



SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY



EFFICIENT AND ROBUST SPEAKER IDENTIFICATION SYSTEM UNDER ADVERSE CONDITIONS

PhD dissertation

Swati Prasad

Aarhus BSS
Aarhus University
CTIF Global Capsule
Future Technologies for Business Ecosystem Innovation (FT4BI)
Department of Business Development and Technology
(BTECH)

2018

PhD Supervisor: Prof. Ramjee Prasad
Aarhus University, Denmark

PhD Committee:

- Associate Professor Sofoklis Kyriazakosh,
Department of Business Development and Technology, Aarhus BSS,
Denmark
- Professor Seshadri Mohan, University of Arkansas at Little Rock,
Arkansas, United States.
- Professor Gregory Yovanof, Athens Information Technology (AIT),
Athens, Greece.

PhD Series: Future Technologies for Business Ecosystem Innovation (FT4BI),
Department of Business Development and Technology (BTECH), School of
Business and Social Sciences, Aarhus University, Herning, Denmark

© Copyright by Swati Prasad

Abstract

Speaker identification system (or voice based biometric), which determines the speaker of a given speech utterance from a group of people, is predicted to have huge importance in near future. Speech being the natural means of communication for humans, fits really well in our requirement for less complex and more secure means of identification, followed by interaction with the electronic devices, than the existing password based methods. But speaker identification system faces poor performance due to a mismatch between the train and the test speech conditions. A mismatch can arise due to factors like differences in handset, transmission channel or microphone; environmental noise; emotional state of the person; voice disguise, etc. It is also referred to as mismatched problem.

In this thesis, mismatched problem, which arises due to environmental noise and voice disguise has been focused. To tackle performance degradation because of environmental noise mismatch, a hybrid method for feature frame selection has been developed. It combines voice activity detection (VAD) and variable frame rate (VFR) analysis methods. The hybrid method efficiently captures the speech part rejecting the non-speech, and the changes in the temporal characteristics of the speech signal considering the signal to noise ratio.

Mismatched problem which arises because of the adoption of voice disguise is less researched and poses threat to the existing speaker identification systems. Therefore, mismatch due to voice disguise has been focused in this

thesis. Robust methods have been developed at the feature, training and the testing level. It has been found that, the use of fixed frame shift, typically of 10ms, leading to a fixed frame rate for acquiring the frames for feature extraction, might not give the best identification accuracy under voice disguise. Therefore, a multiple-model framework which combines features obtained by utilizing three different frame shifts of 3ms, 6ms and 9ms has been developed, and it has shown improved performance over the fixed frame shift method.

Four multistyle training strategies have been investigated for tackling voice disguise mismatch seen in a security conscious organization. It has shown encouraging results over the single style training. Further a fusion framework, utilizing the best two investigated multistyle training strategies has been proposed. It has shown an overall improved performance over single style training, investigated multistyle trainings and the multiple-model methods.

Finally, a method combining multiple frame rates for feature extraction and reliable frame selection at the decision level has also been developed. It has shown an overall better performance over the baseline methods.

Since, voice production is governed by brain, an attempt has also been made to study the brain signal response for motor imagery tasks. This might prove beneficial in future for improving the speaker identification system. A Brain Computer Interface (BCI) to classify motor imagery tasks from the same limb has been studied. For this, a feature selection strategy for brain signal has also been proposed. It consisted of channel selection based on Fisher ratio and time-segment selection by visual inspection. It has shown improved performance over the baseline system.

Dansk Abstrakt

Højttaler identifikation system (eller stemme baseret biometrisk), bestemmer højttaleren af en givet tale udtalelse fra en gruppe mennesker. Det er forudsagt at have stor betydning i den nærmeste fremtid. Tale er det naturlige kommunikationsmiddel for mennesker. Det passer rigtig godt i vores krav til mindre komplekse og sikker identifikationsmidler, efterfulgt af interaktion med de elektroniske enheder, end de eksisterende adgangskode baserede metoder. Men højttaler identifikation systemet står over for dårlige ydeevne på grund af misforhold mellem træning og teste tale miljø. En misforhold kan opstå på grund af faktorer synes godt om forskelle i håndsæt, transmission kanal eller mikrofon; miljøstøj; følelsesmæssige tilstand af personen; stemme forklædning, mv. Det omtales også som mismatchet problem.

I denne afhandling er mismatchet problem, der opstår som følge af miljøstøj og stemme forklædning, blevet fokuseret. At løse nedgang i identifikationsnøjagtighed på grund af miljøstøj-fejlpasning, en hybrid metode til plukning af talrammen er blevet udviklet. Den kombinerer stemme aktivitets detektering og variabel ramme sats analyse metoder. Hybrid metoden indfanger effektivt taledelen, og afviser ikke-talen, og ændringerne i talesignalets tidsmæssige egenskaber under hensyntagen til signal til støjforholdet.

Mismatchet problem, der opstår på grund af vedtagelsen af stemme forklædning, er mindre undersøgt og udgør en trussel mod de eksisterende højttaler identifikations systemer. Derfor har mismatch problem på grund af stemme

forklådning været fokuseret i denne afhandling. Robuste metoder er udviklet på tre niveauer, nemlig featureudtræk, træning og test. Det har vist sig, at brugen af fast ramme skift, typisk 10ms, hvilket fører til en fast billedfrekvens til ramme plukning muligvis ikke giver den bedste identifikations nøjagtighed under stemme forklådning. Derfor er der udviklet en tilgang, der kombinerer flere modeller. Her er der udviklet en model, som udnyttede funktioner opnået ved anvendelse af tre forskellige rammeforskydninger på 3ms, 6ms og 9ms. Det har vist forbedret ydeevne over den fast billedfrekvens metode.

Fire multi-stil trænings strategier er blevet undersøgt for at tackle mismatch på grund af stemme forklådning set i en sikkerhedsbevidst organisation. Det har vist opmuntrende resultater i sammenligning med single-stil træning. Endvidere er der foreslået en fusion ramme, der udnytter de bedste to undersøgte multi-stil trænings strategier. Det har vist en overordnet forbedret ydeevne i forhold til single-stil træning, undersøgt multi-stil træning og mange-model metoder.

Endelig er der udviklet en metode, der kombinerer flere billedfrekvens til ekstraktion af funktioner og pålidelig ramme plukning på beslutningsniveau. Det har vist bedre ydeevne end basislinje metoderne.

Siden stemmeproduktionen styres af hjernen, er der også forsøgt at studere hjernens signalrespons for fantasi af motoraktivitet. Dette kan vise sig at være gavnligt i fremtiden for at forbedre højtaleridentifikations systemet. En hjerne computer grænseflade til klassificering af forestillede motoriske aktiviteter fra samme lemmer er blevet undersøgt. Til dette er der også foreslået en featurevalgsstrategi for hjerne-signal. Det bestod af kanalvalg baseret på fisher ratio og tidssegmentvalg ved visuel inspektion. Det har vist forbedret ydelse i sammenligning med basislinje metoderne.

CV



I am working as Assistant Professor in the Department of Electronics & Communication Engineering at Birla Institute of Technology, Mesra, Ranchi, India, since 2007. I have received my M.E degree in Electrical & Electronics Engineering from Birla Institute of Technology, Mesra, Ranchi, India in 2007. I am a recipient of a B-level certificate in the National Mathematics Olympiad Contest, held in the year 1995, and I have also received the Erasmus Mundus scholarship for pursuing my Ph.D. degree at Denmark. The Ph.D. thesis will be submitted at Aarhus University, Denmark.

My research interest lies in the area of Speaker Identification, Brain Computer Interface (BCI) and Digital Electronics. I have reviewed research papers for Journals like, IEEE Transactions on Industrial Electronics, Computer Speech & Language and Wireless Personal Communications, and for Conferences like, EUSIPCO and ICACCI.

Dansk CV

Jeg arbejder som adjunkt i Department of Electronics & Communication Engineering ved Birla Institute of Technology, Mesra, Ranchi, Indien siden 2007. Jeg har modtaget min ME-grad i Electrical & Electronics Engineering fra Birla Institute of Technology, Mesra, Ranchi, Indien i 2007. Jeg er modtager af et B-niveau certifikat i National Mathematics Olympiad Contest, der blev afholdt i 1995, og jeg har ogs modtaget Erasmus Mundus stipendier til at forflge min Ph.D. Grad i danmark.

Min forskningsinteresse ligger inden for Speaker Identification, Brain Computer Interface (BCI) og Digital Electronics. Jeg har gennemget forskningsdokumenter for tidsskrifter som, IEEE Transaktioner p Industriel Elektronik, Computer Speech og Language & Wireless Personal Communications, og konferencer som, EUSIPCO, ICACCI osv.

Acknowledgements

I want to convey my sincere thanks to my supervisor, Prof. Ramjee Prasad for his vision, encouragement and guidance throughout this Ph.D. study. I also want to thank Prof. Zheng Hua-Tan for his valuable insight and constant guidance related to my work. Through him, I have learned many things about Speaker Identification area.

My special thanks to the whole of Center for TeleInFrastruktur (CTIF), Aalborg University team, particularly, Dr. Neeli R. Prasad & Dr. Rasmus H. Nielsen for facilitating my research work and Susanne Nørrevang, Inga Hauge, Dorthe Sparre & Liselotte Sigh Andersen for their support and timely help in acquiring necessary equipments and facilities for conducting my research work. I want to thank my colleagues Prof. Nisha Gupta, Prof. Vibha Rani Gupta & Dr. Sanjay Kumar at Birla Institute of Technology, Mesra, Ranchi, India for their constant support and guidance in my Ph.D. work.

I want to extend my sincere appreciation to my colleagues and friends at Aalborg, Prateek Mathur, Ambuj Kumar, Sonam Tobgay, Chayapol Kamyod, Dilip Chaudhary and Vrinda Hitendra Kurande, with whom I had many useful discussions related to my work, which eventually helped me in progressing towards my goal. It was because of them that my stay at Denmark was so homely.

Finally, I would like to give my heartily thanks to my family, In-laws and my husband Sawan for their patience, support and encouragement in completing

this work. Last but not the least, my special appreciation to the little ones of my family, Sanjeevani, Manas and Aakash whose innocence filled my heart with happiness and to all those who helped me directly or indirectly in this Ph.D. study.

Contents

1	Introduction	1
1.1	Speaker Identification	1
1.1.1	Applications	4
1.2	Speaker Identification System	5
1.2.1	Performance Evaluation	8
1.3	Challenges Faced in the Speaker Identification - Motivation	8
1.4	Thesis's Objectives	9
1.5	Contribution	9
1.6	Publication List	12
1.7	Structure of the Thesis	13
2	State of the Art	15
2.1	Mismatched Problem	15
2.1.1	Factors Causing the Mismatch:	16
2.1.2	Addressing the Mismatched Problem	17
3	Feature Frame selection - A Hybrid Approach for Environmen- tal Noise Corrupted Speech	23
3.1	Introduction	23
3.2	Hybrid Feature Frame Selection	26
3.3	Experiments and Results	35

3.3.1	Database	35
3.3.2	Average frame rate calculations	36
3.3.3	Speaker Identification Experiments & Results Discussion	38
3.4	Summary	42
4	Multi-Frame Rate based Multiple-Model for Disguised Speech	45
4.1	Introduction	45
4.2	Multi-Frame Rate based Multiple-Model	48
4.2.1	GMM	48
4.2.2	Multi-Frame Rate based Multiple-Model Speaker Identification	49
4.3	Experimental Setups and Results	50
4.3.1	Database	50
4.3.2	Speaker Identification Experiments	51
4.3.3	Results	52
4.4	Summary	57
5	Multistyle Training and Fusion Framework for Disguised Speech	59
5.1	Introduction	59
5.2	The Different Speaking Styles	61
5.3	Multistyle Training Strategies and the Fusion Framework . . .	62
5.4	Experimental Setups and Results	64
5.4.1	Database	65
5.4.2	Speaker Identification Experiments	65
5.4.3	Results	67
5.5	Summary	72
6	Multiple Frame Rates for Feature Extraction and Reliable Frame Selection at the Decision for Disguised Speech	75
6.1	Introduction	75

6.2	Reliable Frame Selection at the Decision	76
6.3	Experimental Setup and Results	78
6.3.1	Database	78
6.3.2	Speaker Identification Experiments	79
6.3.3	Results and Discussions	80
6.4	Summary	85
7	Brain Computer Interface for Classification of Motor Imagery	
	Tasks from the Same Limb Using EEG	87
7.1	Introduction	87
7.2	Dataset	90
7.3	Feature Selection and Extraction Strategy	91
7.4	Experiments and Results	94
7.5	Summary	97
8	Conclusions & Future Work	99
8.1	Conclusions	99
8.2	Future Work	102
	Bibliography	103
	Coauthor Statement	118

List of Figures

1.1	Types of Speaker Recognition System. (a) Speaker Identification. (b) Speaker Verification.	2
1.2	Speaker identification system [12]. (a) Training. (b) Testing/Identification.	7
3.1	Frame selection by the hybrid method [85].	27
3.2	Clean speech utterance frame selection [85].	32
3.3	Babble noise corrupted speech utterance frame selection at SNR of 15dB and 5dB [85].	33
3.4	Car noise corrupted speech utterance frame selection at SNR of 15dB and 5dB [85].	34
3.5	Comparison of the proposed and baseline method's average identification accuracies under various noise scenarios [85]	41
3.6	Comparison of the proposed and the baseline method's average identification accuracies (%) at different SNR values [85]	42
4.1	Comparison of the identification accuracies(%) obtained using speaker models with different frame rates and the baseline for the four speaking style's test data [104].	54
4.2	Comparison of the identification accuracies (%) of the different frame rates models (average), multi-frame rate based multiple-model training and baseline for the four speaking style's test speech data [104]	56

4.3	Average identification accuracy across different speaking styles for the different frame rates, the multi-frame rate based multiple-model training and the baseline method	56
5.1	Norm, Sync and Fast speaking style spectrogram for the speech utterance “If it doesn’t matter who wins, why do we keep score?” [105]	63
5.2	Comparison of the average identification accuracies across the Norm and disguised test speech data for the various speaker identification experiments [106].	69
5.3	Comparison of the identification accuracies of the different multistyle training strategies and the Fusion method for Norm and disguised speech test data [105].	70
5.4	Comparison of the Fusion and the Single style training method for the Norm and voice disguised test speech data [105].	71
5.5	Comparison of Fusion and the multiple model methods for the different voice disguises and Norm speaking style [106].	72
6.1	For the Rsi speech test data, identification accuracies (%) obtained by using different threshold value θ [12]	83
7.1	For sub F, channel Cz graph for the class BH and BL, task onset is at 0s, and the two segments extracted are from -1s to 1s and from 2s to 4s [116]	92

List of Tables

3.1	Parameter A & B selection used in the threshold value for average frame rate calculation in the VFR method [85]	36
3.2	Identification accuracies (%) of the VFR method at different average frame rates for clean and noisy speech [85]	37
3.3	Identification accuracies (%) obtained for the various methods under clean & noisy conditions [85]	40
3.4	Proposed and the baseline methods average identification accuracies (%) at different SNR values [85]	41
4.1	Identification accuracies (%) of the speaker identification experiments which are discussed in Subsection 4.3.2 for the four types of speaking style test data [104].	53
5.1	Identification Accuracies (%) of the various speaker identification experiments for the Norm and disguised speech test data [106].	68
7.1	“Misclassification rate (%) for all the subjects” [116]	96

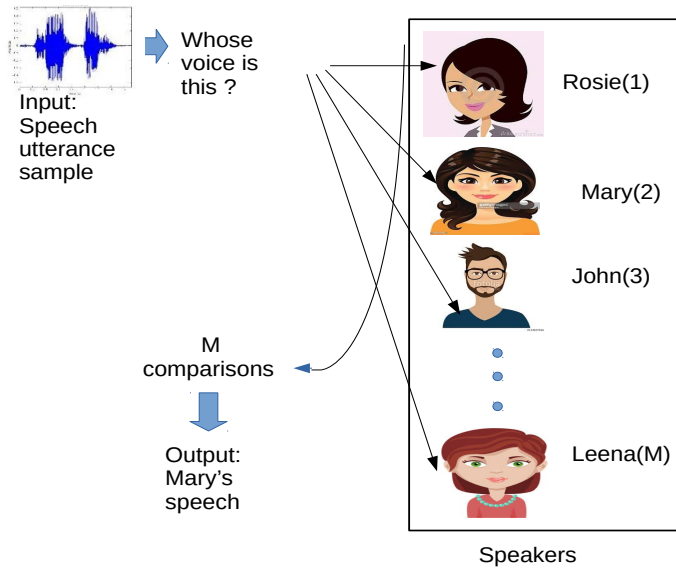
Chapter 1

Introduction

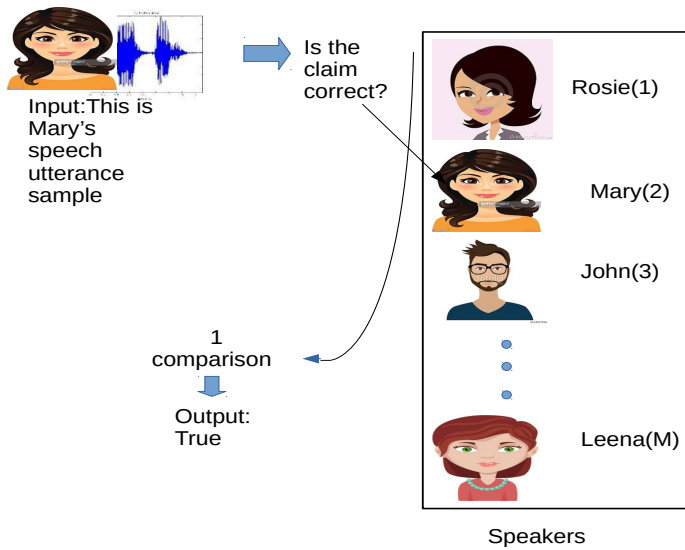
1.1 Speaker Identification

Determining the speaker of a given speech utterance from a group of reference speakers is referred to as *speaker identification*, and a machine providing this task is referred to as speaker identification system. It is also commonly known as voice based biometric [1, 2].

Speaker identification comes under the broad category of *speaker recognition*, which in general, tackles the problem of finding out the speaker of a given speech sample. Speaker recognition is mainly classified into two categories: *speaker identification* and *speaker verification* as shown in Fig. 1.1 below. In speaker identification, the speaker of the given speech sample can be anyone amongst the M reference speakers. Determining the correct speaker requires M comparisons for similarity measure of the given speech sample with each of the M reference speakers. The speaker which gives the highest similarity measure value is decided as the speaker of the given speech sample. Therefore, an increase in the number of reference speakers, means, more number of



(a) Speaker Identification



(b) Speaker Verification

Figure 1.1: Types of Speaker Recognition System. (a) Speaker Identification. (b) Speaker Verification.

comparisons which can lead to a decrease in the accuracy of the speaker identification system. In speaker verification, the given speech sample is spoken by a particular speaker is already claimed. The problem is to only determine, whether the claim is true or false. It requires only one comparison with the claimed speaker and the accuracy of the system is not affected by an increase in the number of reference speakers [3].

Speaker identification can be termed as *open* or *closed*. In an open speaker identification, the speaker of the given speech sample can be anyone belonging to either the set of M reference speakers or from outside this set. Here, the task is to first determine, whether the given speech sample belongs to the set of M reference speakers, and then only the usual identification process is carried out. This system can therefore take $M+1$ decisions, including the decision that, the given voice sample did not belong to any of the reference speakers and is an outsider. In a closed speaker identification, the speaker of the given speech sample always belongs to the set of M reference speakers.

Speaker identification can be further termed as *text-dependent* and *text-independent*. In text-dependent speaker identification, the words of the given speech sample whose speaker is required, should be from a predefined set of words. On the other hand, for text-independent speaker identification, such constraints are not there and the given speech sample can consist of any words of the speaker's choice. It is easy to understand that, the text-independent speaker identification is more difficult than the text-dependent [4] - [8].

In this thesis, closed text-independent speaker identification system is studied, where the problem is to find the speaker of a given speech utterance from a group of M reference speakers (closed). The given speech utterance consisted of words or sentences of the person's choice and no constraint is put in this regard on the person (text-independent). For convenience, closed, text-independent speaker identification system will be simply referred to as speaker identification

system in the rest of the thesis.

1.1.1 Applications

With the fast advancement in technology, more and more applications requiring speaker identification are being designed. The existing password/token based identification will soon become difficult to manage as the number of passwords to be remembered will become large. Speech, being the natural means of communication for humans, has the potential to be adopted more widely in future for biometric based applications. Using speech instead of password for identification will reduce the complexity with which humans interact with the machines. It will also offer better security, as it cannot be stolen, and will be particularly liked by the elderly members of the society and people with mental disabilities because of its ease of use. People will be more cooperative in providing speech for identification, as compared to other modalities like face and iris, as some communities like, Muslim women, hide their face with cloth as part of their religion and might not be comfortable providing their face for identification, also capturing the iris image, requires laser irradiation, which can pose health problems.

Since, voice based system, like telephone is already in use, speech based biometric for remote applications can also be more easily implemented compared to other biometrics like, face, iris and finger. Moreover, finger, iris and face consist of only physical characteristics, on the other hand, speech consist of two different types of characteristics: “physical” and “learned”. Physical characteristics occur in voice because of the structure of the vocal tract, larynx and voice production organs, whereas, learned characteristics are acquired by a person from his/her environment in a period of time. For example, the accent of an European person will be quite different from an Asian person, even if, they both speak the same sentence of the English language. Because of these

two characteristics, i.e. physical and learned, speech offers much more research possibilities and application areas, which are still not fully explored.

For the communication between humans and machines through voice, we need to be successful in the following:

1. Machines, which can understand, “what is being spoken” i.e. the content of the speech, referred to as speech recognition.
2. At the same time, machines being intelligent enough to identify, “who is speaking”, referred to as speaker recognition.

