SYNOPSIS REPORT ON

APPROACHES FOR ROBUST TEXT-DEPENDENT SPEAKER VERIFICATION UNDER DEGRADED CONDITIONS

A

Thesis submitted by

Ramesh Kumar Bhukya

for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

APRIL, 2019

1 Organization and Outline of the thesis

The objective of this thesis work is to develop a robust *text-dependent speaker verification* (TDSV) system [1] by exploring different techniques to obtain better performance under clean and degraded conditions. To achieve this, three different approaches are proposed. First, using combined temporal and spectral enhancement method, the speech regions embedded in background noise are enhanced. Further, using enhanced speech signal end points are detected for the development of TDSV system [2]. Second approach involves the exploration of features extracted from *Hilbert Spectrum* (HS) of the *Intrinsic Mode Functions* (IMFs) obtained from *Modified Empirical Mode Decomposition* (MEMD) of the speech signal. It is noticed that only first few IMFs are useful and necessary for characterizing speaker information and achieving enhanced performance in combination with *Mel frequency cepstral coefficients* (MFCCs) [3]. The third one utilizes the speech-specific knowledge for the robust end point detection, which is suitable for both clean and degraded conditions [4, 5]. Finally, a combined system is developed in a sequential manner, where robust end point detection is performed on the enhanced speech and then HS of the IMFs obtained from MEMD features are extracted from the regions between the detected end points. The combined method significantly improves the system performance.

First chapter of the thesis - is dedicated on the development of a system for authenticating a person from speech utterance. The chapter discusses the modular representation of the work, motivation behind the implementation and some issues present in the practical deployment [6]. Finally, the chapter ends with a brief description of the organization of thesis.

Second chapter of the thesis - presents a review based on possible directions from the perspective of deployment of the system. Based on the major issues addressed in **first chapter**, the literature review has been carried out in three different directions. These are (a) end point detection, (b) speech enhancement, and (c) exploration of alternative/ complementary features for the improvement of TDSV under degraded conditions. Further, based on the review of different approaches, the organization of the work is explained.

Third chapter of the thesis - aims to *"use enhanced speech for the detection of end points, under degraded conditions"*. There are different combination of experimental studies conducted for the speech enhancement under degraded conditions [7]. Among these, end points detected from the speech signal enhanced by the combination of temporal and spectral enhancement techniques gives an improvement in the performance of TDSV system.

The fourth chapter of the thesis investigates the utility of HS of the speech signal, obtained using MEMD for improving the performance of TDSV system. Features extracted from the instantaneous fre-

quencies and energies of IMFs obtained from MEMD [8]. These extracted features are tested individually and in conjunction with the MFCCs for TDSV task, using *dynamic time warping* (DTW) technique [3].

Fifth chapter of the thesis - attempts to detect begin and end points using some speech specific information for the removal of non-overlapping speech as well as non-speech background degradation. Here, vowel-like regions (VLRs), dominant resonant frequency (DRF), and foreground speech segmentation (FSS), glottal activity detection (GAD), dominant aperiodic regions detection (OBS), and speech duration knowledge (SDK) are used to detect begin and end points accurately. Further, using derived begin and end points, TDSV performance is evaluated using DTW technique.

Sixth chapter of the thesis - combines several noise robust methods explored in different stages of the TDSV system. First the speech signal is enhanced using temporal and spectral enhancement technique. On the enhanced speech, robust end point detection algorithm using speech specific knowledge is applied. From the detected speech regions, features, namely HS of the IMFs, obtained from MEMD and conventional MFCC features are extracted for the development of TDSV system.

Seventh chapter of the thesis- summarizes the work, highlights some of the conclusions made and also points to some future directions of the work.

2 Processing Degraded Speech for Text-Dependent Speaker Verification

Detection of begin and end points of speech are crucial for TDSV system. Energy-based end point detection methods fails under degraded conditions [9]. In this work, first the enhancement of degraded speech is carried out. On the enhnaced speech, the end points are detected. Different combinations of enhancement techniques are experimented to detect end points accurately. Using segmented speech, DTW based TDSV experiments are carried out. Results in Table 1 indicate that among various techniques, the combination of temporal and spectral processing methods followed by end point detection gives improved performance. The performance of TD-2 and TD-3 utterances are shown in Figure 1

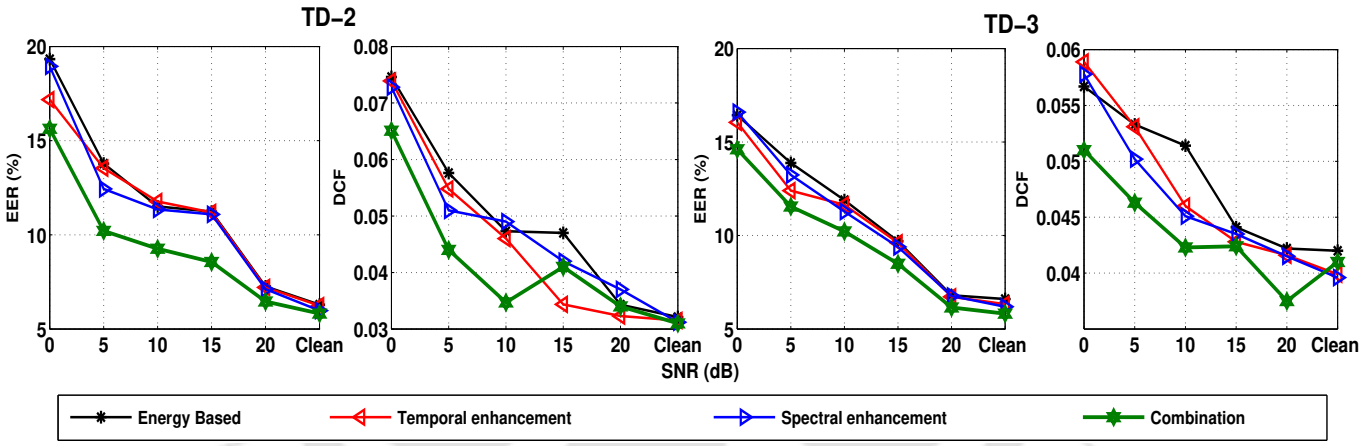
3 Analysis of the Hilbert Spectrum for Speaker Verification

