IMF, LP analysis is done directly on it. The IMFs are obtained from EMD, M2-EMD (D2), and M3-EMD (D2), with $N = 10$ and $M = 9$. Only the first four IMFs are considered for estimating the formants. The performances are evaluated not only for clean speech, but also for speech corrupted by White, HFchannel, and Babble noise, taken from the NOISEX-92 database [159].

Table 3.3: Mean of IR (%) and IE (Hz) of the four principal formants, estimated from 512 files of TIMIT database, under clean conditions. The formants are estimated from the speech signal, and its first four IMFs, derived using EMD, M2-EMD (D2) and M3-EMD (D2). The speech signal is pre-emphasized prior to LP analysis.

Method →	Speech				EMD				M2-EMD (D2)				M3-EMD (D2)			
Formant →	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
IR (%)	80	90	88	82	91	92	89	86	93	97	93	90	91	97	93	90
IE (Hz)	46	85	81	111	42	77	77	101	38	69	72	95	39	69	72	95

Table 3.3 shows the IR and IE of the first four principal formants, as estimated from the speech signal, and from IMFs 1-4, derived from EMD, M2-EMD (D2), and M3-EMD (D2) of the speech signal. The table clearly reflects the superiority of M2-EMD (D2) and M3-EMD (D2) over EMD, in capturing the higher formants - F2, F3, and F4. Figure 3.11 plots the performance metrics, as described above, but for the speech signals corrupted by varying degrees of White, HFchannel, and Babble noise. As is observed in the figure, under White noise, the F1 estimation performance of EMD is better than the EMD variants, both in terms of identification capability (higher IR) and precision (lesser IE). This occurs due to the balancing of the skewed spectrum of speech by White noise. The flat spectrum

of White noise boosts the normally overshadowed higher-frequency spectrum of speech. This makes it easier for EMD to extract higher-frequency information of speech, in its IMFs. Henceforth, the formants identification performance for F2, F3, and F4 are slightly better in the presence of White noise, compared to that under clean conditions, for all the EMD methods. As the lower-order IMFs pick up more high-frequency content, the lower-frequency content of speech shifts more and more towards the higher-order IMFs. As M2-EMD (D2) and M3-EMD (D2) represent the higher-frequency spectrum of speech in a much better way than EMD, under the influence of White noise, the vocal tract filter spectrum of speech spreads beyond the first four IMFs. Hence, more number of IMFs may be required to estimate F1 more accurately. Similar observations also apply when HFchannel noise is used in moderate levels, but the EMD variants still have better precision. An opposite phenomenon occurs in the presence of Babble noise. Babble noise has a more concentrated lower-frequency spectrum, compared to that of White noise and HFchannel noise. As a result, the performance metrics for F3 and F4, in the case of EMD, degrades, as the level of Babble noise increases. Here, the ability of M2-EMD (D2) and M3-EMD (D2) to extract higher-frequency information from speech in a better way, comes into the picture. While the IR rates of F3 and F4 diminishes in the case of EMD, the EMD variants are able to maintain a relatively high IR and a lesser IE. Finally, as far as the speech signal is concerned, it is no match for the combined effect of the four IMFs, obtained from any of the methods.

It is evident from Table 3.3 and Figure 3.11 that cumulatively the first four IMFs of M2-EMD (D2) and M3-EMD (D2) are more effective in detecting the formants, compared to that of EMD, or to pre-emphasized speech, under all conditions. However, the performance metrics shown do not reflect an obvious demerit of the EMD based formants estimation - the presence of multiple formant peaks in the vicinity of one another. As can be observed in Figure 3.8, the segregation of the speech filter spectrum amongst the IMFs of M2-EMD (D2) and M3-EMD (D2) is better than that of EMD, but not perfect. There is still some overlap of spectral information amongst the IMFs. Thus, each IMF, apart from exhibiting a strong peak around a reference formant, also exhibits relatively weaker peaks in the vicinity of some other formants. The challenge is to come up with a process of selecting the best possible peak amongst the multitude of peaks generated by the four IMFs, and get rid of the false alarms, as shown in Figure 3.10. We leave this for future exploration. However, for the sake of comparison, we compare the ability of M2-EMD (D2), DWT (Biorthogonal 2.4 wavelet), and AF-FM

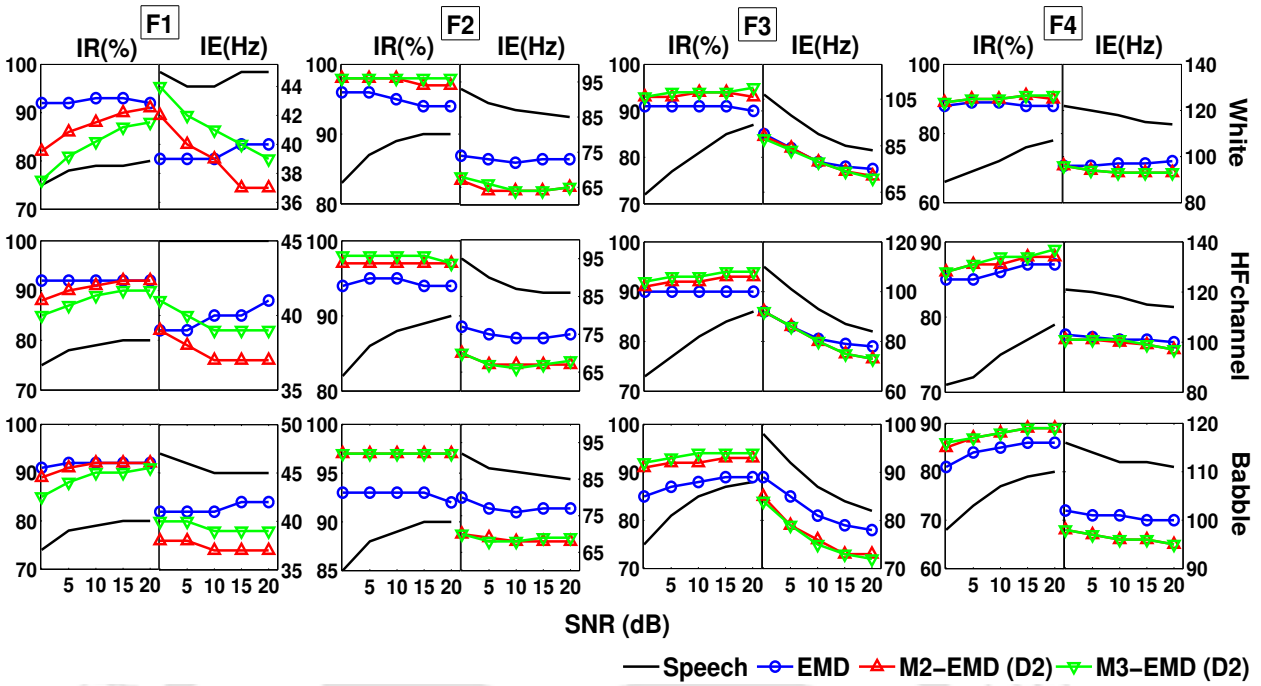


Figure 3.11: **Mean of the identification rate IR (%) and the identification error IE (Hz) of the four principal formants, estimated from 512 files of TIMIT database. The formants are estimated from the speech signal, and its first four IMFs, derived using EMD, M2-EMD (D2) and M3-EMD (D2). The speech signal is pre-emphasized prior to LP analysis. The speech signals are corrupted by White , HFchannel and Babble noise, with SNR varying from 0 to 20 dB, in steps of 5 dB.**

analysis, in detecting the first four formants, for the speech segment corresponding to the LP filter spectra shown in Figure 3.8. For this purpose, apart from IR and IE, another metric - *Number of Spurious Formants* (SF) - is used. SF is defined as -

Number of Spurious Formants (SF) : Let Fi, i=1,2,3,4, be a formant, with its reference formant frequency F_i^r, $i = 1, 2, 3, 4$, corresponding to a voiced frame of a given speech signal. Let $F^{det} = \{F_1^{det}, F_2^{det}, ..., F_D^{det}\}$ denote the set of formant peaks detected. Such a detected formant peak is considered spurious if it does not lie in the range given by $F_i^r \pm \min\{20\ \%\ of\ F_i^r\ ,\ 250\ \text{Hz}\}$, $\forall\ i = 1, 2, 3, 4$, i.e, it does not lie in the vicinity of any of the reference formants. The number of such spurious peaks, out of the set of D peaks, is given by SF.

Table 3.4 shows that M2-EMD (D2) is better than DWT in capturing formants information in it's components. Though AM-FM analysis is able to capture the formants reliably, it produces a large number of spurious formant peaks, resulting from a large number of spurious components, which are not useful for analysis.

Table 3.4: Identification rate IR (%), identification error IE (Hz), and number of Spurious Formants (SF) of the four principal formants, estimated for the speech signal utilized in Figure 3.8. The formants are estimated from the LP filter spectra of the speech signal, the first four IMFs derived using EMD and M2-EMD (D2), the first four high-frequency detail components derived using DWT (Biorthogonal 2.4 wavelet), and 20 AM-FM speech components.

Method →	M2-EMD (D2)				DWT				AM-FM			
Formant →	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
IR (%)	100	100	100	100	0	100	0	100	100	100	100	100
IE (Hz)	81	44	151	89	–	68	–	121	27	7	78	23
SF	19				22				86			

Finally, we need to break the deadlock between M2-EMD (D2), and M3-EMD (D2). Which one should be preferred for speech signal analysis ? When it comes to the ability of extracting IMFs with lesser mode mixing, or in the distribution of the speech spectrum amongst their IMFs, there is no clear winner. However, when it comes to computational time, M2-EMD (D2) is the more efficient one. Table 3.5 shows the time taken in extracting ten IMFs from the ∼3.5 s speech signal (F_s = 8 kHz) that was used in Table 2.2. The IMFs are obtained from EMD, M2-EMD (D2), M3-EMD (D2), and EEMD. $N = 10$ sifting iterations are used to extract the IMFs. In the case of EEMD, only $L = 10$ realizations of White noise are used for the decomposition. Table 3.5 shows the great time advantage obtained for both the EMD variants in comparison to EEMD, for signal decomposition. The major chunk of the time required in EMD is taken by the cubic spline interpolation process. More the number of IPs more is the time required for cubic spline interpolation. Hence, M2-EMD (D2) and M3-EMD (D2) consume more time than EMD. On top of that, M3-EMD (D2) consumes more time in extracting IPs than M2-EMD (D2) - it extracts the IPs locations every sifting iteration. M2-EMD (D2), on the other hand, extracts the IPs locations only once per sifting process. This makes it an overall more time-affordable process than M3-EMD (D2), with little additional time overhead with respect to EMD. As such, we designate M2-EMD (D2) as the final *Modified Empirical Mode Decomposition* algorithm, or MEMD.

Table 3.5: Computational time of EMD, M2-EMD (D2), M3-EMD (D2) and EEMD, in decomposing a speech signal of around 3.5 seconds duration. Ten ($M = 9$) IMFs are extracted from EMD and EEMD, where $N = 10$ sifting iterations are used per sifting process. Only $L = 10$ White noise realizations are used for EEMD. The algorithms are implemented in the GUI mode of MATLAB, on a machine having an Intel i5 quad-core processor of 3.2 GHz clock frequency, and 4 GB RAM.

Method	EMD	M2-EMD (D2)	M3-EMD (D2)	EEMD
Time (s)	0.83	1.11	1.36	15.62

3.8 Conclusion

In this work, we have made an investigation on the effect of modifying the process of extracting the IPs from the standard EMD process, on the IMFs of the speech signal. It was found that a second order differential of either the *outer residue* or the *inner residue* generates more IPs, in a way that benefits the EMD process, extracting IMFs with lesser mode mixing, and better reflecting the normally subdued higher-frequency content of the speech signal. These modified EMD processes are denoted as M2-EMD (D2) and M3-EMD (D2), respectively. Contrary to this, when the speech signal was used to generate more IPs, the EMD process failed altogether. The dyadic filterbank nature of EMD was disrupted, and the EMD process got entangled in a loop, generating similar higher-frequency IMFs. A study of the distribution of the formants of voiced speech amongst the IMFs, was done using LP analysis, which revealed that the formant structure is not limited to the first two IMFs only, particularly in the presence of noise. In the case of M2-EMD (D2) and M3-EMD (D2), the range of IMFs carrying the formants information is much larger than that of EMD, owing mainly to the better segregation of frequency scales, and an enhanced representation of the higher-frequency content of the speech signal in their IMFs.

To demonstrate the benefit of effectively segregating the speech spectrum, the LP-filter magnitude spectra of the first four IMFs were used for the task of estimating the four principal formants of speech. The results reflect the superiority of the EMD variants over EMD, in identifying the formants, with better precision. Thus, the better segregation of speech spectrum in the EMD variants provides us with signal components which could be more useful in many speech signal processing tasks. The proposed variants also provide a better trade-off, compared to other techniques like DWT and AM-FM analysis, in extracting components of the speech signal which reflect its vocal tract resonances, without producing redundant components. Additionally, the proposed variants are time-efficient. M2-EMD (D2), in fact, has only a small additional time cost with respect to EMD, and hence is the most suitable out of the proposed EMD variants, for speech processing applications. As such, it is finally designated as the *Modified Empirical Mode Decomposition* (MEMD) algorithm.

4

Analysis of the Source and System characteristics in the IMFs

Contents

Outline

The objective of this chapter is to study the ability of EMD/MEMD/ICEEMDAN in decomposing the speech signal into meaningful components which can represent its source (glottis) and system (vocal tract) characteristics. Different *phones* or *speech sounds* have different characteristics and waveform shape. As EMD (hence its variants) decomposes the signal based on its waveform characteristics only, it is necessary to evaluate how the IMFs pertaining to different speech signals represent their latent characteristics. For this purpose, at first, synthetic speech signals, representing voiced and unvoiced speech, are generated based on the source-filter theory. The pitch or fundamental frequency of the excitation source, and the resonant frequencies and bandwidths of the vocal tract resonators, are varied to synthesize a wide variety of synthetic speech signals. The correlations of the IMFs of the speech signal with the glottal excitation signal, and the individual resonators, are evaluated. This enables us to observe how well the speech signal can be decomposed into its constituent signals, with respect to its changing inherent characteristics. Extending this study to natural speech, the *Low-Time Liftered* (LTL) and *High-Time Liftered* (HTL) *cepstra* are evaluated for different phones of the TIMIT database, separately for male and female speakers. For any particular phone, the LTL cepstrum of the speech signal is compared with that of each of its IMFs. The same experiments are done for HTL cepstrum. These experiments enable us to evaluate the ability of EMD/MEMD/ICEEMDAN in segregating the vocal tract characteristics of the speech signal from its glottal source characteristics. Further, comparisons are made among the LTL and HTL cepstra of the IMFs of normal speech and that of telephone quality speech, obtained by applying *telephone channel codecs* on normal speech. This enables us to evaluate how well the IMFs can extract latent information from the speech signal when its waveform shape is modified (by codecs) without altering the information content (intelligibility) of the speech signal.

4.1 Introduction

As discussed in Section 1.1, the source-filter theory models any short segment of a speech signal as the output of exciting a cascade of resonators with an impulse train (voiced speech), or a random noise signal (unvoiced speech). The resonators represent the cavities that are formed in the vocal tract during the production of the *quasi-stationary* speech segment. In the case of voiced speech, they are also referred to as the *formants*. The resonators are cumulatively known as the *system*. During the

production of voiced speech, the quasi-periodic movement of the vocal folds (in the glottis) releases *"puffs of air"* into the vocal tract. The impulse train (further shaped by the low pass filter nature of the glottis) represents this phenomenon. The periodicity of the impulse train is known as the pitch or fundamental frequency [2,13,14]. During the production of unvoiced speech, the vocal folds do not exhibit periodic vibration. Hence, random noise is used as the excitation signal in this case, with the glottis representing an all-pass filter. The excitation and the glottis, cumulatively, are known as the *source*. Different phones or speech sounds are produced depending on the nature and behaviour of the glottal source and the vocal tract. In general, the vocal tract is relatively constricted in the case of unvoiced speech, compared to the case of voiced speech. Hence, the resonators in the case of unvoiced speech are generally concentrated at higher freqencies, compared to that of voiced speech [2, 13, 14]. Both the system and the source are key to the production of the speech signal. Hence, they characterize the speech signal. Obviously, they also represent important speaker signatures - no two speakers can produce the same speech signal for the same sound uttered.

As observed in Chapter 2, there are strong similarities between some of the IMFs of the speech signal with the EGG signal. This indicates the ability of EMD, and its variants, in manifesting the glottal source characteristics. Similarly, as has been discussed in Chapters 2 and 3 (using LP analysis), the lower-order IMFs do represent the system characteristics. However, EMD (hence its variants) blindly decomposes the signal based on its waveform shape only. Though the underlying speech production mechanism remains more or less the same, the characteristics and waveform shapes of different phones vary. As such, it is necessary to analyze how well the source and system are manifested in the IMFs of different speech signals, corresponding to different phones.

In Section 3.7, it was observed (using LP analysis) that, cumulatively, the IMFs can represent the vocal tract resonances in a better way than that of the speech signal itself, even under noisy conditions. However, as is expected, the performance of detection of the resonances diminished with the addition of noise. The addition of noise to the speech signal effects its waveform shape, and simultaneously its intelligibility. Hence, the capability of the IMFs to manifest the system characteristics diminished. However, the information embedded in the speech signal may not be solely attributed to its waveform shape. That is why, in communication technology, speech signals transmitted after application of *telephone channel codecs* [160], are received with a seemingly little loss in intelligibility. Though the speech waveform shape changes, the critical information within it is not lost. As such, it would be

interesting to evaluate the characteristics of the IMFs of telephone quality speech as opposed to that of original natural speech. This investigation may be particularly useful as speech is an important aspect of communication technology.

Henceforth, as the preceding discussion dictates, this chapter is dedicated to investigating the capability of the IMFs of the speech signal in manifesting its latent characteristics, as the nature and waveform shape of the speech signal changes. For this purpose, we will first perform controlled experiments on synthetic speech signals, and then extend the investigation to natural and telephone quality speech. The rest of the chapter is organized as follows : Section 4.2 explores how effectively the synthetic speech signals are broken down into their constituents in the form of IMFs. The effects of varying the resonant frequencies, their corresponding bandwidths, and the glottal excitation, are studied. Section 4.3 investigates the source and system separation in the IMFs of natural speech. For this purpose, different phones from the TIMIT [128] corpus are considered, separately, for male and female speakers. The effect of telephone channel codecs, on the separation of source and system characteristics in the IMFs, is also investigated. Section 4.4 concludes this work.

4.2 Decomposition of synthetic speech

In Section 2.1.4, the ability of EMD to segregate a signal composed of a binary mixture of sinusoids has been discussed. The decomposition of a signal resulting from the convolution of multiple signals is a much more complex problem, as convolution involves multiplication, summation, and shifting, to produce the final output. As discussed in Section 1.1, the speech signal is considered to be produced by the convolution of multiple signals comprising the excitation signal, and the impulse responses of the glottis, the resonators, and the lips. Thus, if EMD/MEMD/ICEEMDAN is able to segregate these multiple signals responsible for producing the speech signal, it would be truly useful for speech signal analysis. The ideal objective of AM-FM analysis, i.e, to extract finite but meaningful components of the speech signal, would be all but complete. This motivates us to examine whether the IMFs (obtained from EMD/MEMD/ICEEMDAN) of a small segment of a speech signal can represent the resonances and the glottal source which were key to its production.

To study the decomposition abilities of EMD (and its variants), firstly, we consider speech signals synthetically generated using the source-filter theory. For voiced speech (vowel-like sounds), in general, for every 1000 Hz of the DFT spectrum of the speech signal, a resonance peak is observed. For unvoiced

speech (consonant-like sounds), the resonators are not as well specified [2,13,14]. Nevertheless, in our study, four resonators are used in generating the speech signal, with $F_s = 8$ kHz. The speech signal, $s(n)$, is given by,

$$s(n) = e(n) * g(n) * v_1(n) * v_2(n) * v_3(n) * v_4(n) * r(n) \ , \tag{4.1}$$

$$S(z) = E(z)G(z)V_1(z)V_2(z)V_3(z)V_4(z)R(z) \ , \tag{4.2}$$

$$e(n) = \begin{cases} \sum_{k=0}^{K_i-1} \delta(n - k\frac{F_s}{F_0}) \ , \text{voiced speech} \\ \mathcal{N}(0,1) \ , \text{unvoiced speech} \end{cases} \tag{4.3}$$

$$V_r(z) = \frac{1}{(1 - p_r z^{-1})(1 - p_r^{\star} z^{-1})} \ , \ p_r = (1 - \frac{B_r}{F_s}) e^{j2\pi \frac{f_r}{F_s}} \ , \ r = 1,2,3,4 \ , \tag{4.4}$$

$$G(z) = \begin{cases} \frac{1}{(1-0.98z^{-1})^2} \ , \text{voiced speech} \\ 1 \ , \text{unvoiced speech} \end{cases} \tag{4.5}$$

$$R(z) = 1 - 0.98z^{-1} \ , \tag{4.6}$$

Using the above formulation, voiced and unvoiced speech signals of 1 s duration are synthesized. In equation 4.3, K_i represents the number of impulses in the 1 s signal, for a particular value of pitch or fundamental frequency, F_0. In equation 4.4, f_r represents the analog resonant (center) frequency of the r^{th} resonator, and B_r its corresponding analog bandwidth. A synthetic speech signal so generated is decomposed into $M+1 = 10$ IMFs, separately by EMD, MEMD, and ICEEMDAN. We then analyze and compare the IMFs corresponding to a 20 ms segment/frame at the middle (centered at 0.5 s) of the speech signal, with the actual constituents of the speech segment. The constituents of the segment are given by,

$$g^e(n) = e(n) * g(n) \leftrightarrow G^e(z) = E(z)G(z) \tag{4.7}$$

$$v_r^e(n) = e(n) * v_r(n) \leftrightarrow V_r^e(z) = E(z)V_r(z) \ , \ r = 1,2,3,4 \ , \tag{4.8}$$

$$r^e(n) = e(n) * r(n) \leftrightarrow R^e(z) = E(z)R(z) \tag{4.9}$$

Figure 4.1 shows a 20 ms segment of two synthetic speech signals (one voiced, other unvoiced), and their constituents. The DFT magnitude spectra of the signals are also shown. For both the speech signals, f_rs are considered at 500 Hz, 1500 Hz, 2500 Hz, and 3500 Hz, respectively. For the voiced speech signal, the corresponding bandwidths (B_rs) are considered as 100 Hz, 200 Hz, 300 Hz, and 400 Hz, respectively. This actuates to the lower-frequency resonators having sharper peaks, as is

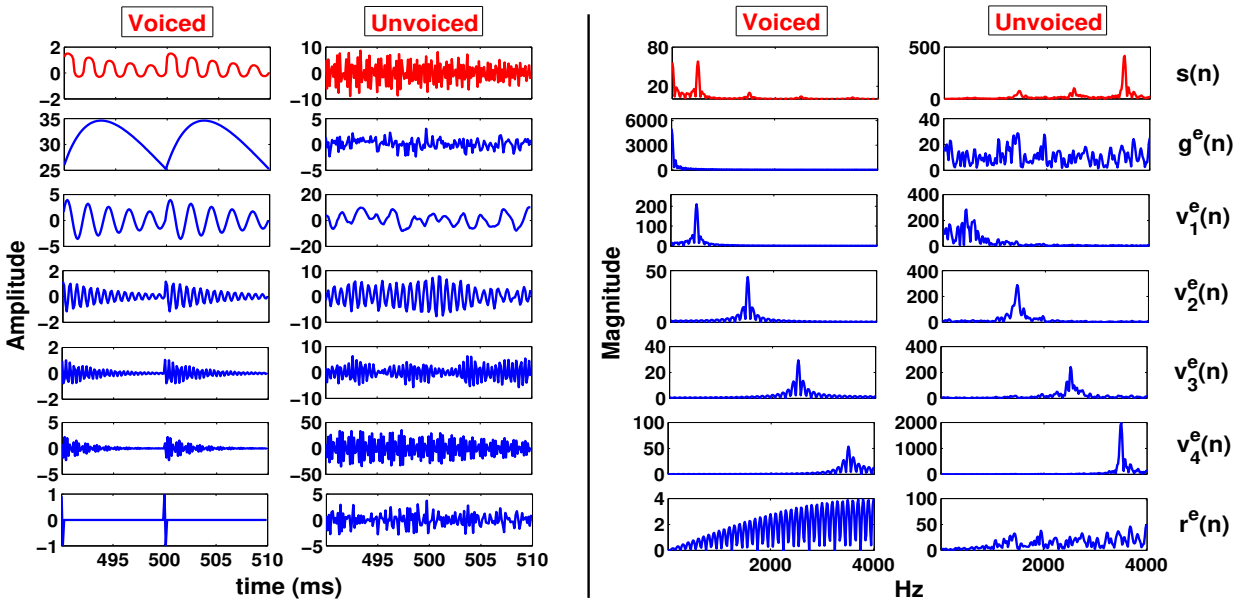


Figure 4.1: **The first column shows a 20 ms segment of a synthetic voiced speech signal, and its constituents. The second column shows a 20 ms segment of a synthetic unvoiced speech signal, and its constituents. The third and fourth columns shows the DFT magnitude spectra of the signals represented in the first and second columns respectively.**

generally observed in natural voiced speech. For the unvoiced speech signal, the B_rs are considered as 400 Hz, 300 Hz, 200 Hz, and 100 Hz, respectively. This actuates to having stronger higher-frequency resonators. Though this may not strictly be true in the case of natural unvoiced speech, we assume so for our hypothetical experiments. $F_0 = 100$ Hz is considered for the voiced speech signal. As can be observed from the figure, voiced speech exhibits periodic nature, whereas unvoiced speech is noise-like.

Figure 4.2 presents the first seven IMFs (extracted using EMD) of the voiced and unvoiced speech segments shown in Figure 4.1. Our objective is to find the correlation or similarity of each IMF with the different constituents of the synthetic speech signal. In particular, we are interested in the glottal characteristics and the vocal tract resonances. Lip radiation characteristics are not considered in this study. Thus, given a voiced/unvoiced speech segment, $s(n)$, and their IMFs, $h_k(n)$, $k = 1, 2, ..., M+1 = 10$ (obtained from EMD/MEMD/ICEEMDAN), the following *maximum correlation coefficients* are evaluated.

$$C[g^e(n), h_k(n)] = \max_{m, m \geq 0} \frac{\sigma[g^e(n+m), h_k(n)]}{\sqrt{\sigma^2[g^e(n)]\sigma^2[h_k(n)]}} \ , \ k = 1, 2, ...M+1 = 10 \tag{4.10}$$

$$C[v_r^e(n), h_k(n)] = \max_{m, m \geq 0} \frac{\sigma[v_r^e(n+m), h_k(n)]}{\sqrt{\sigma^2[v_r^e(n)]\sigma^2[h_k(n)]}} \ , \ r = 1, 2, 3, 4 \ , \ k = 1, 2, ...M+1 = 10 \tag{4.11}$$

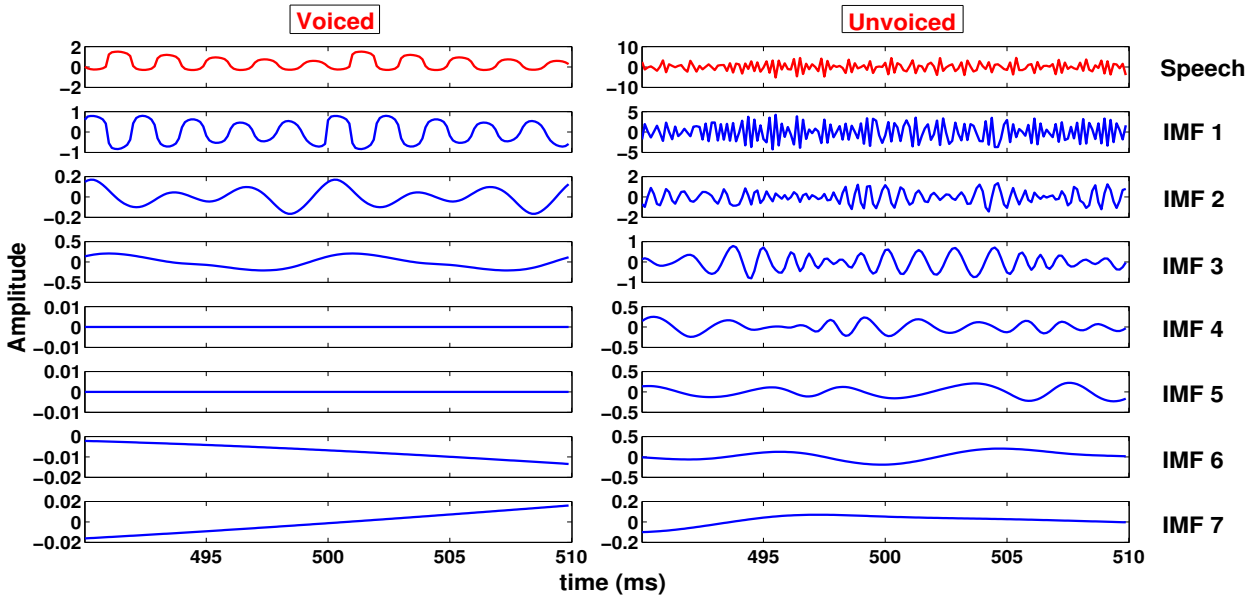


Figure 4.2: **The first column shows a 20 ms segment of a synthetic voiced speech signal, and its first seven IMFs. The second column shows a 20 ms segment of a synthetic unvoiced speech signal, and its first seven IMFs. The IMFs are obtained using EMD.**

In the above two equations $\sigma[.\ ,\ .]$ denotes the cross-covariance operator. Given any two arbitrary signals, $x(n)$ and $y(n)$, the cross-covariance operator, at a lag m, is defined as,

$$\sigma[x(n+m), y(n)] = \mathbb{E}[x(n+m) - \mathbb{E}x(n)][y(n) - \mathbb{E}y(n)] \tag{4.12}$$

$$\sigma^2[x(n)] = \sigma[x(n), x(n)]\ ,\ \sigma^2[y(n)] = \sigma[y(n), y(n)] \tag{4.13}$$

In the following subsections, using the maximum correlation coefficients, we aim to study which IMFs capture the glottal characteristics and the individual resonators of the synthetic speech signal. Using different combinations of F_0, f_rs, and B_rs, speech signals are synthesized which reflect male and female genders, and a broad class of voiced and unvoiced speech. As mentioned before, the speech signals are decomposed by EMD, MEMD, and ICEEMDAN to evaluate which method provides the best decomposition. These broad set of experiments also allow us to evaluate the conditions under which the speech signal (within the perimeters of the source-filter theory) can be best decomposed.

4.2.1 Effect of bandwidths of the resonators

As a general rule, voiced speech exhibits stronger and sharper resonances at lower frequencies. For unvoiced speech, though there is no regular pattern, because of vocal tract constrictions sharp and strong resonances are observed at higher frequencies [2, 13, 14]. Voiced speech is also associated with

a periodic excitation, specified by F_0, unlike in the case of unvoiced speech. Generally, lower pitch frequencies ($\approx$ 100 Hz) are associated with male speakers, and higher pitch frequencies ($\approx$ 200 Hz) with female speakers. Hence, assuming that f_ks are kept fixed (also F_0 is fixed for voiced speech) at 500 Hz, 1500 Hz, 2500 Hz, and 3500 Hz, we hypothesize that lesser the bandwidths of the lower-frequency resonators (compared to that of the higher frequency resonators), the speech signal is more voiced. Hence, keeping F_0 constant at 100 Hz (representing male speakers), and 200 Hz (representing female speakers), two sets of voiced speech signals may be obtained by varying the B_rs. Contrary to voiced speech, greater the bandwidths of the lower-frequency resonators (compared to that of the higher-frequency resonators), the speech signal is hypothesized to be more unvoiced. As such, by varying the bandwidths of the resonators, a set of unvoiced speech signals may be synthesized. The bandwidths, B_rs, for the three sets of synthesized speech signals are varied as,

$$B_1 = 250 - 1.5 \times K_B \text{ Hz} \tag{4.14}$$

$$B_2 = 250 - 0.5 \times K_B \text{ Hz} \tag{4.15}$$

$$B_3 = 250 + 0.5 \times K_B \text{ Hz} \tag{4.16}$$

$$B_4 = 250 + 1.5 \times K_B \text{ Hz} \tag{4.17}$$

For voiced speech, the parameter K_B is varied between $[10, 160]$. Higher the value of K_B, greater is the voiced characteristics of the synthesized speech signal. For unvoiced speech, K_B is varied between $[-160, 0]$. Lesser the value of K_B, more is the unvoiced characteristics of the signal.

Having synthesized the speech signals with different degrees of voiced and unvoiced characteristics, they are decomposed into their IMFs. Then, the maximum correlation coefficients are obtained for every speech signal, as discussed in the previous subsection. Figure 4.3 shows the maximum correlation coefficients, for the first seven IMFs extracted using EMD. The last three IMFs are trend-like, and hence left out of the analysis. As can be seen from the figure, for voiced speech, irrespective of the degree of voicedness, in general, the higher-order IMFs (IMFs 3 and beyond) exhibit high similarity with the glottal source characteristics. However, the correlation values for $F_0 = 200$ Hz are higher than that for $F_0 = 100$ Hz. Also, for $F_0 = 200$ Hz, IMF_2 also has significant similarity with the glottal source constituent. This shows that depending on the value of F_0, the subset of the IMFs that carry glottal source characteristics vary. The first resonator constituent, for $F_0 = 100$ Hz, is carried strongly by IMF_1. The rest of the resonators are also carried mainly by IMF_1 but in a much weaker sense.

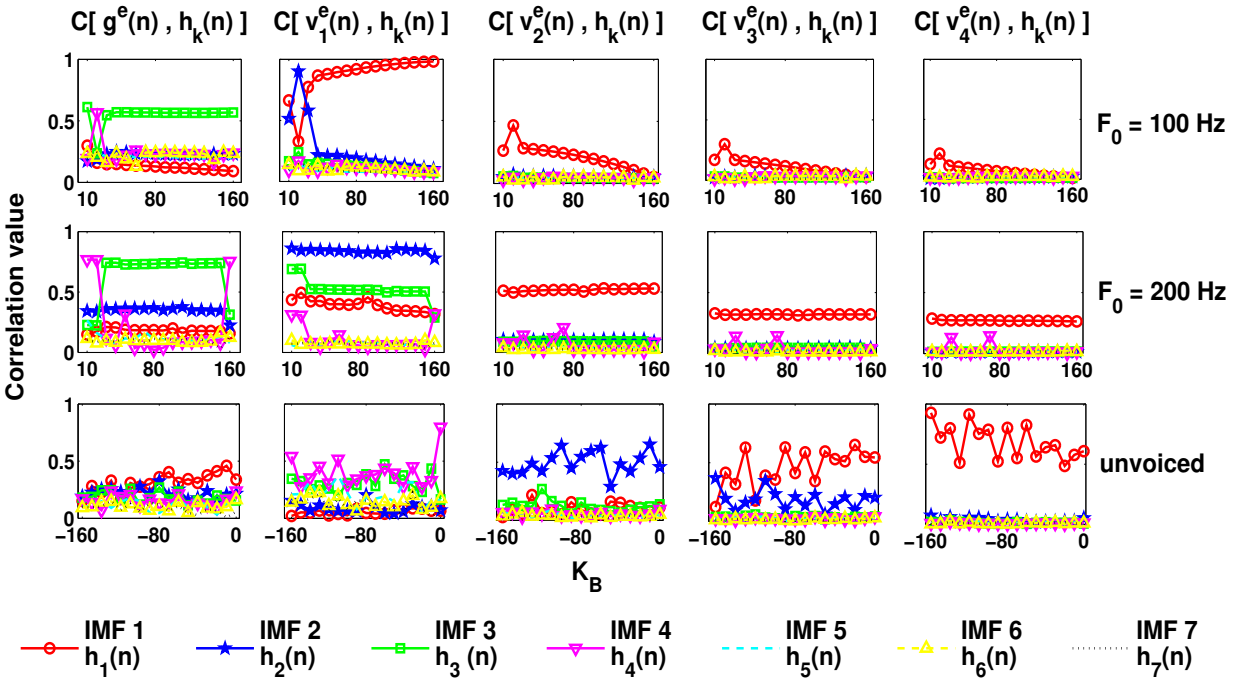


Figure 4.3: Maximum correlation coefficients for synthetic voiced and unvoiced speech signals. For voiced speech signals, two different fundamental frequencies (100 Hz and 200 Hz) are considered. The bandwidths of the four resonators are varied using the K_B parameter. The maximum correlation coefficients are shown for the first seven IMFs, obtained using EMD.

As the degree of voicedness increases, the higher frequency resonators are overshadowed by the first resonator, and their segregation becomes almost impossible. For $F_0 = 200$ Hz, the first resonator is strongly carried by IMF_2, and not IMF_1. The higher frequency resonators are manifested in IMF_1, but in a much stronger way than in the case of $F_0 = 100$ Hz. In the case of unvoiced speech, the glottal characteristics are evenly (and hence weakly) distributed amongst all the IMFs. This is expected as the glottal excitation, in this case, is random noise. IMF_1, in this case, manifests strongly the fourth and third resonators, IMF_2 the second resonator. IMFs 3 and 4 manifest the first resonator in a weak sense.

Thus, the bandwidth, and hence the strength of the resonators, have a strong impact on the decomposition ability of EMD. The stronger resonance peaks are generally carried by IMF_1 alone. In the case of unvoiced speech, and voiced speech at high F_0, the segregation of the constituents into different IMFs, in slightly better, but far from ideal. This may be attributed to the dependence of EMD on the extrema of the signal. For unvoiced speech, and voiced speech at a high F_0 value, the number of extrema present in the speech signal is more. Hence, the decomposition is better.

Figure 4.4 shows the maximum correlation coefficients, for the first seven IMFs extracted using

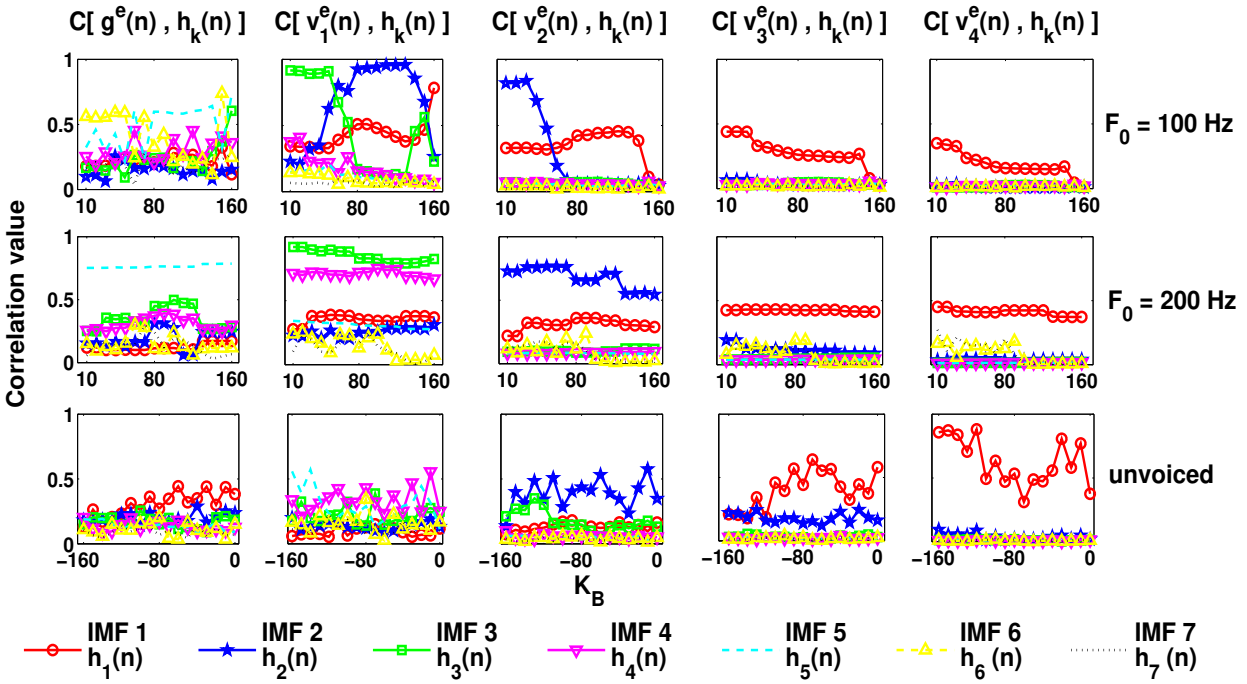


Figure 4.4: Maximum correlation coefficients for synthetic voiced and unvoiced speech signals. For voiced speech signals, two different fundamental frequencies (100 Hz and 200 Hz) are considered. The bandwidths of the four resonators are varied using the K_B parameter. The maximum correlation coefficients are shown for the first seven IMFs, obtained using MEMD.