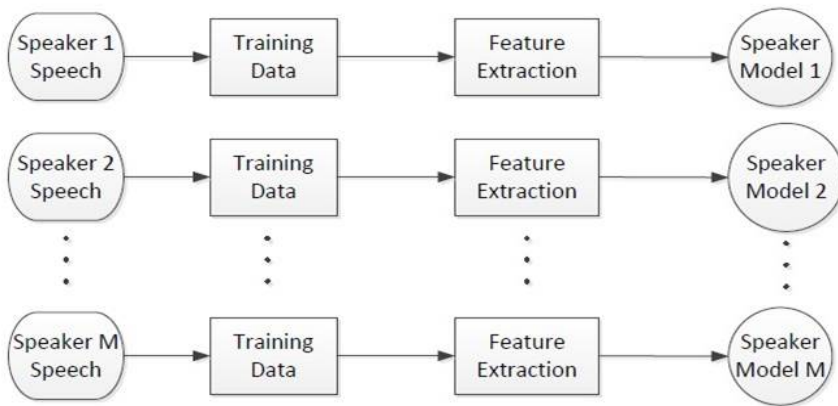
If the above two things, successfully come into reality, it will revolutionize the current electronic consumer market. Imagine an old person giving command to the smart television through voice instead of remote control about the channel he/she wishes to view or about other functionality like volume or colour. The smart television on the other hand, first identify the person as one amongst the authorized users (speaker identification) and then only proceeds for carrying out his/her commands (speech recognition). Using voice in place of remote control, ensures both security and ease of use. It should be noted that, smart television is just one example, there can be many more applications like smart mobile phone, smart door lock, smart washing machines and smart robots, which can utilize this technology. Apart from these, speaker identification finds applications in surveillance, border control, forensic science, identifying the hoax caller, identifying the kidnapper and monitoring the staff in a security conscious organization [9, 10].

1.2 Speaker Identification System

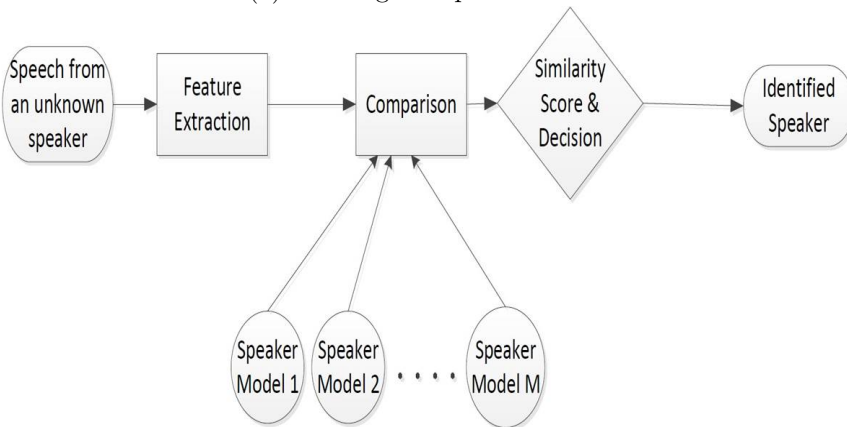
The speaker identification system can be broadly divided into three phases: feature selection and extraction phase, training or enrollment phase and testing

or identification phase [1, 3].

1. Feature selection and extraction phase: This comprises of selecting important parts of the time-domain speech signal from the less important ones and then extracting relevant features from these parts from the irrelevant ones. Relevant features are the ones, which are found in abundance in the speech samples from the same speaker than from different speakers and has the ability to be unique and robust with respect to noise. A list of ideal characteristics of feature is listed in [11]. It says that, ideally
 - A feature should occur frequently and naturally in speech.
 - It should differ from other people's feature but for each speaker it should be consistent.
 - It should not vary with age, emotional state or health of the person.
 - It should be robust to background noise.
 - It should not get modified, if the person deliberately or non-deliberately modify his/her speaking style.
2. Training or enrollment phase: In the training phase, speech samples are collected from each of the M reference speakers. From the collected speech samples of each speaker, features are then selected and extracted. Finally, these features are used to develop a model for each of the speaker, resulting in M speaker models. The modeling process tries to search for a pattern which is unique for every speaker, i.e., the occurrence of the pattern will be more within the speech samples from the same speaker than from different speakers. The training process is shown in the Fig. 1.2(a).
3. Testing or identification phase: In the testing phase, a speech utterance from an unknown speaker will be given. A comparison of this speech utterance with all the M speaker models developed during the training



(a) Training the speaker model



(b) Testing / Recognition

Figure 1.2: Speaker identification system [12]. (a) Training. (b) Testing/Identification.

phase will be carried out for similarity measure. The model scoring the highest similarity measure will be decided as the correct speaker of the given speech utterance. The testing process is shown in the Fig. 1.2(b).

1.2.1 Performance Evaluation

The performance of the speaker identification system is measured using the identification accuracy. It is defined as

$$\begin{aligned} & \text{Identification Accuracy} \\ = & \frac{\text{the number of correctly identified utterances}}{\text{total number of utterances tested from the total speakers}} \\ \times & 100\% \end{aligned} \tag{1.1}$$

1.3 Challenges Faced in the Speaker

Identification - Motivation

Speaker identification research has started long back and in spite of understanding the huge benefits of it for future applications, it is still not considered very reliable and successful in the present scenario. It faces the problem of significant decrease in the performance (identification accuracy), when a mismatch between the training and the test data conditions occur. This is referred to as *mismatched problem*. A mismatch can occur because of environmental noise, noise incorporated by the voice recording device, because of ageing, throat infection, emotional state of the person, etc. A mismatch can also occur when a person intentionally changes his/her voice, referred to as voice disguise, either to hide his/her own information or to sound like a target speaker to steal the target's information [3, 13]. The present study will focus on increasing the identification accuracy, the ultimate goal of the speaker identification system under the mismatched conditions, particularly, under the environmental noise and voice disguise scenarios.

1.4 Thesis's Objectives

Hypothesis: Relevant feature selection and extraction from the speech signal can greatly enhance the efficiency and robustness of the speaker identification system under adverse conditions.

Statement of thesis's objectives:

- To search for feature selection & extraction methods from the speech signal with an aim to increase the efficiency and robustness of the speaker identification system under mismatched conditions.
- To explore and design robust and efficient training methods for the speaker identification system to tackle the mismatched problem.
- To explore and design robust and efficient testing methods for the speaker identification system to tackle the mismatched problem.

1.5 Contribution

Through this thesis work, the following contributions have been made:

1. The performance of the speaker identification system decreases markedly in real life scenario when the test speech data contain environmental noises as well, like the car noise, train noise and street noise, creating a mismatch with the clean train speech data. Due to this, applications requiring mobile phones to provide speech samples for identification do not work efficiently. In an attempt to solve this problem, a hybrid technique combining two frame selection techniques, namely, Voice Activity Detection (VAD) and Variable Frame Rate (VFR) method has been developed. This method has the following salient features:

- It selects the speech part rejecting the silence part from the speech signal.
- It efficiently captures the speech characteristics like vowels which last for a long duration and plosives which occur for short duration in the speech.
- It takes into account the signal to noise ratio.

The developed technique will be particularly useful for accessing remote applications/devices shared by many users. Here, speech for identification can be sent through mobile phones. Once the user is identified as an authorized person, he/she gains access to the applications/devices and can avail personalized services.

2. Few studies dealing with the mismatched condition which occur because of voice disguises are available as compared to the mismatched conditions due to environmental noises or voice recording device/channel variations. Voice disguise happens when a person intentionally modifies his/her voice either to hide his/her own identity or to sound like a target to steal the target's information/resources. This thesis therefore focuses on the mismatched problem arising out of voice disguises and at this end, the following contributions have been made:

- Conventionally, for feature extraction, a frame length of 25-30 ms with a fixed frame shift of half the frame length has been shown to produce the best identification accuracy and robustness against mismatched conditions [14], [15], [16]. But for voice disguise, the usage of this fixed frame shift, providing a fixed frame rate might not produce the best results. This can be due to the fact that speaking rate of the same person can differ. Therefore, different frame shifts, in the range of 1-10 ms keeping the frame length fixed to 25 ms, providing different frame rates has been investigated for

voice disguise. Based on the investigation results, a multi-frame rate based multiple-model method for training the speaker models has been developed, and it has shown improved performance over the conventional fixed frame rate method.

- Security conscious organizations require speaker identification system to identify the person who can adopt voice disguise for leaking sensitive information of the organization. To build speaker identification for this application, four variants of multistyle training have been investigated for voice disguise. A fusion technique, utilizing the two best investigated multistyle trainings out of the four at the decision level has been developed. The fusion technique has shown an overall better and more stable performance compared to the investigated multistyle trainings, multiple-model trainings and the conventional single style trainings.
 - Further, a method has been developed to measure the reliability of the test speech sample. Through this method, reliable frame selection of the test speech sample has been carried out. Finally, a method combining multiple frame rate for feature extraction and reliable frame selection at the decision level has been developed. The method has shown an overall better performance compared to the conventional method.
3. Lastly, as speech production is governed by human brain, information from brain signals may complement the speech features in increasing the robustness of the speaker identification system. With this vision in mind, a Brain Computer Interface (BCI) has also been studied in this thesis. It is used to classify motor imagery tasks from the same limb. To improve the performance of the system, a feature selection strategy consisting of time-segment selection through visual inspection and channel selection through Fisher ratio analysis in the frequency domain has been proposed.

Though in this initial study, we have not recorded brain signal activity for voice, future studies will aim this, and the fusion of the brain signal and the speech information might lead to improved speaker identification system.

1.6 Publication List

Journal

1. S. Prasad, Z.-H.Tan, R. Prasad, “Feature frame selection for robust speaker identification: A hybrid approach”, *Wireless Personal Communications*, pp.1-18, May 2017, Springer, DOI: 10.1007/s11277-017-4544-1. (Chapter 3)
2. S. Prasad, Z.-H.Tan, R. Prasad, “Multiple Frame Rates for Feature Extraction and Reliable Frame Selection at the Decision for Speaker Identification Under Voice Disguise”, *Journal of Communication, Navigation, Sensing and Services (CONASENSE)*, Vol. 2016, no.1, pp-29-44, Jan. 2016, DOI: 10.13052/jconasense2246-2120.2016.003. (Chapter 6)
3. S. Prasad, R. Prasad, “ Fusion Multistyle Training for Speaker Identification of Disguised Speech”, submitted to *Wireless Personal Communications*, Springer (Chapter 5)

Conference

1. S. Prasad, R. Prasad, “Reliable frame selection for robust speaker identification under voice disguise scenario”, *WirelessVITAE 2015*, Hyderabad, Dec. 2015. (Chapter 6)
2. S. Prasad, Z.-H.Tan, R. Prasad, “Multistyle training and fusion for speaker

identification of disguised voice”, *1st International Conference on Communications, Connectivity, Convergence, Content and Cooperation (IC5)*, Mumbai, India, Dec. 2013. (Chapter 5)

3. S. Prasad, Z.-H. Tan, R. Prasad, “Multi-frame rate based multiple-model training for robust speaker identification of disguised voice”, *16th Wireless Personal Multimedia Communications (WPMC)*, Atlantic City, New Jersey, U.S.A., Jun. 2013. IEEE Press. (Chapter 4)
4. S. Prasad, Z.-H. Tan, R. Prasad, A. R. Cabrera, Y. Gu, K. Dremstrup, “Feature selection strategy for classification of single trial EEG elicited by motor imagery”, *14th Wireless Personal Multimedia Communications (WPMC)*, Brest, France, Oct. 2011. IEEE Press. (Chapter 7)

1.7 Structure of the Thesis

The rest of the thesis is organized as follows. The next chapter discusses the state of the art in the speaker identification research focusing the mismatched problem. The Hybrid feature frame selection method, which combines the VFR and the VAD methods is presented in chapter 3, along with the experimental evaluations under various environmental noise scenarios. The multi-frame rate based multiple model for speaker identification of disguised speech is described in chapter 4. Chapter 5 discusses the different variants of the multistyle training strategies and the fusion technique, which is developed based on it, for the speaker identification under voice disguise scenario. The method to measure the reliability of the test speech sample and based on which, the multiple frame rate feature extraction and reliable frame selection at the decision level is discussed in chapter 6. BCI for the classification of motor imagery tasks from the same limb is discussed in chapter 7. Finally, chapter 8 concludes the thesis with a discussion on future work.

Chapter 2

State of the Art

2.1 Mismatched Problem

Speaker identification system is predicted to have huge importance in the coming future. With humans inclining more and more towards electronic equipments for their daily activities, a need for secure and easier method of interaction with these equipments is foreseen. For humans, speech is the most natural means of communication and can be easily provided for identification, as compared to other modalities like face or iris. Some communities, like, Muslim women are supposed to cover their face with a cloth as religious belief and might not be comfortable providing their face for recognition. Iris requires use of laser for capturing the image, which can have some health risks. Since, telephony system is already in use, the implementation of the speaker identification system for secure exchange of information from a distance can also be much easily achieved. Therefore, biometric based on voice for secured communication with the machines will be most likely preferred over other biometrics like finger, face and iris based biometric.

Though research in speaker identification started long back, it still faces poor performance because of many factors. One of the most common problem seen in the speaker identification research is mismatched problem. When the training and the test conditions are similar, it is referred to as matched conditions. Under matched conditions, speaker identification system is observed to have really good performance. In [17], temporal variations of pitch in the speech was used for speaker recognition, 97% identification accuracy was found. But when a dissimilarity between the training and the test conditions are seen, referred to as mismatched problem, the performance of the speaker identification system significantly decreases. One example of the mismatched condition can be seen, when the speaker models are trained using speech data recorded in a quite environment, whereas, the test data contains environmental noises, like, train noise, car noise and street noise as well [3, 13].

2.1.1 Factors Causing the Mismatch:

Mismatched condition can occur because of various factors. Some of them are given below [9]:

1. Speaker based mismatch: The same person can speak a particular word/sentence in different ways depending upon the situation. It is also called as mismatch due to intra/within speaker variability. Following can be the reason for having differences in voice from the same person:
 - When a person is doing some stress related work, like riding a heavy vehicle such as aeroplane, truck or train [18, 19].
 - The emotional state of the person like anger, grief, excitement [20, 21].
 - Speaking louder than normal in the presence of background noise, also known as Lombard effect [19, 22].

- When the person intentionally alters his/her voice to hide their own identity or to sound like a target to steal the target’s resources, called voice disguising [23]. Electronics devices were also used for voice disguising. A typical example is a Voice Changer software [24]
 - When the person is suffering from some illness, like throat infection or due to ageing [25].
2. Environment based mismatch: This mismatch occurs because of background noise like train, car and street noise [3, 13], reverberation [26] and because of microphones which are placed at a distance [27].
 3. Voice recording device based mismatch: This mismatch occurs because of differences in handset (mobile phones, landlines, cordless phone), transmission channel and microphone [28, 29].

2.1.2 Addressing the Mismatched Problem

Early speaker identification research focused on the telephone handset or communication channel based mismatch because during this time the telephone system were mostly fixed and not movable. But with the invention of mobile phones and smart phones, the focus of the mismatch has changed to the environmental/background noise. With the advancement of the technology, many other types of mismatch occurred, listed in the subsection 2.1.1, because of which the performance of the speaker identification system suffered. Several robust methods have been tried to mitigate the mismatch arising due to these factors. The robust methods can be applied at the different levels of the speaker identification system, namely, Feature level, Speaker model level (Training) and the Score/Decision (Testing) level.

- Feature level: In this, the method directly works on the speech signal. A close match between the train and the test speech conditions are tar-

geted. It tries to select and extract robust features with respect to channel/background variations. .

It is desirable to remove the non-speech part from the speech signal called the Voice Activity Detection (VAD), as it carries no speaker specific information and in the presence of noise degrades the identification accuracy markedly. To achieve this, identification of speech and silence part in the signal is required. Several algorithms have been developed for this, which were mostly based on the amplitude level, short term energy and zero crossing rate of the speech signal [30]. In [31], a Gaussian statistical model based VAD is presented, which employed decision directed parameter estimation method for likelihood ratio test. A VAD which is based, not only on Gaussian model, but also on complex Laplacian and Gamma probability density function has been utilized in [32] for improving the performance. In [33], power spectral deviation utilizing Teager energy has been used for the VAD. It has shown better performance over conventional methods in various noisy environment. A VAD based on the long-term pitch divergence, in which bionic decomposition of the speech signal is done, is presented in [34]. It has shown better performance over 6 analyzed VAD algorithms.

Robust features were also developed. Linear Prediction Cepstral Coefficient (LPCC) [35], which models the human voice production organ became very popular. A modified LPC and wavelet transform combination based speech feature has been tried in [36] and has shown good performance for environmental noise. It has been observed that features which mimic the human hearing system can prove to be more robust. The Mel-Frequency Cepstral Coefficient (MFCC) showed superior performance over various other feature in [37] for speech recognition. A multitaper MFCC which is based on multiple time domain windows with frequency domain averaging has shown good performance in [38] for speaker verification. An auditory based feature extraction method

inspired by the traveling waves in the cochlea [39] and a Karhunen-Loeve transform (KLT) [40] based robust speaker identification system has also been developed. In [41], a binary quantization of the feature vector is carried out in order to increase the robustness of the speaker recognition system under noisy condition. A binary time-frequency (T-F) mask is used in [42], to provide information about whether the noise is stronger than the speaker characteristics in the T-F unit under observation and it has shown good performance under additive noise condition. A fusion of MFCC and statistical pH features was proposed in [43] for speaker verification under environmental noise. A novel feature called the Power Normalized Cepstral Coefficients (PNCC) has been presented in [44] for speaker recognition under noisy environments. This feature uses power-law nonlinearity instead of log nonlinearity which is used in MFCC, and a noise-suppression algorithm which depends on asymmetric filtering. In addition a temporal masking module is also included. It has shown better recognition rates over MFCC and RASTA-PLP. A fusion of Subglottal Resonances (SGRs) and cepstral features was proposed in [45]. It showed that, SGRs can be used as complimentary features with the noise robust PNCC and LPCC features for improving the efficiency of speaker identification under noisy environment. Since, cepstral features have been shown to produce a very high speaker identification accuracy, therefore various normalizations techniques like cepstral mean subtraction [46], cepstral mean and variance [47], relative spectral (RASTA) [48] and feature warping [49] have been applied to these feature for mismatch compensation.

- Speaker model level: In this level, robust methods are developed to model the speaker using the features obtained from the speech signal. Gaussian Mixture Model (GMM) has shown a good performance in [50]. A Gaussian Mixture Model Universal Background Model (GMM-UBM) has been used in [51] for text-independent speaker verification. Here, a background model using a large GMM with 2048 mixtures was developed utilizing the

speech data from 90 persons (45 females and 45 males), and it was then used for all the claimants. This model was evaluated for 1996 NIST speaker recognition evaluation corpus [52]. It has been observed that, a system which used GMM-UBM with Bayesian adaptation of claimant model has produced superior results over UBM with claimant model which was not adapted from the UBM. A boosted slice classifier has been introduced in [53] for robust speaker verification task. A transformation algorithm has been developed in [54] for transforming the speaker model in order to be more robust to acoustic mismatch. It utilizes locally collected stereo database and basically increases the variances of the speaker model by a limit for this purpose. A joint factor analysis (JFA) approach was presented in [55] in order to deal with the session variability. An i-vector framework has been proposed in [56], which combines the channel and the speaker into a single space called the total variability space. Support Vector Machine (SVM) utilizing Gaussian supervectors has also been widely used [57]. A colored noise based multi-condition training technique is proposed in [58], in which noisy speech data is generated using white Gaussian sequence and is then used in training speaker models in order to handle the unknown environmental noise in testing condition. In [59], a universal compensation technique is proposed, which is developed combining the multi-conditioning training and missing feature method to study and achieve the robust speaker recognition under noisy condition. A novel speaker binary key representation space is proposed in [60], in order to make the system adaptable to different environmental condition. Deep neural network which has shown good performance for speech recognition [61] has also been investigated for speaker recognition in [62].

- **Score/ Decision level:** In this level, an utterance called the test utterance from an unknown speaker will be given. A comparison of this utterance with all the speaker models developed in the training phase will be car-

ried out for similarity. The model scoring the highest similarity measure will be the speaker of the given test utterance. Score compensation technique like H-norm, T-norm and Z-norm [3] are attempted in order to address the channel effects under different condition. A special kind of weighting model rank is proposed in [63] to increase the robustness of the system. H-norm and T-norm combination technique HT norm [64] is also used in order to handle mismatched condition during training and testing phase. Several combination techniques from the decision obtained from two classifiers has been investigated in [65]. The two classifiers used differ in their feature set. One utilized the MFCC feature and the other used a new Parametric Feature Set (PFS). A fast scoring algorithm and Advanced Missing Feature Theory (AMFT) has been developed in [66] in order to handle various background noises. In [67], speech recordings from Supreme Court of the United States (SCOTUS corpus) were used for identification under reverberant condition. A 100% identification accuracy has been reported for 1 sec speech data using a combination of Gaussian mixture model and monophone Hidden Markov Model (HMM).

A large number of studies have been carried out relating to channel and environment related mismatch, which is discussed above. Relating to other factors causing mismatch, few studies have been carried out . A mismatch due to shouted and neutral speech has been studied in [68]. The identification accuracy decreased from 100% (matched) to 8.71% (mismatched). To tackle this, MFCC feature compensation through GMM mapping method has been done. Another mismatched condition because of speaking under stress has been reported in [69], it also showed significant decrease in speaker identification accuracy. One of the factor, i.e, voice disguise can pose a serious threat to speaker identification research in future. Some of the research addressing voice disguise have been done [70]- [73], but it needs to be focused, and is one of the goal of this thesis work.

Research studies utilizing brain signals for person identification has also been done [74]. Studies in neuroimaging [75] revealed that, different cortical regions were activated for processing different types of vocal information, like speaker identity and language information. Therefore, brain signals captured while the person is speaking or imagining speaker, can be used for identifying a person. In [76], brain signals from the subjects were captured, when they imagined speaking two syllables /ba and/ku and is then utilized for training and identification task. It has reported an accuracy of 99.76%. Inspired by this, a study on Brain Computer Interface for classifying motor imagery tasks has also been carried out in this thesis. So that in future, the fusion of the speech signals and the brain signals can be investigated for improving the speaker identification under mismatched scenario.

Chapter 3

Feature Frame selection - A Hybrid Approach for Environmental Noise Corrupted Speech¹

3.1 Introduction

In chapter 2, it has been discussed that speaker identification systems face marked degradation in performance under mismatched conditions arising out of many factors. One commonly seen mismatch scenario is when speaker models were built from clean speech data and the test speech data were corrupted

¹This chapter is based on the following article: S. Prasad, Z.-H.Tan, R. Prasad “Feature frame selection for robust speaker identification: A hybrid approach.”, which is submitted to the journal *Wireless Personal Communications*.

with environmental acoustic noises as well. In this chapter, we will be focusing on this particular type of mismatch scenario arising due to environmental/background acoustic noises.