This part of the work explores utility of HS of speech signal, constructed from its IMFs, in characterizing speaker information. The IMFs of speech signal are obtained using a non-linear and non-stationary data analysis technique called MEMD [10]. The HS, which is a representation of the instantaneous frequencies and instantaneous energies of the IMFs, is processed in short time-segments to generate features. The

Table 1: Performance of TDSV systems in terms of EER and DCF using RSR2015 database using different methods.

Method –>	TD-1	Energy Based		Temporal enhancement		Spectral enhancement		Combination	
Database	SNR	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
RSR2015	Clean	7.59	0.0388	7.4	0.0385	7.37	0.0382	6.62	0.0379
	20 dB	8.1	0.0394	8.06	0.0393	7.77	0.039	6.95	0.0382
	15 dB	9.26	0.0404	8.65	0.0398	8.54	0.0395	7.58	0.0389
	10 dB	11.23	0.042	11.13	0.0418	10.48	0.0413	10.27	0.0409
	5 dB	12.51	0.0431	12.47	0.0429	12.37	0.0427	11.91	0.0425
	0 dB	20.29	0.0523	20.76	0.0496	19.47	0.0485	18.95	0.0481

Figure 1: Summary of TD-2 and TD-3 test trials, DTW based TDSV systems performance in terms of the EER and DCF for different experimental setup on RSR2015 database.



experimental results are validated using two databases - the RSR2015 and the IITG. The performance of the TDSV system is evaluated for the HS-based features and their combinations with the 39-dimensional MFCCs. Table 2 represents the performance of DTW-based TDSV for HS-based features and their combination with MFCC, under clean and degraded conditions. Also, it is found that the features extracted from the HS are found to be consistently more effective than *cepstral/ energy-like* feature obtained from the raw IMFs under noisy conditions. The performance metrics evaluated for the TD-2 and TD-3 sentences are shown in Figure 2.

4 Significance of Speech-Specific Knowledge for End Point Detection

This chapter proposes a method using some speech specific knowledge to detect the begin and end points of speech under degraded condition [11]. The method is based on VLRs detected using the combination of excitation source and vocal tract system information. Strength of excitation derived using zero-frequency

Table 2: Performances of the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-1 sentence of the RSR2015 database. The testing utterances are clean as well as corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the MEMD features are kept constant, with $K = 4$.

Technique →		Mel filterbank		MEMD filterbank	MEMD - Hilbert Spectrum						
Feature →		MFCCs	Ext. MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs	MFCCs
SNR	Metric	(39)	(51)	+ cG_4	+ F_4	+ E_4	+ σF_3	+ σE_3	+ ΔF_4	+ ΔE_4	+ Υ_4
Clean	EER (%)	7.59	6.22	**5.03**	6.57	**6.62**	6.91	**6.39**	7.28	**6.71**	**7.2**
	DCF	0.0388	0.0378	**0.0368**	0.0381	**0.0379**	0.0381	**0.0377**	0.0386	**0.038**	**0.0386**
20 dB	EER (%)	8.1	11.91	6.55	6.71	6.67	6.98	**6.48**	7.34	7.4	7.84
	DCF	0.0394	0.0425	0.038	0.0382	0.0379	0.0382	**0.0378**	0.0387	0.0385	0.0391
15 dB	EER (%)	9.26	12.47	11.13	7.77	**6.95**	7.64	7	7.37	7.56	8.06
	DCF	0.0404	0.0429	0.0418	0.039	**0.0382**	0.0387	0.0382	0.0387	0.0387	0.0393
10 dB	EER (%)	11.23	12.9	13.22	13.23	10.31	10.27	10.29	**9.89**	10.34	10.48
	DCF	0.042	0.0433	0.0436	0.0436	0.0409	0.0409	0.0409	**0.0408**	0.041	0.0413
5 dB	EER (%)	12.51	20.93	14.18	14.35	12.6	**12.37**	12.59	13.31	13.51	13.14
	DCF	0.0431	0.0499	0.0443	0.0445	0.0428	**0.0427**	0.0428	0.0436	0.0436	0.0435
0 dB	EER (%)	20.29	23.64	21.04	21.6	20.42	20.76	**20.1**	22.17	20.68	22.03
	DCF	0.0493	0.0522	0.05	0.0505	0.0493	0.0496	**0.049**	0.051	0.0495	0.0508

