The Use of Spectral Information in the Development of Novel Techniques for Speech-Based Cognitive Load Classification

A thesis submitted for the degree of

Doctor of Philosophy

By

Phu Ngoc Le

Supervisor: **Prof. Eliathamby Ambikairajah**

Co-supervisors: **Dr. Julien Epps**

Dr. Eric Choi

School of Electrical Engineering and Telecommunications

The University of New South Wales

January 2012

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: LE

First name: PHU

Other name/s:

Abbreviation for degree as given in the University calendar: PhD

School: Electrical Engineering and Telecommunications

Faculty: Engineering

Title: The Use of Spectral Information in the Development of Novel Techniques for Speech Based Cognitive Load Classification

Abstract 350 words maximum: (PLEASE TYPE)

The cognitive load of a user refers to the amount of mental demand imposed on the user when performing a particular task. Estimating the cognitive load (CL) level of the users is necessary to adjust the workload imposed on them accordingly in order to improve task performance. The current speech based CL classification systems are not adequate for commercial use due to their low performance particularly in noisy environments. This thesis proposes many techniques to improve the performance of the speech based cognitive load classification system in both clean and noisy conditions.

This thesis analyses and presents the effectiveness of speech features such as spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) for CL classification. Sub-systems based on SCF and SCA features were developed and fused with the traditional Mel frequency cepstral coefficients (MFCC) based system, producing an 8.9% and 31.5% relative error rate reduction respectively when compared to the MFCC-based system alone. The Stroop test corpus was used in these experiments.

The investigation into cognitive load information in the form of spectral distribution in different subbands shows that the information distributed in the low frequency subband is significantly higher than the high frequency subband. Two different methods are proposed to utilize this finding. The first method, called the multi-band approach, uses a weighting scheme to emphasize the speech features in low frequency subbands. The cognitive load classification accuracy of this approach is shown to be higher than a system based on a non-weighting scheme. The second method is to design an effective filterbank based on the spectral distribution of cognitive load information using the Kullback-Leibler distance measure. It is shown that the designed filterbank consistently provides higher classification accuracies than other existing filterbanks such as mel, Bark, and equivalent rectangular bandwidth.

A discrete cosine transform based speech enhancement technique is proposed in order to increase the robustness of the CL classification system and found to be more suitable than other methods investigated. This proposed method provides a 3.0% average relative error rate reduction for the seven types of noise and five levels of SNR used. In particular, it provides a maximum of 7.5% relative error rate reduction for the F16 noise (in NOISEX-92 database) at 20 dB SNR.

Keywords: Automatic cognitive load classification, cognitive load information distribution, filterbank designing, multi-band, weighting, speech enhancement.

Declaration relating to disposition of project thesis/dissertation

Signature

Eugene Chen
Witness

21/2/2012
Date

FOR OFFICE USE ONLY

Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

COPYRIGHT STATEMENT

'I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International (this is applicable to doctoral theses only).

I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.'

Signed

Date 21/2/2012

AUTHENTICITY STATEMENT

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.'

Signed

Date 21/2/2012

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ....

Date 21/2/2012

Abstract

The cognitive load of a user refers to the amount of mental demand imposed on the user when performing a particular task. Estimating the cognitive load (CL) level of the users is necessary to adjust the workload imposed on them accordingly in order to improve task performance. The current speech based CL classification systems are not adequate for commercial use due to their low performance particularly in noisy environments. This thesis proposes many techniques to improve the performance of the speech based cognitive load classification system in both clean and noisy conditions.

This thesis analyses and presents the effectiveness of speech features such as spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) for CL classification. Sub-systems based on SCF and SCA features were developed and fused with the traditional Mel frequency cepstral coefficients (MFCC) based system, producing an 8.9% and 31.5% relative error rate reduction respectively when compared to the MFCC-based system alone. The Stroop test corpus was used in these experiments.

The investigation into cognitive load information in the form of spectral distribution in different subbands shows that the information distributed in the low frequency subband is significantly higher than the high frequency subband. Two different methods are proposed to utilize this finding. The first method, called the multi-band approach, uses a weighting scheme to emphasize the speech features in low frequency subbands. The cognitive load classification accuracy of this approach is shown to be higher than a system based on a non-weighting scheme. The second method is to design an effective filterbank based on the spectral distribution of cognitive load information using the Kullback-Leibler distance measure. It is shown that the designed filterbank consistently provides higher classification accuracies than other existing filterbanks such as mel, Bark, and equivalent rectangular bandwidth.

A discrete cosine transform based speech enhancement technique is proposed in order to increase the robustness of the CL classification system and found to be more suitable than other methods investigated. This proposed method provides a 3.0% average relative error rate reduction for the seven types of noise and five levels of SNR used. In particular, it provides a maximum of 7.5% relative error rate reduction for the F16 noise (in NOISEX-92 database) at 20 dB SNR.

Keywords: Automatic cognitive load classification, cognitive load information distribution, filterbank designing, multi-band, weighting, speech enhancement.

Acknowledgements

I would like to express my sincere thanks to my supervisor Professor Eliathamby Ambikairajah for his invaluable guidance, encouragement, and technical support. I would also like to thank to my co-supervisors, Dr. Eric Choi and Dr. Julien Epps for their technical support and help in revising and correcting my technical writing.

From our speech research group, I would like to thank Dr. Vidhyasaharan Sethu and Dr. Tharmarajah Thiruvaran for many valuable discussions as well as their help in proof reading my thesis. I would also like to thank Dr. Mohaddesh Nosratighods, Dr. Bo Yin, Dr. Teddy Gunawan for many technical discussions and valuable suggestions. I wish to thank Mr Tet Yap and Ms Karen Kua for their help in proof reading some parts of my thesis. I would like to extend my thanks to other members of our research group, Dr. Mahmood Akhtar, Dr. Liang Wang, Dr. Ning Wang, Dr. Ronny Kurniawan, and Ms Phyu Khing for their support. I would also like to thank all members of the Image Signal and Information Processing group at UNSW for their friendship and thank Mr. Tom Millet for organizing a warm and friendly working environment for us.

I would like to thank Ms Raji Ambikairajah and Ms Stefanie Brown for their assistance in editing and proof reading this thesis.

I wish to acknowledge the Vietnamese government for funding my research. I also wish to acknowledge the National Information Communication Technology Australia (NICTA) and Graduate Research School at UNSW for the additional funding they provided. This research would not have been possible without all of this financial support. I also wish to thank the School of Electrical Engineering and Telecommunications at UNSW for providing me with travel support to attend conferences.

I wish to acknowledge the International Research Center Multimedia Information Communication and Application (MICA), Vietnam for giving me an opportunity to visit and work for a short-term at their center during my internship.

Finally, I would like to express my sincere thanks to my parents, L. Man and T. Cam, and my sister, L. Tai for their endless love, support and encouragement.

List of publications

Journal paper

1. **Le, P. N**., E. Ambikairajah, J. Epps, V. Sethu, E. H. C. Choi, (2011) "*Investigation of spectral centroid features for cognitive load classification"*, Speech Communication, Vol. 53, Issue 4, April 2011, pp 540-551

Conference papers

1. **Le, P. N.**, V. Sethu, E. Ambikairajah, Kua, J. M. K., (2011) "*Investigation of the Robustness of a Non-Uniform Filterbank for Cognitive Load Classification*", *in* Proc. of the 8th International Conference on Information and Comunication System (ICICS) Singapore, Dec. 2011.

2. **Le, P. N.**, J. Epps, E. Ambikairajah, V. Sethu, (2010) *"Robust Speech-Based Cognitive Load Classification Using a Multi-band Approach*", in Proc. of the Second APSIPA Annual Summit and Conference, Biopolis, Singapore, 2010, pp 400-404.

3. **Le, P. N.**, J. Epps, E. H. C. Choi, and E. Ambikairajah, (2010) "*A study of voice source and vocal tract filter based features in cognitive load classification*," in Proc. of the 20th International Conference on Pattern Recognition, Istanbul Turkey, 2010, pp 4516-4519.

4. **Le, P. N**., E. Ambikairajah, E. H. C. Choi, J. Epps, (2009) *"A Non-Uniform Subband Approach to Speech-Based Cognitive Load Classification" in* Proc. of the 7th International Conference on Information and Comunication System (ICICS), Macau, Dec. 2009.

5. **Le, P. N.**, E. Ambikairajah, V. Sethu, (2008) *"Speech Enhancement Based On Empirical Mode Decomposition***",** in Proc. of the IASTED International Conference on Signal Processing, Pattern Recognition and Applications, February 2008, at Innsbruck, Austria, pp. 207-210.

6. **Le, P. N.**, E. Ambikairajah, E. Choi, (2008) "*An Improved Soft Threshold Method for DCT Speech Enhancement*", *in* Proc. of the Second International Conference on Communication and Electronics, Hoian, Vietnam 2008, pp 268 - 271.

7. **Le, P. N.**, E. Ambikairajah, (2007) “N*on-Uniform Sub-Band Kalman Filtering for Speech Enhancement*”, *in* Proc. of International Conference on Signal Processing and Communication System (ICSPCS), Gold coast Australia, 2007.

8. **Le, P. N.**, E. Ambikairajah, E. Choi, (2009) “*Improvement of Vietnamese Tone Classification using FM and* MFCC *Features*”, presented at the IEEE-RIVF International Conference on Computing and Communication Technologies, Danang, Vietnam 2009, pp 140-143.

Acronyms and Abbreviations

AR	Autoregressive
CL	Cognitive load
DCT	Discrete Cosine Transform
EMD	Empirical Mode Decomposition
ERB	Equivalent Rectangular Bandwidth
FF	Formant frequency
FFT	Fast Fourier Transform
FM	Frequency Modulation
FMFCC	Filter Mel Frequency Cepstral Coefficients
GD	Group Delay
GMM	Gaussian Mixture Model
KL	Kullback-Leibler
IMF	Intrinsic Mode Function
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstral Coefficients
PESQ	Perceptual Evaluation of Speech Quality
SCF	Spectral Centroid Frequency
SCA	Spectral Centroid Amplitude
SDF	Shifted Delta Feature
SI	Spectral Intercept
SMFCC	Source Mel Frequency Cepstral Coefficients
SNR	Signal to Noise Ratio
SS	Spectral Slope
SVM	Support Vector Machines
UBM	Universal Background Model

Contents

List of Figures

List of Tables

Chapter 1: Introduction

In modern society, people are faced with working environments that are increasingly demanding. Task environments are becoming more complex and time constraints are increasing. In environments such as a call center and when driving a vehicle, users often need to manage a large amount of information and can easily become overloaded. In other words, they are unable to process all relevant information necessary to perform the task at hand, which can lead to unproductive or dangerous situations. It is therefore desirable to design a system to extract data related to the workload of the users. This data can then be used to respond intelligently and adaptively based on user information processing capacity in order to avoid an overload situation and improve task performance.

The cognitive load (CL) of a person refers to the amount of mental demand imposed on that person when performing a particular task. It reflects the amount of pressure the person experiences in completing a task. Cognitive load has been closely associated with the limited capacity of the human working memory. It is known that the amount of working memory resources devoted to a particular task greatly affects the task performance. In particular, task performance has been shown to degrade by either overload or underload. This can be attributed to task demands that exceed the available cognitive capacity in the former case, or the inadequate allocation of cognitive resources in the latter [1]. As a result, it is necessary to measure a user's cognitive load, or classify it along an ordinal scale, in order to adjust the workload so that the load experienced is maintained within an optimal range for maximum productivity.

- There are many potential applications for a cognitive load measurement system. For example, transportation vehicles are equipped with an increasing number of functions and services, which drivers are required to understand and operate. Consequently, drivers are subjected to an increasing amount of information such as navigation, traffic information, news, speed limit warnings and parking guidance. This information can be distracting and might place drivers at a very high workload which will have an adverse effect on their driving ability and general road safety. In this respect, real-time measurement of the driver's cognitive load will potentially be very useful in the design and development of intelligent in-vehicle systems. Such systems can adapt to a driver's CL level by controlling the amount of information displayed to them, in order to provide them with the best level of driving support and thus reduce the possibility of an overload situation. For instance, if the driver's load level is very high, the system can stop playing the news to reduce distraction for the driver. It can also play a warning message or recommend the driver to stop and revive if the high load level might not allow them to continue driving safely.

- In computer-based learning, where learning materials are presented by a computer, a student will acquire knowledge through the methods that are most conductive for individual learning such as video, audio, graphics and animation. If student's CL level can be measured in real-time, the computer can adapt to it by changing the presentation of the learning materials to ensure that the student's understanding is maximized. For instance, if the student's cognitive load level is too high, implying that they find it difficult to understand the material presented, the computer can provide supplementary information and examples to help and support them. It can also reduce the presentation speed of the material so that the student will have enough time to process the material better.
- In a call center, the agents are often required to manage a high volume of complex information when answering customer queries and providing customer support. As such, they are under high cognitive load. In cases when the agents' level of cognitive load is very high, they may communicate with the customer ineffectively which may result in the customer dissatisfaction. If the agents' cognitive load level can be measured, the agent support system can reduce or eliminate such problems by transferring phone calls from agents with very high CL to agents with lower CL and hence improve overall customer satisfaction.

Due to potential use in real-world applications, cognitive load measurement has been an active research area in the last couple of decades. Many methods have been proposed to measure the cognitive load level, including methods based on physiological technique, behavioral technique, performance technique, and self-reported subjective ranking of the experienced load level. The method based on speech features that represent cognitive load can be considered as belonging to either physiological or behavioral techniques (see Section 2.3), has attracted the attention of many researchers in the last few years [2-5]. This is because speech data exists in many real-life tasks e.g. telephone conversations and voice control systems, and can be easily collected in non-intrusive and inexpensive ways. In addition, Yin et al. has shown that the cognitive load level can be measured in real-time using frame-based acoustic speech features [5].

1.1 Speech based cognitive load classification

Speech is a natural form of communication for human beings. Although the main objective of speech is to convey linguistic information, this is not the only information conveyed by speech. Other information including speaker identity and mental state related information such as cognitive load is also conveyed in speech [5]. Speech is an acoustic signal, generated by the airflow from the lungs considered to be the voice source which then passes through to the pharynx and the oral and nasal cavities, collectively known as the vocal tract filter. The parameters of the voice source and the vocal tract filter vary according to the content of the utterance to be pronounced as well as the mental state of the speaker. Speech processing research can typically be regarded as the effort to determine the parameters which best convey the information in speech, and then apply that information in a practical system.

As mentioned before, cognitive load characterizes the mental workload of a person. It has been shown that the physiological consequences of the mental workload include respiratory changes e.g. increased respiration rate, irregular breathing and increased muscle tension of the vocal cords and the vocal tract [6]. Increased muscle tension of the speech production organs can adversely affect the quality of speech. This suggests that the cognitive load information can be conveyed in speech, which in turn can be characterized by the parameters of the different components of the human speech production system. This suggests the existence of patterns in speech which characterize the load level being conveyed. These patterns may exist in many types of speech features such as prosodic and acoustic features.

The purpose of an automatic speech-based CL classification system is to extract features that are representative of the patterns in speech that characterize the cognitive state of the speaker and then automatically determine the speaker’s load level using pattern classification techniques. These techniques are used to make decisions about the cognitive load level, based on the chosen features.

The usefulness of cognitive load classification for industrial applications depends on a number of factors. Amongst them, the classification accuracy is a very crucial factor. Since the measured load level is used to adjust the amount of workload imposed on the user, an inaccurate measurement would result in an inappropriate adjustment of workload, and hence degrade the performance of the system. For example, if the actual load level of the user is very high, the workload imposed should be reduced in order to avoid an overload situation. However, should an inaccurate measurement of the level indicate a low load level, the system would increase the amount of workload imposed on the user which can generate a dangerous situation. Furthermore, the cognitive load level is usually measured in working environments such as in airports, over telephone channels, and in cars where speech is corrupted by background noise. This can significantly degrade the performance of the system. These factors suggest that it is crucial to develop a cognitive load classification system that performs well, especially in noisy conditions.

1.2 Thesis objective

The principle objective of this thesis is to propose techniques to improve the performance of an existing automatic cognitive load classification system based on speech features and to increase the robustness of the system under noisy conditions. This objective may be expressed in terms of the following aims:

- To investigate the use of various speech features for an automatic CL classification system, specifically those which are complementary to the Mel frequency cepstral coefficients (MFCC) feature used in the existing systems.
- To investigate the spectral distribution of cognitive load information across different frequency bands.
- To propose techniques to improve the performance of the automatic CL classification system by emphasizing the cognitive load information in the frequency region where it is concentrated. One technique is to develop the system based on subband speech features, called a multi-band system, and then employ weighting schemes to emphasize the subband which contains the most CL information. Another technique is to design effective filterbank to extract spectral features specifically for CL classification by increasing the frequency resolution in the region that contains most of the cognitive load information.
- To introduce speech enhancement methods that will improve the quality of speech in noisy conditions in order to make the cognitive load classification system more robust to noise.

1.3 Organization of the thesis

The remainder of the thesis is organized as follows:

Chapter 2: provides an overview of cognitive load and cognitive load theory, the benefits of CL measurement, the existing techniques used to measure CL and the effect of the variation of load level on speech features. This is followed by a review of speech features that have been used in cognitive load classification and an overview of the classification system itself. Finally, it describes the two cognitive load corpora used in this thesis.

Chapter 3: begins with a description of the source-filter model of human speech production system. This is followed by the implementation of a human listening test to investigate the types of speech cues that are used by humans to identify different cognitive load levels. It then studies the effectiveness of various speech features related to the source only, the filter only or both of these components for CL classification. This study aims to provide a method for designing an effective front-end for the classification system and evaluate which component of the source-filter model contributes more to the characterization of cognitive load. Finally, the effectiveness of the spectral centroid frequency and spectral centroid amplitude features for cognitive load classification and their ability to complement the existing MFCC system are analyzed and presented.

Chapter 4: investigates the performance of different weighting schemes for the multi-band cognitive load classification system. It then studies the effectiveness of the multi-band approach and compares it with the traditional full-band approach. The studies in this chapter are carried out in both clean and noisy conditions.

Chapter 5: studies the effect of varying the spectral feature dimensions on the performance of the classification system in order to find the optimum feature vector dimension producing the highest system classification accuracy. It then investigates the distribution of cognitive load information across different subbands. Finally, this chapter designs effective filterbanks to extract spectral features specifically for cognitive load classification, based on the distribution of cognitive load information. The number of filters in the designed filterbank is chosen in order to optimize the dimension of the spectral feature vector.

Chapter 6: proposes two novel speech enhancement methods (based on Kalman filtering and empirical mode decomposition) and one approach to improve an existing speech enhancement method based on the discrete cosine transform. The effectiveness of these methods, in terms of perceptual evaluation of speech quality (PESQ), is investigated and compared to other traditional speech enhancement methods. In addition, their computation complexities are analyzed. The method providing the best compromise between the quality of enhanced speech and computation complexity will be chosen in order to improve the quality of speech in noisy conditions and make the system more robust to noise.

Chapter 7: summarizes the contributions of the thesis. Finally, it presents possible future research avenues that can be investigated following the results shown in this thesis.

1.4 Major contributions

The major contributions of this thesis are the development of several techniques to improve the performance of automatic speech-based cognitive load classification systems in both clean and noisy conditions. These major contributions, together with other contributions, are summarized below:

- An investigation of the effectiveness of speech features related to either the voice-source or the vocal tract filter for CL classification. It was found that although the features relating to the vocal tract filter are more effective, both types of feature are effective for classifying cognitive load.
- The human listening test carried out on a subset of the Stroop test corpus indicated that the breath pattern, speech rate, the use of fillers, and intonation are among the most important cues that humans use to recognize cognitive load level.
- The spectral centroid features, namely spectral centroid frequency (SCF) and spectral centroid amplitude (SCA), have been investigated for CL classification. It has been shown that they complement the traditional MFCC feature.
- The effect of varying the dimensionality of SCF, SCA and MFCC features to the classification system accuracy was investigated and the optimum dimensionality of the feature vectors was found.
- In the investigation of the distribution of cognitive load information in different frequency bands, it was found that cognitive load information is mainly concentrated in the frequency region (0-1.5) kHz, with the maximum amount of information found in (400-1000) Hz. Furthermore, beyond 1 kHz, the amount of

the information contained in individual subband decreases with respect to frequency.

- Two filterbanks were designed to extract the spectral features specifically for cognitive load classification based on the distribution of cognitive load information. It was shown that the designed filterbanks are more effective than existing filterbanks such as mel, Bark and equivalent rectangular bandwidth.
- In the investigation of the accuracy weighting and signal to noise ratio weighting schemes for a cognitive load classification system based on a likelihood combination multi-band approach, it was found that the accuracy weighting scheme is more effective than the signal to noise ratio and non-weighting schemes.
- The effectiveness of the multi-band approach for classification was investigated. It was found that the multi-band approach produced a higher classification accuracy for the system than the traditional full-band approach.
- Two novel speech enhancement methods were proposed based on two different techniques, namely Kalman filtering and empirical mode decomposition. In addition to this, a separate approach was proposed to improve the effectiveness of an existing speech enhancement method based on the discrete cosine transform (DCT). The proposed improved speech enhancement method based on the DCT was shown to improve the accuracy of the classification system under noisy condtions.

Chapter 2: Automatic cognitive load classification system

This chapter initially presents the basic concepts of cognitive load and the necessity to measure it. This is then followed by an overview of some of the methods currently used to measure cognitive load. A discussion about the impact of cognitive load variation on different aspects of speech is presented. This chapter then overviews the speech features that have been used for CL classification and describes the architecture of the automatic speech-based cognitive load classification system used in this thesis. Several components of the system e.g. feature extraction, normalization and classification are explained. Finally, it concludes with a description of the two cognitive load speech databases used in this thesis.

2.1 Cognitive load

2.1.1 Working memory and its limitation

Working memory is the space in human memory where active cognitive processing occurs [7]. Cognitive processing is defined as the procedures and methods that "control, regulate and actively maintain task-related information" [8]. It is widely known that the capacity of working memory is limited. For instance, early investigations showed that working memory can only hold about seven items of information at a time [9] but recent studies indicate a limit of four items [10]. In addition, information is usually processed at the working memory through organizing, contrasting or comparing, rather than just being held [7]. This further reduces the number of items of information that humans are able to deal with to two or three items. Furthermore, working memory resources are required if there is any interaction between the items held in working memory [11].

2.1.2 Cognitive load theory

Cognitive load refers to the metal demand imposed on a user’s cognitive system, or working memory, while completing a task. Cognitive load theory has been developed by education psychologists in order to design effective instructional strategies which take into account the limitations of human cognitive resources. It is built upon the philosophy of learning and its relationship with the human cognitive system. The basic principles of this theory are based on the assumption that working memory is very limited and a separate long-term memory exists that is virtually unlimited. The learning process involves the construction of schema at the working memory which is then transferred to long-term memory. Schema are hierarchical information networks held in long-term memory that serve as internal, mental representations of the world [12]. If the capacity of working memory is exceeded by the demands of the learning task, learning will be ineffective as the schema cannot be constructed. It is therefore crucial to maintain the level of cognitive load within a suitable range to achieve effective learning and optimum performance.

Learning performance can be degraded by a task with very high or very low levels of cognitive load. Very high levels of cognitive load can degrade performance because the subject does not have sufficient resources to perform the task well. Conversely, very low levels of CL can degrade performance as the subject’s cognitive resources are not engaged in an optimal way [1, 13]. Hence, the effective use of working memory is crucial in achieving optimum learning performance. The aim of cognitive load theory is to provide instructional strategies and learning activities to manage subjects’ cognitive load, such that the use of their working memory resources are optimized [7, 14].

2.1.3 Types of cognitive load

There are three different types of cognitive load: intrinsic, extraneous and germane loads. Intrinsic load refers to the cognitive load created by the structure and complexity of the learning material. The complexity of any given content depends on the level of item or complex interactivity of the material, which is the amount of informational units a learner needs to hold in working memory to comprehend information. This type of load cannot be changed by instruction strategies. The extraneous load is created by the presentation of the task and can be changed by modifying the presentation format. An improved task design can reduce the extra load on working memory. For instance, instructional materials addressing the problem of learning to swim would be more effective, and would produce less extraneous load if they included an appropriate graphic or demo video rather than a text only description. Germane load is caused by the active processing of novel information and schema construction and hence is essential to the learning process.

From a cognitive load perspective, intrinsic, extraneous and germane loads are additive [1]. Therefore it is important to maintain the sum of these loads i.e. the total cognitive load associated with an instructional design, within the limit of working memory for learning to be effective. An illustration of three types of cognitive load on working memory is given in Figure 2.1.

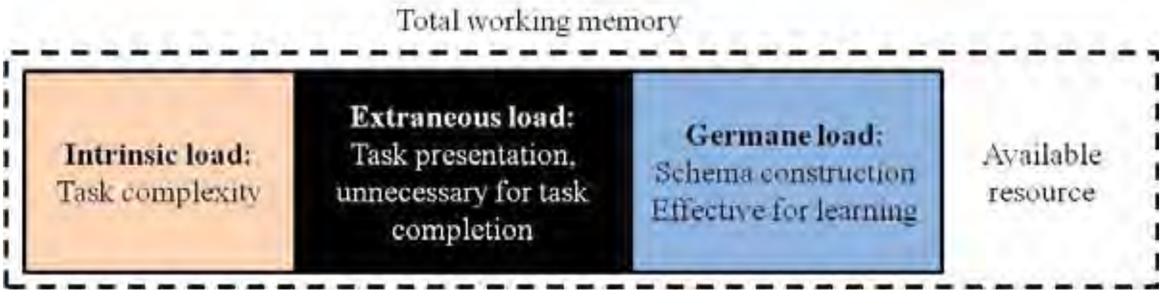

Total working memory
Intrinsic load: Task complexity
Extraneous load: Task presentation, unnecessary for task completion
Germane load: Schema construction Effective for learning
Available resource

Figure 2.1: An illustration of three types of CL on working memory.

Among the three types of cognitive load, it is crucial to ensure that the intrinsic and extraneous loads do not exceed the capacity of working memory. However, the germane load is encouraged. A subject’s learning and understanding of the task will be enhanced by the large available resource of working memory. For tasks with high intrinsic load, it is necessary for a task designer to present the material effectively in order to keep the extraneous load as low as possible to reserve resources for the germane load. However, this may not be very important for a low intrinsic load task as there is plenty of working memory space available for both extraneous and germane loads [15].

Since germane load is necessary for schema construction, which promotes learning and understanding, it can be said that high cognitive load itself does not negatively affect the learning process and task performance. It is high extraneous load that is unnecessary for learning that can degrade the task performance. The objective of cognitive load theory is to design instructional strategies that minimize extraneous CL and promote germane load so that the user’s learning and task performance can be maximized. This can be done by measuring users’ experienced cognitive load and then adapting the user interface and task presentation as per their current cognitive load. This helps to avoid an overload situation where the task demand exceeds the subject’s working memory or an underload situation where the subject is not being involved in the task optimally.

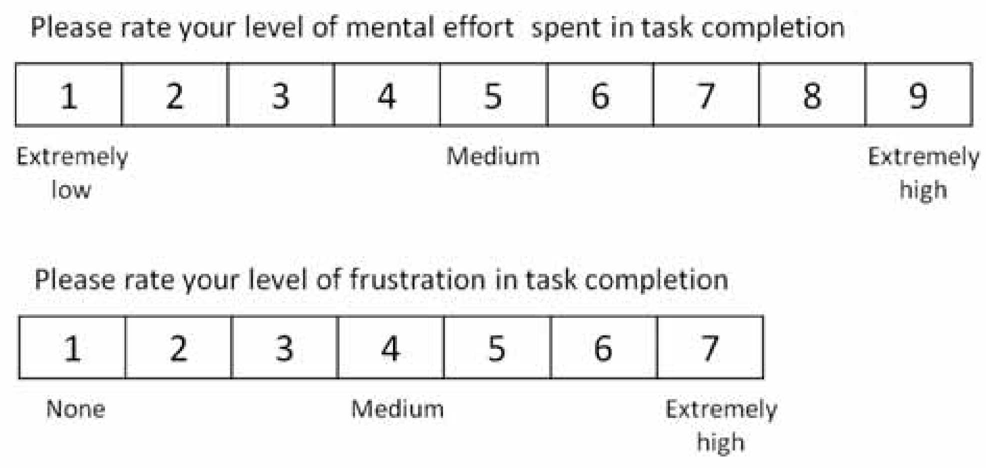
2.2 Overview of cognitive load measurement

Researchers have been attracted to the study of cognitive load measurement for the last couple of decades due to its important role in designing adaptive user interfaces. Numerous methods employing different approaches and measures have been introduced for cognitive load measurement. These methods can be categorized as subjective or self-reporting, physiological, performance-based or behavioral methods [7, 16].

2.2.1 Subjective or self-reporting measures

The subjective or self-reporting measures are estimated by asking users to describe in detail their own perceived load level as induced by the task. They reflect a user's perception of cognitive load by means of introspection. The user is required to perform a self-assessment by answering a set of questions immediately after completing a task.

Subjective measures are based on the assumption that people are able to clarify their cognitive process and report the amount of mental effort expended to perform a task. It has been found that users are able to accurately estimate and report their perceived amount of invested mental effort on a 9-point scale [17-18]. A 7-point rating scale has also been used in other studies [19-20]. However, both empirical and theoretical studies have found that the type of scale used in subjective rating makes no difference [21-22]. Examples of rating scales used in subjective measures estimation are given in Figure 2.2.



Please rate your level of mental effort spent in task completion

1	2	3	4	5	6	7	8	9
Extremely low				Medium				Extremely high

Please rate your level of frustration in task completion

1	2	3	4	5	6	7
None			Medium			Extremely high

Figure 2.2: Examples of 9-point and 7-point self-report rating scales.

In terms of the number of aspects of mental burden that users are required to estimate and report, subjective rating scales can be categorized as one of two types, either unidimensional or multi-dimensional. For a unidimensional scale, users are required to consider only one measure relating to the mental effort spent to perform the task. This type of scale is simple and straight-forward for a user to complete. Rating Scale Mental Effort [23], the Activation scale [24], and the Overall Workload Scale [1] are typical examples of unidimensional scales. Unlike a unidimensional scale, multi-dimensional scales contain more than one measure that users are required to estimate relating to different aspects of mental burden. One of the most popular multidimensional scales used for measuring mental load is the NASA Task Load Index (NASA-TLX) [25]. This scale contains six eleven-point subscales, indicating different aspects of task workload, namely mental demand, physical demand, temporal demand, performance, effort and frustration. The advantage of multi-dimensional scales is that they take into account more specific causes for the load and thus can be more accurate for the purpose of cognitive load estimation. However, the disadvantage is that they rely on the ability of the user to accurately estimate the contribution of different ratings for the cognitive load assuming that users are able to clarify the source of their cognitive load.

Although the use of subjective measures is relatively easy and cost-effective to estimate cognitive load, it is not suitable for implementation in real-world applications since it is highly intrusive and requires time and effort to complete. Furthermore, subjective measures suffer from lack of sensitivity as they only provide a single result for the entire task at completion. However, the effort spent on the task may have changed throughout and hence the cognitive load level of user may have varied during the task. Subjective measures therefore do not reflect the instantaneous cognitive load level. It is also very hard to compare the subjective measures of different users, as the intervals of the scale are unlikely to be consistent across users.

2.2.2 Performance measures

The performance-based methods are categorized into two techniques, namely primary task measurement and secondary task measurement. The primary task measurement are based on the user’s performance of the task being under taken and can include measures relating to task performance such as completion time, critical errors, false starts and latency to response [26-27]. The secondary task or dual-task measurement is based on the performance of a secondary task that is performed concurrently with the primary task. It has also been utilized in research to measure users’ cognitive load [7, 28-30]. Research has found that secondary task measurement is effective for cognitive load measurement as it can indirectly measure the amount of working memory resources being used by the primary task [31]. Performing two tasks at the same time is much more difficult than performing either of these tasks alone. When a user is doing a dual-task, the primary task has first priority when working memory resource is assigned. Therefore, the secondary task performance can be used as a measure of remaining resource not being used by the primary task [31]. The secondary task usually involves a simple activity such as detecting a visual or auditory signal.

Although performance measures are essentially related to the complexity of the task and can be very sensitive to the increase of cognitive load, the use of them as an index of cognitive load has a number of disadvantages particularly because they can be unreliable indicators of load level. It was found in [32] that although two individuals achieve the same level of performance, one expends twice as much as cognitive resources as the other. Furthermore, it is difficult to employ these measures in real-world applications where users' cognitive load level is required to be measured in real-time. It is because performance-based measures are based on features such as completion time and accuracy, which can only be determined after the task has been completed [26].

2.2.3 Physiological measures

The methods used to measure cognitive load levels using physiological measures are based on the assumption that the fluctuation of human cognitive load level is reflected in physiological measures. Numerous measures have been investigated such as heart activity, brain activity e.g. task-evoked brain potentials, eye activity e.g. pupillary response [1], galvanic skin response [33]. The heart-rate measure investigated in [34] was found to be intrusive, invalid, and insensitive to subtle fluctuation in cognitive load. Pupillary response was found to be highly sensitive to fluctuating levels of cognitive load [1]. The effect of cognitive load variation to the pupillary response was investigated in [35] for a group of both young and old participants. It was found that the mean pupil dilation is useful for cognitive load measurement, especially for young participants. In [33], the mean galvanic skin response was found to increase with cognitive load.

Cognitive load measurement using physiological measures has several advantages compared to subjective and performance measures. For example, this method can estimate user cognitive load level automatically due to the subliminal nature of the physiological data being produced. This is unlike the subjective and performance methods which require the involvement of the users. Furthermore, the continuity of physiological data collected from the human body allows a detailed analysis of the fluctuation of cognitive load level while the task is being undertaken. This method can therefore measure cognitive load levels in real-time and is more advanced than the subjective and performance methods.

However physiological methods also have a number of limitations. The main limitation is the intrusiveness caused by the physiological data collection process which requires the attachment of probes, electrodes and monitoring equipment to the user's body. This can interfere with the user's ability to perform the task naturally. Furthermore, similar to other signal processing applications, the physiological data is contaminated by background noise which can reduce the accuracy of the measured cognitive load level and thus it is difficult for this method to be employed in real-life situations.

2.2.4 Behavioral measures

Cognitive load measurement methods using behavioral measures are based on the assumption that users behave and interact differently under different cognitive load levels. Behavioral measures can be used as alternatives to subjective and performance measures and are commonly used in the human computer interaction (HCI) community to assess users' cognitive load for interface evaluation purposes. Various human computer interaction features have been analyzed to clarify the cognitive load state of the user including gaze tracking [36], text input and mouse-click events [37-38] and digital-pen gestures [39].

Unlike the subjective, performance and physiological methods, behavioral methods are objective, non-intrusive and are in real-time as they are based on the data collected from the users while they are performing the task without them knowing that their behavioral data is being recorded. Behavioral methods allow a user to perform the task naturally with minimal interference. These advantages make cognitive load measurement by the behavioral method the most suitable for real-life application systems.

2.3 Cognitive load and speech

The usefulness of speech features for cognitive load measurement has been of interest to many researchers over the last couple of decades [2, 5]. This is because people are required to speak in many real-life tasks such as using telephone and using voice control systems. Speech data can be easily collected in a non-intrusive and inexpensive way. The cognitive load measurement methods based on speech features are non-intrusive, inexpensive and can be performed real-time [5]. As a result, they are more advanced than those based on the subjective, performance and physiological measures.

The impact of cognitive load variation on speech features can be explained by two main reasons. The first is that people tend to communicate in different ways under different levels of CL. For instance, under the high cognitive load caused by the high complexity of a task, they tend to use vocabulary relevant to their feelings such as *hard*, and *difficult* more frequently [40]. Furthermore, they may speak faster because they need to focus on the complex task [41]. For this reason, the variation of the load will affect the linguistic features relating to the content of speech and the dialogue related features of speech (at the word or phrase level). These features are referred to as the high-level features in this thesis. The cognitive load measurement method based on high-level speech features can be categorized as a behavioral method as these features characterize users' behavior. Details of the effect of cognitive load variation on high-level speech features are described in Section 2.3.1. The second reason is based on the assumption that cognitive load is a physiological variable and therefore its variation influences the muscle tension of the vocal cord and the vocal tract of the human speech production system [42]. This in turn affects the prosodic and acoustic features, which are characterized by the vibration rate of the vocal cord and the shape of the vocal tract. These features are referred to as the low-level speech features in this thesis and the cognitive load measurement method based on them can be categorized as a physiological method. Details of the effect of cognitive load variation to low-level speech features are described in Section 2.3.3.

2.3.1 Effect of cognitive load variation on high-level speech features

A number of high-level speech features such as filled pauses, repetitions, silence pause, false starts, disfluencies, response latency and vocabulary categories have been shown to vary according to the fluctuation of cognitive load levels. When the load level increases, it was found that people tend to use words that denote feelings e.g. hard, difficult and heavy more frequently and use prepositions and conjunctions less frequently [40]. Furthermore, the length and frequency of silent pauses are increased [43-44]. This is to be expected because under a difficult task situation and proportionately high cognitive load, people will need more time for problem solving resulting in more silent moments in their speech. Self-correction and false starts, two feature types of disfluencies, have also been identified as indicators of high load [2]. In addition, users tend to engage in self-talk to aid themselves in the problem solving process as the task complexity increases [45]. It was also found in [46-47] that disfluencies and hesitations will occur more frequently in speech under a cognitively demanding task. The correlation between the sentence fragments, consisting of incomplete syntactic structures or ill-formed sentences, and the variation of CL level was investigated in [2]. It was found that under high cognitive load, sentence fragmentation will occur more frequently. Specifically, analysis conducted in their corpus using six types of sentence fragments that are manually detected found that 72% of fragment instances occurred in high cognitive load speech [2].

2.3.2 Human speech production

In order to explain the impact of cognitive load variation on low-level speech features which are characterized by the human speech production system, this subsection briefly describes the human speech production system and the generation of speech.

Speech is a vocalized form of communication for human beings. We use it every day almost unconsciously, without devoting much thought to how it is produced. The human speech production process begins with language processing, where the contents of an utterance are converted into phonetic symbols in the brain’s language center. Following this, three sub-processes take over, including the generation of motor commands for the vocal organs in the brain’s motor center, articulatory movement of the vocal organs based on these motor commands and finally, the emission of air from the lungs. These work together to produce speech [48]. The speech production process is described in Figure 2.3.

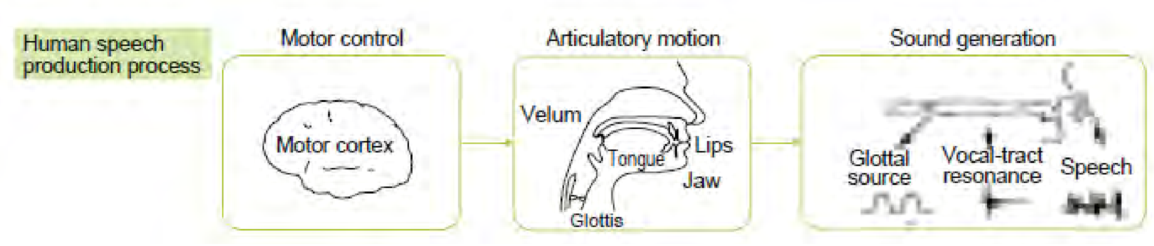


Figure 2.3: Speech production process [48].

Human speech is categorized as voiced (e.g. /aa/) and unvoiced (e.g. /t/). When voiced speech is produced, the airflow from the lungs passes through the opening in the vocal folds, causing them to vibrate. During this vibration, the tension and the elasticity properties of the vocal folds allow them to draw towards each other and separate apart in each vibration cycle. In particular, the air pressure below the folds initially forces the air to flow through the opening of the folds and separates them apart. The velocity of the flow increases the area of constriction and causes a decrease of the air pressure below the folds. This negative pressure will cause the folds to draw towards each other until eventually the opening is closed. The air pressure below the folds then increases to a level sufficient to force the vocal folds to open again. The cycle is repeated until the vocal folds are abducted to produce the phone. The periodic vibration of the vocal folds results in cyclic puffs of air, which is considered to be the sound source. This source is mainly characterized by the fundamental frequency, the rate at which the vocal folds vibrate. When the unvoiced-speech is produced, the vocal folds do not vibrate and the airflow from the lungs passes though a narrow space formed by the tongue inside the mouth. This produces a turbulent flow of air resulting a noise-like sound.

The air stream from the opening of the vocal folds passes though the vocal tract, causing it to resonate. The vocal tract is the combination of all the vocal organs beginning at the opening between the vocal folds and ending at the lips. The resonance characteristics of the vocal tract are determined by the shape of it, which varies when we speak due to movement of the jaws, the tongue and other parts of the mouth. This process enables humans to control the speech sound being produced by changing the position of the vocal organs in their mouth.

2.3.3 Effect of cognitive load variation on low-level speech features

As presented in 2.3.2, the speech production process involves articulator movement. The physiological state that is a response to a perceived high level of task demand i.e. high cognitive load is usually accompanied by specific emotions e.g. fear, anger and anxiety. This causes deviation in the articulator movements which in turn impacts the utterance [49]. Under a high workload task, speaker's respiration rate tends to increase. This increases subglottal pressure during speech, and hence increases the fundamental frequency of voiced speech sections [49]. An increased respiration rate also results in shorter durations of speech between breaths, which affects the articulation rate [42]. In addition, dryness of the mouth in situations of excitement, fear and anger can also affect different aspects of speech production including the muscle activity of the larynx and condition of the vocal cords, which directly affect the volume velocity through the glottis [42]. The effects of heavy task demand on other muscles including those that control the tongue, lips and jaw shaping the resonant cavities of the vocal system also contribute to changes in speech production [42].

Although the impact of load variation on human speech production has not been fully understood, its systematic influence on low-level speech features has been recognized through previous studies. In particular, an increase in load has been associated with an increase in pitch [50-53], reduction in jitter and shimmer [51], increase in the first and fourth formants [53] and decrease in the second formant [54-55]. Other vowel-specific variations in formant frequencies with cognitive load have also been reported [54-55].

Apart from the pitch and formant frequencies, low-level features characterizing the spectral energy distribution have also been found to be indicative of cognitive load. In particular, an increase in cognitive load is reflected by an increase in the spectral energy spread and spectral center of gravity [53], a reduction in the ratio of energy below 500 Hz to energy above it and a decrease in the gradient of energy decay [50]. It has also been suggested that the variability in speech amplitude increases while the speech spectra become flatter under high CL conditions [56].

While both high-level and low-level speech features can potentially be used for cognitive load measurement, their methods of extraction are very different. This in turn affects their ability to develop the cognitive load classification system. Low-level speech features can be extracted automatically and directly from the speech waveform. Therefore, it is possible to develop an automatic cognitive load classification system based on this type of feature. High-level speech features, on the other hand, can only be extracted based on either manual labeling of the speech data or automatic speech recognition. Given that manual labeling is slow and expensive and that automatic speech recognition systems are not yet robust enough for this application, the development of an automatic speech-based cognitive load classification system using high-level features is expected to be difficult. This thesis therefore focuses only on the investigation of the automatic cognitive load classification system based on low-level speech features.

2.4 Automatic speech-based cognitive load classification system

As a pattern recognition system, a speech-based cognitive load classification system consists of a feature extraction module, used to extract relevant features from speech, and a classification module, usually employing a machine learning approach to model and recognize the load specific patterns from these features. In order to improve the robustness of the system, it is necessary to reduce or eliminate variation in patterns due to factors unrelated to cognitive load such as background noise, channel mismatch and speaker variability. The feature extraction module is therefore often combined with other modules such as noise reduction and channel/speaker normalization. The combination of all of these modules is referred to as the front-end. The classification module is referred as the back-end. The general structure of a speech-based cognitive load classification system is shown in Figure 2.4.

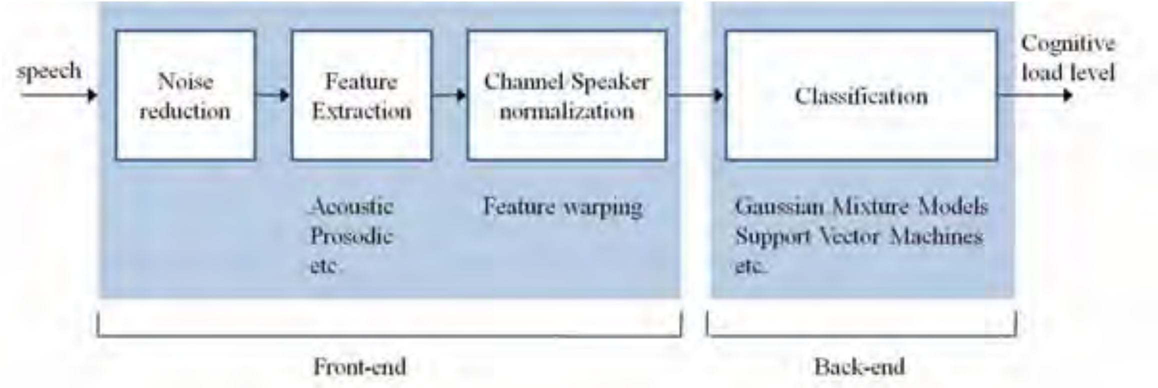


Figure 2.4: The diagram of an automatic speech-based CL classification system.