MEMD. It is clear that the decomposition ability of MEMD is superior to that of EMD. For voiced speech synthesized at $F_0 = 100$ Hz, the glottal characteristics are manifested in the IMFs of higher-order (IMFs 5 and 6), compared to the case of EMD. The separation of the vocal tract resonators into different IMFs is also better achieved, particularly when the degree of voicedness is less (the resonators are of comparable bandwidths). For voiced speech synthesized at $F_0 = 200$ Hz the constituents are better separated than for the case of $F_0 = 100$ Hz. For unvoiced speech the decomposition is similar to that observed in the case of EMD. Of course, as in the case of EMD, for both voiced and unvoiced speech, segregating all the resonators, particularly those with lesser strength (higher bandwidth), is not achieved.

Figure 4.5 shows the maximum correlation coefficients, for the first seven IMFs extracted using ICEEMDAN. The plots show that the decomposition abilities of ICEEMDAN are similar to (and better than) that of MEMD. For voiced speech (considering both $F_0 = 100$, 200 Hz) the higher-frequency resonators have better correlations with IMF_1 than was observed in the case of MEMD. For $F_0 = 100$ Hz, depending on the degree of voicedness, IMFs 1 and 2 show strong correlations with the second resonator. For $F_0 = 200$ Hz, IMF_2 manifests strong correlation with the second resonator,

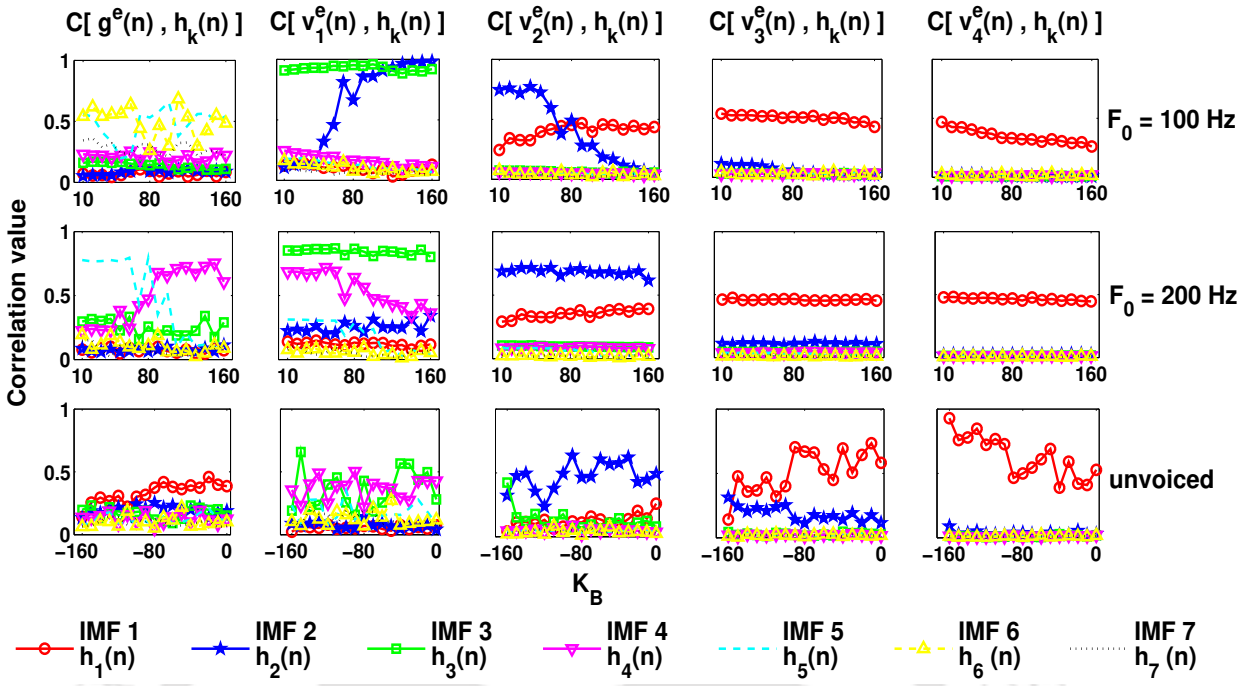


Figure 4.5: Maximum correlation coefficients for synthetic voiced and unvoiced speech signals. For voiced speech signals, two different fundamental frequencies (100 Hz and 200 Hz) are considered. The bandwidths of the four resonators are varied using the K_B parameter. The maximum correlation coefficients are shown for the first seven IMFs, obtained using ICEEMDAN.

irrespective of the degree of voicedness. Again, irrespective of the degree of voicedness and the value of F_0, the first resonator is carried by IMF_3. For unvoiced speech, the maximum correlation coefficients are similar to the case of MEMD.

4.2.2 Effect of frequencies of the resonators

Having observed the effect of variation of bandwidths of the resonators on the decomposition of the speech signal, we move towards studying the effect of variation of the resonant frequencies themselves. For this study, we consider only synthetic voiced speech signals. The f_rs are varied as,

$$f_1 = \begin{cases} 500 - 10 \times K_f \text{ Hz} \ , \ 500 - 10 \times K_f \text{ Hz} > 2 \times F_0 \\ 2 \times F_0 \ , \ \text{otherwise} \end{cases} \tag{4.18}$$

$$f_2 = 1500 - 10 \times K_f \text{ Hz} \tag{4.19}$$

$$f_3 = 2500 - 10 \times K_f \text{ Hz} \tag{4.20}$$

$$f_4 = 3500 - 10 \times K_f \text{ Hz} \tag{4.21}$$

The bandwidths of the resonators are given by $B_r = 0.1f_r$, $r = 1, 2, 3, 4$, i.e, the bandwidths are 10 % of their resonant frequencies. Using these parameters, two sets of voiced speech signals are synthesized, one for $F_0 = 100$ Hz, and other for $F_0 = 200$ Hz. The parameter K_f is varied between $[-45, 45]$ and controls the center frequencies of the resonators. Negative values of K_f result in higher resonant frequencies, which are generally associated with a smaller vocal tract cavities. Conversely, positive values of K_f results in lower resonant frequencies, which are generally associated with larger vocal tract cavities [2, 13, 14].

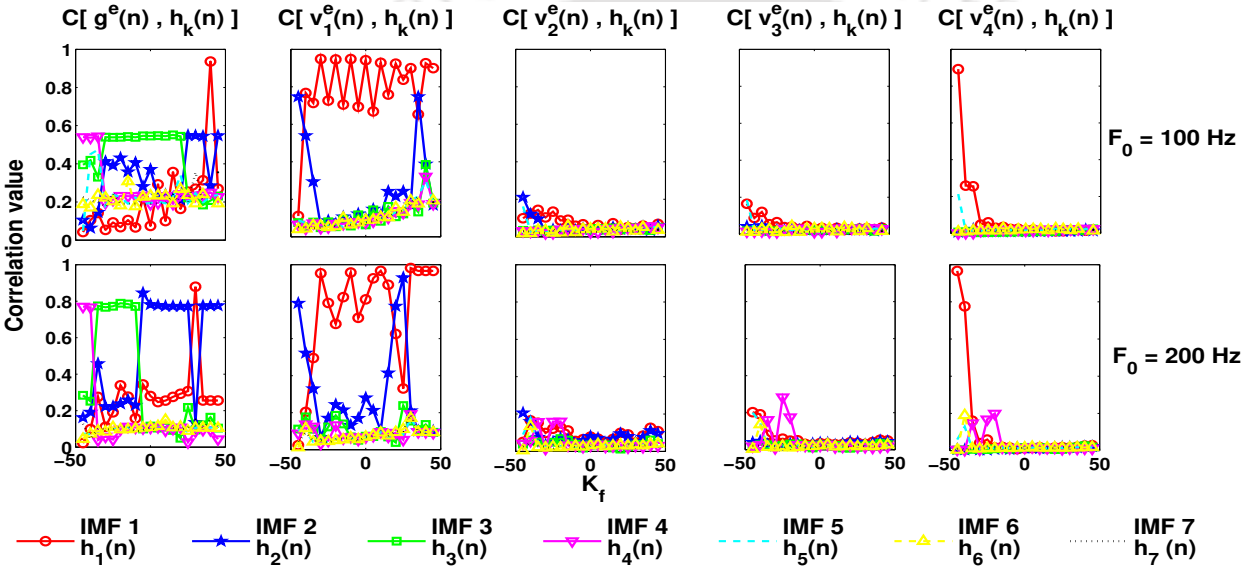


Figure 4.6: **Maximum correlation coefficients for synthetic voiced speech signals. Two different fundamental frequencies (100 Hz and 200 Hz) are considered. The frequencies of the four resonators are varied using the K_f parameter. The maximum correlation coefficients are shown for the first seven IMFs, obtained using EMD.**

The speech signals so synthesized are decomposed into ten IMFs each. The first seven IMFs are used to calculate the maximum correlation coefficients. Figure 4.6 shows the maximum correlation coefficients for the first seven IMFs extracted using EMD. It can be observed from the figure that for both F_0s, when the resonant frequencies are high (K_f is less), IMFs 3 and 4 represent the glottal characteristics the best out of all the IMFs. As the f_ks decrease (K_f is high), IMFs 2 and 1 manifest the glottal characteristics the best. IMF$_1$, in general, manifests the characteristics of the first resonator irrespective of the value of K_f. The first resonator is the strongest and overshadows the rest of the resonators. Only at very low K_f values, when the strength of the first resonator is at its lowest, the fourth resonator (which is farthest from it) is able to project its characteristics on IMF$_1$. Thus having a high-frequency first resonator (and hence larger bandwidth) is beneficial in segregating the

constituents of the speech signal. Of course, extracting the other resonators in the vicinity of the first resonators is extremely difficult.

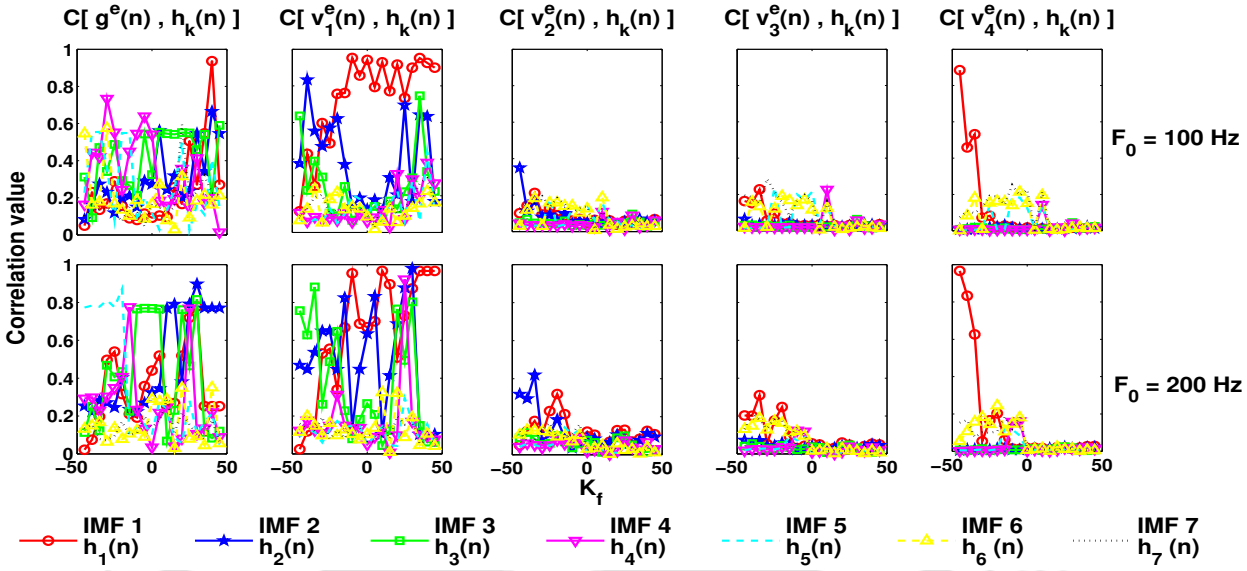


Figure 4.7: **Maximum correlation coefficients for synthetic voiced speech signals. Two different fundamental frequencies (100 Hz and 200 Hz) are considered. The resonant frequencies of the four resonators are varied using the K_f parameter. The maximum correlation coefficients are shown for the first seven IMFs, obtained using MEMD.**

Figure 4.7 shows the maximum correlation coefficients for the first seven IMFs extracted using MEMD. It can be observed from the figure that for both the values of F_0, when the resonant frequencies are high (K_f is less), IMF_4 and beyond represent the glottal characteristics the best out of all the IMFs. In this range of K_f, IMFs 2 and 3 manifest the first resonator. The second and third resonators are not manifested at all, whereas the fourth resonator is represented by IMF_1. As the value of K_f increases, the resonant frequencies decrease, and so their bandwidths. The first resonator subdues the rest of the resonators. Hence, only the first resonator is manifested in IMF_1. As such, IMFs 2 and 3 now strongly manifest the glottal characteristics.

Figure 4.8 shows the maximum correlation coefficients for the first seven IMFs extracted using ICEEMDAN. It can be observed from the figure that compared to EMD and MEMD, ICEEMDAN is a better decomposer. Irrespective of the variations in the f_ks, the glottal source characteristics are manifested by IMF_5 in the case of $F_0 = 100$ Hz, and IMF_4 in the case of $F_0 = 200$ Hz. The first resonator is manifested in IMFs 2 and 3. IMF_1 carries the higher resonators. The correlations are much more significant than in the case of EMD and MEMD.

We may now summarize the observations made from the experiments on synthetic speech :

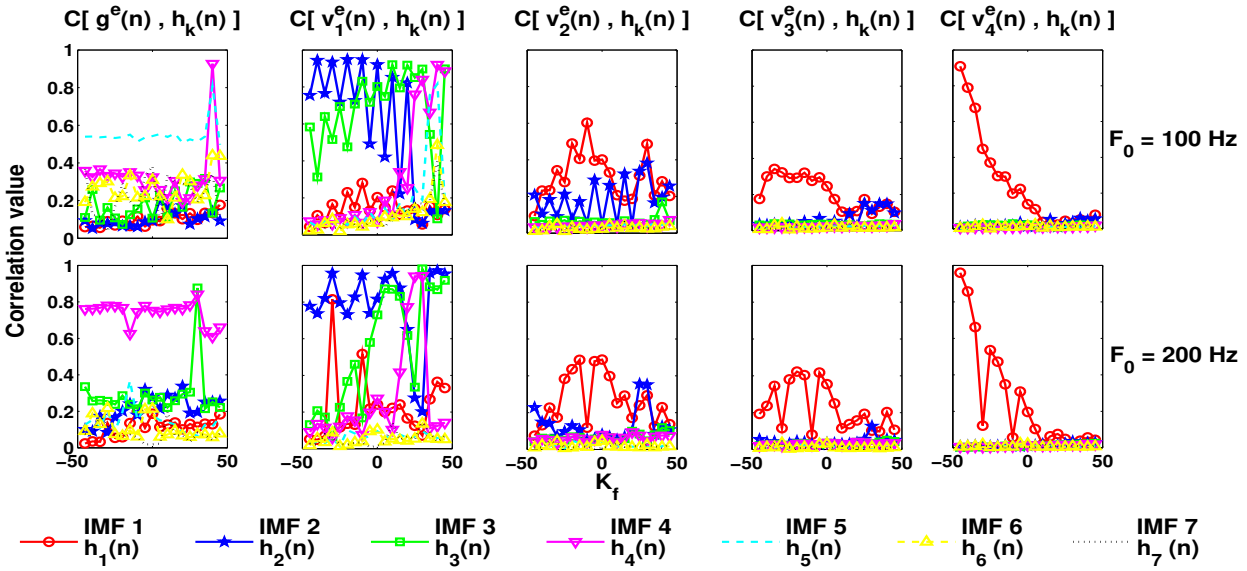


Figure 4.8: Maximum correlation coefficients for synthetic voiced speech signals. Two different fundamental frequencies (100 Hz and 200 Hz) are considered. The frequencies of the four resonators are varied using the K_f parameter. The maximum correlation coefficients are shown for the first seven IMFs, obtained using ICEEMDAN.

- **(i)** Decomposition of the synthetic voiced speech signal into its actual constituents is better achieved when the pitch frequency is high. When the F_0 value is high, the vocal tract resonators and the glottal source are segregated into more number of IMFs, with less overlap.

- **(ii)** In the case of synthetic voiced speech, when the strength of the first resonator overshadows that of the other resonators, it becomes difficult to segregate the higher frequency resonators into different IMFs. Conversely, a better and more equitable decomposition is achieved when the strengths of the higher-frequency resonators are comparable to that of the first resonator.

- **(iii)** In the case of synthetic unvoiced speech, the segregation of the vocal tract resonators into different IMFs is much better achieved compared to that of voiced speech. This is because the higher-frequency resonators, in this case, have more strength than the lower-frequency resonators.

- **(iv)** ICEEMDAN provides the most equitable decomposition, closely followed by MEMD. They are significantly better than EMD in decomposing the speech signal.

4.3 Source-system separation of natural speech based on cepstral analysis

As already discussed in Chapter 1, the source-filter theory is a rather simplistic model, and the conclusions derived from it may not concur with that of natural speech. Hence, having observed the decomposition capabilities of EMD/MEMD/ICEEMDAN on synthetic speech signals, we need to evaluate the observations on natural speech signals. For this purpose, natural speech signals corresponding to six different phones - /ah/, /eh/, /ih/, /ow/, /uh/, /s/ - are considered. The speech signals are taken from the TIMIT [128] corpus. The corpus consists of 6300 speech utterances, spoken by 438 male and 192 female speakers. The utterances are also spread across eight dialects of US English. The signals are downsampled to $F_s = 8$ kHz. The first five phones represent voiced speech (the five english vowels), whereas the last phone represents unvoiced speech. The phones are considered separately for the male and female speakers. Every natural speech signal (corresponding to a phone) is decomposed into $M + 1 = 10$ IMFs using EMD/MEMD/ICEEMDAN. A 20 ms segment/frame at the middle of the signal, and the corresponding segments of their IMFs, are then considered for analysis. The experiments are focussed on two principal objectives :

(i) : Whether different subsets of the IMFs can segregate the source and system characteristics of the speech signal ?

(ii) : When the speech signal is subjected to telephone channel codecs, are the IMFs still able to represent the latent source and system characteristics of the signal ?

Unlike in the case of synthetic speech signals, now we do not have direct access to the constituents of the speech signal. As such, we utilize the well-established technique of *cepstral* or *homomorphic* analysis [2, 13, 14], which is used to represent the source and system characteristics of the speech signal. The following subsections discuss *cepstral* analysis, and its application to the speech signal and its IMFs.

4.3.1 Cepstral analysis of speech

As mentioned earlier, according to the source filter theory, any speech segment, $s(n)$, may be represented as,

$$s(n) = e(n) * g(n) * v(n) * r(n) \leftrightarrow S(z) = E(z)G(z)V(z)R(z) \ , \tag{4.22}$$

$$s(n) = g^e(n) * h^v(n) \leftrightarrow S(z) = G^e(z)H^v(z) \ , \tag{4.23}$$

$$g^e(n) = e(n) * g(n) \leftrightarrow G^e(z) = E(z)G(z) \ , \ h^v(n) = v(n) * r(n) \leftrightarrow H^v(z) = V(z)R(z) \tag{4.24}$$

In the above equations, $v(n)$, and hence $V(z)$, cumulatively represents the cascade of vocal tract resonators. The speech signal is assumed to be the result of the convolution of two primary constituents - (i) the source, $g^e(n)$, and (ii) the system, $h^v(n)$. Based on this formulation, the DFT magnitude spectrum of the speech segment is utilized to find it cepstrum, as,

$$S(f) = G^e(f)H^v(f) \rightarrow \boxed{\log|.|} \rightarrow \log|S(f)| = \log|G^e(f)| \ + \ \log|H^v(f)| \tag{4.25}$$

$$\hat{s}(n) = \hat{g}^e(n) \ + \ \hat{h}^v(n) \leftrightarrow \log|S(f)| = \log|G^e(f)| \ + \ \log|H^v(f)| \tag{4.26}$$

In the above equations, f represents the analog frequencies corresponding to the discrete frequencies of the DFT spectrum. The signal, $\hat{s}(n)$, is called the cepstrum of $s(n)$. The cepstrum consists of two components, $\hat{g}^e(n)$ and $\hat{h}^v(n)$, which are approximations of the source and system characteristics of the speech segment.

The first few cepstral coefficients are generally associated with the system characteristics, and the higher-order cepstral coefficients are associated with the source characteristics. A simple windowing technique, called *liftering*, is used to segregate the cepstral coefficients. When the first few components are retained, it is called *Low-Time Liftering* (LTL). In our case, coefficients 1-14 are retained using LTL for representing the system characteristics. The 0^{th} coefficient, $\hat{s}(0)$, is left out as it represents basically the energy of the speech segment, and may overshadow the other coefficients. Thus, if

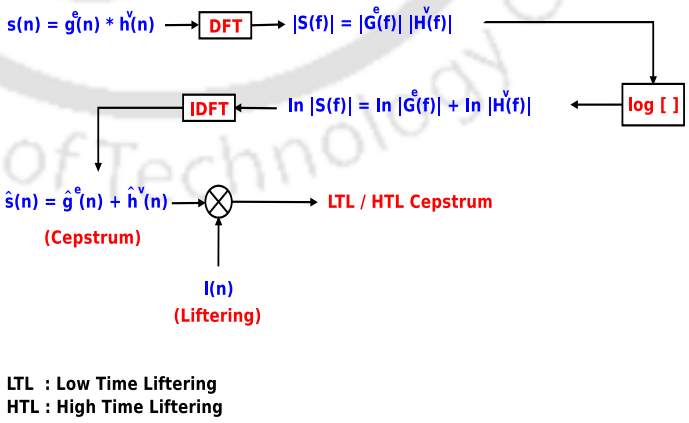


Figure 4.9: Schematic diagram showing the process of extracting the cepstrum of the speech signal.

$\sqcap[s(n)]$ represents the operation of extracting the cepstral coefficients of $s(n)$ followed by LTL, we have,

$$\sqcap[s(n)] = \hat{s}(n)l(n) \ , \ l(n) = \begin{cases} 0 \ , n = 0 \\ 1 \ , 1 \leq n \leq 14 \\ 0 \ , n \geq 15 \end{cases} \tag{4.27}$$

For representing the source characteristics, all the coefficients after the 14^{th} coefficient are considered. This is called *High-Time Liftering* (HTL). Let $\sqcup[s(n)]$ represent the operation of extracting the cepstral coefficients of $s(n)$, and then applying HTL. We have,

$$\sqcup[s(n)] = \hat{s}(n)l(n) \ , \ l(n) = \begin{cases} 0 \ , 0 \leq n \leq 14 \\ 1 \ , n \geq 15 \end{cases} \tag{4.28}$$

Figure 4.9 shows the schematic block diagram of the process of extracting the cepstrum (LTL or HTL) of the speech signal. In the figure, IDFT refers to *Inverse Discrete Fourier Transform* [2,13,14]. Figure 4.10 shows a 20 ms voiced speech segment (corresponding to the phone /ah/ of a male speaker of the TIMIT corpus), and its corresponding cepstrum, LTL cepstrum, and HTL cepstrum. The corresponding DFT spectra ($\mathcal{F}\{.\}$ represents the DFT operation) are also shown. As shown in the figure, cepstrum represents the time-domain manifestation of the logarithmic mapping of the original DFT magnitude spectrum. The LTL cepstrum is the time-domain manifestation of the envelope of the log-magnitude spectrum. The peaks of the spectrum of the LTL cepstrum are a manifestation of the vocal tract resonances. The HTL cepstrum, on the other hand, represents the fluctuations of the log-magnitude spectrum. It represents the harmonics of the pitch frequency of the voiced speech segment. The amplitudes of the harmonics also diminish as their frequencies increase, which is a manifestation of the low-pass filter nature of the glottis. In a nutshell, Figure 4.10 illustrates how the source and system characteristics are split between the HTL and LTL cepstra respectively of the speech segment. Henceforth, we will use cepstral analysis on natural speech signals to evaluate the ability of EMD/MEMD/ICEEMDAN to represent their source and system characteristics.

4.3.2 Source-system separation of natural speech

As observed in our experiments with synthetic speech, the IMFs represent the characteristics of both the glottal source and the vocal tract resonators, albeit in different proportions. The same may be expected in the case of natural speech. Thus, cepstral analysis may also be used on the IMFs of

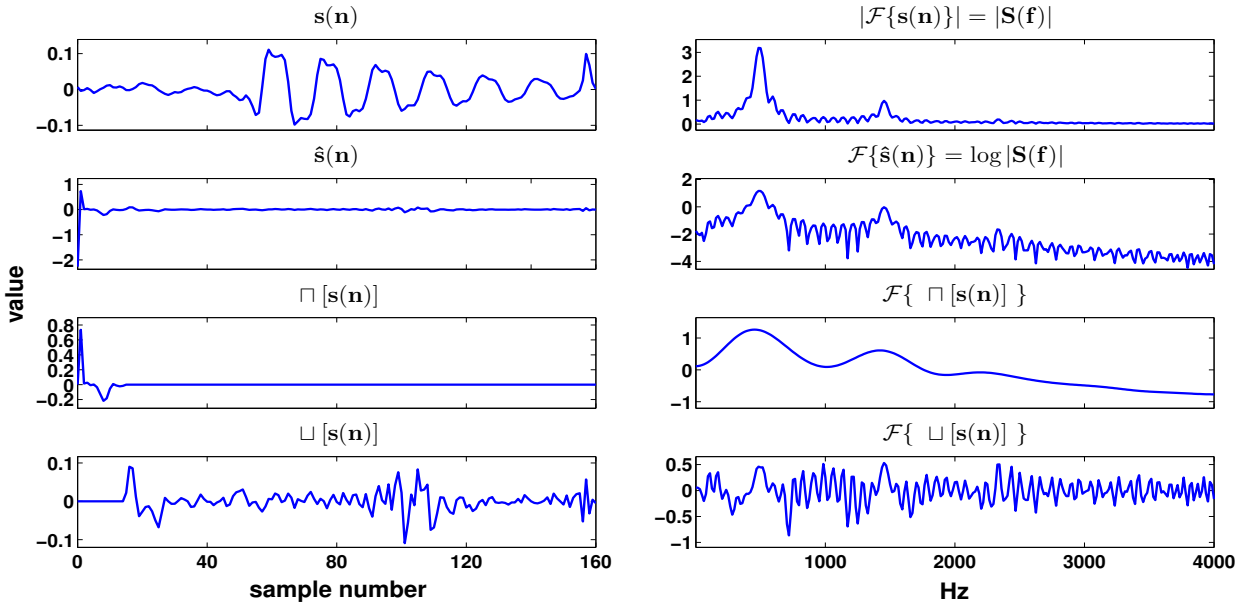


Figure 4.10: **Left column shows a 20 ms (160 samples at F_s = 8 kHz) speech segment, its cepstrum, LTL and HTL cepstrum. The speech segment corresponds to the phone /ah/ of a male speaker of the TIMIT corpus. The corresponding DFT spectra are shown in the right column. $\mathcal{F}\{.\}$ denotes the operation of evaluating the DFT.**

the speech signal to represent the source and system characteristics embedded in them. In general, as we have observed in the case of synthetic speech signals, a single IMF may not capture all the source or system information. It is more reasonable to expect that cumulatively a certain subset of the IMFs would profoundly manifest system characteristics, whereas another subset would strongly exhibit source characteristics. In this subsection, we wish to evaluate if such a separation is feasible, and to what extent.

If we hypothesize that the first $K(\leq M+1=10)$ IMFs of the speech signal prominently manifests its vocal tract characteristics or system information, the LTL cepstrum of the combination of these K IMFs should be similar to that of the speech signal. Hence, the Euclidian distance between the two LTL cepstra should be less. The Euclidian distance between the two LTL cepstra is represented by,

$$\bigcap_K = \sqrt{|| \sqcap [s(n)] - \sqcap [\sum_{k=1}^{K} h_k(n)] ||_2}\ ,\ K = 1, 2, ..., M+1 = 10 \tag{4.29}$$

Conversely, if the higher-order IMFs, starting from IMF_K, carries mainly glottal source information, the HTL cepstrum of their combination should exhibit strong similarity with that of the speech signal. Hence, the Euclidian distance between the two HTL cepstra should be less. The Euclidian distance

between the two HTL cepstra is represented by,

$$\bigcup_K = \sqrt{|| \sqcup [s(n)] - \sqcup [\sum_{k=K}^{M+1=10} h_k(n)] ||_2}\ ,\ K = 1, 2, ..., M+1 = 10 \tag{4.30}$$

Using the above two parameters, we wish to evaluate if there is a certain K^{th} IMF which bifurcates the source and system information of the speech signal.

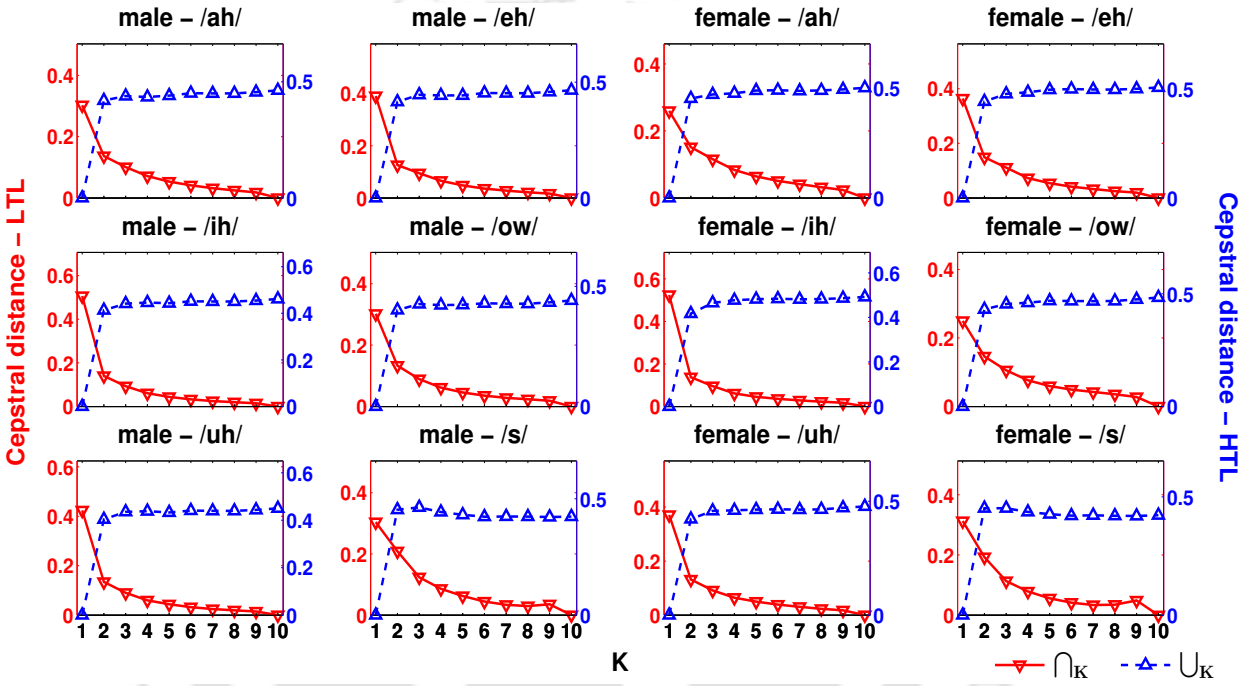


Figure 4.11: Average values of $\bigcap_K$ (Cepstral distance-LTL) and $\bigcup_K$ (Cepstral distance-HTL) for six different phones uttered by the male and female speakers of the TIMIT corpus. The IMFs of the speech signals are obtained using EMD.

Figure 4.11 shows the average values of $\bigcap_K$ (left axis) and $\bigcup_K$ (right axis), for the six phones of the TIMIT corpus. The IMFs, in this case, are extracted using EMD. The two cepstral distances are evaluated separately for the male and female speakers. Considering the male speakers first, we can observe a very similar pattern for both the LTL and HTL cepstral distances across all the five phones representing voiced speech. There is a sharp decrease in the $\bigcap_K$ value from $K = 1$ to $K = 2$, after which the curve saturates. This indicates that most of the system information is contained within the first two IMFs. Of course, with every additional IMF the combination becomes more and more similar to the speech signal, and hence the cepstral distance decreases. For $K = 10$, all the IMFs are combined, which results in the speech signal itself, and thus the distance is 0. Moving on to the $\bigcup_K$ vs. K curves, we observe that for $K > 2$, the curves are flat with minor fluctuations. Thus, not much

can be inferred from these curves. At higher values of K (> 6), the combinations of the IMFs are low-frequency signals which may be even below the pitch frequency of the signal. Hence, these values are not useful for our analysis. In the intermediate range ($K = 3$ to $K = 6$), the cepstral distance remains almost the same, which is an indication that the intermediate IMFs manifest similar glottal source information (pitch frequency or their harmonics). For $K \leq 2$, the combination starts becoming similar to the speech signal. At $K = 1$, the combination is equivalent to the speech signal, and hence the HTL cepstral distance is 0.

The five phones (voiced) uttered by the female speakers show patterns of cepstral distances which are very similar to that of the male speakers. One may argue that the $\bigcap_K$ vs. K curves, for the female speakers, starts saturating after $K = 3$ (instead of $K = 2$ for the male speakers). This is particularly true for the /ah/ and /ow/ phones. Thus, there is a better segregation of the vocal tract resonances in female speech, at least for certain phones. Overall, the observations made for the five voiced speech phones in the case of the female speakers are the same as that of the male speakers.

Thus, in general, considering both the male and female speakers, the LTL cepstral distances show that the first three IMFs, irrespective of the speech phone (voiced), manifest the system characteristics entirely. Thus, irrespective of the differences in the waveform shapes, the lower-order IMFs cumulatively still hold the system information. The distances evaluated on the HTL cepstra do not throw much light on which subset of the IMFs carries source information predominantly. We can only hypothesize that the observations regarding synthetic speech hold true in the case of natural voiced speech, in which case the glottal source characteristics are spread out amongst the higher-order IMFs.

For the phone /s/ (representing unvoiced speech), for both the male and female speakers, the LTL cepstral distances drop gradually (compared to the case of the voiced speech phones). Saturation may be observed after $K = 4$ (instead of $K = 2$ or $K = 3$ for voiced speech phones). Thus, as was observed in the case of synthetic speech signals, the system information is better separated amongst the lower-order IMFs in the case of unvoiced speech than that of voiced speech. The curves for the HTL cepstral distances, again, are inconclusive.

Figure 4.12 is the equivalent to Figure 4.11, except that the IMFs, in this case, are obtained using MEMD. For the male speakers, as one can see from the five voiced speech phones, the $\bigcap_K$ vs. K curves saturate after $K = 3$. Thus, the system information is better segregated amongst the lower-order IMFs

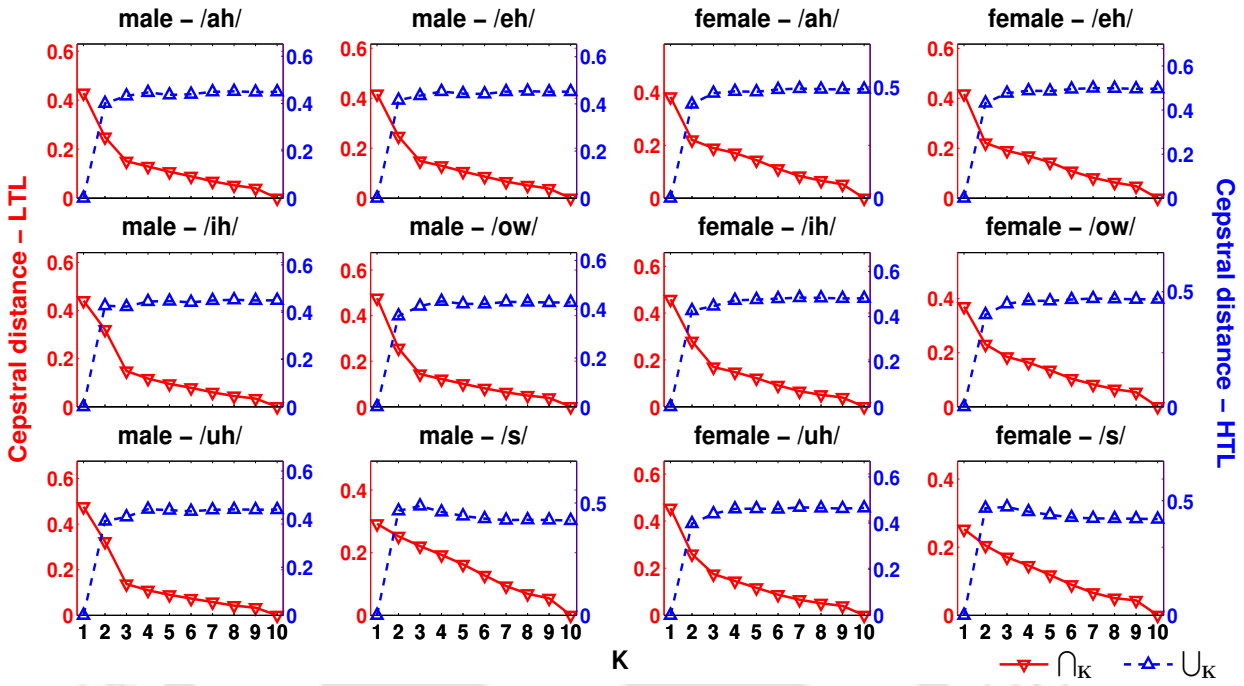


Figure 4.12: Average values of $\bigcap_K$ (Cepstral distance-LTL) and $\bigcup_K$ (Cepstral distance-HTL) for six different phones uttered by the male and female speakers of the TIMIT corpus. The IMFs of the speech signals are obtained using MEMD.

of MEMD than that of EMD (where the curves saturated for $K > 2$). The $\bigcup_K$ vs. K curves are again flat, and hence do not provide much insight into the distribution of source information amongst the IMFs. The curves for the female speakers are similar to that of the male speakers. However, the decline in the LTL cepstral distances is less sharp, particularly for the /ow/ and /uh/ phones. Thus, as observed in the case of EMD, the system information for female speech signals is better distributed amongst the lower-order IMFs of MEMD. For the unvoiced speech phone, /s/, the LTL cepstral distances decline gradually, indicating a very equitable distribution of system information. The $\bigcup_K$ vs. K curves are flat, and hence inconclusive.

Figure 4.13 presents the LTL and HTL cepstral distances for the six phones, where the IMFs of the speech signals are obtained using ICEEMDAN. The observations, in this case, are very similar to that of MEMD. For the five voiced speech phones, the $\bigcap_K$ vs. K curves, in the case of both the male and female speakers, indicate that the system information is predominantly contained in the first three IMFs. For the /s/ phone, again, the LTL cepstral distances decrease gradually, indicating that the system information is more equitably spread amongst the lower-order IMFs. The HTL cepstral distances do not reveal much about the distribution of the source characteristics.