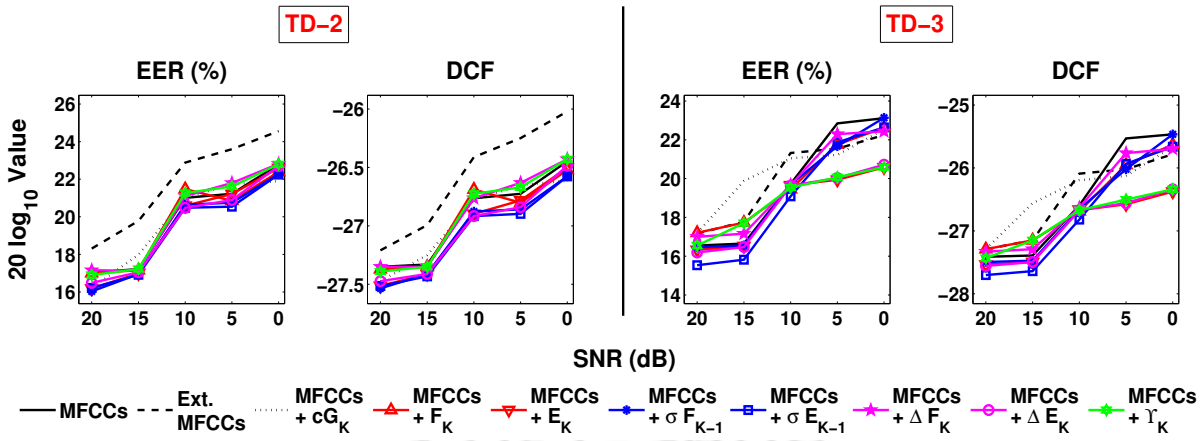


Figure 2: Performances of the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the cG_K seven features are kept constant, with $K = 4$.

filtered signal is used as excitation source information. Whereas, vocal tract system information is obtained from DRF computed from Hilbert envelope of numerator of group delay function (HNGD) [12].

Further, FSS using excitation and vocal tract system information is carried out to remove spurious VLRs in the background speech region [13, 14]. Better localization of the end points is done using more detailed information about excitation source in terms of GAD to detect the sonorant consonants and missed

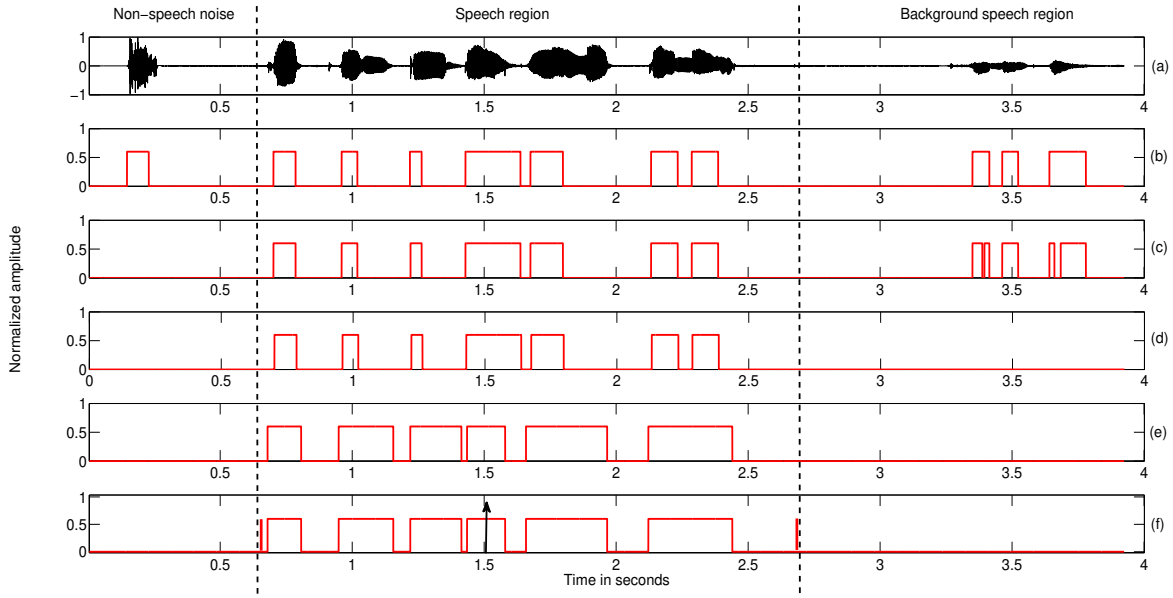


Figure 3: Illustration of the begin and end point detection procedure. (a) Speech signal with non-speech noise and background speech. (b) Detected VLRs. (c) VLRs after removing the non-speech noise using DRF information. (d) VLRs after removing the speech background using foreground speech segmentation. (e) Detected glottal activity regions added to the VLRs. (f) Refined begin and end point using obstruent information. The arrow around 1.5 s shows the center C of the speech utterance. Duration between successive VLRs are less than 300 ms. Therefore, no further modification is made using SDK knowledge. Dotted line shows ground truth manual marking.

Table 3: Performances of the TDSV systems using different end point detection methods evaluated on the TD-1 phrase of the RSR2015 database under degraded speech condition. The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF.