2.4.1 Front-end

2.4.1.1 Feature extraction

A front-end of an automatic cognitive load classification system is designed to extract speech features. These are typically frame-based and are computed from the voiced frames of speech. The feature vector, obtained by concatenating all the feature elements computed in individual frames of an utterance, is referred to as the static feature.

Concatenation of the static and temporal derivatives

The dynamic features which capture temporal information between frames have previously been found to be very useful for cognitive load classification. The concatenation of the dynamic feature into a static feature significantly improves the performance of the classification system, compared to the one based solely on the static feature [5, 57]. The first order derivatives are referred to as delta feature and can be computed based on regression as follows:

$$\Delta C_i(n) = \frac{\sum_{k=-N}^{N} k C_i(n+k)}{\sum_{k=-N}^{N} k^2} \tag{2.1}$$

where $\Delta C_i(n)$ is the delta feature calculated on the n^{th} frame of the i^{th} dimension of the feature C and N specifies the number of frames across which delta features are calculated. Similarly, second order derivatives (the delta-delta features) can be computed using the same equation on the delta feature vector instead of the original feature vector.

Delta and delta-delta features effectively encode the temporal information, however they are limited in their ability to model higher level temporal aspects of speech since they only model the slope of the feature at the current point in time. For instance, with the standard method of calculation using a value of $N = 2$, the delta feature will be an estimate of the slope at the current time based on the values across 5 frames (50 ms if the duration of each frame is 10 ms). Thus, at best, they are only able to incorporate the temporal aspects of speech within a time window of 50 ms. In order to capture the temporal aspect of speech in a longer time window, we need to increase the value of N. However, this will only produce a longer average of the slope and its finer details will be lost.

The shifted delta technique has been proposed as a better alternative for including the temporal information in the speech signal across a longer time window. It was originally proposed for language identification [58]. Shifted delta feature of a frame is obtained by concatenating a number of delta features computed from following frames.

According to the method described in [59], the computation of the shifted delta feature is specified by four parameters: M, D, P, and K. M specifies the number of basic feature streams to use in the calculation. The shifted delta features are computed separately for each of the M feature streams. P is the number of frames from one delta calculation to the next and K is the total number of delta values concatenated together to form the shifted delta feature. For each of the feature streams, the shifted delta feature vector at time n is given by the concatenation of the $\Delta C_i(n, m)$ for $0 \leq m \leq K$, where

$$\Delta C_i(n,m) = \frac{\sum_{d=-D}^{D} dC_i(n+mP+d)}{\sum_{d=-D}^{D} d^2} \tag{2.2}$$

The shifted delta features for each time instance are calculated across a window of $(K\text{-}1)P+2D+1$ frames. For the shifted delta structure used in this thesis where D, P and K are set to 1, 3 and 7 respectively as in [4-5], the shifted delta feature can incorporate temporal information spanning 21 frames, i.e. 210 ms whilst retaining the fine-grained information within that window. This is because a sampling of all the delta values within that window is used. Thus the shifted delta feature allows the inclusion of a much wider range of temporal information than the standard delta or delta-delta features. A diagram showing the method for producing the shifted delta feature is shown in Figure 2.5.

The shifted delta feature of a multi-stream feature is obtained by concatenating the shifted delta feature computed on individual streams. An example of the combination of a three dimensional feature (C_0-C_2) and its shifted delta feature is provided in Figure 2.6.

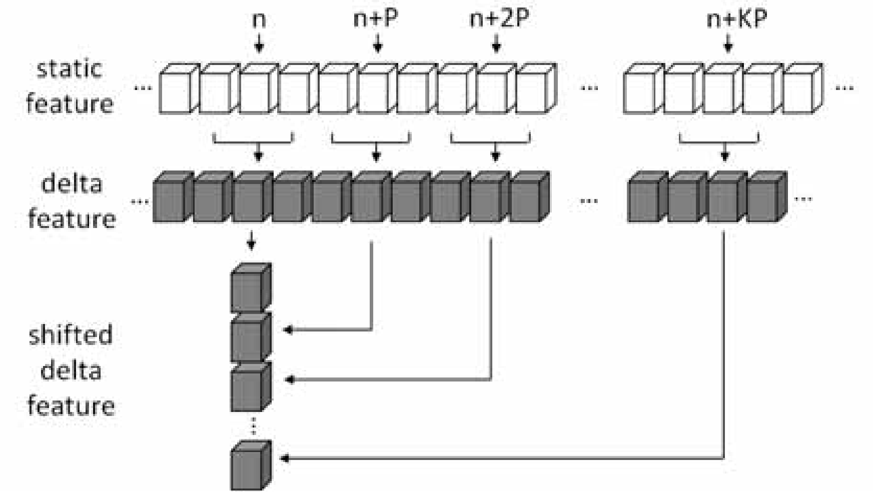
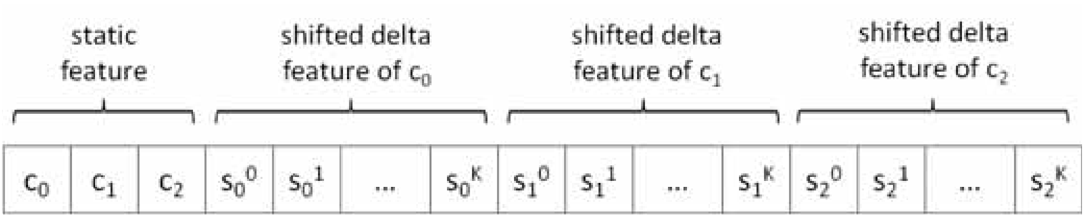


Figure 2.5: Shifted delta feature calculation for a single feature stream at *n*th frame [60].



static feature			shifted delta feature of c_0				shifted delta feature of c_1				shifted delta feature of c_2			
c_0	c_1	c_2	s_0^0	s_0^1	...	s_0^K	s_1^0	s_1^1	...	s_1^K	s_2^0	s_2^1	...	s_2^K

Figure 2.6: Concatenation of the static and shifted delta features.

A number of low-level speech features have been utilized by automatic speech-based cognitive load classification systems to date. In particular, pitch, intensity, and Mel frequency cepstral coefficients (MFCC), have been shown to be effective [5, 53, 57]. In [61], it was shown that the group delay feature, which is based on phase spectrum, can be used to provide additional cognitive load information to the MFCC-based system and improve its performance. In [4], it was indicated that the features based on the voice source are useful for cognitive load classification. The usefulness of formant frequencies was also found in [53-54, 62]. The non-linear Teager energy operator was found to be effective for classifying cognitive load in [63]. Other features including perceptual linear prediction coefficients, spectral center of gravity, spectral energy spread and vowel durations were also found to be useful in cognitive load classification systems [53].

2.4.1.2 Feature warping

In a classification system, the features extracted from speech can be affected by a number of factors such as the short-term channel distortion and speaker variability. A feature normalization technique called feature warping can be used to reduce the effects of these factors and improve the robustness of the system. This technique maps the distribution of a feature stream in a specific time interval to a standardized distribution. In practice, the mapped value of the current feature value is calculated over a sliding window as in [64]

$$m = ipdf\left(\frac{N + 1/2 - R}{N}\right) \tag{2.3}$$

where *m* is the mapped value, *ipdf* is the inverse cumulative distribution function for the normal distribution, *N* is the window length, *R* is the ranking of the value in the descending order of the original feature vector within the sliding window. Figure 2.7 shows an example of the distributions of a feature vector before and after warping.

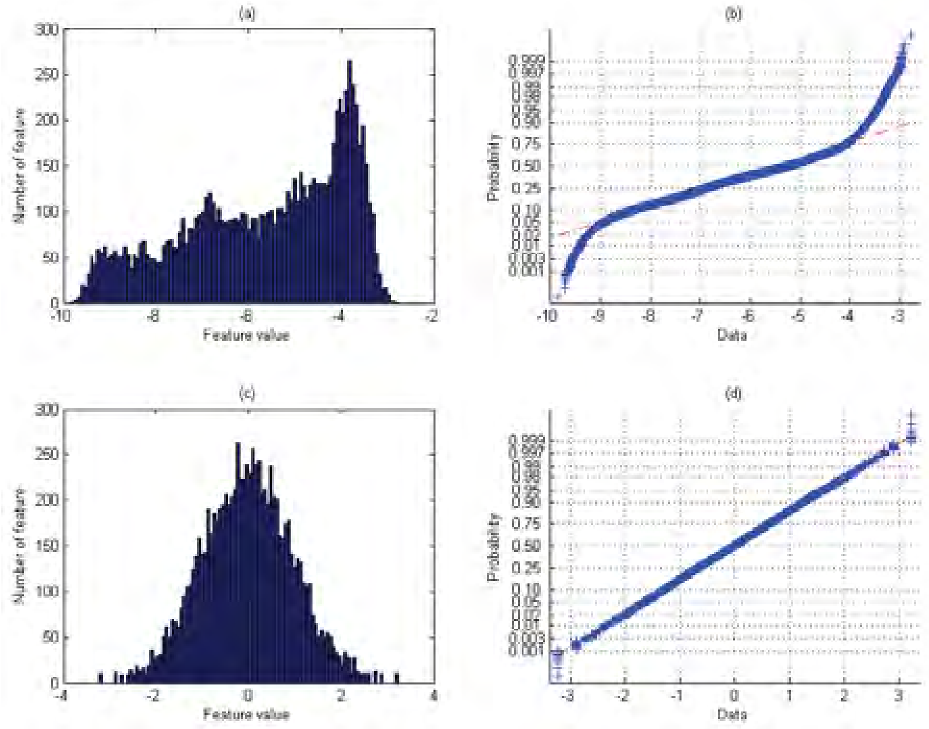


Fig. 2.7: The distribution of a speech feature before warping (a) & (b) and after warping (c) & (d).

2.4.2 Back-end

Given that the cognitive load (CL) specific patterns are contained in the acoustic features extracted by the front-end, the goal of the back-end is to initially model the cognitive load from these patterns and then perform pattern matching to determine the CL level. The Gaussian mixture models (GMMs) and the support vector machines (SVMs) are the two classification methods that have been used in automatic CL classification. Comparing these two methods, GMMs are generative classifiers where the model representing each class is trained individually on a training data set of that class. Generative classifiers do not consider training data from other classes when training the model of one class, thus making the training process of GMMs simple and fast. SVMs, on the other hand, are discriminative classifiers. Training their models takes into account the training data of all classes simultaneously, which makes the training process very complex [65]. Furthermore, SVMs were shown to be less effective than GMMs for CL classification [66]. Hence, Gaussian mixture models were used for all the experiments reported in this thesis.

2.4.2.1 Gaussian mixture model

The Gaussian mixture model (GMM) is a generative classifier used to model the underlying probability density function of speech feature. This model has been widely used as the classifier in many existing classification systems. The basic idea of a GMM is to model the distribution of a feature in the feature space with a number of Gaussian distributions. For instance, the distribution of a single-dimensional feature vector with probability distribution as shown in Figure 2.8a can be described as the sum of three Gaussian distributions with different weights, means, and variances as shown in Fig. 2.8b.

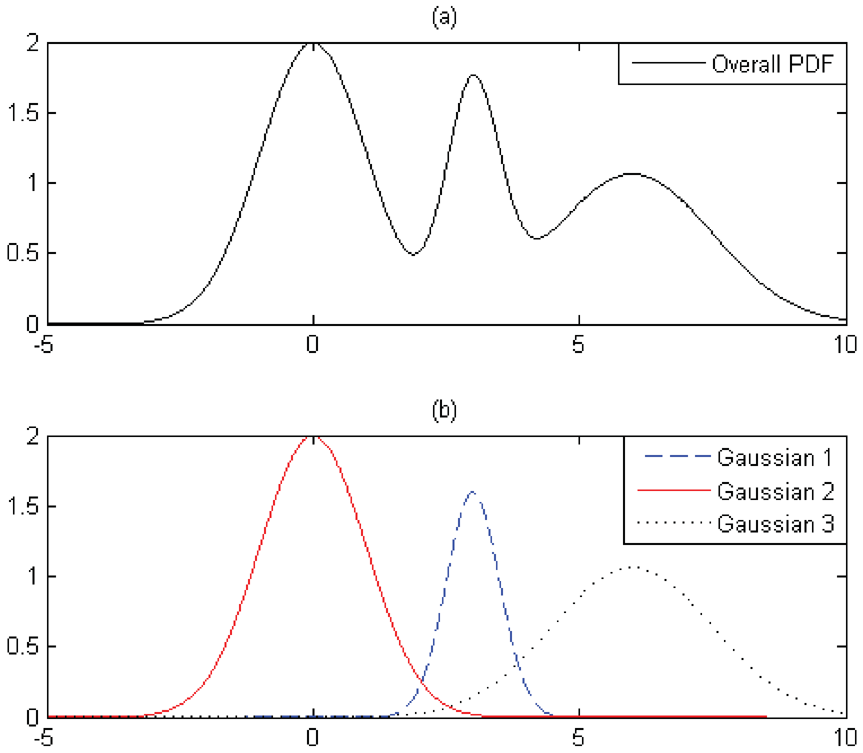


Figure 2.8: (a) Probability distribution of a single-dimensional feature, (b) Three Gaussian components of the distribution shown in (a).

To model more complicated distributions, more Gaussian components are needed. In the case that the feature vector is multi dimensional, the Gaussian mixture model is a mixture of several multivariate unimodal Gaussian densities, expressed as

$$p(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{x}) \tag{2.4}$$

with

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_{\mathrm{i}}|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'(\boldsymbol{\Sigma}_{\mathrm{i}})^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \tag{2.5}$$

where $\boldsymbol{x}$ is the feature vector, ω_i is the weight of the i^{th} mixture satisfying $\sum_{i=1}^{M} \omega_i = 1$, M is the number of mixtures, D is the feature dimension, $\boldsymbol{\Sigma}_{\mathrm{i}}$ and $\boldsymbol{\mu}_i$ are the covariance matrix and mean vector of the i^{th} mixture and $\boldsymbol{\lambda_i} = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{\mathrm{i}}\}$, where $1 \le i \le M$ is the set of GMM parameters. These parameters are estimated based on maximizing the likelihood of the training observation from the estimated model, using the expectation-maximization (EM) algorithm [67].

In cognitive load classification, a universal background model (UBM), which is another GMM, is trained from a large quantity of background speech. The GMMs representing each CL level are adapted from the UBM, using the speech features of the corresponding cognitive load level and based on maximum a posterior (MAP) adaptation [68]. For the features $\boldsymbol{x_t}$ at time $t \in \{1,2,\dots,T\}$ of a particular cognitive load level, the MAP adaptation from the universal background model is as follows [68].

Initially the probabilistic alignment of the feature $\boldsymbol{x_t}$ into the UBM mixture components is computed as

$$\Pr(i, \boldsymbol{x_t}) = \frac{\omega_i p_i(\boldsymbol{x_t})}{\sum_{j=1}^{M} \omega_j p_j(\boldsymbol{x_t})} \tag{2.6}$$

The sufficient statistics for the weight, mean, and variance parameters are then computed as below:

$$n_i = \sum_{t=1}^{T} \Pr(i, \boldsymbol{x_t}) \tag{2.7}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i, \boldsymbol{x_t}) \boldsymbol{x_t} \tag{2.8}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i, \boldsymbol{x_t}) \boldsymbol{x_t}^{\mathbf{2}} \tag{2.9}$$

These sufficient statistics are used to adapt the UBM parameters to obtain the GMM for the corresponding CL level. In order to control the balance between the estimated sufficient statistics and initial UBM parameters, an adaptation coefficient α is used. In addition, a scale factor γ is introduced to satisfy the constraint, $\sum_{i=1}^{M} \widehat{\omega}_i = 1$ on the estimated weights. The adapted weight mean and covariance for the i^{th} mixture are

$$\widehat{\omega}_i = (\alpha_i n_i/T + (1-\alpha_i)\omega_i)\gamma \tag{2.10}$$

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \alpha_i E_i(\boldsymbol{x}) + (1-\alpha_i)\boldsymbol{\mu}_i \tag{2.11}$$

$$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{i}}{}^2 = \alpha_i E_i(\boldsymbol{x}^2) + (1-\alpha_i)(\boldsymbol{\mu}_i + \boldsymbol{\sigma}_i^2) - \widehat{\boldsymbol{\mu}}_i{}^2 \tag{2.12}$$

A fixed relevance factor r is used to calculate the adaptation coefficient, α, as below:

$$\alpha_i = \frac{n_i}{n_i + r} \tag{2.13}$$

This relevance factor is data dependent and has a typical value of 16 [68]. As UBM models the basic distribution of speech feature, the use of this model as the initial distribution for GMMs representing CL levels can improve the precision of the GMMs when training data is limited.

In the testing phase, the log likelihood $\Lambda(X)$ of an utterance $X = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}\}$ is calculated as

$$\Lambda(X) = \log p(X|\boldsymbol{\lambda_l}) - \log p(X|\boldsymbol{\lambda_{UBM}}) \tag{2.14}$$

using the model parameters of the hypothesized cognitive load level $\boldsymbol{\lambda_l}$ and UBM $\boldsymbol{\lambda_{UBM}}$. The cognitive load level that has the highest score is chosen as the load level of the test segment.

A block diagram of an UBM-GMM based CL classification is shown in Figure 2.9.

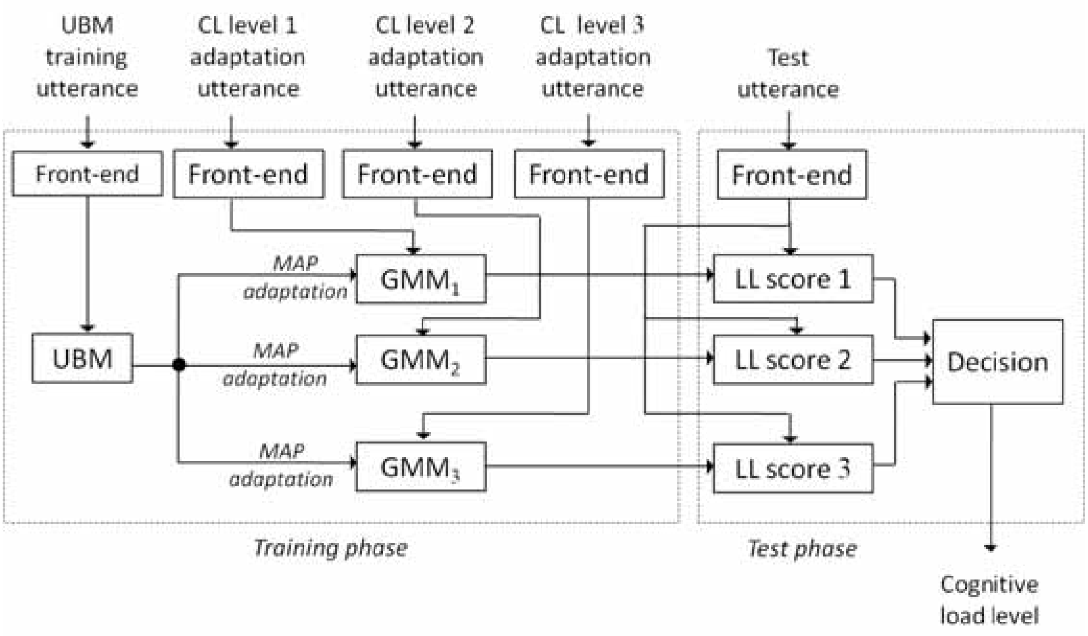


Figure 2.9: Block diagram of an UBM-GMM based CL classification system *(LL: log likelihood).*

2.4.2.2 Fusion method

Cognitive load specific information can be contained in different speech features. An approach using multiple features, where the feature vector is obtained by concatenating several features, can combine the cognitive load information from the various speech features and improve the performance of the classification system. The disadvantage of this approach is that it increases the dimension of the feature vector which consequently increases the amount of data required to estimate the CL model parameters sufficiently. Moreover, a high dimensional feature vector will also lead to an increase in the computational complexity of the system, which in turn increases the processing time of the system. The fusion technique is an alternative approach to combine cognitive load specific information from multiple speech features. In this technique, the likelihood score of the fusion system is obtained by combining the likelihood scores of the cognitive load classification systems based on individual features, as illustrated in Figure 2.10. In this thesis, this is done by using a linear score weighting technique commonly used in speaker recognition and language identification systems [69-70]. Based on this technique, the log likelihood score of the fusion system is obtained as

$$LL = \sum_{i=1}^{N} w_i LL_i \tag{2.15}$$

where LL is the log likelihood score of the fusion system; LL_i is the log likelihood score generated by the i^{th} system based on the i^{th} feature vector; w_i is the positive weighting coefficient that satisfies the constraint $\sum_{i=1}^{N} w_i = 1$. The weighting coefficients are empirically chosen to optimize the performance of the classification system.

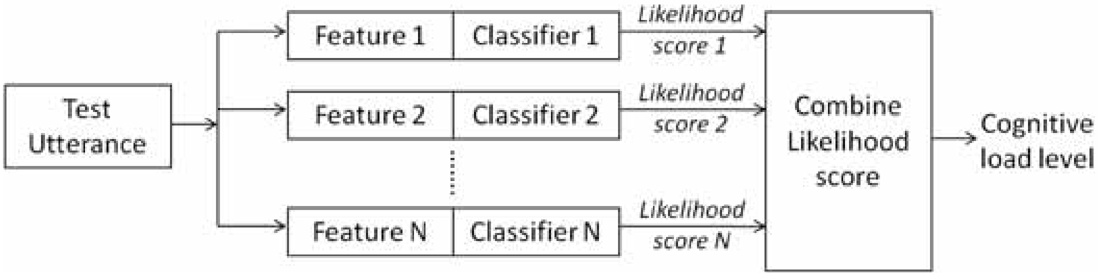


Figure 2.10: Overview of a CL classification system based on fusion technique.

2.4.3 Existing CL classification systems

Possibly the first automatic speech-based cognitive load classification sysem was developed in 2008 by Yin et al. This system used a Gaussian mixture model (GMM)-

based classifier with front-end features consisting of mel frequency cepstral coefficients (MFCC) and prosodic features. When evaluated on a corpus consisting of three cognitive load levels, its classification accuracy was 71.1%.

Since then, numerious features have been used to developed the CL classifciation systems based on a GMM classifier. Table 2.1 summarizes the various front-end features and the corresponding classification accuracy obtained from the experiments performed on the database consisting of three CL levels.

Table 2.1: Summary of various front-end features proposed for cognitive load classification

Author(s)	Front-end features	Classification accuracy
Yin et al. [5]	MFCC, prosodic features	71.1%
Yap et al. [61]	MFCC, prosodic, phase-based features	85.3%
Yap et al. [54]	Formant frequencies	67.9%
Yap et al. [4]	MFCC, prosodic, glottal features	84.4%

It can be seen from Table 2.1 that although many classification systems have been proposed, their performances are not very high. Furthermore, these performances were obtained from the experiements performed in clean conditions where background noise did not exist. In practical scenarios, the cognitive load levels need to be estimated in environments such as in car, at the airport, or in call-center where the speech is corrupted by the background noise. As such, the performance of the above-mentioned systems can be degraded dramatically. Therefore, it is necessary to develop methods to improve the performance of the existing classification systems and to make them more robust to noise.

2.5 Cognitive load speech corpora

The usefulness of a speech feature for cognitive load classification is evaluated based on the accuracy of the classification system using that feature, when tested on a cognitive load speech database. It is therefore necessary to have a standard cognitive load corpus. However, as the study of speech-based cognitive load classification is still in early stages, such a standard corpus is not available. Most of the current research on the cognitive load classification has therefore been conducted on speech corpora collected by researchers themselves [2, 5, 71-73]. Intuitively, the reliability of the results of the final study will be affected by a number of aspects of the cognitive load corpora as described below:

- The number of load levels: human cognitive load is a continuous variable reflecting the amount of mental load imposed on the subject's cognitive system. However, the problem of classifying load levels can only be conducted through a discrete scale containing a limited number of levels obtained by quantizing the continuous scale. A large number of levels would better to describe a human's cognitive load state. This is also desirable for practical cognitive load based systems, as the fine scale would allow the system to adapt better to the cognitive capacity of user and therefore produce higher productivity. However, given that the research of cognitive load classification based on speech features is in early stages, most of the studies in this field are performed on the CL corpora that contain two levels [2, 53, 72-73] or three levels of cognitive load [4, 74].
- Whether the corpus was collected in laboratory or in a real-world environment: results obtained from the use of a real-world corpus would better reveal the actual potential of the system in a real-life application. This is due to the fact that a corpus collected in this fashion is affected by all the factors that can degrade the performance of the system e.g. noise and channel mismatch, which exist in real-life. However, in order to use such corpora, manual segmentation of the speech into different segments corresponding to different cognitive load levels would be required, which can be very time consuming. On the other hand, for a corpus collected in a laboratory different utterances corresponding to different CL levels are recorded separately, hence the manual segmentation process is not required. Furthermore, the factors unrelated to cognitive load i.e. noise and channel mismatch will be low. Results obtained from the use of this type of corpora better reveal the actual ability of speech features in classifying the load level. Due to these advantages, most of the studies on speech based cognitive load classification to date have been carried out on laboratory corpora [2, 4, 53, 72-74].

The cognitive load speech corpora used in all the experiments reported in this thesis are the Stroop test and the Reading Comprehension, which were collected in a laboratory environment [5, 57]. These two corpora have several common features. They contain speech corresponding to three levels of load from fifteen native English speakers, eight of whom are female and seven male. The speech in both corpora is sampled at 16 kHz. The three levels of cognitive load, namely low, medium and high, in these two corpora were induced by asking the speakers to perform three different tasks with the corresponding difficulty level. Details of the collection of these two corpora are described in the next sections.

2.5.1 Collection of the Stroop test database

This database was collected via tasks that were designed based on the 'Stroop test' developed by John Ridly Stroop [75]. In this test, there are two types of tasks namely the reading color name (RCN) task in which participants are asked to read out the words ignoring the font color; and the naming colored word (NCW) task in which the font color of words has to be read out. An example of these two tasks is shown in Figure 2.11. Between these two tasks, the naming colored word is expected to be more difficult as participants need to put in more effort to override the meaning of the text in order to read out the actual font color. This was supported by a study, where a significant delay of task completion was noticed in naming colored word tests compared to reading color name [75]. Speech from NCW is therefore assumed to correspond to a higher cognitive load level than that from RCN.

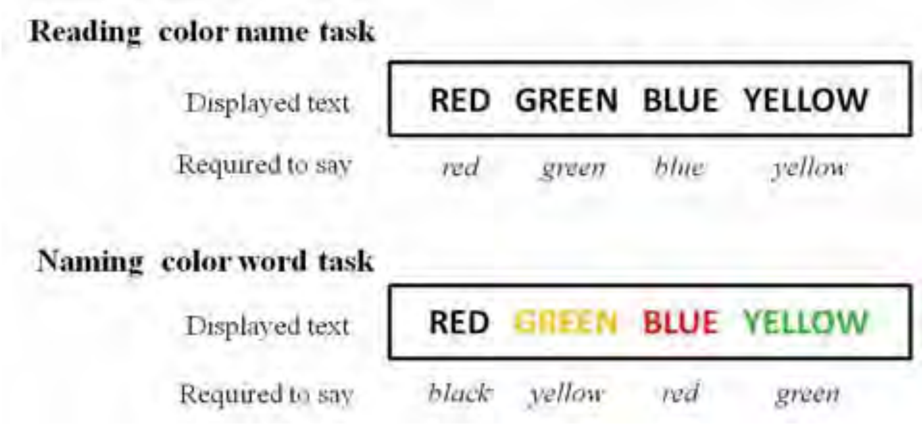


Figure 2.11: An example of two tasks of the Sroop test

Given the nature of the test, the following procedure was proposed to collect three levels of cognitive load speech in the Stroop test corpus [5]:

- Low level of cognitive load speech: participants are required to perform two reading color name tasks:
    - Test 1: all words are written in black.
    - Test 2: all words are written in congruent color i.e. font color is the same as the text meaning).

- Medium level of cognitive load speech: participants are required to perform two naming colored word tasks:
    - Test 1: words are written in either congruent or incongruent color i.e. the font color is different to the text meaning.
    - Test 2: words are written in incongruent color.
- High level of cognitive load speech: participants are required to perform the two naming colored word tasks under a time constraint:
    - Test 1: words are written in incongruent color, appearing only one at a time.
    - Test 2: words are written in incongruent color, appearing consecutively while previous words stay on display.

The speech in this corpus contains four utterances per CL level per subject. The approximate length of each utterance is 15 seconds. In addition to the above-mentioned speech, another task was recorded where the same participants read a story as neutral reference data. The story reading speech consists of approximately 90 seconds of speech per speaker.

2.5.2 Collection of the Reading and Comprehension database

In this database, speech corresponding to each cognitive load (CL) level was recorded by asking the subjects to read out a story of a corresponding level of difficulty and then answer three open ended questions related to the story's content. The difficulty level of the stories is measured using the Lexile scale [76] – a semantic difficulty and syntactic complexity measure scale ranging from 200 to 1700 Lexiles (L), corresponding to the expected reading levels of students from first grade to graduate level. The Lexile ratings of the stories used were 925 L, 1200 L, and 1350 L respectively. The stories contain general knowledge about weather phenomena, household appliances and the functions of the human body to avoid expertise being a factor in the results. The open ended questions are:

- Give a short summary of the story in at least five whole sentences
- What was the most interesting point in this story?
- Describe at least two other points highlighted in this story.

The speech in this corpus contains four utterances corresponding to each cognitive load level for each subject, one from reading the story and three from answering the questions related to that story. The approximate lengths of the utterances that correspond to reading the story for the low, medium, and high cognitive load level are 90 seconds, 140 seconds, and 230 seconds respectively. The approximate length of each answer to the three questions for all three levels of CL is 30 seconds.

Between these two corpora, the Stroop test corpus contains a limited number of color words such as ‘red’, ‘blue’ and ‘green’. This corpus is akin to an isolated speech corpus as the subjects read the words one by one slowly. In addition, there is a speech rate artifact caused by the time constraint for the high CL speech. On the contrary, the Reading and Comprehension corpus contains a significantly larger vocabulary due to the varied content of the long stories, and the open ended nature of the questions. The story reading section of this corpus contains continuous speech and the question answering section consists of spontaneous speech. The high level of phonetic variability in the Reading and Comprehension corpus results in greater variability in speech features for each load level in this corpus compared to the Stroop test corpus. Consequently, classification of the CL levels based on speech from the Reading and Comprehension corpus is more challenging than classification based on speech from the Stroop test corpus.

2.6 Summary

This chapter presented the concept of human working memory, cognitive load and the background of cognitive load theory. Cognitive load theory is based on the assumption that human working memory is limited and its objective is to design the instructional strategies that can exploit working memory resources in an optimal way, in order to achieve the highest task performance. The chapter then described the benefit of cognitive load measurement and reviewed the various techniques that have been used for this purpose. The effect of cognitive load variation to speech features was also presented in this chapter. The human speech production mechanism was then briefly described in order to provide a better understanding of how speech features vary with cognitive load. Section 2.4 then reviewed the speech features that have been used to classify cognitive load levels and described the structure of an automatic CL classification system. The system components of feature extraction, feature normalization and the Gaussian mixture model classifier were described. Finally, it provided details about the Stroop test and the Reading and Comprehension corpora, which are the two cognitive load corpora used in all the experiments for this thesis.

Chapter 3: Investigation of the effectiveness of speech features for cognitive load classification

The human speech production mechanism can be described as a two-stage process, first a sound source generation stage and second a spectral shaping stage. The source-filter model describes the spectral shaping stage as a filtering process where the filter is a model of the vocal tract. This model assumes that the sound source generation and the filtering process are independent of each other [77]. The simplicity of this model is one of its main advantages and it has been used in many different areas of speech processing such as emotion recognition, speaker verification and speech synthesis [78-80]. Based on this model, a speech feature can be categorized as either a source-based feature, a filter-based feature, or a combined feature (describing both source and filter). The impact of cognitive load variation on the human speech production system can occur at the source, the filter or both. While it is very difficult to quantify exactly how much each of these components contributes to the characterization of cognitive load, the effectiveness of source-based or filter-based features in a cognitive load classification system is expected to give an estimation of the relative contribution of these two components in characterizing the load. Finding this relative contribution will help feature selection and the design of a front-end for cognitive load classification.

Given that different features describe different properties of speech, it is possible that the cognitive load information contained in them complement each other. Incorporating the information from several individual features can increase the overall cognitive load information extracted compared to that contained in any one feature. This in turn can improve the performance of the classification system. As such, it is necessary to investigate the effectiveness of different speech features for cognitive load classification.

One of the main aims of this chapter is to analyze the relative contribution of source-based and filter-based features, so this chapter will briefly present the source-filter model of the human speech production system. The details of the human listening test for cognitive load classification are presented to assist in the understanding of the type of speech cues humans used to classify cognitive load. Section 3.3 describes the baseline classification system used to perform the classification experiments in this thesis. The investigation of the effectiveness of various speech features, categorized as either source-based, filter-based or combined, for cognitive load classification is described in Section 3.4. The combined speech features that are presented in Section 3.4 have been utilized in previous studies of cognitive load classification. A novel use of spectral centroid features for cognitive load classification is proposed in Section 3.5. Finally, the comparison and discussion of the effectiveness of different speech features is provided in Section 3.6.

3.1 Source-filter model of human speech production system

The source-filter model presented here models the speech production system, over a short period of time, as a linear time invariant system excited by a sound source.

3.1.1 The source component

The sound sources that produce voiced and unvoiced speech are very different. The source of voiced speech is the pulse train of airflow from lungs created by the periodic vibration of the vocal folds. The spectrum of the glottal source consists of a number of frequency components corresponding to the harmonics of the fundamental frequency of the vibration of the vocal folds. This fundamental frequency will be referred to as F_0 for the rest of this thesis. As a result, increasing the frequency of vibration of the vocal folds will make the spacing between the harmonics in the glottal source spectrum larger while the overall shape of the spectral envelope remains unchanged. An example waveform and spectrum of a glottal source signal is shown in Figure 3.1.

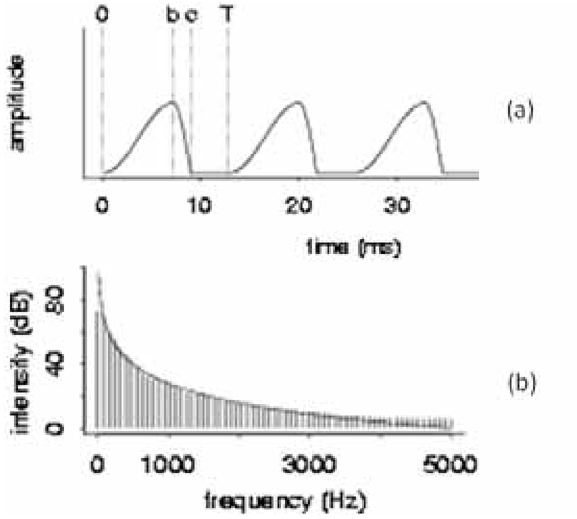


Figure 3.1: (a) Glottal source waveform and (b) the corresponding spectrum [81].

The source of unvoiced speech is turbulent airflow caused by a constriction of the vocal tract at some point. Unlike the periodic excitation of the voiced speech, turbulent airflow contains no dominant periodic component and has a relatively flat spectrum.

3.1.2 The filter component

The vocal tract, which is the region in the speech production system bounded by the glottis and the lips, is modeled as a non-uniform tube whose shape is a function of time. The cross-sectional area of the vocal tract at a particular time instant varies along the vocal tract. This area and hence the shape of the vocal tract is determined by the position of the tongue, lips, jaws and other vocal organs. The shape of the vocal tract changes continuously during speech production, following the movement of these vocal organs to produce the desired speech.

The source waveform travelling through the vocal tract is shaped by the tract's resonance characteristic, determined by its shape which varies with time, to produce speech. However for short durations of typical length 10-20 ms, the shape of the vocal tract and hence its resonance characteristic can be considered as fixed. As a result, speech can be considered to be stationary in the short segments corresponding to these durations. These segments will be referred to as frames of speech in this thesis. The entire speech signal is usually referred to as quasi-stationary as this stationary property only exists in short durations.

Given that the shape of the vocal tract can be assumed to be stable for short durations, the vocal tract can be modeled as a linear time invariant filter in these time periods. The peaks in the magnitude response of this filter reflect the resonant frequencies of the vocal tract and are referred to as the formant frequencies. These will be shown in Figure 3.4 which can be found in Section 3.1.3.

At the mouth/lip opening, the sound waveform ceases to be constrained in its propagation. The effect of this change, referred to as lip radiation, can be approximated as a fixed filter with a 6 dB/octave rising magnifying spectrum. The effect of lip radiation is not accounted for the vocal tract model but is considered as a high-pass filter cascaded to the vocal tract filter as shown in Figure 3.2.

3.1.3 Combining the source and the filter components

According to the source-filter model, speech is produced as the excitation signal *e(n)* being filtered by the vocal tract filter $V(z)$ and the lip radiation filter $R(z)$ as shown in Figure 3.2. The spectrum of speech $S(z)$ can be expressed in the complex frequency domain as:

$$S(z) = E(z)V(z)R(z) \tag{3.1}$$

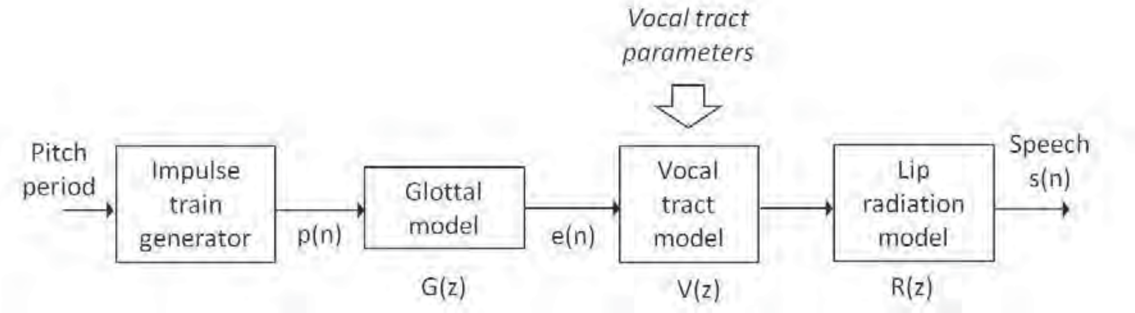


Figure 3.2: The source-filter model for voiced speech production.

As mentioned previously, the effect of the lip radiation is approximated as a 6 dB/octave rise, which can be modeled as a first order high pass filter:

$$\begin{aligned} R(z) &= 1 - \alpha z^{-1} \\ &\approx 1 - z^{-1}, (\alpha \to 1) \end{aligned} \tag{3.2}$$

The excitation signal *e(n)* can either be a periodic glottal pulse sequence for generating voiced speech or random noise for generating unvoiced speech. In the case of voiced speech, the glottal excitation can be considered as the result of the convolution of a train of impulses separated by the pitch period $T_0 = 1/F_0$, where F_0 is the fundamental frequency, and a single glottal waveform. This is another filtering operation where the impulse response of the filter is the single glottal waveform, as illustrated in Figure 3.3.

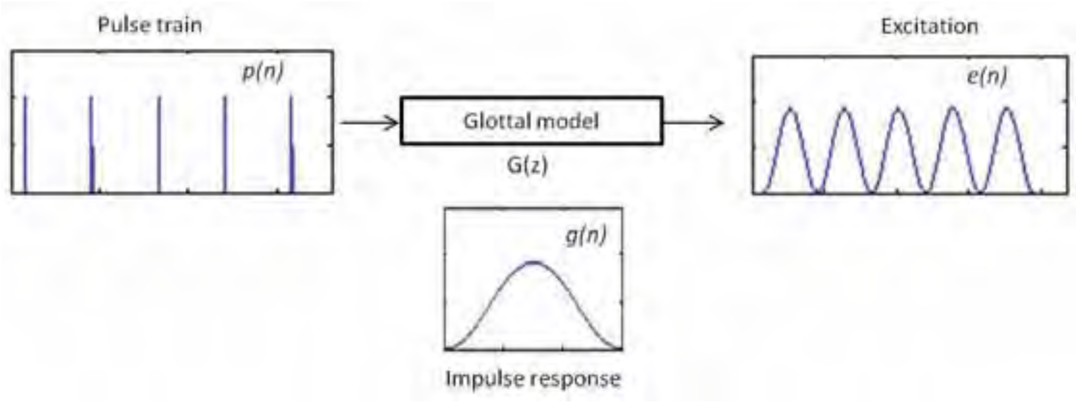


Figure 3.3: Glottal filter model.

From equation (3.1), the spectrum of voiced speech can be expressed as

$$S(z) = P(z)G(z)V(z)R(z) \tag{3.3}$$

where $P(z)$ is the impulse train and $G(z)$ is the response of the glottal transfer function. The vocal tract of voiced speech can be adequately modeled as an all-pole system, expressed as:

$$V(z) = \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-i}} \tag{3.4}$$

where a_i are parameters of the system that can be estimated using linear prediction and p is the order of the system. The system order is usually selected based on the rule of thumb of

$$p = 2 + round(F_s/1000)$$

where F_s is the sampling rate [82].

Estimating glottal waveform (sound source) and parameter of vocal tract filter

The excitation signal or glottal waveform of voiced speech *E(z)* can be obtained from equation (3.1) as

$$E(z) = \frac{S(z)}{V(z)R(z)} \tag{3.5}$$

where the effects of the vocal tract filter and lip radiation are removed from the speech signal. In this thesis, the Iterative Adaptive Inverse Filtering algorithm [83] was used to estimate the glottal waveform and the parameters of the linear prediction model of the vocal tract filter from the speech signal. This method estimates the effect of the voice source on the speech signal as a first order all-pole model. Then by estimating and eliminating the contribution of the voice source, an all-pole model of order $p = 18$ for the vocal tract is computed by linear prediction modeling. The glottal waveform is then obtained by filtering the original speech signal using the inverse of the vocal tract filter and canceling the lip radiation effect by integration.

In order to obtain better estimates of the glottal waveform and the linear prediction model of the vocal tract filter, this computation is executed in a repetitive manner consisting of two phases. In the second phase the glottal source is estimated as a second order all-pole model. The detailed description of the Iterative Adaptive Inverse Filtering algorithm can be found in [83]. Figure 3.4 shows the magnitude spectrum of the phoneme /i/, the corresponding magnitude response of the 18th order all-pole system modeling the vocal tract filter and the corresponding magnitude spectrum of the glottal waveform obtained by applying this algorithm.

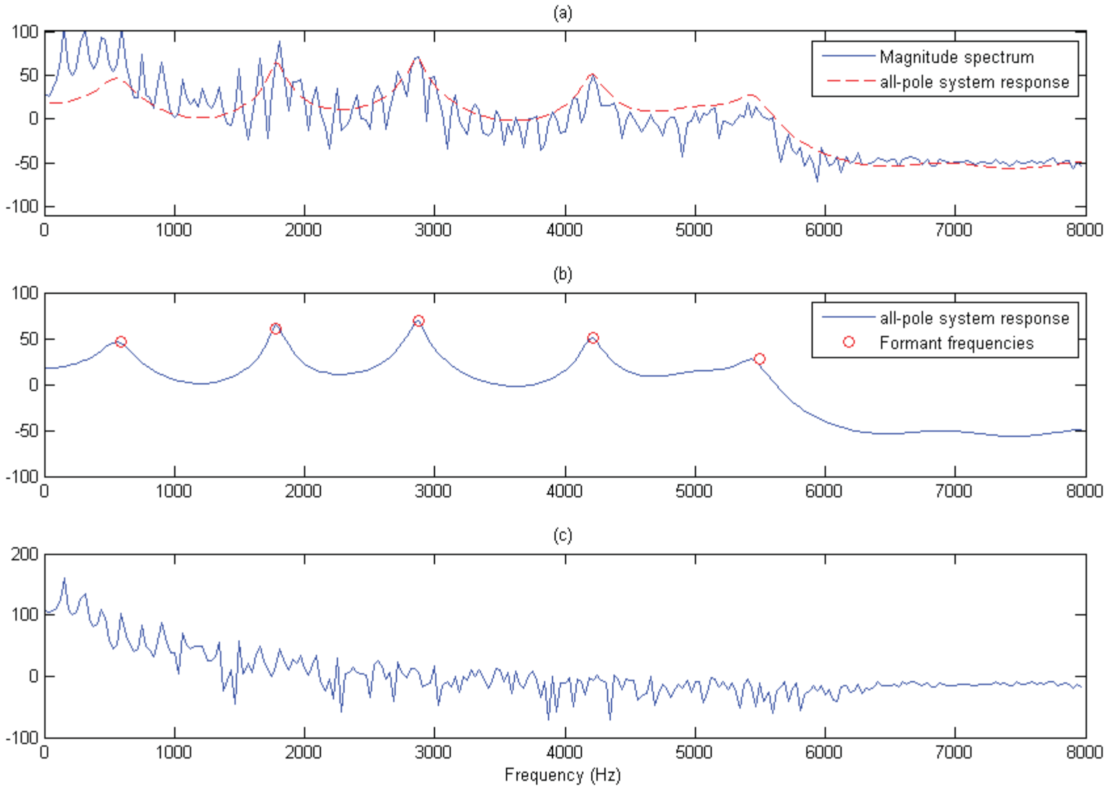


Figure 3.4: (a) Magnitude spectrum of phoneme /i/, (b) the corresponding magnitude response of the vocal tract filter, (c) the corresponding magnitude spectrum of the glottal waveform.