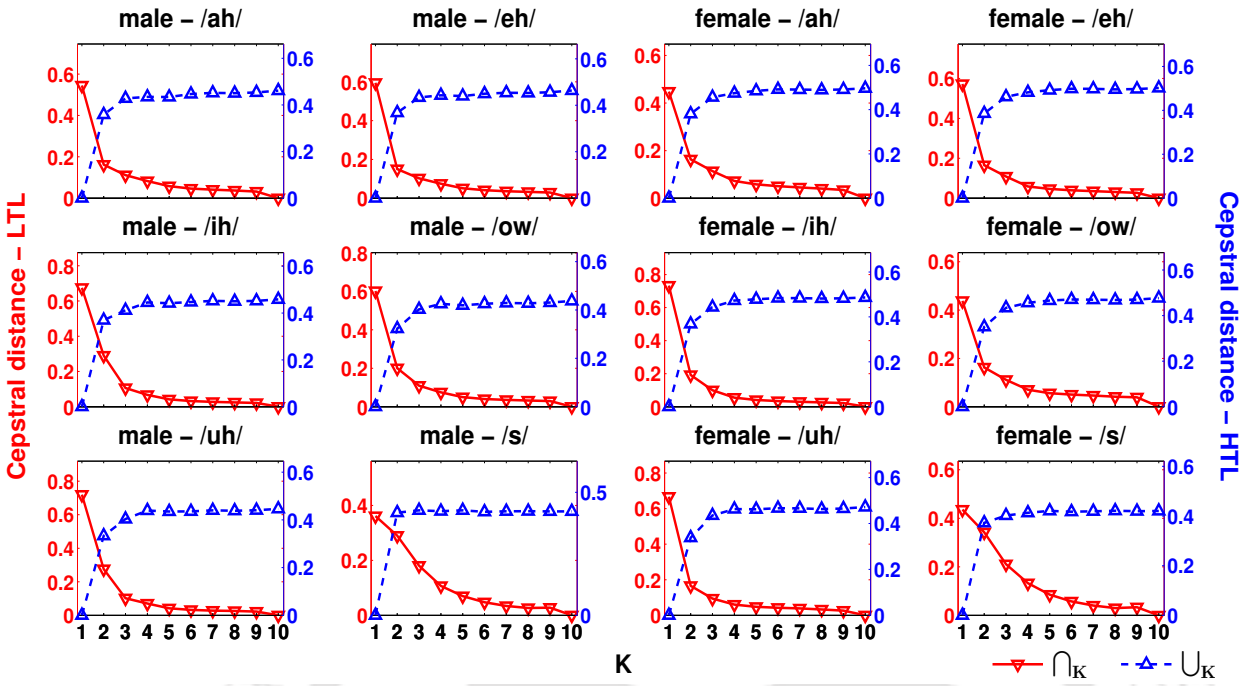


Figure 4.13: Average values of $\bigcap_K$ (Cepstral distance-LTL) and $\bigcup_K$ (Cepstral distance-HTL) for six different phones uttered by the male and female speakers of the TIMIT corpus. The IMFs of the speech signals are obtained using ICEEMDAN.

4.3.3 Source-system separation of telephone quality speech

The information contained in the speech signal may not be merely attributed to its waveform shape. That is why the speech signal can be compressed or modified by coders, transmitted through noisy channels, and still be perfectly interpreted at the receiver. In other words, the critical information regarding the source and system characteristics of the speech signal is still preserved, even when the waveform shape changes. EMD, and its variants, however, act on the signal waveform only. Therefore, it is interesting to verify how well the IMFs are able to represent the source and system information of the speech signal when it is subjected to telephone channel codecs. For this study, the speech signals of the TIMIT corpus are band-pass filtered in accordance with *International Telecommunication Union* (ITU) standards [160]. As is well-known, telephone channel codecs suppress both the lower-frequency spectrum and the higher frequency spectrum. They primarily dilute the pitch and glottal source characteristics embedded in the speech signal. In this thesis, we also refer to ITU standard telephone quality speech as **T-2** type of telephone quality speech. Hence, for every speech segment, $s(n)$, a corresponding ITU-coded speech segment, $s^T(n)$ is obtained. The corresponding IMFs are denoted as $h_k(n)$ and $h_k^T(n)$, respectively, where $k = 1, 2, ..., M + 1 = 10$.

To represent the source and system information contained in the speech signals, and their IMFs, we again evaluate their LTL and HTL cepstra, as discussed previously. However, our objective here is to verify if the source and system information embedded in the k^{th} IMF of $s^T(n)$ is significantly different from that of $s(n)$. For this purpose, the Euclidian distances between the cepstra of the normal and telephone quality speech signal, and that of their corresponding IMFs, are evaluated as follows :

$$\bigwedge^{T} = \sqrt{||\ \sqcap[s(n)] - \sqcap[s^T(n)]\ ||_2}\ ,\ \bigvee^{T} = \sqrt{||\ \sqcup[s(n)] - \sqcup[s^T(n)]\ ||_2} \tag{4.31}$$

$$\bigwedge_k^{T} = \sqrt{||\ \sqcap[h_k(n)] - \sqcap[h_k^T(n)]\ ||_2}\ ,\ \bigvee_k^{T} = \sqrt{||\ \sqcup[h_k(n)] - \sqcup[h_k^T(n)]\ ||_2}\ ,\ k = 1, 2, ...M+1 = 10 \tag{4.32}$$

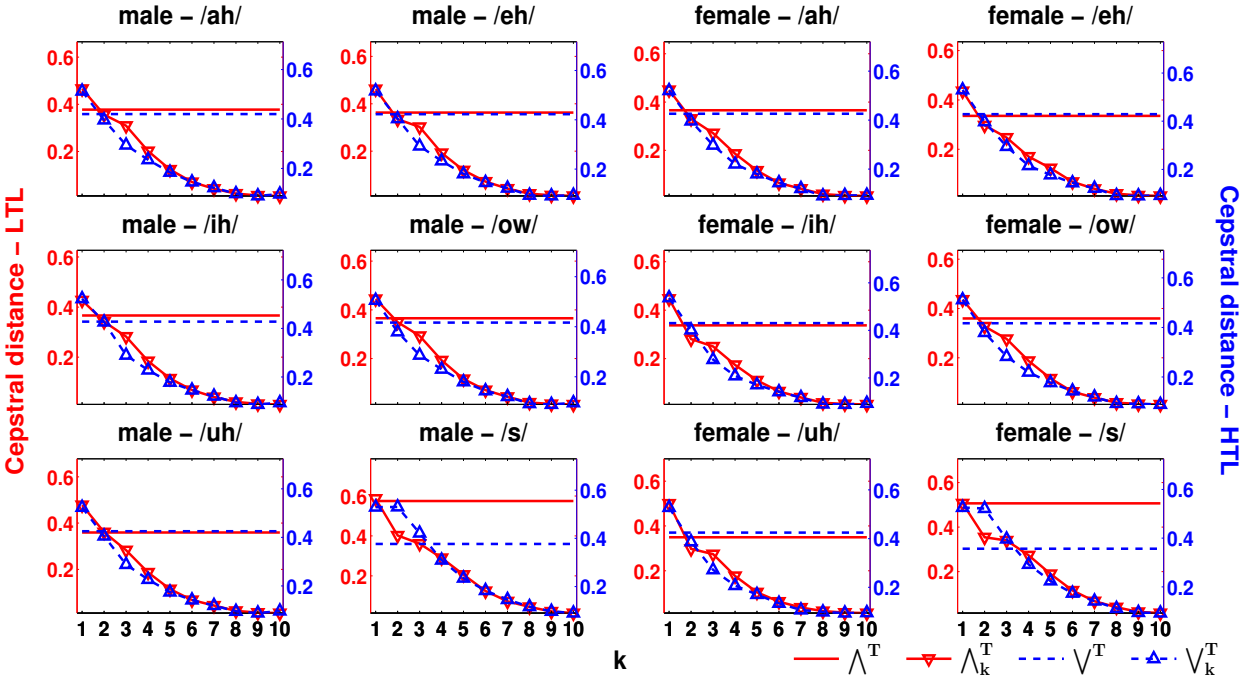


Figure 4.14: Average values of $\bigwedge^T$, $\bigwedge_k^T$ (Cepstral distance-LTL) and $\bigvee^T$, $\bigvee_k^T$ (Cepstral distance-HTL) for six different phones uttered by the male and female speakers of the TIMIT corpus. The IMFs of the speech signals are obtained using EMD.

Figure 4.14 shows the average LTL and HTL cepstral distances for the six phones, for the male and female speakers (separately) of the TIMIT database. The horizontal solid line and dashed line represent the average LTL and HTL cepstral distances, respectively, between the original speech signal, and its T-2 version. They act as benchmarks ($\bigwedge^T$ and $\bigvee^T$) against which the cepstral distances between the individual IMFs (obtained using EMD) may be compared. First, considering the five voiced phones

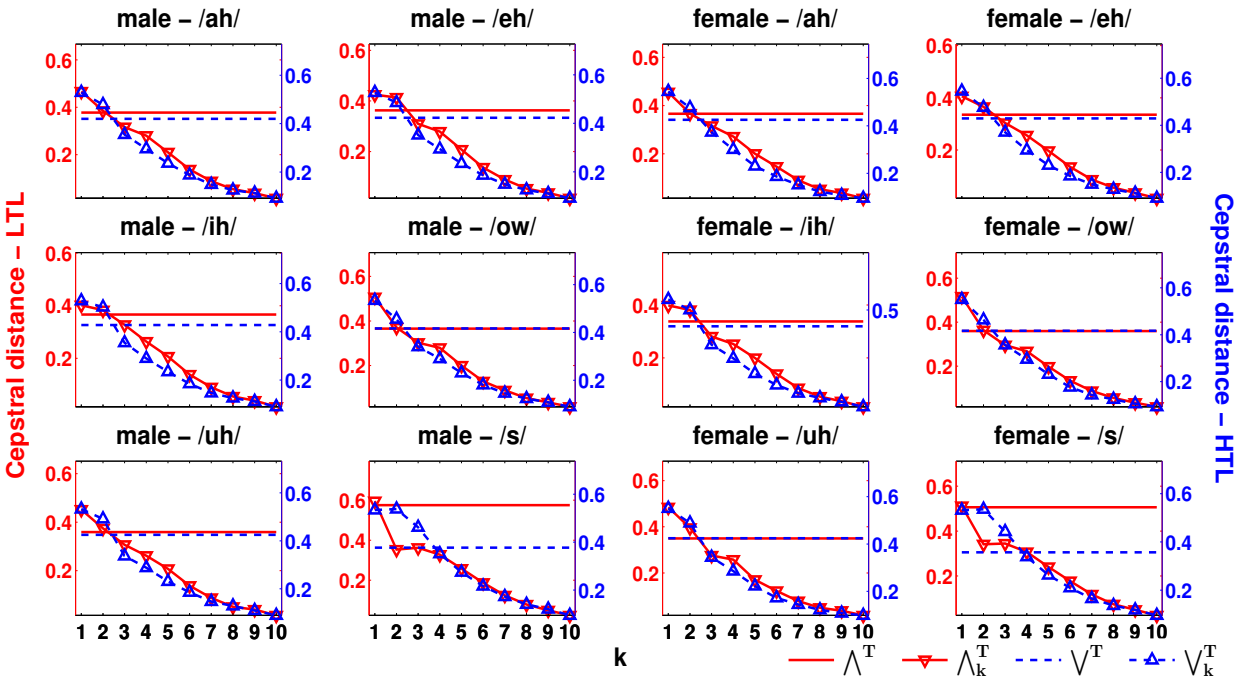


Figure 4.15: Average values of $\bigwedge^T$, $\bigwedge_k^T$ (Cepstral distance-LTL) and $\bigvee^T$, $\bigvee_k^T$ (Cepstral distance-HTL) for six different phones uttered by the male and female speakers of the TIMIT corpus. The IMFs of the speech signals are obtained using MEMD.

uttered by the male speakers, we observe that excepting the first IMF, the LTL and HTL cepstral distances between the corresponding IMFs are lesser than that of their respective benchmarks. The $\bigwedge_k^T$ and $\bigvee_k^T$ values are lower than that of $\bigwedge^T$ and $\bigvee^T$, respectively, for $k \geq 2$. Hence, the first IMF differs the most out of all the IMFs, due to the influence of telephone channel codecs. This is expected as most of the system information pertaining to the higher-frequency spectrum (which is influenced by telephone channel codecs) is manifested in the first IMF (as we have discussed in the previous sub-sections). The high HTL cepstral distances for IMF_1 is just another manifestation of the differences in the waveform shapes. Interestingly, even though the lower-frequency spectrum of speech is affected by telephone channel codecs, the $\bigvee_k^T$ values for the higher-order IMFs are low. Moving onto the voiced speech phones of the female speakers, they replicate the patterns observed in the case of the male speakers. For the unvoiced speech phone, /s/, for both the male and female speakers, the $\bigwedge^T$ value (benchmark) is higher than that of any of the voiced speech phones. This is expected as unvoiced speech is supposed to manifest strong vocal tract resonances at very high frequencies. Hence, suppression of the higher-frequency spectrum by telephone channel codecs is expected to have a significant influence on the LTL cepstrum of the speech signal. Again, barring the first IMF, for

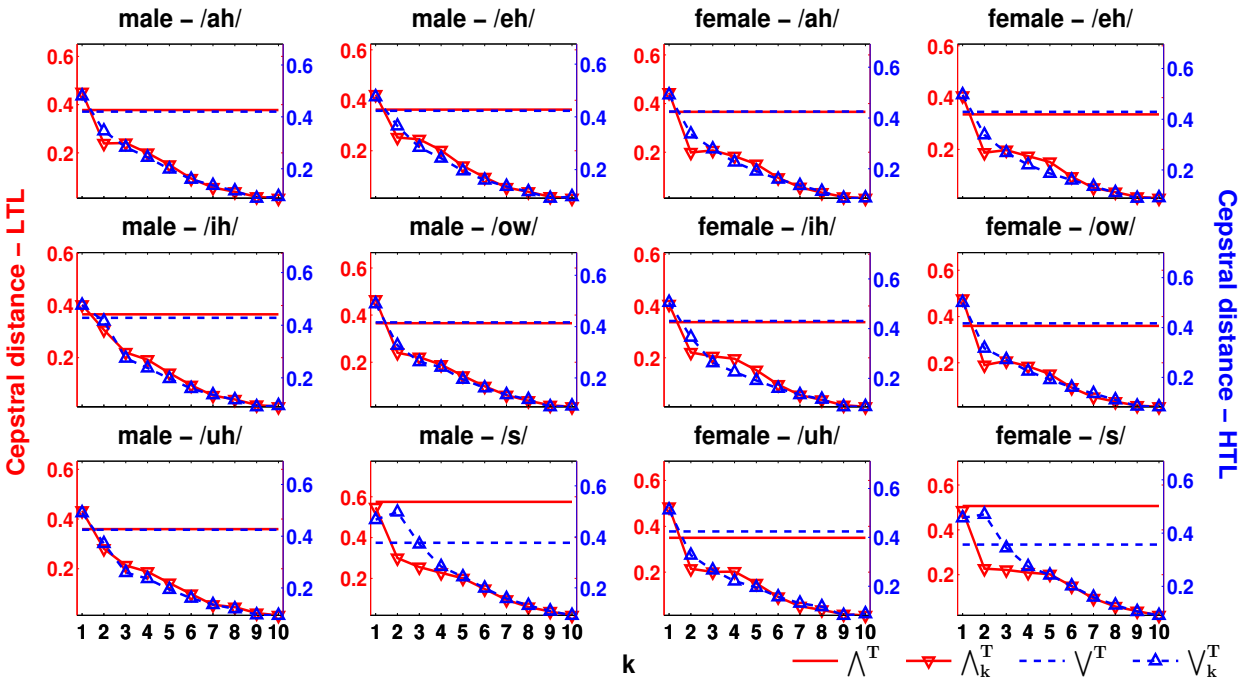


Figure 4.16: Average values of $\bigwedge^T$, $\bigwedge_k^T$ (Cepstral distance-LTL) and $\bigvee^T$, $\bigvee_k^T$ (Cepstral distance-HTL) for six different phones uttered by the male and female speakers of the TIMIT corpus. The IMFs of the speech signals are obtained using ICEEMDAN.

$k \geq 2$, the LTL cepstral distances are below the benchmark. As unvoiced speech is supposed to manifest noise-like source characteristics, the analysis of the $\bigvee_k^T$ vs. k curves are not useful in this case. Nevertheless, the higher-order IMFs manifest lesser cepstral distances.

Figure 4.15 is the equivalent to Figure 4.14, for the IMFs being extracted using MEMD. The observations, for both the male and female speakers, are similar to that of EMD. However, for the voiced speech phones, the LTL and HTL cepstral distances corresponding to the first two IMFs (instead of the first IMF only in the case of EMD) are now above their corresponding benchmarks. For the unvoiced speech phone, /s/, the observations remain the same.

Figure 4.16 shows the LTL and HTL cepstral distances, with the IMFs being obtained using ICEEMDAN. Interestingly, the curves, in this case, are more similar to the case of EMD instead of being similar to that of MEMD. For the voiced speech phones, for both the male and female speakers, only the cepstral distances corresponding to the first IMF is greater than that of the benchmarks. One would have expected that as the behaviour of ICEEMDAN and MEMD are similar, the second IMF would also have higher cepstral distances. Moving on, the HTL cepstral distances of the higher-order IMFs are quite low, and the curves for phone /s/ are more or less the same with respect to EMD and

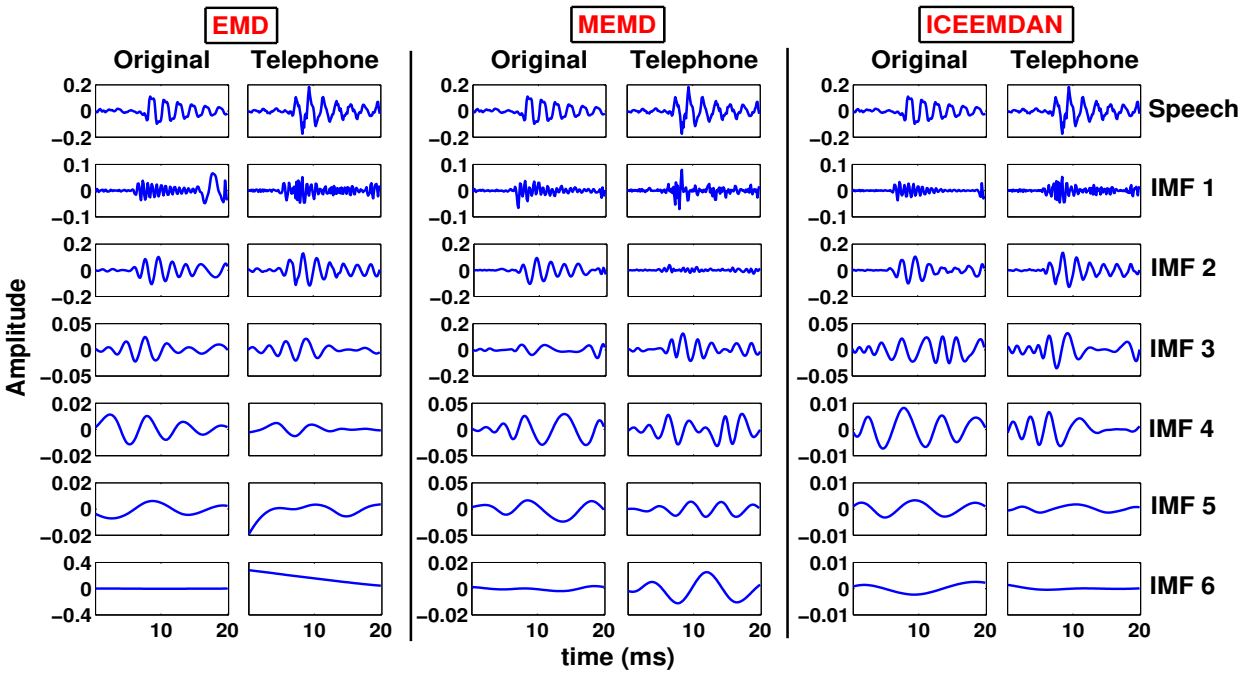


Figure 4.17: The first 6 IMFs obtained from an arbitrary 20 ms voiced speech segment, and its T-2 version (telephone quality speech). The speech segment corresponds to the phone /ah/ of a male speaker. The IMFs are obtained using EMD, MEMD and ICEEMDAN.

MEMD.

The observations made in Figures 4.14, 4.15 and 4.16 complement those made in Section 2.1.4. In Section 2.1.4, the decomposition of a binary mixture of sinusoids using EMD was investigated. It was observed that the EMD decomposition is not adversely affected if the high-frequency component of the binary mixture overrides its low-frequency component. However, when the high-frequency sinusoid is subdued by the low-frequency component, the ability of EMD to segregate the components diminishes. The same observations obviously apply to both MEMD and ICEEMDAN. Hence, considering the current experiments, even though the telephone codecs suppress the lower-frequency spectrum, the higher-order IMFs (obtained from EMD/MEMD/ICEEMDAN) are still able to manifest the source characteristics of the speech signal credibly. Hence, $\bigvee_k^T$ values are low (lower than the benchmark - $\bigvee^T$) even for the higher-order IMFs. The weakening of the higher-frequency spectrum, however, does change the system characteristics manifested in the lower-order IMFs, particularly IMF_1, which manifests most of the higher-frequency resonators.

Having observed the differences in the IMFs, in terms of their source and system characteristics, it is necessary to observe the differences in the actual waveforms themselves, to get the complete picture.

Figure 4.17 shows the first six IMFs of a natural voiced speech signal and its T-2 version. It may be observed, in the case of EMD, that IMFs 2 and beyond have similar waveform shapes for both the original and T-2 speech signals. Mode-mixing is observed in IMF_1 of the original speech signal, which is largely eradicated in the case of the T-2 speech signal. This, again, shows that weakening of the lower-frequency spectrum helps in obtaining a better decomposition. The higher frequency spectrum of the speech signal is better represented (even though ITU codecs suppress the higher-frequency spectrum also). This is because most of the energy in voiced speech is contained in its lower frequency spectrum (due to the low-pass nature of the glottis and the strong low-frequency resonators). The suppression of the lower-frequency spectrum by telephone codecs, as such, has a bigger impact on the slope of the spectrum, than the suppression of the higher-frequency spectrum. The rest of the IMFs of the T-2 speech signal exhibit similar waveform shapes with respect to that of the original speech signal. Moving on to the case of ICEEMDAN, again, the waveform shapes of the IMFs of the T-2 speech signal are similar to that of the original speech signal. Visually, it is difficult to comment on the high-frequency waveforms represented by IMF_1. Moving onto the case of MEMD, it is observed that the first two IMFs of the T-2 speech signal manifest high-frequency fluctuations, whereas only the first IMF of the original speech signal manifests high-frequency fluctuations. The differences in the waveforms of the first two IMFs are quite obvious, and complements the observations of Figure 4.15. The weakening of the lower-frequency spectrum of the speech signal enables MEMD to extract higher-frequency information from the speech signal more proficiently. As such, even IMF_2 manifests high-frequency fluctuations in the case of the T-2 speech signal. The rest of the IMFs of the T-2 speech signal exhibit more similarity with those of the original speech signal.

We may now summarize the observations made from the experiments on natural speech signals :

- **(i)** In the case of the voiced speech signal, the first few IMFs mostly capture the system information, irrespective of what phone or sound it corresponds to. In the case of EMD, for the male speakers, the first two IMFs manifest most of the system information. In the case of the female speakers, the first three IMFs manifest most of the system information. In the case of MEMD and ICEEMDAN, irrespective of the gender or the phone, the first three IMFs manifest most of the system characteristics. Thus, ICEEMDAN and MEMD provide a better and more equitable decomposition of the speech signal.

- **(ii)** In the case of the voiced speech signal, for some phones, a better distribution of the system

information (amongst the first few high-order IMFs) is observed in the case of the female speakers, compared to the male speakers.

- **(ii)** In the case of the unvoiced speech signal, irrespective of the gender, the system information is distributed amongst the first four IMFs, in a very equitable manner. Thus, a better distribution of the system information is observed in the case of unvoiced speech, compared to voiced speech.

- **(iv)** When the speech signal is subjected to telephone channel codecs, apart from the first few lower-order IMFs, which manifest the higher-frequency system characteristics, the rest of the IMFs are relatively unaffected. In the case of EMD and ICEEMDAN, only IMF_1 seems to undergo significant change, whereas in the case of MEMD the first two IMFs show notable changes. As such, the source information seems to be well represented in the higher-order IMFs, even though telephone channel codes suppress the lower-frequency spectrum of the speech signal.

4.4 Conclusion

The principal objective of this chapter was to investigate in detail whether EMD and its variants behave in a consistent manner with respect to changes in the speech waveform. As such, at first, controlled experiments were performed on voiced and unvoiced speech signals, synthesized based on the source-filter theory. Strong correlations were observed between the lower-order IMFs, and the strong resonators used to synthesize the speech signal. The IMFs which immediately succeeded these lower-order IMFs were observed to have strong correlations with the glottal source constituent used to synthesize the speech signal. Overall, a separation of the source and system information was observed in the IMFs, irrespective of the changes in the pitch frequency, and the center frequencies and bandwidths of the resonators. However, the separation of the resonators into different IMFs was better achieved for higher pitch frequencies, and when the higher-frequency resonators are comparable in strength to (or stronger than) the lower-frequency resonators.

Following the observations made on synthetic speech signals, experiments were conducted on natural speech signals, corresponding to different phones or speech sounds. The well-known method of cepstral or homomorphic analysis was used to examine the source and system characteristics manifested in the IMFs. It was observed that the system information is predominantly manifested in the first few lower-order IMFs. A more equitable distribution of the system information was

observed for female speakers (who have higher pitch frequencies than male speakers), and unvoiced speech (which has strong high-frequency resonators). Cepstral analysis, unfortunately, could not provide much insight into the distribution of source information amongst the IMFs. However, as the distribution of system information in the case of natural speech showed strong parallels with that of synthetic speech, the same could be expected in the case of the source characteristics.

The experiments on synthetic and natural speech signals showed that EMD (and its variants) adapts to the nature of the signal. The decomposition is different for voiced and unvoiced speech signals. However, for speech signals with similar characteristics (different voiced speech phones), even though the waveform shapes might look dissimilar, there is strong similarity in the behaviour of the decomposition (how the source and system characteristics are manifested). As speech is an important aspect of communication technology, it is important to investigate the behaviour of the decomposition when the speech signal is subjected to telephone channel codecs. With this objective, the IMFs of the original speech signal, and its telephone quality version were compared with one another. Based on cepstral analysis and the waveform shapes of the IMFs, it was observed that apart from the first few lower-order IMFs (which manifest the higher-frequency system information), the rest of the IMFs of the telephone quality speech signal had strong similarities with that of the original speech signal.

Overall, the experiments performed as part of this chapter show that even though EMD and its variants principally decompose the signal based on its waveform shape, the nature of the decomposition is still relatively consistent. Until the speech signals exhibit starkly different characteristics (voiced speech vs. unvoiced speech), the characteristics of their IMFs are similar. Even for the same speech sound, as long as the principal characteristics (intelligibility) are maintained, even though the waveforms change (original speech vs. telephone quality speech), the IMFs are still notably similar. This indicates that any process developed for the IMFs of any arbitrary speech signal should also be largely applicable to any other speech signal. Based on this observation, in the next two chapters, we would utilize the IMFs of the speech signal for two speech processing applications - (i) *detection of the Glottal Closure Instants of voiced speech*, and (ii) *extracting features from the IMFs for Speaker Verification*. The performances would be evaluated not just for clean natural speech signals, but speech signals affected by noise and telephone channel conditions, and for different speaking styles.

5

Detection of the Glottal Closure Instants using EMD

Contents

Outline

The objective of this work is to explore an alternative to short-time LP analysis, for detecting the *Glottal Closure Instants* (GCIs) of the speech signal. With this objective, the IMFs of the speech signal, obtained from two variants of EMD - ICEEMDAN and MEMD - are investigated for determining its GCIs. ICEEMDAN and MEMD, both effectively curtail *mode mixing*, a drawback of classical EMD. The partial summation of a certain combination of these IMFs, while discarding certain high-frequency and low-frequency IMFs, results in a signal which emphasizes the source information of the speech signal, and could be used for estimating the GCIs. Based on this principle, methods are devised for finding this subset of IMFs and using them effectively for estimating the GCIs in voiced speech. The results, evaluated on the 5-speaker CMU Arctic database, and the 10-speaker APLAWDW database, reveal that the *ICEEMDAN based GCIs Estimation* (IGE), is comparable to the state-of-the-art methods, particularly under noisy, and telephone channel conditions. The results also show that the *MEMD based GCIs Estimation* (MGE) is able to achieve performances similar to that of IGE, particularly for telephone quality speech, and at only a fraction of the time cost of IGE.

5.1 Introduction

The detection of the *Glottal Closure Instants* (GCIs), or the *instants of significant excitation in voiced speech* has been given considerable importance in speech processing [5–11, 21, 39]. They are the instants at which the speech production apparatus is excited by an impulsive signal, generated by the abrupt closure of the vocal folds in the glottis, during the production of voiced speech [5–11, 21, 39]. As has been already discussed, despite the non-linear and non-stationary characteristics of the speech signal, the *source-filter* theory of speech production has remained the basis for speech analysis. Thus, LP analysis, and the LP residual in particular, has remained the cornerstone in a multitude of efforts, for the task of estimating the GCIs [5, 6, 8, 9, 11, 21, 39]. One of the first methods to utilize the LP residual in estimating the GCIs was the *Epoch Filter* (EF) [21]. In this process, the LP residual spectrum is whitened by multiplying with a Hanning window. The large swings around the GCIs, in the whitened LP residual, are further reduced by taking its Hilbert envelope. Following this method, algorithms like the *Group Delay* (GD) [39], *Hilbert Envelope and Group Delay* (HEGD) [5], *Dynamic Programming and Phase Slope Algorithm* (DYPSA) [6], *Speech Event Detection using the Residual Excitation And a Mean-based Signal* (SEDREAMS) [8], *Yet Another GCI detection Algorithm* (YAGA)

[9], and *Integrated Linear Prediction Residual using Plosion Index* (ILPR-PI) [11], all of which depend on LP analysis, have been developed.

The GD method evaluates the average slope of the phase spectrum, called the *phase slope function*, at each time instant of the LP residual, to construct a *sinusoid-like* waveform, the positive going zero-crossings of which coincide with the GCIs. The HEGD and DYPSA try to refine the GD method by minimizing its computational cost, and the number of spurious GCIs estimated from it, respectively. The YAGA is the same as the DYPSA algorithm but combines Wavelet Transform (WT) and LP analysis to derive a cleaner signal than the LP residual as a representation of the excitation source. The SEDREAMS algorithm uses a moving average filter to obtain a *sinusoid-like* waveform from the speech signal, where prospective regions of the GCIs are defined. Within these regions, the LP residual is used to estimate the GCIs. The ILPR-PI combines LP analysis with a non-linear operator, called the Dynamic Plosion Index, for estimating the GCIs. The *Zero Frequency Resonator* (ZFR) [7] is probably the first significant deviation from using LP analysis for estimating the GCIs. By using a cascade of two marginally stable 0-Hz resonators, on the pre-emphasized speech signal, the ZFR obtains an exponentially increasing/decreasing output. When the trend of such an output is removed in short segments of one to two pitch periods, the detrended signal resembles a *sinusoidal* signal, whose positive going zero-crossings coincide with the GCIs.

The above discussion reflects how useful a *sinusoidal representation of the source signal* is for the purpose of detecting the GCIs. In many of the popular methods, as discussed above, a sinusoid-like waveform is used directly or indirectly for estimating the GCIs. As most of these methods rely on LP analysis, they are prone to its limitations. The processing of the speech signal based on the *source-filter theory* makes *short-time processing* necessary, i.e., the speech signal is processed in small segments/frames which are assumed to be produced by a linear and stationary process. Again, when the speech signal is affected by noise, the LP analysis becomes more erroneous, and hence the performances of these techniques diminish noticeably [11]. Even the ZFR, which is free from the limitations of LP analysis, is observed to have a limited performance for telephone quality speech, where the low-frequency spectrum of speech is highly attenuated [11]. Thus, while most of these techniques perform excellently under clean and controlled data conditions, when the speech signal is subjected to external influences they are not as effective [10, 11]. Again, it is generally observed that while a technique performs credibly in a certain condition, like a certain type and level of noise, its

performance diminishes rapidly when the speech signal is affected by a different set of conditions. There is a significant gap in the performances of the techniques from one scenario to another [10, 11].

To summarize the above discussion, the processing of the speech signal under the assumptions of short-time stationarity and linearity is the principal reason for the limitations of the LP residual based techniques. As such, it is our hypothesis that using non-linear and non-stationary signal analysis on the speech signal should provide better performances under varied conditions. Henceforth, this work is focussed on realizing a single objective - *to use some non-linear and non-stationary signal analysis method to obtain sinusoid-like signals directly from the speech signal, and utilize them to develop a robust technique for detecting the GCIs under different scenarios.* Thus, the task here is to develop a technique whose performance does not fluctuate starkly when the speech signal is subjected to diverse types and levels of noise, and telephone channel conditions.

With the above objective in mind, we investigate, in this work, the effectiveness of the various AM-FM components of the speech signal, as obtained from EMD, for the purpose of estimating the GCIs (of the glottal source) of its voiced speech regions. In particular, in this work, we study the utility of the AM-FM components or the IMFs of two variants of EMD, which effectively curtail the phenomenon *mode-mixing* observed in standard EMD, as discussed in Chapter 2. The first variant, ICEEMDAN, discussed in Section 2.2.1, is very effective in reducing *mode-mixing*, but at a large time-cost. The other variant, MEMD [161], proposed in Chapter 3, reduces *mode-mixing* less effectively, but provides a great time-advantage. Moreover, as we have observed in Chapter 4, ICEEMDAN and MEMD exhibit a better ability to extract the latent characteristics of the speech signal compared to EMD. As such, this work proposes two methods (but using a common principle), which utilizes the IMFs obtained from ICEEMDAN and MEMD, respectively, for estimating the GCIs of the speech signal.

The rest of the chapter is organized as follows : Section 5.2 discusses the principle used for detecting the GCIs, using the IMFs of ICEEMDAN and MEMD. Sections 5.3 and 5.4 describe the methods used for detecting the GCIs using the IMFs of ICEEMDAN and MEMD respectively. Section 5.5 presents two methods which mimic the proposed methods, but uses simple band-pass filtering. Section 5.6 presents the performances of the proposed algorithms, and compares them with that of the standard methods. Section 5.7 concludes this work.

5.2 Principle of estimating the GCIs

Let us consider a synthetic voiced speech signal, $s(n)$, constructed using the source-filter theory (using the framework discussed in Section 4.2) at $F_s = 8$ kHz. The input excitation, $e(n)$, is modelled as a uniform train of impulses of unit strength, having $F_0 = 100$ Hz. The vocal tract system is modelled as a cascade of four resonators. The resonant frequencies (f_rs) are considered as 700, 1200, 2400 and 3600 Hz, and the resonators bandwidths (B_rs) are considered as 70, 140, 210 and 280 Hz. The resultant synthetic speech signal sounds like the vowel /a/. Figure 5.1(a) shows the speech signal so constructed, and Figure 5.1(b)-(h) the first seven IMFs derived from it using ICEEMDAN. IMFs 8-10 are not shown in the figure. It is evident from the figure that as the order of the IMF increases, the frequency of the dominant oscillation reflected by it decreases. The *mean frequency* (F^m) of each IMF, which is a *measure of the dominant frequency reflected in the IMF*, is mentioned above the plot of the IMF in Figure 5.1. The mean frequency of IMF$_k$ ($F_k^m,\ k = 1, 2, ...7$) is given by equation (2.28). It is calculated from its power spectrum (squared magnitude spectrum), $S_k(f)$, as,

$$F_k^m = \sum_{f=0}^{F_s/2} \frac{f \times S_k(f)}{\sum_{f=0}^{F_s/2} S_k(f)}\ ,\ k = 1, 2, ..., 7 \tag{5.1}$$

It can be seen that the mean frequencies of the IMFs decrease in a more or less *dyadic filterbank* nature. Correspondingly, as the IMF order increases, the IMFs become more sinusoidal, as seen in Figure 5.1. The vertical dashed lines in Figure 5.1(b)-(l) indicate the impulses of the input excitation, $e(n)$. It can be observed that the IMFs intersect the input excitation, $e(n)$, at different points of their near-sinusoidal curves, i.e., the IMFs have a different phase-shift with respect to the impulse locations. Out of these seven IMFs, the higher-frequency IMFs, IMF$_1$ and IMF$_2$, seem to represent the impulse excitation instants at some of their minima locations. However, there are other minima in the vicinity which could cause ambiguity. Similarly, IMF$_4$ also intersects $e(n)$ at some of its minima locations. However, it has stronger adjacent minima, which, as we will see later, might make the selection of the target minima difficult. IMF$_3$, Figure 5.1(d), as such, provides the best trade-off - it intersects $e(n)$ at its prominent minima, with good accuracy, and with much lesser ambiguity with the adjacent minima, compared to IMFs 1, 2 and 4. As such, IMF$_3$ may be described as an *oscillatory* or an *imperfect sinusoidal* signal, some of whose *minima locations* coincide with the impulse locations of the excitation signal. Thus, given an impulse location, $\{l_m^{ref} = m\frac{F_s}{F_0},\ m \in \mathbb{Z}\}$, IMF$_3$ may be represented

as,

$$h_3(n) \approx a_3(n)\cos[2\pi\frac{f_3^m}{F_s}(n - l_m^{ref}) + \pi] \ , \tag{5.2}$$

where $a_3(n)$ is the time-varying amplitude envelope of $h_3(n)$, and f_3^m the average frequency of the sinusoid over time. It is interesting to note that IMF_5, Figure 5.1(f), represents a near sinusoid with $F_5^m = F_0 = 100$ Hz, which shows the ability of ICEEMDAN to characterize glottal activity [106].