Technique / SNR	Metrics	Energy	FSS	GAD	VLR	VLR+DRF	VLR+DRF +FSS	VLR+DRF +GAD	VLR+DRF +GAD+FSS	VLR+DRF+GAD +FSS+OBS	VLR+DRF+GAD +FSS+OBS+SDK
Clean	EER (%)	7.59	7.3	7.28	6.79	7.59	7.3	6.96	6.2	5.51	**5.37**
	DCF	0.0388	0.0353	0.0406	0.0361	0.0347	0.0353	0.0407	0.0386	0.038	**0.0341**
20 dB	EER (%)	8.17	7.64	7.3	7.64	8.84	7.37	7.08	8.37	6.96	**6.2**
	DCF	0.0394	0.0411	0.0353	0.0411	0.0381	0.0341	**0.033**	0.0341	0.0407	0.0365
15 dB	EER (%)	9.64	8.45	10.37	10.75	11.36	10.85	8.84	9.59	11.55	**7.64**
	DCF	0.0387	0.034	0.0425	0.066	0.0428	0.0428	0.0381	**0.0337**	0.0428	0.0412
10 dB	EER (%)	11.13	12.43	10.55	12.43	11.55	11.55	11.48	12.6	13.36	**9.49**
	DCF	0.0448	0.0522	0.0604	0.0522	0.0735	0.0649	0.0653	0.0648	0.054	**0.0402**
5 dB	EER (%)	16.57	17.15	16.15	15.99	16.57	16.18	17.89	17.68	17.68	**13.37**
	DCF	0.0688	0.0657	0.0671	0.0646	0.0647	0.0707	0.0725	0.0669	0.0648	**0.0509**
0 dB	EER (%)	20.29	17.94	19.84	16.18	17.17	21.64	17.94	17.73	17.73	**16.47**
	DCF	0.0693	0.0684	0.0694	0.0826	0.0758	0.0781	0.084	0.0948	0.0669	**0.0577**

VLRs [15, 16]. To include unvoiced consonants, obstruent region detection is done at the beginning of the first VLR and the end of last VLR [16]. Different steps involved in the end point detection algorithm are depicted in Figure 3. Performance of TDSV experiments using proposed end point detection method is shown in Table 3. Figure 4 shows the performance of the TDSV systems using different EPD methods evaluated on the TD-2 and TD-3 speech utterances.

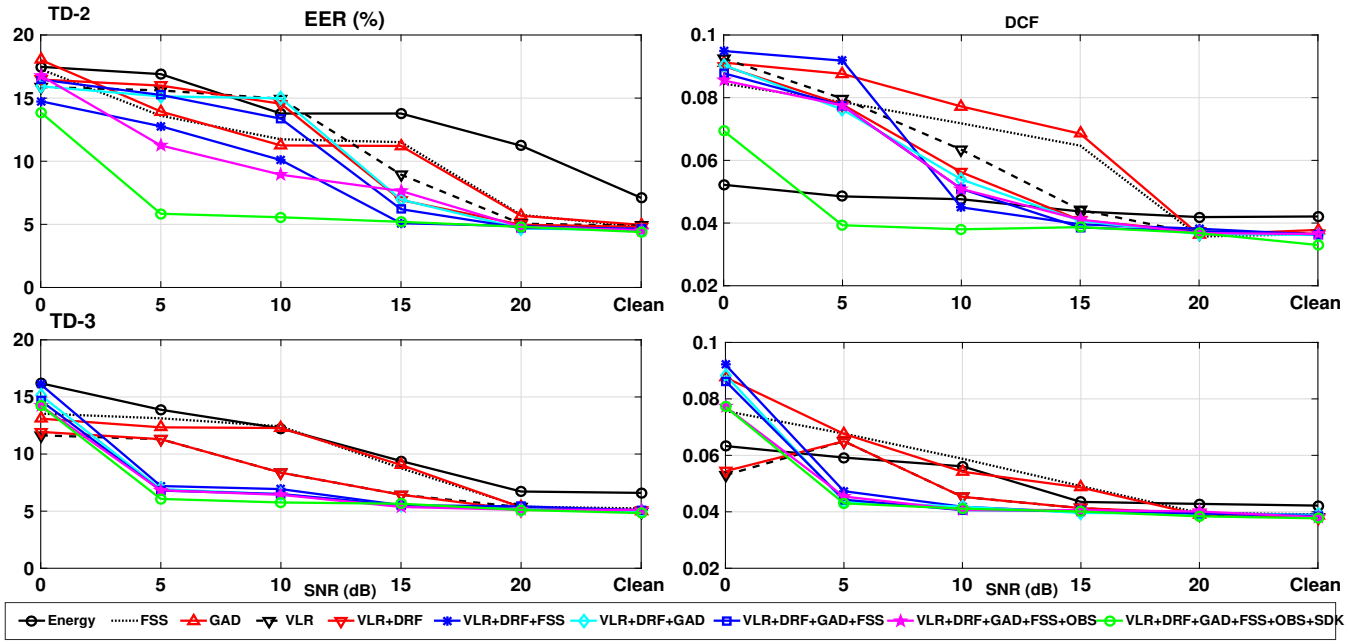


Figure 4: Performances of the TDSV systems using different end point detection methods evaluated on the TD-2 and TD-3 phrases of the RSR2015 database.The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB in steps of 5 dB. The TDSV systems is built using 39-dimensional MFCCs and DTW, and performance is evaluated in terms of EER and DCF.