The fundamental frequency of the analyzed speech is 155 Hz. It can be seen from Figure 3.4c that the magnitude spectrum of the glottal waveform is a sequence of pulses located at multiples of 155 Hz. Generally, the amplitude of these pulses decreases with respect to frequency. This decreasing trend can be approximated as a gain of -12 dB/octave as shown in Figure 3.4c. In fact the glottal model can be considered as a second-order low-pass system. The glottal waveform is obtained from the response of this system to an impulse train at pitch period apart.

3.2 Human listening test

The main aim of the listening test is to investigate what sort of speech cues humans use to identify cognitive load (CL) levels. The speech cues, obtained from the feedback of the participants were informally collected after the participants completed the test. They are potentially useful in designing the front-end for the automatic cognitive load classification system. The listening test was conducted on a subset of the Stroop test corpus with eleven untrained listeners, five males and six females, who have no experience in studies related to cognitive load. The test was only performed on a subset of the corpus in order to reduce the testing time. The Reading and Comprehension corpus was not used here because listeners can be able to identify the CL level of speech from the content of the story rather than from the speech itself. The Stroop test corpus, on the other hand, contains only speech of the color words, e.g., *'red'*, *'blue'*, *'green'* here the linguistic content does not reveal any information about the difficulty level of the task. The subset of the Stroop test corpus used for this listening test consists of speech which was randomly chosen from eight speakers, four of whom were male and four female. Three utterances from the three different CL levels were chosen for each speaker.

3.2.1 Test procedure

Figure 3.5 shows the user interface used in the test. Listeners were asked to complete the test for each speaker, before moving to the next speaker. This method of testing was chosen because listeners were not given training utterances but were asked to rank the cognitive load level solely based on each utterance. For each speaker, three utterances with different cognitive load levels were randomly assigned to the 'Speech A', 'Speech B' and 'Speech C' buttons as illustrated in Figure 3.5. Listeners were permitted to listen to these utterances as many times as they wanted by clicking the three buttons. After listening they were required to specify the cognitive load level of each utterance. The reported classification accuracies are obtained as the percentage of test utterances whose cognitive load levels are correctly estimated.

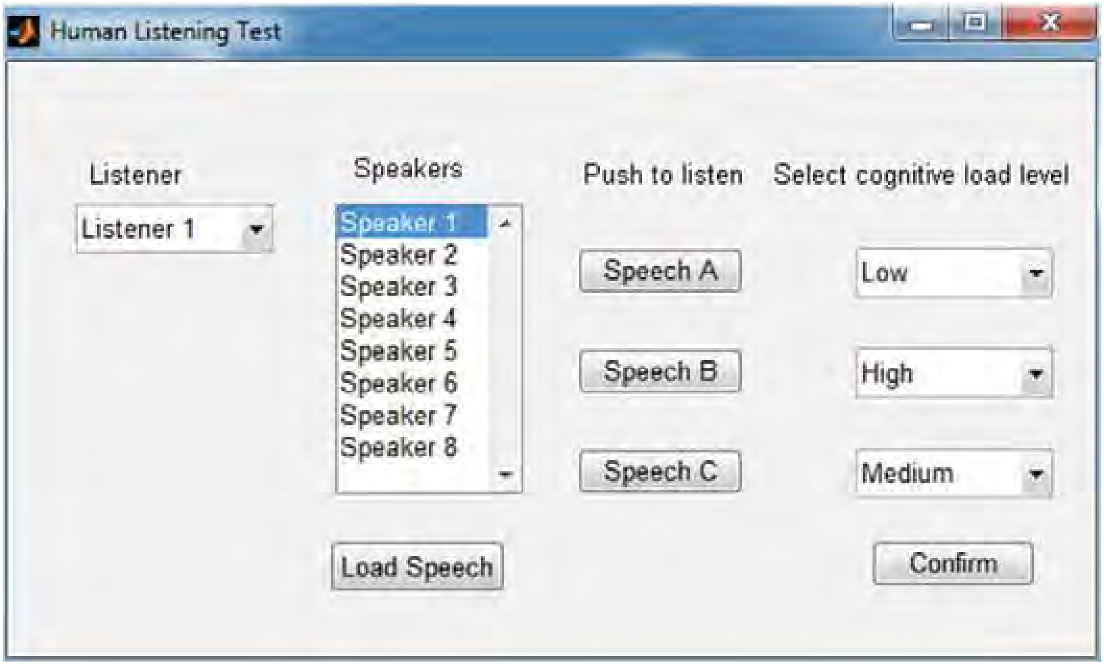


Figure 3.5: The listening test user interface.

3.2.2 Results and discussion

The average accuracy of all eleven listeners was 72.3% and the overall confusion matrix of the test is presented in Table 3.1.

Table 3.1: Confusion matrix of the human listening test.

Actual CL Identified CL	Low	Medium	High
Low	86.4%	5.8%	8.0%
Medium	2.3%	68.2%	29.5%
High	11.4%	26.1%	62.5%
Overall accuracy = 72.3%			

The overall high classification accuracy, as compared to the random level of around 33.3%, indicates that speech does contain cognitive load (CL) information.

It can also be observed from the confusion matrix that the classification accuracy of low cognitive load is significantly higher than that of the medium and high cognitive load. Moreover, the rate of confusion between medium and high load levels of speech is large. This is most probably because the difference in difficulty levels between the low and medium load tasks is much larger than the difference in difficulty levels between the medium and high load tasks. This is understandable given that the low cognitive load speech was collected from the reading color name task. This task is expected to be much easier than the naming color word task on which both the medium and high load speech were collected. Furthermore, the time constraint in the high load task may not have been sufficient to cause a significant difference in terms of the effort subjects need to make in order to perform the high load task compared to the medium load task.

Figure 3.6 shows the classification accuracy of individual listeners in the listening test. This figure indicates that all listeners have the ability to identify the human cognitive load level by listening to their speech, though this ability varies considerably between listeners. The high performance results in the listening test based on the individual or on average strongly suggest that cognitive load cues are contained in speech. These cues can indicate some basic patterns that can be exploited by an automatic speech-based cognitive load classification system.

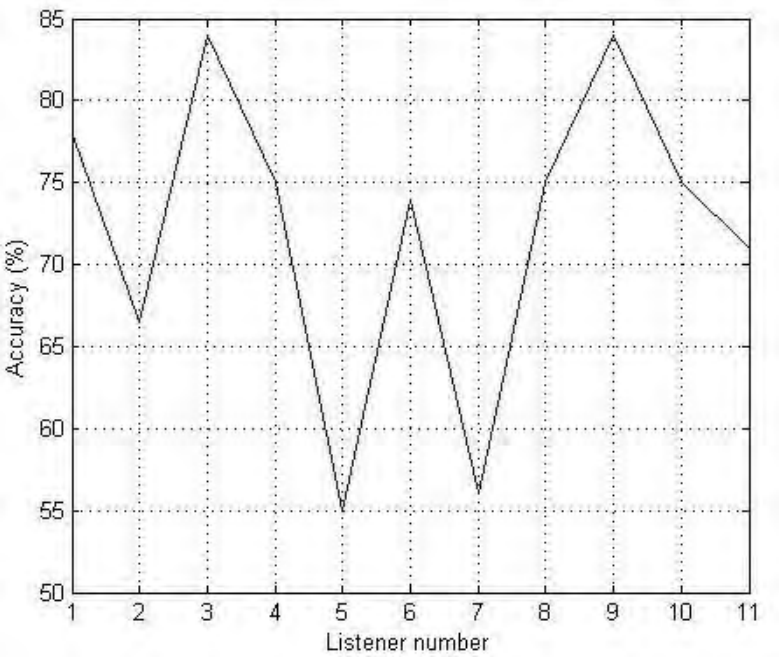


Figure 3.6: Accuracies of individual listener in the listening test.

3.2.3 Speech cues of cognitive load

The feedback from the participants revealed that generally, the cognitive load (CL) levels of speech were identified through numerous speech cues. These are summarized below:

- Low cognitive load speech: this type of speech is uttered naturally and in a relaxed manner, without any confusion or hesitation. This reflects the simplicity of the reading color name task used to collect the speech. The speech cues that the listeners used to identify the low CL level speech are soft breathing, slow and consistent speech rate within an utterance and natural variation in intonation.

- Medium cognitive load speech: this type of speech is uttered unnaturally and with considerable confusion and hesitation. This is indicated by the presence of filler sounds, the parts of speech which are not generally recognized as purposeful or containing formal meaning e.g. ‘uh’ and ‘ah’, in the utterance. Furthermore, the speakers seemed to be nervous when uttering this type of speech, as indicated by heavy breathing. Under medium CL the speech rate is slow and inconsistent, and at times is extremely slow. This might be because speakers were confused and hence spent more time thinking and deciding which words to articulate. One more cue used to identify medium CL speech is less variation intonation, which might be due to the nervousness of the speakers.
- High cognitive load speech: the most distinctive cue indicating this type of speech is the higher speech rate compared to low and medium cognitive load speech. This is most likely due to the time constraint of the task that is used to increase the load level on speakers when collecting the database. The intonation of the high CL speech is flatter than those of the low and medium of cognitive load speech. This might be because of the nervousness of the speakers and the high speech rate of this type of speech.

A summary of speech cues that listeners used to identify the three levels of cognitive load are presented in Table 3.2.

Table 3.2: Speech cues of three CL levels.

Cognitive load level Speech cue	Low	Medium	High
Breath pattern	Soft	Heavy	Heavy
Speech rate (the number of words per unit of time) within utterance	Slow, consistent	Slow, inconsistent	Fast, consistent
Filler sounds e.g. ‘uh’ and ’ah’	Never	Sometimes	Sometimes
Intonation	Naturally varying	Less varying	Flat

From Table 3.2, intonation seems to be a useful speech cue for identifying cognitive load. Intonation is characterized by the pitch contour. A flat intonation is reflected by a level pitch contour, whereas a naturally varying intonation is reflected by a more varying pitch contour. The effectiveness of intonation in cognitive load classification by humans supports the effectiveness of the shifted delta feature of the pitch for automatic CL classification in this thesis (presented in Section 3.4.1.1) and also in previous studies as this feature captures the temporal variation of the pitch contour [3, 5].

3.3 Baseline cognitive load classification system

3.3.1 System setup

The automatic cognitive load classification system used to investigate the effectiveness of speech features presented in this thesis is described in this section. A Universal Background Model – Gaussian Mixture Model (UBM-GMM) based classifier was used as the back-end of the system. Both the Stroop test and Reading and Comprehension corpora were used in these experiments, each comprising of speech recordings collected from fifteen speakers. The speech in these corpora is in the form of short-time utterances. In order to obtain the speech feature, the speech of each utterance is segmented into 25 ms frames with a 15 ms frame overlap. The feature vector of each utterance is then obtained by combining all features extracted from individual voiced frames of speech. The voiced frames are detected from the result of pitch extraction; a speech frame is considered to be voiced if its pitch can be estimated. During the training phase, the speech features extracted from the utterances in the UBM training dataset were used to estimate the parameters of the universal background model. The speech features extracted from training data set utterance for different CL levels are then used to estimate the parameters of the Gaussian mixture model corresponding to these loads using maximum a posteriori adaptation on the universal background model. During the testing phase, the log likelihood of each testing dataset utterance is computed using the speech feature of that utterance based on the equation (2.14). The estimated load level corresponds to the model with the largest log-likelihood score.

3.3.2 Allocation of training and testing data

All the experiments in this thesis were conducted in a speaker independent and text independent manner. Data from speakers used in the test database was not present in the training data and text transcription was not used. For experiments performed on the Reading and Comprehension corpus, data from a set of five speakers was used as the test dataset and data from two other sets with five speakers in each was used as the training dataset, as illustrated in Figure 3.7a. All experiments were performed three times in a round-robin fashion, with each of the three different datasets used to form the test set. All results reported are obtained by averaging over the three instances. This mode of experiment was chosen in order to have a large number of test samples for each test set. When using the Stroop test corpus, leave-one-speaker-out experiments were performed instead. Data from one speaker was used in the testing phase and data from the other fourteen speakers was used in the training phase, as illustrated in Figure 3.7b. Each experiment was performed fifteen times with different speakers used as the test speaker and the results were averaged. This mode of experiment has been used for the Stroop test corpus in previous studies [4-5].

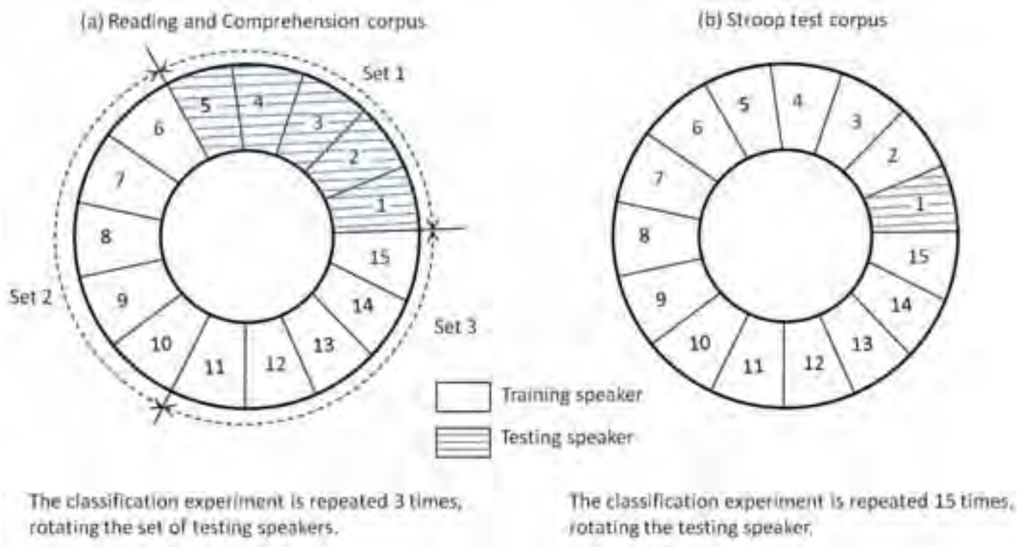

(a) Reading and Comprehension corpus
Set 1
Set 2
Set 3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
(b) Stroop test corpus
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Training speaker
Testing speaker
The classification experiment is repeated 3 times, rotating the set of testing speakers.
The classification experiment is repeated 15 times, rotating the testing speaker.

Figure 3.7: Allocation of training and testing speech data

For experiments conducted on the Reading and Comprehension corpus, the story reading speech of the three cognitive load levels from the training set speakers was used as the universal background model training dataset. The question answering speech data from the training set speakers was used as the GMMs adaptation dataset. The question answering speech data from the test speakers was used to compute the likelihood scores from the Gaussian mixture models. For experiments conducted on the Stroop test corpus, story reading speech data from the training set speakers was used as the universal background model training dataset. The speech corresponding to each of the three CL levels from the training set speakers was used as the GMMs adaptation dataset. Speech corresponding to each of the three cognitive load levels from test speaker, was used to compute the likelihood scores from the Gaussian mixture models.

3.4 The effectiveness of source and filter based features

This section investigates the effectiveness of different speech features, categorized as source-based, filter-based or combined features in classifying cognitive load. The classification performance of these features can be used to evaluate the relative significance of different aspects of the human speech production system in cognitive load classification. To quantify the effectiveness of these features, they were employed individually to perform CL classification and the obtained accuracy was taken as the measure of their effectiveness. The performance of the fusion system based on fusing the classification results of different features was also determined. They were used to evaluate the complementary CL information capacity of different speech features. The classification experiments were conducted under two conditions, either excluding or including the dynamic shifted delta feature vector, in order to investigate the importance of temporal information for the classification.

3.4.1 Source-based features

3.4.1.1 Pitch

Pitch is a feature of the source characterizing the vibration rate of the vocal folds. It is computed for every frame of speech. Amongst the numerous pitch estimation algorithms proposed over the years, the Robust Algorithm for Pitch Tracking (RAPT) proposed by Talkin [84] is one of the most popular algorithms. The pitch feature used for all experiments in this thesis was computed using the RAPT algorithm.

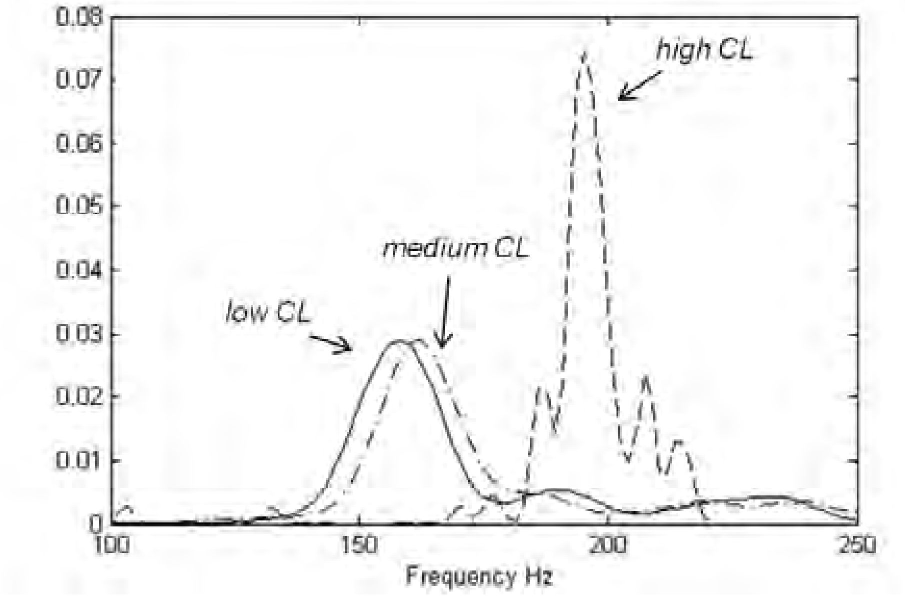


Figure 3.8: Distribution of the pitch of the words '*gray*'.

Figure 3.8 shows an example of the distribution of the pitch values (computed for every 25 ms frames, 15 ms overlapping) of the words '*gray*' in the Stroop test corpus, as spoken by a female speaker under low, medium and high levels of cognitive load. It can be seen from this figure that the pitch increases according to the cognitive load level, which is supported by the results reported in [50-53].

A single pitch (F_0) value is extracted for every frame to test the performance of the pitch feature for classification. The static (original) pitch feature vector is obtained by combining the pitch values of all the frames. Table 3.3 shows the classification accuracy obtained using the pitch feature vector and the combination of pitch and its shifted delta feature vector as the front-end of the automatic CL classification system tested on the Stroop test and the Reading and Comprehension corpora.

Table 3.3: Classification accuracies of the system using pitch.

Corpus	Accuracy (%)	
	Pitch	Pitch and its shifted delta feature
Stroop test	32.8	52.2
Reading and Comprehension	33.3	37.0

The results in Table 3.3 show that although the pitch feature alone did not provide good performance for the classification system, as the classification accuracies obtained are close to the level of selecting the correct CL level (low, medium or high) by chance, the combination of the pitch and its shifted delta feature provided good classification performance. This indicates that temporal variation of the pitch contour is useful for classifying the cognitive load levels and agrees with the usefulness of intonation in the identification of cognitive load level by humans. The classification accuracy obtained for the Stroop test corpus, using the combination of pitch and the shifted delta feature, is significantly higher than that for the Reading comprehensive corpus. This is most probably because the pitch feature in the Reading and Comprehension corpus has greater variability compared to that in the Stroop test corpus due to the high level of phonetic variability in this corpus.

3.4.1.2 Intensity

Intensity characterizes the amplitude of the vocal fold vibration which in turn depends on the pressure of the subglottic airstream. The loudness of speech as perceived by the listener is determined by the sound pressure level of the sound wave at the listener's eardrum. The pressure level depends on the intensity of the speech at the mouth of the speaker and the distance between speaker and listener. In speech analysis, the loudness of recorded speech is determined as the sound pressure level at the microphone. The intensity is dependent on this loudness and the transfer function of the microphone. Therefore in order to use the intensity of speech as a feature to characterize the amplitude of the vocal fold vibrations, it is necessary to ensure that the distance between the speakers and the microphone are fixed for all utterances recorded. This assumption can reasonably be made as all the recording processes of the Stroop test and the Reading and Comprehension corpora occurred in the same recording studio, using a close talk headset. Like pitch, intensity is a single dimension feature with one intensity value per frame. The Praat software [85] is used to extract the intensity feature from the speech in this thesis. The accuracies of the automatic CL classification system using the intensity feature vector as the front-end are provided in Table 3.4.

Table 3.4: Classification accuracies using intensity.

Corpus	Accuracy (%)	
	Intensity	Intensity and its shifted delta feature
Stroop test	32.8	56.9
Reading and Comprehension	34.1	41.5

As seen in Table 3.4, the combination of intensity and its shifted delta feature produced a high classification accuracy for the system although the intensity alone did not perform better than the level of selecting the correct CL level by chance. This suggests that the temporal variation of the intensity is useful in characterizing the cognitive load level. Furthermore, when a combination of the intensity and the shifted delta feature were used, the accuracy obtained on the Stroop test corpus is again much higher than that of the Reading and Comprehension corpus, similar to the results found when using the pitch feature.

Although both intensity and pitch are source-based features, they capture different aspects of the voice source. The cognitive load information contained in these two features can therefore complement each other. As such, the incorporation of the information from these two features should be considered as this can improve the performance of the system. The classification accuracies of the fusion of the intensity-based and the pitch-based systems with and without the shifted delta feature (SDF) are provided in Table 3.5.

Table 3.5: Accuracies of the fusion of pitch-based and intensity-based systems.

Corpus	Accuracy (%)	
	Without SDF	With SDF
Stroop test	39.1	68.6
Reading and Comprehension	36.3	44.4

It can be observed from Tables 3.3, 3.4, and 3.5 that the classification accuracies of the fusion of pitch-based and intensity-based systems are higher than those of the system based on individual features. This shows that the CL information contained in the pitch and intensity features are complementary as expected.

3.4.1.3 Source Mel frequency cepstral coefficients (SMFCC)

This feature is a compact representation of the spectral envelope of the glottal waveform. It is extracted through a filtering process where a series of triangular filters that are equally spaced in the mel frequency scale are used to filter the estimated glottal waveform.

To compute the SMFCCs, a windowing process is applied to the glottal waveform to segment it into short-time frames. The magnitude spectrum is computed for these glottal frames and it is then multiplied by the magnitude responses of the filters to compute the average energies within each filter band. The logarithm of these energies is then taken in order to reduce their dynamic range. The discrete cosine transform (DCT) is then applied to this results, and finally the SMFCCs are obtained as the first *N* DCT coefficients. This allows the reduction of the dimension of the feature vector at the cost of detailed information about the magnitude spectrum. The SMFCC feature extraction process is summarized in the block diagram shown in Figure 3.9. In this section, twenty filters were used to compute the spectrum energies and the number of DCT coefficients used is *N = 12*.

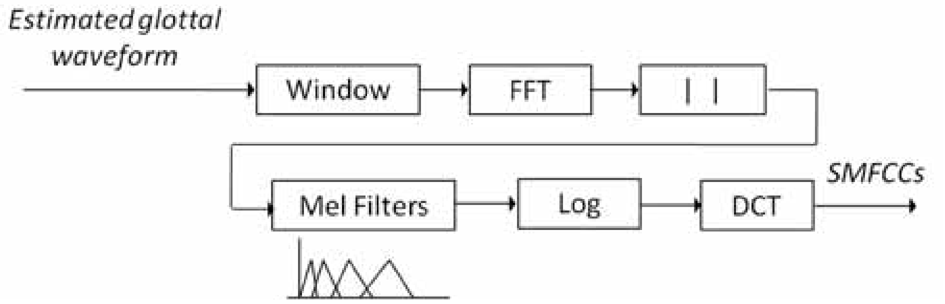


Figure 3.9: Block diagram of SMFCCs extraction.

Figure 3.10 shows the magnitude spectra of glottal waveforms computed on two 25 ms segments of speech of the phoneme /uw/ spoken by a female speaker under two different load levels. The variation of the spectrum of the glottal waveform due to changes in the load level can be observed from this figure. For instance, the distance between adjacent pulses of the spectrum (which is equal to the pitch F_0), of the low CL is less than that of the high CL. This again indicates that pitch F_0 increases when the CL level increases, which agrees with the observation in Figure 3.8.

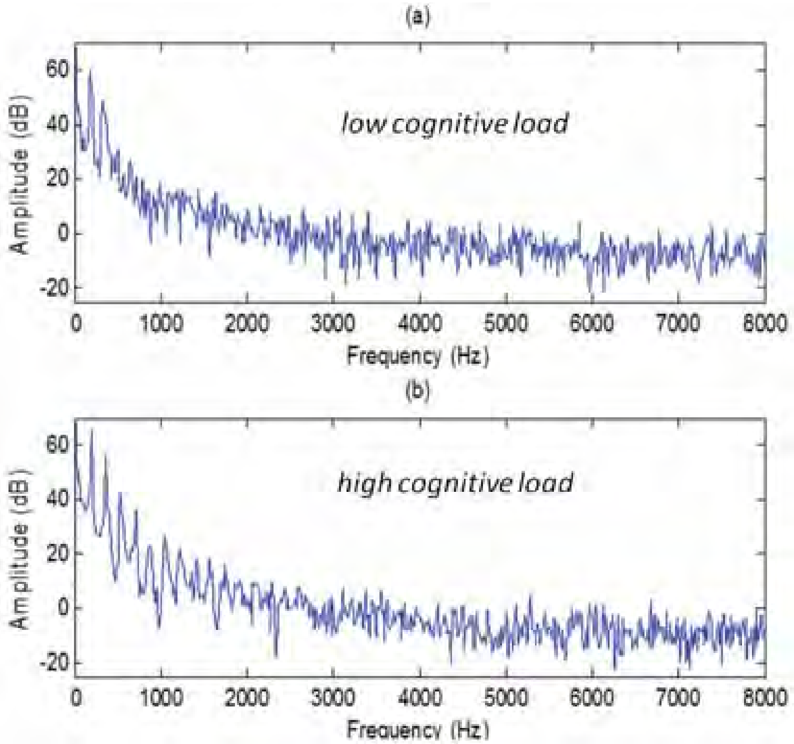


Figure 3.10: Magnitude spectrum of the glottal waveform of the phoneme /uw/ spoken under low CL (a) and high CL (b).

Furthermore, the distribution of the first coefficients of SMFCC computed from the words '*gray*' spoken by a female speaker under three different load levels in the Stroop test corpus is shown in Figure 3.11. It can be seen from this figure that discrimination of SMFCC for different levels of CL exists and hence this feature can be used to classify cognitive load levels.

The classification accuracies when the SMFCC feature is used as front-end feature in the classification system, with and without concatenating its shifted delta feature, are shown in Table 3.6.

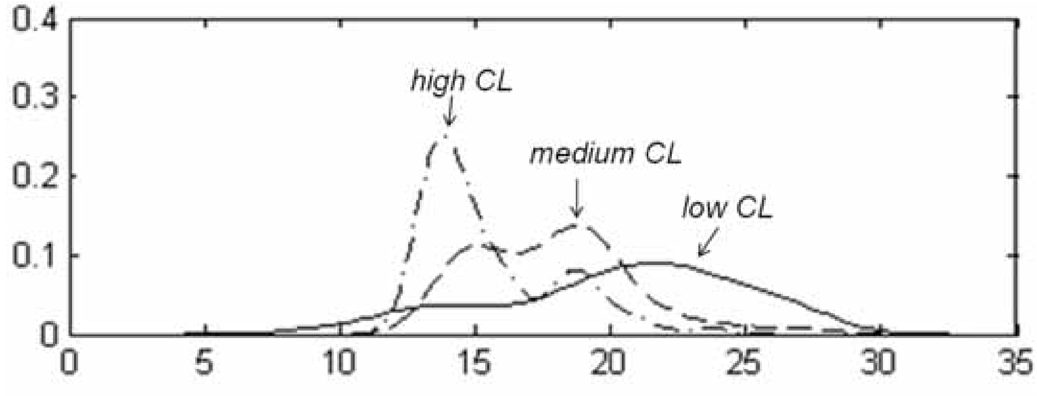


Figure 3.11: Distribution of the first SMFCC of the word '*gray*' for low, medium and high CL.

Table 3.6: Classification accuracies using SMFCC.

Corpus	Accuracy (%)	
	SMFCC	SMFCC and its shifted delta feature
Stroop test	38.2	58.9
Reading and Comprehension	34.2	38.5

It can be seen from Table 3.6 that all the systems based on the SMFCC feature have a high classification accuracy, except for the system using the static SMFCC feature on the Reading and Comprehension corpus. In addition to the usefulness of the other two voice source related features namely pitch and intensity, this indicates that voice source related speech features are effective for CL classification.

3.4.2 Filter-based features

3.4.2.1 Formant frequencies

When the excitation signal passes through the vocal tract it is shaped by the resonance characteristic of the vocal tract. This produces a number of peaks in the magnitude spectrum of the speech signal that are located at the resonant frequencies of the filter modeling the vocal tract, as shown previously in Figure 3.4. These peaks are dominant spectral components of speech and are referred as speech formants.

In terms of human speech production modeling, the vocal tract is modeled as an all-pole system or filter whole parameters can be estimated by linear prediction analysis. The formant frequencies can then be estimated from the magnitude response of the all-pole system. In this thesis, the formant frequencies are extracted using the Wavesurfer/Snack toolkit [86] and the order of the all-pole system used is 10.

The classification accuracies of the automatic cognitive load classification system using the frequency of the first three formants, with and without the concatenation of their shifted delta feature (SDF), as front-end features are provided in Table 3.7.

Table 3.7: Classification accuracies using formant frequency.

Corpus	Accuracy (%)	
	Formant frequency	Formant frequency and its SDF
Stroop test	57.4	75.6
Reading and Comprehension	31.9	45.9

It can be seen from Table 3.7 that the formant frequency provided high performance for all the classification systems, except for the system using only the formant frequency on the Reading and Comprehension corpus. This shows that formant frequencies are useful for the classification.

Furthermore, as formant frequency is a filter-based feature while pitch and intensity are the source-based features, the information contained in the formant frequency can complement that contained in pitch and intensity. Incorporating the cognitive load information from the formant frequency with that from either the pitch or intensity can improve the performance of the system. The classification accuracies of the fusion of the formant-based system with either the pitch-based or intensity-based system, with and without using the shifted delta feature (SDF), are provided in Tables 3.8 and 3.9 respectively.

Table 3.8: Accuracies of fusion of formant-based and pitch-based systems.

Corpus	Accuracy (%)	
	Without SDF	With SDF
Stroop test	60.3	75.6
Reading and Comprehension	37.0	46.7

Table 3.9: Accuracies of fusion of formant-based and intensity-based systems.

Corpus	Accuracy (%)	
	Without SDF	With SDF
Stroop test	64.1	78.2
Reading and Comprehension	38.5	51.9

The results in the Tables 3.3, 3.4, 3.7, 3.8 and 3.9 show that the classification accuracies of the fusion of the formant-based system with that of either pitch-based or intensity-based systems are higher than the classification accuracy of the systems based on individual speech features. This shows that incorporating the information from a filter-based feature e.g. formant frequencies, and a source-based feature e.g. pitch or intensity, increases the amount of cognitive load information, due to the fact that the information contained in these features is complementary. Furthermore, it is interesting to see that the improvement of accuracy when fusing the pitch-based and the formant-based systems is less than that when fusing the intensity-based and formant-based systems. This is most probably because the pitch feature is more correlated to the formant feature than the intensity feature.

3.4.2.2 Filter Mel frequency cepstral coefficients (FMFCC)

The FMFCC feature is a compact representation of the spectral envelope of the vocal tract filter. In order to compute this feature, the spectral envelope of the vocal tract filter is estimated from the linear prediction coefficients which were obtained from the implementation of Iterative Adaptive Inverse Filtering algorithm mentioned in Section 3.1.3. This spectral envelope is then passed through a filterbank of twenty mel-scale filters. Then FMFCCs are then obtained as the first twelve coefficients of the discrete cosine transform of the logarithm of the filter output energies.

Figure 3.12 shows the spectral envelope of the vocal tract filters computed on a 25 ms segment of the phoneme /uw/ uttered by a female speaker under two different load levels. The variation of the spectral envelope of the vocal tract filter due to cognitive load can be observed from this figure. For instance, the second formant frequency decreases when the load increases, as found in [54]. Moreover, the bandwidths of the formants of the low CL speech are larger than those of the high CL speech.

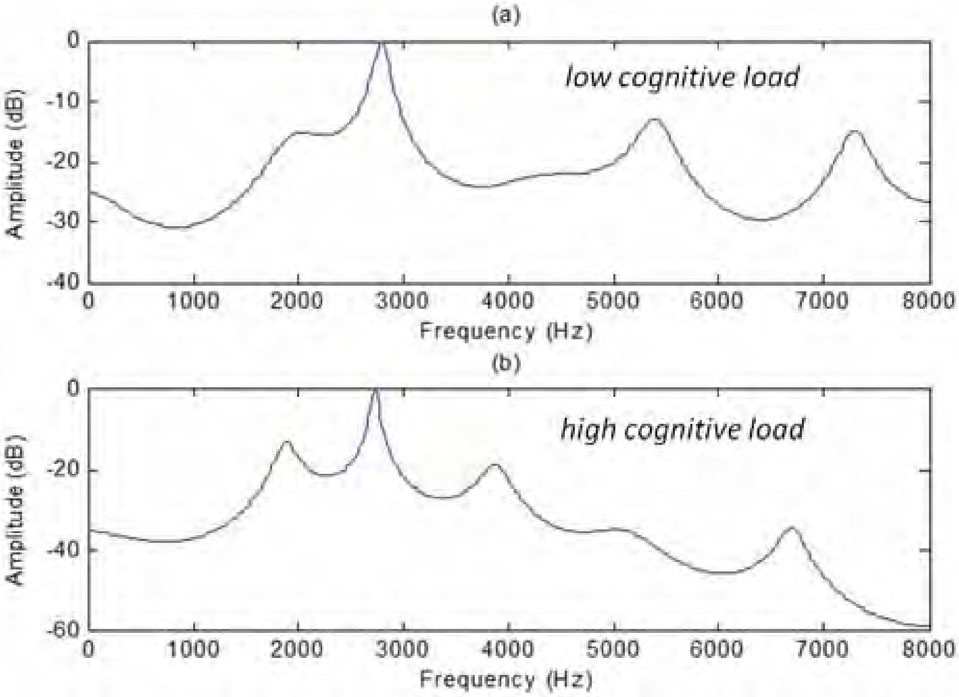


Figure 3.12: Spectral envelope of vocal tract filter of phoneme /uw/ uttered under low CL (a) and high CL (b).

The distribution of the first FMFCC computed on the words '*gray*' spoken by a female speaker with three different levels of cognitive load in the Stroop test corpus are shown in Figure 3.13. It can be seen from this figure that there is discrimination of FMFCC of different levels of cognitive load. Therefore, this feature can be useful for the classification.

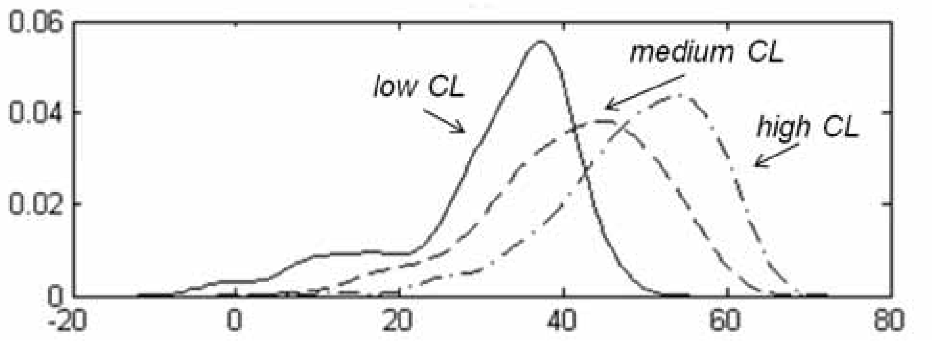


Figure 3.13: Distribution of the first FMFCC of the word '*gray*' for low, medium and high CL.

The classification accuracies obtained when FMFCC features were used as front-end features in the classification system are presented in Table 3.10. Furthermore, as FMFCC captures the variation of the spectral envelope of the vocal tract filter while SMFCC captures the variation of the spectral envelope of the source excitation signal, the cognitive load information contained in these two features can be complementary. It is therefore expected that incorporating the information from these two features can improve the performance of the classification system. The classification accuracies of the fusion of the FMFCC-based and SMFCC-based systems, with and without the shifted delta features (SDF), are presented in Table 3.11.

Table 3.10: Classification accuracies using FMFCC.

Corpus	Accuracy (%)	
	FMFCC	FMFCC and its shifted delta feature
Stroop test	41.1	70.4
Reading and Comprehension	36.2	47.4

Table 3.11: Accuracies of fusion of SMFCC-based and FMFCC-based systems.

Corpus	Accuracy (%)	
	Without SDF	With SDF
Stroop test	43.2	76.9
Reading and Comprehension	39.1	52.6

It can be seen from Table 3.10 that all the classification systems based on the FMFCC feature produced high accuracy. This suggests that there is CL information contained in the FMFCC features and that they are useful for the classification. Furthermore, it can be observed from Tables 3.6, 3.10, and 3.11 that the fusion of SMFCC-based and FMFCC-based systems produces higher classification accuracy than systems based on individual features. This indicates that the information in these two features is complementary, as expected.

3.4.3 Combined features

3.4.3.1 Mel frequency cepstral coefficients (MFCCs)

The MFCCs are a compact representation of the speech spectral envelope estimated based on the subband spectral powers obtained using a series of filters. The computation of this feature follows exactly the same steps as the computation of SMFCC as shown in Figure 3.9, except that the input is a speech signal rather than the glottal waveform. As MFCCs are computed from the magnitude spectrum of speech, the information from both the voice source and the vocal tract filter is contained in this feature. This feature can therefore be considered as a combined feature. The accuracy of the CL classification system when MFCCs are used at the front-end are presented in Table 3.12.

Table 3.12: Classification accuracies using MFCCs.

Corpus	Accuracy (%)	
	MFCCs	MFCCs and their shifted delta feature
Stroop test	45.6	74.3
Reading and Comprehension	40.0	60.7

The results in Table 3.12 indicate that the accuracies of all MFCC-based classification systems are significantly higher than the level of selecting the correct CL level by chance. The MFCC feature is therefore very useful for classifying the load levels.

3.4.3.2 Spectral slope and spectral intercept

The spectral slope and spectral intercept are computed from a linear approximation to the speech spectrum. The spectral slope characterizes how quickly energy drops as frequency increases and the spectral intercept is an approximation of the energy at zero frequency. In this thesis, a straight line that best fits the magnitude spectrum of speech in the least square sense is first estimated. The spectral slope (a) and spectral intercept (b) are then obtained as the slope and the intercept of this straight line as shown in Fig. 3.14.

The accuracy of the classification system obtained when the combination of the spectral slope and spectral intercept is used at the front-end of the system is shown in Table 3.13.

Table 3.13: The accuracies using combination of spectral slope (SS) and intercept (SI).

Corpus	Accuracy (%)	
	SS and SI	SS, SI and their shifted delta features
Stroop test	40.4	63.7
Reading and Comprehension	41.5	46.7

The results in Table 3.13 indicate that the combination of spectral envelope and spectral intercept are useful for CL classification. Furthermore, like other features, the temporal information of this feature is very important for the classification.

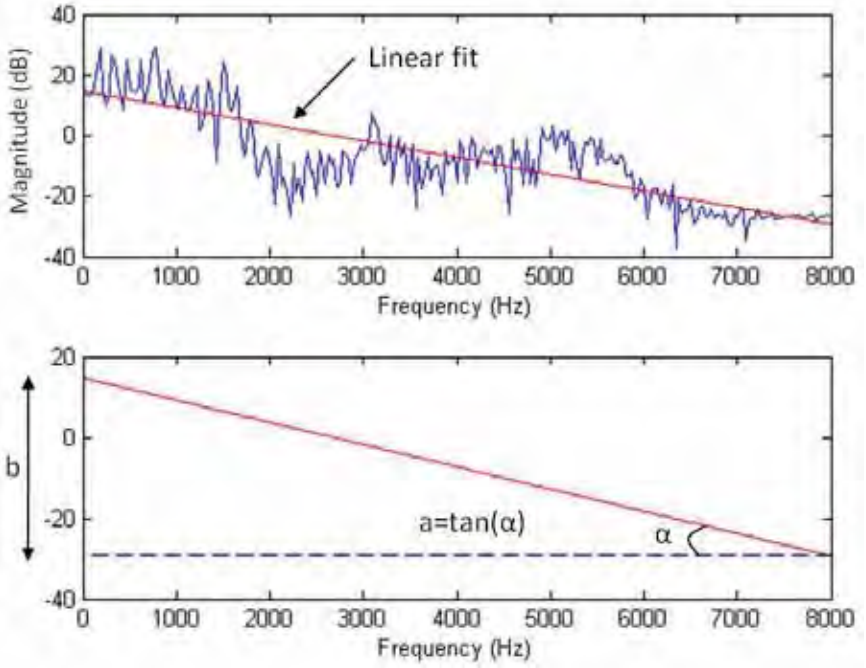


Figure 3.14: The estimation of spectral slope and spectral intercept features.

3.4.3.3 *Group delay feature (GD)*

The MFCC feature is extracted from the magnitude spectrum of speech. The information of the phase spectrum is therefore ignored by using this feature. The group delay feature, on the other hand, represents the spectral phase [87-88]. Cognitive load information contained in the group delay feature is therefore expected to complement that contained in the MFCC feature. The group delay feature therefore can be used to improve the performance of the MFCC-based cognitive load classification system.

In order to compute the group delay feature, linear prediction analysis is used to estimate the coefficients of the all-pole model of the vocal tract filter. The phase response $\phi(f)$ of this all-pole model is then estimated. The group delay $G(f)$ is obtained as the negative of the differentiation of this phase response as

$$G(f) = -\frac{d\phi(f)}{df} \tag{3.6}$$

At this stage, the dimension of the group delay is very large. The discrete cosine transform (DCT) is applied to the original group delay so that a compact representation of the group delay can be obtained as the first ten DCT coefficients. The process of group delay feature extraction is briefly described in Figure 3.15.

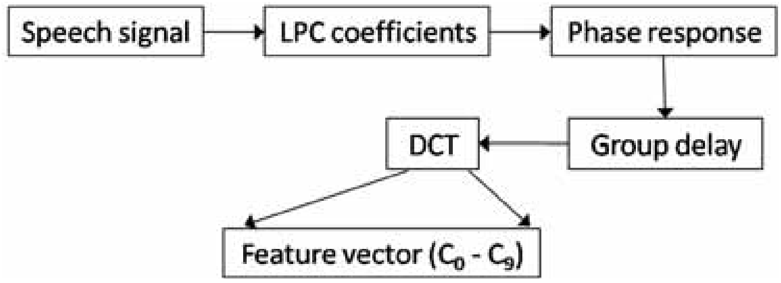


Figure 3.15: Extraction of the group delay feature [89].

The classification accuracies obtained when using group delay as the front-end are reported in Table 3.14. In addition, the classification accuracies of the fusion of the group delay feature based and the MFCC-based systems are provided in Table 3.15.

Table 3.14: Classification accuracies using group delay feature (GD).

Corpus	Accuracy (%)	
	GD	GD and its shifted delta feature
Stroop test	40.7	72.0
Reading and Comprehension	43.7	52.5

Table 3.15: Accuracies of fusion of group delay and MFCC features.

Corpus	Accuracy (%)	
	Without SDF	With SDF
Stroop test	46.5	76.3
Reading and Comprehension	45.2	62.6

It can be seen from Table 3.14 that group delay feature provided high accuracy for the classification system. This indicates that there is the cognitive load information contained in group delay feature and that it is useful for the classification system. Furthermore, the accuracies presented in Tables 3.12, 3.14, and 3.15 indicate that the fusion of the MFCC-based system and the group delay based system produces higher classification accuracy than the systems based on individual features. This shows that the cognitive load information contained in the group delay and the MFCC features are complementary.