Speech signals are non-stationary signals. They are cut into smaller segments called frames, typically in the range of 25-30 ms long size with a frame shift of half the frame size in which they exhibit quasi-stationary behaviour for processing. But this method of frame making from the speech signal, especially when environmental noises are present is inefficient. Frame making should take into account the following points in case of mismatch due to environmental noise:

1. Speech signal contains both fast changing and steady state regions. Fast changing regions such as plosives occur for a very short period and therefore, more frames are needed from these regions to capture its characteristics properly. On the other hand, steady state regions, such as, vowel occur for a longer duration and therefore less number of frames are needed from these regions, so that extra addition of same type of speech characteristics can be avoided.
2. Speech signal also contains both speech and non-speech regions. The effect of frame making from the non-speech part when dealing with clean speech is not severe and may have a positive effect on the speaker identification accuracy. But, in the presence of environmental noise, non-speech part may greatly decrease the speaker identification accuracy. Therefore, non-speech part should be avoided for frame making.
3. Depending on the signal-to-noise ratio, a frame of the speech signal can be termed as reliable or unreliable. The unreliable frames should be discarded.
4. The average total number of frames made per second (frame rate) should be approximately equal to the traditional method, otherwise it will re-

quire more storage space and take more time for processing, which might not be suitable for some applications.

A frame selection method based on the average frame energy above a threshold has been done in [77]. A feature frame selection method [78], in which minimum redundancy between selected frames but maximum relevance to speaker models were targeted. In [79], a distance metric based frame selection, which is performed on the spectrum of the speech instead of the time-domain has been done. The norm of delta-MFCC vector has been used in [80] for detecting spectral changes like steady state regions and transient regions. Voice activity detection (VAD) for separating speech and non speech part has been done in [81] using a novel likelihood ratio sign test which took into account long term speech information. For improving VAD, spectral subtraction speech enhancement is applied before energy based VAD in [82]. A feature frame selection based on the weights assigned by two models: one from speech and the other from noise has been presented in [83]. These studies took into account the above referred points either individually or in combination but the joint study of all the above four referred points for the speaker identification application under environmental noise is not seen. Therefore, this chapter attempts to study the speaker identification system under environmental noise scenario by taking into account all the above four referred points.

A variable frame rate (VFR) analysis method in which frame selection depended on speech signal characteristics and was also based on signal to noise ratio (SNR) weighted energy distance method has shown good performance for speech recognition [84]. In speech recognition, instead of identifying the speaker of the utterance, the words in the utterance are determined. This chapter investigates this SNR weighted energy distance based VFR method [84] for the speaker identification application under environmental noise. Further, this method has been combined with the statistical model based VAD [31] method for proposing the hybrid frame selection method for speaker identification un-

der environmental noise.

The rest of the chapter is organized as follows. The next section describes the hybrid feature frame selection method. Section 3.3 presents the experiments conducted and discusses the results. Finally, Section 3.4 presents the summary of the chapter.

3.2 Hybrid Feature Frame Selection

The hybrid feature frame selection method simply combines the frames selected by two methods for feature extraction. The first method called the VFR analysis method captures the frames based on the changes in the temporal characteristics of the speech signal. It also takes into account the signal to noise ratio during frame selection for the reliability measurement. The second method called the VAD method selects only the speech part of the signal for frame making, rejecting the non-speech part of the signal. Fig. 3.1 shows the speaker identification system using the hybrid frame selection method for feature extraction.

The VFR and the VAD methods are briefly presented below:

- VFR Based Frame Selection

The VFR analysis method selects frames according to the changes in the temporal characteristics of the speech signal. In this, dense frames are first created by using the fixed frame rate (FFR) method with a frame size of 25ms and a very small frame shift of 1ms. Distances between two adjacent frames are then calculated as the difference in energy. To measure the reliability of the frame, these distances are further weighted by the signal to noise ratio. Finally, frames with accumulated SNR weighted energy distance above a particular threshold will be selected and the rest

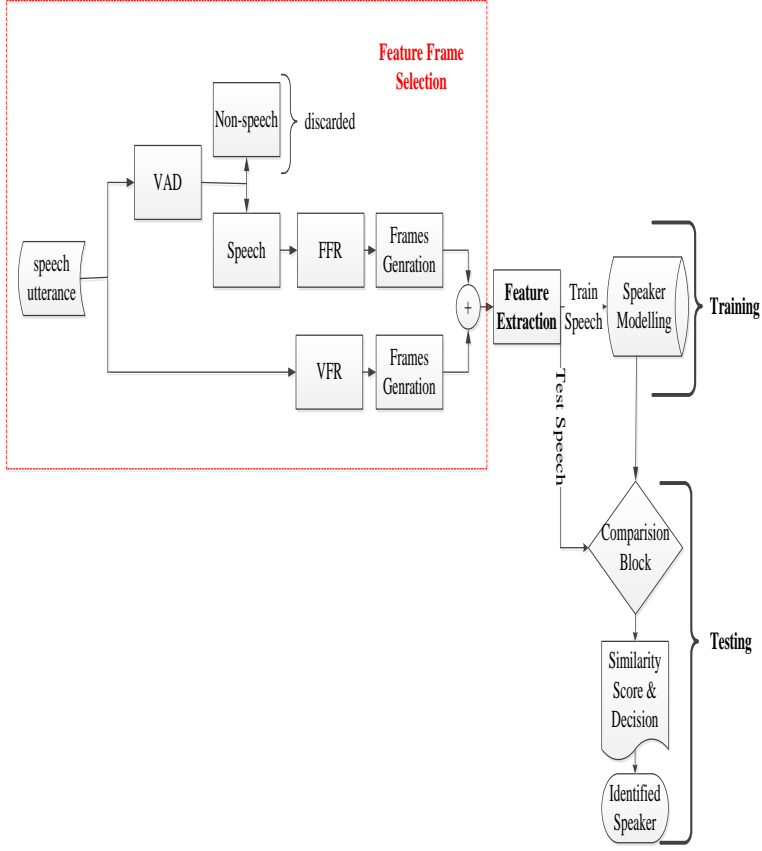


Figure 3.1: Frame selection by the hybrid method [85].

will be discarded. This process created more number of frames for fast changing regions and fewer frames for steady state speech regions.

The low-complexity VFR frame selection method [84, 85] can be implemented using the following steps:

1. Dense frames of the speech signal are created using the FFR method by keeping the frame size as 25ms and frame shift as 1ms.
2. The *a posteriori* SNR weighted energy distance of the frames are

calculated as:

$$D_{SNR}(t) = |\log E(t) - \log E(t-1)| \times SNR_{post}(t) \quad (3.1)$$

where, $\log E(t)$ is the logarithmic energy of frame t , $SNR_{post}(t)$ is the *a posteriori* SNR value of the frame t and is given by:

$$SNR_{post}(t) = \log \frac{E(t)}{E_{noise}} \quad (3.2)$$

Calculation of the *a posteriori* SNR value of the frame t is simpler than the *a priori* SNR value.

A priori SNR value is given by:

$$SNR_{prio}(t) = \log \frac{E_{speech}(t)}{E_{noise}} \quad (3.3)$$

which requires an additional step of speech estimation from noisy speech. In contrast, the *a posteriori* SNR value calculation did not have to do this step as it utilizes the energy of the noisy speech directly, making it less complex. Here, E_{noise} is the noise energy of the frame t and is considered same for all frames. It is estimated by taking the average energy of the initial 10 frames which are assumed to be non-speech only and it approximately corresponds to 34ms of the utterance.

3. The threshold function is calculated as :

$$T = \overline{D_{SNR}(t)} \times f(\log E_{noise}) \quad (3.4)$$

where, $\overline{D_{SNR}(t)}$ is the average of the weighted energy distances calculated in Step 2) over the whole utterance. The function f is the sigmoidal function, given by:

$$f(\log E_{noise}) = A + \frac{B}{1 + e^{-2(\log E_{noise} - 13)}} \quad (3.5)$$

The constant value of 13 is used to make the turning point of the sigmoid at an *a posteriori* SNR value between 15 and 20 dB. The parameters A and B determine the frame rate and its value calculation is discussed in subsection 3.3.2.

4. In this step, frames are finally selected or rejected based on an updated accumulative distance. *A posteriori* SNR weighted energy distances are accumulated as:

$$Acc(i) = Acc(i - 1) + D_{SNR}(i) \quad (3.6)$$

for frames $i=1,2,3,\dots$. The accumulative distance $Acc(t)$ of frame t is compared with the threshold T : if $Acc(t) > T$, frame t is selected and Acc value is reset. The calculation of accumulated distance is again started from frame $i=t+1$. This process is continued until a decision on all the frames have been made.

- VAD Based Frame Selection

Gaussian statistical model based VAD [31,85] is used to select the active speech part of the utterance. Here, Likelihood Ratio Test (LRT) is used for decision making of speech and non-speech part.

1. A binary hypothesis can be made assuming additive noise as follows:

$$H0: Z(t) = N(t) \quad (\text{Speech absence})$$

$$H1: Z(t) = S(t) + N(t) \quad (\text{Speech presence})$$

$Z(t)$, $N(t)$ and $S(t)$ represents the noisy speech, noise, and speech, respectively at frame t and is given by the following k -dimensional Discrete Fourier Transform (DFT) coefficients :

$$\begin{aligned} Z(t) &= [Z_0(t), Z_1(t), \dots, Z_{k-1}(t)]^T, \\ N(t) &= [N_0(t), N_1(t), \dots, N_{k-1}(t)]^T \quad \text{and} \\ S(t) &= [S_0(t), S_1(t), \dots, S_{k-1}(t)]^T \end{aligned}$$

2. The probability density function conditioned on $H0$ and $H1$ are given as follows, when $Z(t)$, $N(t)$, and $S(t)$ are considered as asymptotically independent Gaussian random variables.

$$p(Z(t) | H0) = \prod_{j=0}^{k-1} \frac{1}{\pi \lambda_{n,j}} \exp \left(- \frac{|Z_j(t)|^2}{\lambda_{n,j}} \right) \quad (3.7)$$

$$p(Z(t) | H1) = \prod_{j=0}^{k-1} \frac{1}{\pi(\lambda_{n,j} + \lambda_{s,j})} \exp \left(- \frac{|Z_j(t)|^2}{(\lambda_{n,j} + \lambda_{s,j})} \right) \quad (3.8)$$

Here, $\lambda_{n,j}$ and $\lambda_{s,j}$ represents the variances of N_j and S_j , respectively.

3. The likelihood ratio for the j th frequency bin is:

$$\Lambda_j \equiv \frac{p(Z_j(t) | H1)}{p(Z_j(t) | H0)} = \frac{1}{1 + \xi_j} \exp \left(\frac{\gamma_j \xi_j}{1 + \xi_j} \right) \quad (3.9)$$

where, $\xi_j \equiv \frac{\lambda_{s,j}}{\lambda_{n,j}}$ represents the *a priori* signal-to-noise ratio and is estimated by the decision directed method [31]. $\gamma_j \equiv \frac{|Z_j(t)|^2}{\lambda_{n,j}}$ represents the *a posteriori* signal-to-noise ratio.

4. The final decision about presence and absence of speech is determined by the geometric mean of the individual frequency bands,

$$\log \Lambda = \frac{1}{k} \sum_{j=0}^{k-1} \log \Lambda_j \geq_{H0}^{H1} \eta \quad (3.10)$$

where, η represents a preset threshold.

Lastly, Speech frames are made using the active speech part selected by the VAD. The FFR method is used for frame making and it utilizes a frame size of 25ms with a frame shift of 10ms.

Frame selection by the VFR, VAD and the proposed hybrid technique are shown in Figures 3.2 to 3.4 for the clean speech utterance, babble noise corrupted speech utterance and the car noise corrupted speech utterance, respectively. In the Figures, the first panel shows the time-domain waveform of the speech utterance in which the decision about the active speech part and the non-speech part made by the VAD is shown as a pulsed waveform. In the pulsed waveform, the “0 level” shows the non-speech part and the “above 0

level” shows the active speech part. Panel 2 shows the wideband spectrogram of the speech utterance. Panel 3 shows the frame selection from the speech utterance utilizing the FFR method in which 25ms frame size with 10ms frame shift is utilized. The vertical lines represent frame selection at a particular instant. Panel 4 depicts the frame selection from only the active speech part selected by the VAD as shown in Panel 1. For frame selection FFR method is utilized and a frame size of 25 ms with a frame shift of 10ms is used. Panel 5 shows the frame selection by the VFR method and Panel 6 shows the frame selection by the hybrid technique (referred to as Proposed). The speech utterance used for these figures is the utterance of the number “73” i.e. “seventy three”.

It can be observed from the figures that, instead of selecting frames at a fixed frame rate from the whole utterance (Panel 3), VAD selects frames at a fixed rate but only from the active speech part of the speech signal (Panel 4). VFR (Panel 5) tries to capture more frames at the transient regions, fewer frames at the steady state regions and no frames at the non-speech regions. Observing Fig. 3.3, it can be seen that VFR performed better than the VAD in rejecting the non speech part before the start of the babble noise corrupted utterance. For the car noise corrupted speech utterance (Fig. 3.4), VAD performed better than the VFR in capturing the speech part. Therefore, combining the frames selected by the VAD and the VFR method as in hybrid technique may improve speaker identification accuracy under environmental noise condition, as it will be adding up the different and complementary characteristics of the individual methods.

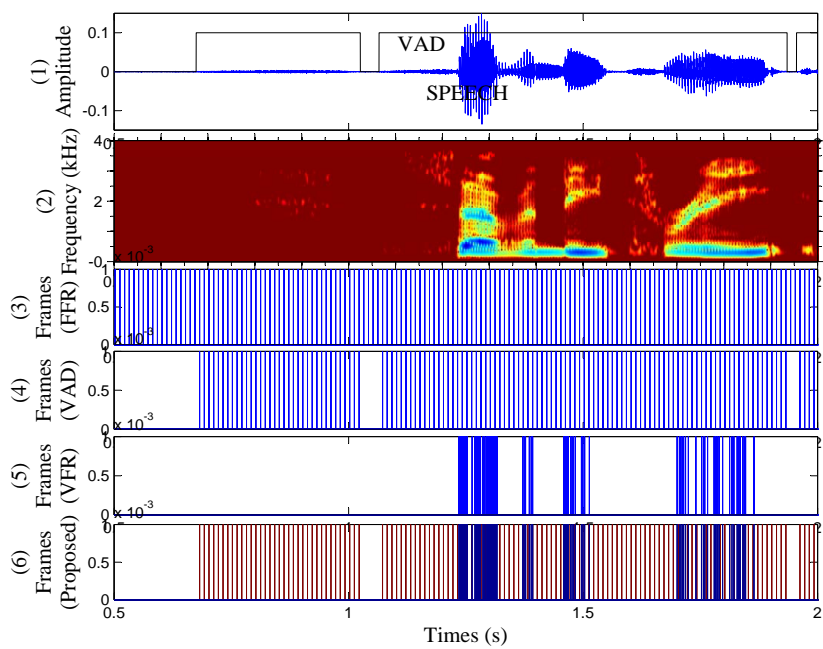
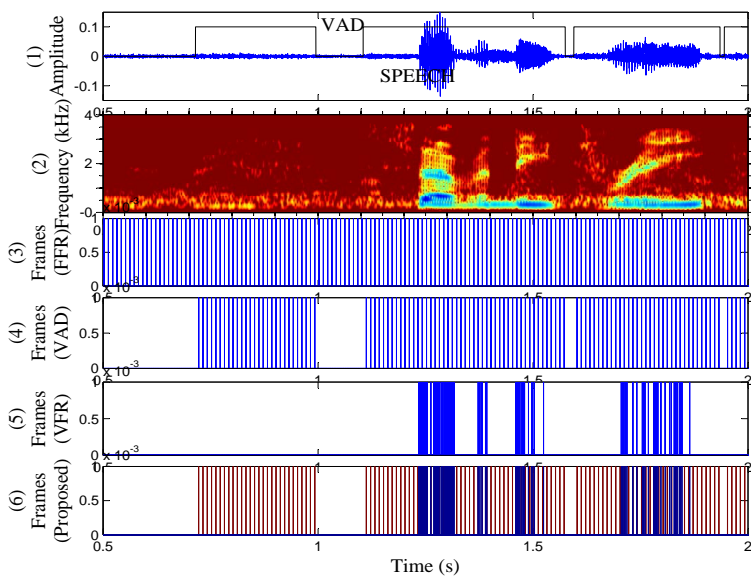
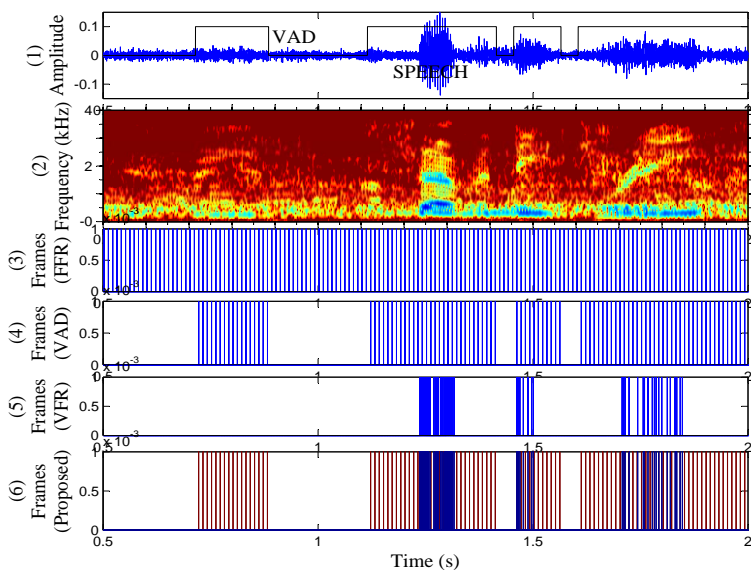


Figure 3.2: Clean speech utterance frame selection [85].

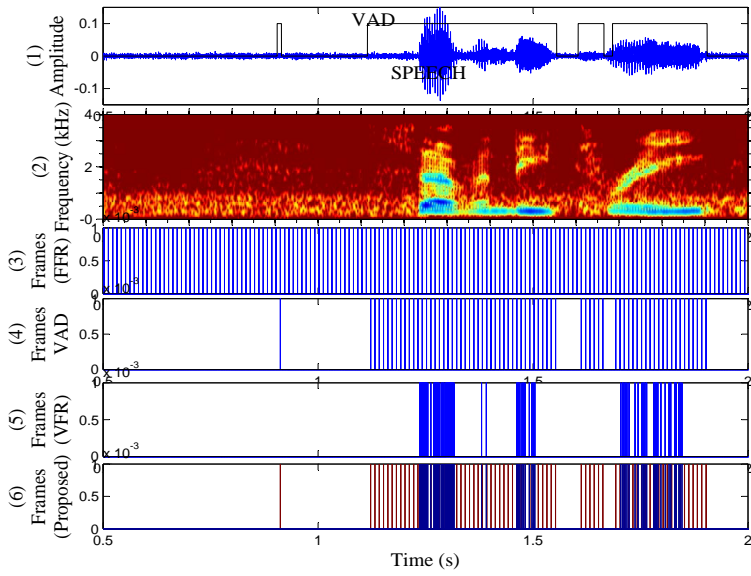


(a) 15dB SNR

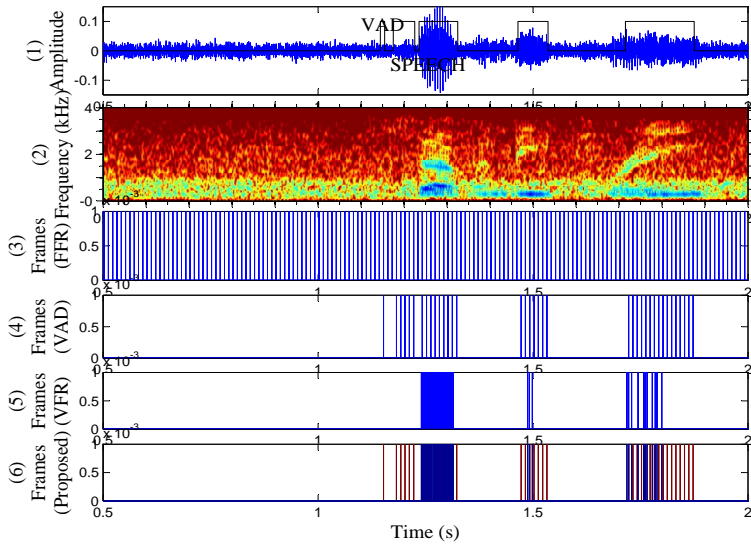


(b) 5dB SNR

Figure 3.3: Babble noise corrupted speech utterance frame selection at SNR of 15dB and 5dB [85].



(a) 15dB SNR



(b) 5dB SNR

Figure 3.4: Car noise corrupted speech utterance frame selection at SNR of 15dB and 5dB [85].

3.3 Experiments and Results

The database used for conducting the different experiments has been briefly described in subsection 3.3.1. The value of the parameter A and B used in the threshold function (Equation 3.5), which decides the frame rate in the VFR frame selection method needs to be calculated. This is carried out in subsection 3.3.2. Finally, subsection 3.3.3 presents the different speaker identification experiments conducted for evaluating the hybrid feature frame selection method and discusses the results obtained.

3.3.1 Database

YOHO database [86] and Aurora II database [87] were used to build the noisy and clean YOHO database for conducting the various experiments of this chapter.

YOHO database consisted of speech recordings in a quiet office environment from 138 speakers (106 males and 32 females). Though some office noise was present but it is considered as Clean YOHO speech only. Eight different types of environmental noise, namely, Babble, Exhibition, Restaurant, Airport, Car, Street, Subway and Train were taken from the Aurora II database and is artificially added to the clean YOHO speech at four different SNRs of 5dB, 10dB, 15dB and 20dB for generating the noisy YOHO database. For addition of the noise to the Clean YOHO speech, an equal length signal is randomly cut from the noise signal (assumed to be very large in comparison to the Clean YOHO speech utterance) and is then added to the speech utterance at the desired SNR. For doing this task FaNT software is utilized [88].

In YOHO database, the train and the test speech data were collected separately in three month's time period. For the collection of training data, 4 recording sessions were conducted. Each session collected 24 speech utterances

(approx. 5s long) per speaker. Therefore, for training each speaker model, 96 utterances were used making a total of 480s of speech. For test data collection, 10 recording sessions were conducted, collecting 4 unseen utterances from each speaker. Therefore, a total of 40 utterances per speaker were collected for testing. For evaluating the different speaker identification systems, 5520 test utterances of approx. 5s length from all the speakers were utilized.