5 Robust Text Dependent Speaker Verification

In this work, we explore various noise robust techniques at different stages of a TDSV system [17]. Speech-specific knowledge-based technique is used for robust end point detection under degraded conditions [4]. We also explore a combined temporal and spectral speech enhancement technique prior to the end points detection for enhancing speech regions embedded in noise [18]. From the segmented speech, HS of the IMFs obtained from MEMD and MFCCs are extracted. Using these features TDSV system is developed. Further, results showed that the combination of speech enhancement, followed by speech-specific knowledge-based end point detection and augmentation of HS-based features with MFCCs, shows significant improvement in the performance of TDSV system. Table 4 summarizes the results obtained from energy-based and speech-specific knowledge based end point detection approaches. The performance of the system and the performance metrics are evaluated for the TD-2 and TD-3 utterances are shown in Figure 5. It is also found from Table 5 that the use of speech enhancement prior to signal and feature level compensation results in further improvement in performance for the low SNR cases. The performance metrics evaluated for the TD-2 and TD-3 utterances are shown in Figure 6.

Table 4: The performance of the TDSV system using robust end point detection followed by MEMD feature extraction. The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The testing utterances are corrupted with Babble noise, with SNR varying from 20-0 dB, in steps of 5 dB. The dimensions of the MEMD features are kept constant, with $K = 4$. The numbers shown in brackets are obtained by using energy-based VAD and without brackets are obtained by using speech specific knowledge based robust end point detection method.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs	Ext. MFCCs	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +
SNR	Metric	39 dim	51 dim	cG_4	F_4	E_4	σF_3	σE_3	ΔF_4	ΔE_4	Υ_4
20 dB	EER (%)	6.81 (8.1)	6.55 (11.91)	6.46 (6.55)	6.33 (6.71)	6.44 (6.67)	6.5 (6.98)	6.92 (6.48)	6.3 (7.34)	6.76 (7.4)	*6.25* (7.84)
	mDCF	0.0442 (0.0394)	*0.0332* (0.0425)	0.0415 (0.038)	0.0444 (0.0382)	0.0447 (0.0379)	0.0452 (0.0382)	0.0474 (0.0378)	0.0453 (0.0387)	0.0453 (0.0385)	0.0429 (0.0391)
15 dB	EER (%)	7.23 (9.26)	6.9 (12.47)	6.93 (11.13)	7.12 (7.77)	*5.94* (6.95)	7.49 (7.64)	6.08 (7)	7.34 (7.37)	7.37 (7.56)	6.63 (8.06)
	mDCF	0.0464 (0.0404)	0.0453 (0.0429)	*0.0408* (0.0418)	0.0483 (0.039)	0.0439 (0.0382)	0.0472 (0.0387)	0.0462 (0.0382)	0.0465 (0.0387)	0.0463 (0.0387)	0.0434 (0.0393)
10 dB	EER (%)	7.5 (11.23)	7.38 (12.9)	11.23 (13.22)	7.26 (13.23)	*6.89* (10.31)	8.1 (10.27)	7.35 (10.29)	8.32 (9.89)	7.13 (10.34)	7.35 (10.48)
	mDCF	0.0488 (0.042)	0.046 (0.0433)	*0.0418* (0.0436)	0.0476 (0.0436)	0.0442 (0.0409)	0.0488 (0.0409)	0.0473 (0.0409)	0.0494 (0.0408)	0.0485 (0.041)	0.0485 (0.0413)
5 dB	EER (%)	13.42 (12.51)	10.73 (20.93)	*10.21* (14.18)	11.05 (14.35)	11.25 (12.6)	10.32 (12.37)	12.52 (12.59)	10.32 (13.31)	14.09 (13.51)	13.56 (13.14)
	mDCF	0.048 (0.0431)	0.0453 (0.0499)	0.0444 (0.0443)	*0.0441* (0.0445)	0.0487 (0.0428)	0.0476 (0.0427)	0.0447 (0.0428)	0.0479 (0.0436)	0.0459 (0.0436)	0.0458 (0.0435)
0 dB	EER (%)	22.93 (20.29)	*16.9* (23.29)	18.43 (21.04)	18.76 (21.6)	20.04 (20.42)	21.56 (20.76)	22.22 (20.1)	20.86 (22.17)	21.04 (20.68)	18.71 (22.03)
	mDCF	0.0476 (0.0493)	0.0485 (0.0522)	0.051 (0.05)	0.0527 (0.0505)	*0.0453* (0.0493)	0.0486 (0.0496)	0.0489 (0.049)	0.0556 (0.051)	0.0464 (0.0495)	0.0488 (0.0508)