3.4.3.4 Frequency modulation (FM)

The frequency modulation (FM) feature is motivated by the AM-FM model of speech signals which was inspired by evidence of such modulation in speech production. According to the AM-FM model, the speech signal $s[n]$ is expressed as the sum of all resonances [90]

$$s[n]=\sum_{k=1}^{K}a_k[n]\cos(\phi_k[n]) \tag{3.7}$$

where K is the total number of resonances, $a_k[n]$ and $\phi_k[n]$ are the amplitude and the phase of the k^{th} resonance respectively, and n is the sample index. A series of band pass filters is used to isolate these resonances and extract the frequency modulation feature corresponding to each resonance. The output of the k^{th} filter can be expressed in the form of the AM-FM model as [90]

$$p_k[n]=a_k[n]\cos\left(\frac{2\pi f_{ck}n}{f_s}+\frac{2\pi}{f_s}\sum_{r=1}^{n}q_k[r]\right) \tag{3.8}$$

where $q_k[n]$ is the FM component, f_s is the sampling frequency f_{ck} is the center frequency of the k^{th} band pass filter.

Numerous methods to estimate FM from speech have been proposed over the years [91]. Among them, the method based on the second-order all-pole model has been shown to produce a considerably more consistent estimate [91]. In this method a second-order all-pole resonator, a simple but effective model to characterize the band pass filter, is used to model the FM component in each filter band. The parameters of this resonator are estimated using linear prediction and the frequency modulation feature is obtained from the pole angle of the estimated all-pole resonator. The frequency modulation features used for all the experiments reported in this thesis are extracted using this second-order all-pole model method. A bank of twenty one Gabor filters is used to isolate the resonators [92].

The accuracies of the classification system when frequency modulation features are used as the front-end are shown in Table 3.16. This suggests that cognitive load level is characterized by the FM features. However, compared to the MFCC-based system, the FM-based system is less accurate which shows that FM features are less important than MFCC for the classification. Furthermore, the accuracies of the fusion of FM-based and the MFCC-based systems, with and without using the shifted delta features (SDF), are provided in Table 3.17.

Table 3.16: Accuracies using frequency modulation (FM) feature.

Corpus	Accuracy (%)	
	FM	FM and its shifted delta feature
Stroop test	45.0	51.1
Reading and Comprehension	36.3	40.7

Table 3.17: Accuracies of fusion of FM-based and MFCC-based systems.

Corpus	Accuracy (%)	
	Without SDF	With SDF
Stroop test	49.1	75.0
Reading and Comprehension	43.2	63.0

It can be seen from Table 3.16 that the frequency modulation feature provided high performance for the system. Furthermore, it can be seen from Tables 3.12, 3.16 and 3.17 that the fusion of the MFCC-based and FM-based systems produce higher classification accuracy than those of the systems based on individual features. This suggests that the information in the FM feature complements that in the MFCC feature.

3.5 The effectiveness of spectral centroid features

The effectiveness of MFCC, as presented in Section 3.4.3.1, suggests that information related to the spectral envelope is useful for cognitive load (CL) classification. However, MFCC does not completely characterize the spectral envelope as some details about the speech spectrum, such as the spectral energy distributions within filter subbands, are not captured by MFCC. This is because in this feature, information in each subband is represented by a single value that represents the total spectral power contained in that subband. Spectral centroid features can be used to capture more information about these subband spectral distributions and hence can be useful for classifying the CL levels.

The spectral centroid frequency (SCF) is an estimate of the ‘center of gravity’ of the spectrum within each subband. Originally proposed as a feature for speech recognition systems, it has been reported that the SCF is a formant-like feature, as it provides the approximate location of formant frequencies within the subbands [93]. However, this feature can be estimated easily and reliably, unlike the formant feature [93]. Also since features based on formant frequencies have been recognized to be effective for CL classification, as presented in Section 3.4.2.1, we can expect that the SCF will also prove to be useful for cognitive load classification systems. In addition to spectral centroid frequency, the use of another feature termed spectral centroid amplitude (SCA), which is the weighted average magnitude spectrum in the subband, is also proposed in this section.

3.5.1 Feature extraction

The spectral centroid features are extracted from speech frames as follows: Let $s[n]$ represent a speech frame of length N in the time domain where $n \in [0, N\text{-}1]$ and let $S[f]$ represent the spectrum of this frame. Then $S[f]$ can be divided into M subbands by using a series of Gabor filters [92] whose frequency responses are $W_m[f]$, where $m \in [1, M]$.

Assume that the m^{th} subband has a lowest frequency l_m and a highest frequency u_m. Each of the two spectral centroid features can be calculated from $S[f]$ for the m^{th} subband as follows.

The spectral centroid frequency (SCF) is computed as the weighted average frequency for a given subband, where the weights are the normalized energy of each frequency component in that subband, expressed as

$$SCF_m = \frac{\sum_{f=l_m}^{f=u_m} f \left| W_m[f] S[f] \right|}{\sum_{f=l_m}^{f=u_m} \left| W_m[f] S[f] \right|} \tag{3.8}$$

The final SCF vector of each frame is obtained by concatenating all the $\mathrm{SCF_m}$.

The spectral centroid amplitude (SCA) is the weighted average magnitude spectrum in the subband, with the frequency serving as weights, as shown in equation (3.9) [94].

$$SCA_m = \frac{\sum_{f=l_m}^{f=u_m} f \left| W_m[f] S[f] \right|}{\sum_{f=l_m}^{f=u_m} f} \tag{3.9}$$

A feature vector is obtained by concatenating all $\mathrm{SCA_m}$ in that frame, then a logarithm is applied to reduce the dynamic range of the feature vector. The discrete cosine transform (DCT) is then applied to obtain the final SCA feature vector. The use of the DCT is intended to decorrelate the feature vector, as it does when conventionally used in computing MFCCs. The computation of the SCA is expressed as

$$SCA(k) = \frac{1}{\sqrt{M}} \mu_k \sum_{m=0}^{M-1} \log(SCA_m) \cos\left(\frac{\pi}{2M} (2m+1) k \right) \tag{3.10}$$

where $k = 0, \ldots, M-1$; $\mu_0 = 1$; $\mu_k = \sqrt{2}$ for $1 \le k \le M-1$

In the case of the SCF the DCT is not applied similar to [93] as it is a frequency based feature. This decision is supported by the choice not to apply the DCT to another similar frequency based feature called the frame-averaged frequency modulation feature in [95]. Henceforth, SCF_m and SCA_m will refer to the feature values in each subband, whereas SCF and SCA will refer to the final spectral centroid feature vectors.

As the spectral centroid frequency and spectral centroid amplitude features are computed from the spectrum of speech containing information from both the voice source and vocal tract filter, they are considered to be combined features. The stages involved in the computation of the spectral centroid features are illustrated in Figure 3.16.

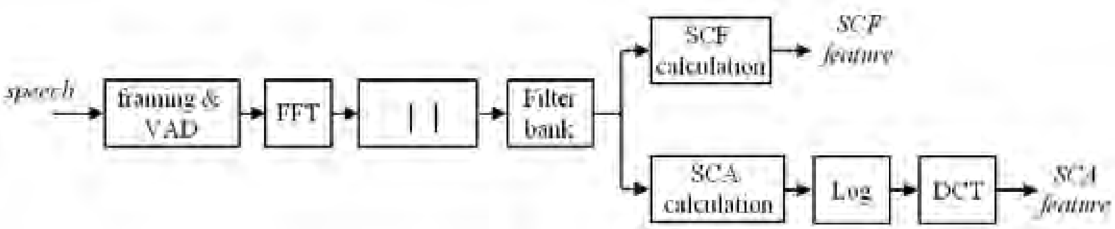


Figure 3.16: Block diagram of SCF & SCA feature extraction [94].

3.5.2 Complementary behavior between spectral centroid and MFCC features

As previously mentioned, the MFCCs are computed from the total energy in each subband and hence will only reflect the variation of the total energy in a subband. However, there are instances where the distribution of energy within each subband varies but the total energy does not and MFCCs will not reflect this. The use of frequency as weights for computing the SCA_m allows for the variations in the spectral energy distribution in these instances to be reflected in the SCA_m values, as shown in Figure 3.17. The energies of the two spectra shown in this figure are the same but the SCA_m values are different.

As explained previously, the spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) features capture different aspects of the spectral distribution in each subband and are therefore expected to complement each other. The complementary nature of these features is illustrated in Figure 3.18, which shows the spectral centroid features corresponding to different examples of synthetic spectra in two different subbands. These spectra comprise of straight lines with varying slopes. It can be observed that the resultant variations in SCF and SCA are very different. Moreover, there are regions of the energy-slope plane where one of the two features varies more than the other. In Figure 3.18, lines of constant energy correspond to constant MFCC feature values. Hence MFCCs cannot distinguish between them, by way of contrast with the SCA and SCF. The use of the frequency as the weight also makes the SCA values in different subbands very different as shown in Figures 3.18c and 3.18d, though these subbands have the same spectral distributions.

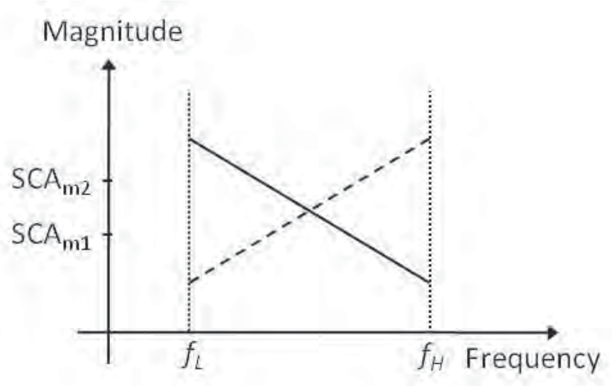


Figure 3.17: Example of the spectra in the mth subband $[f_L, f_H]$.

The solid line is the spectrum 1 and the dashed line is the spectrum 2 after [94].

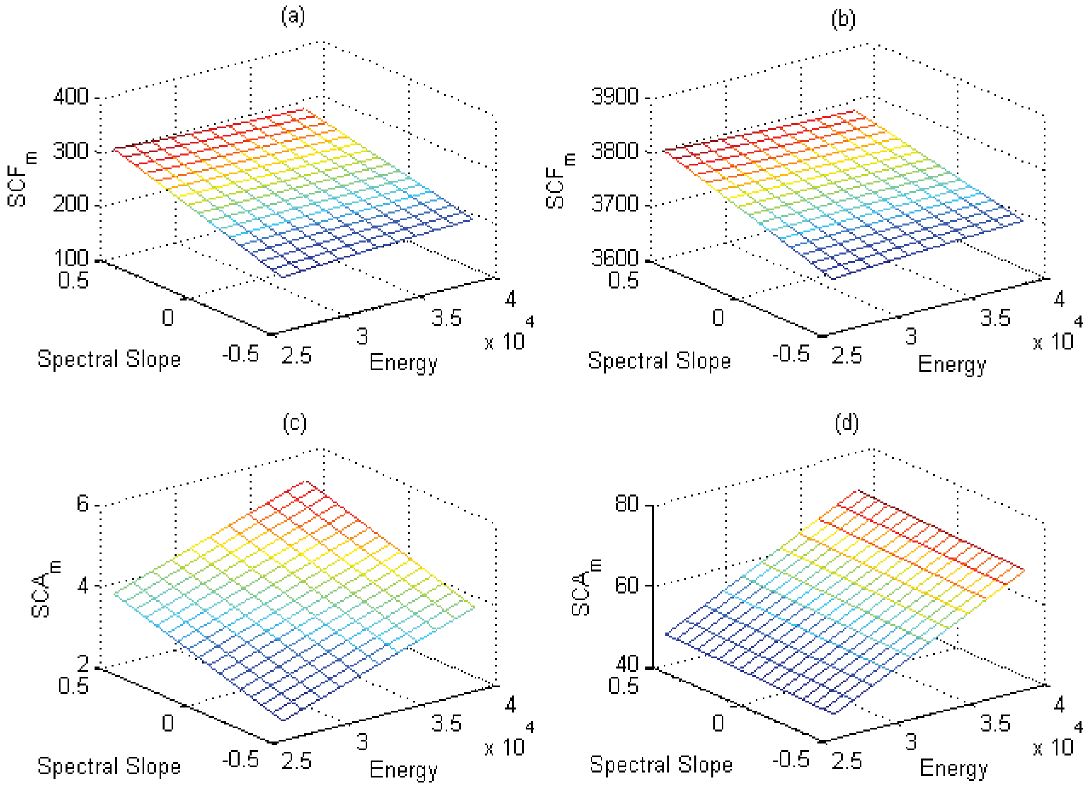


Figure 3.18: The variation of the SCF_m and SCA_m in two subbands (a) & (c) for the low frequency subband, and (b) & (d) for the high frequency subband.

3.5.3 Cognitive load (CL) discrimination ability of spectral centroid features

Figure 3.19 shows the spectral centroid features as well as the spectral envelopes for low, medium and high cognitive load levels extracted from an utterance of the vowel */ey/* spoken by a female speaker in the Stroop test corpus. In this example, the spectral centroid features were extracted by splitting the speech spectrum into six non-overlapping equally spaced subbands in the mel scale. The number of subbands was chosen to make visualization simple. It can be observed that the roll off in the spectral envelope is steeper for high CL. Since the SCA_m are computed as the weighted average spectral power in each subband, this large negative slope results in the value of the high frequency SCA_m for high cognitive load being substantially lower than that of low cognitive load. On the other hand, the low frequency SCA_m of high cognitive load is larger than that of low cognitive load. In addition to the differences in spectral slopes, it can also be observed that the spectral power distributions in the individual subbands are different, which results in the SCF_m in each subband varying between different cognitive load levels.

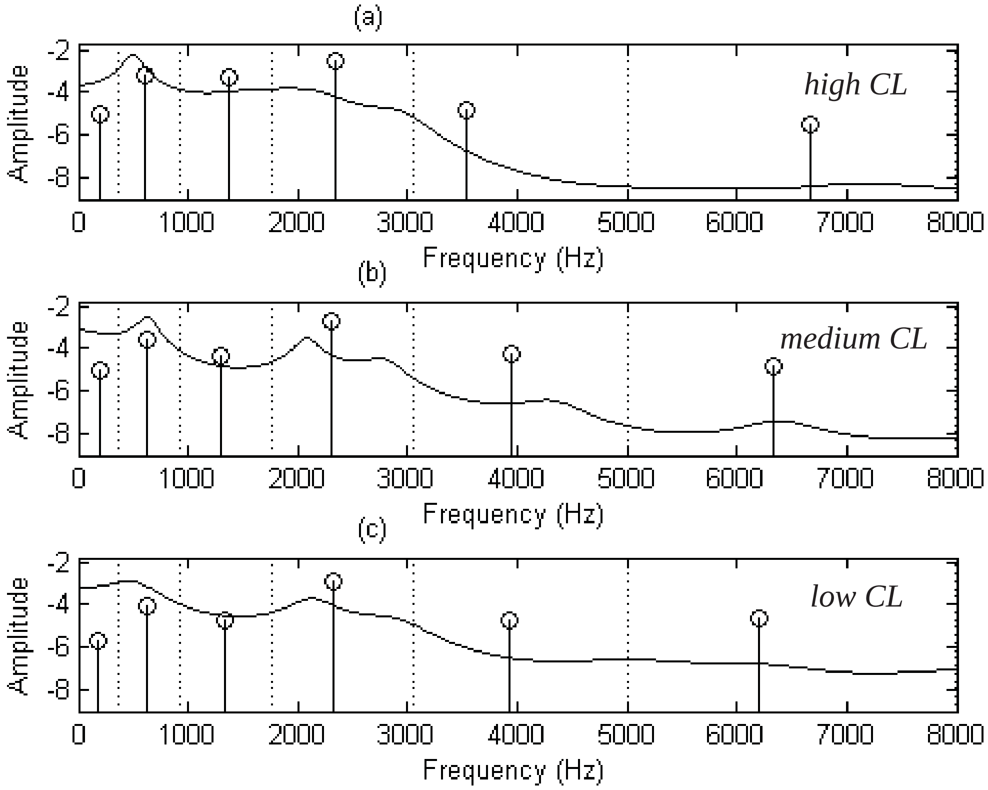


Figure 3.19: Subband spectral centroid frequencies (SCF_m), subband spectral centroid amplitudes (SCA_m), and linear predictive spectral envelope of the vowel /ey/ under (a) high CL, (b) medium CL and (c) low CL. SCF_m are shown by locations of the stems, SCA_m are shown by the amplitude of the stems, the subband boundaries are shown by the dotted vertical lines, and the spectral envelope is shown by the solid continuous curve.

Figure 3.20 shows the statistical spread of the six coefficients of spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) features, computed from speech of the word ‘blue’ spoken by a female speaker in the Stroop test corpus. In this figure, the thick bar extends from the 15th to the 85th percentile, the thin bar extends from the 5th to the 95th percentile, and the middle strip indicates the median of the distribution. The potential for discrimination between different cognitive load levels can be observed from this figure.

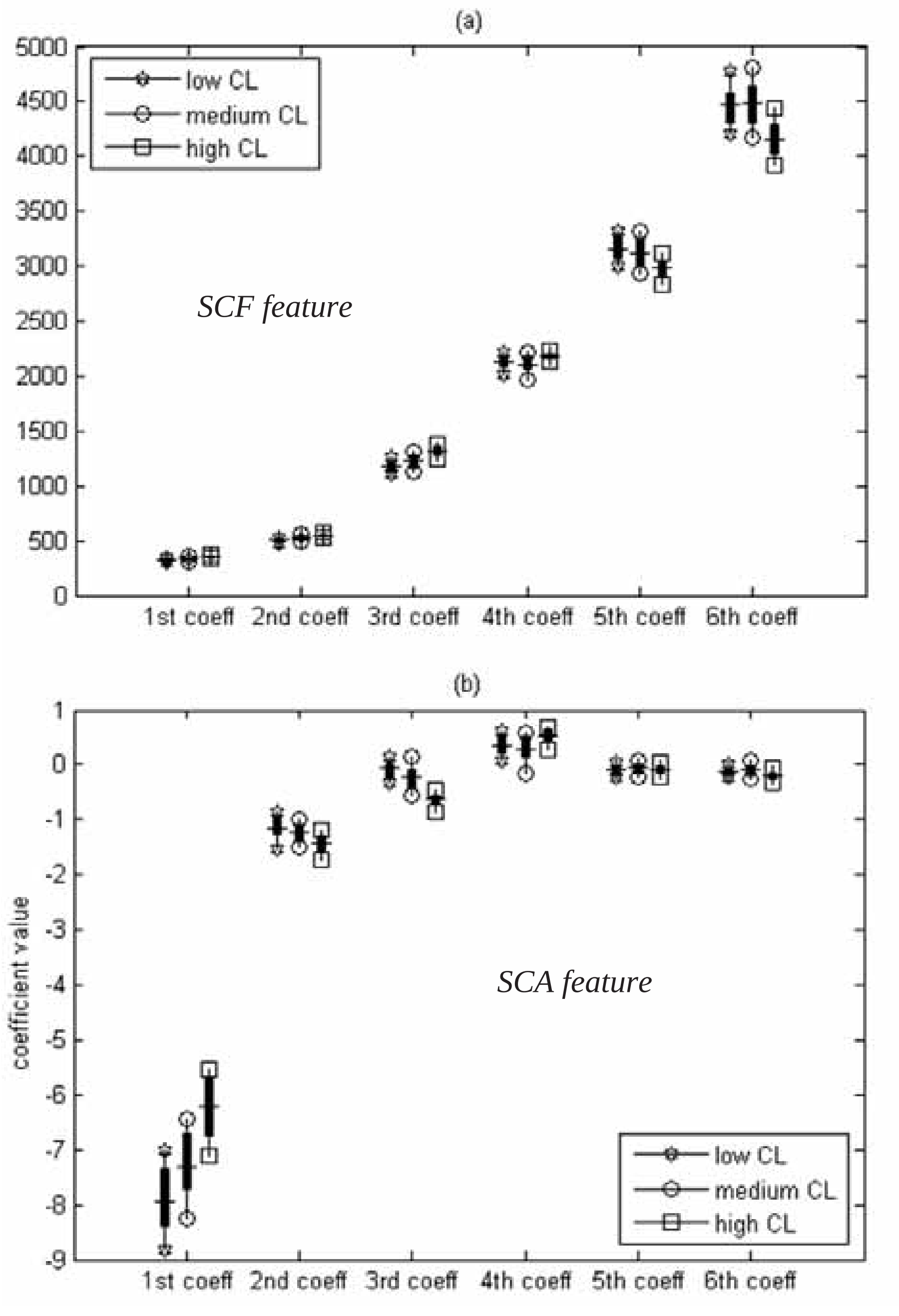


Figure 3.20: Statistical variation of the six coefficients of (a) SCF and (b) SCA over the three levels of CL speech of the word ‘blue’. The thick bar extends from the 15th to the 85th percentile and the thin bar extends from the 5th to the 95th percentile. The middle strip indicates the median.

3.5.4 Performance of the spectral centroid features

The spectral centroid features used in the classification experiments in this section are computed using a bank of twelve Gabor filters which are equally spaced in the mel scale. The number of filters was chosen to be 12 so that the SCF and SCA feature vectors have the same dimensions as the other cepstral-based feature vectors e.g. MFCC, SMFCC and FMFCC. This will make the comparison of the effectiveness of these features more straightforward.

The classification accuracies obtained when the spectral centroid features are used individually as the front-end of the classification systems are reported in Table 3.18. The classification accuracies of the fusion system of these two features, and of each of these two features with the MFCC features, with and without concatenating with the shifted delta feature (SDF), are also presented in this table.

Table 3.18: The accuracies using individual SCF, SCA, and fusion between SCF and SCA, SCF and MFCC, SCA and MFCC.

Feature	Corpus	Accuracy (%)	
		Without SDF	With SDF
SCF	Stroop test Reading and Comprehension	48.0 44.4	73.7 57.0
SCA	Stroop test Reading and Comprehension	54.4 34.1	80.9 48.9
Fusion SCF and SCA	Stroop test Reading and Comprehension	59 47.4	82.7 58.5
Fusion SCF and MFCC	Stroop test Reading and Comprehension	54.1 49.8	76.3 63.0
Fusion SCA and MFCC	Stroop test Reading and Comprehension	56.8 42.1	82.4 62.2

It can be observed from Tables 3.18 and 3.12 that both spectral centroid features yield comparable classification accuracy for the system compared to the traditional MFCC feature. Furthermore, the fusion of SCF-based and SCA-based systems produced higher accuracy compared to the systems based on individual SCF or SCA features. This indicates that the cognitive load information contained in the spectral centroid frequency and spectral centroid amplitude features is complementary, as discussed in Section 3.5.2. It can also be seen from these tables that the fusion of either SCF-based or SCA-based systems with an MFCC-based system consistently outperforms the MFCC based system. This implies that the information contained in the spectral centroid features is complementary to that contained in the MFCC feature. Fusion of the SCF-based and

SCA-based systems with the MFCC-based system reduces the relative error rate by 8.9% and 31.5% respectively when compared to the MFCC-based system on the Stroop test corpus. In addition, the corresponding relative error rate reductions obtained on the Reading and Comprehension corpus are 5.9% and 3.8%.

3.6 Comparison and discussion of performance of different speech features

Table 3.19: Summary of accuracies of different speech features, with and without using the shifted delta feature (SDF).

Category	Feature	Accuracy (%)			
		Stroop test		Reading and Comprehension	
		No SDF	With SDF	No SDF	With SDF
Source-based	Pitch (F_0)	32.8	52.2	33.3	37.0
	Intensity	32.8	56.9	34.1	41.5
	Source MFCC (SMFCC)	38.2	58.9	34.2	38.5
Filter-based	Formant frequencies (FF)	57.4	75.6	31.9	45.9
	Filter MFCC (FMFCC)	41.1	70.4	36.2	47.4
Combined	**MFCC**	**45.6**	**74.3**	**40.0**	**60.7**
	Spectral slope and intercept (SSI)	40.4	63.7	41.5	46.7
	Group delay (GD)	40.7	72.0	43.3	52.5
	Frequency modulation (FM)	45.0	51.1	36.3	40.7
	Spectral centroid frequency (SCF)	**48.0**	**73.7**	**44.4**	**57.0**
	Spectral centroid amplitude (SCA)	**54.4**	**80.9**	**34.1**	**48.9**
Fusion	Pitch and formant frequency	60.3	75.6	37.0	46.7
	Intensity and formant frequency	64.1	78.2	38.5	51.9
	Pitch and Intensity	39.1	68.6	36.3	44.4
	SMFCC and FMFCC	43.2	76.9	39.1	52.6
	MFCC and group delay	46.5	76.3	45.2	62.6
	MFCC and frequency modulation	49.1	75.0	43.2	63.0
	SCF and MFCC	54.1	76.3	49.8	63.0
	SCA and MFCC	56.8	82.4	42.1	62.2
	SCF and SCA	59.0	82.7	47.4	58.5

The classification accuracies of all the speech features used in this chapter are summarized in Table 3.19 for ease of comparison. Many interesting trends can be observed from this table.

Both source-based and filter-based features are useful for classifying cognitive load levels. This means that both the voice source and the vocal tract filter are important in characterizing cognitive load variation. Therefore, an effective front-end of the classification system needs to utilize both source-based and filter-based features. However, it can be noted that filter-based features e.g. formant frequencies (FF) and filter Mel frequency cepstral coefficients (FMFCC) were superior to the source-based features e.g. the pitch (F_0), intensity and source Mel frequency cepstral coefficients (SMFCC). The better performance of the filter-based features when compared to the source-based features suggests that in the source-filter model of human speech production system, the filter is more important than the source in characterizing variations of the load level.

The combined speech features e.g. MFCC, SCF and SCA, perform better than the source-based features e.g. SMFCC, or filter-based features e.g. FMFCC for the classification. This comparison is straightforward as these features are computed in the frequency domain and have the same dimension. This is probably because the combined features capture the information comprehensively as they are estimated from the speech spectrum which includes information from both the source and the filter. Conversely, the features estimated from either source or filter only may lack comprehensiveness in capturing CL information. This again suggests that cognitive load is characterized by both the source and filter components of the human speech production system.

It was also observed that the concatenation of a feature with its shifted delta feature (SDF) consistently improves the performance of the system. This suggests that temporal information of the speech features is very important for classifying the load level. Because of this, all the research in the remainder of this thesis is performed with the combination of the speech features and their shifted delta features.

It can be seen that fusion systems based multiple features consistently produced higher classification accuracy than the single feature systems. This is probably because different speech features characterize different aspects of speech. Hence the cognitive load information contained in them can be complementary. Consequently, incorporating the information from different speech features can increase the amount of the cognitive load information to the system and improve its performance.

It should be noted that there is a consistent trend in classification accuracies of different speech features on the two corpora. That is any speech features that perform well on the Stroop test corpus also perform well on the Reading and Comprehension corpus even though these two corpora are collected under very different conditions. This consistency indicates that the experiments conducted are independent of variations in databases. In addition, when the shifted delta features are used, the classification accuracies obtained on the Stroop test corpus are significantly higher than the corresponding accuracies obtained on the Reading and Comprehension corpus. This is probably due to the greater variability of speech features in the Reading and Comprehension corpus due to the higher level of phonetic variability when compared to the Stroop test corpus.

Among the features used in this investigation, MFCC, SCF and SCA provide the highest accuracies for the system when the speech features and their shifted delta features are used, presented in bold in Table 3.19. These features are computed from the spectrum of speech and are henceforth referred to as spectral features in this thesis. Due to their outperformance, they are chosen for further study in the remainder of this thesis.

3.7 Summary

This chapter initially carried out a human listening test on a subset of the Stroop test corpus. The high accuracy of this test implies that cognitive load specific patterns exist in speech features and hence it is possible for an automatic speech-based cognitive load classification system to classify the load level of a speech segment based on these patterns. Furthermore, it was found that the breath pattern, speech rate, the insertion of filler sounds e.g. ‘uh’ and ‘ah’, and the intonation of the utterance are the most important speech cues used by humans to identify cognitive load levels. The usefulness of the intonation in this test supports the effectiveness of the shifted delta feature of pitch for an automatic speech-based cognitive load classification system in this chapter and in previous studies [3, 5].

The results of the investigation of the effectiveness of source-based and filter-based speech features indicated that although the filter-based features are more effective, both of these features are effective for cognitive load classification. This suggests that in the source filter model of human speech production, the filter is more important than the source in characterizing the variation of the load level. Nevertheless, an effective classification system should utilize both types of the features.

The use of the spectral centroid features (SCF and SCA) for CL classification was proposed. Fusion of these two features or either each of them with the MFCC feature consistently provided higher accuracy than systems based on individual features. This indicates that cognitive load information is contained in the spectral centroid features. These features are complementary to each other and are complementary to the MFCC feature. Fusion of the SCF-based and SCA-based systems to the MFCC-based system reduced the relative error rate by 8.9% and 31.5% respectively when compared to the MFCC-based system on the Stroop test corpus. The corresponding relative error rate reductions obtained on the Reading and Comprehension corpus are 5.9% and 3.8%.

The spectral features, namely MFCC, spectral centroid frequency and spectral centroid amplitude have been shown to provide the highest classification accuracies for the system compared to all of the other features used. This motivates the use of these features for the study of cognitive load classification in the remainder of this thesis. In addition, the temporal information of the speech features captured by the shifted delta features was shown to be very important for classifying the load level. Hence, the speech features are always combined with their shifted delta features when used for classification experiments in the rest of this thesis.

Chapter 4: Multi-band approach for cognitive load classification

4.1 Introduction

The study on cognitive load classification presented in Chapter 3 was conducted under clean condition where the speech used in both the training and testing phases is clean, i.e. not corrupted by noise. However, in many real-life applications, cognitive load has to be estimated in practical scenarios where speech is recorded in noisy environments such as in cars, over telephone channels or in air traffic control rooms. Such noisy conditions would cause a mismatch between the feature distribution of the training speech and that of the test speech which consequently can degrade the performance of the cognitive load classification system significantly. Techniques to reduce the effect of noise and improve the performance of the system under noisy conditions are therefore necessary to make it usable in real-life applications.

Speech features from individual subbands are corrupted to varying levels depending upon the type of noise and therefore have different levels of reliability. In this thesis, the approach taken to develop the cognitive load classification system based on features that are extracted from the whole speech spectrum e.g. MFCC, spectral centroid frequency and spectral centroid amplitude is referred to as the full-band approach. This approach may not be the most effective as it disregards the unequal levels of reliability of speech features in different frequency bands. Unlike systems based on the full-band approach, systems based on the multi-band approach, which will be discussed in this chapter, utilize the speech features extracted from different subbands independently. The classification system based on this approach can therefore reduce the effect of noise by de-emphasizing the contribution of speech features from the subbands that are less reliable. It has been shown in studies of speech recognition and speaker recognition that the multi-band approach provides a higher performance than the full-band approach for systems in noisy conditions [96-97].

Furthermore, spectral features such as MFCC and spectral centroid features used in Chapter 3 were extracted through the mel filterbank. This filterbank was chosen because it is commonly used in other classification systems such as speech recognition, speaker recognition and emotion recognition [95, 98-99]. However, in speaker recognition it has been shown that speaker specific information is not distributed according to the mel frequency scale and that the mel filterbank is not the optimal filterbank for speaker recognition [100]. Similarly in cognitive load classification, the load information may not be distributed according to the mel frequency scale. This implies that the amounts of cognitive load information contained in different mel subbands (the subbands have the same widths in mel scale) may be different as will be presented in Section 4.2.2.2. Therefore even in clean conditions, it may be possible to improve the performance of the cognitive load classification system based on a multi-band approach by emphasizing the contribution of speech features in the subbands containing more cognitive load information. However, this cannot be achieved using the full-band approach as the speech features used in this approach are extracted from the whole spectrum.

The main aim of this chapter is to investigate the effectiveness of the multi-band approach and compare it with the full-band approach for a cognitive load classification system. This chapter initially analyzes the spectral distribution of cognitive load information in different mel subbands through classification experiments. It then studies the effectiveness of the multi-band approach for classification under clean conditions. This is followed by an investigation of the reliability of speech features for classification in different subbands under noisy conditions. This chapter then looks at the effectiveness of different weighting schemes for a multi-band cognitive load classification system. Finally, it investigates the effectiveness of the multi-band approach and compares it with the full-band approach for cognitive load classification under noisy conditions.

4.2 Motivation for using a multi-band approach

Two of the most important aspects affecting the effectiveness of speech features in a subband under noisy conditions are how severely these speech features are corrupted by noise and how much cognitive load (CL) information is contained in them.

4.2.1 Advantage of multi-band over full-band approach

4.2.1.1 Effect of band-limited noise

An illustration of the impact of band-limited noise, i.e. noise limited to a frequency region smaller than the bandwidth of speech, on the full-band and subband speech features is shown in Figure 4.1. This figure plots the mean square error of the cepstral coefficients computed on clean speech and noisy speech using both the full-band and multi-band approaches. The noisy speech in this example is obtained by contaminating clean speech with a 100 Hz sinusoid. This sinusoid is used to simulate a band-limited noise as its power only exists at 100 Hz. For the full-band approach, six cepstral coefficients were computed from speech with a bandwidth of 8 kHz. For the multi-band approach, three cepstral coefficients were extracted from each of the two mel subbands of speech spectrum. It can be seen from Figure 4.1 that in the full-band approach all the cepstral coefficients are affected by noise. This is indicated by the large mean square errors. For the multi-band approach, the three cepstral coefficients extracted from the second subband are almost unaffected by the band-limited noise, as indicated by the very small mean square errors. Thus under the effect of band-limited noise, the multi-band approach can be more effective than the full-band approach as it can emphasize (by assigning more weight to) the three cepstral coefficients from the second subband and de-emphasize (by assigning less weight to) three cepstral coefficients from the first subband.

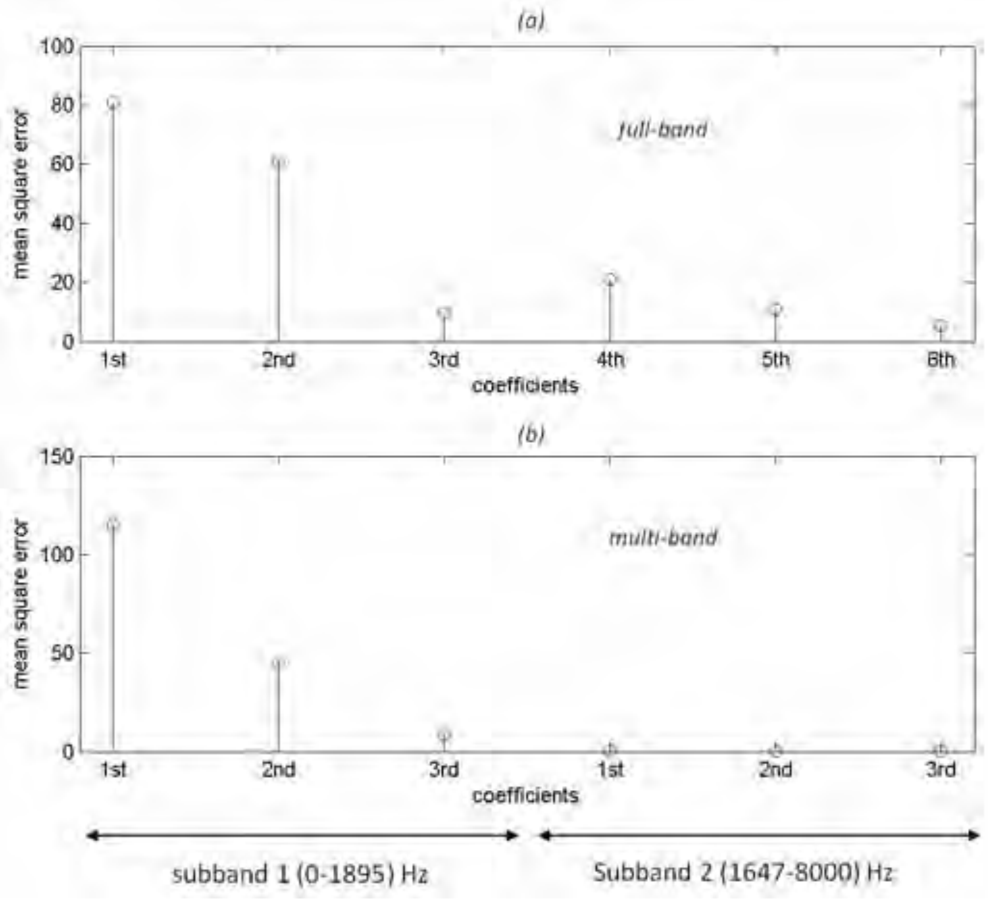


Figure 4.1: Mean square error of cepstral coefficients of clean and noisy speech computed based on (a) full-band and (b) multi-band approaches.

4.2.1.2 Effect of different types of noise

Unlike band-limited noise, other types of noise are distributed over the whole bandwidth of the speech signal. Under the effect of these noise types, the severity to which speech features in a subband are affected depends on the amount of noise present in that subband. Table 4.1 shows the distribution of spectral power of different noise types from the NOISEX-92 dataset in different subbands. These subbands were obtained by splitting the 8 kHz bandwidth of the speech signal into two subbands whose bandwidths are equal in the mel scale.

Table 4.1: The distribution of noise power.

Noise type	The distribution of noise power in different subbands (%)	
	Subband 1 (0-1895) Hz	Subband 2 (1647-8000) Hz
Pink	49.9	54.0
White	24.0	89.0
Leopard	88.8	12.4
Factory	58.3	45.1
F16	56.3	46.4
Buccaneer	50.7	54.4
Babble	81.8	21.5

It can be observed from Table 4.1 that the distributions of noise power in the two subbands are very different. This suggests that speech features in different subbands will be corrupted by noise to different levels. For example, for noise types such as leopard, factory, F16 and babble, the amount of noise in the first subband is larger than that in the second subband. Therefore, under these noise conditions, the speech features in the first subband will be affected more. For the noise types such as pink, white and buccaneer, the amount of noise in the second subband is larger than those in the first subband. Hence under these noise conditions, the speech features in the second subband will be affected more. This in turn suggests that speech features in different subbands may have different levels of reliability. By emphasizing the speech features in a reliable subband and de-emphasizing features in a less reliable one, the performance of a classification system based on the multi-band approach can be improved.

4.2.2 Variation of CL information in different subbands

The effectiveness of the spectral features does not reveal any information about how cognitive load information is distributed in different subbands. Determining how cognitive load information is distributed can be useful in improving the performance of a classification system based on a multi-band approach. This can be done by emphasizing speech features in the subbands containing large amounts of cognitive load (CL) information. From the perspective of classification, the amount of CL information contained in a subband is proportional to the performance of the system using speech features in that subband. That is, speech features extracted from a subband containing a larger amount of cognitive load information should produce a classification system with a higher accuracy. This section investigates the distribution of CL information in different subbands by carrying out classification experiments using subband speech features. The obtained accuracies are then used as a measure of the amount of cognitive load information contained in individual subbands.

4.2.2.1 Subband based feature extraction

The cepstral coefficients are used in this chapter to perform CL classification. In order to compute the subband cepstral coefficients, a mel filterbank consisting of twenty four filters is split into a number of sub-filterbanks. Each sub-filterbank has the same number of consecutive filters from the original filterbank and two consecutive sub-filterbanks have no filters in common. The sub-filterbanks obtained cover the frequency regions whose widths are equal in the mel frequency scale. The cepstral coefficients of the individual subbands are then obtained as the first few discrete cosine transform coefficients of the log energies of the filter outputs in corresponding subbands. Two approaches used to split the bandwidth of the speech signal into subbands employed in this thesis are the two subband (2-band) and three subband (3-band) approaches. An illustration of subband feature extraction for the 2-band approach is shown in Figure 4.2.

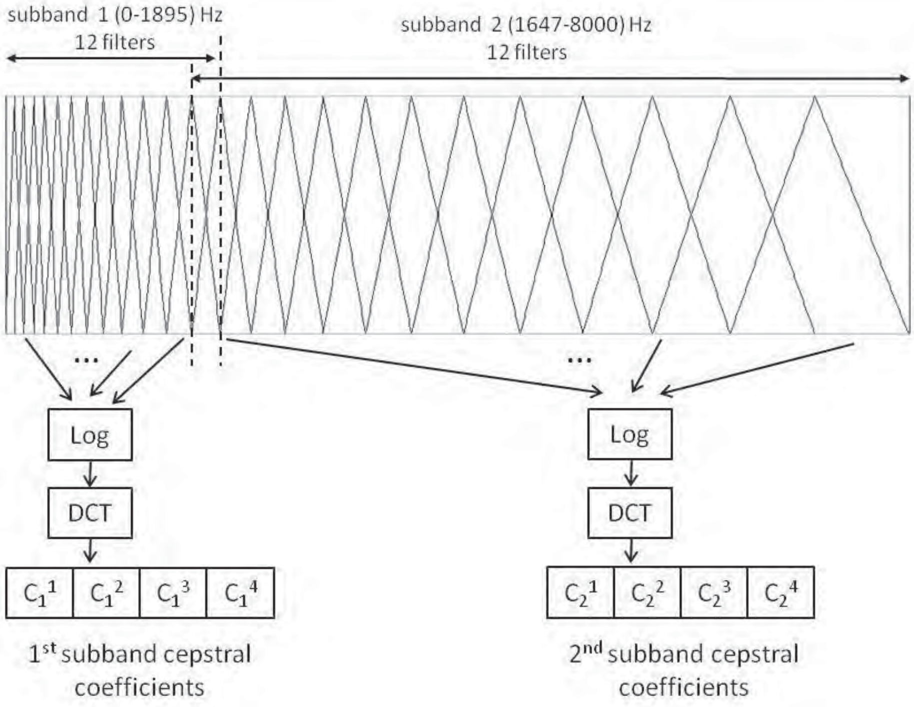

subband 1 (0-1895) Hz
12 filters
subband 2 (1647-8000) Hz
12 filters
...
...
Log
DCT
Log
DCT
C_1^1 C_1^2 C_1^3 C_1^4
C_2^1 C_2^2 C_2^3 C_2^4
1st subband cepstral coefficients
2nd subband cepstral coefficients

Figure 4.2: Extracting cepstral coefficients for the 2-band approach.

The frequency regions of the subbands for the 2-band and 3-band approaches are:

2-band approach: (0-1895), (1647-8000) Hz

3-band approach: (0-1034), (868-3184), (2811-8000) Hz

In the full-band approach, the discrete cosine transform (DCT) was applied to the log energies of the output of all twenty four filters and the first twelve DCT coefficients were used as the cepstral coefficients.

The number of filters and cepstral coefficients for each subband of the 2-band and 3-band approaches as well as the full-band approach are shown in Table 4.2.

Table 4.2: The number of filters and cepstral coefficients of multi-band and full-band approaches.

	Multi-band		Full-band
	2-band	3-band	
Number of filters in each subband	12	8	24
Number of cepstral coefficients in each subband	4	3	12

4.2.2.2 Distribution of CL information in different mel subbands

The accuracies obtained from classification experiments using the subband cepstral coefficients of the 2-band and 3-band approaches performed on the Stroop test corpus are presented in Figure 4.3 as a function of frequency.

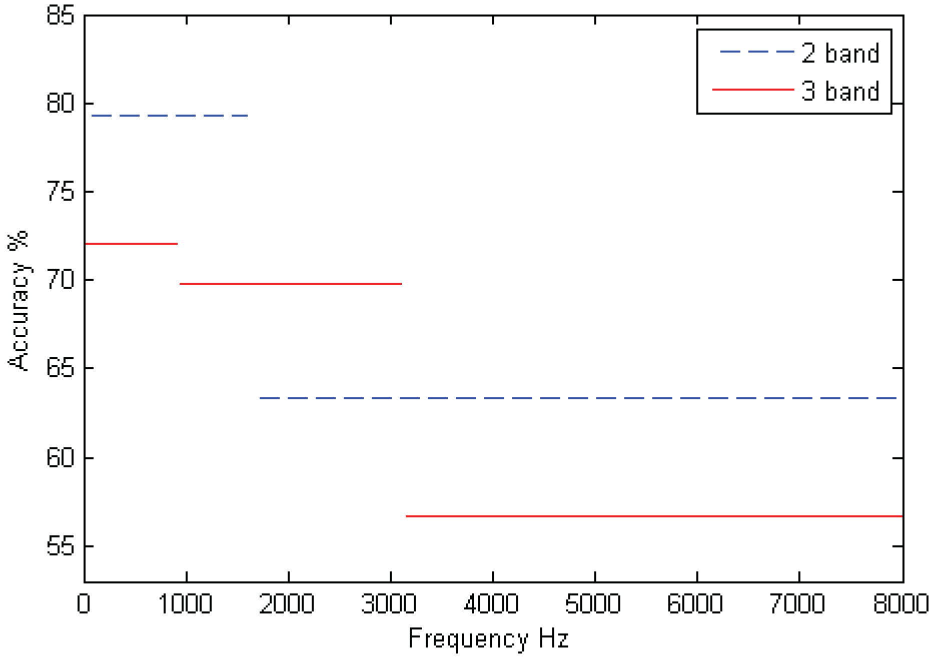


Figure 4.3: Classification accuracy of the subband features, for the clean speech of the Stroop test corpus.