3.3.2 Average frame rate calculations

The number of frames selected per second (frame rate) by the VFR analysis method can be controlled by varying the parameters A and B used in the sigmoid function for the calculation of the threshold value (Equation 3.5). To understand the effects of the parameters A and B corresponding to different average frame rate on the speaker identification accuracy, different values of the parameters A and B have been chosen as shown in Table 3.1. Using these A and

Table 3.1: Parameter A & B selection used in the threshold value for average frame rate calculation in the VFR method [85]

Parameters		Average frame rate (Hz)
A	B	
12	2.5	50
9	2.5	60
7	2	70
5	2	80
4	1.5	90
3	1.5	100

Table 3.2: Identification accuracies (%) of the VFR method at different average frame rates for clean and noisy speech [85]

Noise	SNR (dB)	Average frame rate(Hz)					
		100Hz	90Hz	80Hz	70Hz	60Hz	50Hz
		A=3, B=1.5	A=4, B=1.5	A=5, B=2	A=7, B=2	A=9, B=2.5	A=12, B=2.5
Clean	—	97.98	98.21	98.33	97.86	97.98	98.33
Babble	20	96.31	95.83	96.31	96.55	96.67	95.95
	15	92.86	92.74	93.21	93.21	93.45	92.98
	10	80.36	81.55	81.79	82.86	81.43	82.50
	5	55.36	56.19	52.62	57.98	59.40	56.55
Car	20	94.52	94.17	94.52	94.17	93.69	93.81
	15	84.88	85.60	85.83	85.24	85.95	84.17
	10	63.69	64.05	63.45	63.93	65.36	64.88
	5	40.95	41.43	37.36	41.07	41.79	38.10
<i>Average</i>		78.55	78.86	78.14	79.20	79.52	78.59

B values, speaker identification experiments were conducted on a smaller set of 21 speakers of the YOHO database for Clean, Babble and Car noise corrupted test speech data. The identification accuracies obtained are tabulated in Table 3.2. From the table, it can be observed that, varying the parameters A and B did not result in a significant change in the average speaker identification accuracy across the clean and noisy test speech. The value of A=9 and B= 2.5 gave the best average identification accuracy across the Clean and the noisy test speech. Since, the aim here is to get the highest speaker identification accuracy, therefore, for all further experiments involving VFR analysis method,

either individually or in hybrid, the parameter value for A and B will be chosen as 9 and 2.5, respectively, corresponding to 60 Hz frame rate. The proposed hybrid frame selection method combines the frames selected by the VFR and the VAD method. With the parameter value of A=9 and B=2.5 corresponding to 60 Hz frame rate of the VFR analysis method, average frame rate of the proposed hybrid frame selection method was found out to be approx. 110 Hz. The value of 110 Hz is considered optimal, which is near to the conventional FFR analysis of 100Hz. Therefore, almost same storage space will be required.

3.3.3 Speaker Identification Experiments & Results Discussion

The *a posteriori* SNR weighted energy distance based VFR method [84], [89] has shown better performance than other VFR methods [90], [91] for the speech recognition application. Therefore, in this chapter, *a posteriori* SNR weighted energy distance based VFR method has been investigated for the speaker identification experiment and is also taken as one of the baselines. It is called as Bsln-VFR. The standard Gaussian statistical model based VAD [31] is taken as the second baseline system. It is called as Bsln-VAD. Apart from these two the conventional FFR analysis method, which employs no robustness method has also been included for the performance comparison. It is called as Bsln-no robustness method. Speaker identification experiments were conducted for the total 138 speakers for the Bsln-VFR, Bsln-VAD, Bsln-no robustness and the hybrid feature frame selection method (called Proposed HVV), for the clean and noisy YOHO test data at 4 different SNRs. Twelve MFCC excluding the 0^{th} coefficient were extracted from the frames as features and 64 components Gaussian mixture model were used for speaker modeling. The model which gave the highest likelihood measure of the test utterance has been decided as the speaker of the test utterance.

The identification accuracies obtained for the Bsln-no robustness, Bsln-VFR, Bsln-VAD and the Proposed HVV method under various noise scenarios and clean condition are tabulated in Table 3.3.

From the table, it can be observed that the Proposed HVV method has shown better performance than the baseline methods for all noise scenarios across different SNRs except for the Babble noise corrupted speech at 5dB SNR. When the average of the identification accuracies for the different noise scenarios have been taken, the Proposed HVV showed an absolute improvement of 5.54% and 9.50% and a relative improvement of 9.79% and 18.05% from the Bsln-VFR and Bsln-VAD method, respectively. For the SNR of 5dB, the Proposed HVV has shown an absolute improvement of 6.45% and 5.47% and a relative improvement of 40.29% and 20% for the Street and Train noise, respectively over the Bsln-VFR method. Proposed HVV has also achieved a good performance over the Bsln-VAD at 5dB SNR for the Babble, Restaurant and Airport noise. It has shown an absolute improvement of 20.53%, 16.79% and 18.45% and a relative improvement of 127.28%, 102.69% and 81.35% for the Babble, Restaurant and Airport noise, respectively over the Bsln-VAD.

It can be observed that, for Clean condition, Bsln-VAD has performed better than the Proposed HVV. It can also be observed that, Proposed HVV has performed better than the Bsln-VFR for Clean condition. Since, Proposed HVV combines Bsln-VAD and Bsln-VFR methods, it can be concluded that, under clean condition, the effects of Bsln-VFR is more on the Proposed HVV as compared to the Bsln-VAD. However, under noisy conditions Proposed HVV performed better than both Bsln-VAD and Bsln-VFR. Figure 3.5 shows comparisons between the different methods for various noise scenarios considered in this chapter. The graph is obtained by averaging the identification accuracies across the four SNRs for a particular noise type. It also confirms the better performance of the Proposed HVV method over the Bsln-VFR and the Bsln-VAD methods.

Table 3.3: Identification accuracies (%) obtained for the various methods under clean & noisy conditions [85]

Noise	SNR (dB)	Bsln-no ro- bustness	Bsln- VAD	Bsln- VFR	Proposed HVV
Babble	20	67.71	88.81	88.25	91.91
	15	46.63	73.51	81.91	86
	10	22.79	46.86	65.05	70.61
	5	6.35	16.13	36.82	36.66
average		35.87	56.33	68.01	71.29
Exhibition	20	37.13	78.25	75.91	84.16
	15	17.71	54.19	55.64	65.25
	10	7.78	26.65	26.63	32.72
	5	3.05	9.48	7.81	10.71
average		16.41	42.14	41.5	48.26
Restaurant	20	70.59	88.74	88.51	91.79
	15	50.38	72.34	82.12	85.97
	10	24.6	45.45	63.67	69.76
	5	6.76	16.35	29.35	33.14
average		38.08	55.72	65.93	70.17
Airport	20	60.66	88.74	88.67	91.82
	15	39.49	75.69	82.04	86.64
	10	18.91	52.79	67.31	72.44
	5	6.64	22.68	37.13	41.13
average		31.43	59.98	68.88	73.01
Car	20	44.53	87.27	84.3	89.14
	15	26.11	73.06	70.07	78.57
	10	11.71	49.77	46.37	56.11
	5	3.44	22.62	21.08	26.76
average		21.45	58.18	55.46	62.65
Street	20	46.12	82.23	79.08	86.46
	15	27.5	63.71	64.96	73.93
	10	12.78	39.99	40.48	51.16
	5	5.02	18.33	16.01	22.46
average		22.86	51.07	50.13	58.5
Subway	20	29.15	74.6	73.77	80.98
	15	11.86	48.95	53.77	60.77
	10	4.21	20.74	26.63	30.82
	5	1.12	7.32	8.56	9.92
average		11.78	37.9	40.69	45.63
Train	20	52.96	86.55	86.68	90.52
	15	30.75	73.48	77.27	82.81
	10	12.69	52.62	57.43	64.3
	5	3.83	26.42	27.36	32.83
average		25.07	59.77	62.19	67.62
Total Average		25.34	52.64	56.6	62.14
Clean		91.32	96.74	91.78	94.98

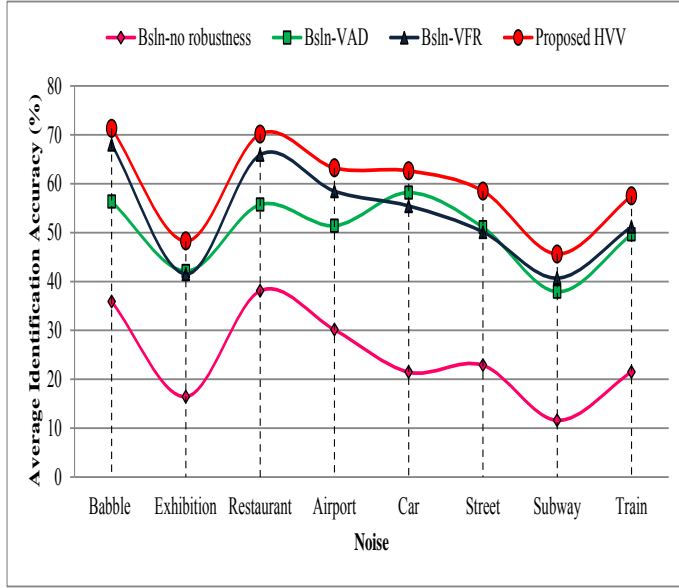


Figure 3.5: Comparison of the proposed and baseline method’s average identification accuracies under various noise scenarios [85]

Table 3.4: Proposed and the baseline methods average identification accuracies (%) at different SNR values [85]

SNR (dB)	Bsln-no robustness	Bsln- VAD	Bsln- VFR	Proposed HVV
20	51.11	84.4	83.15	88.35
15	31.3	66.87	71.03	77.49
10	14.43	41.86	49.2	55.99
5	4.53	17.42	23.02	26.72

Figure 3.6 shows the comparisons of the different methods of frame selection at the four SNR values. The graphs are obtained by averaging the identification accuracies across the different noise scenarios for a particular SNR value and is also tabulated in Table 3.4. Here also, Proposed HVV method performed

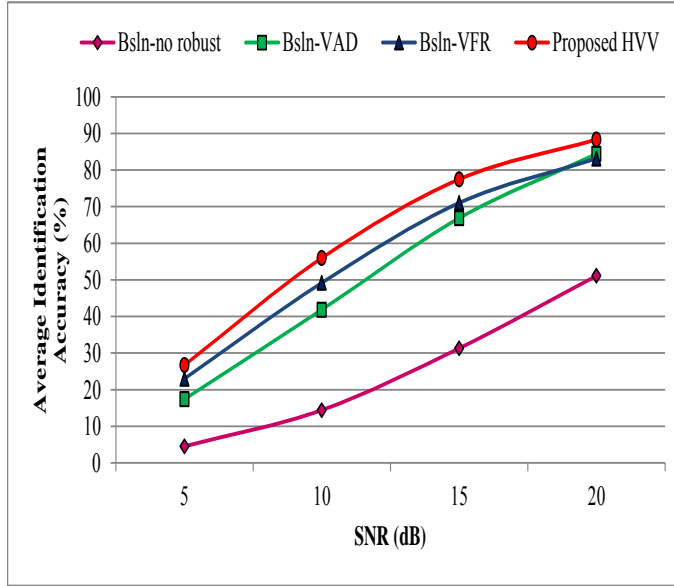


Figure 3.6: Comparison of the proposed and the baseline method's average identification accuracies (%) at different SNR values [85]

better than the other methods. For the SNR of 5dB, Proposed HVV obtained an absolute improvement of 3.7% and 16.1% and a relative improvement of 9.3% and 53.34% over the Bsln-VFR and Bsln-VAD method, respectively.

3.4 Summary

This chapter proposes a hybrid feature frame selection technique for speaker identification. It is based on the speech signal characteristics, signal to noise ratio and speech and non-speech region of the signal. To achieve this, hybrid technique combines the voice activity detection (VAD) method and variable frame rate analysis (VFR) method. The hybrid feature frame selection technique has shown better performance over the conventional fixed frame rate (FFR) frame selection and the individual VAD and VFR methods of frame

selection for various noise scenarios.

The number of frames per second (frame rate) made in the hybrid feature frame selection technique is approximately equal to the traditional FFR method. Frame rate can also be controlled by selecting suitable parameter values in the threshold function. This flexibility is beneficial for optimizing the number of speech frame selection according to different applications.

Chapter 4

Multi-Frame Rate based Multiple-Model for Disguised Speech ¹

4.1 Introduction

The performance of the speaker identification system degrades when a mismatch between the training and the testing speech data occurs due to a modification in the person's voice. Modification in a person's voice can occur unintentionally or intentionally. Modification in the voice due to factors like soar throat, emotional state, change in weather, old age are categorized as unintentional. In intentional modification, a person deliberately tries to modify

¹This chapter is based on the following published article: S. Prasad, Z.-H.Tan, R. Prasad, "Multi-frame rate based multiple-model training for robust speaker of disguised voice", *16th Wireless Personal Multimedia Communications (WPMC)*, Atlantic City, New Jersey, U.S.A., Jun. 2013. IEEE Press.

his/her voice either to hide his/her own identity or to be recognized as a target speaker to gain access to their information. Intentional voice modification by a person is also referred to as voice disguise and it is quite frequently encountered in the forensic science area besides other areas. Further, voice disguise can be done electronically, for example, by using the software “voice changer” [24] or non electronically, for example, by speaking fast, adopting a foreign accent, imitation and whisper.

Various studies addressing voice disguise by utilizing electronic means has been done [92], [93]. But a very limited number of studies were found on the non-electronic voice disguise case. A study in the forensic science area [94] has revealed that non-electronic voice disguise is more common in crimes than the electronic voice disguise. Therefore, non-electronic voice disguise requires attention and is the focus of the present chapter. From here onwards, non-electronic voice disguise will be simply referred to as voice disguise only.

In this chapter, three different types of voice disguises/speaking styles together with the normal speaking style has been selected for the study. Out of the three different voice disguises used in the study, two are variants of the imitative style, namely, synchronous and repetitive synchronous imitation and one is the fast speaking non-imitative style.

Earlier works on non-electronic voice disguise was done by Endreas and his group [73] and they showed that, disguisers succeed in varying the formant structure of their voice but they failed in adapting it to the formant structure of a target speaker. The effects of ten different types of voice disguises on the Forensic Automatic Speaker Recognition System is reported in [14]. It is found that, different types of voice disguises degrades the system’s performance by different degrees.

An imitator usually tries to copy the target speaker’s voice by modifying the prosodic elements. A study [95] has been conducted investigating the effects of

professionally imitated voice on the prosodic speaker identification system. It is found that F0 range outperformed the other eleven voice source and prosodic features investigated for identification. It has also been showed that it was easier to copy the target speaker on the basis of the whole sentence than by words. In [15], various voice modifications were tried by the speakers on their own wish and its effects on the Gaussian Mixture Model (GMM) based speaker identification system were reported. It showed significant decrease in performance when the GMM speaker model were trained utilizing only the normal speaking style. The robustness of a new speech feature, namely, “Pyknogram frequency estimate coefficients (*pykfec*)” against voice disguise has been studied in [16] and it showed an overall positive effect on identification accuracy.

All the above studies mostly used a frame size ranging from 10-30 ms (vocal tract information) with a fixed frame shift of half the frame size for extracting speech features for speaker modeling. This setting is preferred because many state of the art speaker identification system has shown good performance with this [3, 16, 50]. But when voice disguise is present, chances of differences in the speaking rate of the training and the test speech data are high and the use of the fixed frame rate for speech feature extraction might not give the best results.

Therefore, this chapter investigates the usage of different frame rates for feature extraction and speaker modeling and its effects on the speaker identification performance under voice disguise case is investigated. It is achieved here by keeping the frame size fixed at say, 25 ms for capturing the vocal tract information but by changing the frame shifts in the range varying between 1-10 ms. Based on the outcomes of the investigation, further, a multi-frame rate based multiple-model speaker identification system is proposed and is showing promising results over the baseline systems which used single frame rate for feature extraction and speaker modeling.

The rest of the chapter is organized as follows. The next section presents the multi-frame rate based multiple-model speaker identification system. The experimental setups and the results obtained were presented in section 4.3. Finally, section 4.4 gives a summary of the chapter.

4.2 Multi-Frame Rate based Multiple-Model

In this section, the conventional GMM based speaker identification method is briefly explained followed by the proposed multi-frame rate based multiple-model method.

4.2.1 GMM

The speaker models are built utilizing the Gaussian mixture density of a speaker ‘sp’ given by

$$p(\vec{x}|\lambda sp) = \sum_{i=1}^N w_i g_i(\vec{x}) \quad (4.1)$$

where, \vec{x} is a D-dimensional feature vector, w_i are the mixture weights with the constraint $\sum_{i=1}^N w_i = 1$ and $g_i(\vec{x})$ are the Gaussian component densities given by:

$$g_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \sqrt{\det C_i}} \times \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)^T C_i (\vec{x} - \vec{\mu}_i)\right\} \quad (4.2)$$

with mean vector $\vec{\mu}_i$ and covariance matrix C_i . Therefore, the complete Gaussian mixture density can be parameterized by:

$$\lambda sp = \{w_i, \vec{\mu}_i, C_i\} \quad i = 1, \dots, N \quad (4.3)$$

Given a training speech data ϕ_{sp} of speaker ‘sp’ and $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_R\}$ being the extracted feature vectors from the speech frames with a fixed frame size and frame shift, the training of speaker model ‘sp’ aims to estimate the parameters of λsp by maximizing the likelihood function:

$$p(X|\lambda sp) = \prod_{t=1}^R p(\vec{x}_t|\lambda sp) \quad (4.4)$$

In the speaker identification from a group of M speakers, the speaker model with the maximum log-likelihood criteria for a test utterance will be decided as the true speaker of the test utterance. The maximum log-likelihood criteria will be given by [96] :

$$\widehat{SP} = \arg \max_{1 \leq sp \leq M} \sum_{t=1}^R \log p(\vec{x}_t|\lambda sp) \quad (4.5)$$

4.2.2 Multi-Frame Rate based Multiple-Model Speaker Identification

In this, the speech features from the training dataset ϕ_{sp} of speaker ‘sp’ are extracted in a different way from the one discussed above. First, multiple copies (say Q) of the training dataset ($\phi_{sp_1}, \phi_{sp_2}, \dots, \phi_{sp_Q}$) were generated. From each Q training datasets of the speaker sp , frames were then made by keeping the frame size fixed to 25ms but by varying the frame shifts across the Q training datasets, resulting in different frame rates for each. Q GMM speaker models are then developed ($\lambda sp_1, \lambda sp_2, \dots, \lambda sp_Q$) utilizing the speech features obtained from each of the Q training speech dataset of the speaker sp .

The decision rule of the speaker identification for a given test utterance \vec{x}_t is given by:

$$\widehat{SP} = \arg \max_{1 \leq sp \leq M} \sum_{j=1}^Q \sum_{t=1}^R \log p(\vec{x}_t | \lambda sp_j) \quad (4.6)$$

4.3 Experimental Setups and Results

The database, the different speaker identification experiments conducted and the results obtained are discussed below:

4.3.1 Database

The CHAINS corpus consist of both male and female speech recordings of 36 speakers [97]. All the speech recordings were carried out in two sessions with a gap of about two months period. The first recording session was conducted in a quiet office environment utilizing the microphone U87 condenser. The second recording session was carried out in a sound proof booth with AKG C420 headset condenser microphone. The bulk of the speakers belonged to same dialect, raising the difficulty of speaker identification and they provided the speech recording in six different speaking styles. Out of the six different speaking styles, four were selected for the experiments and is briefly explained below:

1. Normal speaking (Norm):

The speakers spoke the given texts/sentences in a speaking style with which they usually interact with others in their daily life. It belonged to the first recording session.

2. Synchronous speaking (Sync):

Two speakers spoke the given texts/sentences in synchrony with each other. This produced a change in the timing of the speech production units to be relatively equal but slower than the Norm speaking style [98].

This speaking style is considered as an imitative style type but is not significantly different from the Norm speaking style. It also belonged to the first recording session.

3. Fast speaking (Fast):

The speakers recorded the texts/sentences at a much higher speed than the Norm speaking style. An example of the fast speaking has been played for the speaker so that they get an idea about how much fast to speak. It comes under the non-imitative speaking style and it belonged to the second recording session.

4. Repetitive synchronous Imitation speaking (Rsi):

A recorded sentence in the target speaker’s voice was played in a repeating loop and the speakers joined the repeating loop after the second loop and spoke the sentence in synchrony for about 6 times in which they tried their best to mimic the speaking style of the target. The penultimate recording is then kept. It produced a close match to the target speaker’s voice in timing and intonation. It was originally developed as a pedagogical tool for teaching prosody [99]. It is also considered as an imitative style and it belonged to the second recording session.

For speaker modelling, ~ 70 sec Norm speech data was used per speaker. For testing, ~ 30 sec unseen speech data of all the speaking styles were utilized. During testing, the 30 sec data were broken into ~ 5 sec long utterances making a total of 6 utterances per speaker and a grand total of 216 test utterances from all the speakers for each speaking style.

4.3.2 Speaker Identification Experiments

Speaker identification experiments were conducted for speaker models developed utilizing features which are extracted from frames generated using differ-

ent frame rates and for multiple-model method. For the performance evaluation, the GMM based speaker models which utilized a frame size of 25ms and a frame shift of half the frame size has been used as the baseline. Therefore, speaker identification experiments were conducted for the following three cases:

1. Baseline system, where speaker models are made utilizing frames of frame size 25ms with a single fixed frame shift of 10ms. Let us call this system as Bslne-10ms.
2. Speaker models obtained utilizing frames of fixed frame size i.e. 25ms but by varying the frame shift in the range 1ms-9ms. Therefore, 9 speaker models were made, and for each, speaker identification experiments were conducted. Let us call these models as 1ms, 2ms, ..., 9ms.
3. Speaker models obtained using the multi-frame rate based multiple model method described in Section 4.2.2. Here, three copies of the training dataset has been made. From each copy, frames were made utilizing one of the following frame shifts i.e. 3ms, 6ms and 9ms resulting in 3 different models for each speaker. Let us call this system as MFR-mul.

Twelve liftered MFCC excluding the 0^{th} coefficient were used for feature extraction from the 25ms frames. Sinusoidal liftering is done [100] in order to have similar magnitude values for low and high order cepstral. Cepstral mean removal was applied to the coefficients to compensate for channel variations. 64 component GMM were used for speaker modeling.

4.3.3 Results

The identification accuracies obtained for the speaker identification experiments conducted for the different voice disguised test data and normal test data for the three cases mentioned in Subsection 4.3.2 has been tabulated in

Table 4.1: Identification accuracies (%) of the speaker identification experiments which are discussed in Subsection 4.3.2 for the four types of speaking style test data [104].