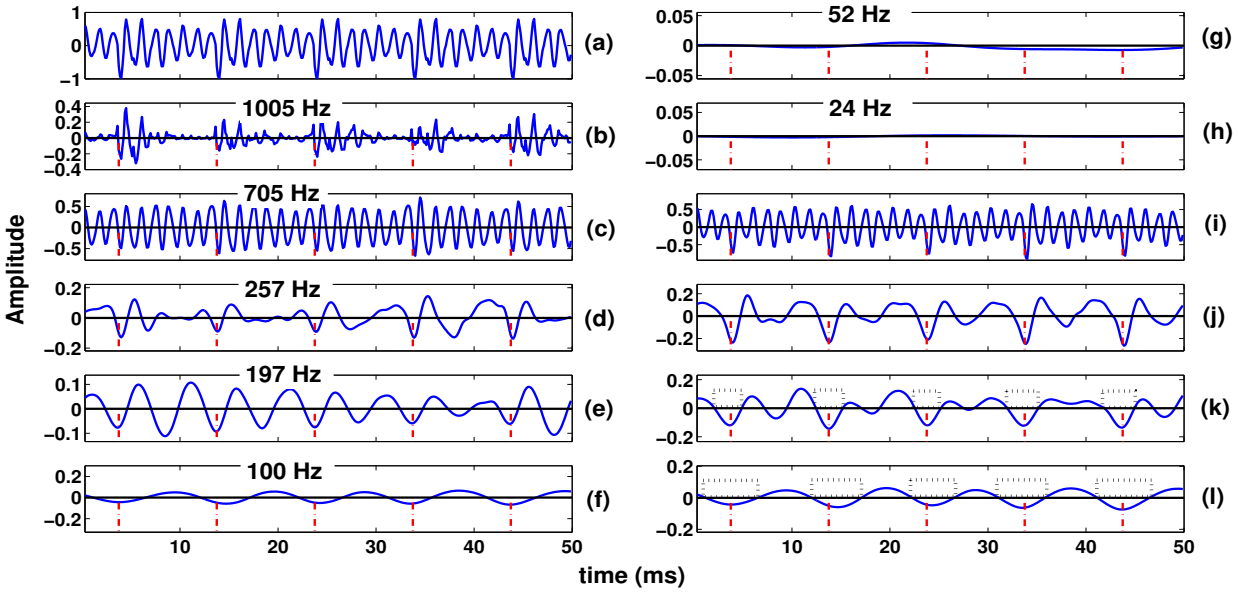


Figure 5.1: **(a) A segment of a synthesized vowel-like speech signal $s(n)$ having $F_0 = 100$ Hz ; (b)-(h) are IMFs 1-7 respectively obtained from ICEEMDAN of $s(n)$. IMFs 8-10 are not shown ; (i) sum of IMFs 2-6 ; (j) sum of IMFs 3-6 ; (k) sum of IMFs 4-6 ; (l) sum of IMFs 5-6 ; Dashed vertical lines indicate the impulse excitation $e(n)$. Dotted rectangles in (k) and (l) indicate regions of search for the GCIs.**

Figure 5.1(i)-(l) represent the sum of the IMFs $2-6$, $3-6$, $4-6$, and $5-6$, respectively. From a different perspective, Figure 5.1(i)-(l) also represent the sum of IMFs having different frequency content. Thus, Figure 5.1(l) represents the sum of the IMFs having mean frequency 50 Hz $\leq F^m \leq F_0 = 100$ Hz. Figure 5.1(k) represents the same for 50 Hz $\leq F^m \leq 2 \times F_0 = 200$ Hz, Figure 5.1(j) for 50 Hz $\leq F^m \leq 5 \times F_0 = 500$ Hz, and Figure 5.1(i) for 50 Hz $\leq F^m \leq 8 \times F_0 = 800$ Hz. As normal human speech does not have a pitch frequency below 50 Hz, hence the IMFs below 50 Hz, which are trend-like low amplitude signals, are left out of the summation. It can be seen that just like IMF_3, the sum of IMFs starting from IMF_3, Figure 5.1(j), manifests the impulse locations at its prominent minima, but with potentially less spurious minima in the neighbourhood. Thus, the signal represented by Figure 5.1(j) provides a better candidate for estimating the impulse locations of $e(n)$ than that by

IMF_3 of Figure 5.1(d). The challenge now is to identify the minima which coincide with the impulse train, amongst the multitude of minima in the auxiliary signal. For this purpose, we need to look at Figure 5.1(k),(l). It can be observed that in either of these two signals, the impulse locations lie within the regions defined by their positive and negative zero-crossings, as represented by dotted rectangles. Either of these signals, thus, represents a low-resolution signal for identifying the impulse excitations. Of course, in the case of Figure 5.1(k), some spurious regions will also be detected. Such regions need to be eliminated by a post processing mechanism, described later. However, in principle, once such regions are identified, the signal represented by Figure 5.1(j) may be used for high-resolution analysis. The locations of the minimum values of the signal in Figure 5.1(j), within the identified regions, gives the instants of impulse excitation.

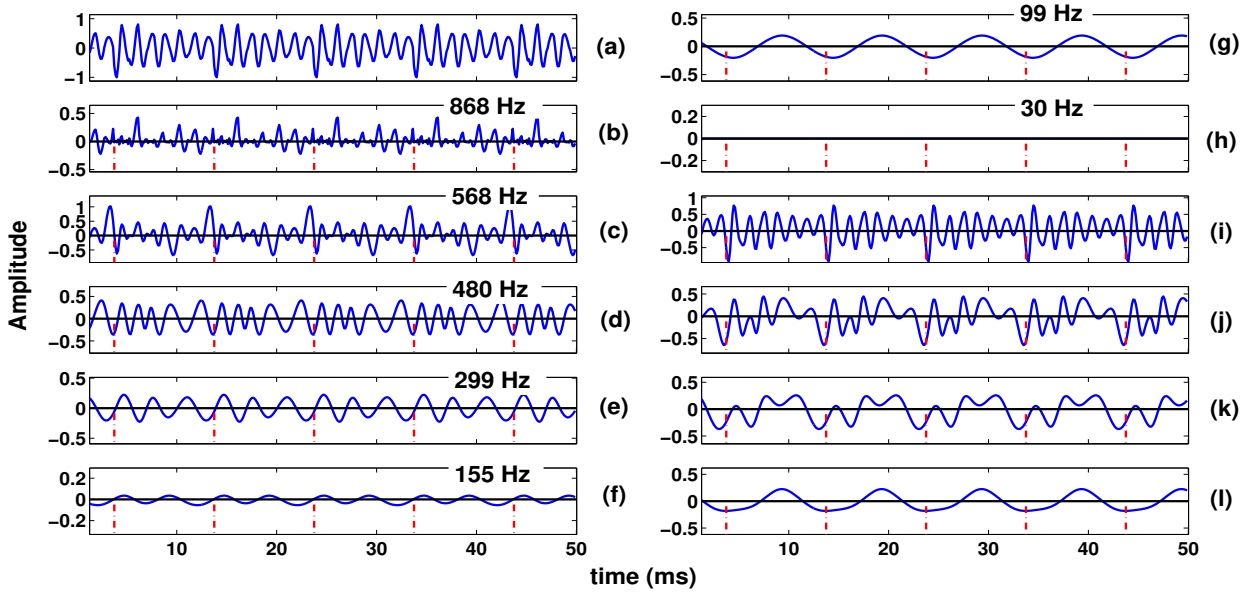


Figure 5.2: **(a) A segment of a synthesized vowel-like speech signal, $s(n)$, having $F_0 = 100$ Hz ; (b)-(h) are IMFs 1-7 respectively obtained from MEMD of $s(n)$. IMFs 8-10 are not shown ; (i) sum of IMFs 2-6 ; (j) sum of IMFs 3-6 ; (k) sum of IMFs 4-6 ; (l) sum of IMFs 5-6 ; Dashed vertical lines indicate the impulse excitation $e(n)$.**

Figure 5.2 represents the equivalent figure of Figure 5.1, but here the decomposition of the synthetic speech signal is done by MEMD. Figure 5.2(a) represents the synthetic speech signal, and Figure 5.2(b)-(h) its first seven IMFs. Figure 5.2(i)-(l) represent the sum of the IMFs $2-6$, $3-6$, $4-6$, and $5-6$, respectively. The vertical dashed lines in Figure 5.2(b)-(l) represent $e(n)$. As is observed in the figure, the lower-order IMFs segregate the higher-frequency content of the speech signal from its lower-frequency content (reflected in the higher-order IMFs). The noisy high-frequency nature of IMF_1 makes it difficult to estimate the excitation instants from it. Among the more sinusoidal IMFs,

IMF_2 and IMF_3 seem to represent the impulse excitations at some of their minima locations. IMF_3, Figure 5.2(d), has a more sinusoidal nature, and it is easier to associate this observation with it. Again, the sum of IMFs $3-6$, Figure 5.2(j), represents a signal some of whose minima coincide with (or lies in close proximity to) the excitation instants. Additionally, this signal (compared to IMF_3) has less spurious minima in the neighbourhood of the minima which may represent the excitation instants. The surrounding spurious minima are also of lesser strength. As such, the signal representing the sum of IMFs $3-6$, Figure 5.2(j), may be considered a better candidate than IMF_3 for estimating the impulse locations of $e(n)$. It is important to note at this point that this signal also represents the sum of the IMFs having mean frequencies in the range 50 Hz $\leq F^m \leq 5 \times F_0 = 500$ Hz.

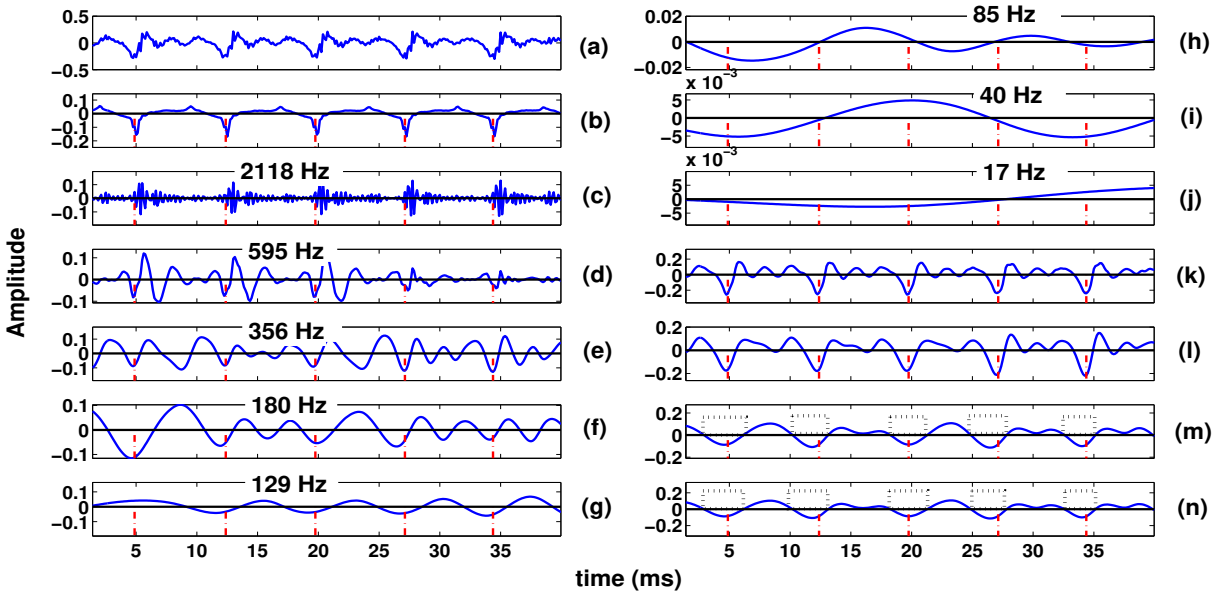


Figure 5.3: **(a) A natural speech signal $s(n)$ having pitch frequency $F_0^{ref} = 116$ Hz ; (b) dEGG corresponding to $s(n)$. The negative peaks indicate the GCIs (c)-(j) are IMFs 1-8 respectively obtained from ICEEMDAN of $s(n)$. IMFs 9 and 10 are not shown ; (k) sum of IMFs 2-6 ; (l) sum of IMFs 3-6 ; (m) sum of IMFs 4-6 ; (n) sum of IMFs 5-6 ; Dashed vertical lines indicate the GCIs. Dotted rectangles in (k) and (l) indicate regions of search for the GCIs.**

Having seen how the IMFs can be used as time-domain multi-resolution signals for estimating the excitation impulses of the synthetic speech signal, we proceed towards investigating the IMFs of the natural speech signal. Figure 5.3 shows the same plots as that of Figure 5.1, but the IMFs here are extracted from a randomly selected speech file/utterance ($F_s = 8$ kHz) from the CMU-Arctic database [158]. Figure 5.3(a) shows the speech signal. Figure 5.3(b) shows the *difference* EGG (dEGG) signal, corresponding to the speech signal. The dEGG signal is the first difference of the EGG signal, recorded simultaneously with the speech signal. A peak detection algorithm is used to detect the large

negative peaks of the dEGG signal, as shown in Figure 5.3(b), which indicate the *instants of significant excitation of the vocal tract* [7]. These are taken as the true or reference GCIs, and indicated by the vertical dashed lines in Figure 5.3(b)-(n). The average distance between a pair of consecutive reference GCIs gives us the *reference time-period*, the inverse of which gives the *reference pitch frequency* (F_0^{ref}). For the given speech signal, $F_0^{ref} = 116$ Hz. Figure 5.3(c)-(j) represents the first eight IMFs of the speech signal. As in the synthetic speech case, the mean frequencies of the IMFs, listed above their plots, decrease monotonically with the IMF order. Correspondingly, the IMFs assume more sinusoidal shapes. It can be seen from Figure 5.3 that the GCIs correspond to different points in the near-sinusoidal IMFs. Amongst them, the signal represented by IMF_3, Figure 5.3(e), manifest the GCIs at some of its minima, with the least ambiguity - it provides the best candidate for estimation of the GCIs, amongst all the IMFs. Figure 5.3(k)-(n) represent the sum of IMFs $2-6$, $3-6$, $4-6$, and $5-6$, respectively. It may be observed that like IMF_3, the sum of IMFs 3-6, represented by Figure 5.3(l), also intersects the GCIs at some of its minima locations. It, however, has less spurious minima, and hence is a better candidate for detecting the GCIs than IMF_3 itself. From a frequency domain perspective, Figure 5.3(l) represents the sum of IMFs having 50 Hz $\leq F^m \leq 5 \times F_0^{ref}$. Similarly, Figure 5.3(m) represents the sum of IMFs having 50 Hz $\leq F^m \leq 2 \times F_0^{ref}$, and Figure 5.3(n) the sum of IMFs having 50 Hz $\leq F^m \lesssim F_0^{ref}$, F_m^5 being only slightly higher than $F_0^{ref} = 116$ Hz. As in the synthetic case, either of the signals represented by Figure 5.3(m),(n) can be used for estimating the *regions of search* for the GCIs. Once such regions are identified, the signal in Figure 5.3(l) can be used to estimate the exact locations of the GCIs.

Figure 5.4 is the equivalent of Figure 5.3, except that the IMFs are obtained from MEMD. Figure 5.4(a),(b) represent the natural speech signal, and its dEGG signal, respectively. Figure 5.4(c)-(j) represent the first eight IMFs. Figure 5.4(k)-(n) represent the sum of IMFs $2-6$, $3-6$, $4-6$, and $5-6$, respectively. The vertical dashed lines in Figure 5.4(b)-(n) indicate the reference GCIs obtained from the dEGG signal. One may observe IMF_6, Figure 5.4(h), having $F_6^m = 123$ Hz $\approx F_0^r = 116$ Hz, which manifests the periodicity of the glottal source. Among the IMFs shown, IMF_3 seems to have its prominent minima in close proximity to the reference GCIs. Again, it may be observed that the sum of IMFs $3-7$, Figure 5.4(l), also has its prominent minima coinciding with the reference GCIs, just as in the case of IMF_3. However, it has lesser spurious minima compared to that of IMF_3, and as such may be considered more suitable for detecting the GCIs. Again, this signal, as in the synthetic case, represents

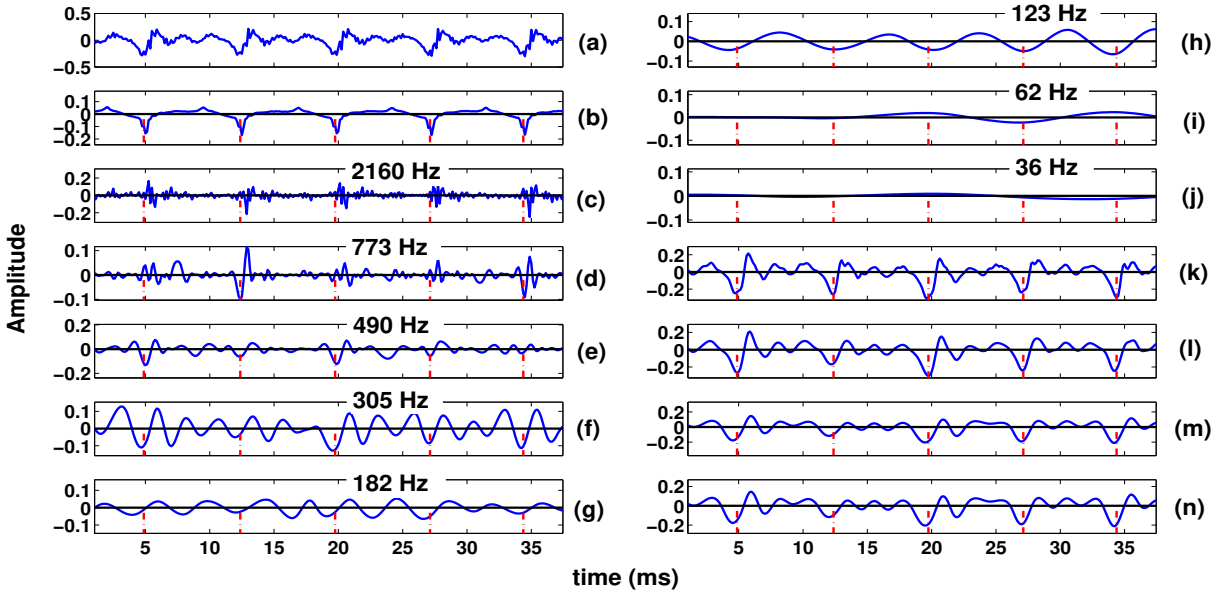


Figure 5.4: **(a) A natural speech signal, $s(n)$, having pitch frequency $F_0^{ref} = 116$ Hz ; (b) dEGG corresponding to $s(n)$. The negative peaks indicate the GCIs (c)-(j) are IMFs 1-8 respectively obtained from MEMD of $s(n)$. IMFs 9 and 10 are not shown ; (k) sum of IMFs 2-7 ; (l) sum of IMFs 3-7 ; (m) sum of IMFs 4-7 ; (n) sum of IMFs 5-7 ; Dashed vertical lines indicate the GCIs.**

the sum of the IMFs having mean frequencies in the range 50 Hz $\leq F^m \leq 5 \times F_0^r$. However, while MEMD provides a significant time-advantage with respect to ICEEMDAN, it is comparatively less effective in reducing *mode-mixing*. As such, the sum of the lower-frequency IMFs is not very effective in providing us the *regions of search* for the GCIs, unlike in the case of ICEEMDAN. Therefore, in the case of MEMD, we would require to devise an alternate methodology for finding the *regions of search* of the GCIs.

Thus, in the case of both ICEEMDAN and MEMD, the partial sum of the IMFs, having their mean frequencies in the range 50 Hz $\leq F^m \leq 5 \times F_0^r$, provides us with a *sinusoid-like* signal, some of whose minima manifest the GCIs. We may call this signal, resulting from the partial sum of the IMFs, as the *Source Enhanced and De-trended Speech* (SEDS) signal. Additionally, it is worth mentioning that the fact that the SEDS signal manifests the GCIs complements the observations made in Chapter 4, where the first few lower-order IMFs were found to predominantly capture system information, leaving the succeeding IMFs to manifest the source characteristics of the speech signal. Moreover, as per the observations made in Chapter 4, the principle for estimating the GCIs, discussed above, should be largely consistent across variations of voiced speech signals. Further, as the proposed principle is dependent on the pitch frequency of the speech signal, it may be considered adaptive to the pitch

induced variations in the decompsoition of the signal (observed in Chapter 4).

5.3 Procedure for ICEEMDAN based GCIs Estimation (IGE)

Based on the principle described in the previous section, the following procedure is employed to detect the GCIs from the speech signal.

- **(i)** Decompose $s(n)$ using ICEEMDAN. Construct the *de-trended speech* signal, $s_d(n)$.

$$s(n) = \sum_{k=1}^{10} h_k(n),\ s_d(n) = \Big\{ \sum_k h_k(n) | F_k^m > 50Hz \Big\} \tag{5.3}$$

- **(ii)** Estimate the average pitch frequency, F_0^{est}, from $s_d(n)$. In this work, the *Robust Alogithm for Pitch Tracking* (RAPT) alogithm [162] is used to estimate the pitch frequencies of voiced frames of the de-trended speech signal. Frames of 20 ms, with a 10 ms frameshift (50 % overlap between frames), are used. F_0^{est} represents the average pitch frequency over all the voiced frames.

- **(iii)** Sum up the components with 50 Hz $\leq F^m \leq 2 \times F_0^{est}$.

$$s_R(n) = \sum_k h_k(n)\ \forall \left\{ k \mid 50 \text{ Hz} \leq F_k^m \leq 2 \times F_0^{est} \right\} \tag{5.4}$$

- **(iv)** Detect the negative-going and positive-going zero-crossings of $s_R(n)$. The region from a given negative zero-crossing to the closest positive zero-crossing to its right, is a *region of search*. Let $\{R_r,\ r = 1, 2, ..., I_{ige}\}$ represent the regions of search, I_{ige} being the total number of regions.

- **(v)** Discard the first IMF, and sum up the rest of the components with 50 Hz $\leq F^m \leq 5 \times F_0^{est}$. This provides us with the SEDS signal.

$$s_e(n) = \sum_{k \geq 2} h_k(n)\ \forall \left\{ k \mid 50 \text{ Hz} \leq F_k^m \leq 5 \times F_0^{est} \right\} \tag{5.5}$$

- **(vi)** In any r^{th} region of search (R_r), find the location of minimum amplitude in $s_e(n)$. These locations provide the initial estimates of the GCIs, $\{l_r^i,\ r = 1, 2, ..., I_{ige}\}$.

Figure 5.5 shows the GCIs estimated using the above process from an arbitrary speech file/utterance. The dashed stem lines indicate the GCIs estimates obtained by the aforementioned process. However, as seen in Figure 5.5(d), along with the actual GCIs, some spurious GCIs are also estimated. The reason for such spurious estimates is the *imperfect sinusoidal nature* of the IMFs.

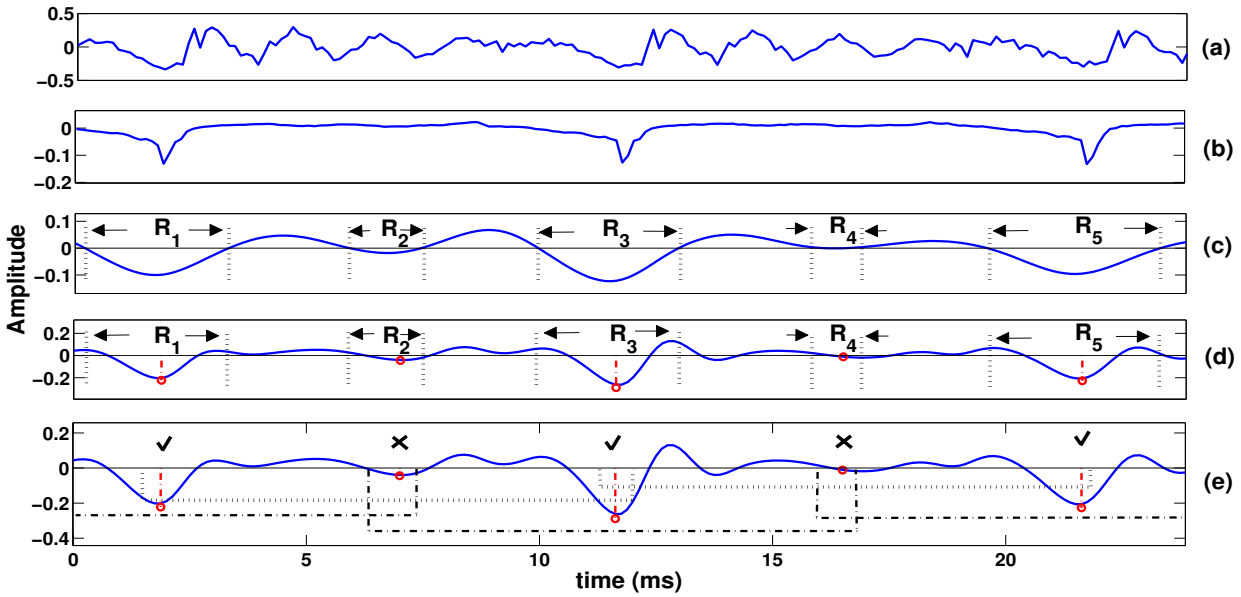


Figure 5.5: **(a) Segment of a speech signal from the CMU-Arctic database ; (b) dEGG signal corresponding to the speech signal ; (c) $s_R(n)$ obtained from the IMFs of ICEEMDAN. The search regions are denoted by $\{R_r, r = 1, ..., 5\}$; (d) $s_e(n)$. The estimated GCIs within $\{R_r, r = 1, ..., 5\}$ are denoted by dashed stem lines ; (e) $s_e(n)$. Spurious and correct GCI estimates are indicated by cross and ticks respectively. Dotted rectangles indicate the L_r regions of estimated spurious GCIs. Dash-dotted rectangles indicate the L_r regions of estimated correct GCIs.**

The existence of false estimates is a familiar problem in many GCI estimation methods, and needs further refinement [5, 6, 9, 39]. To eradicate such false estimates, we apply the following simple methodology.

- **(i)** For the r^{th} estimated GCI, l_r^i, consider a region,

$$l_r^i - W/2 < L_r < l_r^i + W/2,\ W = \nu \frac{F_s}{F_0^{est}},\ \nu \in \mathbb{R}^+ \tag{5.6}$$

  The size of the window, W, is determined by the factor ν, and the effect of this parameter is discussed in the Results section.

- **(ii)** Within any p^{th} region, L_p, find the GCIs estimated within it, i.e, find the set,

$$\{(l_r^i, v_r^i) \mid (l_r^i, v_r^i) \in L_p,\ r = 1, 2, ..., I\},\ v_r^i = s_e(l_r^i) \tag{5.7}$$

  Compare the amplitudes, v_r^i, of this set. If the amplitude v_p^i is the minimum amplitude in this set, then l_p^i is considered as a legitimate GCI, otherwise it is rejected as a spurious one. May the *final estimated GCIs* be denoted by $\{l_r^{est},\ r = 1, 2, ..., F_{ige} \leq I_{ige}\}$.

Figure 5.5(e) shows how the above process works. Let us consider the first (leftmost) GCI estimate. There are two GCIs estimates within the region L_1 specified by the window W. The amplitude, v_1^i, of $s_e(n)$ at the given GCI location (for which the region is defined), l_1^i, is lesser than that of the other GCI estimated within the region. Hence, l_1^i is considered a correct GCI estimate. Now, let us consider the second GCI estimate. Clearly, it is a spurious one, and hence marked X. There are two other GCIs estimates within the region L_2, apart from the spurious estimate. Amongst these three GCIs estimates, since the amplitude, v_2^i, of $s_e(n)$ at the spurious GCI location, l_2^i, is not the minimum, it is rejected. Next, we consider the third GCI estimate, which is a correct estimate. There are two more GCIs estimates within the region L_3, apart from the correct estimate. In this case, the amplitude, v_3^i, of $s_e(n)$ at the given GCI location, l_3^i, is the minimum amongst the amplitudes at all the three GCIs estimates within L_3. Hence, it is accepted as a correct estimate. The last two estimates are verified in the same fashion.

The entire process may be termed as *ICEEMDAN based GCIs Estimation* (IGE).

5.4 Procedure for MEMD based GCIs Estimation (MGE)

While MEMD provides a significant time-advantage with respect to ICEEMDAN, it is comparatively less effective in reducing mode-mixing. As such, the sum of lower-frequency IMFs, like that of Figure 5.1(k),(l) and Figure 5.3(m),(n), is not very effective in providing us the *regions of search* for the GCIs, unlike in the case of IGE. Therefore, we devise an alternate methodology for finding the *regions of search* of the GCIs, based on which the SEDS will provide the exact estimates of the GCIs.

- **(i)** Decompose the speech signal $s(n)$ using MEMD. Construct the *de-trended speech* signal, $s_d(n)$.

$$s(n) = \sum_{k=1}^{10} h_k(n), \;\; s_d(n) = \Big\{ \sum_{k} h_k(n) | F_k^m > 50 \text{ Hz} \Big\} \tag{5.8}$$

- **(ii)** Estimate the average pitch frequency, F_0^{est}, from $s_d(n)$. In this work, RAPT is used to estimate the pitch frequencies of voiced frames of the de-trended speech signal. Frames of 20 ms, with 50 % overlap, are used. F_0^{est} represents the average pitch frequency over all the voiced frames.

- **(iii)** Let $s_h(n) = \Delta\Delta s_d(n)$, where Δ indicates a difference operation. Let n_d denote the minima locations of $s_h(n)$. Construct the envelope, $e_d(n)$, of $s_d(n)$, by using cubic spline interpolation, with the points of interpolation being $\{n_d, s_d(n_d)\}$.

- **(iv)** Detect the negative-going and positive-going zero-crossings of $e_d(n)$. The region from a given negative zero-crossing to the closest positive zero-crossing to its right, provides a prospective *initial region of search*. Let $\{R_r^i,\ r = 1, 2, ..., I_{mge}\}$ represent the initial regions, I_{mge} being the total number of regions.

- **(v)** In every $\{R_r^i,\ r = 1, 2, ..., I_{mge}\}$, find the point of minimum amplitude in $e_d(n)$, (n_r^i, v_r^i), where, $v_r^i = e_d(n_r^i)$.

- **(vi)** For a given $(n_r^i, v_r^i),\ r = 1, 2, ..., I_{mge}$, consider a window, L_r, given by

$$n_r^i - W/2 < L_r < n_r^i + W/2,\ W = \nu \frac{F_s}{F_0^{est}},\ \nu \in \mathbb{R}^+ \tag{5.9}$$

  The size of the window, W, is determined by the factor ν, and the effect of this parameter is discussed in the results section.

- **(vii)** Within any given p^{th} window, L_p, find the set $\{(n_r^i, v_r^i) \mid (n_r^i, v_r^i) \in L_p,\ r = 1, 2, ..., I_{mge}\}$. Compare the amplitudes, v_r^i, of this set. If v_p^i is the minimum amplitude in this set, then the region, R_p^i, is considered as a legitimate region of search, otherwise it is rejected as a spurious one. May the *final set of regions of search* be denoted by $\{R_r^f,\ r = 1, 2, ..., F_{mge} \leq I_{mge}\}$.

Figure 5.6 demonstrates the working of the above process. As may be observed from Figure 5.6(c), the signal, $e_d(n)$, represents a symmetric smooth envelope of the given speech signal, Figure 5.6(a). There are five *initial regions of search* in $e_d(n)$, out of which three regions contain the GCIs (indicated by the vertical dashed lines), and the other two are spurious. The five *minimum value points* corresponding to these five regions are indicated by arrows in Figure 5.6(d). Centered around each *minimum value point*, a symmetric window of width W is considered. For the first (leftmost) point, (n_1^i, v_1^i), the window contains the points $\{(n_1^i, v_1^i),\ (n_2^i, v_2^i)\}$. As $v_1^i = \min\{v_1^i, v_2^i\}$, R_1^i is considered as a legitimate region. For the second point, (n_2^i, v_2^i), the window contains the points $\{(n_1^i, v_1^i),\ (n_2^i, v_2^i),\ (n_3^i, v_3^i)\}$. As $v_2^i \neq \min\{v_1^i, v_2^i, v_3^i\}$, R_2^i is considered as a spurious region. In the same manner, the legitimacy of the remaining regions are evaluated. Finally, three legitimate regions remain, $\{R_1^f, R_2^f, R_3^f\} = \{R_1^i, R_3^i, R_5^i\}$.

Having derived the regions, $\{R_r^f,\ r = 1, 2, ..., F_{mge}\}$, where the GCIs are ought to be contained, the following procedure is applied :

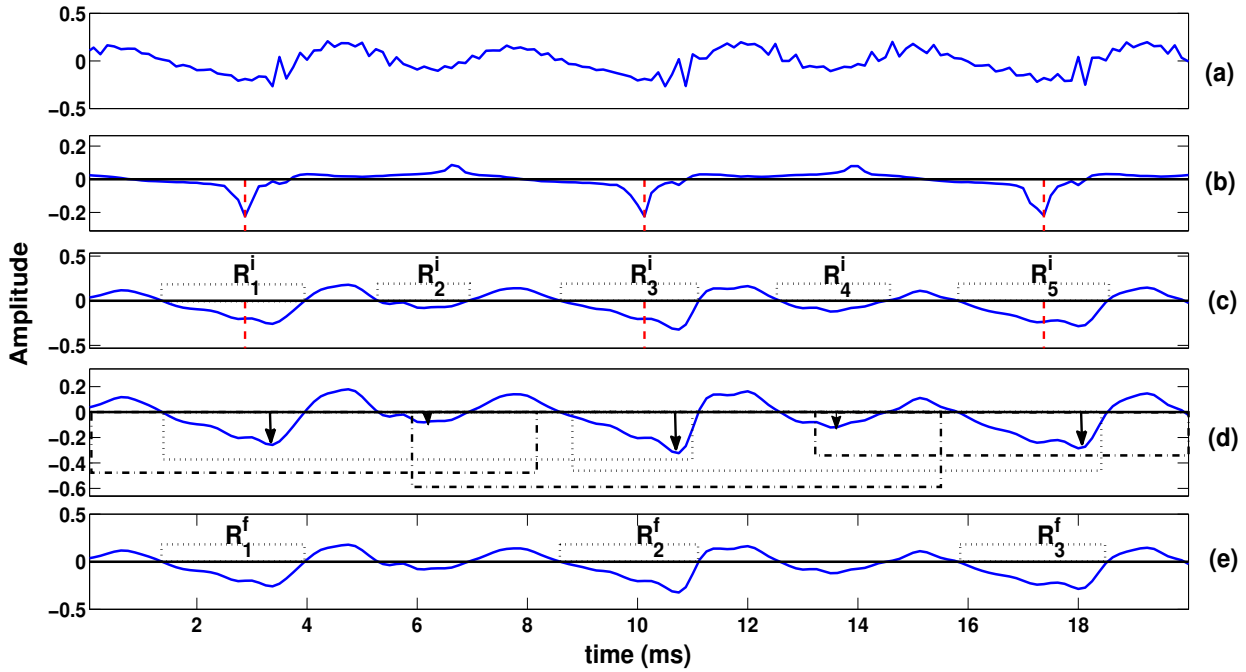


Figure 5.6: **(a) Segment of a speech signal from the CMU-Arctic database ; (b) dEGG signal corresponding to the speech signal. The dashed vertical lines indicate the reference GCIs ; (c) $e_d(n)$. The dashed vertical lines indicate the reference GCIs. The dotted rectangles indicate the *initial regions of search*, $\{R_r^i,\ r = 1, ..., 5\}$; (d) $e_d(n)$. The vertical arrows indicate the *points of minimimum amplitude*, $\{(n_r^i, v_r^i),\ r = 1, ..., 5\}$, corresponding to the initial regions. The dashed rectangles indicate the windows, L_1^i, L_3^i, and L_5^i. The dotted rectangles indicate the windows, L_2^i, and L_4^i ; (e) $e_d(n)$. The dotted rectangles indicate the *final regions of search*, $\{R_r^f,\ r = 1, ..., 3\}$.**

- **(i)** Discard the first IMF, and sum up the rest of the components with 50 Hz $\leq F^m \leq 5 \times F_0^{est}$. This results in the SEDS signal, $s_e(n)$.

$$s_e(n) = \sum_{k \geq 2} h_k(n) \ \forall \left\{ k \mid 50 \leq F_k^m \leq 5 \times F_0^{est} \right\} \tag{5.10}$$

- **(ii)** In the regions of search, $\{R_r^f,\ r = 1, 2, ..., F_{mge}\}$, find the corresponding locations of minimum amplitude in $s_e(n)$. These locations, $\{l_r^{est},\ r = 1, 2, ..., F_{mge}\}$, are the estimates of the GCIs.

Figure 5.7(d) shows the SEDS signal, $s_e(n)$, corresponding to the given speech signal. The *final regions of search*, as estimated from the speech envelope signal, $e_d(n)$, shown in Figure 5.7(c), provide the time intervals where the GCIs are contained. Then, within these intervals (shown by dotted rectangles), the points of minimum value (shown by circles) of the SEDS signal may be associated with the GCIs.

This complete process of estimating the GCIs is termed as *MEMD based GCIs Estimation* (MGE).

5.5 Mimicing IGE/MGE by band-pass filtering

The IGE and MGE algorithms use summations of subsets of the IMFs, which resemble band-pass filtered signals. As such, one may be critical of using IGE/MGE at all if the same or similar performances could be obtained by simply band-pass filtering the speech signal. As discussed in Chapter 2, the bandwidth of the power spectrum an IMF is proportional to its center or dominant frequency. As the order of an IMF increases, its dominant frequency decreases, it becomes more narrowband, and hence looks more sinusoidal [87, 109–111, 163]. Based on this observation, we design two algorithms that mimic the IGE and MGE processes, but which uses simple band-pass filtering of the DFT spectrum of the speech signal, instead of using the IMFs obtained from ICEEMDAN and MEMD. The basic idea is to evaluate whether by band-pass filtering the DFT spectrum at different center frequencies (with associated bandwidths), we may estimate the GCIs. As Fourier analysis is limited to signals generated by linear and stationary processes, the DFT spectrum is evaluated for speech segments/frames of 20 ms, with a frameshift of 10 ms. *OverLap and Add* (OLA) [2,13,14] is then used to obtain the necessary signals. The two algorithms are named BPF-IGE and BPF-MGE, respectively. They are described below :

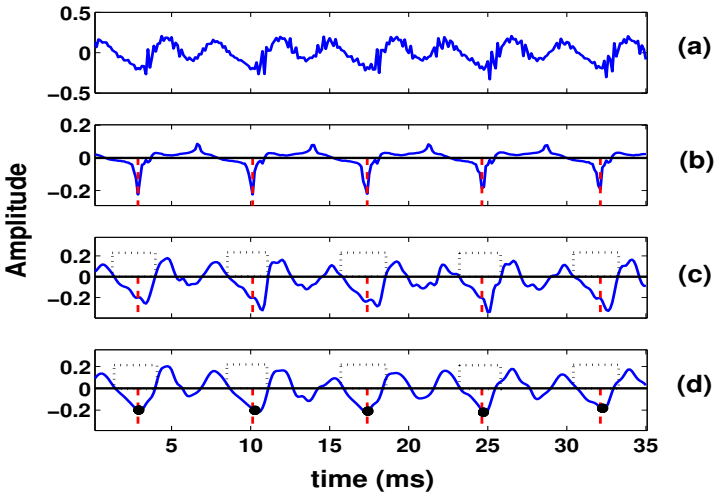


Figure 5.7: **(a) Segment of a speech signal from the CMU-Arctic database ; (b) dEGG signal corresponding to the speech signal. The dashed vertical lines indicate the reference GCIs ; (c)** $e_d(n)$**. The dashed vertical lines indicate the reference GCIs. The dotted rectangles indicate the** *final regions of search* **; (d)** $s_e(n)$**. The dotted rectangles indicate the** *final regions of search*. **The circles indicate the minimum amplitude points, within the** *final regions of search*, **which give the estimates of the GCIs.**

5.5.1 Procedure for BPF-IGE

The steps of the BPF-IGE algorithm may be summarized as :

- **(i)** Let $s^i(n),\ i \in \mathbb{N}$, be a segment/frame of $s(n)$. A framesize of 20 ms, with frameshift of 10 ms, is considered. Obtain the DFT spectrum of $s^i(n)$. Obtain the de-trended speech frame, $s^i_d(n)$,

by high-pass filtering the DFT spectrum.

$$s^i(n) \leftrightarrow S^i(f)\ ,\ s_d^i(n) \leftrightarrow S_d^i(f) = \begin{cases} 0\ ,\ |f| < 50 \text{ Hz} \\ S^i(f)\ ,\ 50 \text{ Hz} \leq |f| \leq \frac{F_s}{2} \end{cases} \tag{5.11}$$

Use OLA to obtain the de-trended speech signal, $s_d(n)$, from $s_d^i(n),\ i \in \mathbb{N}$.

- **(ii)** Obtain the pitch estimate, F_0^{est}, from $s_d(n)$, as in the case of IGE.
- **(iii)** For the i^{th} frame, the signal, $s_R^i(n)$, is obtained by band-pass filtering $s^i(n)$ around its first two pitch harmonics, as,

$$s_R^i(n) \leftrightarrow S_R^i(f) = \begin{cases} S^i(f)\ ,\ |f| \in (1 \pm 10\%) \times kF_0^{est}\ \forall\ k = 1, 2 \\ 0\ ,\ \text{otherwise} \end{cases} \tag{5.12}$$

Use OLA to obtain $s_R(n)$ from $s_R^i(n),\ i \in \mathbb{N}$.