The results in Figure 4.3 indicate that the speech features in the low frequency subbands make the classification system significantly more accurate than those in the high frequency subbands for both 2-band and 3-band approaches. This suggests that the low frequency subbands contain a significantly larger amount of cognitive load information than the high frequency subbands. Therefore, the multi-band approach can be expected to be more effective than the full-band approach for classifying the load levels, even in clean conditions, because it can emphasize the speech features in the low frequency bands in order to improve the performance of the system.

4.3 Multi-band classification system

4.3.1 Overview of multi-band system

The cognitive load classification system based on the multi-band approach, henceforth called a multi-band system, utilizes the speech features computed from different subbands independently. The number of subbands used is one of the important parameters of the system. While a larger number of subbands may allow more flexibility in isolating the effects of band-limited noise, it would also result in a narrower bandwidth for each subband. This will decrease the amount of cognitive load information contained in each subband, which in turn reduces the cognitive load discrimination ability of the subband speech features. For speech recognition, it has been shown that a larger number of subbands produces lower accuracy than a smaller number of subbands [101]. Thus in this chapter, the multi-band system is developed based on two subband (2-band) approach. Furthermore, the performance of the three subband (3-band) multi-band system is presented in the last section of the chapter (Section 4.5.4) to consolidate the effectiveness of the multi-band approach.

There are two methods that can be used to develop a classification system based on the multi-band approach, namely likelihood combination and feature combination [97].

4.3.1.1 Likelihood combination

In the likelihood combination method the cepstral coefficients in each subband are used in the training phase to estimate a subband cognitive load statistical model. During the testing phase, the likelihood score is estimated independently for each individual band, as shown in Figure 4.4a. As the subband spectrum has less variation than the full-band spectrum, we need fewer cepstral coefficients to describe the subband spectrum, when compared to the full-band spectrum. The statistical models of subband cepstral coefficients therefore can be modeled more effectively than those of the full-band cepstral coefficients.

From these subband likelihood scores, the overall likelihood score of the multi-band system is obtained as follows

$$LL = \sum_{i=1}^{N} \omega_i LL_i \tag{4.1}$$

where *LL* is the log likelihood score of the multi-band system, LL_i and ω_i are the log likelihood score and the weighting coefficient of the i^{th} subband respectively. $\omega_i \in [0, 1]$ and satisfies $\sum_{i=1}^{N} \omega_i = 1$. *N* is the number of subbands in the system. In this scheme, large weighting coefficients are used to emphasize the recognition results from reliable subbands and small weighting coefficients are used to de-emphasize the recognition results from less reliable subbands in order to improve the performance of the classification system.

Weighting schemes:

- **Accuracy weighting scheme**: this weighting scheme emphasizes the classification result of the subband that produces higher performance for the CL classification system. The weighting coefficient of the i^{th} subband is determined as the normalized classification accuracy of that subband and is expressed as:

$$\omega_i = \frac{Accuracy_i}{\sum_{i=1}^{N} Accuracy_i} \tag{4.2}$$

  where $Accuracy_i$ is the accuracy of system based on the cepstral coefficients of the i^{th} subband, computed from clean speech. The effectiveness of the accuracy weighting scheme will be investigated in this study for both noisy and clean conditions.

- **Signal to noise ratio (SNR) weighting scheme**: this weighting scheme emphasizes the classification results of subbands that have a higher SNR. The weighting coefficient of the i^{th} subband is determined as the normalized signal to noise ratio of that subband that is expressed as:

$$\omega_i = \frac{SNR_i}{\sum_{i=1}^{N} SNR_i} \tag{4.3}$$

where SNR_i is the signal to noise ratio of the i^{th} subband. SNR_i is computed as the ratio of the speech power to the noise power of the i^{th} subband. The noise power and speech power are estimated as the sum of noise power density and speech power density respectively. The noise power density is estimated as the mean of the power densities in non-speech frames and the speech power density is estimated by subtracting the estimated noise power density from the noisy speech power density. The effectiveness of the signal to noise ratio weighting scheme will be investigated in noisy conditions.

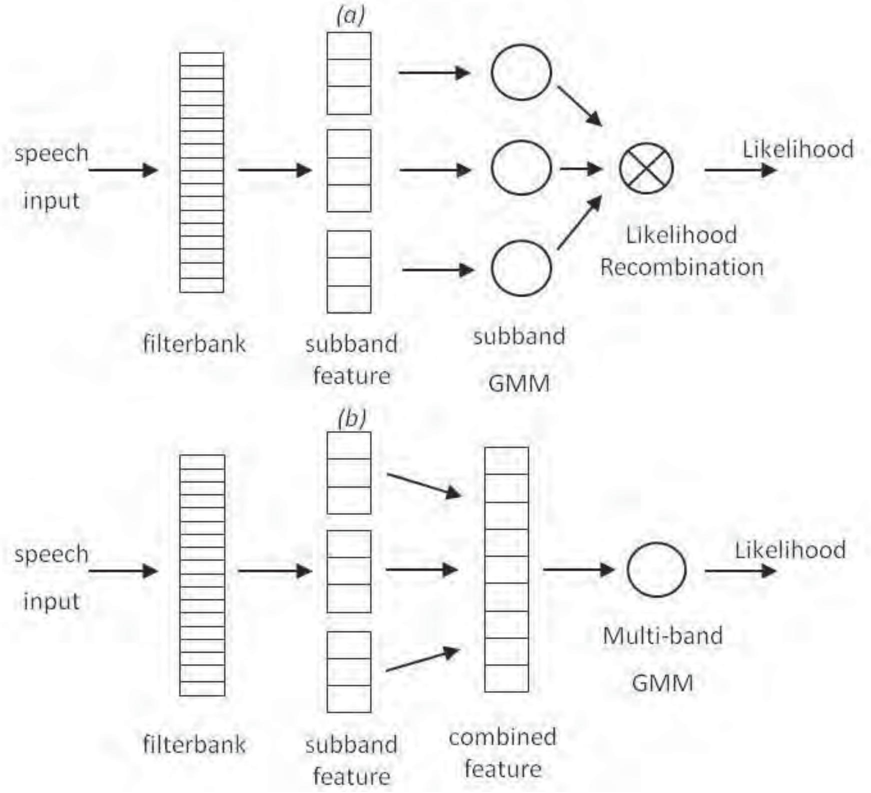


Figure 4.4: Multi-band CL classification system based on (a) likelihood combination, and (b) feature combination [97].

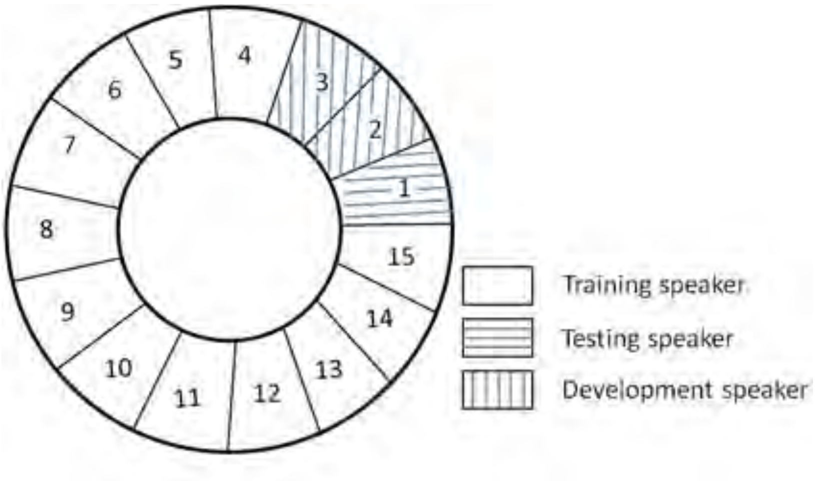
Non-weighting scheme: In this scheme, the weighting coefficients of all subbands are equal (no relative weighting). The accuracy of the system based on a non-weighting scheme is used as the reference for comparison with the system based on the accuracy and SNR weighting schemes.

4.3.1.2 Feature combination

Unlike the likelihood combination method, the feature combination method does not assign weighting coefficients to different subbands. Instead, the acoustic features extracted from individual subbands are concatenated into a single vector which is then used as the input to the classifier, as illustrated in Figure 4.4b. Under the effect of band-limited noise, the feature combination method is expected to be more effective than the full-band approach as it is able to isolate the effect of noise in a few feature components. In the feature combination method, the statistical model can utilize joint cognitive load information in adjacent subbands as it models the combined features in all subbands. In this sense, the feature combination method is more advanced than the likelihood combination method, where the joint information is not exploited by the statistical models, as they are modeled independently in different subbands.

4.3.2 Classification experiment setup for multi-band approach

The classification system used in this chapter to evaluate the performance of the speech features was described in Section 3.3.1. In order to estimate the weighting coefficients for the likelihood combination approach, the Stroop test corpus is split into three datasets, namely training, testing and development datasets. The weighting coefficients are estimated from the development dataset. As the weighting scheme is not applied for the feature combination method and full-band approaches, the development dataset is not used in the classification experiments based on these approaches. Among the fifteen speakers in the Stroop test corpus, data from twelve speakers was used as the training dataset, data from two speakers was used as the development dataset and data from one speaker was used as the testing dataset. This is illustrated in Figure 4.5. Cross fold validation is performed i.e. each experiment was performed fifteen times with different speakers used as the testing and development speakers each time. The overall accuracy of the system is then obtained as the average of the accuracies for each fold.



The classification experiment is repeated 15 times,
rotating the testing and development speakers.

Figure 4.5: Allocation of training, testing, and development dataset.

The training speech used for the experiments carried out in noisy conditions is clean and only the testing speech is noisy. This is in line with practical scenarios as the training speech can be recorded in the laboratory which is almost unaffected by noise. Noisy speech is obtained by adding noise from the NOISEX-92 database to clean speech at five SNR levels: 0, 5, 10, 15 and 20 dB. The NOISEX-92 database was created by the Speech Research Group at Carnegie Mellon University [102]. Amongst the fifteen noises present in this database, a subset of seven common noises were used in this thesis: babble, pink, white, leopard, factory, F16, and buccaneer. These noises were resampled from 19.98 kHz to 16 kHz for compatibility with the Stroop test corpus speech signals. Each set of experiments in this study (for one noise and one SNR) is repeated 35 times (7 noises x 5 SNRs). Due to the large number of experiments to be carried out, this chapter was carried out only on the Stroop test corpus.

4.3.3 Estimation of weighting coefficients for likelihood combination

The weighting coefficients for different subbands of the likelihood combination are estimated for individual testing speakers independently from the corresponding development dataset.

- Accuracy weighting scheme: the weighting coefficients of this scheme are estimated using equation (4.2) where the *Accuracy*$_i$ are computed for individual testing speakers where the corresponding development dataset is used in the testing phase. Figure 4.6 plots the accuracy weighting coefficients of different subbands for different testing speakers. It can be seen from this figure that the weighting coefficients of the first subband are larger than those of the second subband for most of the testing speakers. This means that the accuracy weighting scheme will emphasize the speech features in the low frequency subband and de-emphasize the speech features in the high frequency subband.

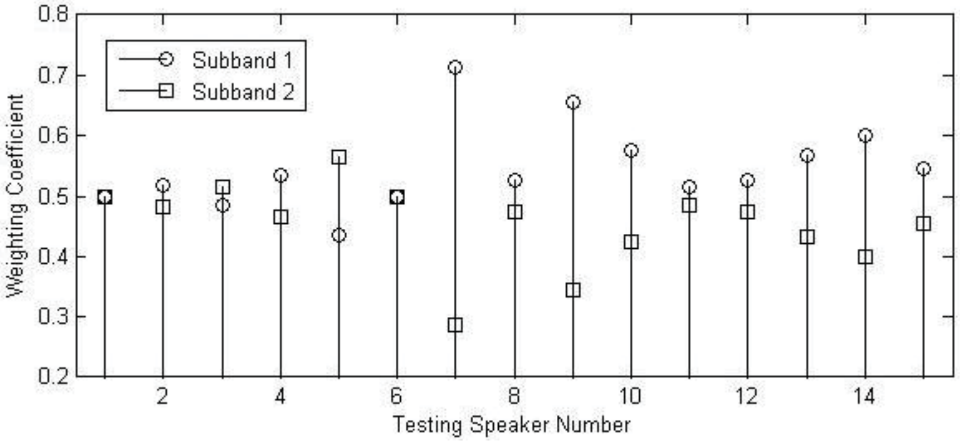


Figure 4.6: Weighting coefficients of accuracy weighting scheme.

- Signal to noise ratio (SNR) weighting scheme: the weighting coefficients of this scheme are estimated using equation (4.3) computed on the development dataset. Figure 4.7 plots the SNR weighting coefficients average across all testing speakers and all SNR levels for the two different subbands and the seven different noise types. It can be observed from this figure that according to this weighting scheme, the weighting coefficients of the first subband are significantly larger than those of the second subband for all of the noise types except for the leopard and babble noise. This means that under the effect of all noise types except leopard and babble noise, the SNR weighting scheme for the likelihood combination system will emphasize the speech features in the low frequency subband. For leopard and babble noise, the weighting coefficients of the second subband are larger than those of the first subband. This is probably due to the fact that these noises are mostly distributed in the first subband, as shown in Table 4.1. Therefore under the effect of these two noises, the system based on the signal to noise ratio weighting scheme will emphasize the speech features in the high frequency subband.

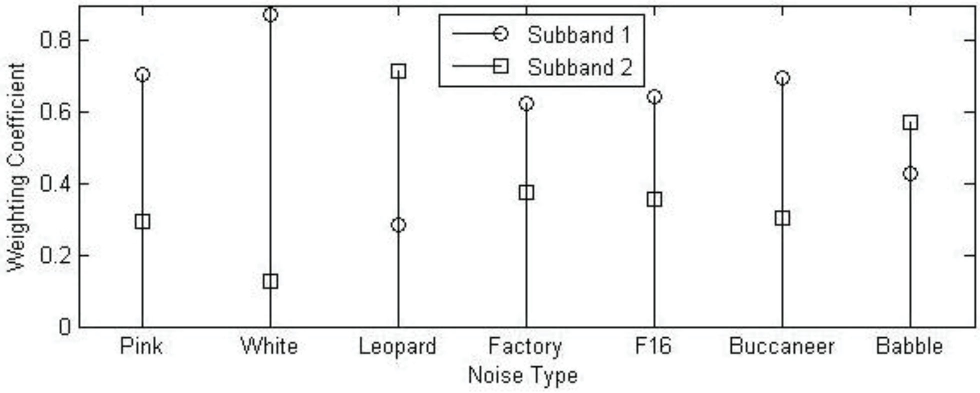


Figure 4.7: Average weighting coefficients of SNR weighting scheme.

4.4 Performance of multi-band approach in clean condition

This section investigates the effectiveness of the multi-band approach based on two subbands and compares it with that of the full-band approach under clean conditions by comparing the results of cognitive load classification experiments. For the multi-band system based on the likelihood combination method, the accuracy weighting and non-weighting schemes are used to combine classification results of individual subbands. The SNR weighting scheme is not used here as the speech used in these experiments is clean. The results of these experiments are shown in Table 4.3.

Table 4.3: Accuracy of multi-band and full-band approaches in clean condition.

System	Multi-band (2-band)				Full-band
	Likelihood combination		Feature combination		
	Non-weighting	Accuracy weighting			
Classification accuracy (%)	76.9	78.2	80.0		75.9

The results in Table 4.3 indicate that all multi-band systems yield higher classification accuracies than the full-band system. It is interesting to see that even the likelihood combination approach based on a non-weighting scheme results in better performance than the full-band approach. This might be because the statistical models of the subband features are more accurate than the full-band model due to the lower dimensionality of the subband features.

When the accuracy weighting scheme was applied, the performance of the likelihood combination system was better than that of the non-weighting scheme and hence much better than the full-band system. As found in Section 4.2.2.2, the first subband is significantly more useful than the second subband for cognitive load classification. Moreover, the accuracy weighting scheme assigns more weight to the low frequency subband and less weight to the high frequency subband. The higher performance of the accuracy weighting scheme compared to the non-weighting scheme therefore indicates that we can improve the performance of the multi-band classification system based on the likelihood combination by emphasizing speech features in the low frequency region.

Furthermore, the feature combination approach results in higher performance when compared to the full-band approach and the likelihood combination approach. The better performance of the feature combination approach compared to the likelihood combination approach might be because the statistical model in the feature combination utilizes the joint cognitive load information between adjacent subbands more effectively than the likelihood combination approach.

In terms of system performance, the multi-band systems based on likelihood combination with accuracy weighting scheme and feature combination method produced a 9.5% and 17% relative error rate reduction respectively, when compared to the traditional full-band system.

4.5 Performance of multi-band approach under noisy conditions

4.5.1 Reliability of subband speech features

As discussed in Section 4.2, the performance of the system under noisy conditions based on the likelihood combination approach can be improved by assigning weighting coefficients to speech features in different subbands according to their reliability. It is therefore necessary to evaluate the reliability of speech features in different subbands. From the point of view of a classification system, the reliability of a feature is proportional to the accuracy of the classification system using that feature. In this section subband speech features are used to perform classification experiments in noisy conditions and the obtained accuracy is used as a measure of their reliability. In particular, the cepstral coefficients computed in different subbands are used to perform the classification experiments in noisy conditions where the test speech was contaminated by various noise types with SNR varying from 0 dB to 20 dB in steps of 5 dB. The average accuracies of the subband speech features across all levels of SNR of individual noise are plotted in Figure 4.8.

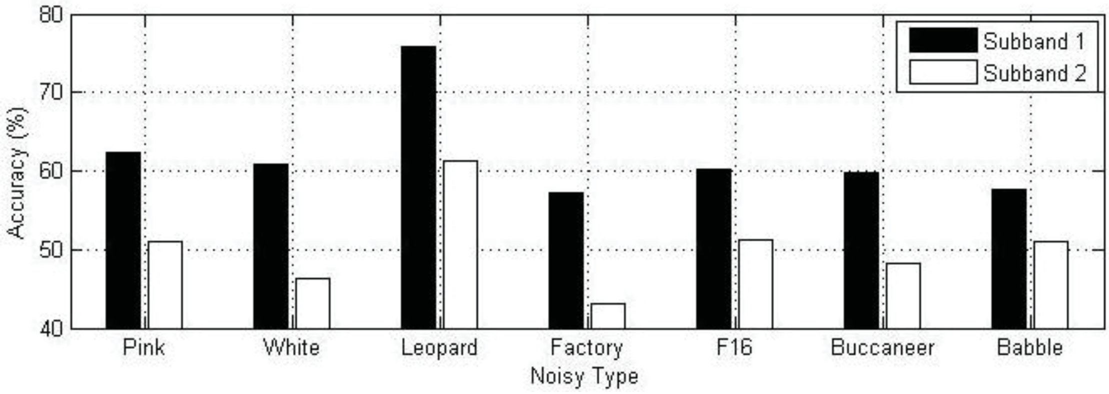


Figure 4.8: Average accuracies of subband speech features.

It can be observed from Figure 4.8 that the average accuracy of the first subband is always significantly larger than that of second subband for all of the noise types tested. This indicates that in noisy conditions the speech features in the low frequency band are significantly more reliable than those in the high frequency band with regards to cognitive load classification. The superiority of the low frequency subband compared to the high frequency subband is most probably because it contains the most significant amount of cognitive load information as described in Section 4.2.2.2.

4.5.2 Weighting schemes for likelihood combination

This section investigates the effectiveness of the accuracy weighting and SNR weighting schemes and compares them with that of the non-weighting scheme for cognitive load classification based on the likelihood combination method. It is done by carrying out the classification experiments with these weighting schemes under the effect of seven different noise types and at five levels of SNR. The accuracies obtained from these experiments are used to evaluate the effectiveness of different weighting schemes. The average classification accuracy across all SNRs obtained from these experiments for individual noise types is shown in Table 4.4.

The results in Table 4.4 show that both accuracy and SNR weighting schemes provided higher accuracy for the multi-band classification system based on likelihood combination than the non-weighting scheme. Since both accuracy and SNR weighting schemes assign a larger weight to the low frequency subband than the high frequency subband, the higher accuracies of these weighting schemes compared to the non-weighing scheme indicate that the emphasis of speech features in the low frequency subband can improve the performance of the multi-band cognitive load classification system based on likelihood combination under noisy conditions.

Table 4.4: The average accuracies of different weighting schemes.

	Average classification accuracy (%)		
Weighting scheme / Noise type	Accuracy weighting	SNR weighting	Non-weighting
Pink	60.5	62.1	60.0
White	57.9	59.9	56.9
Leopard	72.9	67.9	72.3
Factory	54.5	55.8	53.2
F16	59.9	60.3	59.6
Buccaneer	59.9	59.5	60.1
Babble	56.8	56.4	56.4
Average	**60.3**	**60.3**	**59.8**

The accuracy and SNR weighting schemes provided equal performances for the multi-band classification system based on likelihood combination. Furthermore, the performance of these two weighting schemes is higher than that of the non-weighting scheme. The accuracy weighting scheme will be used to develop the multi-band cognitive load classification system based on likelihood combination in the next section.

4.5.3 Comparison of the effectiveness of multi-band and full-band approaches

Cognitive load classification experiments in noisy conditions were carried out to compare the effectiveness of the 2-band multi-band approach for both likelihood combination and feature combination methods with that of the full-band approach. The accuracies averaged over all SNRs of these experiments for individual noise types are shown in Table 4.5.

The average classification accuracies presented in the last row of Table 4.5 show that all the multi-band approaches provided higher accuracies than the full-band approach. This suggests that the multi-band approach is more robust to noise than the full-band approach for cognitive load classification. The likelihood combination and the feature combination systems produced 3.9% and 9.9% relative error rate reduction respectively compared to the full-band system. In addition, between the two multi-band approaches, feature combination provided a higher accuracy than likelihood combination. This is consistent with the results obtained under clean conditions.

Table 4.5: The accuracies of multi-band and full-band approaches in noisy conditions.

	Average classification accuracy (%)		
System	Multi-band		
Noise type	Feature combination	Likelihood combination (accuracy weighting)	Full-band
Pink	64.0	60.5	58.0
White	62.2	57.9	54.3
Leopard	74.9	72.9	71.0
Factory	56.9	54.5	55.8
F16	61.0	59.9	57.8
Buccaneer	58.8	59.9	53.8
Babble	61.8	56.8	60.2
Average	**62.8**	**60.3**	**58.7**

4.5.4 Performance of the multi-band system based on three subbands

In order to consolidate the effectiveness of the weighting schemes and the multi-band approach, the classification experiments were performed on a three subband (3-band) multi-band system. The accuracies averaged over all SNRs are presented in Table 4.6.

Table 4.6: The average accuracies of the 3-band multi-band systems.

System / Noise type	Average accuracy (%)				
	Multi-band				Full-band
	Feature combination	Likelihood combination			
		Accuracy weighting	SNR weighting	Non-weighting	
Pink	62.9	62.4	62.4	62.0	58.0
White	59.7	61.7	60.0	61.8	54.3
Leopard	75.9	73.3	69.2	73.1	71.0
Factory	57.0	56.9	57.7	56.9	55.8
F16	60.2	57.6	56.9	57.8	57.8
Buccaneer	55.8	59.2	58.8	57.4	53.8
Babble	65.5	61.9	61.5	60.9	60.2
Average	**62.4**	**61.9**	**60.9**	**61.4**	**58.7**

The accuracies averaged across noise types presented in the last row of Table 4.6 show that all multi-band systems based on 3-band approach provide higher accuracy than the full-band system. This again indicates the effectiveness of the multi-band approach for cognitive load classification. Furthermore, the accuracy weighting scheme is more effective and the signal to noise ratio weighting scheme is less effective than the non-weighting scheme.

The average accuracies across all SNRs obtained from the experiments using the subband speech features of the three bands approach are shown in Figure 4.9. It can be seen from this figure that the second subband is more reliable than the other subbands for classifying cognitive load in noisy conditions. Furthermore, the weights of the accuracy weighting and SNR weighting schemes shown in Figure 4.10 indicate that the accuracy scheme emphasizes the second subband and the SNR scheme emphasizes the first subband. These observations further support the hypothesis that the emphasis of the speech features in a more reliable subband can improve the performance of the cognitive load classification system in noisy conditions.

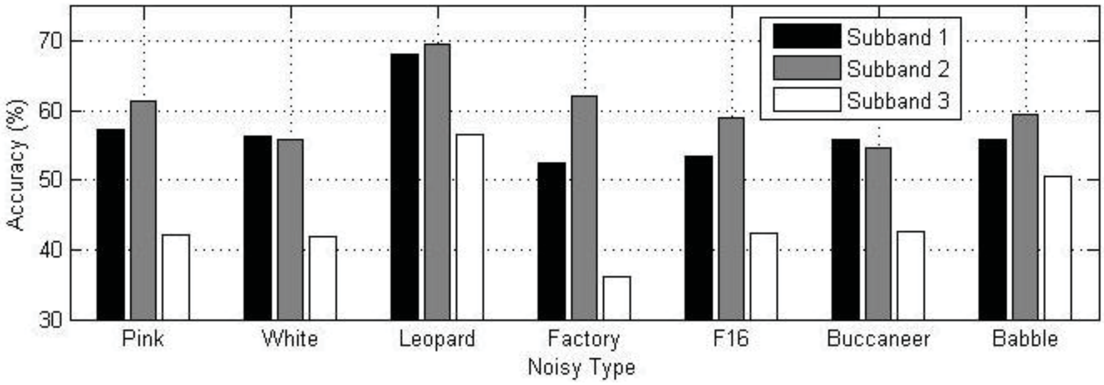


Figure 4.9: Average accuracies of subband features of 3-band approach.

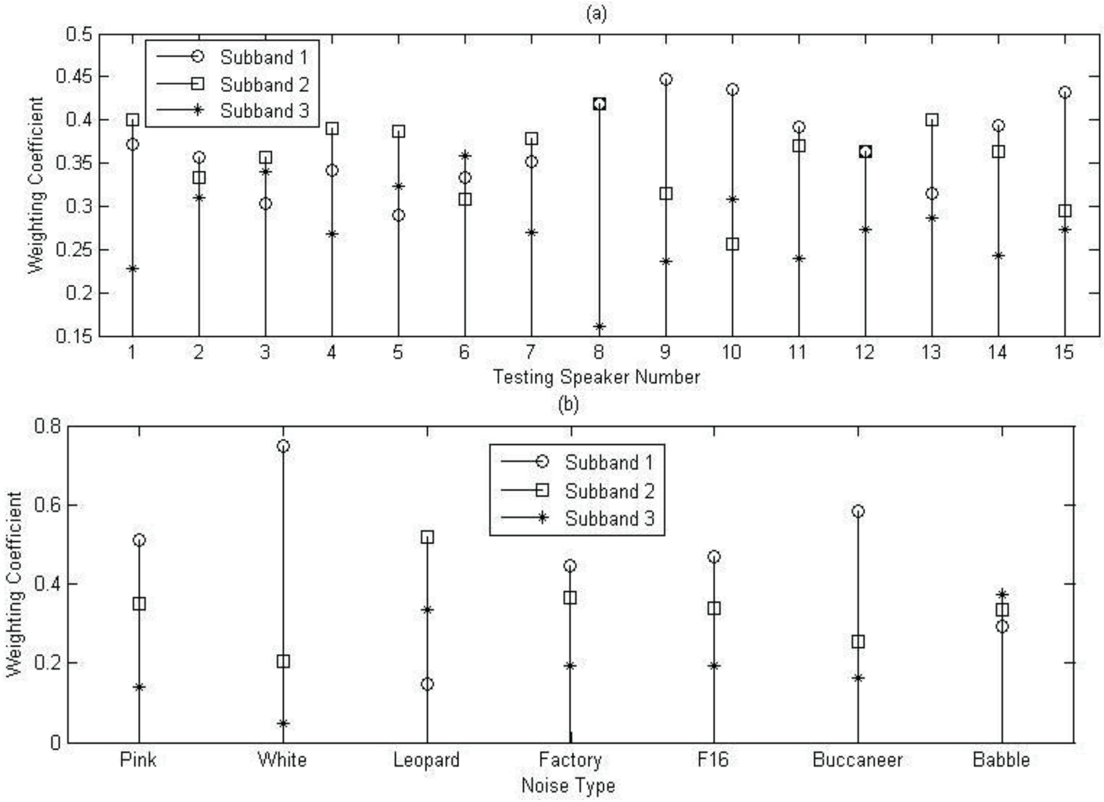


Figure 4.10: (a) Accuracy weighting coefficients and (b) SNR weighting coefficients averaged across all testing speakers and SNR levels of the 3-band approach.

4.6 Summary

This chapter found that the low frequency mel subbands are more important for cognitive load classification than the high frequency mel subbands. It then investigated the effectiveness of different weighting schemes and proposed the use of a multi-band system to emphasize speech features in the low frequency band in order to improve the performance of the system.

In clean conditions the proposed accuracy weighting scheme is found to produce higher classification accuracy when compared with the non-weighting scheme for the multi-band classification system based on likelihood combination. This indicates that assigning more weight to emphasize speech features in the low frequency band can improve the performance of the multi-band system. Furthermore, both likelihood combination and feature combination multi-band systems produce higher performance than the full-band system. In particular the likelihood combination based on accuracy weighting scheme and the feature combination systems reduce the relative error rate by 9.5% and 17% respectively compared to the traditional full-band system.

Under noisy conditions it was found that the proposed accuracy and SNR weighting schemes produce higher accuracy than the non-weighting scheme for the 2-band system. As these two weighting schemes emphasize the low frequency subband, we can conclude that the low frequency region is more important than the high frequency region for cognitive load classification, not only in clean conditions but also in noisy conditions. Furthermore, both likelihood combination and feature combination of multi-band system were found to provide higher classification performance than the traditional full-band system.

Although the effectiveness of the multiband approach was evaluated only on cepstral features, it is reasonable to expect that this approach is also effective for the classification system using other spectral features such as spectral centroid frequency and spectral centroid amplitude features. This is because when analyzing the distribution of cognitive load information contained in these features, it was also found that the low frequency subbands are more important than the high frequency subbands for cognitive load classification as will be described in Section 5.3.2.

Chapter 5: Investigation of cognitive load information distribution and filterbank design

5.1 Introduction

The study presented in Chapter 4 found that the spectral information specific to cognitive load is mainly distributed in the low frequency mel subbands. Therefore, the use of weighting schemes has been proposed to emphasize the speech features in the low frequency subband in order to improve the performance of the cognitive load classification system. Weighting schemes are not the only avenue of utilizing the cognitive load information for the classification system development. An alternative is to emphasize the spectral information in the low frequency region. In the case of the spectral features, namely cepstral coefficients, spectral centroid frequency and spectral centroid amplitude, this can be done by increasing the frequency resolution of the front-end filterbanks by allocating more filters in the low frequency region.

Furthermore, the significant difference of the amounts of spectral information contained in different mel subbands, as found in Chapter 4, implies that the mel filterbank may not be an optimal filterbank and also motivates the study of designing effective filterbanks specific for cognitive load classification by choosing the filter bands according to the distribution of cognitive load information in order to improve the performance of the system.

The design of an effective filterbank is constrained by the number of filters in the filterbank. As the number of filters in a filterbank used in front-end feature extraction corresponds to the feature dimension of the spectral features, this chapter initially investigates the effect of varying the dimension of the features on the performance of the classification system in order to determine the dimensions that produce the highest performance. The distribution of cognitive load information along the frequency bands is analyzed by quantifying the amount of spectral information contained in every uniformly divided subband. The filterbank that can effectively extract the specific spectral features for cognitive load classification is then designed by allocating the center frequencies and bandwidths of its filters according to the distribution of cognitive load information. The number of filters in the designed filterbank is chosen to produce the optimal spectral feature.

5.2 The effect of varying the feature dimension of the spectral features

5.2.1 Hypothesis

The spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) features capture the shape of the speech spectrum within individual subbands covered by the filters of the filterbank used to extract these features. The widths of these subbands depend on the number of filters of the filterbank. A small number of filters would result in very large subbands and hence the extracted spectral features may not be able to capture the details of spectral information properly. On the other hand, a large number of filters would result in a narrow bandwidth and therefore the extracted spectral features in adjacent subbands may have redundant details due to adjacent band correlation. Another drawback for a large number of filters is the increase in the number of modeling parameters in the Gaussian mixture model due to the high dimensionality of the feature vectors. In this case the training data may not be sufficient to estimate all the parameters of the model. The dimension of the spectral centroid features is equal to the number of filters used to extract them. This suggests that the spectral centroid features with large or small dimensions can degrade the performance of the classification system.

The MFCC feature captures the approximation of the spectral envelope via the filterbank which consists of 20 filters. The dimension of this feature is N ($1 \leq N \leq 20$), which is equal to the number of discrete cosine transform (DCT) coefficients where the DCT is applied to the log of the spectral energies at the output of the filters.

5.2.2 System performance with different feature dimensions

This section investigates the effect of varying the feature dimension of the spectral features (MFCC, SCF and SCA) extracted using the mel filterbanks to the performance of the cognitive load classification system. The purpose of this investigation is to determine the dimensions that produce the best performance. This is done empirically by varying the dimension of these features from 2 to 20 in steps of 2. The dimensions of SCF and SCA are varied by changing the number of filters used to extract them. The dimension of the MFCC is varied by changing the number of DCT coefficients used to represent it. These features were used to perform the classification experiment. The dimensionality producing the best performances is determined from the classification accuracies. The accuracies of the classification system using the individual MFCC, SCF and SCA features in the front-end are showed in Figure 5.1. The accuracies obtained by fusing the classification results of the SCF-based and SCA-based systems were also determined and showed in this figure.

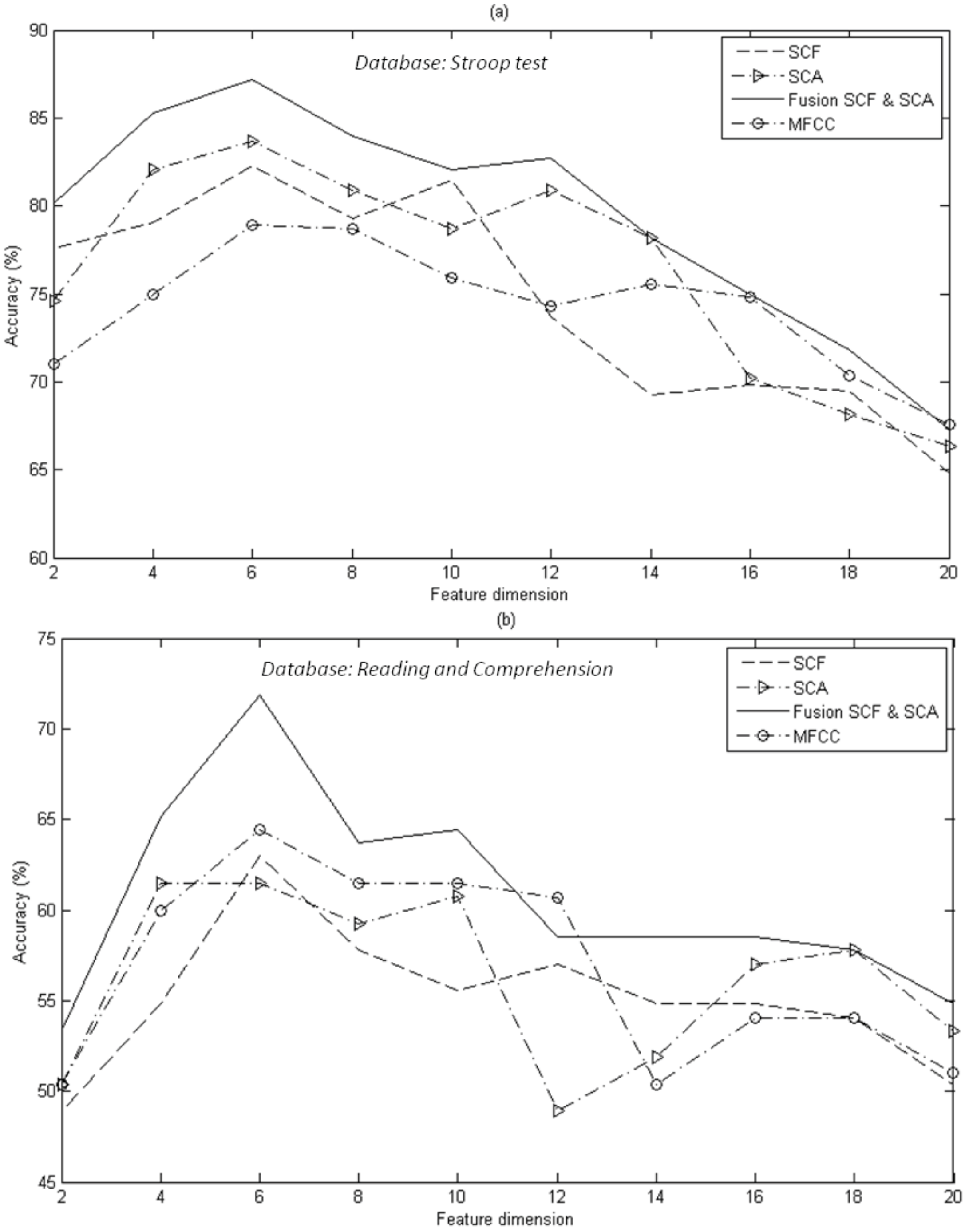


Figure 5.1: Performance of the spectral features with various dimensions evaluated on (a) Stroop test, and (b) Reading and Comprehension corpora.

The results in Figure 5.1 show that all the accuracy curves initially increase and then decrease as the feature dimension increases. This is in line with the hypothesis that the spectral features with large or small dimensions degrade the performance of the system. Interestingly, all the curves reach their peaks at the feature dimension of six. This indicates that six dimensional spectral features produce the highest performance.

5.2.3 Evaluation of the correlation of SCF and SCA

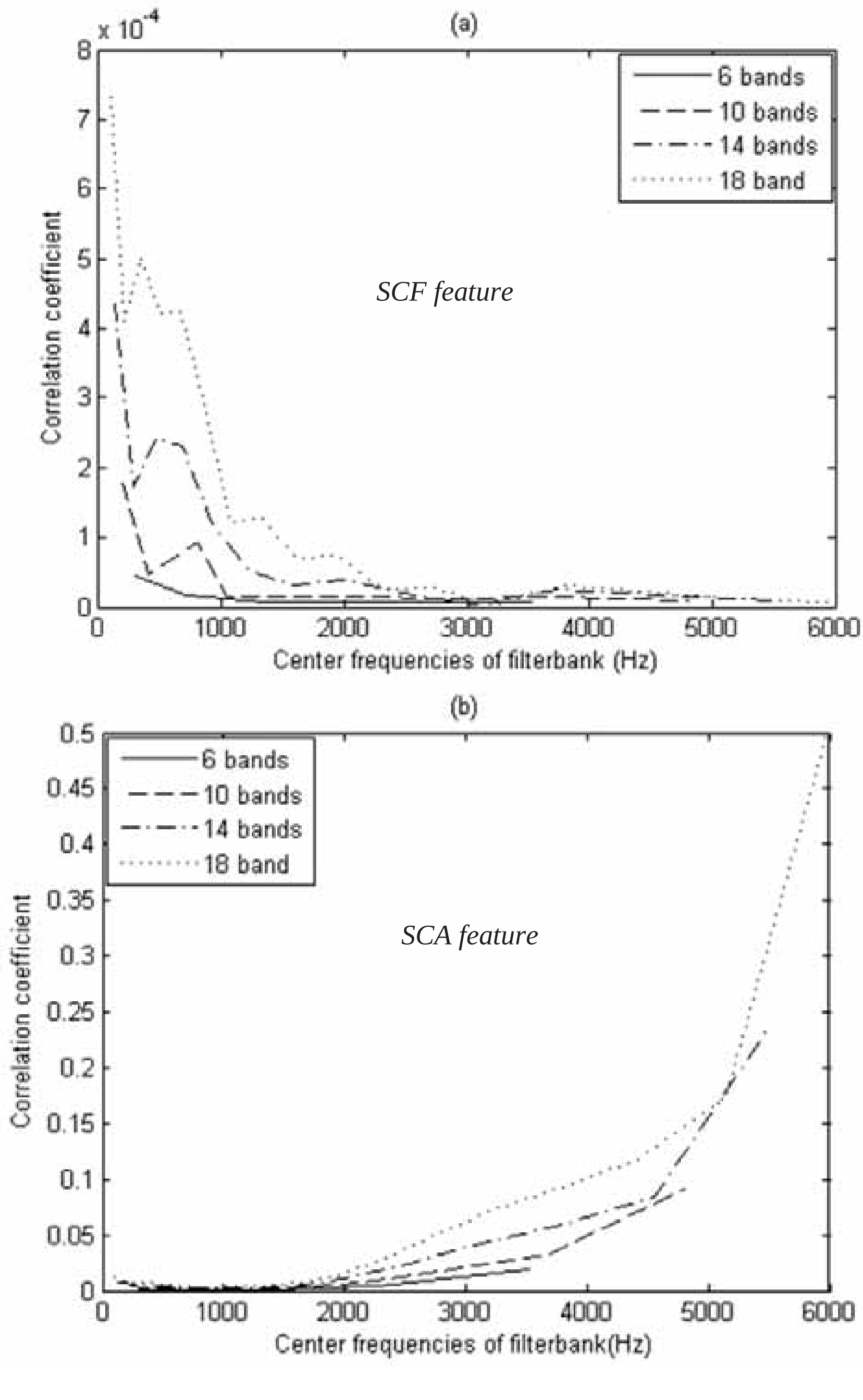


Figure 5.2: Correlation coefficients of adjacent bands of SCF (a) and SCA (b).

As an attempt to explain the performance degradation of the spectral centroid features with large dimensions, the correlation coefficients of the adjacent frequency bands of SCF and SCA features were computed separately over the two corpora. The correlation coefficient ρ between the k^{th} band feature and its adjacent band feature was calculated as

$$\rho_{k,k+1} = \frac{E\{(X_k - \mu_k)(X_{k+1} - \mu_{k+1})\}}{\sigma_k \sigma_{k+1}} \tag{5.1}$$

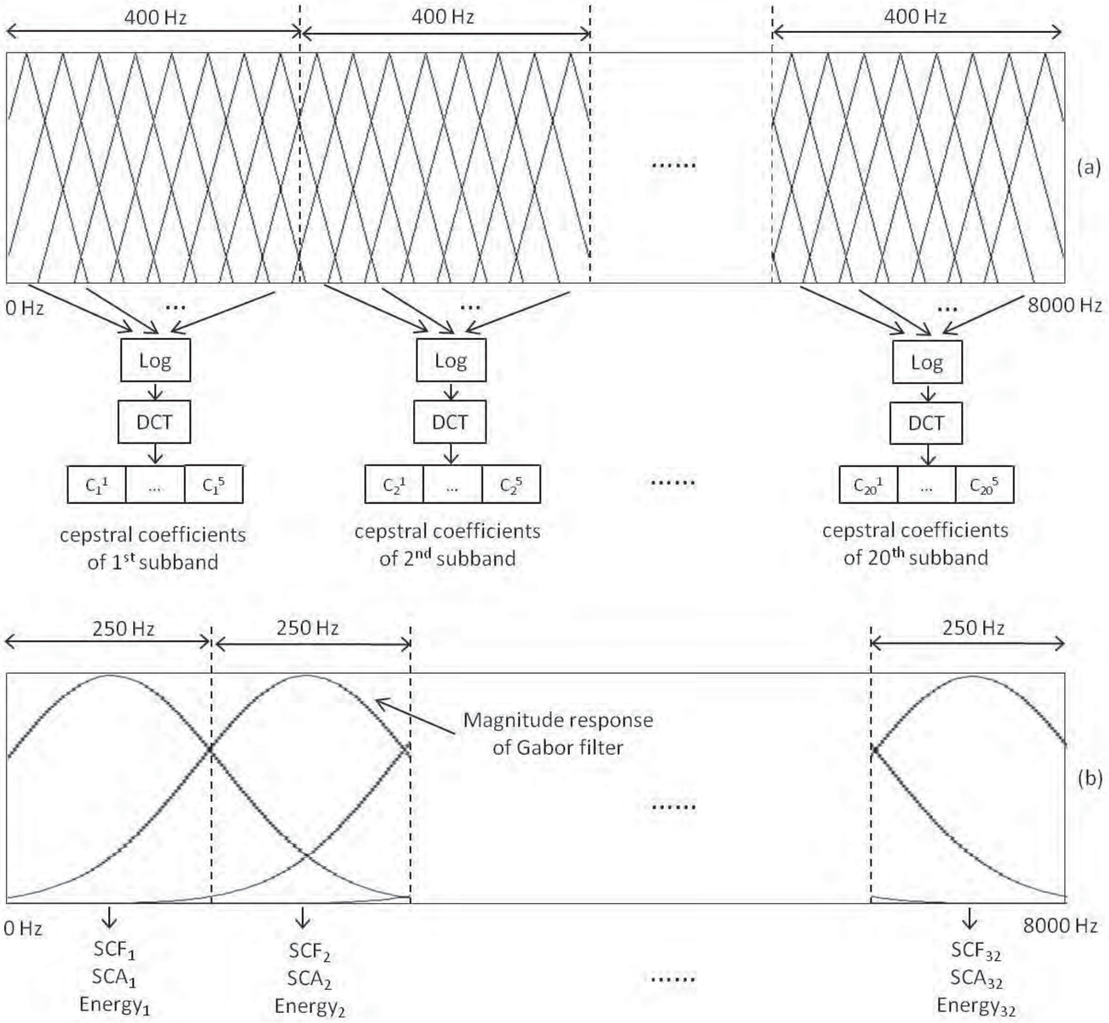
where X_k is the feature vector in the k^{th} band, μ_k and σ_k are the mean and variance of X_k respectively.