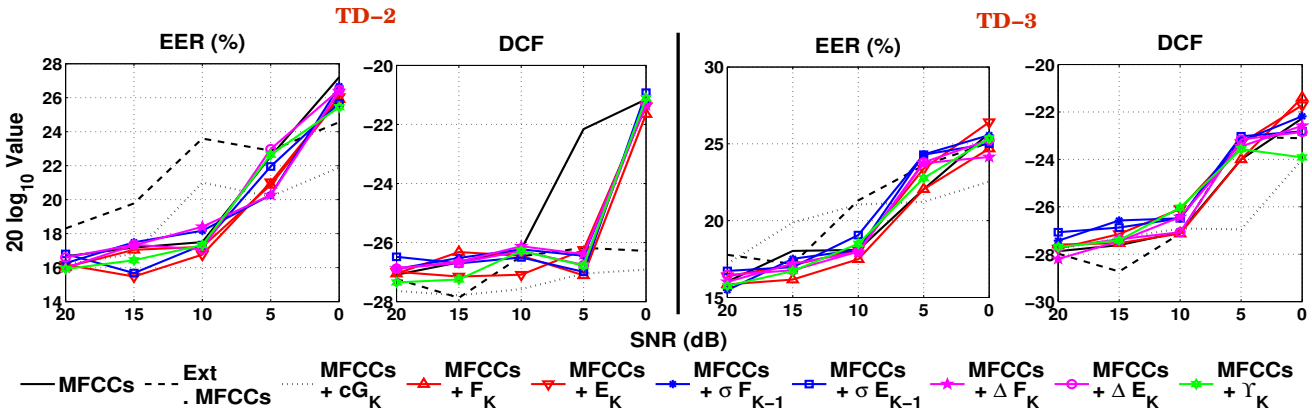


Figure 5: The performance of the TDSV system using robust end point detection followed by MEMD feature extraction. The results are shown for the 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the cG_K seven features are kept constant, with $K = 4$.

Table 5: The performance of the TDSV system using speech enhancement followed by robust end point detection and extraction of MEMD features. Results are shown for 39-dimensional MFCCs, 51-dimensional Ext. MFCCs, and the combinations of the 39-dimensional MFCCs with the cG_K feature and each of the seven HS experimental feature obtained from MEMD. The performance metrics are evaluated for the TD-1 speech utterance of the RSR2015 database. The testing utterances are corrupted with Babble noise, with SNR varying from 20-0 dB, in steps of 5 dB. The dimensions of the MEMD features are kept constant, with $K = 4$.

Technique		Mel filterbank		MEMD filterbank	MEMD-Hilbert Spectrum						
Feature		MFCCs	Ext. MFCCs	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +	MFCCs +
SNR	Metric	39 dim	51 dim	cG_4	F_4	E_4	σF_3	σE_3	ΔF_4	ΔE_4	Υ_4
20 dB	EER (%)	7.3	5.37	5.08	5.08	5.12	5.37	5.08	5.08	5.51	5.83
	mDCF	0.0353	0.0341	0.0341	0.033	0.0338	0.0341	0.033	0.0341	0.038	0.0365
15 dB	EER (%)	10.37	7.59	11.55	11.36	10.75	10.85	8.84	11.55	8.45	7.01
	mDCF	0.0571	0.0337	0.0428	0.0428	0.066	0.0428	0.0381	0.0428	0.044	0.0441
10 dB	EER (%)	10.55	12.6	13.36	11.55	12.43	11.55	11.48	13.36	10.75	11.22
	mDCF	0.0704	0.0648	0.054	0.0535	0.0522	0.0649	0.0635	0.054	0.066	0.0482
5 dB	EER (%)	13.15	17.68	17.68	16.57	16.18	16.18	17.89	17.68	16.53	15.49
	mDCF	0.0725	0.0669	0.0648	0.0647	0.0707	0.0707	0.0725	0.0648	0.0617	0.0602
0 dB	EER (%)	17.24	17.73	17.73	17.17	21.64	21.64	17.94	17.73	17.68	20.14
	mDCF	0.0864	0.0748	0.0669	0.0758	0.0781	0.0781	0.084	0.0669	0.082	0.0777

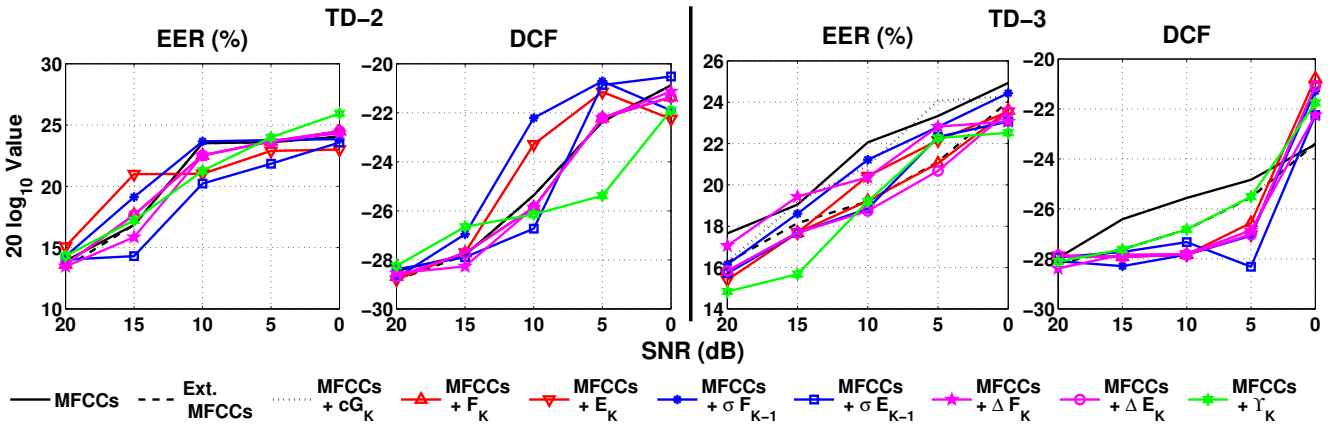


Figure 6: The performance of the TDSV system using speech enhancement followed by robust end point detection and extraction of MEMD features. Results are shown for 39-dimensional MFCCs with the cG_K feature and each of the seven experimental features. The performance metrics are evaluated for the TD-2 and TD-3 sentences of the RSR2015 database. The testing utterances are corrupted by Babble noise, with SNR varying from 20-0 dB. The dimensions of the cG_K seven features are kept constant, with $K = 4$.