- **(iv)** Obtain the regions of search, $\{R_r,\ r = 1, 2, ..., I_{ige}\}$, from $s_R(n)$ as in the case of IGE.
- **(iv)** For the i^{th} frame, the signal, $s_e^i(n)$, is obtained by band-pass filtering $s^i(n)$ around its first five pitch harmonics, as

$$s_e^i(n) \leftrightarrow S_e^i(f) = \begin{cases} S^i(f)\ ,\ |f| \in (1 \pm 10\%) \times kF_0^{est}\ \forall\ k = 1, 2, 3, 4, 5 \\ 0\ ,\ \text{otherwise} \end{cases} \tag{5.13}$$

Use OLA to obtain the SEDS signal, $s_e(n)$, from $s_e^i(n),\ i \in \mathbb{N}$.

- **(iv)** Implement the rest of the steps of the IGE algorithm to obtain the GCIs estimates, $\{l_r^{est},\ r = 1, 2, ..., F_{ige}\}$.

5.5.2 Procedure for BPF-MGE

The steps of the BPF-MGE algorithm may be summarized as :

- **(i)** Let $s^i(n),\ i \in \mathbb{N}$, be a segment/frame of $s(n)$. A framesize of 20 ms, with a frameshift of 10 ms, is considered. Obtain the DFT spectrum of $s^i(n)$. Obtain the de-trended speech frame, $s_d^i(n)$, by high-pass filtering the DFT spectrum.

$$s^i(n) \leftrightarrow S^i(f)\ ,\ s_d^i(n) \leftrightarrow S_d^i(f) = \begin{cases} 0\ ,\ |f| < 50 \text{ Hz} \\ S^i(f)\ ,\ 50 \text{ Hz} \leq |f| \leq \frac{F_s}{2} \end{cases} \tag{5.14}$$

Use OLA to obtain the de-trended speech signal, $s_d(n)$, from $s_d^i(n),\ i \in \mathbb{N}$.

- **(ii)** Implement the rest of the steps of the MGE algorithm till the final set of legitimate regions, $\{R_r^f,\ r = 1, 2, ..., F_{mge}\}$, is obtained from $s_d(n)$.

- **(iii)** For the i^{th} frame, the signal, $s_e^i(n)$, is obtained by band-pass filtering $s^i(n)$ around its first five harmonics, as

$$s_e^i(n) \leftrightarrow S_e^i(f) = \begin{cases} S^i(f)\ ,\ |f| \in (1 \pm 10\%) \times kF_0^{est}\ \forall\ k = 1, 2, 3, 4, 5 \\ 0\ ,\ \text{otherwise} \end{cases} \tag{5.15}$$

Use OLA to obtain the SEDS signal, $s_e(n)$, from $s_e^i(n),\ i \in \mathbb{N}$.

- **(iv)** Estimate the GCIs, $\{l_r^{est},\ r = 1, 2, ..., F_{mge}\}$, from $s_e(n)$, as in the case of MGE algorithm.

It may be noted that, in the BPF-IGE and BPF-MGE algorithms, f represents the analog frequencies of the corresponding digital frequencies of the DFT spectrum of the signal under consideration.

Having described the procedures of the proposed algorithms, we now proceed towards evaluating their performances over standard speech databases, and under different scenarios.

5.6 Results and Discussion

For the experiments performed using ICEEMDAN (described in Section 2.2.1), $\epsilon_0 = 0.2$ is considered. The number of iterations in a complete sifting process is determined by the *local-global stopping criterion*, described in Section 2.1.4. The maximum number of iterations per sifting process is not allowed to exceed 15, i.e., $N \leq 15$ [105, 116, 130, 131]. In this work, $M = 9$, and only $L = 20$ White noise realizations are used. For the experiments performed using MEMD, $N = 10$, and $M = 9$ are considered.

Table 5.1: Description of the databases used. The CMU-Arctic database consists of five speakers, and the APLAWDW database consists of ten speakers.

Database	CMU-Arctic (5 speakers)					APLAWDW (10 speakers)
Speakers	BDL (male)	JMK (male)	SLT (female)	EDX (male)	KDT (male)	All (5 male , 5 female)
Number of Utterances	∼1100	∼1100	∼1100	∼1900	∼500	∼50 per speaker
Average Duration	∼3 s	∼3 s	∼3 s	∼1 s	∼2 s	∼3 s

Databases : Two databases, the CMU-Arctic database [158], and the APLAWDW database [164], are considered for this work. The CMU-Arctic database consists of five speakers, and the APLAWDW

database consists of ten speakers. The databases are described in Table 5.1. All utterances (from both the databases) considered in this work are of $F_s = 8$ kHz.

In order to evaluate the performances of the proposed IGE and MGE algorithms five objective metrics have been used [5–11]. They are described below :

- **Identification Rate (IR)** : The percentage of glottal or larynx cycles in which only one GCI is estimated.
- **Miss Rate (MR)** : The percentage of glottal cycles in which only no GCI is estimated.
- **False Alarm Rate (FAR)** : The percentage of glottal cycles in which more than one GCI is estimated.
- **Identification Accuracy (IA)** : The standard deviation of the timing or identification error, ζ, from the actual reference GCI.
- **Accuracy to $\pm$ 0.25 ms (IA$'$)** : The percentage of GCIs which are within a timing error bound of $\pm$ 0.25 ms.

Figure 5.8 aids in understanding the metrics used for evaluation. Out of the five metrics, IR, MR, and FAR reflect the reliability of the estimation technique, whereas IA and IA$'$ represent the accuracy of the technique. Higher values of IR and IA$'$, and lower values of MR, FAR, and IA are desirable. Using these five metrics, the proposed algorithms are compared with five competitive state-of-the-art algorithms - DYPSA, YAGA, SEDREAMS, ZFR, and ILPR-PI.

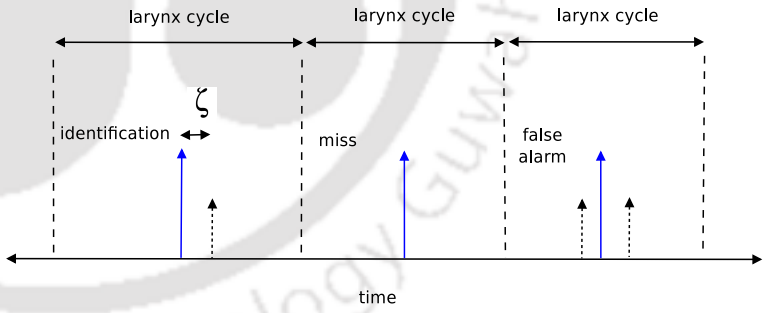


Figure 5.8: Characterization of the estimates of the GCIs showing three larynx or glottal cycles, with examples of each possible outcome from the estimates [5–11]. The solid arrows indicate the reference GCIs, obtained from the dEGG signal. The dotted arrows indicate the GCIs estimated. ζ is the *identification error*.

The performance of the various algorithms are evaluated not only for clean speech, but speech degraded by White, Babble, and HFchannel noise [159], with the SNR being varied from 0 to 20 dB, in steps of 5 dB. Apart from this, the performance for speech affected by telephone channel conditions is also evaluated. Two types of telephone quality speech are considered in this work :

- **T-1** : The T-1 type of telephone quality speech has been derived by band-pass filtering the speech signal between 300 - 3400 Hz, using the VOICEBOX toolbox [165]. The magnitude response of the filter is given by a raised cosine function in the range of 0-300 Hz and 3400-4000 Hz, and unity between 300 - 3400 Hz [11, 165].

- **T-2** : The T-2 type of telephone quality speech has been derived by band-pass filtering the speech signal, followed by encoding, in accordance with ITU standards [160].

5.6.1 Effect of the parameter ν on the performance of estimating GCIs

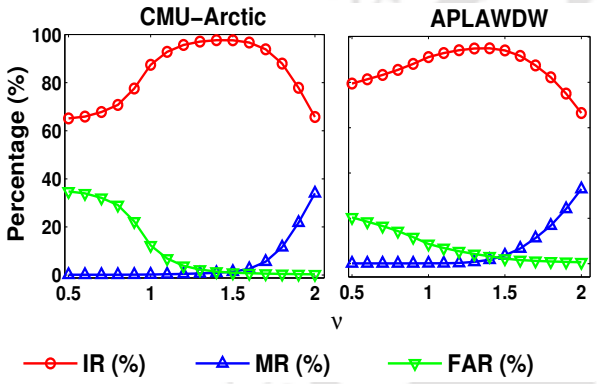


Figure 5.9: Effect of the window size parameter, ν, in estimating the GCIs using the IGE algorithm. Clean speech signals from the CMU-Arctic and APLAWDW databases are used.

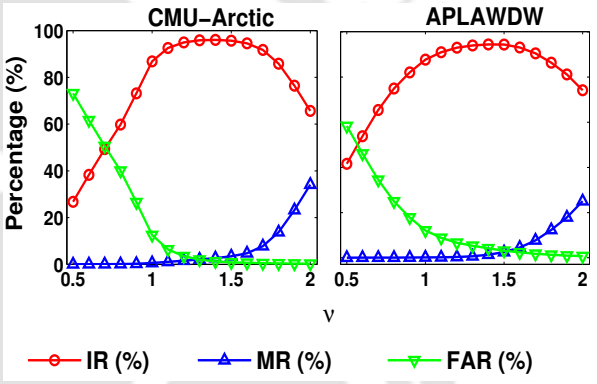


Figure 5.10: Effect of the window size parameter, ν, in estimating the GCIs using the MGE algorithm. Clean speech signals from the CMU-Arctic and APLAWDW databases are used.

As discussed in the previous section, the existence of false alarms or false estimates of the GCIs is almost inevitable in most of the GCIs estimation algorithms. In the case of IGE and MGE too, such estimates (regions in the case of MGE) are obtained initially along with correct estimates (regions in the case of MGE). For eliminating such estimates (regions in the case of MGE), we employ a shifting window of size W, which is dependent on an adjustable parameter ν, and the F_0 of the signal. Hence, the effect of the parameter, ν, on the overall performance of the IGE and MGE algorithms, needs to be studied, and its optimum range needs to be found out. It also needs to be investigated if the optimum range of ν varies for the two databases. For this purpose, the IR, MR, and FAR for the two algorithms are obtained for different values of $\nu \in [0.5, 2]$, for the two databases separately. Only clean speech signals are used for this study.

Figure 5.9 shows the plots of the three metrics for IGE, and Figure 5.10 shows the same for MGE. Separate plots are shown for the CMU-Arctic and APLAWDW databases. As can be observed from

Figure 5.9, the IRs for both the databases obtain saturation somewhere between $\nu = 1.1$ to $\nu = 1.5$. The same is observed for MGE, as shown in Figure 5.10. In this range, for both the algorithms, and both the databases, the FAR drops sharply, with little increase in MR. This means that (assuming F_0^{est} is a reasonably accurate estimate of the actual F_0) the window size should be slightly larger than the *pitch period* (the inverse of F_0), to eliminate the false estimates/regions without adversely affecting the overall performance of the algorithm. Larger values of ν results in a large window, which not only eliminates false estimates/regions, but also the correct ones. Smaller values of ν, on the contrary, are not very efficient in eliminating the false estimates. It is also important to notice that the optimal range of ν is nearly the same for both the databases. Hence, the window size may be considered independent both in terms of the methodology employed to estimate the GCIs, and the change in the dataset.

As such, using $\nu \in [1.1, 1.5]$ should give us results which are close to the best possible performances of the algorithm. Henceforth, $\nu = 1.4$ is used for estimating the GCIs by both IGE and MGE in this work, under all conditions.

5.6.2 Performances under clean and telephone channel conditions

The performances of IGE, MGE, and the five state-of-the-art algorithms are first evaluated for clean speech signals of the CMU-Arctic and APLAWDW databases. Table 5.2 shows the five performance metrics for the algorithms. The five speakers of the CMU-Arctic database are shown individually for better analysis. The best performances are highlighted in bold-font. As can be observed from the table, among the state-of-the art algorithms YAGA, SEDREAMS, and ILPR-PI provide consistent performance across the different speakers of the CMU-Arctic database. DYPSA and ZFR, on the other hand, show some fluctuations in performance with respect to the change in speakers. The performance of DYPSA drops starkly for EDX, and that of ZFR drops for KDT. The performances of IGE are relatively invariant to speaker change, similar to that of YAGA, SEDREAMS, and ILPR-PI. The performances of MGE are worse for the KDT and EDX speakers, compared to that of the other speakers, even though the variation in performance is not as alarming as that of the DYPSA and the ZFR. The main differences between IGE and MGE, for these two speakers, are in IR and MR. Otherwise, IGE and MGE are neck to neck in the other metrics. The low FAR values prove how well the method for eliminating spurious GCIs work, for both IGE and MGE. Overall, the performances of IGE and MGE are competitive with the state-of-the-art algorithms for every speaker

Table 5.2: Robustness of the GCIs estimation methods, according to the five performance metrics, for clean speech signals.

Database	Metric	IGE	MGE	DYPSA	YAGA	SEDREAMS	ZFR	ILPR-PI
BDL (CMU)	IR(%)	97.59	97.63	94.96	97.82	97.51	97.12	**98.52**
	MR(%)	0.68	1.49	2.7	0.71	1.12	1.41	**0.62**
	FAR(%)	1.73	0.89	2.34	1.47	1.37	1.47	**0.86**
	IA(ms)	0.33	0.44	0.52	0.41	0.4	0.41	**0.31**
	IA'(%)	71.34	63.34	78.6	84.5	83.8	82.62	**87.3**
JMK (CMU)	IR(%)	**98.93**	98.9	97.5	98.76	98.56	95.83	98.7
	MR(%)	**0.35**	0.4	1.4	0.55	0.48	3.81	0.59
	FAR(%)	0.72	0.71	1.1	0.69	0.96	**0.36**	0.71
	IA(ms)	0.53	0.53	0.57	0.51	0.54	0.72	**0.35**
	IA'(%)	58.67	57.61	73.56	74.52	73.6	36.76	**83.7**
SLT (CMU)	IR(%)	**99.39**	99.28	97.15	98.15	98.45	98.96	98.83
	MR(%)	**0.1**	0.23	1.75	0.47	0.31	0.49	0.56
	FAR(%)	0.52	**0.49**	1.1	1.38	1.24	0.55	0.61
	IA(ms)	**0.23**	0.24	0.56	0.39	0.41	0.32	0.34
	IA'(%)	77.63	75.79	67.16	**81.23**	74.9	78.8	80.78
EDX (CMU)	IR(%)	96.74	94.19	80.03	95	98.16	91.5	**97.85**
	MR(%)	0.93	3.63	2.5	**0.9**	0.8	7.1	0.9
	FAR(%)	2.33	2.18	17.47	4.1	1.04	1.4	**1.25**
	IA(ms)	0.79	0.97	0.66	0.7	**0.53**	0.8	0.53
	IA'(%)	36.87	34.07	82.1	83.32	85.2	50.58	**85.6**
KDT (CMU)	IR(%)	94.49	91.7	95.1	96.51	96.9	84.67	**97.12**
	MR(%)	4.65	7.73	2.1	0.91	1.12	8.86	**0.21**
	FAR(%)	0.86	**0.56**	2.8	2.58	1.98	6.47	2.67
	IA(ms)	0.93	1.05	0.56	0.58	0.53	0.83	**0.37**
	IA'(%)	50.92	33.23	82.66	88.15	84.12	38.12	**90.15**
APLAWDW	IR(%)	95.05	94.46	94.5	**96.9**	96.8	96.75	95.5
	MR(%)	1.72	1.35	2.5	**1.1**	1.8	1.7	1.9
	FAR(%)	3.23	4.18	3	2	**1.4**	1.55	2.6
	IA(ms)	**0.39**	0.42	0.8	0.69	0.62	0.75	0.6
	IA'(%)	64.86	59.28	68.15	77.5	78.12	49.21	**80.15**

case. The performances of IGE and MGE for the APLAWDW database are similar to the state-of-the art algorithms. The only parameter, one may argue, in which the IGE and MGE lag behind with respect to the other algorithms is IA′. This is a direct result of the effect of mode-mixing, as it cannot be completely eliminated. Thus, while some GCIs estimates are extremely accurate, certain others are a little far from the reference GCIs. However, it must be noted that the IA′ values of IGE and MGE are still competitive with respect to the state-of-the-art algorithms.

Having observed the performances of IGE and MGE over clean speech signals, we now investigate whether they could perform credibly when the speech signals are affected by telephone channel conditions. Under telephone channel conditions, the low-frequency spectrum of the speech

Table 5.3: Robustness of the GCIs estimation methods, according to the five performance metrics, for two types of telephone quality speech (T-1 and T-2)

Type	Database	Metric	IGE	MGE	DYPSA	YAGA	SEDREAMS	ZFR	ILPR-PI
T-1	CMU-Arctic	IR(%)	**91.1**	89.77	88.29	45.81	72.05	54.36	79.21
		MR(%)	5.39	7.14	0.8	0.9	0.12	**0.01**	1.05
		FAR(%)	3.5	**3.09**	10.91	53.29	27.83	45.63	19.74
		IA(ms)	1.03	1.1	0.4	1.28	0.34	**0.25**	1.07
		IA'(%)	33.76	32.98	81.35	27.26	**80.98**	75.68	42.62
	APLAWDW	IR(%)	**86.04**	84.59	91.35	75.33	70.67	72.26	85.52
		MR(%)	2.32	5.27	0.77	0.96	1.42	**0.13**	0.96
		FAR(%)	11.64	10.14	**7.88**	23.71	27.91	27.61	13.51
		IA(ms)	0.9	1.12	0.44	1.46	0.92	**0.76**	0.76
		IA'(%)	29.69	25.97	**79.79**	10.99	44.41	45	59.49
T-2	CMU-Arctic	IR(%)	85.36	**86.74**	80.62	36.3	18.23	17.65	65.82
		MR(%)	9.83	9.98	1.76	1.17	0.49	**0.08**	1.45
		FAR(%)	4.81	**3.28**	17.62	62.53	81.28	82.27	32.73
		IA(ms)	1.49	1.27	0.69	1.05	1.08	**0.58**	0.96
		IA'(%)	20.28	24.88	49.97	34.54	33.05	**55.46**	53.03
	APLAWDW	IR(%)	**85.22**	84.85	86.18	67.18	27.14	35.86	83.6
		MR(%)	6.68	8.55	2.22	**0.78**	2.43	0.62	1.89
		FAR(%)	8.1	**6.6**	11.6	32.04	70.43	63.51	14.51
		IA(ms)	1.13	1.08	0.62	0.74	1.57	1.69	**0.56**
		IA'(%)	17.64	19.2	48.34	53.3	15.55	16.83	**71.36**

signal, which contains the glottal source information, is suppressed. This allows us to evaluate how well the algorithms perform when the information source (the glottal characteristics) from which the algorithms are trying to extract information (the GCIs), is subdued or obscured. Table 5.3 shows the five performance metrics for the two types of telephone quality speech, T-1 and T-2, for the CMU-Arctic and APLAWDW databases, for the seven algorithms. As the table shows, the performances of all the algorithms (particularly in terms of IR) degrade under telephone channel conditions. YAGA, SEDREAMS, and ZFR, all of which performed credibly under clean speech conditions, show significant degradation for both types of telephone quality speech. For T-1 type of telephone quality speech, only IGE, MGE, DYPSA and ILPR-PI provide credible performance (IR $\gtrsim$ 80%). When the IR is itself low, the metrics related to accuracy (IA and IA$'$) become less relevant. In the case of T-2 type of telephone quality speech, the performances of all the algorithms are worse than that of T-1 type. This is expected, as ITU telephone channel codecs suppress the lower-frequency spectrum of speech much more than simple band-pass filtering as in the case of T-1 type of telephone quality speech. For T-2 type, only IGE, MGE, DYPSA, and ILPR-PI provide credible performance, compared to the other algorithms. The performance of ILPR-PI degrades

starkly for the CMU-Arctic database for T-2 type, compared to that of T-1 type.

Thus, out of the five state-of-the-art algorithms, only the DYPSA, which did not perform as well as the other algorithms for clean speech, works credibly in the case of telephone quality speech. Noticeably, both IGE and MGE provide credible performance for clean speech, as well as telephone quality speech.

5.6.3 Performance under noisy conditions

Having evaluated the performances of the algorithms on telephone quality speech, we now proceed towards investigating their performances on speech signals corrupted by noise. The objective here is to investigate how the algorithms perform when external sources (any other signal source not corresponding to the speaker) try to mask or subdue the spectrum of the speech signal.

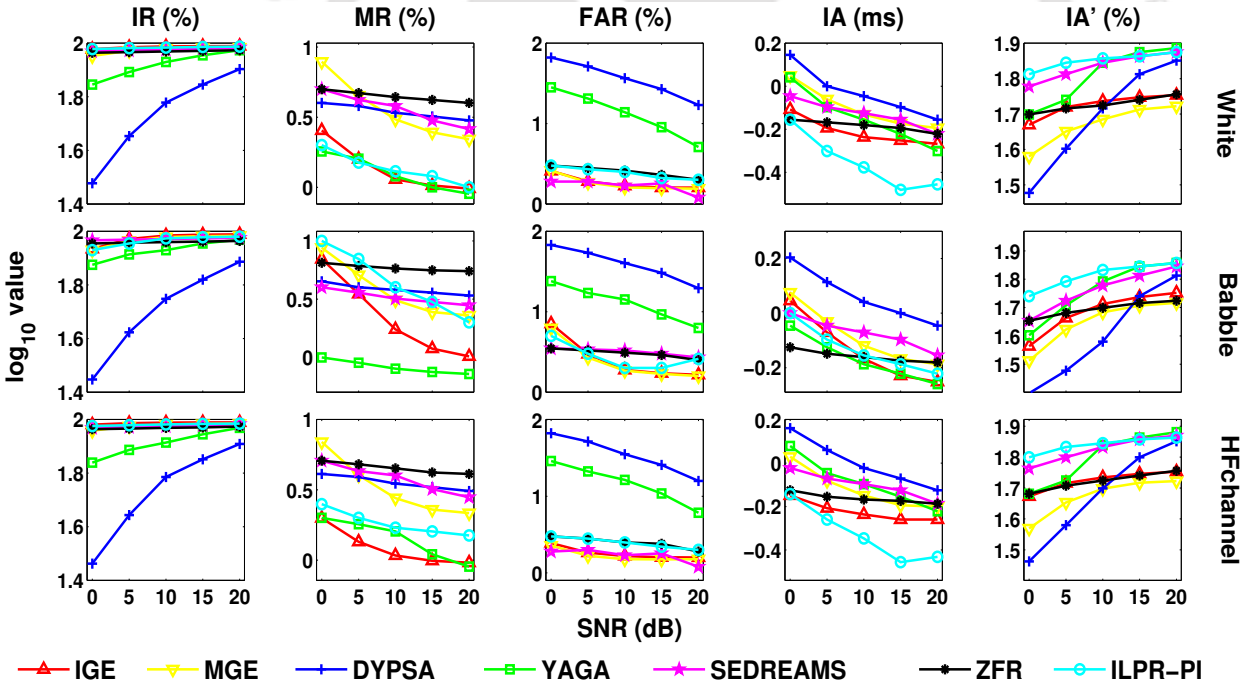


Figure 5.11: **Robustness of the GCIs estimation methods, according to the five performance metrics, for speech signals corrupted by White, Babble, and HFchannel noise. The performance metrics are averaged over all the speech signals of the CMU-Arctic and APLAWDW databases combined. The SNR is varied from 0 dB to 20 dB for each type of noise.**

Figure 5.11 shows the performances of the seven algorithms for White, Babble, and HFchannel noise, averaged over the two databases. For visual clarity, the plots are shown in logarithmic scale. First, let us consider the case of speech signals corrupted by White noise. As can be observed from the plots, all the algorithms are neck to neck for all the metrics, particularly when the level of noise is low

(SNR $\geq$ 15 dB). As the level of noise increases, the IRs of DYPSA and YAGA worsen substantially with respect to the other algorithms. ZFR, which performed poorly under telephone channel conditions, remarkably, shows very little fluctuation with respect to the level of noise, for all the five metrics. For IGE, one may notice how low the MR and FAR are even at high noise levels, resulting in high IR. For MGE, the FAR remains low even at high noise levels, but the MR increases with respect to that of IGE. Again, as in the case of clean speech, the only metric that IGE and MGE might be considered lagging is IA$'$. As mentioned earlier, this is a result of mode-mixing. However, the values of IA$'$ for the two algorithms are close to that of ZFR, and better than that of the DYPSA (for SNR $\leq$ 10 dB).

Next, let us consider the case of HFchannel noise. As the plots in Figure 5.11 show, the performances for HFchannel noise are similar to that of White noise for all the seven algorithms, and the five metrics. As such, the same observations made for White noise are applicable for HFchannel noise. In both these types of noise, IGE and MGE show limited fluctuations in their performances as the level of noise increases. This may be because the SEDS signal is constructed leaving out the first IMF (and other high-frequency IMFs). The first IMF, thus, captures most of the White/HFchannel noise, making IGE and MGE immune to its influence.

Lastly, let us consider the case of Babble noise. As can be observed in Figure 5.11, the performances of the algorithms are worse for Babble noise, than that for White and HFchannel noise. DYPSA, which performed remarkably for telephone quality speech, is also not immune to external low-frequency noise. Apart from DYPSA and YAGA, the rest of the algorithms almost converge to the same performance levels, and there is hardly any difference amongst their performances. The IRs of IGE and MGE are competitive with the best performing algorithms for different levels of noise. The MRs, FARs, and IAs are similarly competitive. The IA$'$ values are now even closer to the best cases, and almost identical to that of ZFR.

Table 5.4: Computational complexity of the seven different methods for detecting the GCIs. Time (in seconds) taken by the seven algorithms for detecting the GCIs from a speech signal, of ~4 s duration. The algorithms are run in MATLAB, in desktop GUI mode, in a machine with 8 GB RAM, using Intel quad-core i7 processor, of 2.9 GHz clock frequency.

Algorithm	IGE	MGE	DYPSA	YAGA	SEDREAMS	ZFR	ILPR-PI
Time (s)	63.17	1.62	0.76	1.19	0.74	2.28	0.54

5.6.4 Computational complexity

Thus, as discussed in the preceding two subsections, considering all conditions - clean, telephone channel, and noise - both IGE and MGE perform competitively with the standard algorithms.

However, the time-costs of the algorithms need to be considered for practical applicability. Table 5.4 lists the time-costs of the seven algorithms for a clean speech signal from the CMU-Arctic database. The algorithms are run in MATLAB, in desktop GUI mode, in a machine with 8 GB RAM, using Intel quad-core i7 processor, of 2.9 GHz clock frequency. The IGE is not time-efficient like the other algorithms. The efficiency of MGE is much better, and comparable to the state-of-the-art methods. The fact that the performance of MGE is not vastly different from that of IGE, while costing only a fraction of the time-cost of the IGE, makes it more suitable for practical applications. However, it is to be noted that the two proposed algorithms for estimating the GCIs from the IMFs do not account principally for their overall time-costs. It is the time taken to extract the IMFs that is the main reason, particularly for ICEEMDAN. It is to be expected that with efficient programming, using parallel processing, the time cost of extracting the IMFs using ICEEMDAN could be significantly reduced. Further, as better and efficient methods of curbing mode-mixing are developed, the IMFs could provide even more accurate estimates of the GCIs, with the same proposed methodologies. Thus, even though currently the large time-cost seems a bottleneck for practical applications, that may not be so in the near future, or even now with efficient programming.

5.6.5 Advantage of IGE/MGE over simple band-pass filtering

Table 5.5: Robustness of IGE, MGE, BPF-IGE, and BPF-MGE in GCIs estimation, according to the five performance metrics, for speech signals under clean and telephone channel conditions. The performance metrics are averaged over all the speech signals of the CMU-Arctic and APLAWDW databases combined.

Condition →	Clean					T-1					T-2				
Metric →	IR (%)	MR (%)	FAR (%)	IA (ms)	IA' (%)	IR (%)	MR (%)	FAR (%)	IA (ms)	IA' (%)	IR (%)	MR (%)	FAR (%)	IA (ms)	IA' (%)
IGE	97.47	0.96	1.57	0.54	57.57	90.73	5.17	4.11	1.02	33.46	85.35	9.58	5.06	1.46	20.08
MGE	96.41	2.17	1.42	0.63	53.03	89.51	6.89	3.6	1.09	32.37	86.63	9.84	3.53	1.25	24.34
BPF-IGE	95.98	3.07	0.95	0.87	51.55	90.53	6.63	2.84	0.97	39.75	87.24	8.41	4.35	1.26	26.35
BPF-MGE	96.56	2.07	1.38	0.68	50.8	91.21	6.21	2.58	1.01	39.09	87.93	8.86	3.21	1.3	25.59

Finally, we proceed towards clarifying whether simply band-pass filtering the speech signal, instead of using subsets of the IMFs, could be just as effective in estimating the GCIs. In other words, we need to compare the performances of IGE and MGE with that of BPF-IGE and BPF-MGE, under different scenarios.

Table 5.5 shows the performances metrics, averaged over the CMU-Arctic and APLAWDW databases, for IGE, MGE, BPF-IGE, and BPF-MGE. As can be observed in the table, for clean

speech, BPF-IGE and BPF-MGE provide almost identical performances as that of IGE and MGE, respectively. While this questions the utility of ICEEMDAN and MEMD, on the other hand this further validates our approach for estimating the GCIs. The accuracy of the estimates (IA), however, is poorer than that of IGE and MGE. Interestingly, even for telephone quality speech (both T-1 and T-2), BPF-IGE and BPF-MGE perform remarkably well. In fact, BPF-IGE and BPF-MGE perform marginally better than IGE and MGE (IR is almost the same, whereas IA and IA' are better). These observations suggest that when the speech signal is relatively clean or uninfluenced, the band-pass filtering of the speech signal performs in an identical fashion as that of using a subset of its IMFs. However, when the speech signal is corrupted by noise, the same may not happen. As such, we evaluate the performances of BPF-IGE and BPF-MGE under White, Babble and HFchannel noise.

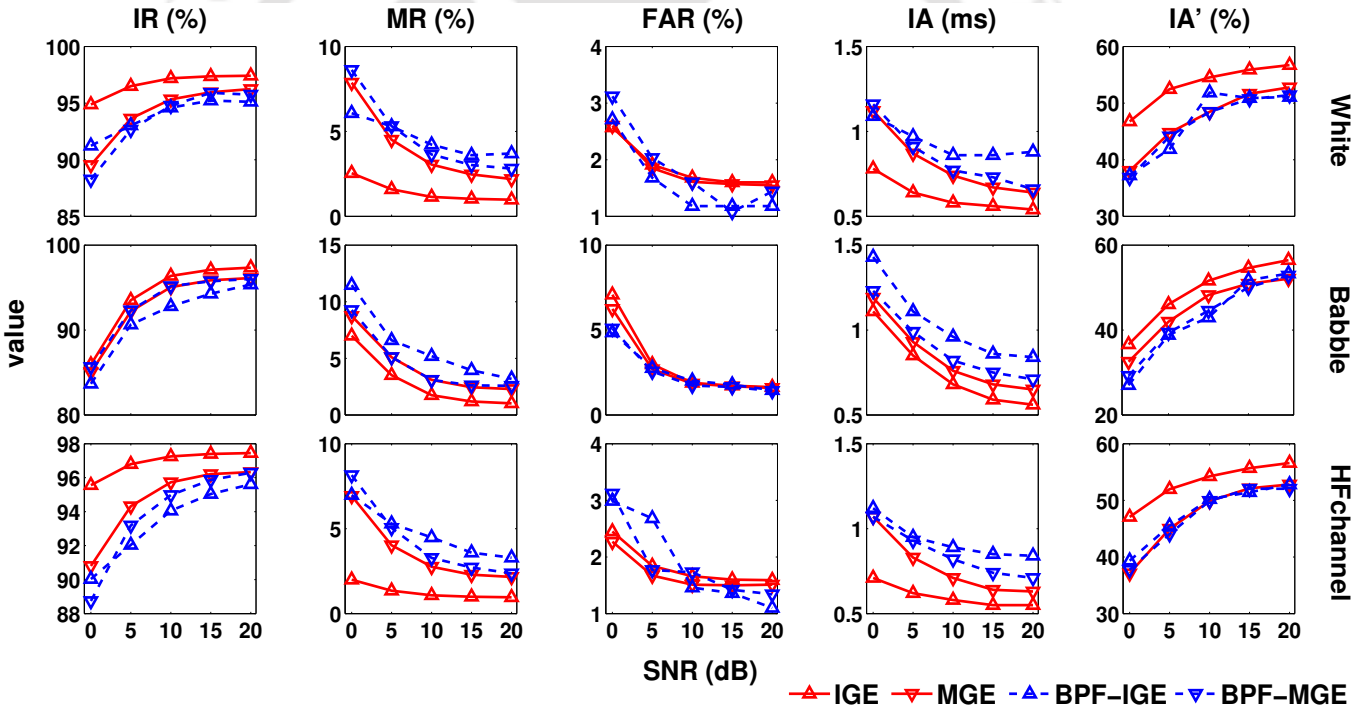


Figure 5.12: **Robustness of IGE, MGE, BPF-IGE, and BPF-MGE, according to the five performance metrics, for speech signals corrupted by White, Babble, and HFchannel noise. The performance metrics are averaged over all the speech signals of the CMU-Arctic and APLAWDW databases combined. The SNR is varied from 0 dB to 20 dB for each type of noise.**

Figure 5.12 presents the performance metrics for IGE, MGE, BPF-IGE, and BPF-MGE, when the speech signals are corrupted by different types (and levels) of noise. In the case of White noise, particularly for SNR $\leq$ 15 dB, IGE stands out with respect to the band-pass filtered algorithms, for all the metrics. The IRs of BPF-IGE and BPF-MGE are, however, similar to that of MGE. However, the

IAs of MGE are marginally better than that of the band-pass filtered algorithms. The performances of the four algorithms under HFchannel noise are similar to that under White noise. IGE clearly stands out. Again, the IRs of the other three algorithms are similar. However, the IAs of MGE are discernibly better than that of the band-pass filtered algorithms. In the case of Babble noise, the IRs of all the four algorithms are similar. However, the IA and IA$'$ values of IGE and MGE are distinctly superior.

To illustrate the differences between the algorithms under noisy conditions, we consider the case of a segment of a speech signal (taken from the APLAWDW database), corrupted by Babble noise at SNR = 0 dB. Figure 5.13(a) shows the speech signal. Figure 5.13(b) shows the SEDS signal, $s_e(n)$, obtained using BPF-IGE. Figure 5.13(c) represents the $s_R(n)$ signal (obtained using BPF-IGE), used for finding the regions of search. The vertical dashed lines indicate the reference GCIs, obtained from the dEGG signal (not shown in the figure). As can be observed in Figure 5.13(b), the rightmost GCI was missed by the BPF-IGE algorithm. This is due to two factors :

(i) The $s_R(n)$ signal, Figure 5.13(c), contains a number of spurious regions, i.e., a number of positive-going and negative-going zero-crossings.

(ii) The $s_e(n)$ signal, Figure 5.13(b), contains a stronger peak to the left of the peak corresponding to the rightmost GCI. Hence, instead of the actual GCI, the peak to its left was considered by the algorithm.

In contrast to the $s_e(n)$ and $s_R(n)$ signals of the BPF-IGE algorithm, we may look at the same signals for the IGE algorithm. They are presented in Figure 5.13(d),(e) respectively. The number of spurious regions in the $s_R(n)$ signal are lesser in this case, compared to that of BPF-IGE. Again, the peaks in $s_e(n)$ corresponding to the legitimate GCIs are stronger than those in its vicinity. As such, both the legitimate GCIs were picked up by

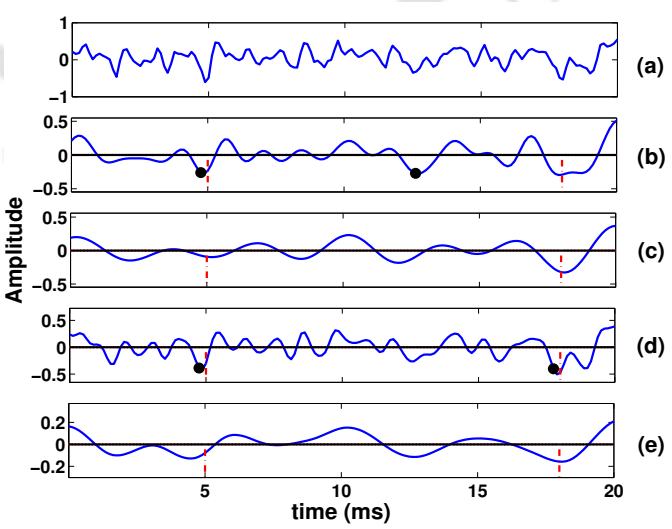


Figure 5.13: **(a) A speech signal, $s(n)$, corrupted by Babble noise at SNR = 0 dB ; (b) $s_e(n)$ obtained using BPF-IGE ; (c) $s_R(n)$ obtained using BPF-IGE ; (d) $s_e(n)$ obtained using IGE ; (e) $s_e(n)$ obtained using IGE. The dashed vertical lines indiacte the reference GCIs. The circles indicate the estimated GCIs.**

IGE. Of course, the accuracy is not as

precise as would have been in the case of a clean speech signal. Similar observations can be made in case of the BPF-MGE and MGE algorithms, and hence not pictorially represented here. Thus, while in hindsight the process of band-pass filtering works for speech uninfluenced by external factors, when the speech signal is corrupted, the true ability of EMD (hence MEMD and ICEEMDAN) are useful.

5.7 Conclusion

This work focussed on detecting the GCIs of the speech signal using non-linear and non-stationary signal analysis, as an alternative to the source-filter theory or LP-analysis based methodologies of doing the same. The motivation behind using non-linear and non-stationary analysis was that state-of-the-art techniques, mainly based on short-time LP analysis, provide inconsistent performances when the speech signal is subjected to external influences. With this motivation, this work investigated the capability of two non-linear and non-stationary signal analysis methods - ICEEMDAN, and MEMD - for reliably extracting the GCIs under varied conditions - clean, noisy and under telephone channel effects. It was observed that both ICEEMDAN and MEMD could extract simple sinusoid-like components (IMFs) from the speech signal, which could be utilized to detect the GCIs. Hence, two simple and uncomplicated processes were developed, called IGE and MGE, for detecting the GCIs from the IMFs derived from ICEEMDAN and MEMD respectively. IGE and MGE are observed to provide comparable performances, under varied conditions, with the state-of-the-art algorithms. The principal drawback of IGE is that it is time-costly. MGE, of course, is a much faster algorithm, but its performance is marginally inferior to that of the IGE. For clean speech, and telephone quality speech, the difference between the two algorithms is marginal - IR of MGE is within 1% of that of the IGE, whereas IA differs by < 0.2 ms between the two algorithms. Again, one may argue that the time-cost of IGE could be drastically reduced by parallel programming and efficient coding. Nevertheless, as EMD improves as a non-linear and non-stationary data analysis method, the proposed principle and methodologies for detecting the GCIs could be expected to provide even more accurate estimates of the GCIs.

6

Analysis of the IMFs for Speaker information

Contents

Outline

The objective of this work is to investigate the utility of the IMFs of the speech signal, generated by the *data-adaptive filterbank* nature of EMD, and its variant, MEMD, in capturing the identity of the speaker. As such, in this work, the IMFs obtained from EMD and MEMD are utilized in the task of text-independent *Speaker Verification* (SV). Three different features are extracted over 20 ms frames, from the IMFs. They are then tested individually, and in conjunction with the MFCCs, for SV. Two corpora - the NIST SRE 2003 corpus and the CHAINS corpus - are used for the experiments. The results evaluated on the NIST SRE 2003 database, using the *i-vector* framework, reveal that the features extracted from the IMFs, in conjunction with the MFCCs, enhances the performance of the SV system. Further, it is observed that only a small set of lower-order IMFs is useful and necessary for characterizing speaker-specific information. The combination of the features with the MFCCs is also found to be useful when *short speech utterances* of $\leq$ 10 s are used for testing. Similarly, the results evaluated on the CHAINS corpus, using the conventional *Gaussian Mixture Model* (GMM) framework, reveal that the features, in combination with the MFCCs, enhance the performance of the SV system, not only for *normal* speech, but also for *fast* and *whispered* speech. Again, it is observed that only the first few IMFs are needed and useful for achieving such enhanced performance. Further, it is observed that the features derived (in combination with the MFCCs) from the small set of IMFs provide better or equivalent performance compared to the same features derived from a large but fixed AM-FM Gabor filterbank. This exhibits the utility of the *adaptive characteristics* of the *compact* and *concise* EMD/MEMD filterbank.