The variation of the average correlation coefficients of the SCF and SCA against its center frequency of the band is given in Figure 5.2 for the Stroop test corpus. Similar patterns of variation in the correlation were also observed for the features extracted from the Reading and Comprehension corpus. It can be seen from this figure that the correlation coefficient of the spectral centroid features increases when the number of filters, i.e. the feature dimension, increases. This effect causes redundancy in the feature vector which may be one of the reasons for degradation in the performance of the system with a large number of filters.

5.3 The distribution of CL information across different frequency bands

The effectiveness of the spectral features in cognitive load classification not only depends on their dimensions but also on the way the extraction filters are arranged across the speech bandwidth. An effective filterbank is expected to allocate a large number of filters i.e. have a high frequency resolution in the frequency region containing a significant amount of cognitive load information. Therefore in order to design effective filterbanks, it is necessary to know how cognitive load information is distributed across the bandwidth of speech.

The amount of information presented in each frequency band is proportional to the discrimination ability of the speech features in that band. In other words, the speech feature computed from a frequency band containing a significant amount of CL information is more discriminative for cognitive load than the feature computed from another frequency band containing a lesser amount of CL information. In this study the speech features are extracted in individual subbands and their cognitive load discrimination ability is quantified and used as a measure of the amount of cognitive load information distributed in the corresponding subbands. The discrimination ability of the feature is estimated at three different stages of the classification process: namely feature, model and classification stages. It is estimated at the feature stage based on the separation between the distributions of the feature for different cognitive load levels. It is estimated at the model stage by using the average of the pairwise distances between different statistical models representing different cognitive load levels. Finally, it is estimated at the classification stage based on the accuracy of the system.



400 Hz
400 Hz
400 Hz
......
(a)
0 Hz
...
...
...
8000 Hz
Log
Log
Log
DCT
DCT
DCT
C_1^1 ... C_1^5
C_2^1 ... C_2^5
......
C_{20}^1 ... C_{20}^5
cepstral coefficients of 1st subband
cepstral coefficients of 2nd subband
cepstral coefficients of 20th subband
250 Hz
250 Hz
250 Hz
Magnitude response of Gabor filter
......
(b)
0 Hz
8000 Hz
SCF_1 SCA_1 $Energy_1$
SCF_2 SCA_2 $Energy_2$
......
SCF_{32} SCA_{32} $Energy_{32}$

Figure 5.3: An illustration of the feature extraction of (a) subband cepstral coefficients, and (b) subband SCF, SCA, and energy

In this section, the amount of cognitive load (CL) information distributed in a subband is evaluated through the discrimination ability of the subband cepstral coefficients, SCF, SCA and energy. The subband energy is computed as the sum of the magnitude spectrum in a subband, using equation (3.9) with the weighting coefficient *f* = 1. All of these features are extracted from individual subbands, however the extraction of the cepstral coefficients is different to that of the spectral centroid and energy features. For spectral centroid and energy features, a single feature is extracted from each subband using a filter. The widths of the subbands need to be small so that the features can sufficiently capture the detailed variation of the speech spectrum. A number of cepstral coefficients are extracted from each subband using a series of filters covering that subband. As such, cepstral coefficients can capture the detailed variation of the speech spectrum within a larger subband. The analysis of the CL information distribution using the cepstral coefficients was performed on 400 Hz subbands and that using the SCF, SCA and energy was performed on 250 Hz subbands. Figure 5.3 shows an illustration of the extraction of cepstral coefficients, SCF, SCA, and energy in individual subbands.

5.3.1 Analysis on cepstral coefficients

In order to extract the subband cepstral coefficients, the 8 kHz speech spectrum was first split into twenty uniform subbands of width 400 Hz. The spectrum in each subband was then passed through a series of eight triangular filters, equally spaced in the linear frequency scale for that subband (Figure 5.3a). The triangular filter was chosen as it is commonly used to extract cepstral coefficients for speech recognition [99]. Finally, the discrete cosine transform (DCT) was applied to the log of the power spectrum at the outputs of the filters and the first five DCT coefficients were obtained as the cepstral coefficients for the corresponding subband.

5.3.1.1 Feature-based measure

A feature is considered to be highly discriminative of CL if there is a large separation between the distributions of this feature for different cognitive load levels. In the view of pattern recognition, the speech feature for each cognitive load level is considered as a class and the separation of different classes can be measured using the Fisher ratio, which is defined as the ratio between the variance between classes to the mean of the variances within individual classes [103]. The Fisher ratio for a single dimension feature vector *x* is mathematically expressed as [103]

$$Fisher\ ratio = \frac{\frac{1}{M}\sum_{i=1}^{M}(u_i - u)^2}{\frac{1}{M\,N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(x_i^j - u_i\right)^2} \tag{5.2}$$

where M is the number of classes; N is the number of speech frames; x_i^j is a single dimensional feature of the j^{th} speech frame of cognitive load level i where $i = 1, 2, 3, \ldots, M$ and $j = 1, 2, 3, \ldots, N$; u_i and u are the average subband features of load level i and all load levels respectively, which are defined as

$$u_i = \frac{1}{N}\sum_{j=1}^{N} x_i^j \; ; \qquad u = \frac{1}{M\,N}\sum_{i=1}^{M}\sum_{j=1}^{N} x_i^j \tag{5.3}$$

According to the Fisher criterion, a feature with larger values of Fisher ratio will be more discriminative. The Fisher ratio computed for the cepstral coefficients in different subbands is graphed against the center frequency of a particular subband in Figure 5.4.

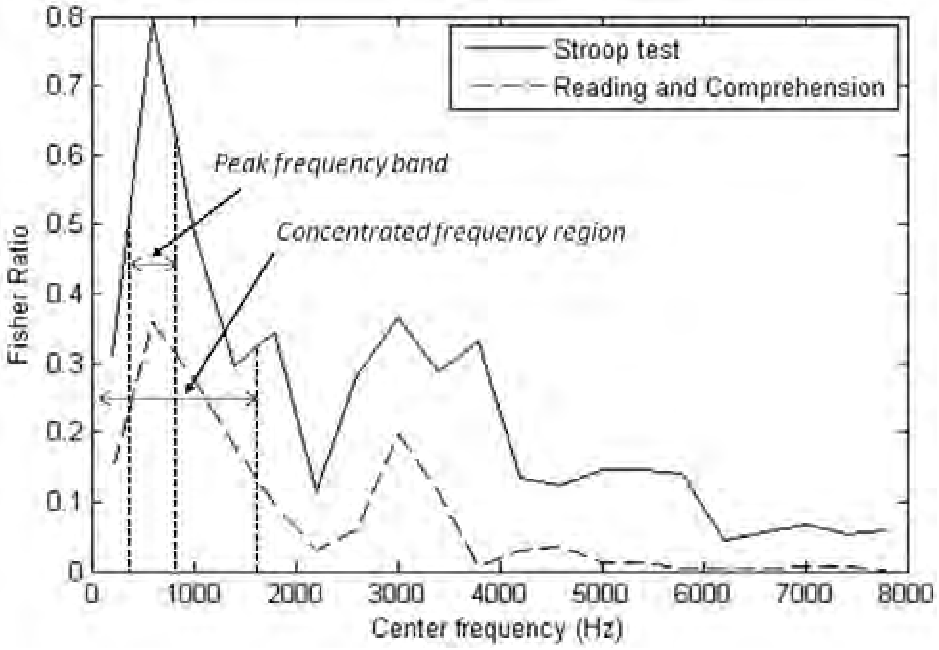


Figure 5.4: Fisher ratio of subband cepstral coefficients.

It can be observed from Figure 5.4 that both Fisher ratio curves of the two corpora initially increase at low frequencies and then gradually decrease as the last frequency band is approached. This suggests that the cognitive load information is concentrated in a specific frequency region. For the convenience of comparing with the results of other methods, the frequency region over which CL information is mainly distributed and the frequency band where the Fisher ratio curves reach their maxima (peak frequency band), are roughly estimated and presented in Table 5.1.

Table 5.1: Concentrated frequency region and peak frequency band of CL according to Fisher ratio curves of cepstral coefficients.

Stroop test		Reading and Comprehension	
CL information concentrated frequency region (Hz)	Peak frequency band (Hz)	CL information concentrated frequency region (Hz)	Peak frequency band (Hz)
(0-1600)	(400-800)	(0-1600)	(400-800)

5.3.1.2 Model-based measure

At the model stage, each cognitive load level is represented by the probability distribution model of the acoustic feature of that level. The probability distribution model used in this work to represent each load level is a Gaussian mixture model. The CL discrimination ability of a speech feature can be quantitatively evaluated through the dissimilarity between different GMMs. The features producing larger dissimilarity between different models are more discriminative. The Kullback-Leibler (KL) distance is used in this work to measure the dissimilarity of GMMs modeling the distributions of cepstral coefficients at two different cognitive load levels in an individual subband [104]. The KL distance is commonly used to measure the distance between two probabilistic models in an information-theoretic sense [105]. The KL distance between two GMMs can be approximated by [104]

$$KL(f \parallel g) \approx \sum_{i=1}^{N} \alpha_i KL(f_i \parallel g_i) \tag{5.4}$$

where *f* and *g* are the two GMMs considered, *N* is the number of mixtures in each model, α_i is the weight of the i^{th} mixture, f_i and g_i are the i^{th} mixture of *f* and *g*. $KL(f_i \parallel g_i)$ is the distance between the i^{th} mixtures of *f* and *g*, which is expressed as [106]

$$KL(f_i \parallel g_i) \approx 0.5\left(\vec{\mu}_i^g - \vec{\mu}_i^f\right)^T \left(\frac{1}{\Sigma^g} + \frac{1}{\Sigma^f}\right)\left(\vec{\mu}_i^f - \vec{\mu}_i^g\right) + 0.5tr\left(\frac{\Sigma^f}{\Sigma^g} + \frac{\Sigma^g}{\Sigma^f} - 2I\right) \tag{5.5}$$

where Σ is the covariance matrix, μ is the mean vector of the Gaussian model and *I* is the identity matrix.

The pairwise Kullback-Leibler (KL) distance in equation (5.4) is computed for every two GMMs and then the final Kullback-Leibler distance is obtained by averaging all pairwise distances. This final Kullback-Leibler distance will be referred to as KL distance for the remainder of this thesis for simplicity. The KL distance of the subband cepstral coefficients is graphed against the center frequency of the subbands as in Figure 5.5.

Figure 5.5 shows that the KL distance curve of the Reading and Comprehension corpus monotonically decreases from the first band to the last band and the curve of the Stroop test corpus initially increases slightly to the second band and then decreases to the last frequency band. These observations indicate that cognitive load information is mainly distributed in the low frequency region and the amount of the information contained in individual subbands decreases with respect to frequency. The frequency regions in which cognitive load information is concentrated with respect to the KL distance curves and the frequency bands where the KL distance curves reach their maxima are roughly determined and presented in Table 5.2. It can be seen that these frequency regions are relatively consistent with those obtained from the analysis of the Fisher ratio, presented in Table 5.1.

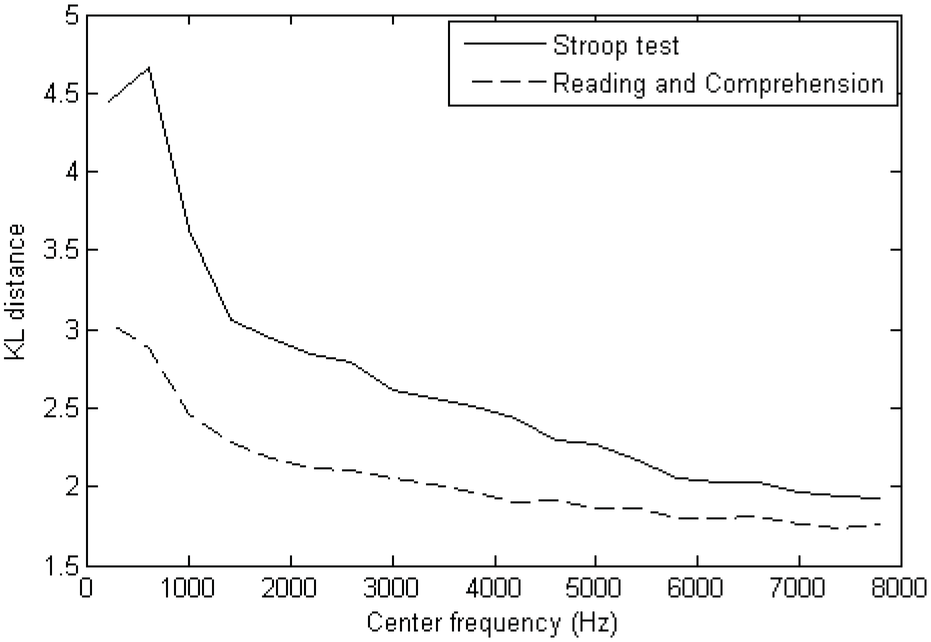


Figure 5.5: KL distance of subband cepstral coefficients.

Table 5.2: Concentrated frequency region and peak frequency band of CL according to the KL distance curves of cepstral coefficients.

Stroop test		Reading and Comprehension	
CL information concentrated frequency region (Hz)	Peak frequency band (Hz)	CL information concentrated frequency region (Hz)	Peak frequency band (Hz)
(0-1600)	(400-800)	(0-1600)	(0-400)

5.3.1.3 Performance based measure

At the classification stage, the discriminative ability of a speech feature can be evaluated using the accuracy of the classification system based on that feature. That is, speech features producing higher classification accuracy are more discriminative. In this section a series of cognitive load classification experiments were conducted by using the cepstral coefficients in individual subbands as features. The obtained accuracies are used to evaluate the cognitive load discriminative ability of the speech features in the corresponding subbands. The accuracies obtained from these experiments are shown in Figure 5.6.

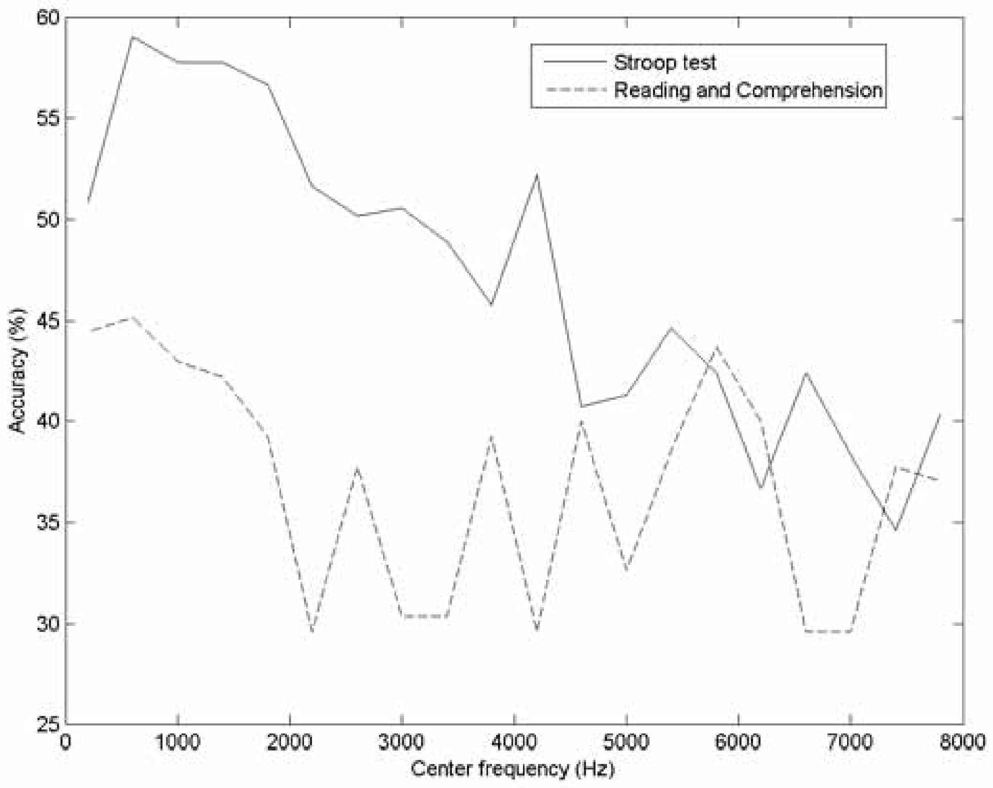


Figure 5.6: Classification accuracies of subband cepstral coefficients.

It can be seen from Figure 5.6 that the accuracy curves have high values in the frequency region below 2 kHz. Both curves reach their maxima at the second subband ranging from 400 Hz to 800 Hz which is called the peak frequency band. These observations indicate that the cognitive load information is mainly distributed in the low frequency region and that the 400 Hz to 800 Hz frequency band contains the largest amount of the information as presented in Table 5.3. These observations are consistent with those of previous analysis using the Fisher ratio and the KL distance. Beyond the peak frequency band, the accuracy curve of the Stroop test corpus tends to decrease to the last frequency band. This suggests that the amount of cognitive load information contained in individual subbands decreases with respect to frequency. The accuracy curve of the Reading and Comprehension corpus decreases to approximately 2 kHz and then varies randomly close to 33.3% which happens to be the probability of determining the correct CL level by chance. This can be attributed to the fact that when the cepstral coefficients of these subbands were used to train the models, the three Gaussian models obtained were not well separated to represent the three different load levels. This claim is supported by the very small KL distances of the Reading and Comprehension corpus beyond 2 kHz as shown in Figure 5.5.

Table 5.3: Concentrated frequency region and peak frequency band according to the accuracy curves of cepstral coefficients.

Stroop test		Reading and Comprehension	
CL information concentrated frequency region (Hz)	Peak frequency band (Hz)	CL information concentrated frequency region (Hz)	Peak frequency band (Hz)
(400-2000)	(400-800)	(0-1600)	(400-800)

5.3.2 Results from the analysis on SCF, SCA, and energy

In this section, analysis of cognitive load information distribution is performed on the other three spectral features: spectral centroid frequency (SCF), spectral centroid amplitude (SCA) and energy, to consolidate the results of cepstrum coefficient analysis. In order to extract the subband SCF, SCA, and energy, the speech spectrum was decomposed into thirty two subbands of width 250 Hz using thirty two uniformly spaced Gabor bandpass filters (Figure 5.3b). The Gabor filter is commonly used to extract spectral features [94]. The thirty two subbands are chosen herein in order to have high resolution in the spectral domain. The spectral centroid and energy features are computed in individual subbands and their discrimination abilities for cognitive load are quantitatively analyzed at the feature, model, and classification stages.

The results from the analysis of these features using the Fisher ratio, KL distance, and classification accuracy measures are shown in Figures 5.7, 5.8, and 5.9 respectively. These figures again indicate that the cognitive load information distribution initially increases to a peak and then decreases in the high frequency bands. These observations are consistent with the results obtained from the analysis on the cepstral coefficients. Furthermore, the information distribution curves of the SCA and energy are very similar.

This can be due to the fact that both of these features relate to the magnitude spectrum in each subband. The frequency region in which cognitive load information is mainly distributed and the peak frequency band obtained from the analysis on the spectral centroid frequency and spectral centroid amplitude features are presented in Table 5.4.

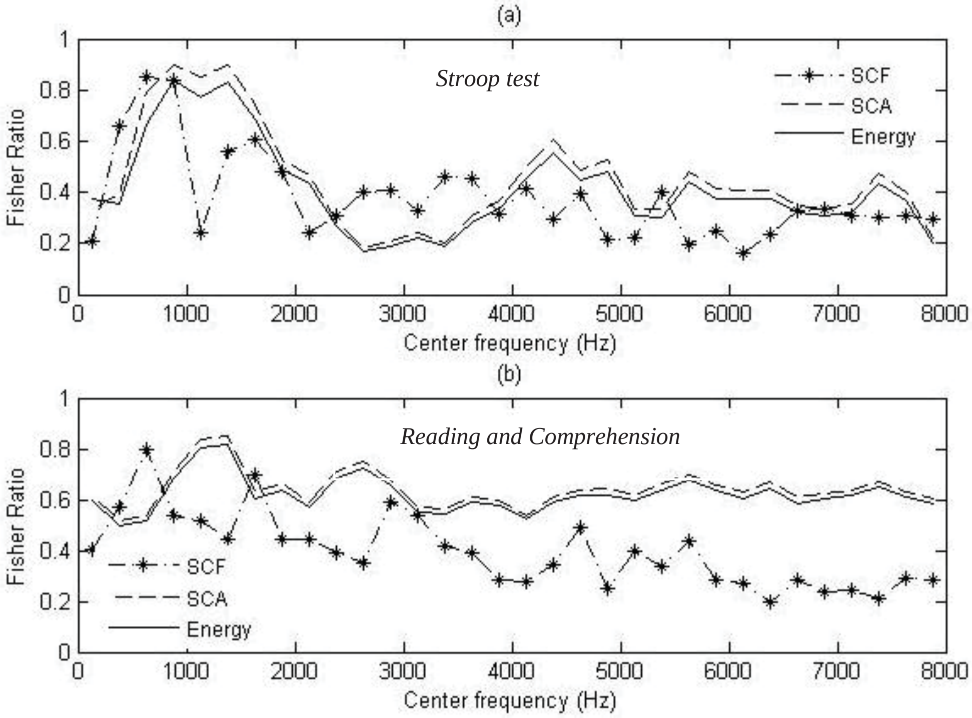


Figure 5.7: Fisher ratio of subband SCF, SCA, and energy features computed across (a) Stroop test corpus, and (b) Reading and Comprehension corpus.

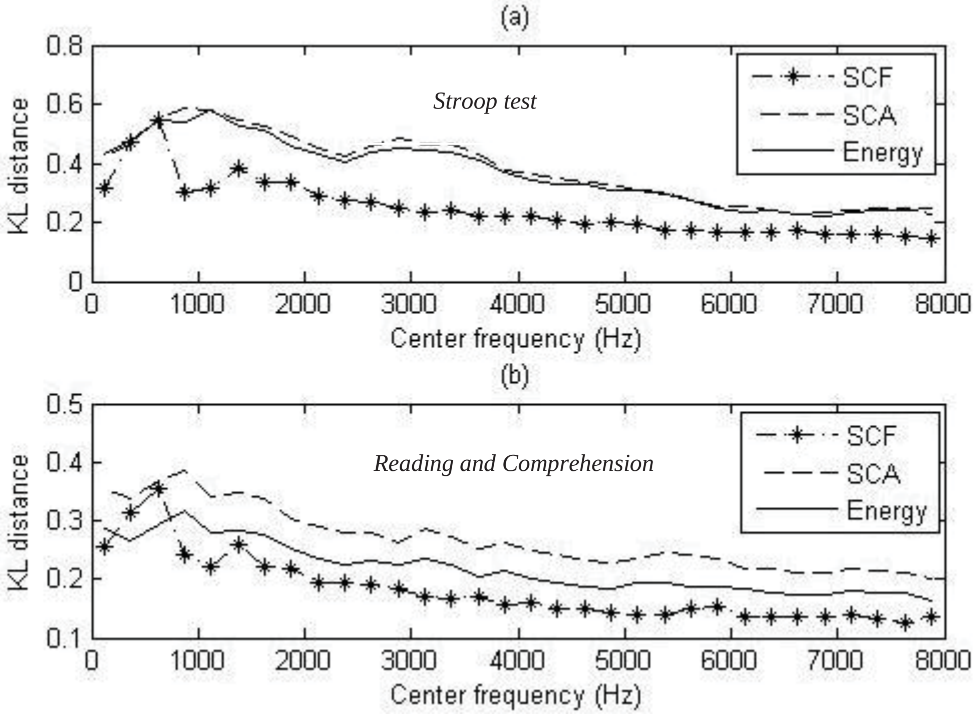


Figure 5.8: KL distance of subband SCF, SCA, and energy computed across (a) Stroop test corpus, and (b) Reading and Comprehension corpus.

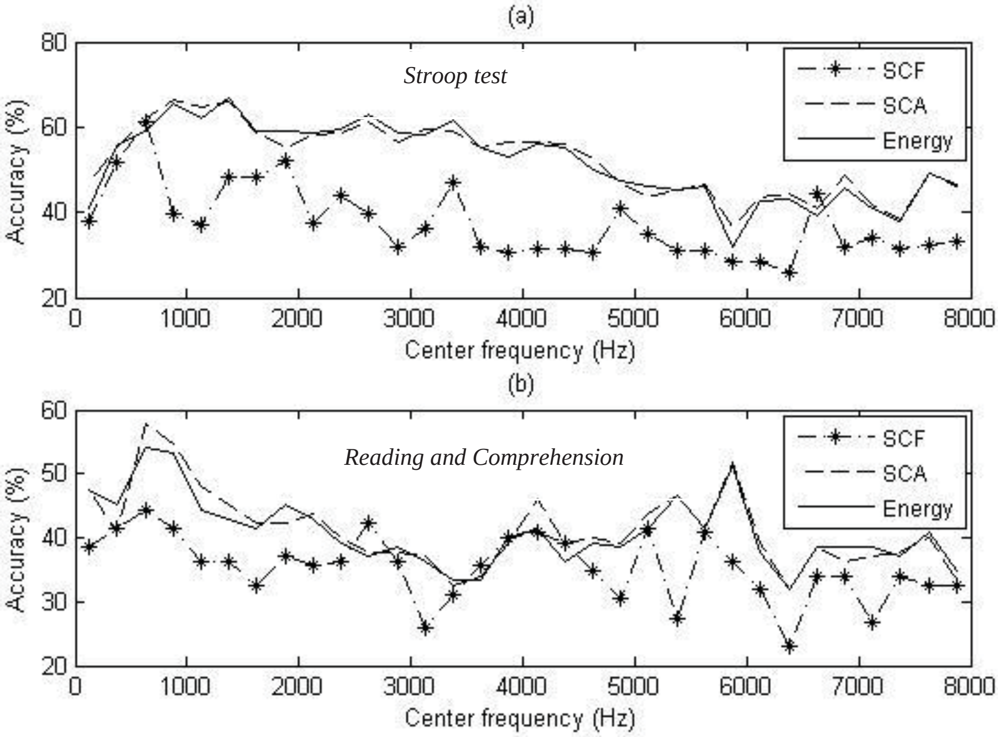


Figure 5.9: Classification accuracies of subband SCF, SCA, and energy evaluated on (a) Stroop test corpus, and (b) Reading and Comprehension corpus.

5.3.3 Spectral distribution of CL information

The frequency regions in which cognitive load information is mainly distributed and the peak frequency band containing the largest amount of CL information obtained by using different analysis methods and different spectral features are tabulated in Table 5.4 for a comprehensive analysis. The results obtained by analyzing the energy feature are not reported herein as they are same as those obtained by analyzing the SCA features.

It can be observed from Table 5.4 that the results obtained by using three different analysis methods are highly consistent. The overall results for concentrated frequency region and peak band for each feature are determined by at least two of the three analysis methods. The overall peak frequency band for the SCA feature performed on the Reading and Comprehension corpus is left blank as the results obtained by the three different methods are not very consistent.

According to the results in Table 5.4, the region from 0 Hz to 1.5 kHz is the frequency region in which CL information is mainly concentrated for speech sampled at 16 kHz as it is the intersection of the overall concentrated regions of all three features for both corpora. This is illustrated in Figure 5.10a. The exception is the SCA feature of the Stroop test corpus which indicated (250-1500) Hz. Furthermore, the peak frequency bands of the cepstral coefficients, SCA, and SCF are (400-800) Hz, (750-1000) Hz, and (500-750) Hz respectively. The union of these three regions is (400-1000) Hz in which all the CL information distribution curves reach their maximum, as shown in Figure 5.10b. This indicates that the band from around 400 Hz to 1 kHz contains the largest amount of CL information for speech sampled at 16 kHz.

Table 5.4: Concentrated frequency region and peak frequency band of CL information using cepstral coefficients, SCA and SCF features.

Feature	Analysis method	Stroop test		Reading and Comprehension	
		CL information concentrated frequency region (Hz)	Peak frequency band (Hz)	CL information concentrated frequency region (Hz)	Peak frequency band (Hz)
Cepstral coefficients	Fisher Ratio	(0-1600)	(400-800)	(0-1600)	(400-800)
	KL distance	(0-1600)	(400-800)	(0-1600)	(0-400)
	Classification accuracy	(400-2000)	(400-800)	(0-1600)	(400-800)
	Overall	(0-1600)	(400-800)	(0-1600)	(400-800)
SCA	Fisher Ratio	(250-1750)	(750-1000)	(0-1500)	(1250-1500)
	KL distance	(250-1500)	(750-1000)	(0-1500)	(750-1000)
	Classification accuracy	(500-1500)	(750-1000)	(500-1500)	(500-750)
	Overall	(250-1500)	(750-1000)	(0-1500)	---
SCF	Fisher Ratio	(0-1500)	(500-750)	(0-1500)	(500-750)
	KL distance	(0-1500)	(500-750)	(0-1500)	(500-750)
	Classification accuracy	(0-2000)	(500-750)	(0-1500)	(500-750)
	Overall	(0-1500)	(500-750)	(0-1500)	(500-750)
Intersection/union frequency region (Hz)		(0-1500)	(400-1000)	(0-1500)	(400-1000)

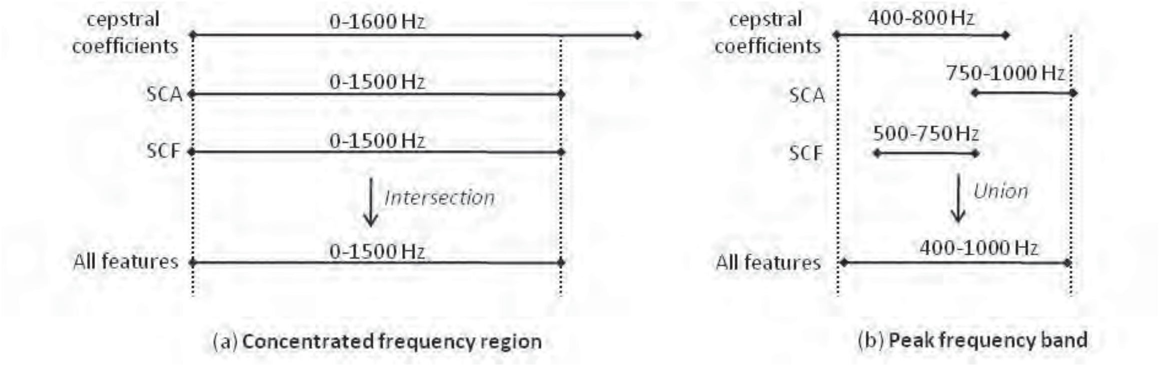


Figure 5.10: Determining the concentrated frequency region and peak frequency band for all features

It is interesting to see in Table 5.4 that the observations of the frequency region in which CL information is concentrated and the peak frequency band are highly consistent between both corpora, although the speech in these two corpora was collected in very different methods. This indicates that the frequency region in which CL information is mainly concentrated and the peak frequency band found in this study are data independent.

5.4 Filterbank design for CL classification

The analysis in Section 5.3 shows that CL information is mainly concentrated in the frequency region between 0 Hz and 1.5 kHz. It is hypothesized that if the front-end filterbanks used to capture the spectral features have a high frequency resolution, i.e. contain a large number of filters, in this region, the CL information in this region can be emphasized. This in turn can improve the performance of the classification system based on the spectral features. This section designs the filterbanks to extract the spectral features specifically for cognitive load classification by systematically allocating the center frequencies and bandwidths of their filters in such a way that a larger number of filters are allocated in (0-1.5) kHz according to the spectral information of cognitive load contained in it. An example of a triangular filterbank with the center frequency and bandwidth of its second filter i.e. fc_2 and BW_2 are illustrated in Figure 5.11.

Among the three measures used to quantify the distribution of CL information, i.e. KL distance, Fisher ratio and classification accuracy, the KL distance measure is expected to describe the distribution of spectral information more accurately than the other two measures. This is because this measure estimates the distance between GMMs which is based on the modeling of the distribution of the speech features. The Fisher ratio, on the other hand, is computed solely from the means and the variances of the feature distributions while ignoring the shape of the feature distributions, which is an important factor in classification. Since KL distance considers the mean, variance and shape of a speech feature distribution, it would be more reliable than the Fisher ratio in measuring the distance between classes. Furthermore, the disadvantage of using the classification accuracy score to represent the distribution of cognitive load information is that it varies randomly if the models of different CL levels are not sufficiently separate, as shown in Figures 5.6 and 5.9b and discussed in Section 5.3.1.3. For these reasons, the distribution of cognitive load information based on Kullback-Leibler distance is used as the basis to allocate the center frequencies and bandwidths of the filters in this study.

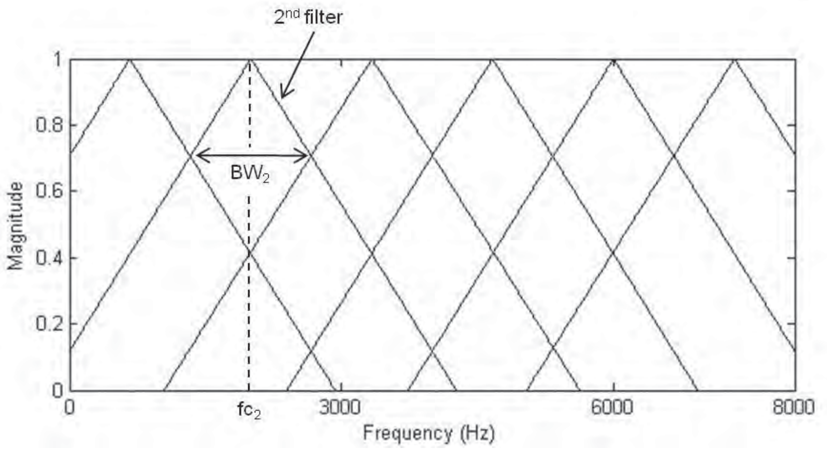


Figure 5.11: An example of a triangular filterbank with 2nd filter's center frequency fc_2 and bandwidth BW_2.

The effectiveness of the filterbank designed in this study is evaluated by comparing with the mel, Bark, equivalent rectangular bandwidth (ERB) and Hertz filterbanks. The mel, Bark, and ERB are perceptually motivated filterbanks whose frequency resolution is high in the low frequency region and low in the high frequency region. For the Hertz filterbank, the filters are uniformly allocated along the bandwidth of the speech signal.

In this study, filterbanks are designed independently on two different corpora based on the CL information distribution estimated for each corpus. All the filterbanks are designed in such a way to produce spectral features with six dimensions because Section 5.2 found that six dimensions produce highest accuracy for the classification system.

5.4.1 Procedure to allocate center frequencies and bandwidths of the filters

The procedure to allocate the center frequencies and the bandwidths of the filters of filterbanks in individual frequency regions according to the amount of CL information distributed in the corresponding frequency regions is presented below:

1. The original Kullback-Leibler (KL) curve is generated by connecting all the elements of the KL distance,

   $$KL = \{(f_{c1}, KL_1), (f_{c2}, KL_2), \dots, (f_{cN}, KL_N)\} \tag{5.6}$$

   where f_{ci} and KL_i are the center frequency and KL distance of the i^{th} subband respectively, and N is the number of subbands.
2. The curve is normalized such that the area below it is one unit. This is done because the actual value of Kullback-Leibler distance is not as important as the variation of it for this design.
3. The normalized curve is divided into N subbands whose areas under the curve are equal. This division creates narrower subbands in the frequency region with a larger Kullback-Leibler distance.
4. The center frequencies and the bandwidths of the filters are allocated as the centers and the widths of the subbands mentioned in step 3. This will ensure the bandwidth of a filter in the frequency region with a larger KL distance is smaller than that in the frequency region with smaller KL distance. As such, more filters are allocated in the frequency region containing more CL information.

In order to illustrate the procedure described above, an example of allocating the center frequencies and bandwidths of the filters of a filterbank based on this procedure is shown in Figure 5.12. The number of filters in this example is ten, chosen for visualization convenience. The Kullback-Leibler (KL) distance curve in this example was computed from the subband cepstral coefficients of the Reading and Comprehension corpus, as shown in Figure 5.5. It can be seen from Figure 5.12 that the bandwidth of the filter (the width of the subband) increases with respect to the frequency, i.e. the number of filters of the designed filterbank in the low frequency region is larger than that in the high frequency region. Therefore it is expected that this filterbank can emphasize the CL information in the low frequency region. However, it can also be seen from Figure 5.12 that the bandwidths of the filters in the low frequency region are only slightly smaller than those in the high frequency region. As a result, the number of filters in the low frequency region is only slightly more than that in the high frequency region. This filterbank may hence not be very effective for the classification as it may not sufficiently emphasize the cognitive load information in the low frequency band. The reason that the bandwidths of the filters in the low frequency region are not very small is because the gradient of the KL distance curve is small as seen in Figure 5.12.

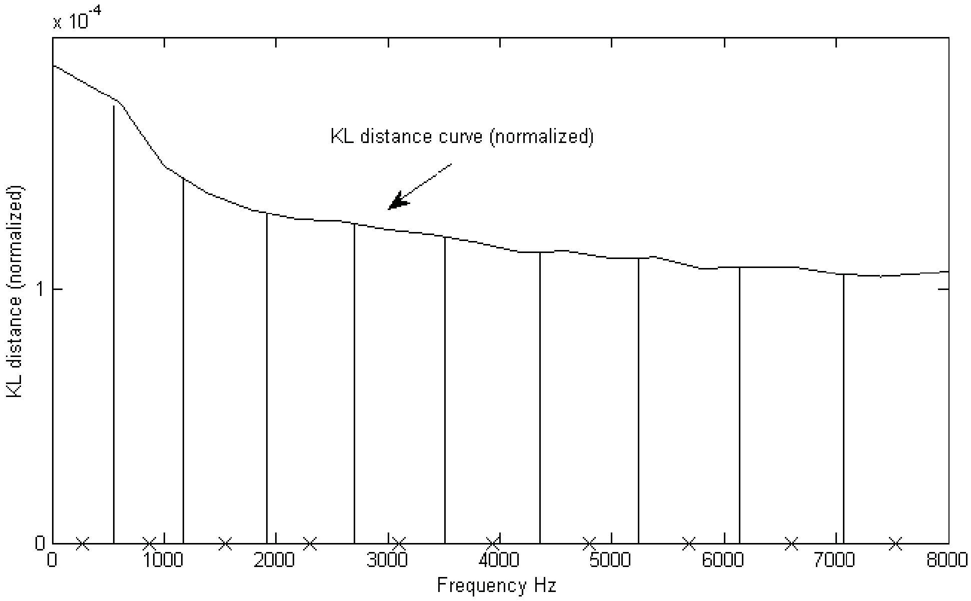


Figure 5.12: Allocation of center frequencies and bandwidths of a filterbank consisting of ten filters with a KL distance curve. The center frequency of each filter is marked by ‘x’ and the bandwidth is indicated by the adjacent vertical lines.

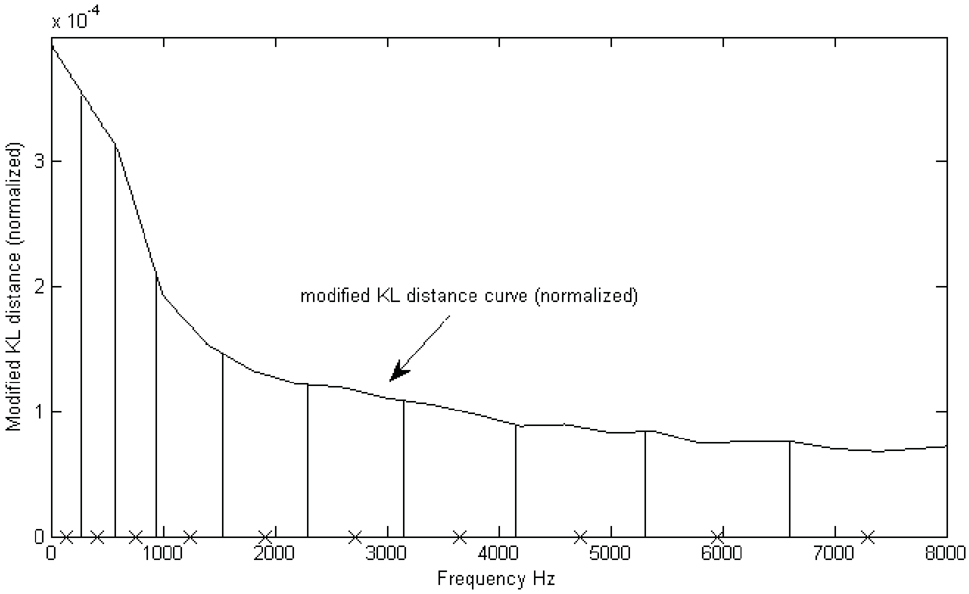


Figure 5.13: Allocation of center frequencies and bandwidths of the filterbank consisting of ten filters with a modified KL distance curve with α = 3. The center frequency of each filter is marked by ‘x’ and the bandwidth is indicated by the adjacent vertical lines.

In order to further decrease the bandwidth of the filters in the low frequency region to allocate a larger number of filters in this region, an approach to modify the Kullback-Leibler distance curve in such a way that its gradient is increased was proposed. The modified curve is obtained by connecting all the modified KL distance (KL_{mod}), which is expressed as

$$KL_{mod} = \{(f_{c1}, KL_1{}^{\alpha}), (f_{c2}, KL_2{}^{\alpha}) \dots, (f_{cN}, KL_{\mathrm{N}}{}^{\alpha})\} \qquad (5.7)$$

with $\alpha > 1$ to increase the gradient of the modified curve. This modified distance curve is used to allocate the center frequencies and bandwidths of the filters, as described in steps 2–4 of the above-mentioned procedure. An illustration of allocated center frequencies and bandwidths for ten filters of a filterbank based on the modified Kullback-Leibler distance curve with $\alpha = 3$ is shown in Figure 5.13. It can be observed from Figures 5.12 and 5.13 that by increasing α, the gradient of the Kullback-Leibler distance curve increases. This reduces the bandwidth of the filters in the low frequency region and increases the bandwidth of the filters in the high frequency region. Therefore, a larger number of filters are allocated in the low frequency region and a smaller number of filters are allocated in the high frequency region. This can lead to the filterbank having more emphasis on the speech features in the low frequency region. However, a very large value of α would result in a very small number of filters in the high frequency region and can cause the filterbank to not capture the cognitive load information in this region sufficiently. As a consequence, this can degrade the performance of the classification system. These explanations suggest that there should be a range of good values of α for the design that provides a compromise between capturing the information in both low and high frequency regions. The filterbanks designed with those values of α are expected to produce high performance for the system.

The value of α used in this study was empirically chosen by carrying out two sets of classification experiments. The first set of experiments was carried out with the value of α varied from 1 to 10 with a step size of 1. Preliminary experiments had found that α beyond this range degrades the system performance. This set of experiments was used to determine the range of α producing high classification accuracy. The second set of experiments was carried out to fine tune the value of α found in the first set of experiments by varying that α with a step size of 0.1. The α value producing the highest classification accuracy in the second set of experiments will be chosen for allocating the center frequencies and bandwidths of the filters of the designed filterbanks.

5.4.2 Designing filterbank to extract cepstral coefficients

In this section, a filterbank consisting of twenty triangular filters used to extract cepstral coefficients for each corpus. The filterbank was designed using the procedure described in Section 5.4.1 where the KL_i in equation (5.7) is the KL distance computed on the cepstral coefficients of the i^{th} subband.