Test Data	Frame Shift									Proposed	Baseline
	1ms	2ms	3ms	4ms	5ms	6ms	7ms	8ms	9ms	MFR-mul	Bsline-10ms
Norm	99.07	99.54	99.54	100	99.07	100	99.54	99.07	99.07	99.07	99.07
Sync	94.91	95.37	95.83	96.3	94.44	94.91	95.37	94.91	96.3	95.37	94.91
Fast	84.98	85.92	87.32	83.1	84.98	85.45	85.92	85.92	85.45	88.26	85.45
Rsi	76.85	76.85	79.17	74.54	78.24	76.85	77.78	76.39	74.07	81.94	72.69
<i>Average</i>	88.95	89.42	90.47	88.49	89.18	89.30	89.65	89.07	88.72	91.16	88.03

Table 4.1. From the table, it can be observed that varying the frame shift in the range 1 ms-9 ms for the Norm speaking test data showed either slight improvement in identification accuracy or remained equal to 99.07% of Bslne-10 ms system. A maximum of 0.93% relative improvement is observed for the 4 ms and 6 ms case compared to the Bslne-10 ms. For the Sync speech test data, similar observations have been found with a maximum of 1.46% relative improvement in the identification accuracy for the 4 ms and 9 ms case compared to the Bslne-10 ms. However, for the 5 ms case 0.49%, decrease in identification accuracy is observed compared to the Bslne-10 ms. Fast speech

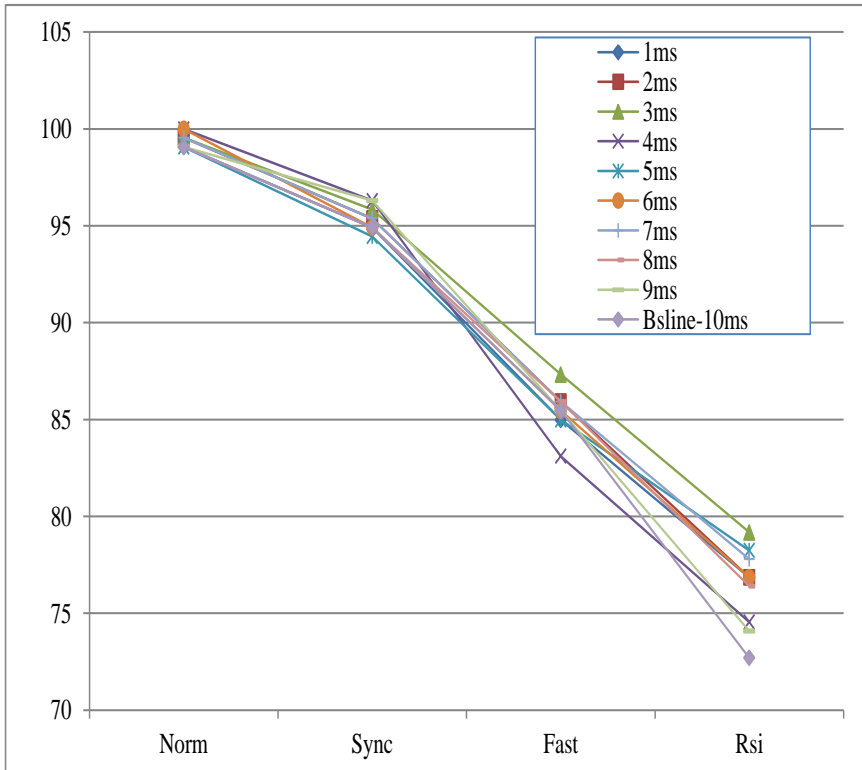


Figure 4.1: Comparison of the identification accuracies(%) obtained using speaker models with different frame rates and the baseline for the four speaking style's test data [104].

test data has seen more number of declines in the identification accuracies, namely for 1 ms, 4 ms and 5 ms case. A maximum relative improvement of 2.18% has been observed for 3 ms case compared to the Bslin-10 ms. The best results amongst all test data has been observed for the Rsi speech. Here, the identification accuracy has always seen improvement and a maximum of 9.71% relative improvement has been observed compared to the Bslin-10 ms system. Fig. 4.1 depicts the comparison of the systems utilizing different frame rates for speaker modeling (1 ms-9 ms) and Bslin-10 ms for disguised and normal test speech in graphical form. 3 ms system outperformed all others systems for Fast and Rsi speech.

Encouraged by the results obtained by varying the frame shifts (Case 2), MFR-mul system is developed. Comparing MFR-mul and Bslin-10 ms, it can be observed that for Norm speech test data, both systems performed equally, but for Sync, Fast and Rsi test data, MFR-mul outperformed the Bslin-10 ms system. It showed a relative improvement of 0.48% for Sync, 3.29% for Fast and 12.72% for Rsi speech as compared to Bslin-10ms. Compared with the different frame rate models (1 ms- 9 ms), MFR-mul always performed better for the fast and Rsi test speech data.

Fig. 4.2 shows a comparison of the identification accuracies of the three systems (Different frame rate models: 1 ms-9 ms, MFR-mul and Bslin-10 ms) for the different test speech data. For different frame rate models (1 ms-9 ms), identification accuracy has been calculated by averaging across a single speaking style. It also confirms the superior performance of the MFR-mul system over the others. Fig. 4.3 shows the comparison of the different systems by taking an average of the identification accuracies of all the test speech data. Here also MFR-mul outperformed the others.

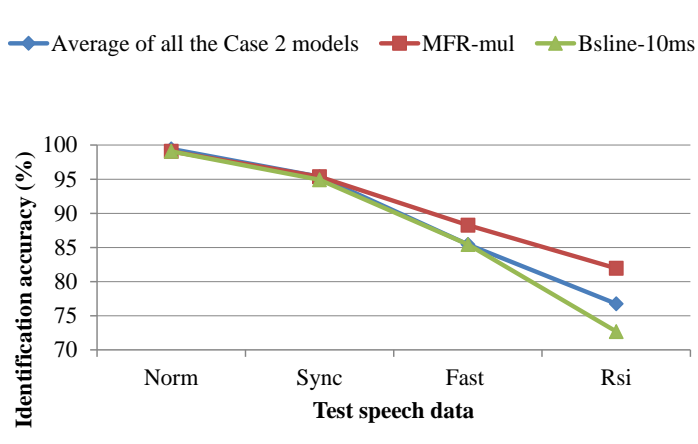


Figure 4.2: Comparison of the identification accuracies (%) of the different frame rates models (average), multi-frame rate based multiple-model training and baseline for the four speaking style's test speech data [104]

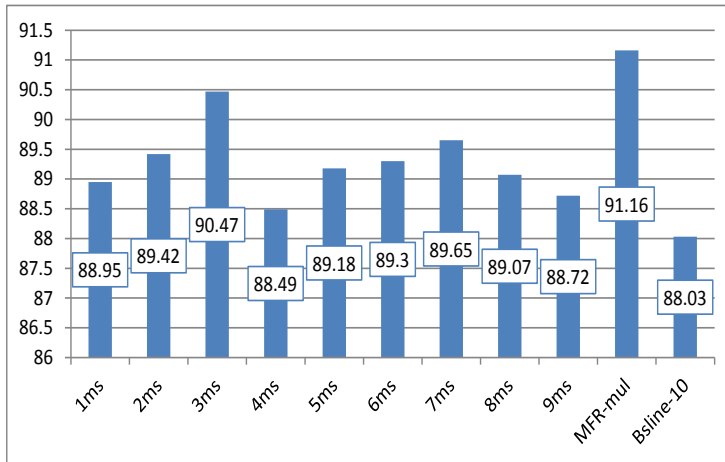


Figure 4.3: Average identification accuracy across different speaking styles for the different frame rates, the multi-frame rate based multiple-model training and the baseline method

4.4 Summary

The effects of using different frame rates for feature extraction and speaker modeling on the speaker identification performance is investigated in this chapter for the case, when speaker models were trained from normal speaking style data but the test data contains disguised speech. Three types of disguised speech from the CHAINS corpus, namely, fast (Fast), synchronous (Sync) and repetitive synchronous imitation (Rsi) were used for the experiments. Out of these three, Sync and Rsi comes under the imitative style type and Fast is the non-imitative style voice disguise. Experimental results showed that, the use of different frame rates may improve the speaker identification performance for disguised speech. Based on these observations, further, a multi-frame rate based multiple-model is proposed and it has outperformed the conventional GMM system utilizing the single frame rate method on an average across the different speaking styles test data. For the Rsi voice disguise, it gave the best identification accuracies.

Chapter 5

Multistyle Training and Fusion Framework for Disguised Speech¹

5.1 Introduction

Chapter 4 investigated different frame rates for speaker modeling under the voice disguised mismatched condition. Here, speaker models were built from the normal speaking style training data and the test data consisted of different speaking styles/disguised speech data. In this chapter, a different approach

¹This chapter is based on the following published article: S. Prasad, Z. -H. Tan, R. Prasad, “Multistyle training and fusion for speaker identification of disguised voice”, *1st International Conference on Communications, Connectivity, Convergence, Content and Co-operation (IC5)*, Mumbai, India, Dec. 2013 & submitted article: S. Prasad, R. Prasad, “Fusion multistyle training for speaker identification of disguised speech”, *Wireless Personal Communications*.

is used. Here, different multistyle training strategies for speaker modeling are investigated for the disguised speech. In multistyle training, instead of using only normal speaking style speech data for training, other easily produced speaking styles, like fast speaking and synchronous speaking are also utilized for training the speaker model. This approach can be implemented for security conscious organization for monitoring the employees, where chances of leakage of information through the employee by adopting voice disguise over phone or even otherwise is high.

Early research on multistyle training method was done by Lippmann [19] for speech recognition. He utilized, five easily produced speaking styles (normal, fast, clear, loud and question-pitch) for speech model training. It showed improved performance for speech recognition of stress speech, produced during a workload task. A colored noise based multicondition training method is explored for noise robustness [58] because it has been noted that many environmental noises contain colored spectra. In [59], multicondition training data were generated by artificially adding white noise at different SNRs to multiple copies of clean training data. Speaker models made using this multicondition training data showed improved performance for speaker identification under unknown environmental noise types. A multiple-model frame work has been proposed in [101] for handling noisy speech in speech recognition task. In this, specific models for a particular noise type with a specific SNR value are trained, and for recognition, the best model matching the test speech was selected from others. It has reported superior performance than the multicondition training method.

Inspired by these works, this chapter investigates different multistyle training strategies for voice disguised test speech data. For multistyle training of speaker models, three easily produced speaking styles, namely, normal (Norm), synchronous (Sync) and fast (Fast) speaking styles were used. For testing the models, unseen speech data from all the above mentioned speaking styles

(Norm, Sync and Fast) are used. In addition to this, the repetitive synchronous imitation (Rsi) speaking style/disguised speech which was not used for training is also used for the testing. Important insights were drawn from the results obtained. This led to the development of the fusion multistyle training framework.

The rest of the chapter is organized as follows. The next section gives a brief discussion on the disguised speech/speaking style which are available with the speaker. The different multistyle training strategies investigated in this chapter together with the proposed fusion multistyle training method are presented in section 5.3. The dataset, the speaker identification experiments conducted and the results obtained are explained in section 5.4. Finally, section 5.5 summarizes the chapter.

5.2 The Different Speaking Styles

A wide range of options are available with the speakers to modify their speaking style. For example, speakers can bring a change in their voice by raising or lowering the pitch, chewing something (like chewing gum), cheek pulling, lip protrusion, whisper, mimicry, adopting foreign accent, using a different dialect, objects in mouth and objects over mouth [102]. Predicting the full range of voice disguises which can be used by the speakers in advance is very difficult because it is dependent on the person's ingenuity, thereby, making the speaker identification task more difficult.

For the experiments conducted in this chapter, four speaking styles, namely, Normal (Norm), Synchronous (Sync), Fast and Repetitive Synchronous (Rsi) Imitation were used, whose full details can be found in [97] and is briefly explained in Subsection 4.3.1

5.3 Multistyle Training Strategies and the Fusion Framework

The different multistyle training strategies investigated in this chapter for the speaker identification performance under mismatched conditions arising due to disguised speech are presented first, followed by the description of the proposed fusion framework.

1. Multistyle Training Strategies

The speaking styles Norm, Sync and Fast explained briefly in Subsection 4.3.1 were used for multistyle training. The spectrogram of these three speaking styles for a speech utterance can be seen in Fig 5.1. It can be observed that Norm and Sync speaking styles did not appear very different but Fast speaking resulted in a quite different spectrogram than the Norm speaking style. The following four types of multistyle training strategies were investigated in this chapter:

- Multistyle Training I (Mul-I):

The training utterances from the three speaking styles Norm, Sync and Fast were mixed randomly. This was then used for feature extraction and speaker modeling. A frame size of 25ms with a frame shift of 10ms was utilized for feature extraction.

- Multistyle Training (Mul-II):

All the training utterances from the Norm speaking style were simply concatenated with all the utterances from the Fast speaking style followed by all the utterances from the Sync speaking style. The training set obtained was then utilized for feature extraction and speaker modelling. A frame size of 25ms with a frame shift of 10ms was utilized for feature extraction.

- Multistyle Training III (Mul-III):

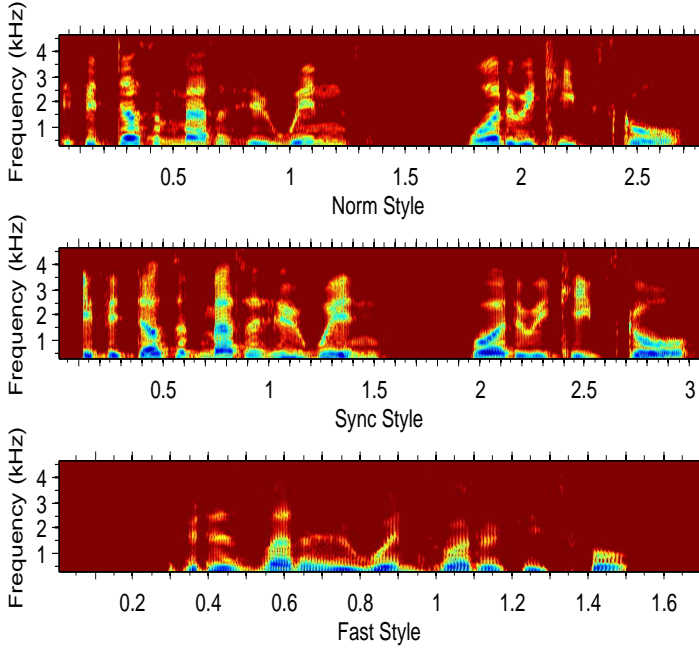


Figure 5.1: Norm, Sync and Fast speaking style spectrogram for the speech utterance “If it doesn’t matter who wins, why do we keep score?” [105]

First utterance of Norm speaking style was concatenated with the first utterance of the Fast speaking style followed by the first utterance of the Sync speaking style. The same way of concatenation was used for the second utterance and the process was continued till all the utterances come to an end. The resultant training set was then utilized for feature extraction and speaker modeling. For this also, a frame size of 25ms with a frame shift of 10ms was used for feature extraction.

- Multistyle Training IV utilizing higher frame rate for feature extraction from Fast Speaking style (Mul-IV):

This training strategy is similar to the Mul-III training strategy except that for feature extraction from the Fast speaking style, a

higher frame rate was used. This was achieved by utilizing a frame size of 25ms with a smaller frame shift of 3ms.

For the speaker identification task from a group of M speakers for a given test utterance $X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_R\}$. The decision rule is based on the maximum log-likelihood rule [96], which is given by the following equation:

$$\widehat{Spk} = \arg \max_{1 \leq spk \leq M} \sum_{t=1}^R \log p(\vec{x}_t | \lambda_{spk}) \quad (5.1)$$

2. Fusion Framework

To explore the diversity of the different multisyle training strategies, the fusion framework combines the multistyle training strategies at the decision level for identifying the correct speaker from a group of M speakers. In this chapter, two multistyle training strategies Mul-II and Mul-IV were combined by the following combination by maximum rule:

$$\widehat{Spk} = \arg \max_{1 \leq spk \leq M} \max_{1 \leq j \leq 2} \sum_{t=1}^R \log p(\vec{x}_t | \lambda_{spkj}) \quad j = \{\text{Mul-II}, \text{Mul-IV}\}. \quad (5.2)$$

where λ_{spkj} represents the speaker model obtained from utilizing the multistyle training strategy j

5.4 Experimental Setups and Results

The database used for the experiments, the different speaker identification experiments conducted in this chapter and the results obtained are discussed in this section.

5.4.1 Database

The CHAINS corpus [97] consisting of 36 speakers, mostly from the same dialect is selected to carry out the experimental evaluations. The speakers provided their voice in 6 different speaking styles for this database. Out of the six speaking styles, four speaking styles, namely, Norm, Sync, Fast and Rsi were used and a brief explanation of each of these styles were given in Subsection 4.3.1. All the speakers recorded their speech in two different sessions separated by 2 month’s period. The first session was carried out in a quiet office environment consisting of some noise utilizing the microphone U87 condenser. The second session was carried out in a sound proof booth utilizing AKG C420 headset condenser microphone.

The speaker models were built utilizing ~ 70 s training speech data per speaker. For testing ~ 30 s unseen speech data per speaker were used. For testing, 5 sec long utterances were used. Therefore, 6 utterances per speaker and a total of 192 utterances from all the speakers were tested.

5.4.2 Speaker Identification Experiments

Speaker identification experiments were conducted for the different multistyle training strategies and the proposed fusion framework discussed in section 5.3. For performance evaluation, speaker identification experiments employing single style speaker models and multiple model framework were also conducted and are briefly described below:

Single style

GMM speaker models were developed utilizing the training speech data from the single speaking styles of Norm, Sync and Fast resulting in three speaker

models. Let us call these models as NORM, SYNC and FAST, respectively. Speech features were extracted with 25ms frame size with 10ms frame shift.

Multiple model

The single style speaker models obtained above, namely, NORM, SYNC and FAST for the individual speaking styles of Norm, Sync and Fast, respectively, were combined at the decision level to get the multiple model framework. The following three types were utilized in this study:

- Multiple model I (**MM-I**):

The likelihoods “ $p(X|\lambda_{spk})$ ” from the single style speaker model NORM, SYNC and FAST for a test speech utterance X were added and the speaker model which maximizes this value is decided as the true speaker of the test utterance.

- Multiple model II (**MM-II**):

The combination rule is similar to the Fusion method described in Section 5.3. The maximum of the log-likelihoods from the single style models NORM, SYNC and FAST for a test speech utterance were selected and the speaker model which maximizes this value is decided as the true speaker of the test utterance.

- Multiple model III (**MM-III**):

The maximum of the log-likelihoods from the single style speaker models NORM, SYNC and FAST for a speech utterance at the frame level are selected. The speaker model which maximizes the log-likelihood criteria (eqn. 5.1) for the whole test speech utterance is decided as the true speaker of the test utterance.

Twelve liftered MFCC feature were extracted excluding the 0th coefficient from the speech frames for speaker modeling. Cepstral mean removal of the coefficients were taken for channel compensation and 64 component GMMs were utilized for modeling the speakers.

5.4.3 Results

Table 5.1 depicts the speaker identification accuracies obtained for the different multistyle training strategies, proposed Fusion method, single style training method and the multiple-model methods for the test speech data which consisted of the three types of voice disguises along with the Norm speaking style test data. Comparing the different multistyle training strategies (Mul-I, Mul-II, Mul-III and Mul-IV) and the single style training method for speaker modeling (NORM, SYNC and FAST), it can be observed that all multistyle training strategies outperformed the single style training method on an average across the different test speech data. Comparing the different multistyle training strategies with the multiple models (MM-I, MM-II and MM-III), it can be seen that, all multistyle training strategies performed better than the multiple model MM-I but MM-II and MM-III outperformed the multistyle training strategies Mul-I and Mul-II on an average across the different speaking style's test speech data.

Comparing only the different multistyle training strategies, it is interesting to note that different multistyle training strategies showed quite different identification performances for the different speaking style's test speech data and which can prove quite beneficial in improving the robustness of the overall system. Multistyle training Mul-II showed the best performance for the Norm speaking style test data and worst for the Fast speaking style compared to the other multistyle training strategies. On the other, hand Mul-IV showed the best performance for the Fast and Rsi speaking styles and worst for the

Table 5.1: Identification Accuracies (%) of the various speaker identification experiments for the Norm and disguised speech test data [106].

Test	Multistyle				Proposed	Single Style			Multiple Models		
	Mul-I	Mul-II	Mul-III	Mul-IV	Fusion	NORM	SYNC	FAST	MM-I	MM-II	MM-III
Norm	98.61	99.07	98.61	97.69	99.07	99.07	90.74	61.57	93.98	99.07	98.61
Sync	99.07	99.07	99.07	97.69	98.61	95.37	99.07	53.7	94.91	99.07	98.61
Fast	95.31	94.37	97.18	97.65	97.18	84.51	71.83	98.12	94.84	96.71	97.18
Rsi	92.59	93.06	93.52	94.91	95.37	76.85	68.52	93.06	90.74	91.67	93.06
<i>Average</i>	<i>96.40</i>	<i>96.39</i>	<i>97.1</i>	<i>96.99</i>	<i>97.56</i>	<i>88.95</i>	<i>82.54</i>	<i>76.61</i>	<i>93.61</i>	<i>96.63</i>	<i>96.86</i>

Norm and Sync speaking styles. These observations lead to the proposed fusion method in which fusion of the Mul-II and Mul-IV training strategies at the decision level has been done.

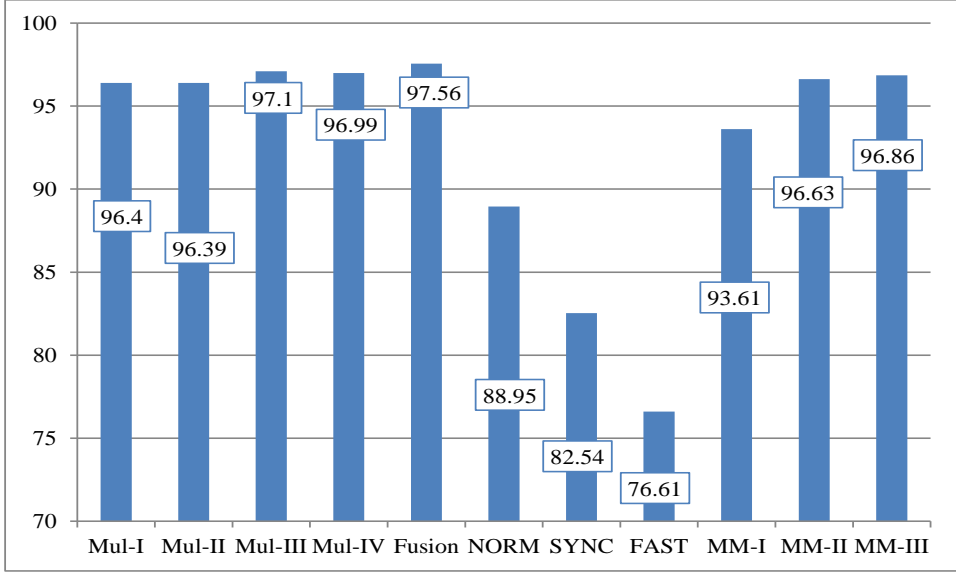


Figure 5.2: Comparison of the average identification accuracies across the Norm and disguised test speech data for the various speaker identification experiments [106].