6.1 Introduction

In the recent times, research in speech based biometric systems and technologies have gained attention. The MFCCs have been the cornerstone for speaker modeling for a long stretch of time [2,13,14]. The MFCCs are obtained by the application of the *Mel filterbank* on the magnitude spectrum of STFT of the speech signal. The Mel filterbank is a non-uniform filterbank, which is based on how the human ear perceives sound using critical bands [2]. The final outcome, the MFCCs, are found to be effective in capturing speaker information in a compact manner. However, the human cognitive system, and its amazing abilities, are far from being completely and properly understood, let alone being replicated by machines. As such, it is quite ambitious to assume that the MFCCs capture all

possible information that could be used by machines for identifying the speaker. In fact, as discussed in Section 1.1.4, this viewpoint is not only valid for speaker recognition but for all speech applications where MFCCs are the principal features of choice.

The performance of the speaker recognition system is found to degrade significantly under different practical challenges [166, 167]. These include degraded acoustic environments, change in speaking styles, short utterances for training and testing, etc., which have always demanded to look for more features representing speaker information. The MFCCs are believed to capture the vocal tract information embedded in the speech signal [2, 13, 14]. As such, features which capture other aspects of speaker information, like the glottal source [168, 169], and long-term information (idiolect, prosody, modulation, etc.) [46, 170–172] have been explored. Apart from these attempts, alternative signal processing approaches (like AM-FM analysis) for modeling speaker information have also been explored, as discussed in Section 1.1.4.

The focus of this work is to investigate the capability of the AM-FM components or the IMFs of the speech signal, as obtained from EMD, and MEMD, in characterizing speaker-specific information. Unfortunately, due to the high time-cost of EEMD (and its variants), they cannot be used in the experiments. As such, the objective of this work is to determine whether the adaptive nature of EMD/MEMD is beneficial in any way for extracting speaker-specific cues. This work investigates whether *energy* or *cepstral-like* features, derived from the IMFs, could represent speaker information in a different manner that the MFCCs. Three different features - *Sum Log Squared Amplitude* [92], *Log Sum Squared Amplitude* [2, 13, 14], and *Entropy* [173] - are extracted over 20 ms frames, from the IMFs, and utilized in the task of text-independent *Speaker Verification* (SV) [2, 13, 14]. The first two features represent cepstral-like features, similar to the MFCCs, but obtained from the filterbank manifested in the IMFs of EMD/MEMD. The Entropy feature is a measure of the information content in the IMFs. We hypothesize that the IMFs are AM-FM signals which vary adaptively with the speech signal, and as such they might capture certain aspects of the signal which a fixed structure like the Mel filterbank may not. Though the Mel filterbank is believed to capture the vocal tract characteristics, it was principally designed from the perspective of speech perception [2, 13, 14]. The IMFs, however, have been shown to capture the vocal tract resonators [122, 147, 161] and the glottal source characteristics [88, 89, 106, 145, 146] of the speech signal (as discussed in Chapter 4), which are production aspects of the speech signal. Just like the perception aspects, the production aspects carry

critical information not only of the speech signal, but also of the speaker uttering the signal. Based on these observations, we expect the features obtained from the EMD/MEMD filterbank to augment the performance of the SV system.

SV represents a binary classification task, where a speaker tests the system against speech utterances pre-recorded from a closed set of speakers [2, 13, 14]. The testing speaker has to test the system against his own pre-recorded speech. If he does so, the SV system, ideally, would recognize the testing speaker as a valid claim. If a speaker from the closed set tests his voice against pre-recorded speech of another speaker, the SV system, ideally, should reject his claim. When there is no specific set of speech sentences/phrases that the testing speaker needs to utter for testing the SV system, it is called a text-independent SV system. This work uses features derived from the speech signal - the MFCCs, and the above mentioned three features - to realize such a SV system.

The principal question of the range of IMFs that is useful for SV is investigated in this paper. The three experimental features are concatenated with 39-dimensional MFCCs, and evaluated in the SV task. The performances of these combinations are compared with 51-dimensional MFCCs to evaluate whether the features really capture speaker information in a different and useful manner. The same three features extracted from the IMFs are also extracted from 20 AM-FM signals obtained from a *20-filter uniform Gabor filterbank* applied to the speech signal. The performance of the SV system for these 20-dimensional features is compared with the features (much lesser dimension than 20) obtained from the IMFs. This enables us to verify the utility of the adaptive nature of the EMD/MEMD filterbank.

Two databases - the NIST SRE 2003 corpus [174], and the CHAINS corpus [41], are used in this work in an attempt to find the relevant answers. From both the databases, standard quality speech utterances, with $F_s = 8$ kHz, are used. The state-of-the-art *i-vector* framework [175] is used for the NIST corpus, whereas the standard *Gaussian Mixture Model* (GMM) [176] is used to train and test the speakers for the CHAINS corpus. Also, extending the experiments, we investigate whether the features derived are useful when applied to *fast* or *whispered* speech [47, 177–180], i.e., when the rate or nature of speech is drastically changed from normal speech. Also, we explore how effective the features derived from the IMFs may be in the case of *limited test-data conditions*, i.e., when *short test-utterances*, of 10 s - 2 s, are used for verifying the claim of the speaker [181, 182]. The performances of the SV system under these scenarios are relevant from the perspective of practical deployment.

The rest of the chapter is organized as follows: Section 6.2 describes the significance of the features extracted from the IMFs for SV. The experimental setup for the experiments is described in Section 6.3. Section 6.4 analyzes the experimental results. Section 6.5 summarizes and concludes this work.

6.2 Significance of the IMFs in characterizing Speakers

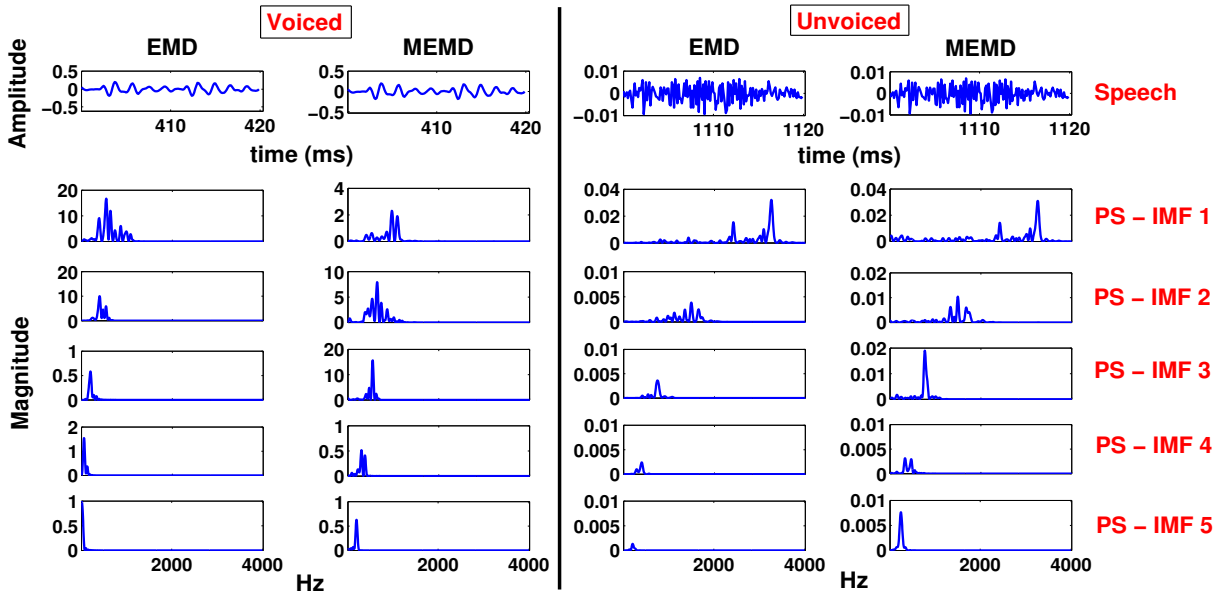


Figure 6.1: Power spectrum (PS), i.e., squared magnitude spectrum, of each of IMFs 1-5 of a 20 ms *voiced speech segment*, **and that of a 20 ms** *unvoiced speech segment*. **The IMFs are generated by EMD (first and third column) and MEMD (second and fourth column).**

As we have observed in Chapter 4, the decomposition provided by EMD (and its variants) exhibits a certain behaviour for voiced speech signals, and a different behaviour for unvoiced speech signals. As a reflection of this phenomenon, we may look at Figure 6.1, which shows the power spectra (squared magnitude spectra) of the first five IMFs, corresponding to two different 20 ms segments of an arbitrary speech signal/utterance, $s(n)$. The topmost plots of the first two columns (from the left) show a *voiced speech segment*, whereas that of the third and fourth columns show an *unvoiced speech segment*. The rest of the plots in each column presents the power spectra of IMFs 1-5, corresponding to the speech segment. The IMFs are obtained from both EMD and MEMD. It may be observed that the power spectra of the IMFs represent different parts of the speech spectrum, as if they have been obtained by bandpass filtering the speech segment. For the two different speech segments, the spectra of the corresponding IMFs are different. *This makes EMD/MEMD akin to using an adaptive filterbank*, as discussed in Chapter 2. For any other speech signal, the spectra of its IMFs will be different for

different parts of the signal. As such, if two speech signals are uttered by two different speakers, their characteristics may be different, and these differences may be exemplified in the IMFs.

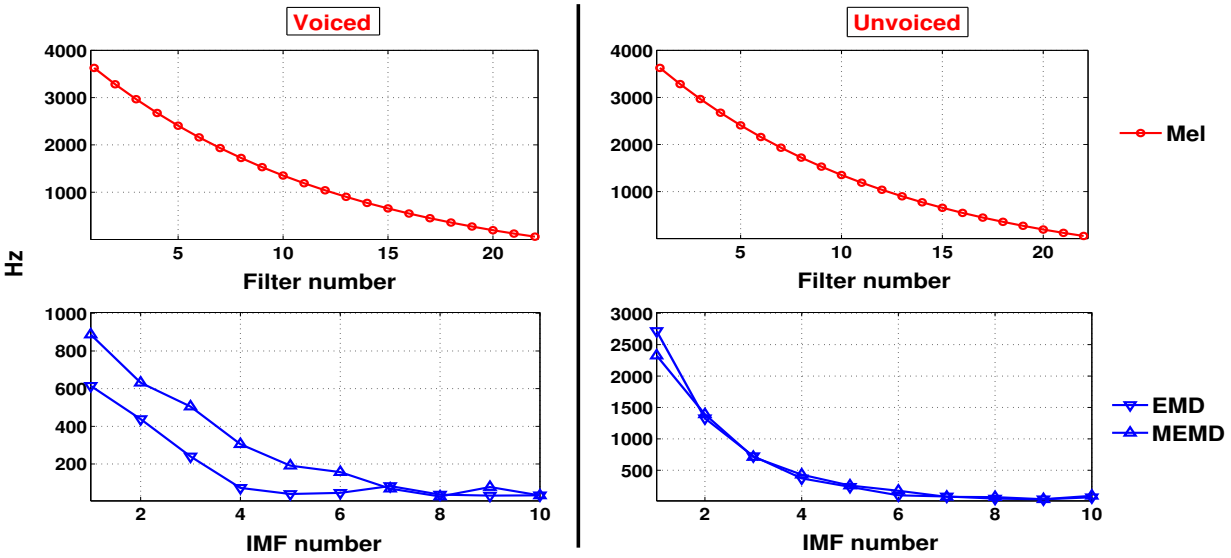


Figure 6.2: [Top left, Top right] : Center frequencies of a 22-filter *Mel filterbank*. [Bottom left, Bottom right] : Mean frequencies of the IMFs corrsponding to a *voiced* and an *unvoiced* speech segment of an arbitrary speech signal.

To explore the utility of the IMFs in SV, we begin by comparing the mean frequencies of the IMFs generated by EMD and MEMD, from the *voiced* and *unvoiced* speech segments, shown in Figure 6.1, with the center frequencies of a 22-filter Mel filterbank. To calculate the mean frequencies, we compute the power spectra of the ten IMFs ($M = 9$), corresponding to each of the speech segments. The mean frequency of IMF_k of a speech segment, $s(n)$, is then calculated from its power spectrum, $S_k(f)$, as per equation (2.28), as,

$$F_k^m = \sum_{f=0}^{F_s/2} \frac{f \times S_k(f)}{\sum_{f=0}^{F_s/2} S_k(f)} \ , \ k = 1, 2, ..., M+1 = 10 \ , \tag{6.1}$$

where, $F_s = 8$ kHz is the sampling frequency of $s(n)$. Figure 6.2 shows the comparison between the Mel center frequencies, and the mean frequencies of the IMFs, obtained from both EMD and MEMD. It is evident that the IMFs of MEMD have higher mean frequencies than that of EMD, particularly in the case of the voiced speech segment, as observed in Chapter 3. It is also clear that the Mel filterbank is a much more precise and detailed filterbank than that represented by EMD or MEMD. The different filters in this detailed filterbank structure capture energies corresponding to different regions of the speech spectrum. As the nature of the signal varies, the energy pattern captured by this filterbank varies. In the case of EMD/MEMD, there is no fixed filterbank structure. Instead the implicit filters

manifested in EMD/MEMD adjust themselves with respect to the signal characteristics. Hence, we can observe significant differences in the center frequencies of the IMFs of the unvoiced speech segment with respect to that of the voiced speech segment. This is how EMD/MEMD captures information from the speech signal in a different manner than the Mel filterbank. As such, we may expect that features extracted from the IMFs could carry different speaker signatures than that carried by the MFCCs.

Besides the differences in the filterbank structures, it is also important to note that the Mel filterbank is principally designed to mimic the band-pass filterbank structure of the human ear. It is a filterbank motivated from the science of speech perception [2]. The EMD/MEMD filterbank is not motivated by speech science. It simply adaptively decomposes the speech signal, which hopefully represents the processes involved in producing the signal. Fortunately, as we have discussed in Chapter 4, the IMFs of the speech signal do manifest the vocal tract resonators and the glottal excitation source, the key components of speech production according to the source-filter theory. The IMFs have, in fact, been used to estimate the formant frequencies [147, 161] and the *pitch* or *fundamental frequency* [88, 89, 145, 146] of the speech signal. Both of these aspects carry important signatures of the speaker. Hence, we may expect the IMFs to be useful in recognizing and discriminating speakers.

6.2.1 Feature Extraction from the IMFs

With the objective of capturing speaker specific information from the IMFs, the speech signal is subjected to a 10-level ($M = 9$) decomposition by EMD/MEMD. Then, the IMFs are processed in short-time segments/frames for extracting features. Let $h_k^i(n)$ represent the i^{th} frame of IMF_k, or $h_k(n)$, of any given speech utterance, $s(n)$, given a framesize of 20 ms and a frameshift of 10 ms. Three features, as described below, are then derived from the IMFs. Let N_f represent the number of samples in a 20 ms frame, and K ($\leq M + 1 = 10$) the number of IMFs from which the features are extracted.

6.2.1.1 Log Sum Squared Amplitude

It is defined as,

$$l_k^i = \log_{10}\Big[\sum_{n=0}^{N_f}\{h_k^i(n)\}^2\Big]\ ,\ k = 1, ..., K \leq M+1\ ;\ L_K^i = [l_1^i, l_2^i, ..., l_K^i]^T\ ;\ L_K = \{L_K^i, \forall i \in \mathbb{N}\} \quad (6.2)$$

Thus, a K-dimensional feature vector is obtained for any given speech frame. The Log Sum Squared Amplitude feature is analogous to the log-energies computed from the Mel filterbank in extracting the MFCCs. The objective here is to capture information that is present in the energies of the IMFs, which are produced by an implicit filterbank structure very different to the Mel filterbank, as discussed earlier.

The feature vector, L_K^i, is further operated upon by *Discrete Cosine Transform* (DCT) [2, 13, 14, 173], resulting in a *refined* feature vector. The DCT of any K-dimensional feature vector extracted from the i^{th} speech frame, $\bar{V}_K^i = [\bar{v}_1^i, \bar{v}_2^i, ..., \bar{v}_K^i]^T$, is given by,

$$\mathrm{c}\bar{v}_p^i = \begin{cases} \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \bar{v}_k^i \cos\left\{\frac{\pi(2k-1)(p-1)}{2K}\right\} \ , \ p = 1 \\ \\ \frac{1}{\sqrt{2K}} \sum_{k=1}^{K} \bar{v}_k^i \cos\left\{\frac{\pi(2k-1)(p-1)}{2K}\right\} \ , \ p = 2, 3, ..., K \end{cases} , \tag{6.3}$$

$$\mathrm{c}\bar{V}_K^i = [\mathrm{c}\bar{v}_1^i, \mathrm{c}\bar{v}_2^i, ..., \mathrm{c}\bar{v}_K^i]^T \ , \ \mathrm{c}\bar{V}_K = \{\mathrm{c}\bar{V}_K^i, \forall i \in \mathbb{N}\} \tag{6.4}$$

The *refined* Log Sum Squared Amplitude feature is, thus, obtained as,

$$\mathrm{c}L_K^i = \mathrm{DCT}\{L_K^i\} = [\mathrm{c}l_1^i, \mathrm{c}l_2^i, ..., \mathrm{c}l_K^i]^T \ , \ \mathrm{c}L_K = \{\mathrm{c}L_K^i, \forall i \in \mathbb{N}\} \tag{6.5}$$

6.2.1.2 Sum Log Squared Amplitude

It is defined as,

$$g_k^i = \sum_{n=0}^{N_f} \log_{10}\left[\{h_k^i(n)\}^2\right] \ , \ k = 1, ..., K \leq M+1 \ ; \ G_K^i = [g_1^i, g_2^i, ..., g_K^i]^T \ ; \ G_K = \{G_K^i, \forall i \in \mathbb{N}\} \tag{6.6}$$

This feature, like the previous feature, also captures the energies embedded in the implicit EMD/ MEMD filterbank. However, the dynamic range of the energies is magnified in this case by taking the logarithm at every sample point. As such, subtle differences in energy might be better exemplified in this feature. This raw feature is further processed by DCT to obtain the *refined* Sum Log Squared Amplitude feature as,

$$\mathrm{c}G_K^i = \mathrm{DCT}\{G_K^i\} = [\mathrm{c}g_1^i, \mathrm{c}g_2^i, ..., \mathrm{c}g_K^i]^T \ , \ \mathrm{c}G_K = \{\mathrm{c}G_K^i, \forall i \in \mathbb{N}\} \tag{6.7}$$

6.2.1.3 Entropy

It is defined as,

$$e_k^i = -\sum_{x \in X} p_k^i(x) \log_2\{p_k^i(x)\} \ , \ k = 1, ..., K \leq M+1 \ ; \ E_K^i = [e_1^i, e_2^i, ..., e_K^i]^T \ ; \ E_K = \{E_K^i, \forall i \in \mathbb{N}\} \tag{6.8}$$

In the above equation, $p_k^i(x)$ is the probability distribution of the amplitudes of $h_k^i(n)$ derived by 16 level histogram quantization in the range $[\max_n\{h_k^i(n)\}, \min_n\{h_k^i(n)\}]$, which results in an alphabet of 16 symbols, given by, $X = \{x_1, x_2, ..., x_{16}\}$.

The Entropy feature is different from the previous two features, and the MFCCs, in that it is not a representation of the energies of the filters constituting the filterbank. It is a measure of the information content of the IMFs, or the implicit EMD/MEMD filters resulting in the IMFs. This feature is further operated upon by DCT, resulting in the *refined* Entropy feature, given by,

$$\mathrm{c}E_K^i = \mathrm{DCT}\{E_K^i\} = [\mathrm{c}e_1^i, \mathrm{c}e_2^i, ..., \mathrm{c}e_K^i]^T \ , \ \mathrm{c}E_K = \{\mathrm{c}E_K^i, \forall i \in \mathbb{N}\} \tag{6.9}$$

6.2.1.4 Motivation behind the features

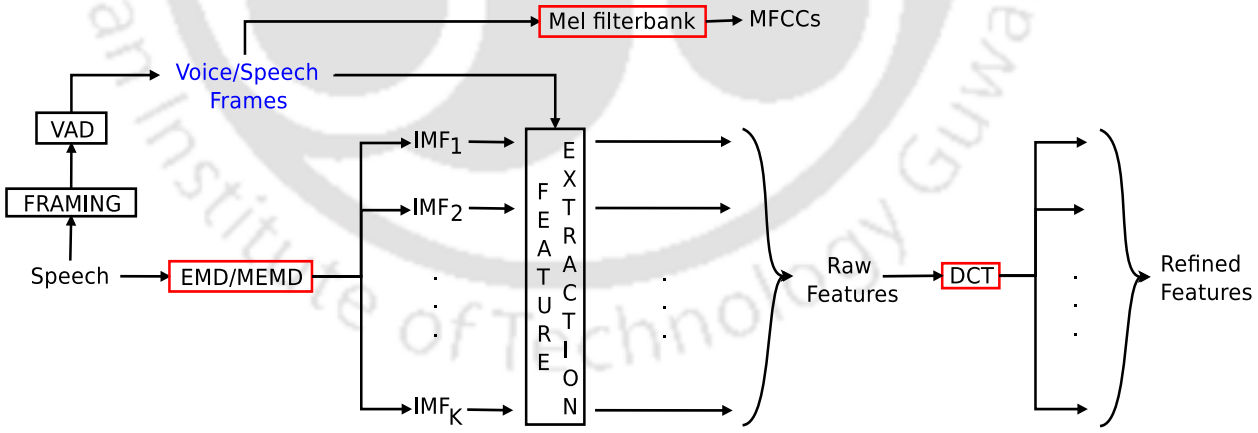


Figure 6.3: **Feature extraction process of the MFCCs, and from the IMFs of Speech. VAD refers to** *Voice Activity Detection.*

Figure 6.3 represents the schematic of the feature extraction process, from the IMFs. The motivation behind the three features directly comes from the objective of this work, as mentioned in the Introduction section. We are investigating the question - *"Does the data-adaptive dyadic filterbank represented by EMD/MEMD capture speaker specific information ?"*. The MFCCs are the

baseline features used in speaker recognition. As such, the features from EMD/MEMD are extracted in a similar fashion as the MFCCs are extracted from the Mel filterbank. Given a speech segment/frame, the MFCCs are obtained from it by the procedure shown in Figure 6.4. Pre-emphasis is done on the speech segment prior to applying the Mel filterbank to reduce the spectral slope of the speech signal. Then, the total energy of the speech segment is segregated in the frequency domain amongst the filters of the Mel filterbank. The logarithm of the energies is taken just to decrease/increase the dynamic range of the energies. DCT is then used to refine the feature vector consisting of these energies. In the case of EMD/MEMD, the total energy of the signal is distributed in the time domain, in its IMFs. So, analogous to the way MFCCs are obtained, the L_K feature represents the log-energies of the speech signal, distributed amongst the implicit filters of EMD/MEMD. Similarly, G_K also represents a measure of the magnitudes or energies of the speech signal belonging to different frequency ranges determined by the adaptive filterbank. The entropy feature, E_K, is used to capture the information content, spread in the adaptive filterbank. DCT is applied to these features, as in the process of extracting the MFCCs, to refine the features in a way that they exhibit their patterns more strongly. This is illustrated in the next subsection.

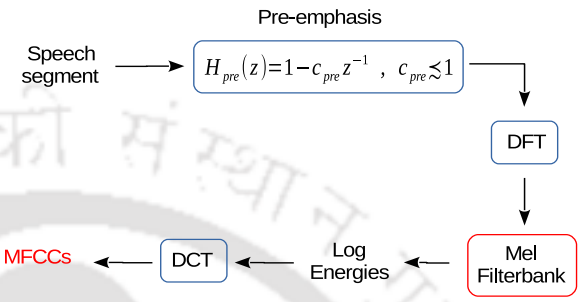


Figure 6.4: **Process of extracting MFCCs from a speech segment.**

6.2.2 Patterns of Feature Variations

To examine whether the IMFs might carry any discriminatory speaker-specific information, six voiced speech utterances, each of 1 s duration, are synthesized, as described in Section 4.2 . Three of the speech utterances have their F_0s centered around 100 Hz, representing male speakers, and three others around 200 Hz, representing female speakers. A cascade of four resonators is used for synthesizing the signals. The resonant frequencies (f_rs) are varied by a margin of 50 Hz within 600-850 Hz, 1100-1350 Hz, 2300-2550 Hz, and 3600-3850 Hz, respectively, for the four resonators. The bandwidths (B_rs) of the resonators are kept fixed at 70, 140, 210 and 280 Hz respectively. The fundamental frequencies, and the formant frequencies, of each of the six signals, are thus given by :

S1 : (95 ; 600, 1100, 2300, 3600) **S4** : (195 ; 750, 1250, 2450, 3750)

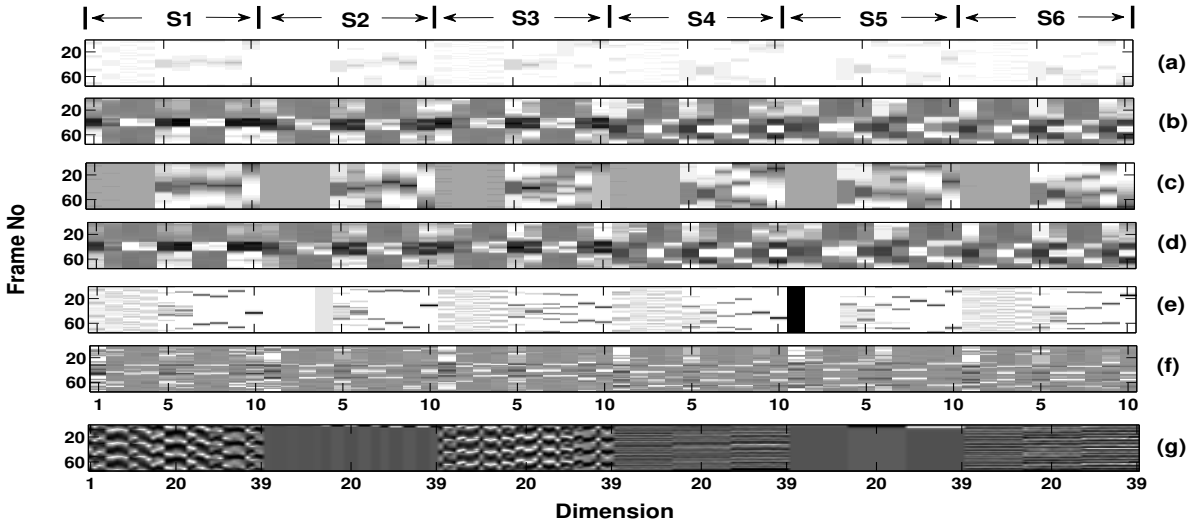


Figure 6.5: **Grayscale images representing the 10-dimensional raw and refined features, derived from the IMFs of EMD, and the 39-dimensional MFCCs, for six different synthetic voiced speech utterances (S1,S2,...S6) representing six different speakers. A framesize of 20 ms with a frameshift of 10 ms is used. (a) L_K, (b) cL_K, (c) G_K, (d) cG_K, (e) E_K, (f) cE_K, and (g) MFCCs. $K = 10$.**

S2 : (100 ; 650, 1150, 2350, 3650) **S5** : (200 ; 800, 1300, 2500, 3800)

S3 : (105 ; 700, 1200, 2400, 3700) **S6** : (205 ; 850, 1350, 2550, 3850)

Figure 6.5 represents, in the form of gray-scale images, the $K = 10$ dimensional feature vectors, derived from the IMFs obtained from EMD, and the 39 dimensional MFCCs feature vector, for all the frames of each of the six synthesized speech sentences. For better visualization, each feature dimension is deducted of its mean, and then normalized to lie in the range of [-1,1]. The MFCCs are also computed over frames of 20 ms, with a frameshift of 10 ms, and consists of 13 cepstral (excluding the 0^{th} coefficient), 13 Δ cepstral (*velocity*), and 13 $\Delta\Delta$ cepstral (*acceleration*) coefficients. The speech signals are pre-emphasized ($H_{pre}(z) = 1 - c_{pre}z^{-1}$, $c_{pre} = 0.98$) prior to extracting the MFCCs. Figure 6.5(a),(b) represent the feature sets L_K and cL_K, respectively. Similarly, Figure 6.5(c),(d) represent the feature sets G_K and cG_K, and Figure 6.5(e),(f) the feature sets E_K and cE_K. Figure 6.5(g) represents the MFCCs.

It is easily observed in Figure 6.5 that the pattern of variations, across the feature dimensions of each of the feature vectors, is different for different speakers. Also, it is evident from the figure that the refined features, derived from EMD, better reflect such patterns than the raw features. This depicts the ability of DCT to spread and enhance the variations of the features across their dimensions. In general, it may also be observed that the patterns of feature variations show more discrimination between the

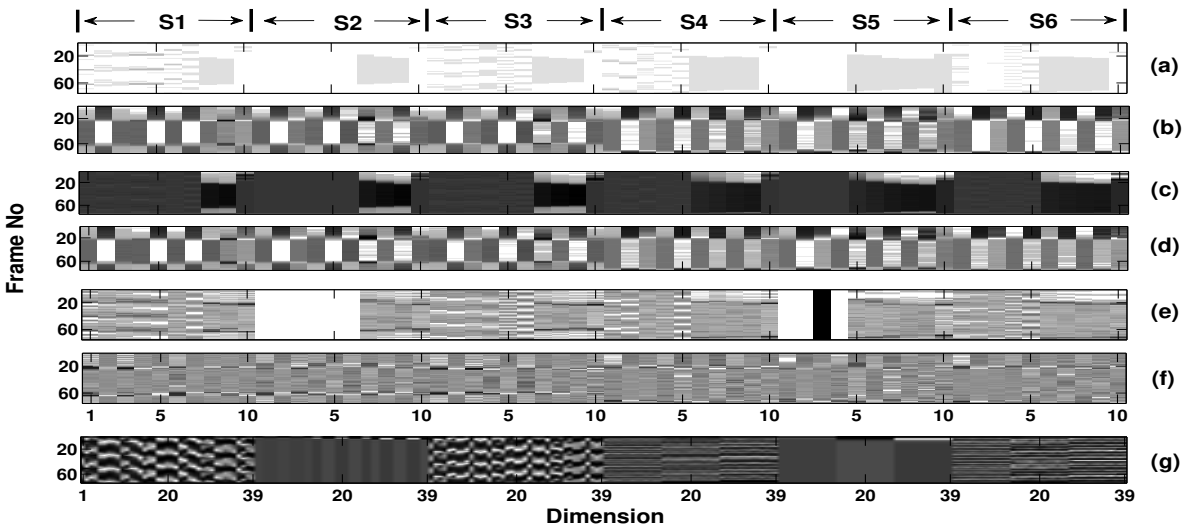


Figure 6.6: **Grayscale images representing the 10-dimensional raw and refined features, derived from the IMFs of MEMD, and the 39-dimensional MFCCs, for six different synthetic voiced speech utterances (S1,S2,...S6) representing six different speakers. A framesize of 20 ms with frameshift of 10 ms is used. (a) L_K, (b) cL_K, (c) G_K, (d) cG_K, (e) E_K, (f) cE_K, and (g) MFCCs. $K = 10$.**

male (S1,S2,S3) and female (S4,S5,S6) speech sentences, and lesser so between the sentences having closer pitch frequencies. This may be attributed to the ability of EMD in characterizing glottal activity in some of its IMFs [106], which we have discussed in Chapters 2 and 4. The same is also observed in the case of the MFCCs.

Figure 6.6 is the equivalent to Figure 6.5, but the features here are derived from the IMFs of MEMD, instead of EMD. The observations of Figure 6.5 are equally applicable to Figure 6.6. However, as may be observed from the comparison of the two figures, the feature variations are much more discernible in the case of MEMD than that of EMD. This may be attributed to reduced mode-mixing in the case of MEMD, and the ability of MEMD to segregate the speech spectrum more effectively. The less overlapping spectral segregation in the case of MEMD, further, enables it to encapsulate the vocal tract resonances in a better way than EMD, as we have already discussed in Chapters 3 and 4.

Figure 6.7 plots the first two dimensions of the three experimental features derived using EMD, and that of the 13 cepstral, 13 Δ cepstral, and 13 $\Delta\Delta$ cepstral dimensions of the 39-dimensional MFCCs. The plots for the features derived using MEMD are similar, and hence not presented here. For better visualization, the mean of each feature-dimension is removed, and then it is normalized to lie in the range of [-1,1]. As is evident from the figure, each of the six synthesized speech utterances, representing six different speakers, form different clusters in the two-dimensional feature space corresponding to each of these features. The plots for the cL_K and cG_K features look similar. This is expected as

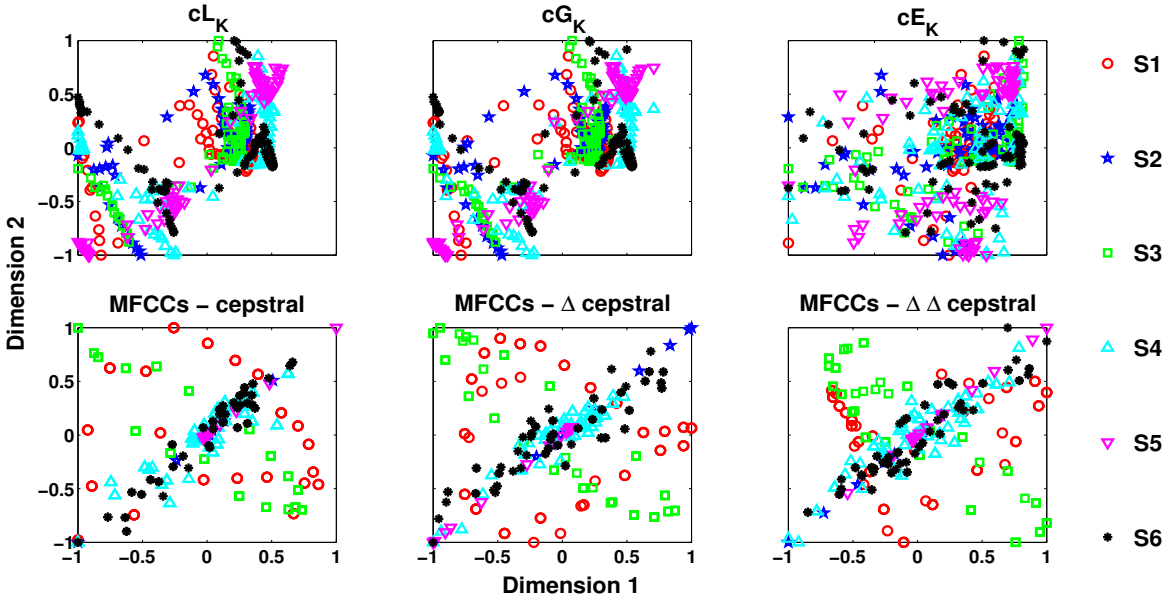


Figure 6.7: X-Y plot of the first two dimensions of each of the following features : cL_K, cG_K, cE_K, MFCCs - cepstral coefficients, MFCCs - Δ cepstral coefficients, and MFCCs - $\Delta\Delta$ cepstral coefficients. The cL_K, cG_K and cE_K features are obtained using EMD, with $K = 10$. The features are plotted for the six synthetic voiced speech utterances, S1-S6, representing six different speakers.

both of them capture energies of the implicit filters of the EMD filterbank. But, there still exists subtle differences between them. The cE_K feature looks quite different from all the other five features represented. More importantly, the EMD features represent different cluster formations, with respect to the MFCCs, for the six utterances. This suggests that they capture speaker specific information in a different way than the MFCCs, and hence may be useful in recognizing speakers in combination with the MFCCs.

With these observations, made under controlled experimental conditions, we now move towards applying the above-mentioned features on natural speech.

6.3 Experimental Setup

All speech utterances considered in this work are of $F_s = 8$ kHz. Three types of filterbanks - the Mel filterbank, the EMD/MEMD filterbank, and a traditional AM-FM filterbank - are used in this work for the experiments. The speech signal is pre-emphasized ($H_{pre}(z) = 1 - c_{pre}z^{-1}$, $c_{pre} = 0.98$) prior to applying the Mel filterbank. Using a 22-filter Mel filterbank, MFCCs and Extended MFCCs (Ext. MFCCs) features are extracted from the speech signal. The MFCCs feature vector is 39-dimensional, and consists of the first 13 cepstral (excluding the 0^{th} coefficient), the first 13 Δ cepstral, and the first

13 $\Delta\Delta$ cepstral coefficients. The Ext. MFCCs feature vector is 51-dimensional, and consists of the first 17 cepstral (excluding the 0^{th} coefficient), the first 17 Δ cepstral, and the first 17 $\Delta\Delta$ cepstral coefficients. The three experimental features, as discussed in the previous section, are extracted from the IMFs of the speech signal. They are tested individually, and in combination with the MFCCs, in the SV system. The dimensions (K) of the three experimental features (in combinations with the MFCCs) are varied to find the subset of the IMFs which are effective in capturing speaker specific information. Again, by comparing the performances of the MFCCs + EMD/MEMD features with the performance of the Ext. MFCCs, we can ascertain whether the EMD/MEMD features are truly useful.

The same three features obtained using EMD/MEMD are also obtained using a large traditional AM-FM filterbank. For this purpose, a 20-filter (close to the number of filters used in the Mel filterbank) Gabor filterbank is used, as shown in Figure 6.8. The Gabor filters are uniformly spaced in frequency, between $0 - \frac{F_s}{2} = 4$ kHz, with each filter having an effective bandwidth of 400 Hz. This results in 20 AM-FM signals from the speech signal, which are then subjected to the same procedure as the IMFs (Figure 6.3), to obtain the same three experimental features from them. However, the dimensions of the feature vectors obtained from traditional AM-FM analysis are kept fixed at $K = 20$. These feature sets are also tested individually, and in combination with the MFCCs, as in the case of the feature sets obtained from the IMFs. Comparing the performances of these MFCCs + AM-FM features with the MFCCs + EMD/MEMD features allows us to evaluate the usefulness of the small but adaptive EMD/MEMD filterbank in capturing speaker specific information.

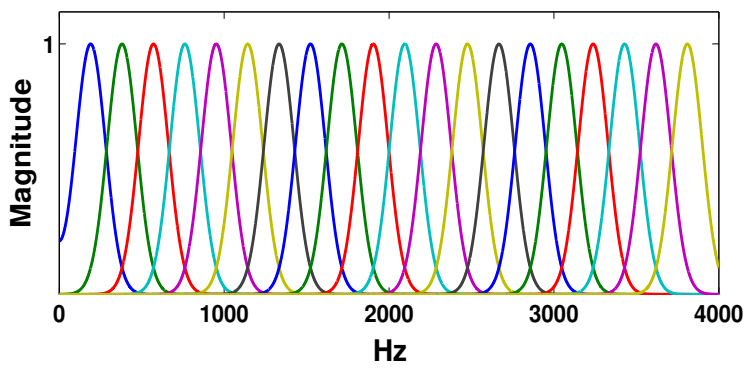


Figure 6.8: A 20-filter uniform Gabor filterbank. Each filter has an effective bandwidth of 400 Hz.

Two corpora - the NIST SRE 2003 corpus [174], and the CHAINS corpus [41] - are used in this study. Using the NIST corpus, the features are tested not only under sufficient data conditions (normal length test speech utterances), but also under limited data conditions. The limited test dataset is obtained by truncating the original test speech utterances to the first 10 s, 5 s, 3 s and 2 s of their actual durations. Using the CHAINS corpus, the features are tested under three different

speech delivery modes - *normal* or *solo*, *fast*, and *whispered* - produced by the same speakers.