6 Summary

In this thesis, we focused on developing a robust TDSV system suitable for under the degraded conditions. In the pre-processing stage, a robust end point detection using speech specific knowledge is used instead of energy based voice activity detection. In addition, we also explored application of combined temporal and spectral speech enhancement technique for the detection of end points in a better way. In the feature extraction stage, features are extracted from HS of the IMFs obtained from MEMD augmented with

MFCC features. Further, results showed that the combination of speech enhancement followed by speech-specific knowledge-based end point detection and augmentation of HS-based features with MFCCs, showed significant improvement in the performance of TDSV system under low SNR cases.

6.1 Contributions of the thesis

The work presented in this thesis proposes TDSV system for practical scenarios using noise robust techniques. The contributions of the thesis are the following:

- Investigating combined temporal and spectral speech enhancement technique for TDSV task under practical scenarios.
- Investigating the HS features in combination with MFCCs for TDSV task.
- Proposing a method for detecting the begin and end points using speech specific knowledge.
- Showing that the various noise robust techniques at different stages of a TDSV system further improves the system performance.

7 List of Publications

Journal Publications

- Published Papers:

  (i) Ramesh K. Bhukya, S. R. M. Prasanna and Biswajit Dev Sarma, "Robust Methods for Text-Dependent Speaker Verification", **Circuits, Systems, and Signal Processing, (Springer)**, vol. 38, no. 11, pp. 5253–5288, November 2019.

  (ii) Ramesh K. Bhukya, Biswajit Dev Sarma and S. R. M. Prasanna, "End Point Detection Using Speech Specific Knowledge for Text-Dependent Speaker Verification", **Circuits, Systems, and Signal Processing (Springer)**, vol. 37, no. 12, pp. 5507 - 5539, May 2018.

  (iii) Rajib Sharma, Ramesh K. Bhukya and S. R. M. Prasanna, "Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification", **Speech Communication (Elsevier)**, vol. 96, pp. 207-224, February 2018.

  (iv) Banriskhem K Khonglah, Ramesh K. Bhukya and S. R. M. Prasanna, "Processing Degraded Speech for Text-Dependent Speaker Verification," **International Journal of Speech Technology (Springer)**, vol. 20, pp. 839-850, December 2017.

  (v) Rajib Sharma, S. R. M. Prasanna, Ramesh K. Bhukya and R.K.Das, "Analysis of the Intrinsic Mode Functions for Speaker Information", **Speech Communication (Elsevier)**, vol. 91, pp. 1-16, July 2017.

- Manuscripts Under Review

  (i) Ramesh K. Bhukya and S. R. M. Prasanna, "Some Issues in the Practical Deployment of the Text-Dependent Speaker Verification", **Under Review in**, Journal of Signal Processing Systems, Springer, submitted in October 2018.

  (ii) Ramesh K. Bhukya, S. R. M. Prasanna and Sarfaraz Jelil, "Text-Dependent Speaker Verification Using DTW-GMM Based Two-tier Authentication", **Under Review in**, Expert Systems with Applications, Elsevier, submitted in November 2018.

  (iii) Ramesh K. Bhukya and S. R. M. Prasanna, "End Point Detection, Speech Enhancement and Feature Extraction - A Review", **to be submitted in IETE Technical Review**.

  (iv) Ramesh K. Bhukya and S. R. M. Prasanna, " Approaches for Robust Methods for End-to-End Text-Dependent Speaker Verification Under Practical Conditions", **to be submitted in**, **Circuits, Systems, and Signal Processing, (Springer)**.

  (v) Ramesh K. Bhukya and S. R. M. Prasanna, "Deep Learning Approaches for End-to-End Robust Text-Dependent Speaker Verification Under Degraded Conditions", **to be submitted in**, **IET Biometrics**.

  (vi) Ramesh K. Bhukya and S. R. M. Prasanna, " Robust Voice Liveness Detection and Text-Dependent Speaker Verification Using Throat Microphone", **to be submitted in**, **IET Signal Processing**.

Conference Publications

- Published Paper and Accepted Publication:

  (i) S. Dey, S. Barman, Ramesh K. Bhukya, R.K.Das, Haris B. C, S. R. M. Prasanna and R. Sinha, "Speech Biometric Based Attendance System", *in Proc. National Conf. on Communication (NCC)*, IITK, Kanpur, India, February 2014.

  (ii) D. Mahanta, A. Paul, Ramesh K. Bhukya, R.K.Das, R.Sinha and S. R. M. Prasanna, "Warping Path and Gross Spectrum Information for Speaker Verification Under Degraded Condition", *in Proc. National Conf. on Communication (NCC),* IITG, Guwahati, India, 2016.