The Fusion method has shown the best performance outperforming all other methods considered in this chapter when an average of the different speaking style's test speech data were considered. A graph showing the comparison of the proposed Fusion method with the other methods is shown in Fig 5.2. The graph shows the average identification accuracies of the different speaking style's test data and it clearly shows the better performance of the Fusion method. Compared to the different multistyle training strategies, the Fusion method has achieved the best performance for the Rsi voice disguise outperforming Mul-I by $\sim 2.78\%$, Mul-II by $\sim 2.31\%$, Mul-III by $\sim 1.85\%$ and Mul-IV by $\sim 0.46\%$. A comparison of the different multistyle training strategies and the

Fusion method for the different speaking style's test data is also depicted in Fig 5.3. From Fig 5.3, it can be observed that the Fusion method has showed a more stable performance across the different speaking style's test data compared to the other multistyle training strategies. Compared with the single

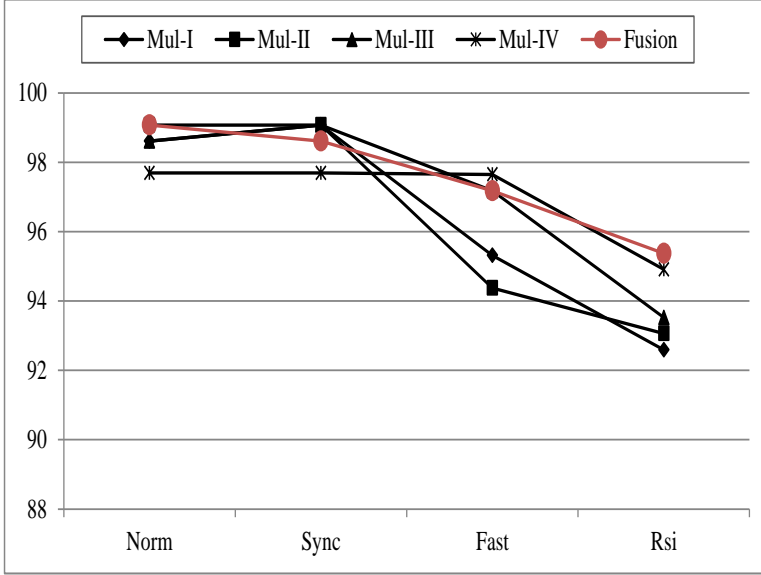


Figure 5.3: Comparison of the identification accuracies of the different multi-style training strategies and the Fusion method for Norm and disguised speech test data [105].

style training methods NORM, SYNC and FAST, like all multistyle training strategies, the Fusion method has also shown a significantly better performance on an average across the different speaking style's test speech data. A comparison of the Fusion and the single style training methods is also shown in Fig 5.4. Fusion method again showed a stable performance across the different speaking style's test data compared to the single style training methods. Compared with the multiple model methods. For the Rsi voice disguised test data, Fusion showed an improvement of $\sim 4.63\%$ from MM-I, $\sim 3.7\%$ from MM-II and $\sim 2.31\%$ from MM-III. For the Sync voice disguised test data MM-II per-

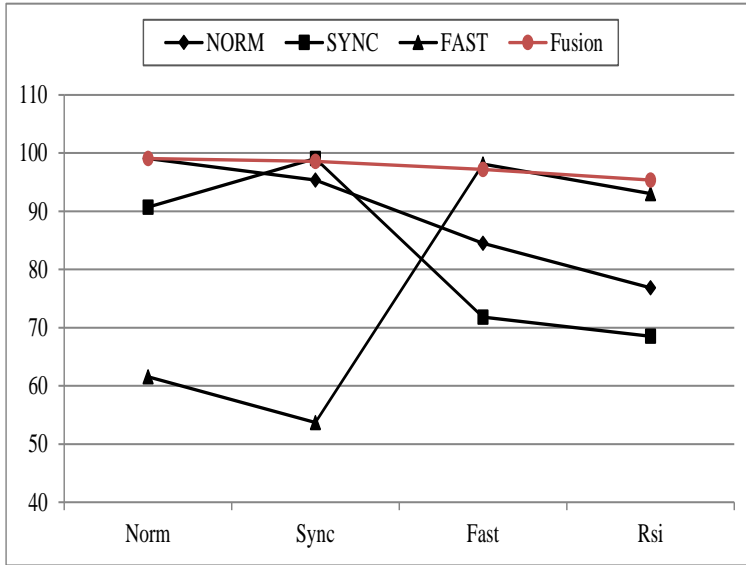


Figure 5.4: Comparison of the Fusion and the Single style training method for the Norm and voice disguised test speech data [105].

formed better than the Fusion by ~ 0.48 %. For Fast and Norm test speech data, Fusion either showed improvement or performed equally. A comparison of the different multiple models and the Fusion method is also presented in Figure 5.5. In this case also the proposed Fusion method has shown a more stable performance than the other methods.

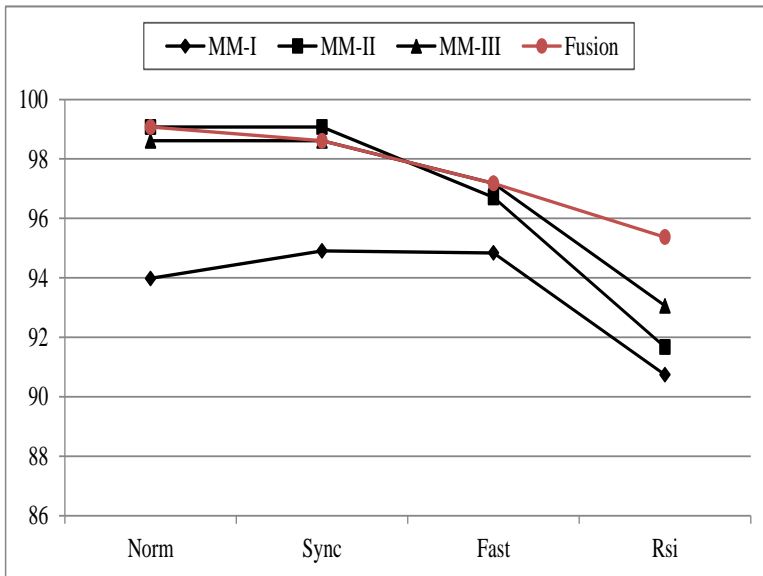


Figure 5.5: Comparison of Fusion and the multiple model methods for the different voice disguises and Norm speaking style [106].

5.5 Summary

Four different types of multistyle training strategies obtained from the speaking styles normal, Synchronous and fast training data were investigated for the speaker identification accuracy under the voice disguised scenario. The three voice disguised test data used were the unseen Sync and Fast test data, and the repetitive synchronous imitation which is not used during the training. All the speech data were taken from the CHAINS corpus. The multistyle training strategies has given useful insights for improving the robustness of the speaker identification system. A fusion method is therefore proposed combining the two multistyle training strategies at the decision level. The fusion method has shown the best performance on an average across the different voice disguised test data compared to the different multistyle training strategies , single style

training and the multiple-model methods considered in this chapter. The fusion method has also shown a more stable performance for the different voice disguises compared to the other methods.

Chapter 6

Multiple Frame Rates for Feature Extraction and Reliable Frame Selection at the Decision for Disguised Speech¹

6.1 Introduction

In Chapter 4, we have observed that utilizing a different value of frame shift instead of the usual 10ms frame shift for frame making from the speech signal

¹This chapter is based on the following published article: S. Prasad, Z.-H. Tan, R. Prasad, “Multiple frame rates for feature extraction and reliable frame selection at the decision for speaker identification under voice disguise” , *CONASENSE*, no.1, pp-29-44, Jan. 2016.

for feature extraction may lead to a positive effect on the speaker identification accuracy under the voice disguised scenario. Here, only normal speech data from speakers were utilized for training the models. Different speaker models were made by varying the frame shift in the range 1ms-9ms. Frame shift 3ms showed the best performance amongst the other frame shift models. In this chapter, not only normal speaking style but other speaking styles, namely, synchronous and fast speaking are also utilized for training the speaker models. For frame making from the training speech data, a frame shift of 3ms is utilized. The value of 3ms frame shift is used because it showed the best performance for the normal speaking style based speaker models presented in Chapter 4. Further, during the testing or at the decision level, a method has been developed which performs reliable frame selection from the test speech utterance. Only these reliable frames will then participate in the final decision making process.

This chapter is organized as follows. The next section describes the reliable frame selection method at the decision. Section 6.3 presents the database, the different speaker identification experiments conducted and discusses the results. The last section summarizes the chapter.

6.2 Reliable Frame Selection at the Decision

During the testing phase in the speaker identification system, a speech utterance from an unknown speaker is given. The task is to find out the speaker of the given speech utterance from a group of M people.

Let us assume that the speech utterance is represented by a feature vector sequence $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_T\}$. Conventionally, the speaker model which maximizes the posterior probability for the feature vector X is decided as the true speaker of the given speech utterance. It is represented by the following

equation [96]:

$$\hat{S} = \arg \max_{1 \leq s \leq M} \sum_{t=1}^T \log p(\mathbf{x}_t | \lambda_s) \quad (6.1)$$

where \mathbf{x}_t is a D-dimensional feature vector and T is the total number of feature vector in the given speech utterance. $p(\mathbf{x}_t | \lambda_s)$ is the Gaussian mixture density of speaker 's', which is a linear weighted sum of N component densities $b_j(\mathbf{x}_t)$

$$p(\mathbf{x}_t | \lambda_s) = \sum_{j=1}^N w_j b_j(\mathbf{x}_t) \quad (6.2)$$

Here, w_j are the mixture weights with $\sum_{j=1}^N w_i = 1$.

In the proposed method the speaker of the given speech utterance is found from a group of M people in a different way, which is described below [103]:

1. Instead of finding the speaker of the whole speech utterance, the speaker of each frame is determined, which is given by the following decision rule:

$$\hat{S} = \arg \max_{1 \leq s \leq M} p(\mathbf{x}_t | \lambda_s) \quad (6.3)$$

2. In this step, how reliable is the decision of step 1 about the speaker is calculated. For this the distance between the probability measure of the identified speaker with the rest of the speakers is calculated as follows:

$$D_s = p(\mathbf{x}_t | \lambda_{\hat{S}}) - p(\mathbf{x}_t | \lambda_s) \quad k = \{1, 2, 3, \dots, M\} - \{\hat{S}\} \quad (6.4)$$

In this way (M-1) distances will be calculated. The larger the distance D_s will be, the more confidence will be in the decision.

3. The distances which are calculated in step 2 are now compared with a threshold value θ . If $D_s > \theta$, the frame is kept otherwise it is discarded.
4. After step 3, all the unreliable frames of the test speech utterance will be discarded and the remaining frames will only participate in the final decision making.

5. The final decision rule for the speech utterance which now consist of only reliable feature vector sequence is same as equation 6.1.

6.3 Experimental Setup and Results

In this Section, the database used for the experiments, the various speaker identification experiments conducted and the results obtained are discussed.

6.3.1 Database

CHAINS database is used to conduct the experiments. The details of this database can be found in [97]. The database consisted of speech recordings in different speaking styles from 36 speakers consisting of both males and females. The speakers mostly belonged to the same dialect which made the identification task tougher and they provided the speech recordings in two sessions. The two sessions were separated by a period of two months, the first session speech recordings were carried out in a quiet office environment with some office noise utilizing the microphone Neumann U87 and the second in a sound proof booth using the microphone AKG C420. Four different types of speaking styles were utilized for the experiments, namely, normal (Norm), synchronous (Sync), fast (Fast) and repetitive synchronous imitation (Rsi) speaking.

Norm and Sync speech recordings were done in the first session and Fast and Rsi in the second. The use of different microphone and speaking styles provided a good mismatch both in channel and style suitable for the present mismatch problem experimentation.

For the training of speaker models, three speaking styles, namely, Norm, Sync and Fast were utilized. For testing, unseen speech recordings from Norm, sync and fast were used. Additionally, Rsi speaking which was not used for

training is also included for the testing. For training ~ 70 sec speech data per speaker were utilized. For testing, three, ~ 10 sec speech utterances per speaker i.e. total 30 sec speech data per speaker were utilized from Norm, Sync and fast speaking. For Rsi speaking, four, ~ 10 sec utterances per speaker were used.

6.3.2 Speaker Identification Experiments

Speaker models were developed utilizing 64 component Gaussian mixture models. 12 liftered MFCC excluding the 0^{th} coefficient were used as speech features. Liftering is done to rescale the higher and lower order cepstral so that they have similar magnitudes. Cepstral mean removal is also utilized for channel compensation.

The following speaker identification experiments were conducted:

1. Two speaker models were made for each speaker in the group of M speakers. One utilized a frame size of 25ms and a frame shift of 10ms for feature extraction and the other utilized a frame size of 25ms with a frame shift of 3ms for feature extraction.

During testing, similarly two sets of frames were made. One with 25ms frame size and 10ms frame shift and the other using 25ms frame size and 3ms frame shift. Reliable frame selection from these two sets were carried out using the method described in Section 6.2. For the final decision, reliable frames selected from these two sets were simply combined for feature extraction, and the decision making is done by the same method using the decision rule of equation 6.1. Let us call this system as Proposed.

For evaluating the proposed system, the following two baseline system were considered:

2. The baseline system utilized the Norm, Sync and Fast speaking styles for speaker modeling. The frame size of 25ms with a frame shift of 10ms were

used for feature extraction and speaker modeling. This resulted in three speaker models for each speaker which was developed utilizing the Norm, Sync and fast speaking style speech data. This typical value of 25ms for frame size and 10ms for frame shift have been utilized in many research studies [14], [16], [96], which uses different database, and is found to give the best identification accuracy. Let us call this baseline system as Bsln1.

3. In chapter 3, it has been shown that varying the frame rate by keeping the frame size fixed and changing the frame shift may improve the speaker identification accuracy under voice disguise scenario. The frame shift of 3ms has shown the best performance. Here, speaker modeling utilized only Norm speaking style. For this baseline, speaker models utilized three speaking style, namely, Norm, Sync and Fast and for feature extraction a frame size of 25ms with a frame shift of 3ms were used. In this way, three speaker models for each speaker were made using Norm, Sync and Fast speaking. Let us call this baseline system as Bsln2.

6.3.3 Results and Discussions

Table 6.1 shows the identification accuracies obtained for the Bsln1 and the Bsln2 systems for the different speaking style test speech data. The test speech data which are mismatched both in style and channel with the training speech data are shown with a superscript of star and the test speech data which are matched in channel are shown without any superscript of star. For example, for the training speech data of Norm (first row of the Table 6.1), the test speech data, namely, Fast and Rsi are mismatched both in style and channel and are shown with a superscript of star and the test speech data, namely, Norm and Sync are matched in channel and therefore are shown without any superscript of star.

From the Table 6.1, it can be observed that, when the test speech data are

Table 6.1: For the normal and disguised speech test data, identification accuracies (%) obtained for Bsln1 & Bsln2 system where speaker models are trained using different speaking style’s speech data [12].

Train			
Speech	Test Speech	Bsln1	Bsln2
Norm	Norm	100	100
	Sync	100	100
	Fast*	90.74	92.59
	Rsi*	77.78	81.94
<i>Average</i>		92.13	93.63
Sync	Norm	96.30	95.37
	Sync	100	100
	Fast*	76.85	78.70
	Rsi*	72.22	68.37
<i>Average</i>		86.34	85.61
Fast	Fast	99.07	100
	Rsi	95.83	93.75
	Norm*	66.67	67.59
	Sync*	60.19	62.96
<i>Average</i>		80.44	81.08

mismatched both in speaking style and channel, the identification accuracies decreased markedly for both Bsln1 and the Bsln2 systems. The performance declined more for Sync and Fast speaking style train data as compared to the Norm training data.

Comparing the Bsln1 and the Bsln2 systems, it can be observed that, for the Norm speaking style train speech data, Bsln2 performed better than the Bsln1 on an average across the different speaking style test data. Bsln2 showed a

relative improvement of 1.63% over Bsln1. Bsln2 performed equally with Bsln1 for the test speech that are matched in channel but for test speech mismatched both in style and channel, Bsln2 performed better than Bsln1. Bsln2 showed a relative improvement of 2.04% and 5.35% over the Bsln1 for Fast and Rsi test speech, respectively.

For Sync speaking style train speech data, Bsln1 performed better than the Bsln2 on an average across the different speaking style test speech data. Bsln1 showed a relative improvement of 0.85% over Bsln2.

For Fast speaking style train speech data, again Bsln2 performed better than the Bsln1 on an average across the different speaking style test speech data. Bsln2 showed a relative improvement of 0.80% over Bsln1.

Comparing Bsln1 system for the different train speech data, it can be observed that, Norm train speech data performed the best on an average across the different speaking style test speech data. Bsln1 system trained using Norm speaking style train data showed a relative improvement of 6.7 % and 14.5% over Bsln1 system trained using Sync and Fast training speech data, respectively. Similar kind of observation is found for Bsln2 system. Bsln2 system trained using Norm speaking style train speech data showed a relative improvement of 9.4% and 15.5% over Bsln2 system trained using Sync and Fast speaking speaking style train speech data. From these observation it can be concluded that, if a single speaking style has to be used for training the speaker model under voice disguise, Norm speaking style gives the best performance over Sync and Fast. Therefore, in the Proposed system, only Norm speaking style train speech data has been used for modeling the speakers.

For the reliable frame selection from the test speech signal used in the Proposed system, the value of the threshold θ needs to be determined. One way of calculating the threshold θ is by observing all the distances D_s (refer to Section 6.2) for a given test speech utterance. Selecting a larger value of

θ compared to the minimum of all the distances D_s found for the test speech utterance, may lead to the rejection of a large number of frames, which might remove frames which carry some important speaker specific information. These rejected frames will now not be able to participate in the decision making about the speaker of the test speech utterance and may result in a decrease in the identification accuracy. Therefore an optimum value of θ needs to be used and is tough to estimate. Further, the threshold value θ should vary for different test speech utterances. But for this initial study, θ value is kept fixed and is determined by observing the identification accuracies found by testing only the Rsi test speech utterances and speaker models trained using only the Norm speaking style train speech data. The frame size of 25ms with a frame shift of 10ms is utilized for feature extraction.

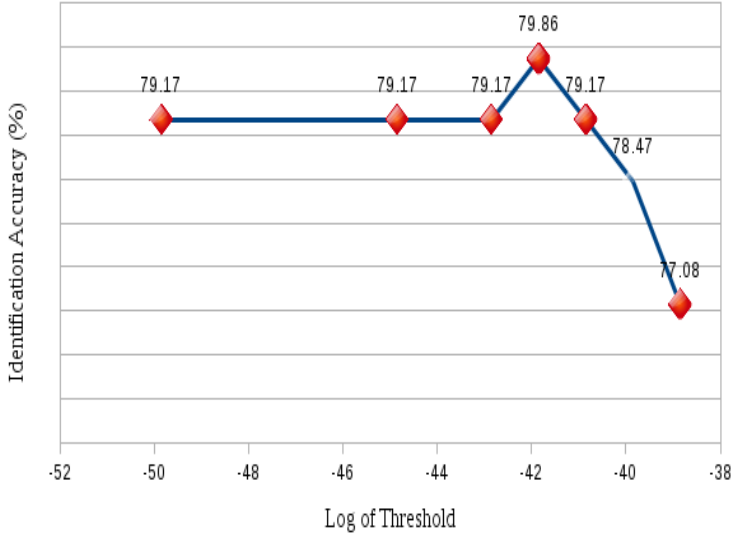


Figure 6.1: For the Rsi speech test data, identification accuracies (%) obtained by using different threshold value θ [12]

For calculating the fixed θ value, different θ values were chosen by observing the minimum of all the distances D_k from a few Rsi test speech utterances.

The identification accuracies obtained for each of these θ value for the Rsi test speech is plotted and is shown in Fig. 6.1. For more clarity log of θ is plotted instead of θ . The θ value which gave the highest identification accuracy, i.e., $1.411\text{e-}42$ is decided as the θ value for the conduction of the experiments of the Proposed system.

Table 6.2: Identification accuracies (%) of the baseline and the proposed systems for Norm and disguised speech test data [12].

Test			
Speech	Bsln1	Bsln2	Proposed
Norm	100	100	100
Sync	100	100	100
Fast*	90.74	92.59	92.59
Rsi*	77.78	81.94	83.33
<i>Average</i>	92.13	93.63	93.98

Table 6.2 shows the identification accuracies for the Bsln1, Bsln2 and the Proposed system for the different speaking style test speech data. From the table, it can be observed that the Proposed system has performed slightly better than the Bsln1 and the Bsln2 on an average across the different speaking style test speech data. For test speech matched in channel, the Proposed system performed equally with the Bsln1 and the Bsln2. For the Fast speech, mismatched both in speaking style and channel, Proposed system performed better than Bsln1 but did not do better than the Bsln2. For the Rsi test speech data, mismatched both in speaking style and channel, the Proposed system achieved a relative improvement of 4.35% and 1.70% over the Bsln1 and the Bsln2, respectively.

6.4 Summary

This chapter studies the effect of reliable frame selection at the decision level for voice disguised test speech. For frame selection from the test speech signal, frames were made at two different frame rates utilizing a fixed frame size but varying the frame shift. It showed an overall improvement over the baseline methods and performed well for the disguised test speech of repetitive synchronous imitation (Rsi). Future studies will aim at involving more types of disguised speech and development of a better algorithm for the threshold value selection utilized in the reliable frame selection method.

Chapter 7

Brain Computer Interface for Classification of Motor Imagery Tasks from the Same Limb Using EEG ¹

7.1 Introduction

The production of the speech signal is related to our thoughts and thereby brain signals. The study of the brain signals might give useful insights for improving the speaker identification system. An attempt has been made in this

¹This chapter is based on the following published article: S. Prasad, Z.-H. Tan, R. Prasad, A. R. Cabrera, Y. Gu, K. Dremstrup, “Feature selection strategy for classification of single trial EEG elicited by motor imagery”, *14th Wireless Personal Multimedia Communications (WPMC)*, Brest, France, Oct. 2011. IEEE Press.

direction by studying a Brain-Computer Interface (BCI) system. BCI connects the humans with the computer not through the conventional ways but through brain signals. Two devices through which brain signals can be acquired from the humans for interface with the computer are: Electroencephalogram (EEG) and Electrocorticogram (ECoG). EEG is a non-invasive method for recording the electric signals from the brain and it is easily recorded using electrodes which are placed over the scalp. On the other hand, ECoG is an invasive method, where the electrodes are placed directly over the brain surface for recording the brain activity. It can pose serious health problems to the humans. Therefore, out of these two methods, EEG is more popular in BCI research as it is easy to use, non-invasive, and is cheaper too [107]. It has been found that imagination of different tasks shows discriminative changes in the brain signals. This property of the brain signal proves, highly beneficial to the people suffering from neuro-muscular diseases, like, brain stem stroke and amyotrophic lateral sclerosis. In this, the person is in a locked-in state where his or her feelings and thinking capabilities are intact, but they have problem in movement related tasks, speech or vision, that is, they have problems in expressing their feelings to others and to carry out their day-to-day activities. Through BCI research, such people have now started expressing their needs/feelings to others. BCI tries to understand their intentions through brain signal manipulations and convey it to the outer world [108] [109], [110].