The NIST corpus consists of 356 speakers, for which 356 speech utterances, each of ∼2 mins duration, are provided for training. Out of the 356 speakers, 144 are male, and 122 female. The test dataset consists of 2559 speech utterances, whose durations vary between ∼14 s to ∼44 s. The *i-vector* text-independent SV system is used to evaluate the performance of the various features, on the NIST corpus. Apart from the train and test dataset, the implementation of the *i-vector* system requires a *development dataset*, and an *Universal Background Model* (UBM) dataset. The Switchboard Corpus II dataset [183] is used as the development dataset, and a subset of it is taken as the UBM dataset. In the *i-vector* framework, the *mean supervectors*, obtained from GMM of a given speech utterance, is represented in terms of a much lower dimensional vector, called the *i-vector* [175]. This representation is achieved by using a transformation matrix, T_v, called the *total variability matrix*, which includes all the session and channel variabilities of the development dataset. Henceforth, if S_u is the mean supervector of a given utterance, and i_u its corresponding *i-vector*, then they are related as,

$$S_u = S_{\text{UBM}} + T_v i_u \ , \tag{6.10}$$

where S_{UBM} is the mean supervector of the UBM dataset. The *i-vector* so generated from a given utterance does contain session and channel variabilities [175, 184]. To curb the effects of the data conditions, the *i-vector* is processed by *Linear Discriminant Analysis* (LDA), followed by *Within Class Covariance Normalization* (WCCN) [185]. Finally, given such processed train and test *i-vectors*, $\hat{i}_{tr}$ and $\hat{i}_{ts}$, the verification is performed based on the *cosine kernel score*, given by,

$$i_s = \frac{< \hat{i}_{tr}, \hat{i}_{ts} >}{\sqrt{||\hat{i}_{tr}||_2 \ ||\hat{i}_{ts}||_2}} \tag{6.11}$$

The experimental setup for the i-vector framework may be summarized as,

- (i) A gender independent UBM of 512 mixtures using the UBM dataset of approximately around 10 hours. The UBM dataset is a subset of the development datset, and consists of approximately equal amount of data from male and female speakers.
- (ii) A T_v matrix of 400 columns using development data.
- (iii) 200 dimensional LDA matrix using development data *i-vectors*.
- (iv) Full dimensional WCCN matrix using development data *i-vectors*.

The CHAINS corpus consists of 36 speakers (20 male and 16 female), speaking in different *modes of articulation* or speaking styles. Three different speaking styles - *normal* or *solo*, *fast*, and *whispered* - are considered in this work. For *normal* and *whispered* speech, 4 utterances, amounting to ∼3 mins duration, are taken as training data for each speaker. 33 other utterances, of duration varying from ∼1 s to ∼4 s, are provided for each speaker, which are used for testing. For *fast* speech, the larger duration 4 utterances amount to ∼2 mins data, and the testing files vary between ∼1 s to ∼3 s. Standard GMM (using *Maximum Likelihood* estimation) of 512 mixtures is used for implementing the SV system for the CHAINS corpus. This allows the testing of the features on two different modeling systems, apart from two different databases. The *log-likelihood scores* obtained for the test utterances are used for verification. The male speakers are verified against one another, and the female speakers likewise. Assume that $\lambda^s = \{w_m^s, \mu_m^s, \Sigma_m^s \mid m = 1, 2, ..., 512\}$ represents the D-dimensional feature space of a particular speaker modeled by GMM, where w_m^s, μ_m^s and Σ_m^s represent the mixture weight, mixture mean and mixture covariance (diagonal) respectively of the m^{th} Gaussian mixture. Assume that a D-dimensional feature set $\bar{X} = \{\bar{X}_1, \bar{X}_2, ..., \bar{X}_F\}$, consisting of F feature vectors, are obtained from an arbitrary utterance for testing against the model. Then, the log-likelihood score is obtained as,

$$\log P(\bar{X}/\lambda^s) = \sum_{i=1}^{F} \log \Big\{ \sum_{m=1}^{512} w_m^s \, \mathcal{N}(\bar{X}_i/\mu_m^s, \Sigma_m^s) \Big\} \, , \tag{6.12}$$

where

$$\mathcal{N}(\bar{X}_i/\mu_m^s, \Sigma_m^s) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_m^s|^{1/2}} \, e^{\left\{-\frac{1}{2}(\bar{X}_i - \mu_m^s)^{\mathrm{T}} (\Sigma_m^s)^{-1} (\bar{X}_i - \mu_m^s)\right\}} \tag{6.13}$$

The log-likelihood score is then used for verifying whether the testing utterance corresponds to the same speaker as the GMM.

All features, whether obtained from the Mel filterbank, the EMD/MEMD filterbank, or the AM-FM filterbank, are extracted from frames of 20 ms, with a frameshift of 10 ms. Also, energy based *Voice Activity Detection* (VAD), i.e., *speech frame vs. non-speech frame discrimination*, is done on the speech signal, and only the feature vectors corresponding to the speech frames having sufficient energy are retained. VAD eliminates the *silence* frames from the analysis, preserving only the voiced and unvoiced frames. For this work, for any given speech utterance, the speech frames having energy greater than 6 % of the average speech frame energy are considered as *voice/speech frames*, and only

the features corresponding to these frames are processed. The entire feature extraction process is schematically shown in Figure 6.3. Finally, the feature vectors are subjected to *Mean Subtraction and Variance Normalization* [186]. All types of data - train, test, UBM, and development - undergo these procedures, before being finally subjected to the *i-vector* or the standard GMM based SV system.

The final results are reported in terms of the *Equal Error Rate* (EER) and the *Detection Cost Function* (DCF). Let $\mathrm{S}c = \{\mathrm{sc}_1, \mathrm{sc}_2, ..., \mathrm{sc}_S\}$ represent the entire set of scores (*cosine kernel scores* or *log-likelihood scores*) obtained after testing against the speaker models. The set Sc is normalized so that $\{\mathrm{sc}_i \in [0,1] \mid i = 1, 2, ..., S\}$. Let $\xi^s \in [0,1]$ be the decision threshold, above which the claim of the test utterance against the model is accepted. Let S_G denote the number of genuine/true claims that have been accepted. Let S_I denote the number of imposter/false claims that have been rejected. Let S_M denote the number of genuine/true claims that have been rejected. Let S_F denote the number of imposter/false claims that have been accepted. As such, $S = S_G + S_I + S_M + S_F$. We may now define EER and DCF as follows :

Equal Error Rate (EER) : For any given threshold, ξ^s, the *Miss Rate* is given as $\mathrm{MR} = \frac{S_M}{S_G+S_M} \times 100$ %. The *False Alarm Rate* is given as $\mathrm{FAR} = \frac{S_F}{S_I+S_F} \times 100$ %. At a particular threshold, $\xi^s = \xi_0^s$, MR = FAR. This error is known as the EER. In other words, EER = MR = FAR, at $\xi^s = \xi_0^s$.

Detection Cost Function (DCF) : For any given threshold, ξ^s, the *Probability of Miss* is given as $P_M = \frac{S_M}{S_G+S_M}$. The *Probability of False Alarm* is given as $P_F = \frac{S_F}{S_I+S_F}$. Two parameters, C_M and C_F, assign costs to the event of a *miss* (a genuine claim rejected) and that of a *false alarm* (an imposter claim is accepted), respectively. Also, an *a priori* probability, P_T, is assigned, which assumes that out of all the test claims against a speaker, only a fraction P_T are genuine claims. The cost parameter of the DCF, at any given ξ^s, is given by,

$$C_{\xi^s} = C_M \times P_M \times P_T + C_F \times P_F \times (1 - P_T) \tag{6.14}$$

The DCF is then given by,

$$\mathrm{DCF} = \min_{\xi^s \in [0,1]} C_{\xi^s} \tag{6.15}$$

In this work, $C_M = 10$, $C_F = 1$, and $P_T = 0.1$ are considered, for both the databases.

6.4 Results and Analysis

In this section, we present the experimental results for the text-independent SV systems implemented on the NIST and the CHAINS corpora, for the features derived from the Mel filterbank, EMD/MEMD, and the AM-FM filterbank. The performances of the experimental features are evaluated individually, and in combinations with the MFCCs.

6.4.1 Performances of the features using the *i-vector* SV system

Table 6.1: Performances of the MFCCs, the Ext. MFCCs, and the 10-dimensional EMD/MEMD features, on the NIST SRE 2003 corpus.

Technique →	Mel filterbank		EMD			MEMD		
Feature →	MFCCs	Ext. MFCCs	cG_K	cE_K	cL_K	cG_K	cE_K	cL_K
Dimension →	39	51	K = 10			K = 10		
EER (%)	2.76	2.66	28.14	34.55	29.67	24.66	34.19	26.38
DCF	0.0485	0.0484	0.5328	0.6556	0.5617	0.4639	0.6474	0.4993

Table 6.2: Performances of the MFCCs, the combinations of the 10-dimensional EMD/MEMD features with the MFCCs, and the Ext. MFCCs, on the NIST SRE 2003 corpus.

Technique →	Mel filterbank		EMD			MEMD		
Feature →	MFCCs	Ext. MFCCs	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K
Dimension →	39	51	39 + { K = 10 }			39 + { K = 10 }		
EER (%)	2.76	2.66	2.98	**2.62**	2.89	2.93	**2.53**	2.94
DCF	0.0485	0.0484	0.0541	**0.0484**	0.0534	0.0528	**0.0473**	0.0554

Table 6.1 shows the SV performances of the 39-dimensional MFCCs, the 51-dimensional Ext. MFCCs, and the three refined features derived from all ten IMFs of EMD and MEMD. It is clear that the features derived from the IMFs alone cannot match the performance of the MFCCs. The *dyadic filterbank* structure of EMD/MEMD is compact and adaptive, but lacks the detail of the 22-filter *Mel filterbank*. Due to the reduced mode mixing, the features derived from MEMD perform better than those derived from EMD, though much inferior to the MFCCs and the Ext. MFCCs. Also to be noted that the performance of the Ext. MFCCs is marginally better than that of the MFCCs, though it

contains an additional 12 dimensions. Table 6.2 lists the performances of the 10-dimensional refined features when they are concatenated with the MFCCs. As is observed from the table, the refined Entropy feature enhances the performance of the SV system, particularly in the case of MEMD. This 49-dimensional combination outperforms the Ext. MFCCs as well, which suggests that they represent speaker specific information more efficiently than the higher dimensions of the Ext. MFCCs. The other two features do not enhance the performance of the *i-vector* SV system. However, we investigate further to find out if there is an optimum set of IMFs which could enable these two features to augment the performance of the SV system, and to further improve the performance of the system for the Entropy-MFCCs combination.

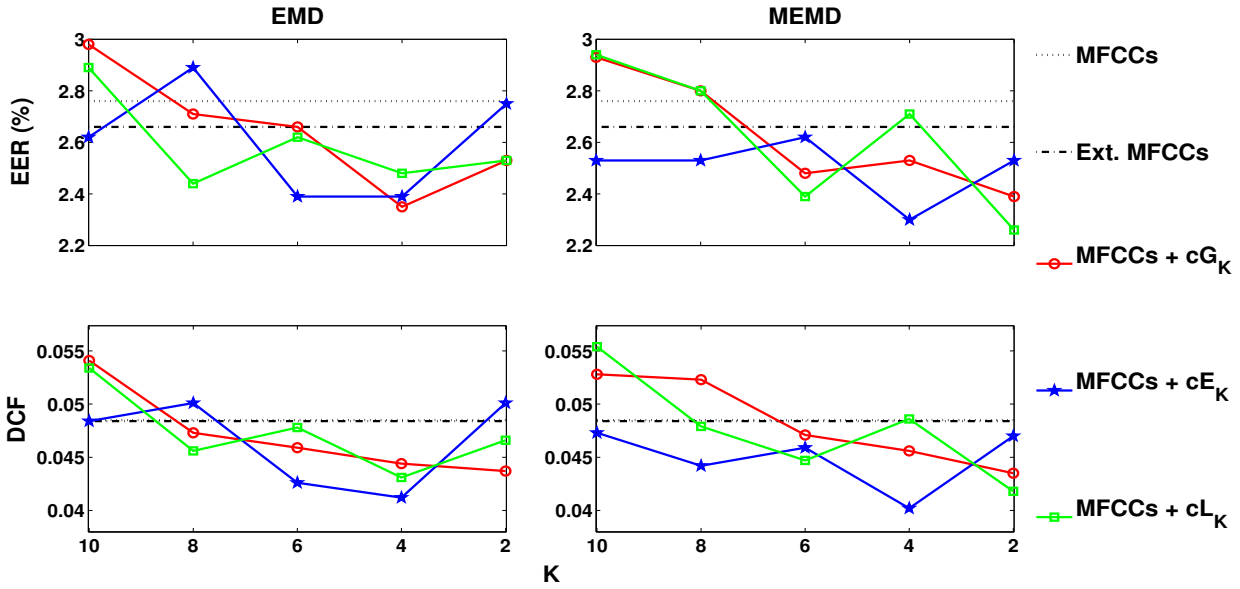


Figure 6.9: **Performances of the combinations of the EMD/MEMD features with the 39-dimensional MFCCs, on the NIST SRE 2003 corpus. The dimensions of the EMD/MEMD features are decreased from 10 to 2.**

Figure 6.9 plots the performances of the EMD/MEMD features, in combinations with the MFCCs, as the dimensions of the EMD/MEMD features are reduced from 10 to 2, in steps of 2. As is observed in the figure, all three features improve the performance of the SV system when their dimensions are reduced, in general for $K \leq 6$. If we assume that EMD/MEMD behaves as a perfect dyadic filterbank, then for 8 kHz speech, IMF_k will have its mean frequency, $F_k^m \leq \frac{4000}{2^{j-1}}$ Hz. Practically, beyond IMF_6, the IMFs have their dominant frequencies below the human pitch frequency range, which might be the reason why the inclusion of such IMFs is insignificant for speaker modeling. It may also be observed that the MEMD features provide similar (or better) performances as the EMD features, at a lesser

dimension. All the features, particularly cG_K and cE_K, achieve their best performance when $K \leq 4$, indicating that they primarily capture speaker specific information from the IMFs representing the system characteristics of the speech utterances, as discussed in Chapter 4. These feature combinations outperform the Ext. MFCCs as well, which suggests that the EMD/MEMD filterbank is indeed useful in capturing speaker specific information.

Table 6.3: Performances of the best combinations of the 39-dimensional MFCCs with the EMD/MEMD features, on the NIST SRE 2003 corpus. The test data duration is varied from Normal (≥14s) to 2s.

Technique →		Mel filterbank		EMD				MEMD		
Feature →		MFCCs	Ext. MFCCs	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K	MFCCs + cL_K	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K
Dimension →		39	51	39 + { K = 4 }	39 + { K = 4 }	39 + { K = 4 }	39 + { K = 8 }	39 + { K = 2 }	39 + { K = 4 }	39 + { K = 2 }
Test Data	Metric									
Normal	EER (%)	2.76	2.66	2.35	2.39	2.48	2.44	2.39	2.3	**2.26**
	DCF	0.0485	0.0484	0.0444	0.0412	0.0431	0.0456	0.0435	**0.0402**	0.0418
10 s	EER (%)	6.59	5.74	6.05	**5.64**	5.69	5.96	6.5	6.05	6.14
	DCF	0.1183	0.1047	**0.103**	0.1055	0.1044	0.1122	0.1146	0.1074	0.1106
5 s	EER (%)	11.74	10.79	11.2	**10.29**	11.02	11.29	11.97	11.65	11.38
	DCF	0.2217	0.2012	0.1994	**0.1892**	0.2052	0.2064	0.223	0.2149	0.2117
3 s	EER (%)	18.93	**16.71**	16.89	16.89	17.29	17.61	18.02	17.07	18.25
	DCF	0.352	0.311	**0.3026**	0.3175	0.3253	0.3231	0.3375	0.316	0.3388
2 s	EER (%)	23.85	**21.59**	21.95	22.54	22.31	23.17	24.12	22.4	23.08
	DCF	0.448	**0.403**	0.4091	0.4218	0.4177	0.4348	0.4445	0.4209	0.4357

Table 6.3 lists the performances of the best combinations of the MFCCs + EMD/MEMD features, under limited data conditions. Experiments under limited data conditions are desirable from the viewpoint of the speaker, who would like to test the system using a password or a short codeword rather than lengthy utterances [181, 182]. As can be seen from the table, the performance of the MFCCs degrades starkly as the data available for testing becomes less abundant. The combinations of the features derived from the IMFs with the MFCCs are observed to enhance the performance of the SV system, for all the cases. The best performance for each limited data condition is highlighted in bold. It is observed that the EMD features provide slightly better performance than the MEMD features, under limited test data conditions. For the cG_K and cL_K features, this may be attributed

to the fact that the MEMD features use a smaller dimension than the EMD features, which may not be helpful when the test dataset is small. When abundant data is available for testing, the higher dimensions are redundant, but when the dataset is limited, the higher dimensions may be useful. The improved performances observed for the Ext. MFCCs also indicate this assertion. The Ext. MFCCs are observed to provide a significant improvement in the performance of the SV system, by means of its additional 12 dimensions. The performances of the MFCCs + EMD/MEMD features are not far-off, particularly considering the fact that the EMD/MEMD features add much fewer dimensions.

Table 6.4: Performances of the three experimental features, derived from traditional AM-FM analysis, and their combinations with the 39-dimensional MFCCs, on the NIST SRE 2003 corpus. The test data duration is varied from Normal (≥14s) to 2s.

Technique →		Mel filterbank		AM-FM analysis : 20 filter uniform Gabor filterbank					
Feature →		MFCCs	Ext. MFCCs	cG_K	cE_K	cL_K	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K
Dimension →		39	51	$\{K = 20\}$			$39 + \{K = 20\}$		
Test Data	Metric								
Normal	EER (%)	2.76	2.66	5.1	19.02	4.47	**2.57**	2.66	2.84
	DCF	0.0485	**0.0484**	0.0943	0.3544	0.082	0.0491	0.0497	0.0508
10 s	EER (%)	6.59	**5.74**	10.12	25.79	9.8	6.73	6.23	6.59
	DCF	0.1183	**0.1047**	0.19	0.4867	0.1792	0.1265	0.1127	0.1239
5 s	EER (%)	11.74	**10.79**	16.62	32.16	16.4	12.2	11.38	12.06
	DCF	0.2217	**0.2012**	0.3106	0.5987	0.3062	0.2288	0.2135	0.2258
3 s	EER (%)	18.93	**16.71**	22.58	36.72	23.04	18.74	17.34	19.15
	DCF	0.352	**0.311**	0.4218	0.6904	0.4276	0.348	0.3257	0.3558
2 s	EER (%)	23.85	**21.59**	28	38.93	28.27	23.76	22.18	24.07
	DCF	0.448	**0.403**	0.5254	0.7372	0.5321	0.4447	0.4618	0.4538

Finally, in Table 6.4, the performances of the three experimental features, cG_K, cE_K, and cL_K, extracted from the fixed AM-FM filterbank, are enlisted. A comparison of the performances of the isolated features in Table 6.4 with that in Table 6.1 shows that the large detailed Gabor filterbank is significantly better than the small concise adaptive filterbank represented by EMD/MEMD. However, the performances of the isolated features obtained from the Gabor filterbank, particularly that of cL_K and cG_K, are still quite inferior to that of the MFCCs, which highlights the significance of the

MFCCs. In combination with the MFCCs, only the refined Entropy feature, cE_K, shows consistent improvement in the performance of the SV system. The other two features, in combination with the MFCCs, result in degradation. A comparison of the performances enlisted in Table 6.4 with that in Table 6.3 shows that the three features extracted from the IMFs (in combination with the MFCCs) provide a better performance than the same features extracted from the traditional AM-FM filterbank, and at a much smaller dimension. This shows that the adaptive filterbank of EMD/MEMD, though concise, is more capable of complementing the Mel filterbank than the detailed but fixed filterbank used in traditional AM-FM analysis.

6.4.2 Performances of the features using the GMM based SV system

Table 6.5: Performances of the combinations of the EMD/MEMD features with the 39-dimensional MFCCs, on the CHAINS corpus. Speech utterances of *normal* or *solo* speaking style are used. The dimension of the EMD/MEMD features are decreased from 8 to 2.

Technique →		Mel filterbank		EMD			MEMD		
Feature →		MFCCs	Ext. MFCCs	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K
Dimension	Metric								
39 + $\{K = 8\}$	EER (%)	7.74	6.31	7.32	7.49	7.32	7.58	7.66	7.49
	DCF	0.1443	0.1175	0.1348	0.1421	0.1331	0.1399	0.1395	0.1405
39 + $\{K = 6\}$	EER (%)	-	-	7.07	7.57	6.99	**7.07**	7.49	**7.15**
	DCF	-	-	0.1336	0.1405	0.1259	**0.1316**	0.1411	**0.1328**
39 + $\{K = 4\}$	EER (%)	-	-	**6.82**	7.58	**6.65**	7.58	7.49	7.58
	DCF	-	-	**0.1274**	0.1419	**0.1234**	0.1402	0.1396	0.143
39 + $\{K = 2\}$	EER (%)	-	-	7.24	**7.49**	7.41	7.99	**7.07**	7.83
	DCF	-	-	0.137	**0.1352**	0.1337	0.1507	**0.1333**	0.1398

Table 6.5 shows the performance of the GMM based text-independent SV system on the CHAINS corpus, where *normal* mode of speech articulation is used by the speakers. The dimensions of the EMD/MEMD features are decreased from $K = 8$ to $K = 2$ to find out the optimum set of IMFs for which the features provide the best performances, in combination with the MFCCs. As already ascertained from the experiments in the NIST corpus, the lower-frequency IMFs are not particularly useful, and hence $K > 8$ is not considered. As can be observed from the table, all the combinations provide an improvement in the performance of the SV system. The best performances, again, are

observed for $K \leq 4$ for the EMD features, and $K \leq 6$ for the MEMD features. For the EMD features, the performances of the features for $K = 4$ and $K = 6$ dimensions are very similar. The table indicates that for the refined Entropy feature the higher-frequency IMFs are more useful, whereas the other two features, cL_K and cG_K, also capture glottal source information present in the intermediate IMFs. Again, the Ext. MFCCs, by means of its additional 12 dimensions, provides considerable improvement. The best performances of the MFCCs + EMD/MEMD features are, however, not far off, considering that much fewer dimensions are added.

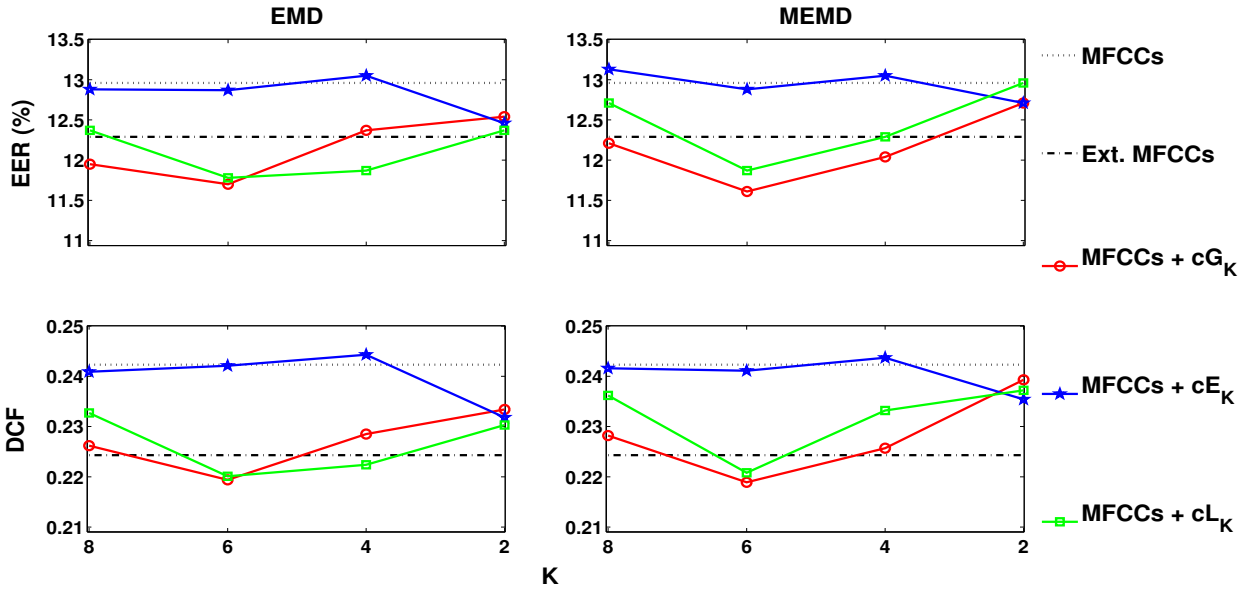


Figure 6.10: **Performances of the combinations of the EMD/MEMD features with the 39-dimensional MFCCs, on the CHAINS corpus. Speech utterances of *fast* speaking style are used. The dimension of the EMD/MEMD features are decreased from 8 to 2.**

Figures 6.10 and 6.11 plot the performances of the combined features, for varying dimensions, for *fast* and *whispered* modes of articulation, respectively. One can easily notice the significant drop in performances of the MFCCs and the Ext. MFCCs in these scenarios, compared to *normal* speech. For *fast* speech, all the combinations, particularly for $K \leq 6$, provide performance enhancement compared to the MFCCs. The cG_K and cL_K features (in combination with the MFCCs), at $K = 6$, in fact outperform the Ext. MFCCs. Interestingly, for *whispered* speech, the Ext. MFCCs perform poorly compared to the MFCCs. In this scenario, all the combinations, with $K \leq 6$, provide performance enhancement compared to the MFCCs and the Ext. MFCCs. A comparison of the performances in Table 6.4, and Figures 6.10 and 6.11, show that the MEMD features exhibit a consistent pattern compared to the EMD features. In the case of MEMD, the cL_K feature always

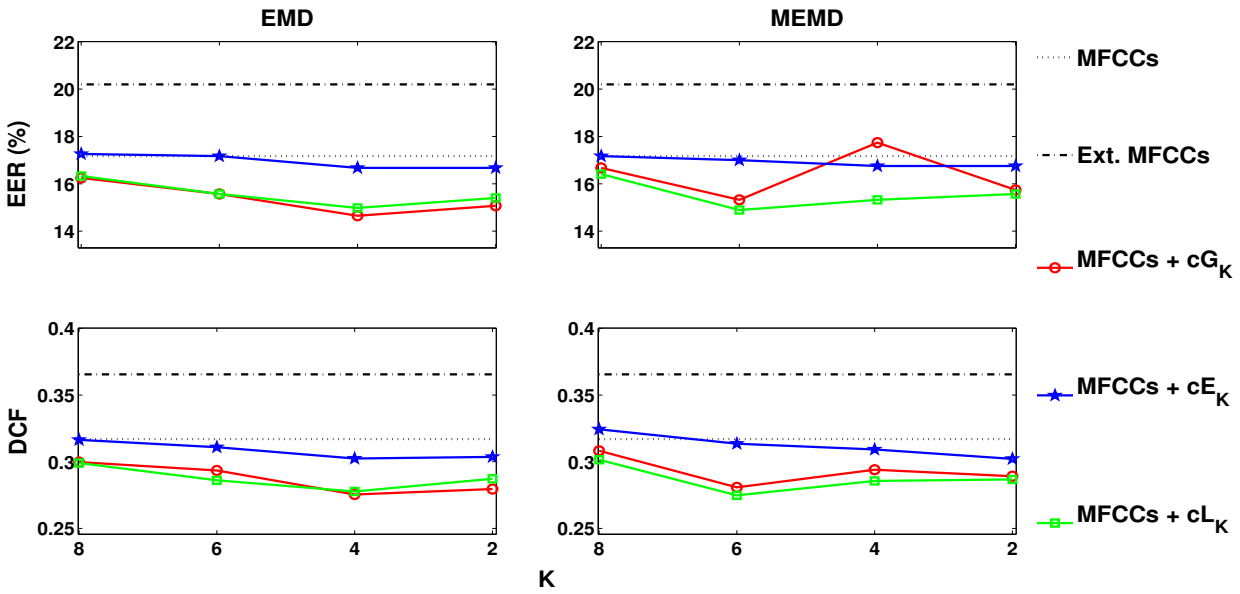


Figure 6.11: Performances of the combinations of the EMD/MEMD features with the 39-dimensional MFCCs, on the CHAINS corpus. Speech utterances of *whispered* speaking style are used. The dimension of the EMD/MEMD features are decreased from 8 to 2.

gives optimum performance at $K = 6$, in combination with the MFCCs, for all three types of speech articulations. So, does the cG_K feature. Similarly, the cE_K feature provides its optimum performance at $K = 2$. In the case of the EMD features, the optimum performances are observed at different dimensions, as the type of articulation changes. This is an assertion of the fact that the IMFs of MEMD have more stable characteristics, with reduced mode mixing. As ascertained from the experiments in the NIST corpus, the individual features are not capable of matching the performance of the MFCCs, and hence they are not explored for the CHAINS corpus.

The improvements shown in Figures 6.10 and 6.11 are relevant from the point of view of practical implementation of the SV system. As is evident from the figures, when the rate or nature of speech differs from the normal mode of speech the SV system suffers. This is because if the speakers are unco-operative, or speak at a fast rate, or in a lower register (whispering), say to maintain privacy, the vocal tract resonances and the glottal source vibration differ significantly from the normal speaking pattern [47, 177–180]. As such, the improvements observed in Figures 6.10 and 6.11 are encouraging, and more experiments might be considered in the future in this regard.

Finally, for the sake of comparison, the performances of the three experimental features (in combination with the MFCCs), derived from the 20 filter uniform Gabor filterbank, are presented in Table 6.6. As is evident from the table, the large filterbank does not perform consistently for

different types of speech. While the performance degrades for normal speech, there is some improvement in the case of *fast* and *whispered* speech. Even under these circumstances, the best performances of the features derived from the IMFs are comparable to (or better than) the performances of the same features obtained using the traditional AM-FM filterbank. This, again, confirms the effectiveness of the adaptive filterbank manifested in the small set of lower-order IMFs.

6.5 Conclusion

This work investigated the utility of the time-domain components or the IMFs of the speech signal, obtained from the adaptive AM-FM analysis technique - EMD, and its variant MEMD, in characterizing speakers for the task of SV. It was observed that the IMFs of EMD/MEMD could be utilized to extract features, which, in combination with the MFCCs, could augment the performance of the SV system. Individually, however, the IMFs were not effective in discriminating speakers. It was observed that only a small set of IMFs, up to the first six IMFs, were useful in enhancing the performance of the SV system. In general, the experiments on the NIST SRE 2003 corpus, using the *i-vector* system, showed that the IMFs are not only useful for characterizing speakers when sufficient test data is available, but also when there is a scarcity of data required for verifying a claim. The experiments on the CHAINS corpus, using standard GMM to train and test the speakers, again, showed that the IMFs are useful not only for SV when the speakers exhibit *normal* mode of speech,

Table 6.6: Performances of the combinations of the three experimental features, derived from traditional AM-FM analysis, with the 39-dimensional MFCCs, on the CHAINS corpus. Speech utterances of *normal* or *solo*, *fast*, and *whispered* speaking styles are used.

Technique →		Mel filterbank		AM-FM analysis : 20 filter uniform Gabor filterbank		
Feature →		MFCCs	Ext. MFCCs	MFCCs + cG_K	MFCCs + cE_K	MFCCs + cL_K
Dimension →		39	51	39 + { K = 20 }		
Style	Metric					
Normal	EER (%)	7.74	**6.31**	9.18	9.09	8.75
	DCF	0.1443	**0.1175**	0.1632	0.1674	0.1648
Fast	EER (%)	12.96	12.29	**11.95**	14.48	12.62
	DCF	0.2423	0.2243	**0.2225**	0.2671	0.2346
Whispered	EER (%)	17.17	20.2	14.81	20.54	**13.8**
	DCF	0.317	0.3655	0.2791	0.3807	**0.2613**

but also when *fast* and *whispered* speaking styles are used by the speakers. Overall, the MEMD features, owing to reduced mode mixing in its IMFs, provide a consistent pattern of performance compared to the EMD features as the type of articulation varies.

7

Summary and Conclusions

Contents

Outline

This chapter, the final chapter of this thesis, summarizes the experiments done as part of this thesis, the results accumulated from them, and the conclusions arrived at after the analysis of the results. Towards the end, some criticism of the work done in the thesis, and future directions of work are also mentioned for researchers who would like to use EMD for speech processing applications.

7.1 Summary

The objective of this thesis was to develop and utilize the technique of EMD for the purpose of adaptive AM-FM analysis of speech. As such, this thesis commenced with a study of the requirement of AM-FM analysis in speech processing. **The first chapter of the thesis - Introduction** - was devoted to this cause. The first chapter discussed the limitations of conventional short-time analysis of speech, specifically STFT, LP analysis, and the Mel filterbank. It then introduced, and discussed in detail the standard mechanism of AM-FM analysis of speech. The chapter concluded by discussing the limitations of traditional AM-FM analysis, and the improvements required to make AM-FM analysis more effective for speech processing.

Resuming from the conclusion of the first chapter, **the second chapter of the thesis - Empirical Mode Decomposition : A Review** - discussed the technique of EMD as an *adaptive* AM-FM analysis method for speech processing. This chapter was completely devoted to studying the characteristics of EMD, and its limitations or undesirable properties (specifically *mode-mixing*). The chapter continued by discussing the noise-assisted advanced variants of EMD, and their advantages and disadvantages over standard EMD, DWT and traditional AM-FM analysis. Parallely, the chapter showcased the ability of EMD (and its noise-assisted variants) as a method for performing adaptive AM-FM analysis. The chapter concluded by laying out the framework of this thesis.

The third chapter of the thesis - Modified Empirical Mode Decomposition - attempted to fulfill the first objective of the thesis. The objective was to *"reduce mode-mixing but at a much lesser time-cost than the noise-assisted EMD variants"*. Based on careful observation of the EEMD algorithm, experiments are undertaken to increase the number of IPs in the standard EMD process. This resulted in three different *modified* EMD algorithms, out of which two are able to successfully reduce mode-mixing. These algorithms were compared with DWT and standard AM-FM analysis. Out of the two successful modified EMD algorithms, the one with the lesser time-cost was declared

as the *Modified Empirical Mode Decomposition* (MEMD) algorithm.

The fourth chapter of the thesis - Analysis of the Source and System characteristics in the IMFs - attempted to fulfill the second objective of the thesis. The objective was to *"investigate the distribution of the source and system characteristics of the speech signal in its IMFs"*. The EMD process (and hence its variants) decomposes the signal based on its waveform characteristics. The information contained in the speech signal, however, may not be attributed to its waveform shape only. As such, a telephone quality speech signal is highly intelligible even though its waveform shape differs from the original speech signal. With this viewpoint, investigations were done using synthetic speech signals, natural speech signals, and telephone quality versions of natural speech signals. The objective was to find out the behaviour of EMD, MEMD, and ICEEMDAN to speech signals corresponding to different waveform shapes, but whose underlying production mechanism was more or less the same. A wide range of voiced and unvoiced speech signals, corresponding to different speech sounds or phones, were considered for the study. The well-accepted technique of *cepstral* or *homomorphic* analysis was utilized to study the source and system characteristics manifested in the IMFs.

The fifth chapter of the thesis - Detection of the Glottal Closure Instants using EMD - catered to the third objective of the thesis. The objective was to *"detect the GCIs of the voiced regions of the speech signal, using its IMFs"*. The broad idea was to provide an alternative to the state-of-the-art methods of estimating the GCIs, which are dependent on conventional short-time analysis of speech. Based on the observation that most of the popular methods use *sinusoid-like* signals for estimating the GCIs, this work investigated if the IMFs of the speech signal could be utilized for the task. A principle or framework was proposed based on experimental observation of the *mean frequencies* of the IMFs of synthetic and natural speech signals, for detecting the GCIs from their IMFs. Two similar methods were then developed, one using the IMFs of ICEEMDAN, and the other using the IMFs of MEMD. The two methods, called IGE and MGE, respectively, were tested under normal, noisy, and telephone channel conditions.

The sixth chapter of the thesis - Analysis of the Intrinsic Mode Functions for Speaker information - catered to the fourth objective of the thesis. The objective was to *"show that the small but adaptive filterbank manifested in the IMFs capture speaker specific information that can complement the Mel filterbank"*. Mimicking the process of extracting the MFCCs from the speech signal, three different features were obtained from the IMFs of the signal. The features were then

tested individually, and in combination with the MFCCs, in the task of SV. The features were tested using the *i-vector* based SV syatem, and the standard GMM based SV system, on two different corpora. The features were tested not only under normal test conditions, but also under limited test data conditions. The features were also tested for fast and whispered speaking styles, apart from normal speech articulation.

7.1.1 Conclusions

The important conclusions made at the end of each chapter may be summarized as :

- **Chapter 1 :** AM-FM analysis aims to provide an analysis of the speech signal without being limited to the assumptions of short-time stationarity and linearity. It is an alternative to the conventional methods of speech processing like STFT, LP analysis, and Mel filterbank.
  - AM-FM analysis represents the speech signal as a combination of a small and finite number of time-varying AM-FM signals, representing its resonances or centers of energy.
  - Traditional AM-FM analysis is obtained using MDA, which provides a fixed analysis dependent on the design of a large filterbank.
  - Along with useful speech components, a multitude of unwanted AM-FM signals are also generated using MDA. Hence, an adaptive, concise, and meaningful AM-FM analysis technique is desired.

- **Chapter 2 :** EMD and its noise-assisted variants have the ability to circumvent much of the limitations of MDA, and provide an adaptive, concise, and meaningful AM-FM analysis of the speech signal.
  - EMD decomposes the speech signal without requiring any pre-knowledge about the characteristics of the signal, i.e., it requires no *a priori* basis, unlike techniques like WT.
  - EMD exhibits the phenomenon of mode-mixing, which may inhibit its utility in different applications. Also, the decomposition of the signal into its actual components, using EMD, is dependent upon the separation between their frequencies and their relative strengths.
  - Noise-assisted EMD variants, namely EEMD and its advanced versions (like ICEEMDAN), reduce mode-mixing but at the expense of time.

- EMD and its noise-assisted variants seem to manifest the vocal tract resonances and the glottal source information of the speech signal in its IMFs. The segregation of the vocal tract resonances seems to be better in the case of the noise-assisted variants.

- **Chapter 3 :** Increasing the number of IPs in the EMD algorithm results in reduced mode-mixing, and a better segregation of the speech spectrum amongst the IMFs of the speech signal.
  - Increasing the number of IPs in the *inner residue* signal (from its number of extrema) partially mimics the EEMD algorithm.
  - The resulting algorithm, called MEMD, reduces mode-mixing, but does not have a high time-cost like EEMD or its variants.
  - The speech spectrum is more equitably segregated amongst the IMFs of the speech signal, in the case of MEMD, compared to that of EMD. The higher-frequency formants of voiced speech are also better represented in the IMFs of MEMD.

- **Chapter 4 :** EMD, MEMD, and ICEEMDAN can segregate the speech signal into its system and source characteristics in different subsets of its IMFs.
  - The decomposition is adaptive to different types of speech - voiced and unvoiced.
  - The behaviour of the decomposition is similar for different voiced speech sounds or phones. The decomposition of telephone quality speech is also similar to that of original speech.
  - A more equitable segregation of the source and system characteristics is observed in the case of ICEEMDAN, followed by MEMD and EMD.

- **Chapter 5 :** Non-linear and non-stationary analysis, based on ICEEMDAN and MEMD, can be used for detecting the GCIs, just like conventional methods primarily based on LP analysis.
  - Discarding the extreme high-frequency (lower-order) IMFs, the partial summation of the rest of the IMFs results in a signal (SEDS) that manifests the GCIs in some of its minima locations.
  - With the aid of another signal, from which broader regions of search of the GCIs may be estimated, the SEDS can be successfully used to estimate the GCIs.