5.4.2.1 Filterbank design

The classification accuracies obtained from the coarse tune experiment to find α are shown in Figure 5.14. It can be observed from this figure that both very small and very large values of α degrade the performance of the CL classification system, as explained in Section 5.4.1. Furthermore, the approximate region of α producing high system performance on the Stroop test and Reading and Comprehension corpora are $3 < \alpha < 5$ and $6 < \alpha < 8$ respectively.

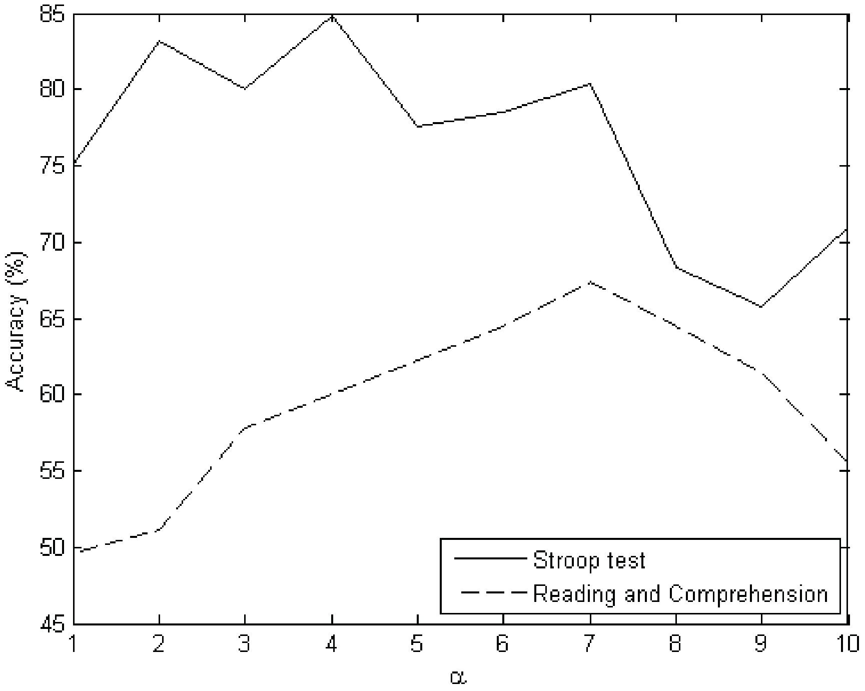


Figure 5.14: Classification accuracy of cepstral coefficients extracted using the designed filterbanks with various values of α.

In the second set of experiments for the Stroop test and Reading and Comprehension corpora, the value of α was varied with a step of 0.1 between 3 and 5 and between 6 and 8 respectively. It was found from these experiments that the value of α producing the highest classification accuracy when performed on the Stroop test corpus was $\alpha = 4$ and for the Reading and Comprehension corpus was $\alpha = 7.6$. These values of α were used to designed the filterbank for extracting the cepstral coefficients for CL classification. The center frequencies and bandwidths of the designed filterbanks, mel, Bark, ERB and Hertz filterbanks are shown in Figure 5.15.

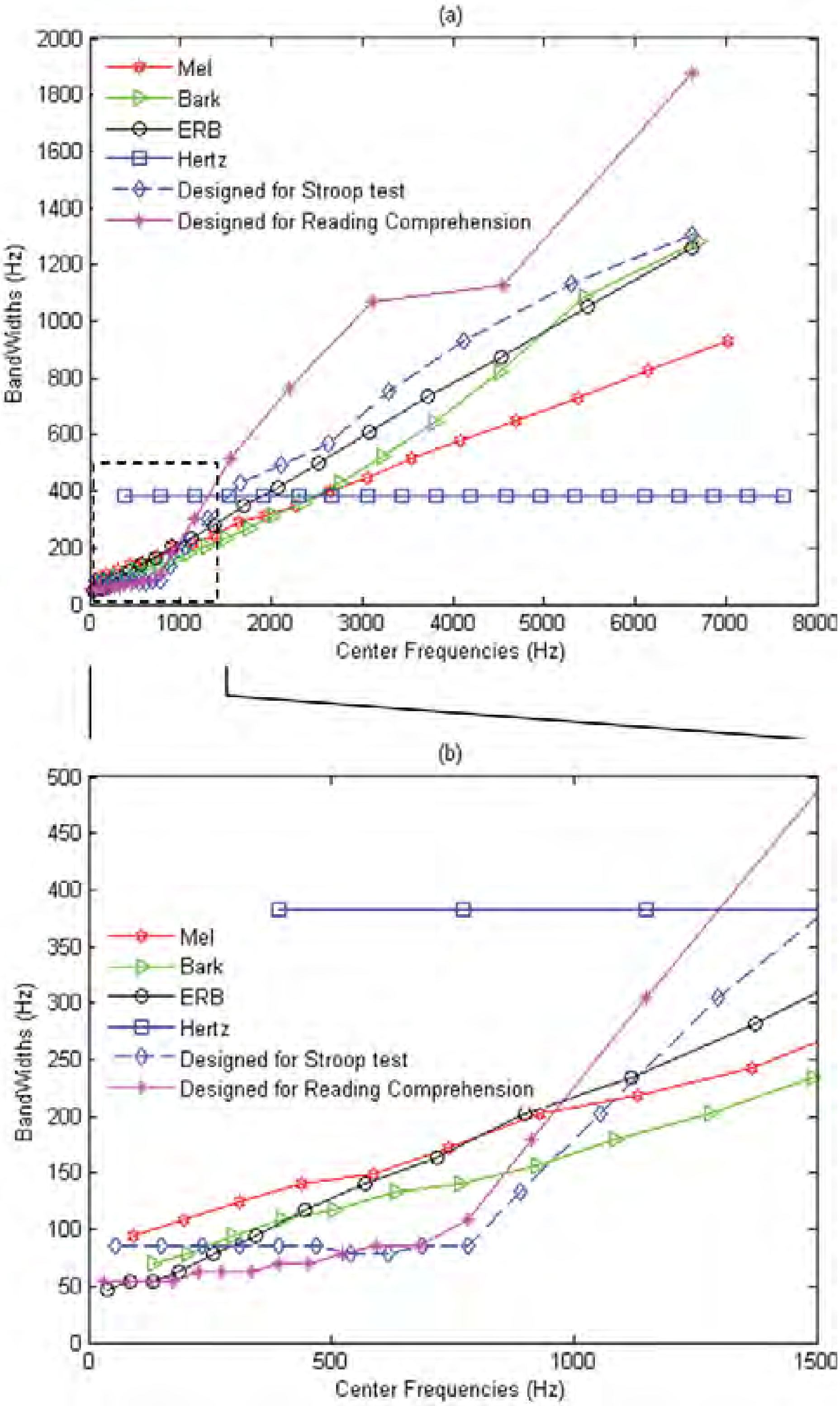


Figure 5.15: (a) Center frequencies and bandwidths of different filterbanks used to extract cepstral coefficients and (b) The magnified view in the region (0-1.5) kHz.

The number of filters allocated in the region from 0 Hz to 1.5 kHz of different filterbanks is shown in Table 5.5 for ease of frequency resolution comparison in this region. It can be seen from this table that the designed filterbanks have a larger number of filters in the frequency region (0-1.5) kHz in which cognitive load information is concentrated, as found in Section 5.3, than the other filterbanks. The designed filterbank therefore can capture CL information more effectively than the other filterbanks.

Table 5.5: The number of filters in the region (0-1.5) kHz of various filterbanks.

Filterbank	Designed for Reading and Comprehension	Designed for Stroop test	Mel	Bark	ERB	Hertz
Number of filters	15	13	9	11	12	3

5.4.2.2 Performance of the designed filterbanks

In order to evaluate the effectiveness of the designed filterbanks for cognitive load (CL) classification and for comparison with the mel, Bark, ERB and Hertz filterbanks, cepstral coefficients are extracted using these filterbanks for CL classification. The classification accuracies obtained from these experiments are shown in Table 5.6.

Table 5.6: Accuracies of cepstral coefficients based on different filterbanks.

Filterbank		Designed	Mel	Bark	ERB	Hertz
Accuracy (%)	Stroop test	84.8	78.9	79.8	83.9	75.6
	Reading and Comprehension	71.1	64.5	65.2	68.1	48.9

As shown in Table 5.6, the designed filterbanks consistently produce the highest performance for the system compared to all other filterbanks used. This supports the hypothesis that cognitive load information can be captured more effectively and the performance of the system can be improved by allocating a larger number of filters of the filterbank in the frequency region from 0 Hz to 1.5 kHz in which CL information is mainly concentrated. It is also clear that the perceptual filterbanks are not the optimal filterbanks for the classification. Furthermore, among the three perceptual filterbanks mel, Bark and ERB, the ERB filterbank produced the highest accuracy and the mel filterbank produced the lowest accuracy for the classification system. This may be because among these three filterbanks the ERB filterbank has the largest number of filters and the mel filterbank has the smallest number of filters in the frequency region (0-1.5) kHz. These observations further support the hypothesis that increasing the frequency resolution of the filterbank in the low frequency region can improve the performance of the classification system.

Compared to the mel, Bark and ERB filterbanks evaluated on the Stroop test corpus, the designed filterbank provides a relative error rate reduction of 28.0%, 24.8% and 5.6%. The corresponding relative error rate reductions for the Reading and Comprehension corpus are 18.6%, 17% and 9.4% respectively.

The Hertz filterbank consistently produces the lowest performance compared to all filterbanks used. This is most probably due to the fact that this filterbank has the lowest number of filters in the frequency region from 0 Hz to 1.5 kHz, as seen in Figure 5.15 and Table 5.5.

5.4.3 Designing a filterbank to extract spectral centroid features

In this section, a filterbank consisting of six Gabor filters used to extract the spectral centroid features (SCF and SCA) for each cognitive load corpus is designed by allocating the center frequencies and bandwidths. Six filters were chosen because this produces spectral centroid features with six dimensions, shown to be the dimension providing the highest performance for the system according to the investigation in Section 5.2. The procedure to allocate the center frequencies and bandwidths of the filters of the filterbanks was described in Section 5.4.1. The KL_i in equation (5.7) is obtained by normalizing the Kullback-Leibler distance curves of the spectral centroid frequency and spectral centroid amplitude features to have the unit area under these curves and then averaging them. As the fusion of the SCF-based and SCA-based systems always outperforms the individual systems, the objective of designing the filterbanks in this section is to produce the best results under the fusion of the SCF-based and SCA-based systems.

5.4.3.1 Filterbank design

The classification accuracies obtained from the first set of experiments to coarse tune α, i.e. α varied from 1 to 10 with a step size of 1, are shown in Figure 5.16. It can be seen from this figure that the performance of the system does not vary much with respect to α. Furthermore, the fused system consistently provides higher accuracies than those systems based on individual SCF and SCA features. The range of α that produces the highest performance for the fusion system on the Stroop test corpus is $2 < \alpha < 6$ and for the Reading and Comprehension corpus is $7 < \alpha < 9$.

In the second set of experiments to fine tune α, the value of α was varied from 2 to 6 and from 7 to 9, with a step of 0.1 accordingly. It was found from these experiments that the value of α producing the highest performance for the system performed on the Stroop test corpus was $\alpha = 2.6$ and the Reading and Comprehension corpus was $\alpha = 7.1$. These values of α were used to allocate the center frequencies and bandwidths of the designed filterbanks.

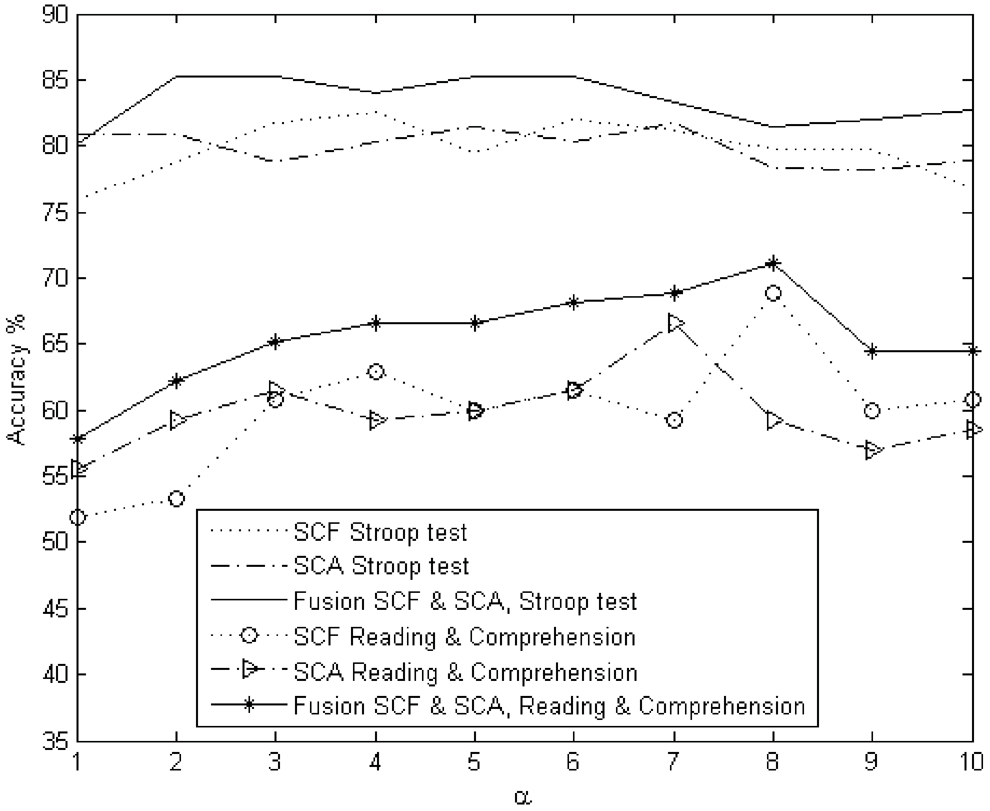


Figure 5.16: Classification accuracies of SCF and SCA extracted using the designed filterbanks with various value of α.

The center frequencies and bandwidths of the designed filterbanks, mel, Bark, ERB and Hertz filterbanks are shown in Figure 5.17. It can be observed from this figure that the designed filterbank for the Reading and Comprehension corpus allocates a larger number of filters in the region from 0 Hz to 1.5 kHz than other filterbanks. Furthermore, unlike the perceptual filterbanks that have the largest frequency resolution in the lowest frequency region, the designed filterbanks have the largest resolution in the region approximately (400-1000) Hz, containing the most significant amount of cognitive load information, as found in Section 5.3.3. The designed filterbanks therefore can capture the cognitive load information contained in the spectral centroid features more effectively than the mel, Bark, ERB and Hertz filterbanks.

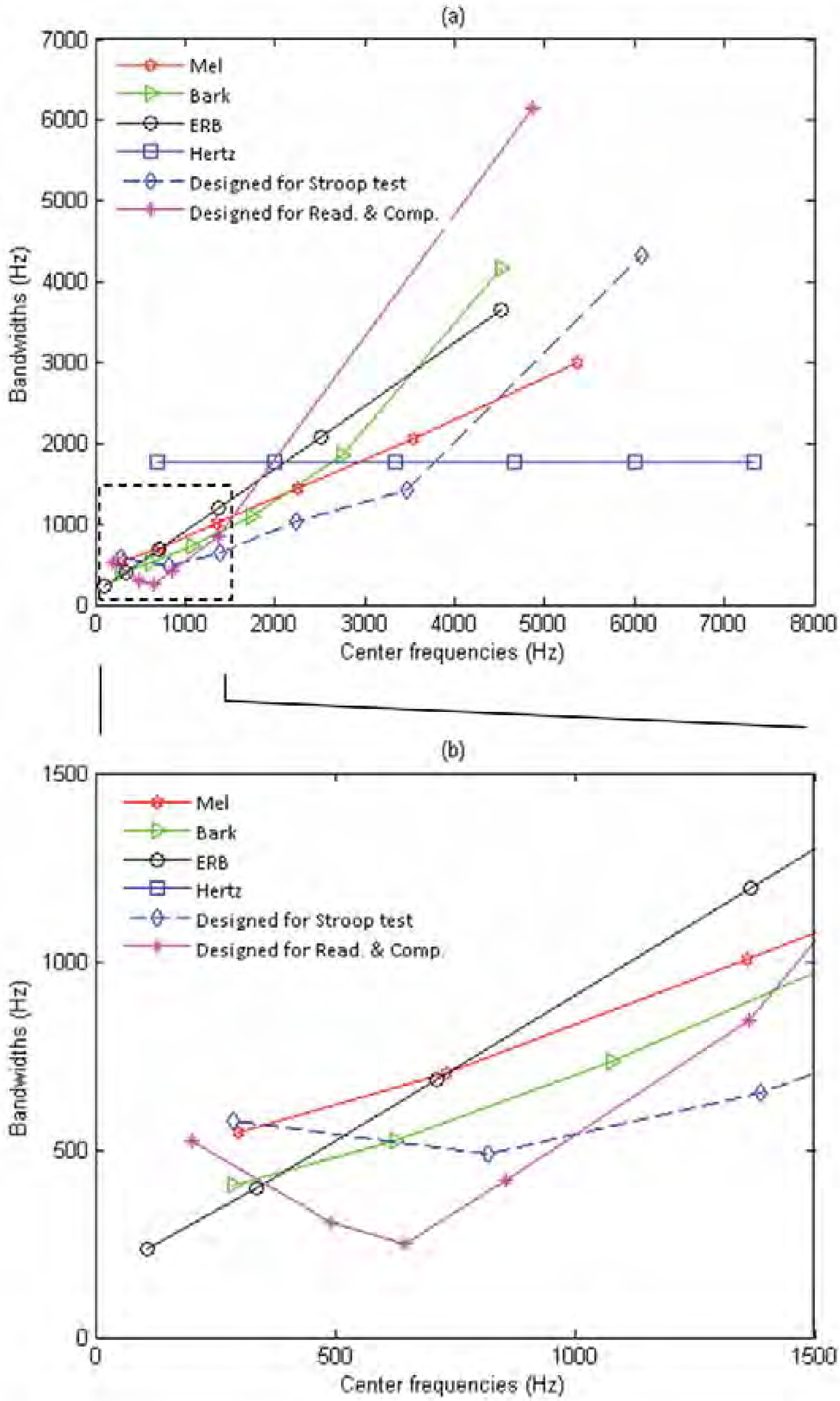


Figure 5.17: (a) Center frequencies and bandwidths of different filterbanks used to capture the SCF and SCA features, (b) The magnified view in the region (0-1.5) kHz.

5.4.3.2 *Performance of the designed filterbanks*

In order to evaluate the effectiveness of the designed filterbanks and compare it with the mel, Bark, ERB and Hertz filterbanks, the spectral centroid features extracted using these filterbanks were used to perform the classification experiments. The classification accuracies of these experiments are shown in Table 5.7.

Table 5.7: Classification accuracies of SCF and SCA extracted using different filterbanks.

	Filterbank		Designed	Mel	Bark	ERB	Hertz
Accuracy (%)	Stroop test	SCF	78.2	82.0	84.3	82.8	71.3
		SCA	84.3	83.7	83.9	84.3	75.9
		Fusion SCF&SCA	87.8	87.2	84.6	86.5	77.6
	Reading and Comprehension	SCF	62.2	63.0	68.9	65.2	44.4
		SCA	72.6	61.5	64.4	63.7	47.4
		Fusion SCF&SCA	74.8	71.9	70.4	69.6	48.9

It is observed from Table 5.7 that among all the filterbanks used, the designed filterbanks provide the highest accuracies for the SCA-based system and the fusion of SCF-based and SCA-based system. However, the system based on the SCF feature extracted using the designed filterbanks did not outperform those using mel, Bark and ERB filterbanks.

Compared to the mel, Bark and ERB filterbanks, evaluated on the Stroop test corpus, the designed filterbank provides a relative error rate reduction of 4.7%, 20.8% and 9.6% respectively for the fusion of SCF-based and SCA-based systems. The corresponding relative error rate reductions for the Reading and Comprehension corpus are 10.3%, 14.9% and 17.1% respectively.

The effectiveness of the designed filterbanks for the fusion of SCF and SCA can be explained by the fact that they provide the highest frequency resolution at approximately (400-1000) Hz. As a result, they can capture cognitive load information more effectively than the other filterbanks. Furthermore, the designed filterbanks consistently provide the highest performance for an SCA-based system. However, they did not provide better performance than the mel, Bark and ERB filterbanks for an SCF-based system. This might be because the SCF is a frequency-based feature which captures the approximate location of the local maxima of the magnitude spectrum in subbands, unlike the cepstral coefficients and SCA features which are amplitude-based. That is the high frequency resolution in the frequency region from 400 Hz to 1 kHz of the designed filterbanks result in narrow bandwidth of their filters in this frequency region. This narrow bandwidth may not be adequate to extract the SCF effectively within that frequency region. This can explain the lower performance of the SCF features computed using the designed filterbanks. This conclusion is partly supported by the study reported in [107], indicating that the frequency modulation features, which are also the frequency-based features, that are extracted from a very narrow bandwidth filter are not effective for speaker recognition.

Similar to the investigation on the cepstral coefficients presented in Section 5.4.2, the Hertz filterbank consistently provided the lowest performance. This is most probably because this filterbank has the least number of filters in the frequency region from 0 Hz to 1.5 kHz and therefore does not capture the CL information effectively.

The higher frequency resolution in the low frequency region and the outperformance of the designed filterbanks in this section and in Section 5.4.2 compared to the other existing filterbanks suggest that by emphasizing the speech features in the low frequency region, which contain significant amount of cognitive load information, the performance of the cognitive load classification system can be improved.

The performance of the classification system was improved in Chapter 4 by assigning a larger weight to the lower frequency subbands for emphasizing the speech features extracted from these subbands. In this chapter, the system performance is improved by increasing the number of filters in the lower frequency subbands. Both of them re-iterate the fact that the CL information is concentrated in lower subbands.

5.4.4 Performance of designed filterbanks in noisy conditions

In this section, the effectiveness of the designed filterbanks for cognitive load classification in noisy conditions is investigated. Noisy speech is generated by adding noise from the NOISEX-92 dataset to clean speech at five levels of signal to noise ratio (SNR): 0, 5, 10, 15 and 20 dB. A subset of seven noises from NOISEX-92 is used in this study: babble, pink, white, leopard, factory, F16 and buccaneer. Due to time constraints, all the experiments in this study were carried out on the Stroop test corpus only. The classification system used in this study is same as the full-band system described in Section 4.3.2.

Table 5.8 shows the average accuracy computed across all signal to noise ratios for different noise types of the classification system using cepstral coefficients extracted using the designed filterbank (Section 5.4.2.1) and the other filterbanks. It can be seen from this table that among all the filterbanks used, the designed filterbank provides the highest performance for the system under the effect of four of the seven noise types tested namely pink, leopard, factory and buccaneer. Furthermore for white and F16 noise, the proposed filterbank is the second best filterbank for the classification. Computed for all noisy conditions, i.e. seven noise types and five SNRs, the cepstral coefficients extracted using the designed filterbank provide the highest performance, compared to all the filterbanks tested. As the designed filterbank emphasizes the speech features in the low frequency region compared to the other filterbanks, these observations indicate that even in noisy conditions, the speech features in the low frequency region are very important for cognitive load classification.

Table 5.8: Average accuracy of cepstral coefficients in noisy conditions (maximums for individual noise types in bold).

	Accuracy (%)							
Noise types / Filterbanks	Pink	White	Leopard	Factory	F16	Buccaneer	Babble	Average
Designed	**65.2**	61.0	**82.3**	**59.8**	61.9	**61.2**	64.9	**65.2**
Mel	64.0	58.4	78.0	58.7	**64.7**	60.0	66.3	64.3
Bark	63.8	**62.6**	80.1	58.7	61.7	61.1	**67.3**	65.0
ERB	63.5	57.6	81.3	59.5	60.4	58.0	65.0	63.6
Hertz	53.6	49.5	70.7	49.0	57.3	50.3	55.5	55.1

Table 5.9: Average accuracy of the fusion of SCF-based and SCA-based systems in noisy conditions (maximums for individual noise types in bold).

	Accuracy (%)							
Noise / Filterbanks	Pink	White	Leopard	Factory	F16	Buccaneer	Babble	Average
Designed	63.3	56.3	**85.9**	63.2	65.6	59.4	**73.6**	66.8
Mel	64.1	58.5	83.7	62.6	63.2	59.4	71.0	66.1
Bark	**67.2**	60.4	82.6	64.6	**67.1**	63.7	71.7	68.1
ERB	**67.2**	**61.9**	84.6	**65.4**	66.4	**64.0**	71.3	**68.7**
Hertz	56.7	49.9	76.5	55.1	57.8	54.9	60.8	58.8

The average accuracy across all SNRs of the fusion of SCF-based and SCA-based systems, where the SCF and SCA features are extracted using the filterbank designed for the Stroop test corpus (Section 5.4.3.1) and the other filterbanks is shown in Table 5.9. It can be seen from this table that among all the filterbanks tested, the designed filterbank yields highest performance under the effect of the leopard and babble noise. However, it produces lower performance for other noise types. Averaged over all noisy conditions, the designed filterbank produces a higher performance than the mel and Hertz filterbanks but has a lower performance than the Bark and ERB filterbanks.

5.5 Summary

This chapter has investigated the effect of varying the feature dimensions of the spectral features (SCF, SCA, and MFCC) on the performance of the cognitive load classification system. It was found that spectral features with six dimensions produce the highest performance. Furthermore, a very small or a very large feature dimension was found to degrade the performance of the system, which is partly explained through an analysis of correlation of adjacent subbands of the spectral centroid features.

The analysis of cognitive load information distribution using cepstral coefficients, SCF, SCA and energy at the feature, model and classification stages has consistently revealed that cognitive load information is concentrated in the frequency region from 0 Hz to 1.5 kHz, reaching a peak in the region from 400 Hz to 1 kHz. Beyond 1 kHz the amount of cognitive load information contained in individual subband decreases with respect to frequency. This implies that an effective cognitive load classification system needs to emphasize the speech features in the frequency region (0-1.5) kHz.

Different filterbanks were designed for each corpus to extract the cepstral coefficients and spectral centroid features for cognitive load classification by allocating the center frequencies and bandwidths of their filters in such a way as to have high frequency resolution around (0-1.5) kHz. In clean conditions, it was found that the designed filterbanks consistently provided a higher performance than three traditional perceptual filterbanks mel, Bark and ERB and the Hertz filterbank for the systems based on cepstral coefficients, spectral centroid amplitude feature, and fusion of spectral centroid frequency and spectral centroid amplitude features, evaluated on both corpora. In particular, when evaluated on the Reading and Comprehension corpus, the cepstral coefficients computed with the designed filterbank provide a 18.6%, 17% and 9.4% relative error rate reduction compared to the mel, Bark and ERB filterbanks respectively. The corresponding relative error rate reductions of the designed filterbank for the fusion system of SCF and SCA are 10.3%, 14.9% and 17.1%. The Hertz filterbank is found to consistently produce the lowest performance for the system.

The success of the designed filterbanks over the other filterbanks indicates that having more filters in the low frequency region can emphasize the cognitive load information in this region and as such improves the performance of the classification system. Although this method of emphasizing the low frequency region is different from how it is emphasized in Chapter 4, where a larger weight is used, in both of these chapters the system performance is improved by emphasizing the low frequency region. These consistent observations strongly suggest that the low frequency region is very important for cognitive load classification.

Furthermore, in noisy conditions, evaluated on the Stroop test corpus, the designed filterbank for extracting the cepstral coefficients is found to produce higher performance for the classification system than the mel, Bark, ERB, and Hertz filterbanks. In addition to this, the filterbank designed to extract the spectral centroid features is found to be more effective than the mel and Hertz filterbanks but less effective than the Bark and ERB filterbanks for the fusion of the SCF-based and SCA-based systems.

Chapter 6: Speech enhancement for cognitive load classification

6.1 Introduction

The performance of the cognitive load classification system significantly degrades in noisy conditions and the system becomes less applicable for industrial implementation. It is therefore necessary to study techniques that reduce the effect of noise. One potential method to achieve this is to employ speech enhancement techniques to preprocess noisy speech before feeding it into the system. It has been shown that the use of speech enhancement can increase the performance of the systems under noisy conditions in speech and speaker recognition applications [108-109]. Furthermore, speech enhancement is also useful in voice communication and hearing aids [110-111].

Due to its wide range of applications, speech enhancement has attracted a large number of researchers studying this area over the last few decades. Many techniques have been introduced for speech enhancement such as spectral subtraction [112], Wiener filtering [113-114], statistical-models [115-116], and Kalman filtering [117-118]. Despite the availability of a large number of methods, speech enhancement is still a challenging problem due to the requirement for very high quality speech in voice communication systems.

Among the methods that have been introduced for speech enhancement, those based on Kalman filtering are known to produce low musical tone and less distortion for enhanced speech [119]. Previous studies apply Kalman filtering in full-band noisy speech [117-118]. In this study, we propose a subband Kalman filtering method where the Kalman filtering is applied to subband speech.

Empirical mode decomposition (EMD) has recently been developed as a tool for the analysis of non-stationary signal [120]. Empirical mode decomposition decomposes any signal into zero mean oscillating components, known as intrinsic mode functions. It has been shown previously that when white noise contaminated speech is decomposed by empirical mode decomposition, the speech and noise components are separated reasonably well [121]. In other words, speech dominates in some intrinsic mode functions, while noise dominates in the others. Hence in this study, we propose a speech enhancement method by applying weighting functions in individual intrinsic mode functions that are subject to the distribution of speech and noise.

In addition, the methods based on the discrete cosine transform (DCT) are known to be effective due to the high energy compaction ability of DCT [122]. A simple but effective method to enhance speech is to apply soft thresholding on noisy speech in order to suppress noise in the DCT domain [123-124]. In this method, a noisy speech frame in the DCT domain is split into a number of subframes. These subframes are categorized as either a signal-dominant subframe or a noise-dominant subframe. In previous studies, the thresholding process is applied only to noise-dominant subframes [123-124]. This can result in a large amount of noise remaining in the enhanced speech as the signal-dominant subframes are not de-noised. In this study, an improved soft thresholding method is proposed by applying the appropriate thresholds for both noise-dominant and signal-dominant subframes.

This chapter initially proposes two novel speech enhancement methods: a non-uniform subband Kalman filtering method and an empirical mode decomposition based method. It then proposes an approach to improve the existing soft thresholding for DCT based speech enhancement method. The effectiveness of each of these methods is compared with other traditional speech enhancement methods using the objective measure of perceptual evaluation of speech quality (PESQ). Their effectiveness is then compared to each other in terms of their PESQ and processing time. Finally, the most suitable method for cognitive load classification is chosen and its usefulness in improving the performance of the cognitive load classification system under noisy conditions is investigated.

6.2 Proposed speech enhancement methods

The quality of enhanced speech can be evaluated by using either subjective or objective measures. Subjective measures are obtained from the human listening tests to estimate the speech quality based on rating scales [125]. Although this measure is very reliable, using it to evaluate speech quality is expensive and time consuming. As such, objective measures are often utilized to judge speech quality. Many objective measures have been introduced such as Itakura-Saito distortion, Articulation Index, signal to noise ratio (SNR), segmental SNR, and PESQ [125]. Among these, PESQ has the highest correlation to the subjective measure and has been widely used to evaluate enhanced speech [125]. PESQ scores range from 4.5 for highest quality of speech down to -0.5 for lowest quality of speech. In this section, the performance of speech enhancement methods is evaluated based on the relative improvement of the PESQ of the enhanced speech ($PESQ_{enhanced}$) from that of the noisy speech ($PESQ_{noisy}$), expressed as [126]

$$\delta = \frac{PESQ_{enhanced} - PESQ_{noisy}}{PESQ_{noisy}} \times 100\% \tag{6.1}$$

All results reported in this section are obtained by carrying out the speech enhancement experiments on noisy speech, which is obtained by adding noise from the NOISEX-92 noise dataset [102] to clean speech from the EBU SQAM speech dataset [127]. This is done at five different SNRs i.e. 0, 5, 10, 15 and 20 dB. The speech dataset contains six speech files of six speakers, three of whom are female and three male, sampled at 8 kHz. The lengths of the files are between 17 and 20 seconds. A subset of seven noise types from NOISEX-92 noise dataset are used in this study, namely babble, pink, white, leopard, factory, F16, and buccaneer.

6.2.1 Kalman filtering method

6.2.1.1 Kalman filtering for speech enhancement

The noisy speech is expressed as

$$x(n) = s(n) + v(n) \tag{6.2}$$

where $x(n)$, $s(n)$, and $v(n)$ are the noisy speech, clean speech, and noise signals respectively. The Kalman filtering technique assumes that clean speech can be described as an autoregressive (AR) process in which each speech sample is considered as the output of an all-pole linear system driven by an excitation signal $\omega(n)$, which is a zero-mean white Gaussian process with variance σ_ω^2:

$$s(n+1) = \sum_{i=1}^{p} a_i s(n+1-i) + \omega(n) \tag{6.3}$$

The noise is also assumed to be an AR process, expressed as

$$v(n+1) = \sum_{i=1}^{q} b_i v(n+1-i) + u(n) \tag{6.4}$$

wherein $u(n)$ is a white Gaussian process with variance σ_u^2. Let $\boldsymbol{s}(n) = [s(n-p+1) \ldots s(n-1)\ s(n)]^T$, $\boldsymbol{v}(n) = [v(n-q+1) \ldots v(n-1)\ v(n)]^T$, and the AR parameters $\boldsymbol{a}(n) = [a_p \ldots a_2\ a_1]^T$, $\boldsymbol{b}(n) = [b_q \ldots b_2\ b_1]^T$. The equations (6.2), (6.3), and (6.4) can be reformulated in the form of a Kalman filtering process equation and measurement equation in the state space domain as follows

$$\bar{\boldsymbol{s}}(n+1) = \bar{\boldsymbol{F}}(n+1)\bar{\boldsymbol{s}}(n) + \bar{\boldsymbol{g}}\bar{\boldsymbol{\omega}}(n) \tag{6.5}$$

$$x(n) = \bar{\boldsymbol{C}}^{\boldsymbol{T}}\bar{\boldsymbol{s}}(n) \tag{6.6}$$

where

$$\bar{\boldsymbol{s}}(n) = \begin{bmatrix} \boldsymbol{s}(n) \\ \boldsymbol{v}(n) \end{bmatrix}, \ \bar{\boldsymbol{\omega}}(n) = \begin{bmatrix} \omega(n) \\ u(n) \end{bmatrix}, \ \bar{\boldsymbol{C}} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{C}_v \end{bmatrix} \tag{6.7}$$

$$\bar{\boldsymbol{F}}(n+1) = \begin{bmatrix} \boldsymbol{F}(n+1) & 0 \\ 0 & \boldsymbol{F}_v(n+1) \end{bmatrix}, \bar{\boldsymbol{g}} = \begin{bmatrix} \boldsymbol{g} & 0 \\ 0 & \boldsymbol{g}_v \end{bmatrix} \tag{6.8}$$

with

$$\boldsymbol{F}(n+1) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ a_p & a_{p-1} & \dots & a_1 \end{bmatrix}_{p \times p}, \quad \boldsymbol{g} = \boldsymbol{C} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{p \times 1} \tag{6.9}$$

$$\boldsymbol{F}_v(n+1) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ b_q & b_{q-1} & \dots & b_1 \end{bmatrix}_{q \times q}, \quad \boldsymbol{g}_v = \boldsymbol{C}_v = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{q \times 1} \tag{6.10}$$

The enhanced speech $\hat{s}(n)$ is obtained by Kalman filtering as below

$$\boldsymbol{G}(n) = \bar{\boldsymbol{F}}(n+1)\boldsymbol{K}(n)\bar{\boldsymbol{C}}[\bar{\boldsymbol{C}}^{\boldsymbol{T}}\boldsymbol{K}(n)\bar{\boldsymbol{C}}]^{-1} \tag{6.11}$$

$$\alpha(n) = x(n) - \bar{\boldsymbol{C}}^T\bar{\bar{\boldsymbol{s}}}(n|\boldsymbol{x}_{n-1}) \tag{6.12}$$

$$\bar{\bar{\boldsymbol{s}}}(n+1|\boldsymbol{x}_n) = \bar{\boldsymbol{F}}(n+1)\bar{\bar{\boldsymbol{s}}}(n|\boldsymbol{x}_{n-1}) + \boldsymbol{G}(n)\,\alpha(n) \tag{6.13}$$

$$\breve{\boldsymbol{K}}(n) = \boldsymbol{K}(n) - \bar{\boldsymbol{F}}(n+1)^{-1}\boldsymbol{G}(n)\bar{\boldsymbol{C}}^T\boldsymbol{K}(n) \tag{6.14}$$

$$\boldsymbol{K}(n+1) = \bar{\boldsymbol{F}}(n+1)\breve{\boldsymbol{K}}(n)\bar{\boldsymbol{F}}(n+1)^H + \boldsymbol{Q} \tag{6.15}$$

The enhanced speech signal is the output of Kalman filtering after the prediction estimation.

$$\bar{\bar{\boldsymbol{s}}}(n|\boldsymbol{x}_n) = \bar{\boldsymbol{F}}(n+1,n)^{-\boldsymbol{1}}\bar{\bar{\boldsymbol{s}}}(n+1|\boldsymbol{x}_n) \tag{6.16}$$

$$\hat{s}(n) = \boldsymbol{C}_1^T\bar{\bar{\boldsymbol{s}}}(n|\boldsymbol{x}_n), \boldsymbol{C}_{\boldsymbol{1}} = \begin{bmatrix} \boldsymbol{C}^{\boldsymbol{T}} \ \underbrace{0 \dots 0}_{q} \end{bmatrix}^{\boldsymbol{T}} \tag{6.17}$$

wherein $\hat{s}(n)$ is the enhanced speech, $\boldsymbol{G}(n)$ is the Kalman gain and $\boldsymbol{K}(n) = E\{[\boldsymbol{s}(n) - \hat{\boldsymbol{s}}(n|\boldsymbol{x}_{\boldsymbol{n-1}})][\ \boldsymbol{s}(n) - \hat{\boldsymbol{s}}(n|\boldsymbol{x}_{\boldsymbol{n-1}})]^T\}$ is the predicted state-error correlation matrix. $\boldsymbol{Q}$ is a sparse matrix with only two nonzero elements. That is $\boldsymbol{Q}(p,q) = \sigma_\omega^2$ and $\boldsymbol{Q}(p+q,p+q) = \sigma_u^2$.

6.2.1.2 Traditional full-band Kalman filtering method

Traditional speech enhancement methods apply Kalman filtering to the full-band speech [117] and hence it is referred to as full-band Kalman filtering in this thesis. In order to enhance speech using the Kalman filtering method, the coefficients of the autoregressive model, $\boldsymbol{a}(n) = [a_p \dots a_2\ a_1]^T$ need to be estimated in advance. The quality of the enhanced speech and the computation complexity of the method based on Kalman filtering are dependent on the order of this model, p. In fact, the order p determines how accurately an autoregressive linear prediction model describes the spectral envelope of a signal. An example of the spectral magnitude and spectral envelope of a 25 ms speech segment with p = 3 and p = 10 is given in Figure 6.1. It can be seen from this figure that the spectral envelope with larger order describes the spectral magnitude more accurately.

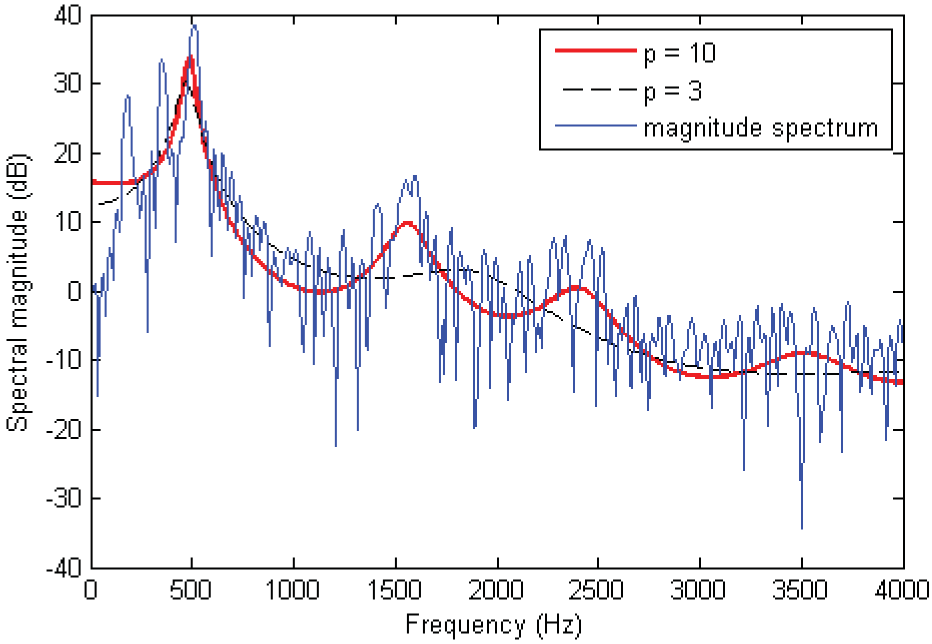


Figure 6.1: Magnitude spectrum of a speech segment and magnitude response of its AR models with different orders.

6.2.1.3 Proposed non-uniform subband Kalman filtering

It is well-known that the frequency resolution of human hearing system is non-uniform. This is usually described by the critical band or Bark scale. Based on the this property, some auditory filterbank models have been extensively researched and are used in subband speech enhancement [128-130]. In this section, a non-uniform subband Kalman filtering method is proposed. In this method, the speech signal is initially decomposed into different subbands using a gammatone filterbank whose frequency response matches the frequency response of the critical band model of the human auditory system [131]. Kalman filtering is then applied to individual subband signals independently. The final enhanced speech signal is obtained by combining the enhanced subband speech signals. As the subband spectral envelope has less variation than the full-band envelope, lower-order AR models are sufficient and hence only lower-order Kalman filtering will be required. This can help to reduce the computational complexity of the proposed non-uniform subband Kalman filtering method as depicted in Figure 6.2.

The noisy speech is initially decomposed into subbands, using a filterbank with M analysis filters. Assuming that the m^{th} analysis filter has impulse response h_m, then the subband speech of the m^{th} band is

$$x_m = x * h_m \tag{6.18}$$

where the symbol ‘*’ represents the convolution operator.

The Kalman algorithm is then applied to subband speech x_m. The outputs of the Kalman filters $\hat{s}_m$ are then passed through the corresponding synthesis filter g_m to obtain the reconstructed subband signal u_m. The final enhanced speech signal is then obtained by combining the reconstructed signals in all the subbands, given as:

$$\hat{s}(n) = \sum_{m=1}^{M} u_m(n) \tag{6.19}$$

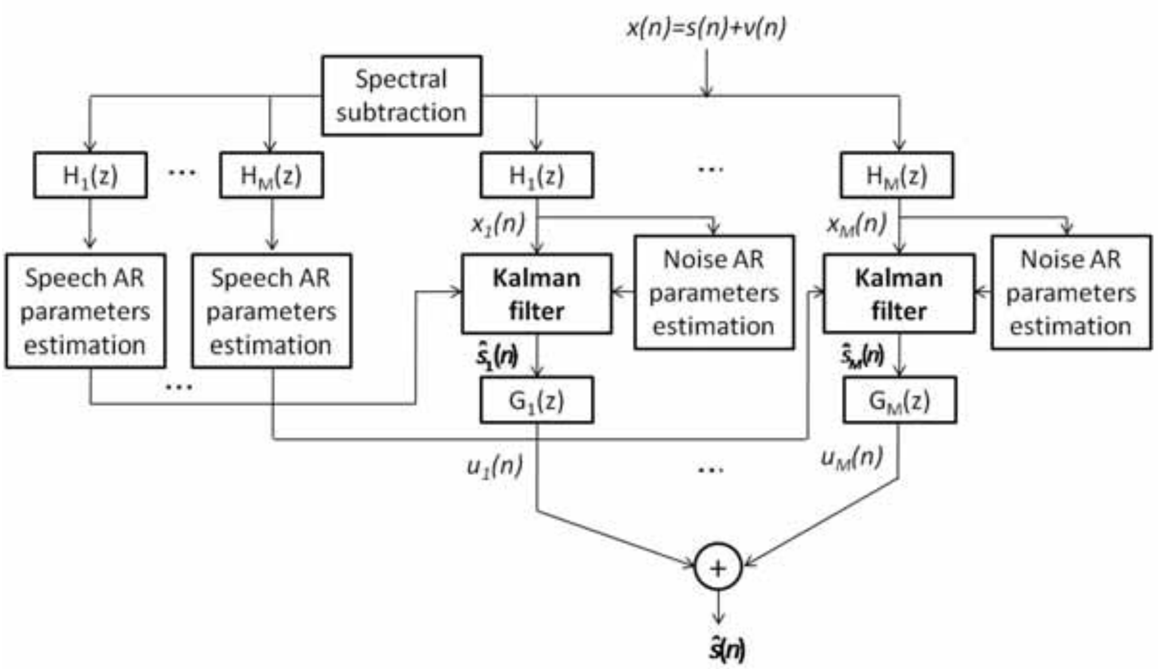


Figure 6.2: Diagram of the proposed subband Kalman filtering method.