  (iii) A. Paul, D. Mahanta, R.K.Das, Ramesh K. Bhukya and S. R. M. Prasanna, " Presence of Speech Region Detection Using Vowel-like Regions and Spectral Slope Information", *in Proc. IEEE India International Conference (INDICON)*, IITR, Roorkee, India, December 2017.

References

[1] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[2] B. K. Khonglah, R. K. Bhukya, and S. R. M. Prasanna, "Processing degraded speech for text dependent speaker verification," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 839–850, 2017.

[3] R. Sharma, R. K. Bhukya, and S. R. M. Prasanna, "Analysis of the hilbert spectrum for text-dependent speaker verification," *Speech Communication*, vol. 96, pp. 207–224, 2018.

[4] R. K. Bhukya, B. D. Sarma, and S. R. M. Prasanna, "End point detection using speech-specific knowledge for text-dependent speaker verification," *Circuits, Systems, and Signal Processing*, pp. 1–33, 2018.

[5] S. Dey, S. Barman, R. K. Bhukya, R. K. Das, B. C. Haris, S. R. M. Prasanna, and R. Sinha, "Speech biometric based attendance system," in *National Conference on Communications*, 2014.

[6] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, 2017.

[7] P. Krishnamoorthy, "Combined temporal and spectral processing methods for speech enhancement," Ph.D. dissertation, Ph. D. dissertation, Indian Institute of Technology Guwahati, Assam, India, 2009.

[8] R. Sharma and S. R. M. Prasanna, "A better decomposition of speech obtained using modified empirical mode decomposition," *Digital Signal Processing*, vol. 58, pp. 26–39, 2016.

[9] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[10] R. Sharma, S. R. M. Prasanna, R. K. Bhukya, and R. K. Das, "Analysis of the intrinsic mode functions for speaker information," *Speech Communication*, vol. 91, pp. 1–16, 2017.

[11] 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.

[12] B. D. Sarma and S. R. M. Prasanna, "Analysis of spurious vowel-like regions (vlrs) detected by excitation source information," in *India Conference (INDICON), 2013 Annual IEEE*. IEEE, 2013, pp. 1–5.

[13] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, 2009.

[14] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 8, pp. 2552–2565, 2011.

[15] K. Ramesh, S. R. M. Prasanna, and R. K. Das, "Significance of glottal activity detection and glottal signature for text dependent speaker verification," in *Signal Processing and Communications (SPCOM), 2014 International Conference on*. IEEE, 2014, pp. 1–5.

[16] B. D. Sarma, S. R. M. Prasanna, and P. Sarmah, "Consonant-vowel unit recognition using dominant aperiodic and transition region detection," *Speech Communication*, vol. 92, pp. 77–89, 2017.

[17] R. K. Bhukya, S. R. M. Prasanna, and B. D. Sarma, "Robust methods for text-dependent speaker verification," *Circuits, Systems, and Signal Processing*, pp. –, Under Review.

[18] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

Organization of the Thesis

(i) **Introduction**

1.1 Objective of the Thesis
1.2 Introduction
1.3 Speaker Verification System
1.4 A Glance Towards Text-Dependent Speaker Verification
1.5 Motivation for the Present Work
1.6 Issues in the development of a TDSV system
1.7 Organization of the Thesis

(ii) **Text-Dependent Speaker Verification Under Degraded Conditions - A Review**

2.1 Objective
2.2 Introduction
2.3 Methods for End Point Detection
2.4 Different Methods for Speech Enhancement
2.5 Different Features for Speaking Modeling
2.6 Summary and Scope for Present Work

(iii) **Processing Degraded Speech for Text-Dependent Speaker Verification**

3.1 Objective
3.2 Introduction
3.3 Speech Enhancement Techniques
3.4 Development of Text-Dependent Speaker verification
3.5 Experimental Results and Discussions
3.6 Summary and Conclusion

(iv) **Analysis of the Hilbert Spectrum for Speaker Verification**

4.1 Objective
4.2 Introduction
4.3 EMD, MEMD and HS
4.4 Utilizing the Constituents of the HS for Characterizing Speakers
4.5 Experimental Setup
4.6 Results and Analysis
4.9 Summary and Conclusion

(v) **Significance of Speech-Specific Knowledge for End Point Detection**

5.1 Objective
5.2 Introduction
5.3 Significance of Speech Specific Knowledge for End Point Detection
5.4 Robust End Point Detection Using Speech Specific Knowledge
5.5 Experimental Evaluation of Proposed End Point Detection
5.6 Development of Text-Dependent Speaker Verification
5.7 Experimental Evaluation
5.8 Summary and Conclusion

(vi) **Exploration of Various Noise Robust Techniques for Speaker Verification**

5.1 Objective

5.2 Introduction

5.3 Robust End Point Detection

5.4 Robust Features from Hilbert Spectrum

5.5 Speech Enhancement Techniques for TDSV System

5.6 Experimental Setup

5.7 Results and Analysis

5.8 Summary and Conclusion

(vii) **Summary and Conclusions**

6.1 Summary of the Present Work

6.2 Contributions of the Present Work

6.3 Directions for Future Research