A BCI system can be broadly divided into 5 parts: signal acquisition, signal pre-processing, feature extraction, classification and controlling interfaces [107]. In signal acquisition, the brain activity in the form of electrical signals are captured from the human brain. The electrical signals are generated by the interactions of the neurons present in the brain. Instead of directly feeding the acquired brain signals to the feature extraction part, it is pre-processed in which mainly removal of artifacts like eye blinking and environment noise are carried out. Feature extraction part tries to search for a compact feature set

which uniquely represent the task thought in the brain. The classification stage decides which task has been thought out of the two or more tasks based on the given brain features. After the classification, it can be used to carry out applications like controlling a robot to do different tasks or to write letters in the controlling interface stage.

Out of the 5 above mentioned parts of the BCI, “feature selection and extraction” from the brain signals holds a very important role in reducing the misclassification rate of a BCI system and is the focus of the present chapter. Some of the channels used for acquiring the brain signals may be noisy or irrelevant in relation to a particular motor imagery tasks. Therefore, selection of relevant channels from the total channel is required. In [111] channel selection is carried out utilizing support vector machine (SVM) and in [112] using genetic algorithms for reducing the misclassification rates. Two methods were used in [113] to choose a smaller subset of features, one was based on the information theory approach and the other utilized genetic algorithm. The genetic algorithm outperformed the information theoretic approach. Different data segment related parameters like length of the segment, the number of trials and the starting position of the segment were included in [114] for feature extraction from the brain signals. This chapter studies the motor imagery task from the same limbs. The brain-signals (EEG) data set used in this study, has been previously used for identification of motor imagery tasks, but a very limited effort has been put on feature selection and extraction for improving the classification accuracy [115]. Therefore, this chapter deals with improving the classification accuracy by proposing a feature selection strategy which consists of two steps: time-segment selection through visual inspection and channel selection through Fisher-ratio analysis in the frequency domain.

The rest of the chapter is organized as follows. The next section presents the EEG data set used in this study. Section 7.3 describes the feature selection strategy and feature extraction. Section 7.4 discusses the experiments and

results. Finally, Section 7.5 gives a summary of the chapter.

7.2 Dataset

Subjects were asked to imagine isometric plantar flexion of the right foot at different target torque (TT) and rate of torque (RTD). When the subjects were imagining these movements, their brain signals were captured using the EEG cap in which electrodes were mounted as per 10-20 system [115]. Torque applied to achieve the maximum contraction of the isometric plantar flexion is referred to as Maximal Voluntary Contraction Torque (MVCT). It is measured here by taking the average of three MVCT values. Based on the MVCT, four different types of motor imagery tasks from the right limb can be defined. TT can be ‘low’ or ‘high’. 30% of MVCT is referred to as low and 60% of MVCT as high. Similarly, RTD can be ‘ballistic’ or ‘moderate’ in accordance with how fast the TT is achieved. Achieving TT as fast as possible is termed as ballistic and in approximately 4s is termed moderate. Therefore, the following 4 types of tasks can be defined.

- BH: Imagining Ballistic movement to reach High TT.
- BL: Imagining Ballistic movement to reach Low TT.
- MH: Imagining Moderate movement to reach High TT.
- ML: Imagining Moderate movement to reach Low TT.

The subjects were given visual cues on the computer screen about which task out of the above four to be imagined and when to start the imagination process. 9 subjects in the age group of 22-33 years participated in the data collection process. The full details of the tasks and the description of the dataset can be found in [115]. In this chapter, the classification of only two

types of motor imagery tasks out of the above 4, namely, BH and BL were considered and the EEG recordings for the task were collected from 6 subjects.

7.3 Feature Selection and Extraction Strategy

This section discusses the proposed feature selection strategy and the feature extraction method from the EEG signal.

- **Feature Selection**

From the EEG signal obtained by the motor imagery tasks, the following two types of feature selection were made:

1. Segment selection from the time-domain EEG Signal

In this, the time-domain EEG signal from a particular channel for the individual classes BH and BL of all the 6 subjects were averaged and plotted as shown in Fig. 7.1. After observing all such figures, 2 segments were selected from the total signal. Both segments were 2s long. The first segment starts 1s before the “onset of the task imagination (represented by 0 in the time axis of the Fig. 7.1)” and ends after 1s of the “onset of the task imagination”. The other starts after 2s of the “onset of the task imagination”. The two segments selected were also depicted in the Fig. 7.1. These 2 segments represented the best 2s discriminative parts in the total time-domain EEG signal for the 2 classes BH and BL.

2. Channel selection from a total of 32 channels used for capturing the EEG signal.

All the 32 channels, which were used for capturing the brain activity during a motor imagery task might not contribute equal information about the imagined task. Some might contain other prominent information like eye blinking (termed artifact). Therefore, a smaller

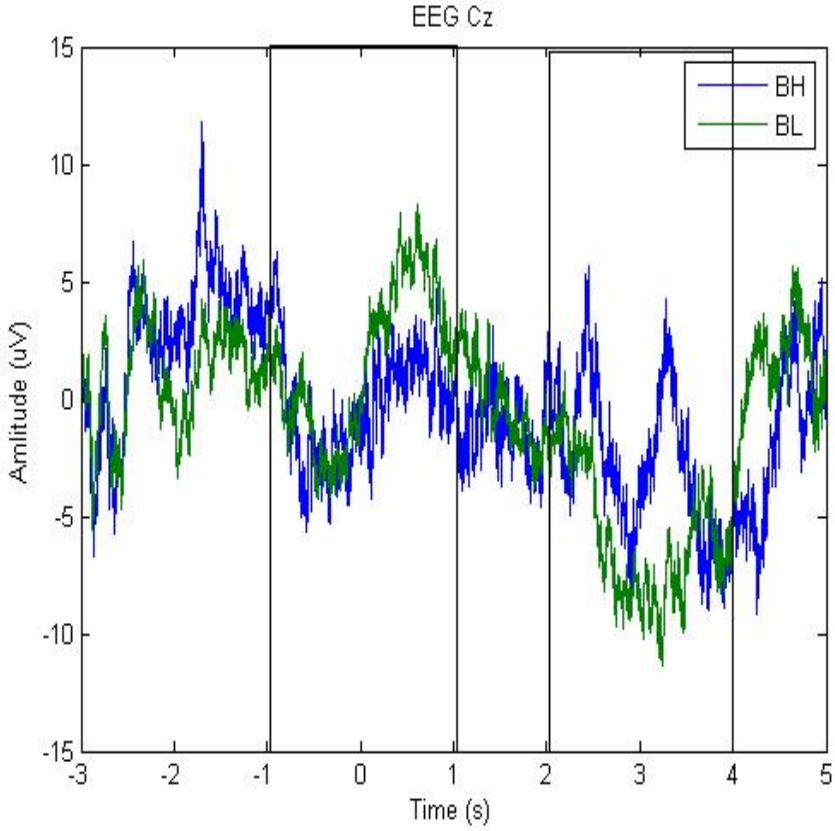


Figure 7.1: For sub F, channel Cz graph for the class BH and BL, task onset is at 0s, and the two segments extracted are from -1s to 1s and from 2s to 4s [116]

subset of channel selection which are more relevant to the imagined task may improve the classification accuracy. To perform channel selection Fisher criterion is used. The discrete wavelet transform (DWT) (briefly described below) of the EEG signal from a particular channel, say “j”, for the BH and BL class will be used to calculate the amount by which BH and BL class differs utilizing the Fisher’s ratio given by [119]:

$$FRatio_j = \frac{(m_{BH_j} - m_{BL_j})^2}{\sigma_{BH_j}^2 - \sigma_{BL_j}^2} \quad (7.1)$$

Here, m_{BH_j} and m_{BL_j} represent the mean of the samples from the BH and BL class, respectively and σ_{BH_j} and σ_{BL_j} represent the variance of the samples from the BH and BL class, respectively, for the j^{th} channel.

The channel with the highest value of the FRatio shows the highest amount of discrimination between the two class, BH and BL and is considered more reliable than other channels for classification. Based on the FRatio calculation, four subsets of channels out of the 32 channels were selected for the classification. They are the best 9, best 8, best 7 and best 6 channels.

- **Feature Extraction**

The EEG signal obtained after the feature selection step is band pass filtered in the range 7-30Hz and notch filtered at 50Hz. The DWT of the signal is then obtained. DWT gives the coefficients of the mapping of the signal into a group of basis functions obtained by the translation and scaling of the mother wavelet. Here, orthogonal wavelet is utilized. In the Multiresolution Analysis (MRA), the mother wavelet $\psi(t)$ is related to the high pass filter $g(n)$ and scaling function $\phi(t)$ is related to the low pass filter $h(n)$ by the following equations [117] [116]. :

$$\psi(t) = \sqrt{2}\sum_n g(n)\phi(t - n) \quad (7.2)$$

$$\phi(t) = \sqrt{2}\sum_n h(n)\phi(t - n) \quad (7.3)$$

Also, for orthogonal wavelets, the mother wavelet can be obtained from the $h(n)$ as $g(n)$ can be given in terms of $h(n)$ as:

$$g(n) = (-1)^{1-n}h(1 - n) \quad (7.4)$$

For the selection of the optimal mother wavelet, 21 $h(n)$ filters of length 4, parameterized with a value in the range $-\pi, \pi$ are used and is described in [118]. The DWT utilizing the mother wavelet $\psi(t)$ gives a set of detail spaces $d_x(j, k)$, partly localized in time and frequency. Instead of directly using these detail spaces as feature, we will calculate marginals $m_x(j)$ defined below, which will make it insensitive to time.

$$m_x(j) = \sum_{k=0}^{\frac{N}{2^j-1}} c_x(j, k) \quad j = 1, 2, \dots, J \quad (7.5)$$

$$c_x(j, k) = \frac{|d_x(j, k)|}{\sum_{j=1}^J \sum_{k=0}^{\frac{N}{2^j-1}} |d_x(j, k)|} \quad (7.6)$$

Here, $d_x(j, k) = \langle x(t), \psi_j(j, k)(t) \rangle$. The j term is related to the scale and k term is related to the translation by $\psi_{j,k} = 2^{-\frac{j}{2}} \psi(2^{-j}t - k)$. J is the highest decomposition level and $J = \log_2 N$, N is the total number of samples in the signal x .

After the calculation of the marginals for all the 21 mother wavelets for a particular channel. Each of these 21 feature space is individually used for classification. The mother wavelet which achieved the lowest misclassification rate is used for calculation of the feature space for the rest of the experiments.

7.4 Experiments and Results

Experiments for classifying the two classes of the motor imagery tasks, namely BH and BL, were conducted for the baseline and the proposed feature selection strategy.

The baseline system is based on the research study [115]. Here, band pass filtering of the signal in the 0.1Hz to 0.75Hz and notch filtering at 50Hz is

used. EEG signals from the 7 preselected channels, namely, F3, F4, C3, Cz, C4, P3 and P4 were used for feature extraction. Features were extracted using the DWT, which is described in the Section 7.3. This system is referred to as Baseline.

Using the proposed feature selection strategy which employed time-segment selection and channel selection by Fisher ratio, experiments were conducted utilizing the best 9, best 8, best 7 and best 6 channels as described in Section 7.3, and are referred to as best 9, best 8, best 7, and best 6 channels respectively. For performance evaluation, experiments were also conducted for all channels, which means, apart from the channel selection step by the Fisher ratio, rest all remained the same as described in Section 7.3 in all channel case.

Classification is carried out using the support vector machines, and is described in [121] for both the baseline and the proposed systems. Support vector machine is selected because of its good generalization property, insensitivity to over training and curse of dimensionality [120]. Here, Gaussian kernel function is used and a 3-fold cross validation is utilized for all the experiments.

The misclassification rate, i.e,

$$\frac{\text{the number of wrongly identified test instances}}{\text{total number of test instances}} \times 100\% \quad (7.7)$$

for the different systems discussed above employing the feature selection strategy and the baseline system are tabulated in Table 7.1. The test instances were taken from all the 6 subjects, namely, A, B, C, D, E and F.

From the table, it can be observed that, all the systems employing the proposed feature selection strategy outperformed the Baseline system on an average across the subject's misclassification rate. An absolute improvement of 7.5% in the average misclassification rate of all the subjects for the Best 7 channel is achieved over the Baseline system.

Comparing the different systems which have employed the proposed feature

Table 7.1: “Misclassification rate (%) for all the subjects” [116]

Subject	Proposed System with Feature Selection Strategy						Baseline
	best 9 channels	best 8 channels	best 7 channels	best 6 channels	7 channels of the Baseline	All channels	
sub A	13.63	12.12	12.2	9.09	28.78	24.24	33.33
sub B	44.79	42.7	44.79	48.95	50	50	40.62
sub C	2.56	1.28	1.28	0	6.41	2.56	8.97
sub D	0	0	0	2.77	0	0	8.33
sub E	15.47	16.66	16.66	13.09	16.66	16.67	22.61
sub F	33.33	31.25	27.08	33.33	27.08	41.67	33.33
<i>Average</i>	18.30±17.55	17.34±16.86	17.00±16.9	17.87±19.24	21.49±17.94	22.52±20.32	24.53 ± 13.57

selection strategy in full, i.e the Best 9, Best 8, Best 7 and Best 6 channel, it can be seen that the average misclassification rate has not shown much difference. The Best 7 channel showed the best performance amongst the 4.

Comparing the performance of the Best 7 channel with the All channels and 7 Preselected channel of the Baseline. The Best 7 channel achieved an absolute improvement of 5.52% over the All channel and 4.79% over the 7 Preselected channel of the Baseline. This points to the importance of feature selection, particularly channel selection before classification for reducing the misclassification rate

7.5 Summary

This chapter presents a feature selection strategy for the classification of the EEG signal in one of the two classes, namely, Ballistic High and Ballistic Low of the motor imagery tasks from the same limbs. It consisted of time-segment selection through visual inspection and channel selection through Fisher ratio calculation in the frequency domain. The experimental results suggest that, instead of using all channels for the classification, a smaller and more relevant subset of the channels related to the task can perform better. Therefore, feature selection and extraction is an important step for improving the classification accuracy of the BCI. Further, the EEG signal can be combined with the speaker identification task for helping in the decision making process, by providing another opinion about the speaker identity through brain signal.

Chapter 8

Conclusions & Future Work

This chapter presents the conclusions followed by a discussion on the future work.

8.1 Conclusions

This study developed efficient and robust speaker identification system under mismatched conditions. A mismatched condition occurs, when the training speech data conditions differ from the testing speech data conditions. For example, the training speech data has been recorded in a clean (noise free) office environment and the testing speech data contains environmental noise, like car noise, train noise and street noise as well.

Different types of mismatch are possible. A mismatch can occur due to environmental noise, voice disguise, handset or channel variations, emotional state of the person and if the person is having some throat infection. This study focuses on the mismatch that occurs because of the environmental noise

and voice disguise.

1. Mismatch due to environmental noises:

For test speech data containing environmental noises, a hybrid feature frame selection method from the time-domain speech signal has been developed. The hybrid technique takes into account the signal to noise ratio and it combines two different types of feature frame selection method, namely, voice activity detection (VAD) and the variable frame rate analysis (VFR) method which complements each other. It has been experimentally found that, hybrid technique efficiently captures the

- speech part rejecting the non-speech part and
- the changes in the temporal characteristics of the speech signal, like vowel and plosives.

under various noise scenarios. It also provides the flexibility to adjust the frame rate (number of frames selected per second). This flexibility will prove beneficial for handling different types of speaker identification applications.

2. Mismatch due to voice disguise:

When a person intentionally alters his/her own voice, either to sound like a target to steal the target's personal information or to hide their own identity, voice disguise occurs. To tackle the mismatch that occurs due to voice disguise, the following methods have been developed.

- In speaker identification, the normal convention is to cut the speech signal into shorter segments called frames of 25-30 ms length size with a fixed frame shift of half the frame size i.e 10-15 ms for feature extraction. Typically, a frame size of 25 ms with a frame shift of 10 ms is used. The use of this typical value has shown good results in many research studies involving mismatched conditions.

This study investigated different frame shifts, ranging from 1-10 ms, keeping the frame size fixed at 25 ms for speaker identification under voice disguise case. It has been found that, changing the frame rate can lead to an increase in the speaker identification accuracy under voice disguise. Further, a multiple model frame work has been developed. It combines features, that are extracted utilizing three different frame shifts. This resulted in improved accuracy over the fixed frame shift method.

- In a security conscious organization, chances of the usage of voice disguise for information leakage is quite high. Speaker identification system can be employed in such organizations to identify the suspects. To make the speaker identification system robust for this purpose, the effects of multistyle training has been explored. Four different types of multistyle training strategies have been developed. It has shown encouraging results over the single style training. Further, a fusion framework has been proposed, in which, out of the four, the best two training strategies has been used. It showed better performance over single style, investigated multistyle and the multiple-model methods.
- Inspired by the experimental results obtained under voice disguise, a multiple frame rate for feature extraction and reliable frame selection at the decision level has also been proposed. It has showed an overall better performance over the baseline method.

In addition to the above, a feature selection strategy for a different field of Brain Computer Interface (BCI) has also been developed. It consists of two step:

- time segment selection by visual inspection and
- channel selection utilizing Fisher ratio analysis in the frequency domain.

Feature selection in this way achieved an improvement over the baseline method.

8.2 Future Work

Future work will involve

- development of efficient speech enhancement techniques which can be combined with the proposed hybrid technique of this study for further improving the speaker identification under environmental noise.
- exploring and collecting different voice disguised speech samples from speakers and developing a dataset for facilitating speaker identification research for disguised speech.
- testing the efficiency of the proposed techniques of this study for disguised speech on the above developed dataset.
- the merging of the two field, i.e., how voice production affects the brain signals and how this knowledge can be utilized for developing more efficient and robust speaker identification system under adverse conditions.

Bibliography

- [1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Commun.*, vol. 52, no. 1, pp. 12-40, 2010.
- [2] A. K. Jain, A. Ross, S. Prabhakar, “An introduction to biometric recognition,” *IEEE Trans. Circuits Systems Video Technol.*, vol. 14, no. 1, pp. 4-20, 2004.
- [3] R. Togneri and D. Pullella, “An overview of speaker identification: Accuracy and Robustness Issues”, *IEEE Circuits Syst. Mag.*, vol.11, no.2. pp. 23-61, May 2011.
- [4] J . Campbell, “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.
- [5] G. R. Doddington, “Speaker recognition-Identifying people by their voices,” *Proc. IEEE*, vol. 73, no. 11, pp. 1651-1664, 1985.
- [6] B. S. Atal, “Automatic recognition of speakers from their voices,” *Proc. IEEE*, vol. 64, no. 4, pp. 460-475, 1976.
- [7] H. Gish and M. Schmidt, “Text-independent speaker identification,” *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 18-32, 1994.

- [8] D. O'Shaughnessy, "Speaker Recognition," *IEEE ASSP Magazine*, pp.4-17, Oct. 1986.
- [9] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans," *IEEE Signal Process. Mag.*, pp.74-99, Nov. 2015.
- [10] S. D. Deshmukh and M. R. Bachute, "Automatic Speech and Speaker Recognition by MFCC, HMM and Vector Quantization." *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 3, no.1, 2013.
- [11] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, no. 68, p. 2044, 1972.
- [12] S. Prasad, Z.-H.Tan, R. Prasad, "Multiple Frame Rates for Feature Extraction and Reliable Frame Selection at the Decision for Speaker Identification Under Voice Disguise", *Journal of Communication, Navigation, Sensing and Services (CONASENSE)*, Vol. 2016, no.1, pp-29-44, Jan. 2016, DOI: 10.13052/jconasense2246-2120.2016.003.
- [13] R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, pp. 58, 1996.
- [14] C. Zhang and T. L. Tan, "Voice disguise and automatic speaker recognition", *Forensic Sci. Inter.*, Elsevier, vol. 175, pp. 118-122, 2008.
- [15] S. S. kajarekar, H. Bratt, E. Shriberg and R. Leon, "A study of intentional voice modifications for evading automatic speaker recognition", *IEEE Odyssey*, San Juan, Puerto Rico, 2006.
- [16] M. Grimaldi and F. Cummins, "Speech style and speaker recognition: a case study", *INTERSPEECH*, Brighton, U.K, Sept.,2009.
- [17] B. S. Atal, "Automatic speaker recognition based on pitch contours." *J. Acoust. Soc. Amer.* vol. 52, no. 6B, pp. 1687-1697,1972.

- [18] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, no. 1-2, pp. 151-173, Nov. 1996.
- [19] R. Lippmann, E. Martin, D. B. Paul, "Multi-style training for robust isolated-word speech recognition," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, Apr. 1987.
- [20] M. V. Ghiurcau, C. Rusu, and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," *Proc. ICASSP*, Prague, Czech Republic, 2011
- [21] N. Jawarkar, R. Holambe, and T. Basu. "Text-independent speaker identification in emotional environments: a classifier fusion approach," *Frontiers in Computer Education*, pp. 569-576, 2012.
- [22] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 366-378, Feb. 2009.
- [23] R. Rodman and M. Powell, "Computer recognition of speakers who disguise their voice," *Proceedings of the International Conference on Signal Processing Applications and Technology (ICSPAT)*, Dallas, Texas, USA, Oct. 2000.
- [24] "AV Voice Changer Software Diamond 7.0 of AVSOFT CORP." Online: <http://www.audio4fun.com/voice-over.htm>, accessed on 14 Mar 2013.
- [25] F. Kelly, A. Drygajlo and N. Harte, "Speaker verification with long-term ageing data," *Proc. Int. Conf. Biometrics*, 2012.
- [26] S. O. Sadjadi and J. H. L. Hansen, "Blind spectral weighting for robust speaker identification under reverberation mismatch," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 22, no.5, pp. 937-945, May. 2014.