  - The proposed algorithms using ICEEMDAN and MEMD, captioned IGE and MGE, respectively, provide credible estimates, not only for clean speech signals but also for speech signals corrupted by different types of noise, and influenced by telephone channel conditions.

- **Chapter 6 :** The adaptive EMD/MEMD filterbank captures speaker-specific information in a different manner than the fixed Mel filterbank, and is useful in enhancing the performance of the SV system.
  - In combination with the 39-dimensional MFCCs, the cepstral or energy-like features extracted from the IMFs enhance the performance of the SV system.
  - The features extracted from the IMFs, at lower dimensions (≤ 6), are more effective than the 12 higher dimensions of the 51-dimensional Ext. MFCCs. They are also more effective than the same features obtained from a 20-dimensional uniform Gabor filterbank.
  - The features extracted from the IMFs augment the performance of the SV system not only under sufficient test data conditions, but also under limited test data conditions. They are also effective for fast and whispered speaking styles, apart from normal speaking style.

7.2 Contributions

The major experimental contributions of this thesis are as follows :

- Development of a *modified* EMD algorithm, captioned MEMD, which reduces mode-mixing, and is faster than the current noise-assisted EMD methods.
  - Show that the MEMD algorithm reduces mode-mixing present in the EMD algorithm.
  - Show that the MEMD algorithm provides a better distribution of the formants structure of *voiced* speech in its IMFs than the EMD algorithm.
  - Show that the MEMD algorithm is faster than the noise-assisted EMD methods.

- Study the ability of EMD, MEMD, and ICEEMDAN, in extracting the source and system characteristics of the speech signal.

  - Investigate, using synthetic speech signals, the effect of change of *pitch* or *fundamental frequency*, *bandwidths* and *center* or *resonant frequencies* of the vocal tract resonators, on the decomposition of the speech signal.
  - Investigate, using *cepstral* or *homomorphic* analysis, the source and system characteristics represented in the IMFs of natural speech signals corresponding to different *phones* or *speech sounds*.
  - Investigate the capability of the IMFs of speech signals in representing their latent source and system characteristics, when the speech signals are subjected to telephone channel codecs.

- Propose a principle/framework for detecting the GCIs of voiced speech without using short-time LP analysis, and which can provide reliable GCIs estimates under varied conditions.
  - Develop a method for detecting GCIs from the IMFs of speech obtained using ICEEMDAN.
  - Develop a method for detecting GCIs from the IMFs of speech obtained using MEMD.
  - Show that the performances are consistent under clean, noisy, and telephone channel conditions, and comparable with the state-of-the-art methods.

- Show that the adaptive filterbank nature of EMD/MEMD captures speaker specific information that can complement the Mel filterbank.
  - Show that only a small subset of the IMFs are useful in augmenting the performance of text-independent SV system.
  - Show that *cepstral* or *energy-like* features, obtained from a small subset of the IMFs, are more useful than the higher dimensions of the MFCCs in capturing speaker characteristics.
  - Show that a small subset of the IMFs is more useful than a large but fixed *Gabor filterbank*.
  - Show that the IMFs are useful not only for *normal* speech, but also for *fast* and *whispered* speaking styles, and under *insufficient test data* conditions.

7.3 Criticism

No work is perfect. Certainly not this thesis. As such, from a self-reflection perspective, in this section, we would like to discuss some of the aspects of this thesis which may be criticized.

This thesis tries to interlink and balance two research domains - the first part of the thesis primarily deals with non-linear and non-stationary signal analysis, and the second part is focussed on speech processing. The first part led to the development of the MEMD algorithm. The second part led to the development of methods for speech processing applications. It is not unfair to argue that the thesis could have been solely based on either of the two domains. Instead of focussing on EMD alone, we could have researched and experimented with other EMD-like techniques, and could have steered the thesis solely in the direction of time-frequency analysis. Conversely, the thesis could have been focussed solely on one or more speech processing applications, experimenting not only with features, but also on the modeling or machine-learning aspects. Nevertheless, since the thesis interlinks two domains, it may seem to be voluminous or even ad-hoc for some readers, in spite of our best efforts. On a positive note, however, the thesis does not simply apply the latest signal processing tools to churn out better performances. There is a genuine attempt to understand the technique, and then apply it to certain tasks which are likely to be realized based on our understanding of the technique.

The first part of the thesis is based on EMD. It proposes some subtle modifications to the EMD algorithm to develop a better method, captioned MEMD. The basic advantage of MEMD is that it is time-efficient compared to the noise-assisted EMD methods. If the noise-assisted methods were fast enough, there was no requirement of MEMD. Again, EMD (hence its variants) does not have a solid mathematical footing. That is why, in the recent past, EMD-like methods like *Multivariate* EMD [187], *Synchrosqueezed Wavelet Transform* [188], *Empirical Wavelet Transform* [189], *Variational Mode Decomposition* [190] have been developed. These new techniques are attempts to mimic the EMD algorithm, with a better mathematical framework, reduced mode-mixing, and in a time-efficient manner. As such, one may be critical of basing this thesis on EMD (and its variants) alone. However, it is to be noted that in spite of the availability of these new techniques (and even after they were developed), the original EMD algorithm (and its noise-assisted variants) continues to be refined and vastly utilized. The first reason is that EMD is the basis of all the relatively modern techniques. As such, the MEMD algorithm, which highlights the importance of the IPs in the EMD algorithm, could be useful for further development of newer versions of the EMD algorithm, and even EMD-like

methods. Secondly, while EMD, and its noise-assisted variants, lack in mathematical capacity, they are very easy to implement. Even though they have no tunable parameters, their behaviour is quite consistent (as we have seen in Chapter 4). Not having a mathematical framework sometimes can be a blessing in disguise ! Moreover, as we have seen in Chapter 2, analyzing the signal decomposition provides sufficient understanding of the EMD and noise-assisted EMD algorithms, and the lack of mathematical structure is kept aside for the benefits they provide.

The second part of this thesis is based on the application of EMD to certain speech processing tasks. Again, owing to the recently developed EMD-like methods, one may object to the experiments conducted based on EMD alone. One could question - *Would the latest EMD-like techniques provide even better performances ? Would the methodologies proposed work for them ?"*. Our hypothesis is that the experimental findings would be relatable to even the latest techniques, as they are inspired from EMD. Thus, the principle of detecting the GCIs (Chapter 5), or the features derived from the IMFs (Chapter 6), should be relevant for the latest and futuristic algorithms which are based on EMD. However, one is allowed to be critical of such a hypothesis, and only when the experiments are performed we would learn the truth. One thing is certain - our experimental findings would be insightful and helpful for anyone who would like to utilize EMD (hence its variants) or EMD-like techniques for speech processing.

7.4 Directions for future work

The following tasks may be taken up using EMD as an adaptive AM-FM analysis method.

- In Chapter 6, only cepstral-like features were extracted from the IMFs and utilized for the SV task. However, in Chapter 1, it was pointed out that one of the limitations of the MFCCs is that it only captures the energy of the signal, and not its phase. As such, the features extracted from the Hibert spectrum may be explored for the SV task. The Hilbert spectrum not only represents the instantaneous amplitudes, but also the instantaneous frequencies (derived from the instantaneous phases) of the IMFs, and hence might prove useful in the SV task. The Hilbert spectrum may be particularly explored for fast and whispered speech articulations, where the performances of the SV system have a large room for improvement.

- In Chapter 6, the IMFs were utilized for text-independent SV, where there was sufficient amount of data available to model the feature space of any individual speaker. In the event of

inconveniences in data collection, *text-dependent* SV may be preferred, which only requires data of a few seconds, corresponding to a fixed phrase. The utility of the IMFs (and its Hilbert spectrum) for such a system where there is a scarcity of training data may be investigated. Further, in the case of practical deployment of any SV system, it would be difficult to test the system under clean conditions. As such, the utility of the IMFs (and its Hilbert spectrum) under noisy testing conditions may be an interesting study.

- In Chapter 5, the IMFs were utilized to detect the GCIs of the speech signal. However, the proposed methods require an *a priori* F_0 estimate of the speech signal. Henceforth, further exploration may be done to avoid the requirement of F_0 estimate of the speech signal in estimating the GCIs. Conversely, as F_0 estimation and GCIs estimation are both related to the source characteristics of the signal, one may attempt to do both simultaneously using a single process. In other words, the dependency on other techniques, like RAPT, may be eliminated.

- In Chapter 3, the first four IMFs were utilized to show that, cumulatively, they could provide better estimates of the formants compared to the speech signal itself. However, using multiple signals for estimating the formants also resulted in spurious estimates. Henceforth, development of a full-proof algorithm for estimanting the formants could be taken up as future work.

- As discussed in Chapter 2, and in the previous section, the principal demerit of EMD (hence its variants) is the absence of a mathematical theory. As such, development of a mathematical framework for EMD, enabling a better control and understanding of its properties, is an open area of research. More work done in this regard would bring more credibility and acceptability to it, in the research domain, particularly in speech processing.

Bibliography

[1] H. Teager and S. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech production and speech modelling*. Springer, 1990, pp. 241–261.

[2] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*. Springer Science & Business Media, 2008.

[3] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication*, vol. 50, no. 4, pp. 312–322, 2008.

[4] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, 1986.

[5] K. Sreenivasa Rao, S. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *Signal Processing Letters, IEEE*, vol. 14, no. 10, pp. 762–765, 2007.

[6] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 34–43, 2007.

[7] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1602–1613, 2008.

[8] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals." in *Interspeech*, 2009, pp. 2891–2894.

[9] M. R. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 82–91, 2012.

[10] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 994–1006, 2012.

[11] A. Prathosh, T. Ananthapadmanabha, and A. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 2471–2480, 2013.

[12] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception." in *INTERSPEECH*, 2003.

[13] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. Prentice-hall Englewood Cliffs, 1978, vol. 100.

[14] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, no. 1, pp. 1–194, 2007.

[15] R. S. Holambe and M. S. Deshpande, *Advances in Non-Linear Modeling for Speech Processing*. Springer Science & Business Media, 2012.

[16] L. Cohen, *Time-frequency analysis*. Prentice Hall PTR Englewood Cliffs, NJ:, 1995, vol. 1406.

[17] B. Boashash, *Time frequency analysis*. Gulf Professional Publishing, 2003.

[18] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[19] J. F. Kaiser, "Some observations on vocal tract operation from a fluid flow point of view," *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, pp. 358–386, 1983.

[20] S. McLaughlin and P. Maragos, "Nonlinear methods for speech analysis and synthesis," *Advances in nonlinear signal and image processing*, vol. 6, p. 103, 2006.

[21] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 4, pp. 309–319, 1979.

[22] H. A. Murthy, "Algorithms for processing Fourier transform phase of signals," Ph.D. dissertation, PhD dissertation, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, India, 1992.

[23] B. Yegnanarayana, K. Madhu Murthy, and H. A. Murthy, "Applications of group delay functions in speech processing," *J. Inst. Elect. Telecommun. Eng*, vol. 34, pp. 20–29, 1988.

[24] K. M. Murthy and B. Yegnanarayana, "Effectiveness of representation of signals through group delay functions," *Signal Processing*, vol. 17, no. 2, pp. 141–150, 1989.

[25] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[26] L. D. Alsteris and K. K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, no. 6, pp. 727–736, 2006.

[27] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 1. IEEE, 1994, pp. I–137.

[28] X. Lu and J. Dang, "Physiological feature extraction for text independent speaker identification using non-uniform subband processing," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–461.

[29] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard, "Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification," *EURASIP Journal on Advances in Signal Proc.*, vol. 2011, pp. 8:1–8:14, 2011.

[30] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard, "Evolutionary Cepstral Coefficients," *Applied Soft Computing*, vol. 11, no. 4, pp. 3419–3428, 2011.

[31] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 1990, pp. 381–384.

[32] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3245–3265, 1993.

[33] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On separating amplitude from frequency modulations using energy operators," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 2. IEEE, 1992, pp. 1–4.

[34] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *Signal Processing, IEEE Transactions on*, vol. 41, no. 10, pp. 3024–3051, 1993.

[35] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, 1996.

[36] W. J. Hardcastle and A. Marchal, *Speech production and speech modelling*. Springer Science & Business Media, 1990, no. 55.

[37] H. Teager, "Some observations on oral air flow during phonation," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 5, pp. 599–601, 1980.

[38] R. Polikar, "The wavelet tutorial," 1996.

[39] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 325–333, 1995.

[40] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 3, pp. 456–459, 1994.

[41] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: Characterizing individual speakers," in *Proc of SPECOM*, vol. 6, 2006, pp. 431–435.

[42] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 456–465, 1997.

[43] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoustical science and technology*, vol. 26, no. 1, pp. 16–26, 2005.

[44] K. Honda, T. Kitamura, H. Takemoto, S. Adachi, P. Mokhtari, S. Takano, Y. Nota, H. Hirata, I. Fujimoto, Y. Shimada *et al.*, "Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling," *Computer methods in biomechanics and biomedical engineering*, vol. 13, no. 4, pp. 443–453, 2010.

[45] C. Jankowski Jr, T. Quatieri, and D. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 325–328.

[46] B. Yegnanarayana, S. Prasanna, J. M. Zachariah, and C. S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 575–582, 2005.

[47] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1097–1111, 2008.

[48] M. S. Deshpande and R. S. Holambe, "Speaker identification based on robust AM-FM features," in *Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on*. IEEE, 2009, pp. 880–884.

[49] M. S. Deshpande and R. S. Holambe, "AM-FM based robust speaker identification in babble noise," *environments*, vol. 6, no. 10, p. 19, 2011.

[50] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. NewYork: Macmillan Publishing, 1993.

[51] M. Sahidullah and G. Saha, "A novel windowing technique for efficient computation of MFCC for speaker recognition," *Signal Processing Letters, IEEE*, vol. 20, no. 2, pp. 149–152, Feb 2013.

[52] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[53] W. Zheng, M. Xin, X. Wang, and B. Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *Signal Processing Letters, IEEE*, vol. PP, no. 99, pp. 1–1, 2014.

[54] M. Reyes-Vargas, M. Sánchez-Gutiérrez, L. Rufiner, M. Albornoz, L. Vignolo, F. Martínez-Licona, and J. Goddard-Close, "Hierarchical clustering and classification of emotions in human speech using confusion matrices," in *Lecture Notes in Artificial Intelligence*. Springer, 2013, vol. 8113, pp. 162–169.

[55] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[56] C.-L. Huang, S. Matsuda, and C. Hori, "Feature normalization using MVAW processing for spoken language recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, Oct 2013, pp. 1–4.

[57] Z. Qin, W. Liu, and T. Wan, "A bag-of-tones model with MFCC features for musical genre classification," in *Advanced Data Mining and Applications*, ser. Lecture Notes in Computer Science, H. Motoda, Z. Wu, L. Cao, O. Zaiane, M. Yao, and W. Wang, Eds. Springer Berlin Heidelberg, 2013, vol. 8346, pp. 564–575.

[58] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proceedings of the SPECOM*, vol. 1, 2005, pp. 191–194.

[59] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[60] M. Skowronski and J. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1774–1780, Sept 2004.

[61] H. Yeganeh, S. Ahadi, S. Mirrezaie, and A. Ziaei, "Weighting of Mel Sub-bands Based on SNR/Entropy for Robust ASR," in *Signal Processing and Information Technology, 2008. ISSPIT 2008. IEEE International Symposium on*, Dec. 2008, pp. 292–296.

[62] X. Zhou, Y. Fu, M. Liu, M. Hasegawa-Johnson, and T. Huang, "Robust Analysis and Weighting on MFCC Components for Speech Recognition and Speaker Identification," in *Multimedia and Expo, 2007 IEEE International Conference on*, July 2007, pp. 188–191.

[63] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 4625–4628.

[64] Z. Wu and Z. Cao, "Improved MFCC-Based Feature for Robust Speaker Identification," *Tsinghua Science & Technology*, vol. 10, no. 2, pp. 158–161, 2005.

[65] Bŏril, H. and Fousek, P. and Pollák, P., "Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition," in *Proc. of INTERSPEECH 2006 - ICSLP*, Pittsburgh, Pennsylvania, September 2006, pp. 381–384.

[66] B. Zamani, A. Akbari, B. Nasersharif, and A. Jalalvand, "Optimized discriminative transformations for speech features based on minimum classification error," *Pattern Recognition Letters*, vol. 32, no. 7, pp. 948–955, 2011.

[67] L. Burget and H. Heřmanský, "Data Driven Design of Filter Bank for Speech Recognition," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Springer, 2001, pp. 299–304.

[68] T. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 297–302.

[69] T. Bäck, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford, UK: Oxford University Press, 1996.

[70] L. Vignolo, H. Rufiner, D. Milone, and J. Goddard, "Genetic optimization of cepstrum filterbank for phoneme classification," in *Proceedings of the Second International Conference on Bio-inspired Systems and Signal Processing (Biosignals 2009)*. Porto (Portugal): INSTICC Press, 14-17 Enero 2009, pp. 179–185.

[71] C. Charbuillet, B. Gas, M. Chetouani, and J. Zarader, "Optimizing feature complementarity by evolution strategy: Application to automatic speaker verification," *Speech Communication*, vol. 51, no. 9, pp. 724–731, 2009.

[72] H. M. Torres and H. L. Rufiner, "Clasificación de fonemas mediante paquetes de onditas orientadas perceptualmente," in *Anales del 1er Congreso Latinoamericano de Ingeniería Biomédica, Mazatlán 98*, vol. 1, México, November 1998, pp. 163–166. [Online]. Available: http://fich.unl.edu.ar/sinc/sinc-publications/1998/TR98

[73] H. M. Torres and H. L. Rufiner, "Automatic speaker identification by means of Mel cepstrum, wavelets and wavelets packets," in *Proceedings of the Chicago 2000 World Congress IEEE EMBS*, July 2000, paper No. TU–E201–02. [Online]. Available: http://fich.unl.edu.ar/sinc/sinc-publications/2000/TR00

[74] A. Dabin, D. H. Milone, and H. L. Rufiner, "Onditas perceptualmente diseñadas para el reconocimiento automático del habla," in *Proc. 7th Argentine Symposium on Artificial Intelligence*, Rosario, Argentina, 2005, pp. 249–260. [Online]. Available: http://fich.unl.edu.ar/sinc/sinc-publications/2005/DMR05

[75] L. D. Vignolo, D. H. Milone, and H. L. Rufiner, "Genetic wavelet packets for speech recognition," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2350–2359, 2013.

[76] L. Vignolo, H. Rufiner, and D. Milone, "Multi-objective optimisation of wavelet features for phoneme recognition," *IET Signal Processing*, March 2016. [Online]. Available: http://digital-library.theiet.org/content/journals/10.1049/iet-spr.2015.0568

[77] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *Signal Processing Letters, IEEE*, vol. 6, no. 10, pp. 259–261, 1999.

[78] D. Dimitriadis, P. Maragos, and A. Potamianos, "Modulation features for speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–377.

[79] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *Signal Processing Letters, IEEE*, vol. 12, no. 9, pp. 621–624, 2005.

[80] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM signal decomposition with application to speech analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, 2011.

[81] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM-FM separation using auditory-motivated filters," *IEEE transactions on speech and audio processing*, vol. 5, no. 5, pp. 465–480, 1997.

[82] P. Cosi, "Evidence against frame-based analysis techniques," *Proceedings of NATO Advance Institute on Computational Hearing*, pp. 163–168, 1998.

[83] Z.-H. Tan and I. Kraljevski, "Joint variable frame rate and length analysis for speech recognition under adverse conditions," *Computers & Electrical Engineering*, vol. 40, no. 7, pp. 2139–2149, 2014.

[84] C.-S. Jung, K. J. Han, H. Seo, S. S. Narayanan, and H.-G. Kang, "A variable frame length and rate algorithm based on the spectral kurtosis measure for speaker verification," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[85] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1783–1786.

[86] N. E. Huang, "Empirical Mode Decomposition and Hilbert spectral analysis," 1998.

[87] N. E. Huang and S. S. Shen, *Hilbert-Huang transform and its applications*. World Scientific, 2005, vol. 5.

[88] H. Huang and J. Pan, "Speech pitch determination based on Hilbert-Huang transform," *Signal Processing*, vol. 86, no. 4, pp. 792–803, 2006.

[89] G. Schlotthauer, M. Torres, and H. Rufiner, "Voice fundamental frequency extraction algorithm based on Ensemble Empirical Mode Decomposition and entropies," in *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany*. Springer, 2010, pp. 984–987.

[90] N. Chatlani and J. J. Soraghan, "EMD-based filtering (EMDF) of low-frequency noise for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1158–1166, 2012.

[91] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "Pathological voice analysis and classification based on Empirical Mode Decomposition," in *Development of multimodal interfaces: active listening and synchrony*. Springer, 2010, pp. 364–381.

[92] J.-D. Wu and Y.-J. Tsai, "Speaker identification system using Empirical Mode Decomposition and an artificial neural network," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6112–6117, 2011.

[93] K. Khaldi, A.-O. Boudraa, A. Bouchikhi, and M. T.-H. Alouane, "Speech enhancement via EMD," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, no. 1, p. 873204, 2008.

[94] A. Bouchikhi and A.-O. Boudraa, "Multicomponent AM-FM signals analysis based on EMD B-splines ESA," *Signal Processing*, vol. 92, no. 9, pp. 2214–2228, 2012.

[95] K. Khaldi and A.-O. Boudraa, "On signals compression by EMD," *Electronics letters*, vol. 48, no. 21, pp. 1329–1331, 2012.

[96] K. Khaldi and A. Boudraa, "Audio watermarking via EMD," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 675–680, 2013.

[97] K. Khaldi, A.-O. Boudraa, and A. Komaty, "Speech enhancement using Empirical Mode Decomposition and the Teager-Kaiser energy operator," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 451–459, 2014.

[98] K. Khaldi, A.-O. Boudraa, B. Torresani, and T. Chonavel, "HHT-based audio coding," *Signal, image and video processing*, vol. 9, no. 1, pp. 107–115, 2015.

[99] K. Khaldi, A.-O. Boudraa, and M. Turki, "Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement," *IET Signal Processing*, vol. 10, no. 1, pp. 69–80, 2016.

[100] J.-C. Cexus and A.-O. Boudraa, "Nonstationary signals analysis by Teager-Huang Transform (THT)," in *Signal Processing Conference, 2006 14th European*. IEEE, 2006, pp. 1–5.

[101] K. Khaldi, M. T.-H. Alouane, and A.-O. Boudraa, "A new EMD denoising approach dedicated to voiced speech signals," in *Signals, Circuits and Systems, 2008. SCS 2008. 2nd International Conference on*. IEEE, 2008, pp. 1–5.

[102] K. Khaldi, A.-O. Boudraa, M. Turki, T. Chonavel, and I. Samaali, "Audio encoding based on the Empirical Mode Decomposition," in *Signal Processing Conference, 2009 17th European*. IEEE, 2009, pp. 924–928.

[103] K. Khaldi, A.-O. Boudraa, B. Torresani, T. Chonavel, and M. Turki, "Audio encoding using Huang and Hilbert transforms," in *Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on*. IEEE, 2010, pp. 1–5.

[104] R. Sharma, K. Ramesh, and S. Prasanna, "Analysis of electroglottograph signal using ensemble Empirical Mode Decomposition," in *India Conference (INDICON), 2014 Annual IEEE*. IEEE, 2014, pp. 1–6.

[105] M. A. Colominas, G. Schlotthauer, and M. E. Torres, "Improved complete ensemble EMD: A suitable tool for biomedical signal processing," *Biomedical Signal Processing and Control*, vol. 14, pp. 19–29, 2014.

[106] R. Sharma and S. R. M. Prasanna, "Characterizing glottal activity from speech using Empirical Mode Decomposition," in *National Conference on Communications 2015 (NCC-2015)*, Mumbai, India, Feb. 2015.

[107] P. Flandrin, "Some aspects of Huang's Empirical Mode Decomposition, from interpretation to applications," in *Int. Conf. Computat. Harmonic Anal. CHA*, vol. 4, 2004.

[108] M. A. Colominas, G. Schlotthauer, and M. E. Torres, "An unconstrained optimization approach to Empirical Mode Decomposition," *Digital Signal Processing*, vol. 40, pp. 164–175, 2015.

[109] Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the Empirical Mode Decomposition method," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 460, no. 2046, pp. 1597–1611, 2004.

[110] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical Mode Decomposition as a filter bank," *Signal Processing Letters, IEEE*, vol. 11, no. 2, pp. 112–114, 2004.

[111] P. Flandrin and P. Goncalves, "Empirical Mode Decompositions as data-driven wavelet-like expansions," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 04, pp. 477–496, 2004.

[112] G. Rilling, P. Flandrin, and P. Gonçalves, "Empirical mode decomposition, fractional gaussian noise and Hurst exponent estimation." in *ICASSP (4)*, 2005, pp. 489–492.

[113] P. Flandrin, P. Gonçalves, and G. Rilling, "EMD equivalent filter banks, from interpretation to applications," *Hilbert-Huang transform and its applications*, pp. 57–74, 2005.

[114] G. Rilling and P. Flandrin, "on the influence of sampling on the Empirical Mode Decomposition." in *ICASSP (3)*, 2006, pp. 444–447.

[115] G. Rilling, "Décompositions Modales Empiriques," Ph.D. dissertation, PhD thesis, Ecole normale supérieure de Lyon, 2007.

[116] G. Rilling, P. Flandrin, P. Goncalves *et al.*, "On Empirical Mode Decomposition and its algorithms," in *IEEE-EURASIP workshop on nonlinear signal and image processing*, vol. 3. NSIP-03, Grado (I), 2003, pp. 8–11.

[117] N. E. Huang, M.-L. C. Wu, S. R. Long, S. S. Shen, W. Qu, P. Gloersen, and K. L. Fan, "A confidence limit for the Empirical Mode Decomposition and Hilbert spectral analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 459, no. 2037, pp. 2317–2345, 2003.

[118] G. Wang, X.-Y. CHEN, F.-L. Qiao, Z. Wu, and N. E. Huang, "On intrinsic mode function," *Advances in Adaptive Data Analysis*, vol. 2, no. 03, pp. 277–293, 2010.

[119] Z. Wu and N. E. Huang, "Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method," *Advances in adaptive data analysis*, vol. 1, no. 01, pp. 1–41, 2009.

[120] G. Rilling and P. Flandrin, "One or two frequencies? the Empirical Mode Decomposition answers," *Signal Processing, IEEE Transactions on*, vol. 56, no. 1, pp. 85–95, 2008.

[121] A. Bouzid and N. Ellouze, "Empirical Mode Decomposition of voiced speech signal," in *Control, Communications and Signal Processing, 2004. First International Symposium on*. IEEE, 2004, pp. 603–606.

[122] A. Bouzid and N. Ellouze, "Voiced speech analysis by Empirical Mode Decomposition," in *Advances in Nonlinear Speech Processing*. Springer, 2007, pp. 213–220.

[123] Y. Chen and M. Q. Feng, "A technique to improve the Empirical Mode Decomposition in the Hilbert-Huang transform," *Earthquake Engineering and Engineering Vibration*, vol. 2, no. 1, pp. 75–85, 2003.

[124] R. Deering and J. F. Kaiser, "The use of a masking signal to improve Empirical Mode Decomposition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 4. IEEE, 2005, pp. iv–485.

[125] Y. Kopsinis and S. McLaughlin, "Improved EMD using doubly-iterative sifting and high order spline interpolation," *EURASIP Journal on Advances in Signal processing*, vol. 2008, p. 120, 2008.

[126] J.-R. Yeh, J.-S. Shieh, and N. E. Huang, "Complementary Ensemble Empirical Mode Decomposition: A novel noise enhanced data analysis method," *Advances in Adaptive Data Analysis*, vol. 2, no. 02, pp. 135–156, 2010.

[127] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4144–4147.

[128] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[129] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[130] "http://perso.ens-lyon.fr/patrick.flandrin/EMD.html." [Online]. Available: http://perso.ens-lyon.fr/patrick.flandrin/EMD.html

[131] "http://www.bioingenieria.edu.ar/grupos/ldnlys/index.htm." [Online]. Available: http://www.bioingenieria.edu.ar/grupos/ldnlys/index.htm

[132] D. Wang, D. Miao, and C. Xie, "Best basis-based wavelet packet entropy feature extraction and hierarchical eeg classification for epileptic detection," *Expert Systems with Applications*, vol. 38, no. 11, pp. 14 314 – 14 320, 2011.

[133] A. R. Ferreira da Silva, "Approximations with evolutionary pursuit," *Signal Processing*, vol. 83, no. 3, pp. 465–481, 2003.

[134] R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, 1992.

[135] N. Saito and R. Coifman, "Local discriminant bases and their applications," *Journal of Mathematical Imaging and Vision*, vol. 5, no. 4, pp. 337–358, 1995.

[136] H. Rufiner and J. Goddard, "A method of wavelet selection in phoneme recognition," in *Proceedings of the 40th Midwest Symposium on Circuits and Systems*, vol. 2, Aug 1997, pp. 889–891.

[137] N. E. Saeedi and F. Almasganj, "Wavelet adaptation for automatic voice disorders sorting," *Computers in Biology and Medicine*, vol. 43, no. 6, pp. 699 – 704, 2013.

[138] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 474 – 485, 2007.

[139] P. Flandrin, P. Gonçalves, G. Rilling *et al.*, *Detrending and denoising with Empirical Mode Decompositions*. Citeseer, 2004.

[140] A.-O. Boudraa, J.-C. Cexus *et al.*, "Denoising via Empirical Mode Decomposition," *Proc. IEEE ISCCSP*, vol. 4, 2006.

[141] Y. Kopsinis and S. McLaughlin, "Development of EMD-based denoising methods inspired by wavelet thresholding," *Signal Processing, IEEE Transactions on*, vol. 57, no. 4, pp. 1351–1362, 2009.

[142] T. Hasan and M. K. Hasan, "Suppression of residual noise from speech signals using Empirical Mode Decomposition," *Signal Processing Letters, IEEE*, vol. 16, no. 1, pp. 2–5, 2009.

[143] G. Tsolis and T. D. Xenos, "Signal denoising using Empirical Mode Decomposition and higher order statistics," *Int J Signal Proc Image Proc Pattern Recog*, vol. 4, pp. 91–106, 2011.

[144] I. Hadhami and A. Bouzid, "Speech denoising based on Empirical Mode Decomposition and improved thresholding," in *Advances in Nonlinear Speech Processing*. Springer, 2013, pp. 200–207.

[145] Z. Yang, D. Huang, and L. Yang, "A novel pitch period detection algorithm based on Hilbert-Huang transform," in *Advances in Biometric Person Authentication*. Springer, 2005, pp. 586–593.

[146] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "A new algorithm for instantaneous F0 speech extraction based on Ensemble Empirical Mode Decomposition," in *Proc. of 17th Eur. Sign. Proces. Conf*, 2009, pp. 2347–2351.

[147] H. Huang and X.-x. Chen, "Speech formant frequency estimation based on Hilbert-Huang transform," *JOURNAL-ZHEJIANG UNIVERSITY ENGINEERING SCIENCE*, vol. 40, no. 11, p. 1926, 2006.

[148] M. Kaleem, B. Ghoraani, A. Guergachi, and S. Krishnan, "Pathological speech signal analysis and classification using Empirical Mode Decomposition," *Medical & biological engineering & computing*, vol. 51, no. 7, pp. 811–821, 2013.

[149] B. Mijovic, M. Silva, B. Van den Bergh, K. Allegaert, J.-M. Aerts, D. Berckmans, S. Van Huffel *et al.*, "Assessment of pain expression in infant cry signals using Empirical Mode Decomposition," *Methods Inf Med*, vol. 49, no. 5, pp. 448–452, 2010.

[150] D. Jhanwar, K. K. Sharma, and S. Modani, "Classification of environmental background noise sources using Hilbert-Huang transform," *International Journal of Signal Processing Systems Vol*, vol. 1, 2013.

[151] B. K. Khonglah, R. Sharma, and S. Mahadeva Prasanna, "Speech vs music discrimination using Empirical Mode Decomposition," in *Communications (NCC), 2015 Twenty First National Conference on*. IEEE, 2015, pp. 1–6.

[152] X. Li and X. Li, "Speech emotion recognition using novel HHT-TEO based features," *Journal of Computers*, vol. 6, no. 5, pp. 989–998, 2011.

[153] L. He, M. Lech, N. C. Maddage, and N. B. Allen, "Study of Empirical Mode Decomposition and spectral analysis for stress and emotion classification in natural speech," *Biomedical Signal Processing and Control*, vol. 6, no. 2, pp. 139–146, 2011.

[154] T. Hasan and J. H. Hansen, "Robust speaker recognition in non-stationary room environments based on Empirical Mode Decomposition." in *INTERSPEECH*, 2011, pp. 2733–2736.

[155] M. K. I. Molla, K. Hirose, and N. Minematsu, "Robust voiced/unvoiced speech classification using Empirical Mode Decomposition and periodic correlation model." in *INTERSPEECH*, 2008, pp. 2530–2533.

[156] Z. Lu, B. Liu, and L. Shen, "Speech endpoint detection in strong noisy environment based on the Hilbert-Huang transform," in *Mechatronics and Automation, 2009. ICMA 2009. International Conference on*. IEEE, 2009, pp. 4322–4326.

[157] M. K. Islam Molla, S. Das, M. E. Hamid, and K. Hirose, "Empirical mode decomposition for advanced speech signal processing," *Journal of Signal Processing*, vol. 17, no. 6, pp. 215–229, 2013.

[158] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[159] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[160] B. Challenge, "The blizzard challenge 2009," 2012.

[161] R. Sharma and S. M. Prasanna, "A better decomposition of speech obtained using modified Empirical Mode Decomposition," *Digital Signal Processing*, vol. 58, pp. 26 – 39, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200416300975

[162] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[163] R. Sharma, L. Vignolo, G. Schlotthauer, M. Colominas, H. L. Rufiner, and S. Prasanna, "Empirical Mode Decomposition for adaptive AM-FM analysis of speech: A review," *Speech Communication*, vol. 88, pp. 39 – 64, 2017. [Online]. Available: //www.sciencedirect.com/science/article/pii/S0167639316302370

[164] "http://www.commsp.ee.ic.ac.uk/ sap/resources/aplawdw/." [Online]. Available: http://www.commsp.ee.ic.ac.uk/~sap/resources/aplawdw/

[165] M. Brookes, "Voicebox," *Speech Processing Toolbox for Matlab, Department of Electrical & Electronic Engineering, Imperial College*, 2009.

[166] H. Beigi, *Speaker Recognition: Advancements and Challenges*. INTECH Open Access Publisher, 2012.

[167] A. Neustein and H. A. Patil, *Forensic speaker recognition: Law enforcement and counter-terrorism*. Springer Science & Business Media, 2011.

[168] S. R. M. Prasanna, C. Gupta, and B. Yegananarayana, "Extraction of speaker specific information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

[169] R. K. Das and S. Mahadeva Prasanna, "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America*, vol. 140, no. 1, pp. 184–190, 2016.

[170] A. Fazel and S. Chakrabartty, "An overview of statistical pattern recognition techniques for speaker verification," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 62–81, 2011.

[171] D. Petrovska-Delacrétaz, A. El Hannani, and G. Chollet, "Text-independent speaker verification: state of the art and challenges," in *Progress in nonlinear speech processing*. Springer, 2007, pp. 135–169.

[172] Q. Jin and T. F. Zheng, "Overview of front-end features for robust speaker recognition," *Proc. APSIPA*, 2011.

[173] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[174] *"The NIST Year 2003 Speaker Recognition Evaluation Plan", NIST , Feb 2003.*

[175] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788 –798, May 2011.

[176] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, 2006.

[177] Q. Jin, S.-C. S. Jou, and T. Schultz, "Whispering speaker identification," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 1027–1030.

[178] X. Fan and J. H. Hansen, "Speaker identification within whispered speech audio streams," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1408–1421, 2011.

[179] X. Fan and J. H. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4553–4556.

[180] M. Sarria-Paja, T. H. Falk, and D. O'Shaughnessy, "Whispered speaker verification and gender detection using weighted instantaneous frequencies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7209–7213.

[181] Rohan Kumar Das, Abhiram B., S. R. M. Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *Interspeech 2014, Singapore*, 2014, pp. 1836–1840.

[182] Rohan Kumar Das, Debadatta Pati, and S. R. M. Prasanna, "Different aspects of source information for limited data speaker verification," in *National Conference on Communications (NCC) 2015*, 2015.

[183] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[184] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005.

[185] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. of ICSLP*, 2006, p. 14711474.

[186] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr 1981.

[187] N. Rehman and D. P. Mandic, "Multivariate Empirical Mode Decomposition," in *Proceedings of The Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, 2009, p. rspa20090502.

[188] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed wavelet transforms: An Empirical Mode Decomposition-like tool," *Applied and computational harmonic analysis*, vol. 30, no. 2, pp. 243–261, 2011.

[189] J. Gilles, "Empirical Wavelet Transform," *IEEE transactions on signal processing*, vol. 61, no. 16, pp. 3999–4010, 2013.

[190] K. Dragomiretskiy and D. Zosso, "Variational Mode Decomposition," *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2014.

[191] R. Sharma, S. Prasanna, R. K. Bhukya, and R. K. Das, "Analysis of the Intrinsic Mode Functions for speaker information," *Speech Communication, Elsevier*, April 2017.

[192] R. Sharma, S. Prasanna, H. L. Rufiner, and G. Schlotthauer, "Detection of the Glottal Closure Instants using Empirical Mode Decomposition," *Circuits, Systems, and Signal Processing, Springer, under review*, August 2017.

[193] R. Sharma and S. Prasanna, "Analysis of the source and system characteristics in the IMFs," *Circuits, Systems and Signal Processing, Springer, submitted to*, October 2017.

[194] R. Sharma, R. K. Bhukya, and S. Prasanna, "Analysis of the Hilbert spectrum for text-dependent speaker verification," *Speech Communication, Elsevier, under review*, April 2017.

List of Publications

Publications composing the thesis

J : [163] R. Sharma, L. Vignolo, G. Schlotthauer, M.A. Colominas, H.L. Rufiner and S.R.M. Prasanna "Empirical Mode Decomposition for adaptive AM-FM analysis of Speech : A Review,", Speech Communication, Elsevier, December 2016.

J : [161] R. Sharma and S.R.M. Prasanna, "A better decomposition of speech obtained using modified Empirical Mode Decomposition," Digital Signal Processing, Elsevier, July 2016.

J : [191] R. Sharma, S.R.M. Prasanna, R.K. Bhukya, and R.K. Das, "Analysis of the Intrinsic Mode Functions for Speaker Information,", Speech Communication, Elsevier, April 2017.

J : [192] R. Sharma, S.R.M. Prasanna, H. L. Rufiner, and G. Schlotthauer, "Detection of the Glottal Closure Instants using Empirical Mode Decomposition," **under review in**, Circuits, Systems, and Signal Processing, Springer, August 2017.

J : [193] R. Sharma and S.R.M. Prasanna, "Analysis of the Source and System characteristics in the IMFs,", **submitted to**, Circuits, Systems, and Signal Processing, Springer, October 2017.

C : [106] R. Sharma and S. R. M. Prasanna, "Characterizing glottal activity from speech using empirical mode decomposition," in NCC-2015, Mumbai, India, Feb. 2015.

Publications other than the thesis

J : [194] R. Sharma, R.K. Bhukya and S. R. M. Prasanna, "Analysis of the Hilbert spectrum for Text-Dependent Speaker Verification," **under review in**, Speech Communication, Elsevier, April 2017.

C : [104] R. Sharma, K. Ramesh, and S.R.M. Prasanna, "Analysis of ElectroGlottoGraph signal using Ensemble Empirical Mode Decomposition," in INDICON, IEEE, 2014.

C : [151] B. K. Khonglah, R. Sharma, and S.R.M. Prasanna, "Speech vs music discrimination using Empirical Mode Decomposition," in NCC-2015, Mumbai, India, Feb. 2015.

⋆ **J** implies Journal publication ; **C** implies Conference publication