- [27] M. Ji, S. Kim, H. Kim and H-S. Yoon, "Text-independent speaker identification using soft channel selection in home robot environments," *IEEE Trans. on Cons. Elec.*, vol. 54, no. 1, Feb. 2008.
- [28] M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer, Y. Lei, "Improving speaker identification robustness to highly channel-degraded speech through multiple system fusion," *ICASSP*, Vancouver, BC, Canada, May. 2013.
- [29] L. F. Gallardo, S. Mller, M. Wagner, "Human speaker identification of known voices transmitted through different user interfaces and transmission channels," *ICASSP*, Vancouver, BC, Canada, May. 2013.
- [30] Joseph, Shijo M., and Anto P. Babu. "Wavelet energy based voice activity detection and adaptive thresholding for efficient speech coding," *International Journal of Speech Technology*, vol. 19, no.3, pp. 537-550, 2016.
- [31] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Proc. Lett.*, vol. 6, no. 1, pp. 1-3, Jan. 1999.
- [32] J. H. Chang, N.S. Kim, S.K. Mitra, "Voice activity detection based on multiple statistical models", *IEEE Trans. on Sig. Process.*, vol. 54, no. 6, pp. 1965-1976, 2006.
- [33] S.-K. Kim, et al. "Power Spectral Deviation-Based Voice Activity Detection Incorporating Teager Energy for Speech Enhancement," *Symmetry* vol. 8, no. 7, 58, 2016.
- [34] X.-K. Yang, et al., "Voice activity detection algorithm based on long-term pitch information," *EURASIP Journal on Audio, Speech, and Music Processing* vol.2016, no..1, pp. 14, 2016.
- [35] B. Atal and M. Schroeder, "Predictive coding of speech signals," *Proceedings of the 6th Int. Congress on Acoustics*), Tokyo, 1968.

- [36] K. Daqrouq, K. Y. Al. Azzawi, “Average framing linear prediction coding with wavelet transform for text-independent speaker identification system,” *Computers & Electrical Eng.*, vol. 38, no. 6, pp. 1467-1479, Nov. 2012.
- [37] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. on Acoustics, speech and signal proces.*, vol. ASSP-28, pp. 357-366, Aug. 1980.
- [38] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. H-Sandsten, H. Li, “Low-variance multitaper MFCC features: A case study in robust speaker verification,” *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 20, no. 7, pp. 1990-2001, 2012.
- [39] Q. Li and Y. Huang, “Robust speaker identification using an auditory-based feature” *ICASSP*, pp. 4514-4517, Dallas, TX, USA, Mar. 2010.
- [40] C. C. T. Chen, C. T. Chen, P. W. Cheng . “Hybrid KLT/GMM approach for robust speaker identification,” *Electron Lett.* vol.39, no.21, 2003
- [41] Z.-X. Yuan, B.-L. Xu and C. -Z. Yu, “Binary quantization of feature vectors for robust text-independent speaker identification,” *IEEE Trans. on Speech and Audio Process.*, vol. 7, pp. 70-78, 1999.
- [42] Y. Shao and D. Wang, “Robust speaker recognition using binary time-frequency masks,” *ICASSP*, pp. I-I, 2006.
- [43] A. Venturini, L. Zao, and R. Coelho. “On speech features fusion, α - integration Gaussian modeling and multi-style training for noise robust speaker classification.” *IEEE/ACM Transactions on Audio, Speech, and Lang. Process.*, vol. 22, no.12, pp. 1951-1964, 2014
- [44] C. Kim, and R. M. Stern, “Power-normalized cepstral coefficients (PNCC) for robust speech recognition.” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* vol. 24, no.7, pp. 1315-1329, 2016.

- [45] J. Guo, et al. "Robust speaker identification via fusion of subglottal resonances and cepstral features." *The Journal of the Acoustical Society of America*, vol. 141, no.4, pp. EL420-EL426, 2017.
- [46] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 2, pp. 254-272, 1981.
- [47] O. M. Strand and A. Egeberg, "Cepstral mean and variance normalization in the model domain," *ISCA Tutorial and Research Workshop*, 2004.
- [48] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-plp speech analysis technique," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, vol. 1, pp. 121-124, 1992.
- [49] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Crete, Greece, pp. 1-5, 2001.
- [50] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Comm.*, vol. 17, no.1-2, pp. 91-108, 1995.
- [51] Reynolds, D. "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," *Eurospeech*, pp. 963-967, 1997.
- [52] NIST, "March 1996 NIST speaker recognition workshop notebook." NIST administered speaker recognition evaluation on the Switchboard corpus, March 27-28, 1996.
- [53] A. Roy, M. M. Doss, and S. Marcel. "A fast parts-based approach to speaker verification using boosted slice classifiers." *IEEE Transactions on Information Forensics and Security*, vol. 7, no.1, 241-254, 2012.
- [54] H. A. Murthy, et al. "Robust text-independent speaker identification over telephone channels." *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 554-568, 1999.

- [55] S.-C. Yin, R. Rose and P. Kenny, “A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification,” *IEEE trans. Audio, Speech, and Language Process.*, vol. 15, pp. 1999-2010, 2007.
- [56] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech and Lang. Proc.*, vol.19, no. 4, pp. 788-798, May 2011.
- [57] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308-311, May 2006.
- [58] L. Zao, R. Coelho, “Colored noise based multicondition training for robust speaker identification,” *IEEE Signal Process. Lett.*, vol. 18, no. 11, pp. 675-678, 2011.
- [59] M. Ji, T. J. Hazen, J. R. Glass and D. A. Reynolds, “Robust speaker recognition in noisy conditions”, *IEEE Trans. Audio, Speech, Lang. Process.*, vol.15, pp. 1711-1723, 2007.
- [60] X. Anguera and J. F. Bonastre, “A novel speaker binary key derived from anchor models,” in *Eleventh Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, 2010
- [61] G. E. Dahl, D. Yu, L. Deng and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 30-42, 2012.
- [62] Y. Lei, N. Scheffer, L. Ferrer, M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

- [63] K. Markov and S. Nakagawa, "Frame level likelihood normalization for text-independent speaker identification using gaussian mixture models," *ICSLP*, pp. 1764-1767, vol. 3, 1996.
- [64] R. B. Dunn, T. F. Quatieri, D. A. Reynolds and J. P. Campbell, "Speaker recognition from coded speech and the effects of score normalization," in *Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, vol.2, pp. 1562-1567 , 2001.
- [65] D. J. Mashao, M. Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recog.*, vol. 39, pp. 147-155, 2006.
- [66] J. Jung, K. Kim, and M. Y. Kim. "Advanced missing feature theory with fast score calculation for noise robust speaker identification." *Electronics letters*, vol. 46, no. 14, pp. 1027-1029, 2010.
- [67] J. Yuan, and M. Liberman. "Speaker identification on the SCOTUS corpus." *Journal of the Acoustical Society of America*, vol. 123, no.5, 3878, 2008
- [68] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, "Speaker identification from shouted speech: Analysis and compensation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May. 2013.
- [69] J. H. Hansen, C. Swail, A. J. South, R. K. Moore, H. Steeneken, E. J. Cupples, T. Anderson, C. R. Vloeberghs, et al., "The impact of speech under 'stress' on military speech technology," *NATO Project Rep.*, no. 104, 2000.
- [70] H. Kunzel, "Effects of voice disguise on speaking fundamental frequency," *Forensic Linguist.*, vol. 7, no. 2, pp. 150-179, 2000.

- [71] P. Perrot, G. Aversano, G. Chollet, “Voice disguise and automatic detection: review and perspectives”, *Progress in Nonlinear Speech Processing.*, Springer, Berlin/Heidelberg, pp. 101-117, 2007.
- [72] A. Eriksson, P. Wretling, “How flexible is the human voice? - A case study of mimicry”. *Eurospeech* vol. 97 no. 2, 1043-1046, 1997.
- [73] W. Endres, W. Bambach, G. Flosser, “Voice spectrograms as a function of age, voice disguise and voice imitation.” *J. Acoust. Soc. America*, vol. 49, no. 6, 1842-1848, 1971
- [74] A. S. Danko and G. C. Fernandez. “My brain is my passport. Verify me.” *IEEE International Conference on Consumer Electronics (ICCE)*, Las Vegas , NV, USA, 2016.
- [75] P. Belin, S. Fecteau and C. Bedard. “Thinking the voice: neural correlates of voice perception.” *Trends in cognitive sciences*, vol.8, no. 3, 129-135, 2004.
- [76] K. Brigham and B. V. K. V. Kumar. “Subject identification from electroencephalogram (EEG) signals during imagined speech.” *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, Washington, DC, USA, 2010.
- [77] M. Graciarena, S. Kajarekar, A. Stolcke and E. Shriberg, “Noise robust speaker identification for spontaneous Arabic speech,” in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, vol. 4, pp. 245-248, Honolulu, Hawaii, U.S.A, Apr. 2007,
- [78] C.-S. Jung, M. Y. Kim , H. -G. Kang, “Selecting feature frames for automatic speaker recognition using mutual information,” *IEEE Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1332-1340, 2010.

- [79] G. Sarkar, G. Saha, "Real time implementation of speaker identification system with frame picking algorithm," *Procedia Computer Science*, vol. 2, pp. 173-180, 2010.
- [80] R. Saeidi, H. R. S. Mohammadi, R. D. Rodman, and T Kinnunen, "A new segmentation algorithm combined with transient frames power for text independent speaker verification," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (ICASSP)*, Honolulu, HI, USA., 2007
- [81] S. Deng, J. Han, "Likelihood ratio sign test for voice activity detection," *IET Signal Process.*, vol. 6, no. 4, pp. 306-312, 2012.
- [82] M. -W. Mak, H. -B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations", *Computer Speech and Lang.*, vol. 28, pp. 295-313, 2014.
- [83] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata and H. G. Okuno, "Speaker identification under noisy environment by using harmonic structure extraction and reliable frame weighting", in *Proceedings of Interspeech*, Pittsburgh, Pennsylvania, pp. 1459-1462, Apr. 2006.
- [84] Z. -H. Tan, B. Lindberg, "Low complexity frame rate analysis for speech recognition and voice activity detection," *IEEE J. Sel. Signal Process.*, vol. 4, no. 5, pp. 798-807, 2010.
- [85] S. Prasad, Z.-H.Tan, R. Prasad, "Feature frame selection for robust speaker identification: A hybrid approach", *Wireless Personal Communications*, pp.1-18, May 2017, Springer, DOI: 10.1007/s11277-017-4544-1.
- [86] J. P. Campbel, Jr., "Testing with YOHO cd-rom verification corpus," in. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.341-344, Detroit, Michigan, U.S.A, May., 1995.

- [87] H. G. Hirsch, D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *in: Proceedings of ISCA ITRW ASR*, 2000.
- [88] H. G. Hirsch, "FaNT Filtering and noise adding tool", [Online]. Available: <http://aurora.hsnr.de/download.html>.
- [89] Z.-H. Tan, I. Kraljevska, "Joint variable frame rate and length analysis for speech recognition under adverse conditions," *Computers and Electrical Engg.*, vol. 40, pp. 2139-2149, 2014.
- [90] J. M-Guarasa, J. Ordonez, J. M. Montero, J. Ferreiros, R. Cordoba, L.F.D Haro, "Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition," *in: Proceedings of Eurospeech*, Geneva, Switzerland, Sept. 2003.
- [91] Q. Zhu, A. Alwan, "On the use of variable frame rate analysis in speech recognition," *in: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Jun. 2000.
- [92] Q. Jin, A. R. Toth, W. A. Black and T. Schultz, "Is voice transformation a threat to speaker identification," *in ICASSP*, Las Vegas, Nevada, U.S.A, Mar. 2008.
- [93] B. Pellom and J. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," *in ICASSP*, Phoenix, Arizona, Mar. 1999.
- [94] H. Masthoff, "A report on voice disguise experiment," *Forensic Ling.*, vol.3, no. 1, pp. 160-167, 1996.
- [95] M. Farrus, M. Wagner, J. Anguita and J. Hernando, "Robustness of prosodic feature to voice imitation," *in INTERSPEECH*, Brisbane, Australia, Sept., 2008.

- [96] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Proc.*, vol. 3, no. 1, pp. 72-83, 1995.
- [97] F. Cummins, M. Grimaldi, T. Leonard and J. Simko, "The CHAINS corpus: Characterizing individual speakers," in *Proc. SPECOM*, St. Petersburg, Russia, Jun. 2006.
- [98] F. Cummins, "Practice and performance in speech produced synchronously," *J. of Phonetics*, vol. 31, no. 2, pp. 139-148, Oct. 2003.
- [99] H. Larson, "Experiences of large implementation of speech analyzing tools in swedish as second language," in *MATISSE-ESCA/SOCRATES*, London, U.K, 1999.
- [100] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "The HTK book version 3.4," Cambridge University Engineering Department, p.64. Online: <http://htk.eng.cam.ac.uk>, accessed on 23 Sept. 2017.
- [101] H. Xu, Z. -H. Tan, P. Dalsgaard and B. Lindberg, "Robust Speech Recognition Based on Noise and SNR Classification-a Multiple-Model Framework," in *INTERSPEECH*, Lisbon, Portugal, Sept. 2005.
- [102] R.D. Rodman, "Speaker recognition of disguised voices: A program for research," in *proc. Consortium on Speech Technology in Conjunction with the Conference on Speaker Recognition by Man and Machine: Directions for Forensic Applications COST250*, Ankara, Turkey, 1998.
- [103] S. Prasad and R. Prasad, "Reliable frame selection for speaker identification under voice disguise scenario," In *Wireless VITAE*, Hyderabad, India, 2015.
- [104] S. Prasad, Z. - H. Tan and R. Prasad, "Multi-frame rate based multiple-model training for robust speaker identification of disguised voice" In *16th*

International Wireless Personal Multimedia Communications (WPMC),
New Jersey, USA, 2013.

- [105] S. Prasad, Z. -H. Tan, R. Prasad. "Multistyle training and fusion for speaker identification of disguised voice," In *1st International Conference on Communications, Connectivity, Convergence, Content and Cooperation (IC5)*, Mumbai, India, Dec. 2013.
- [106] S. Prasad, R. Prasad, "Fusion multistyle training for speaker identification of disguised speech", submitted to *Wireless Personal Communications*.
- [107] L. F N. Alonso and J. G. Gil, "Brain Computer Interfaces, a Review," *Sensors*, vol-12, pp.1211-1279, 2012.
- [108] J. R. Wolpaw, N. Birbaumer, D. J McFarland, G. Pfurtscheller, T.M. Vaughan., "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, pp. 767-791, 2002.
- [109] D. Garrett, D. A. Peterson, C. W. Anderson, M. H. Thaut., "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification," *IEEE transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no.2 , pp.144, 2003.
- [110] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche and B. Arnaldi, "A review of classification algorithms for EEG-based braincomputer interfaces," *Journal of Neural Engineering*, vol. 4, pp. R1-R13, 2007.
- [111] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer and B. Scholkopf, "Support Vector channel selection in BCI," *IEEE transactions on Biomedical Engineering*, vol. 51, no. 6, 2004.
- [112] Q. Wei, W. Tu, "Channel selection by Genetic algorithms for classifying Single -Trial ECoG during Motor Imagery," *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vancouver, Canada, 2008.

- [113] E. Y. Tov and G. F. Inbar, "Feature Selection for the Classification of Movements from Single Movement-Related Potentials," *IEEE transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 3, Sept. 2002.
- [114] H. Cao, B. Walter, J. Steven and Z. Peng, "Individualization of data-segment-related parameters for improvement of EEG signal classification in brain-computer interface," *Transactions of Tianjin University*, vol. 16, pp. 235-238, 2010.
- [115] Y. Gu, O. F. D. Nascimento, M. -F. Lucas, D. Farina "Identification of task parameters from movement-related cortical potentials," *Medical and Biological Engineering and Computing*, vol. 47, pp. 1257-1264, 2009.
- [116] S. Prasad, Z.-H. Tan, R. Prasad, A. R. Cabrera, Y. Gu, K. Dremstrup, "Feature selection strategy for classification of single trial EEG elicited by motor imagery," *14th Wireless Personal Multimedia Communications (WPMC)*, Brest, France, Oct. 2011.
- [117] A. F. Cabrera, D. Farina and K. Dremstrup, "Comparision of Feature selection and classification methods for a brain- computer interface driven by non-motor imagery," *Medical and Biological Engineering and Computing*, vol. 48, pp. 123-132, 2010.
- [118] D. Farina, O. Nascimento, M. Lucas, C. Doncarli, "Optimization of wavelets for classification of movement-related cortical potentials generated by variation of force-related parameters," *Journal of Neuroscience Methods*, vol. 162, pp. 357-363, 2007.
- [119] P. Xiaomei and Z. Chongxun, "Classification of Left and Right Hand Motor Imagery Tasks Based on EEG Frequency Component Selection," *2nd International Conference on Bioinformatics and Biomedical*, Shanghai, 2008

- [120] A. Bashashati, M. Fatourehchi, R. K. Ward and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *Journal of Neural Engineering*, vol. 4, pp. R32-R57, 2007.
- [121] C. W. Hsu, C. C. Chang, C. J. Lin "A practical guide to support vector classification," Department of Computer Science and Information Engineering, National Taiwan University, 2003.



SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Feature frame selection for robust speaker identification: A hybrid approach
Authors:	Swati Prasad, Zheng-Hua Tan, Ramjee Prasad

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: Wireless Personal Communications, Springer, May 2017, pp. 1-18, doi:10.1007/s11277-017-4544-1.

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A. Has essentially done all the work
- B. Major contribution
- C. Equal contribution
- D. Minor contribution
- E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	B
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	A

Signatures of the co-authors

Date	Name	Signature
	Zheng Hua-Tan	
	Ramjee Prasad	

Date: 12/7/17

In case of further co-authors please attach appendix

Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.



SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Multiple Frame Rates for Feature Extraction and Reliable Frame Selection at the Decision for Speaker Identification Under Voice Disguise
Authors:	Swati Prasad, Zheng Hua-Tan, Ramjee Prasad

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: Journal of Communication, Navigation, Sensing and Services (CONASSENSE), Jan. 2016, doi: 10.13052/jconasense2246-2120.2016.003.

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A. Has essentially done all the work
- B. Major contribution
- C. Equal contribution
- D. Minor contribution
- E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	A

Signatures of the co-authors

Date	Name	Signature
	Zheng Hua-Tan	
	Ramjee Prasad	

Date: 10/7/17

In case of further co-authors please attach appendix

Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.



SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Multistyle training and fusion for speaker identification of disguised voice
Authors:	Swati Prasad, Zheng Hua-Tan, Ramjee Prasad

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: 1st International Conference on Communications, Connectivity, Convergence, Content and Cooperation (IC5), Mumbai, India, Dec. 2013

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A. Has essentially done all the work
- B. Major contribution
- C. Equal contribution
- D. Minor contribution
- E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	B
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	A

Signatures of the co-authors

Date	Name	Signature
	Zheng Hua-Tan	
	Ramjee Prasad	

Date: 10/7/12

In case of further co-authors please attach appendix

Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.



SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITY

Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Multi-frame rate based multiple-model training for robust speaker of disguised voice
Authors:	Swati Prasad, Zheng Hua-Tan, Ramjee Prasad

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: 16th Wireless Personal Multimedia Communications (WPMC), Atlantic City, New Jersey, U.S.A., Jun. 2013. IEEE Press

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A. Has essentially done all the work
- B. Major contribution
- C. Equal contribution
- D. Minor contribution
- E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	A

Signatures of the co-authors

Date	Name	Signature
	Zheng Hua-Tan	
	Ramjee Prasad	

Date: 10/7/13

In case of further co-authors please attach appendix

Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.



SCHOOL OF BUSINESS AND SOCIAL SCIENCES
BSS

Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Feature selection strategy for classification of single trial EEG elicited by motor imagery
Authors:	Swati Prasad, Zheng Hua-Tan, Ranjee Prasad, Alvaro Fuentes Cabrera, Ying Gu, Kim Dremstrup

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: 14th Wireless Personal Multimedia Communications (WPMC), Brest, France, Oct 2011. IEEE Press

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A Has essentially done all the work
- B Major contribution
- C Equal contribution
- D Minor contribution
- E Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	B

Signatures of the co-authors

Date	Name	Signature
	Zheng Hua-Tan	
	Ranjee Prasad	
	Alvaro Fuentes Cabrera	
	Ying Gu	
	Kim Dremstrup	

Date: 10/12/12

In case of further co-authors please attach appendix

Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation

Please see the next page for the remaining co-author's signature.



Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Feature selection strategy for classification of single trial EEG elicited by motor imagery
Authors:	Swati Prasad, Zheng Hua-Tan, Ramjee Prasad, Alvaro Fuentes Cabrera, Ying Gu, Kim Dremstrup

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: 14th Wireless Personal Multimedia Communications (WPMC), Brest, France, Oct. 2011. IEEE Press

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A. Has essentially done all the work
- B. Major contribution
- C. Equal contribution
- D. Minor contribution
- E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	B

Signatures of the co-authors

Date	Name	Signature
	Zheng Hua-Tan	
	Ramjee Prasad	
	Alvaro Fuentes Cabrera	
	Ying Gu	
	Kim Dremstrup	

In case of further co-authors please attach appendix

Date: 10/7/12

Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.

Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Reliable frame selection for robust speaker identification under voice disguise scenario
Authors:	Swati Prasad, Ramjee Prasad

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: 5th International Conference on Wireless Communications, Vehicular Technology, Information Theory, and Aerospace & Electronics Systems (Wireless VITAE), Hyderabad, India, Dec. 2015

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

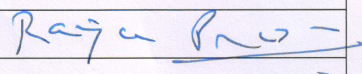
No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A. Has essentially done all the work
- B. Major contribution
- C. Equal contribution
- D. Minor contribution
- E. Not relevant

Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	A

Signatures of the co-authors

Date	Name	Signature
	Ramjee Prasad	

Date: 3-11-17

Signature of the PhD student

In case of further co-authors please attach appendix

*As per policy the co-author statement will be published with the dissertation.

Declaration of co-authorship*

Full name of the PhD student: Swati Prasad

This declaration concerns the following article/manuscript:

Title:	Fusion Multistyle Training for the Speaker Identification of Disguised Speech
Authors:	Swati Prasad, Ramjee Prasad

The article/manuscript is: Published ☐ Accepted ☐ Submitted ☒ In preparation ☐

If published, state full reference:

If accepted or submitted, state journal: Wireless Personal Communications (Springer)

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

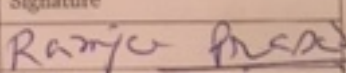
No ☒ Yes ☐ If yes, give details:

The PhD student has contributed to the elements of this article/manuscript as follows:

- A. Has essentially done all the work
- B. Major contribution
- C. Equal contribution
- D. Minor contribution
- E. Not relevant

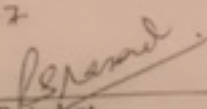
Element	Extent (A-E)
1. Formulation/identification of the scientific problem	A
2. Planning of the experiments/methodology design and development	A
3. Involvement in the experimental work/clinical studies/data collection	A
4. Interpretation of the results	A
5. Writing of the first draft of the manuscript	A
6. Finalization of the manuscript and submission	A

Signatures of the co-authors

Date	Name	Signature
	Ramjee Prasad	

In case of further co-authors please attach appendix

Date: 8/12/17


Signature of the PhD student

*As per policy the co-author statement will be published with the dissertation.