The gammatone filters, which are used as analysis filters, are implemented using FIR filters. The analysis filter for the m^{th} subband is obtained using the following expression,

$$h_m(n) = a_m(nT)^{N-1} e^{-2\pi b BW_m nT} \cos(2\pi b f_{cm} nT) \tag{6.20}$$

where f_{cm} is the center frequency of the m^{th} subband, T is the sampling period, n is the discrete time sampling index and BW_m is the bandwidth of the m^{th} filter. The constant b = 1.65 and values for a_m were selected for each filter such that the filter gain was normalized to 0 dB. The number of subbands M was chosen as 18 for a sampling rate of 8 kHz. In order to achieve perfect reconstruction and linear phase characteristics, the synthesis filters are designed as time-reversed impulse responses of the corresponding analysis filters, i.e. $g_m(n) = h_m(-n)$.

In this study, all of the autoregressive (AR) parameters of speech are estimated from the pre-enhanced speech, which is obtained by partly removing noise in noisy speech using the spectral subtraction method [112], as seen in Figure 6.2. This reduces the noise level in the noisy speech which in turn reduces the estimation error of the estimated parameters. The AR parameters of the proposed subband method are estimated in each subband as Kalman filtering is applied individually to each subband. Given a 25 ms frame of speech that overlaps its neighbor by 15 ms, the autoregressive parameters of speech $\boldsymbol{a}(n) = [a_p \dots a_2\ a_1]^T$for each subband are estimated using the well established Levinson-Durbin algorithm [132]. This algorithm is also used to estimate the AR parameters of the noise $\boldsymbol{b}(n) = [b_q \dots b_2\ b_1]^T$. This is performed in the frames of non-speech segments. In this work, the orders of the full-band autoregressive model of speech and noise are $(p,\ q)$ = (10, 5) as per [118]. The performance of the traditional full-band method is used as a reference to evaluate the performance of the proposed non-uniform subband Kalman filtering method. The orders of the subband AR models of speech and noise used for this are $(p,\ q)$ = (3, 2), which are significantly less than those of the corresponding full-band models. This has the added benefit of reducing the computation complexity of the proposed subband Kalman filtering method.

The average relative improvement in terms of PESQ (δ) across all SNRs tested, i.e. 0, 5, 10, 15, and 20 dB of the enhanced speech using the traditional full-band and proposed subband Kalman filtering methods are reported in Figure 6.3.

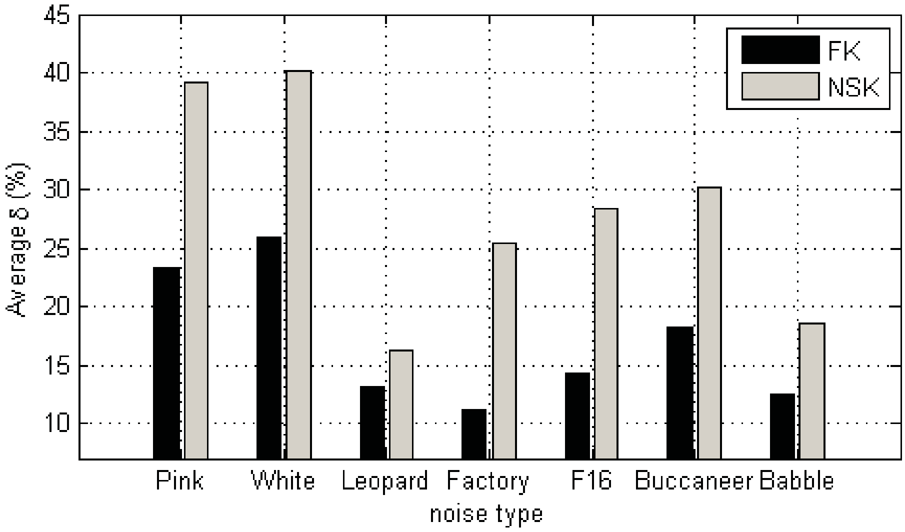


Figure 6.3: Average δ of full-band (FK) and subband Kalman filtering (NSK) methods.

It can be seen from Figure 6.3 that the proposed subband method consistently provides a significantly higher δ than the traditional full-band method. It provides an average improvement of δ over all noise types of 11.4% and a maximum improvement of 15.9% for pink noise compared to the full-band method.

6.2.2 Empirical mode decomposition based method

6.2.2.1 Empirical mode decomposition

Empirical mode decomposition (EMD) was recently pioneered by Huang et. al [120] as a new and powerful data analysis method for non-stationary signals. It is a data-adaptive decomposition method in which any complicated signal can be decomposed into zero mean oscillating components, named intrinsic mode functions (IMFs). These IMFs are signals satisfying two conditions: (1) the number of extrema and the number of zero crossings in the whole signal must differ at most by one; and (2) the mean value of the envelope defined by the local maxima and the envelope defined by the local minima at any point is zero [120]. The second condition modifies the classical global requirement of the first condition to a local one. The sifting process described in [120] was used in this thesis to obtain the IMFs. The sifting process of a signal $x(n)$ is conducted by the following steps:

1. Identify the extrema of $x(n)$, both maxima and minima.
2. Generate the upper envelope $u(n)$ and lower envelope $v(n)$ from the maxima and minima points of $x(n)$ by applying cubic spline interpolation.
3. Determine the mean envelope
$$m_1(n) = (u(n) + v(n))/2$$
4. Determine the new series $h_1(n)$ by removing the low frequency component $m_1(n)$ from signal $x(n)$
$$h_1(n) = x(n) - m_1(n).$$
5. Check if $m_1(n)$ is approximately zero. If so, $h_1(n)$ is the first intrinsic mode function. Otherwise use $h_1(n)$ as a new data set replacing $x(n)$ and repeat steps 1-5 until ending up with an intrinsic mode function.

Once the first intrinsic mode function $h_1(n)$ is derived, $C_1(n) = h_1(n)$, the corresponding residue containing the information about the components of longer periods is determined as:

$r_1(n) = x(n) - C_1(n).$

This residue is considered as a new signal and subject to another sifting process. The procedure is repeated for subsequent residues until $r_N(n)$ is less than a predetermined threshold or is a monotonic function from which no more intrinsic mode functions can be derived. At the end of the decomposition process, the data $x(n)$ will be represented as a sum of N intrinsic mode functions and the last residue signal.

$$x(n) = \sum_{i=1}^{N} C_i(n) + r_N(n) \tag{6.21}$$

Figure 6.6 in Section 6.2.2.2 shows a diagram of empirical mode decomposition by sifting, combined with the process of the proposed speech enhancement method based on empirical mode decomposition.

6.2.2.2 Proposed speech enhancement method based on empirical mode decomposition

The procedure of estimating the intrinsic mode functions presented in Section 6.2.2.1 can be described as a step by step process of subtracting the highest oscillating components. Therefore, the lower order intrinsic mode functions contain higher frequencies and thus have a smaller time scales. In this study, the time scale refers to the distance between two consecutive points where the signal crosses its mean value. The time scale of noise is significantly smaller than that of speech because noise randomly varies while speech is quasi-periodic. This is illustrated in Figure 6.4 which plots a speech and a white noise segment whose means are zero. The instantaneous time scales at particular moments are shown as the distance between two consecutive points where the signals cross zero. It can be seen from this figure that the average time scale of the speech segment is significantly larger than that of the noise segment.

Due to the above-mentioned characteristics, when noisy speech is decomposed the noise components are mainly centered in the lower order intrinsic mode functions [121, 123]. In the case of a noisy speech contaminated by white noise, it was found that the first two intrinsic mode functions mainly contain noise while the third and the fourth ones mainly contain speech [121]. An example of speech contaminated by white noise at 5 dB and its first four intrinsic mode functions is shown in Figure 6.5a. In the case of speech contaminated by other noise such as pink, factory and F16, there is a lack of studies investigating how speech and noise are distributed in the intrinsic mode functions. It is assumed, however, that noise tends to be distributed mainly in lower order intrinsic mode functions and speech tends to be distributed more in higher order functions. This is because the time scale of noise is smaller than that of the speech. In other words, to some extent empirical mode decomposition makes it possible to separate the high frequency noise from the mainly low frequency speech.

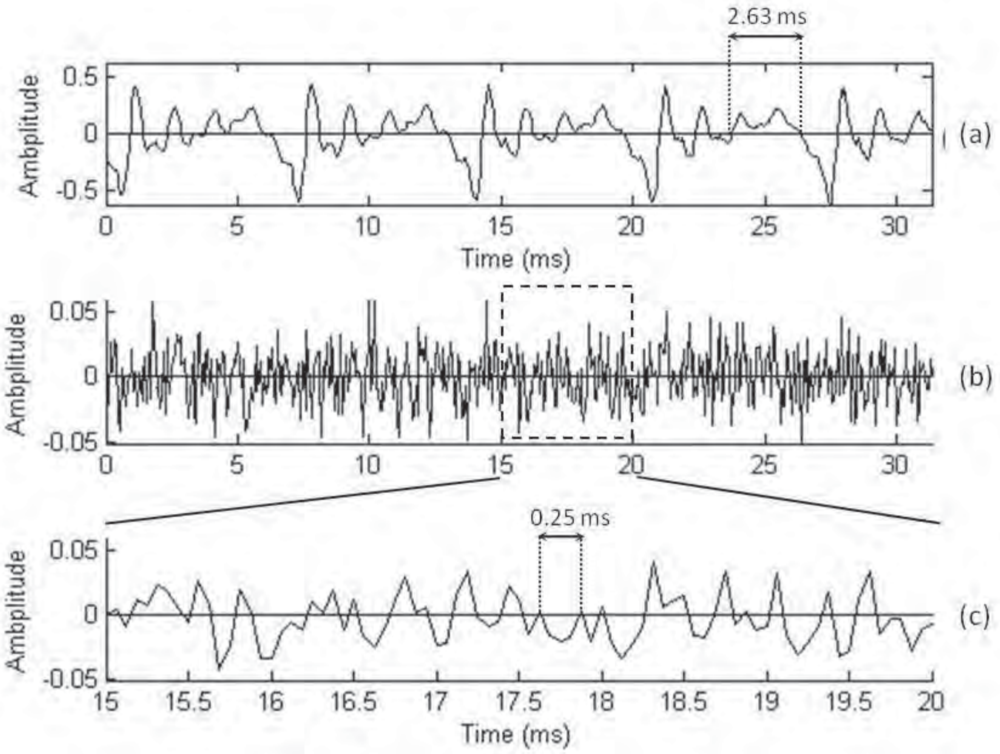


Figure 6.4: An example of (a) speech segment with a time scale of 2.63 ms (b) noise segment and (c) The magnified view of (b) in the region (15-20) ms showing a time scale of 0.25 ms.

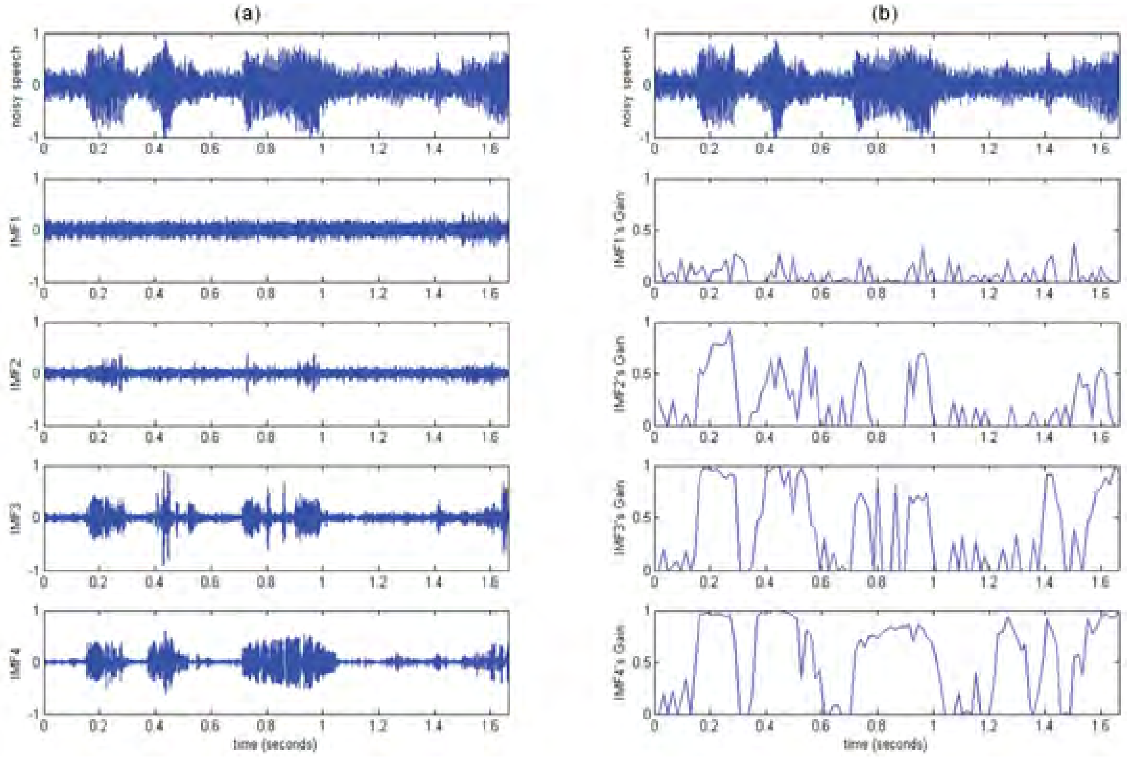


Figure 6.5: (a) Noisy speech and its first four intrinsic mode functions (IMF)
(b) The gains of the first four IMFs.

As discussed, the distributions of speech power and noise power in different order of intrinsic mode functions are very different. We exploit this property in this study by applying a weighting scheme where different weights or gains are applied to different intrinsic mode functions in order to reduce the noise. To achieve this task it is intuitive that a larger gain should be applied to intrinsic mode functions with larger speech power. On the contrary, a smaller gain should be applied to intrinsic mode functions with a smaller speech power. Hence we propose to use the ratio between estimated clean speech power to the estimated noisy speech power computed on each frame of each intrinsic mode functions as the gain as expressed in equation 6.23, in a similar manner to the Wiener filter gain [133]. Let us assume that each intrinsic mode functions (IMF) obtained from the decomposition of noisy speech consists of a clean speech component and a noise component as follow

$$C_i(n) = s_i(n) + v_i(n), \tag{6.22}$$

where C_i is the i^{th} intrinsic mode function of noisy speech; s_i and v_i are the clean speech and noise components of C_i respectively; i = 1,…,*N*, is the IMF index; and *n* is the sample index. *N* is the number of IMFs obtained from the decomposition of noisy speech.

The gain for each intrinsic mode functions is chosen to be

$$k_i(m) = \frac{{\sigma_{si}}^2(m)}{{\sigma_{\mathrm{Ci}}}^2(m)} \tag{6.23}$$

where *m* is the frame index, $\sigma_{Ci}^2(m)$ is the estimated power of the m^{th} frame of the i^{th} IMF of noisy speech, ${\sigma_{si}}^2(m)$ is the estimated power of the clean speech component in the corresponding frame and IMF. Here ${\sigma_{si}}^2(m)$ is estimated by subtracting the noise power from the noisy speech power

$$\sigma_{si}^2(m) = \max\{\sigma_{Ci}^2(m) - \sigma_{vi}^2(m), 0\} \tag{6.24}$$

where $\sigma_{vi}^2(m)$ is the estimated power of the noise component in the m^{th} frame of i^{th} intrinsic mode function. The *max* operation ensures that the estimated clean speech power is always non-negative.

It can be seen from equation (6.23) that the gain $k_i(m) \in [0, 1]$. $k_i(m)$ is zero for a non-speech frame where $\sigma_{si}(m)$ = 0, and $k_i(m)$ is one for a frame where noise does not exist, i.e. $\sigma_{si}(m) = \sigma_{Ci}(m)$.

The weighted intrinsic mode function $\hat{x}_i(n)$ and the residue $\hat{r}_N(n)$ for the m^{th} frame are obtained by weighting the corresponding noisy intrinsic mode function and residue as below

$$\hat{x}_i(n) = k_i(m)\, C_i(n) \tag{6.25}$$

$$\hat{r}_N(n) = k_i(m)\, r_N(n) \tag{6.26}$$

The final enhanced speech signal is obtained by combining the weighted IMFs and the last residue as

$$\hat{s}(n) = \sum_{i=1}^{N} \hat{x}_i(n) + \hat{r}_N(n) \tag{6.27}$$

where N is the number of intrinsic mode functions obtained when performing empirical mode decomposition. The typical value of N used in our experiments is within the range [10, 13]. An outline of the proposed speech enhancement method based on empirical mode decomposition is described in Figure 6.6.

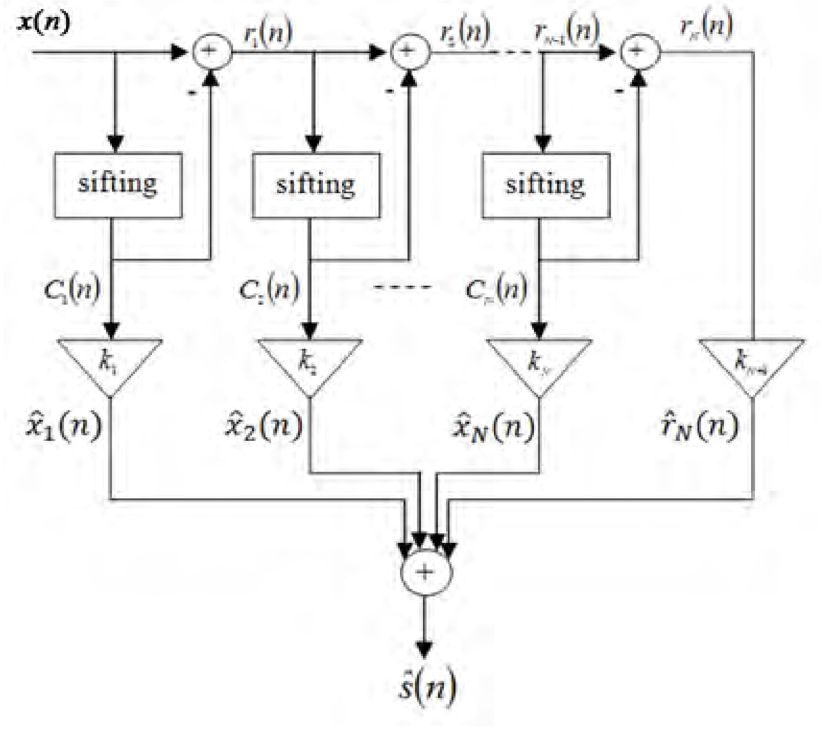


Figure 6.6: Diagram of the proposed empirical mode decomposition method.

An illustration of speech contaminated by white noise at 5 dB and the gain of the first four intrinsic mode functions are shown in Figure 6.5b. It can be seen in this figure that the gain of the first intrinsic mode function is significantly smaller than the gain of the other functions. Furthermore, the gain of the second intrinsic mode function is smaller than those of the third and the fourth functions. As discussed previously, the first and the second intrinsic mode functions mainly contain noise while the third and the fourth intrinsic mode functions mainly contain speech. Assigning small gains to the first and second intrinsic mode functions and large gains to the third and fourth intrinsic mode functions can reduce the noise level in noisy speech.

The effectiveness of the proposed speech enhancement method based on empirical mode decomposition can be seen by observing the waveforms and spectrograms of the clean, noisy, and enhanced speech in Figures 6.7 and 6.8. The speech in this example is contaminated by white noise at 5 dB. It can be seen from Figure 6.7 that most of the noise in the non-speech segments of noisy speech is removed. In addition to this, Figure 6.8 shows that most of the noise in both speech and non-speech segments is removed.

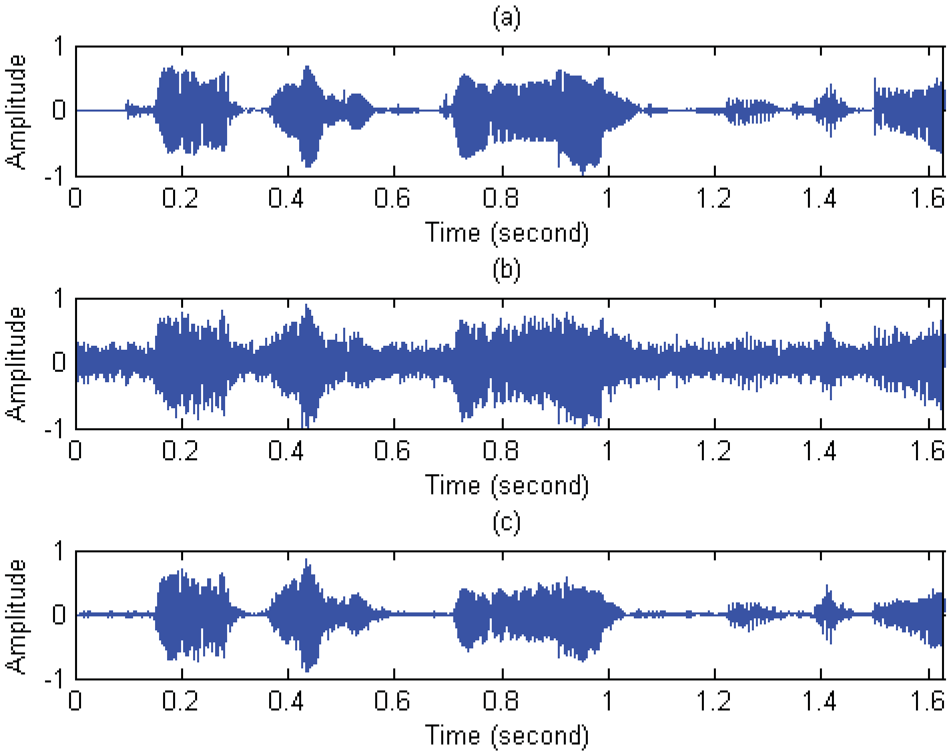


Figure 6.7: The waveforms of (a) clean speech (b) noisy speech and (c) enhanced speech.

The relative improvement of the PESQ (δ) of the enhanced speech based on the proposed empirical mode decomposition method averaged over all SNRs tested is presented in Table 6.1.

Table 6.1: Average δ(%) of the proposed method using empirical mode decomposition

Noise	Pink	White	Leopard	Factory	F16	Buccaneer	Babble	Average
δ(%)	4.3	10.0	4.5	4.7	1.8	9.5	15.7	7.2

It can be seen from Table 6.1 that the proposed speech enhancement method based on empirical mode decomposition consistently produces an improvement of PESQ compared to the noisy speech. In particular, it produces an increase of δ by 7.2% in average and 15.7% in maximum (for speech contaminated by babble noise).

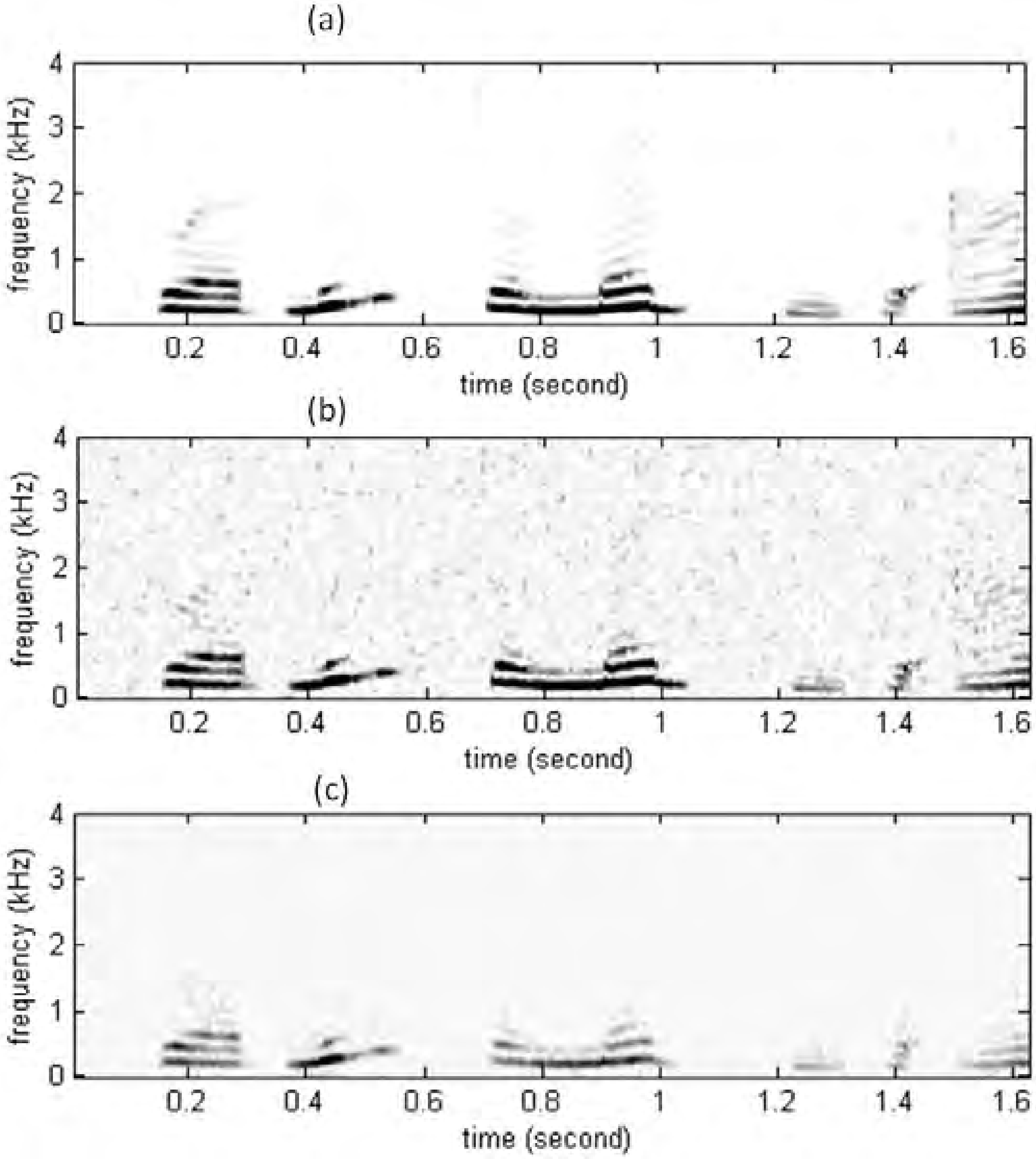


Figure 6.8: The spectrograms of (a) clean speech (b) noisy speech and (c) enhanced speech.

6.2.3 Speech enhancement in DCT domain

In the time domain, a noisy speech signal *x*(*n*) can be considered as a sum of clean speech *s*(*n*) and noise *v*(*n*), as in equation (6.2). For speech enhancement methods based on the discrete cosine transform (DCT), the first step is to convert the speech signal from the time domain to the DCT domain. The k^{th} coefficient of the N-point DCT of signal *x*(*n*) is defined as

$$X(k) = \frac{1}{\sqrt{N}} \mu_k \sum_{n=0}^{N-1} x(n) \cos\left(\frac{\pi}{2N}(2n+1)k\right) \tag{6.28}$$

where $k = 0, \dots, N-1$, $\mu_0 = 1$, $\mu_k = \sqrt{2}$ for $1 \le k \le N-1$.

Noisy speech is expressed in the DCT domain as

$$X(k) = S(k)+V(k) \tag{6.29}$$

where $X(k)$, $S(k)$ and $V(k)$ are the k^{th} DCT coefficients of noisy speech, clean speech and noise signal respectively.

Noisy speech enhancement is performed in the discrete cosine transform domain in order to obtain the DCT coefficient of the enhanced speech $\hat{S}(k)$. The enhanced speech signal in the DCT domain is transformed back to the time domain using the inverse DCT:

$$\hat{s}(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \mu_k \hat{S}(k) \cos\left(\frac{\pi}{2N}(2n+1)k\right) \tag{6.30}$$

6.2.3.1 Traditional soft thresholding method

In traditional soft thresholding method of DCT speech enhancement, the noisy speech signal is segmented into 32 ms frames containing 256 samples. A 512 point DCT is taken on each frame. The 512 DCT coefficients are then divided into 8 subframes consisting of 64 coefficients each as shown in Figure 6.9. Each subframe is categorized to be either signal-dominant or noise-dominant, based on the comparison between signal power and noise power [124]. In particular, a subframe is categorized as a signal-dominant subframe if it satisfies the following condition

$$\frac{1}{64} \sum_{k=1}^{64} \left| X_k \right|^2 \geq \sigma_v^2, \tag{6.31}$$

where σ_v^2 is the noise power, X_k is the k^{th} DCT coefficient in subframe. If this condition is not satisfied the subframe is categorized as noise-dominant.

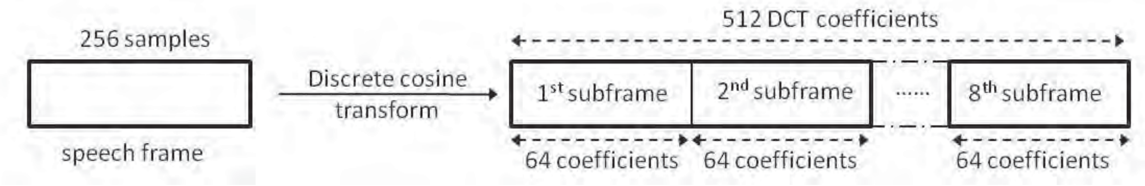


Figure 6.9: An illustration of creating the subframes.

In the analysis of Salahuddin et al. [124], which is performed in the TIMIT database, the absolute values of DCT coefficients in each subframe of noisy and clean speech are sorted in an ascending order. By comparing the sorted absolute DCT coefficients of noisy speech and clean speech they found that for the noise-dominant subframe, the difference between discrete cosine transform coefficients of noisy speech and those of clean speech increases in an approximately linear fashion with respect to the sorted index of the coefficient. Therefore in the traditional soft thresholding method for DCT presented in

[124], the enhanced speech is obtained by applying a linear soft thresholding function to the noisy speech. However for the signal-dominant subframe, it was found that the difference between the discrete cosine transform coefficients of noisy speech and those of clean speech is not significant. Hence in the traditional method, the discrete cosine transform coefficients of the enhanced speech are set to be the same as those of the noisy speech for the signal-dominant subframes. The traditional soft thresholding for DCT speech enhancement can be expressed as

$$\tilde{S}_k = \begin{cases} X_k & \textit{if} \quad \frac{1}{64}\sum_{k=1}^{64}\left|X_k\right|^2 \geq \sigma_v^2 \\ sign(X_k)[\max\{0,(\left|X_k\right| - mj)\}] & \textit{otherwise} \end{cases} \tag{6.32}$$

where $\tilde{S}_k$ is the DCT coefficient of the enhanced speech, j is the sorted index of $|X_k|$ in the subframe, and m is a constant

$$m = \frac{\lambda\,\sigma_v}{\frac{1}{64}\sum_{j=1}^{64} j^2} \tag{6.33}$$

where λ is a constant whose value is 0.8 [124].

6.2.3.2 Proposed improved soft thresholding method

In our analysis conducted on the EBU SQAM and NOISEX-92 datasets, it was observed that for noise-dominant subframes, the difference between the sorted absolute values of DCT coefficients of noisy speech and those of clean speech increases in an approximately linear fashion with respect to the index of the DCT coefficients, as found in [124]. This is illustrated in Figure 6.10a, which plots the average absolute value of DCT coefficients of clean speech and speech contaminated by white noise at 5 dB computed in the noise-dominant subframes across twenty seconds of speech. In our proposed improved soft thresholding method, the DCT coefficients of the enhanced speech are obtained by applying the linear soft thresholding to those of the noise-dominant subframes of noisy speech, in a similar manner to the traditional soft threshold method. For the signal-dominant subframes, it was found that the difference between noisy DCT coefficients and clean DCT coefficients, while less than the different in noise-dominant subframes, is significant. This is illustrated in Figure 6.10b, which shows the average absolute value of DCT coefficients for clean speech and noisy speech at 5 dB, computed in the signal-dominant subframes across twenty seconds of speech. This implies that there is a large amount of noise existing in the signal-dominant subframes. If thresholding is not applied to these signal-dominant subframes, this noise will remain in the resultant enhanced speech, and hence reduce its quality. If thresholding is inappropriately applied, however, the speech signal will be degraded. An appropriate threshold level in signal-dominant subframes should provide a good compromise between noise removal and speech distortion.

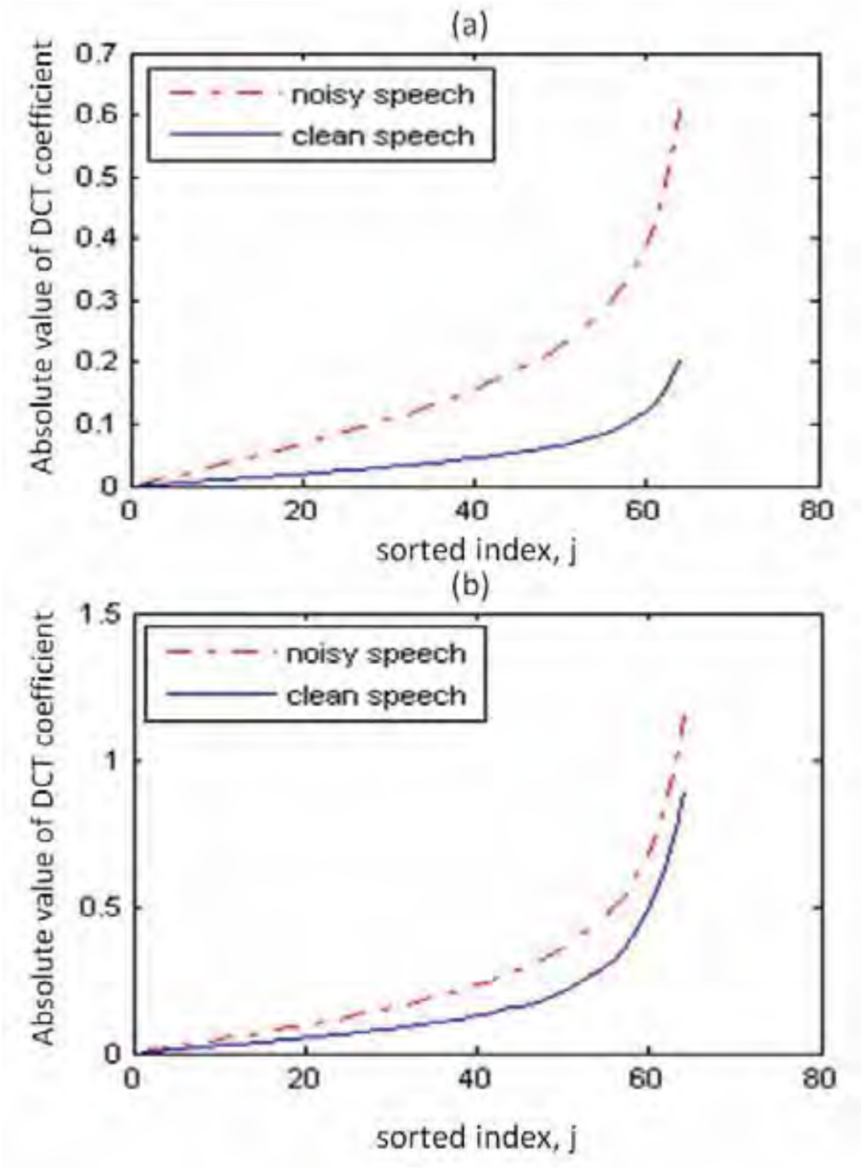


Figure 6.10: Average of the absolute values of DCT coefficients in ascending order of clean and noisy speech of (a) noise-dominant subframes and (b) signal-dominant subframes.

In this study, an improved soft thresholding method is proposed where thresholding is applied to the DCT coefficients in noise-dominant subframes similar to the traditional thresholding method. A soft thresholding is also applied to coefficients in the signal-dominant subframes with a proper threshold level selected so as not to degrade the speech signal. The proposed soft thresholding method is given as follows

$$\tilde{S}_k = \begin{cases} sign(X_k)[\max\{0,(|X_k| - m_s j)\}] & if \quad \frac{1}{64}\sum_{k=1}^{64}|X_k|^2 \geq \sigma_v^2 \\ sign(X_k)[\max\{0,(|X_k| - m_n j)\}] & otherwise \end{cases} \tag{6.34}$$

where m_s and m_n are constants determined as

$$m_s = \frac{\lambda_s \, \sigma_v}{\frac{1}{64}\sum_{j=1}^{64} j^2} \tag{6.35}$$

and
$$m_n = \frac{\lambda_n \, \sigma_v}{\frac{1}{64}\sum_{j=1}^{64} j^2}, \tag{6.36}$$

where λ_s and λ_n are constants that control the amount of noise to be removed from signal-dominant and noise-dominant subframes in the DCT domain respectively.

From equations (6.34 - 36), it can be understood that larger values of λ_s and λ_n will cause more noise being removed and result in more distortion for the speech signal. In this study, λ_n is set to 0.8, according to [123-124]. It is also expected that λ_s should be less than λ_n as the amount of noise that needs to be removed from signal-dominant subframes is less than that from noise-dominant subframes. In this study, the value of λ_s was empirically chosen in order to produce the highest relative improvement of PESQ (δ) for the enhanced speech. The value of λ_s is varied from 0 to 0.8 with the step size of 0.1 and speech enhancement based on the proposed soft thresholding method was carried out for each value of λ_s. The average δ across all SNRs and noise types used of the enhanced speech are graphed against λ_s in Figure 6.11.

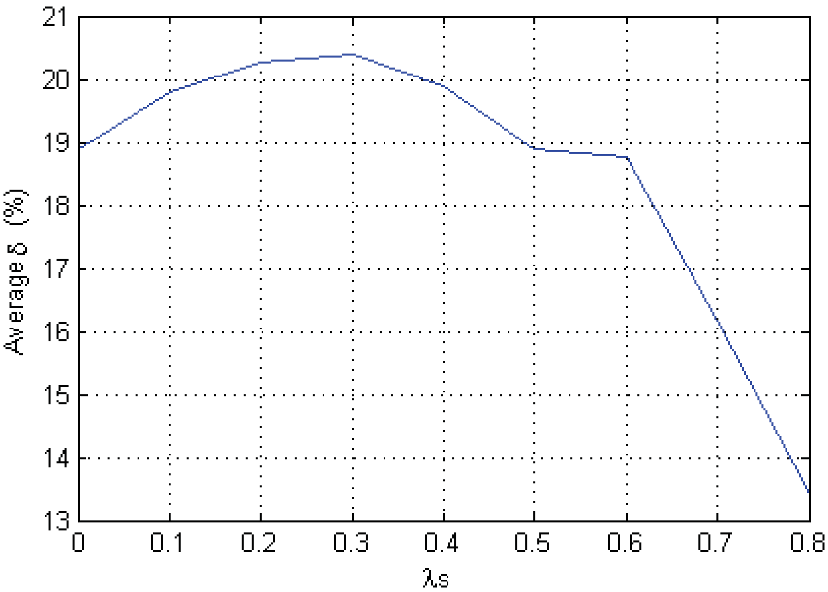


Figure 6.11: Average δ (%) of the proposed thresholding method with various λ_s.

According to this figure, $\lambda_s = 0.3$ produces the highest average δ for the enhanced speech and hence it was chosen to implement the proposed soft thresholding method. The average δ over all SNRs of the enhanced speech using the traditional and proposed improved soft thresholding speech enhancement methods is shown in Figure 6.12.

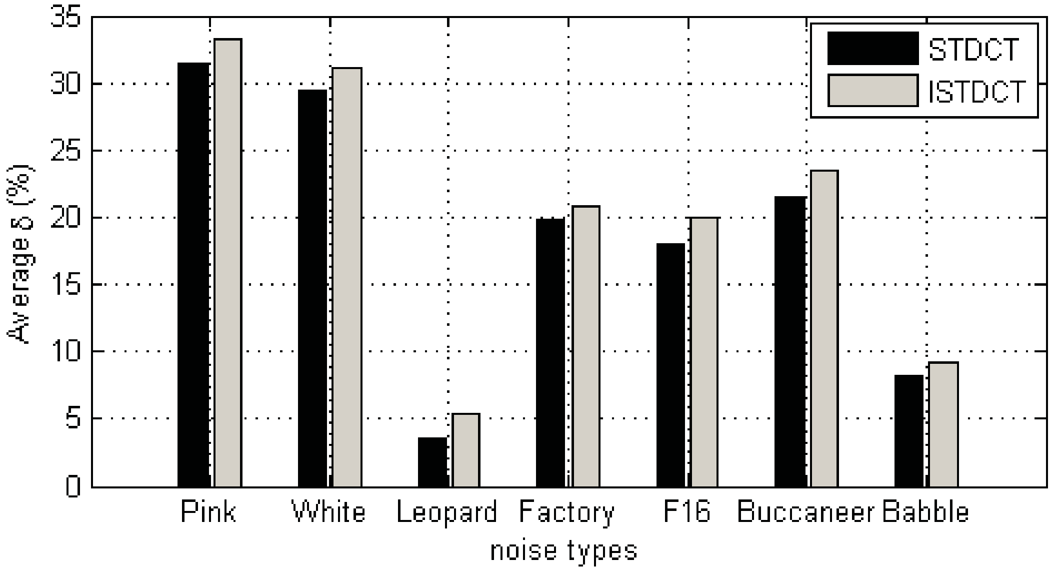


Figure 6.12: Average δ (%) of the traditional soft thresholding DCT (STDCT) and proposed improved soft thresholding DCT (ISTDCT) methods.

It can be seen in Figure 6.12 that the proposed thresholding method consistently provides a higher δ than the traditional thresholding method. The proposed method provides an average δ increase of 1.6% over all noise types used and maximum increase of 2% for buccaneer noise compared to the traditional method.

6.2.4 Comparison of the proposed speech enhancement methods

The effectiveness of a speech enhancement method depends on two important factors, namely the quality of the enhanced speech and the processing time. The purpose of estimating cognitive load (CL) levels is to dynamically adjust the workload imposed on users. Therefore it is necessary to ensure that CL levels are estimated accurately and quickly. This implies that an effective speech enhancement method for cognitive load classification needs to produce high quality enhanced speech in order to improve the accuracy of the system. In addition, the required processing time needs to be short so that the CL level can be recognized in real-time or close to real time.

For the purpose of seeking the best speech enhancement method for a cognitive load classification application, this section compares the effectiveness of the non-uniform subband Kalman filtering method, empirical mode decomposition based method, and improved soft thresholding based on DCT method in terms of the above-mentioned two factors. The average relative improvement of PESQ (δ) under all noisy conditions i.e. under the effect of seven noise types and five signal to noise ratios and processing time as a factor of real time based on Intel Core 2 Duo 2.5 GHz processor of these methods are listed in Table 6.2.

Table 6.2: Average δ(%) and processing time of the three proposed methods.

Methods	δ(%)	Processing time (x real time)
Non-uniform subband Kalman filtering	28.3	203
Empirical mode decomposition	7.2	48.8
Improved soft thresholding DCT	20.4	13.2

It can be seen from Table 6.2 that among the proposed methods, the one based on empirical mode decomposition is the worst method as it produces significantly lower quality of enhanced speech compared to the other two methods and requires considerably longer processing time compared to the improved soft thresholding discrete cosine transform method. Furthermore, the discrete cosine transform method produces 7.9% lower δ than the Kalman filtering method. However, it requires significantly less processing time than the Kalman method. In relative terms, the use of soft thresholding DCT method saves 93.5% processing time compared to the use of subband Kalman filtering method. Taking into account the quality of the enhanced speech and the processing time as discussed above, the improved soft thresholding DCT method seems to be more suitable for the task at hand than the subband Kalman filtering and empirical mode decomposition based methods for the purpose of cognitive load classification.

6.3 Incorporating the thresholding DCT module into CL classification system

This section investigates the effectiveness of the proposed improved soft thresholding DCT speech enhancement method in increasing the performance of the cognitive load classification system under noisy conditions. All the experiments in this section are performed on the Stroop test corpus. Furthermore, only the test speech is noisy and all training data is clean. This is typical in practical scenarios as the training speech can be recorded in relatively noise free conditions. The incorporation of the speech enhancement module to the system is shown in Figure 6.13. The noisy speech in this experiment is obtained by adding the noise from the NOISEX-92 database to the clean speech at five SNR levels namely 0, 5, 10, 15, and 20 dB. A subset of seven types of noise, namely pink, white, leopard, factory, F16, buccaneer, and babble are used in this study. The classification system used in this section is the same as the full-band system described in Section 4.3.2. In addition, the speech features used in all classification experiments in this section are the MFCC features.

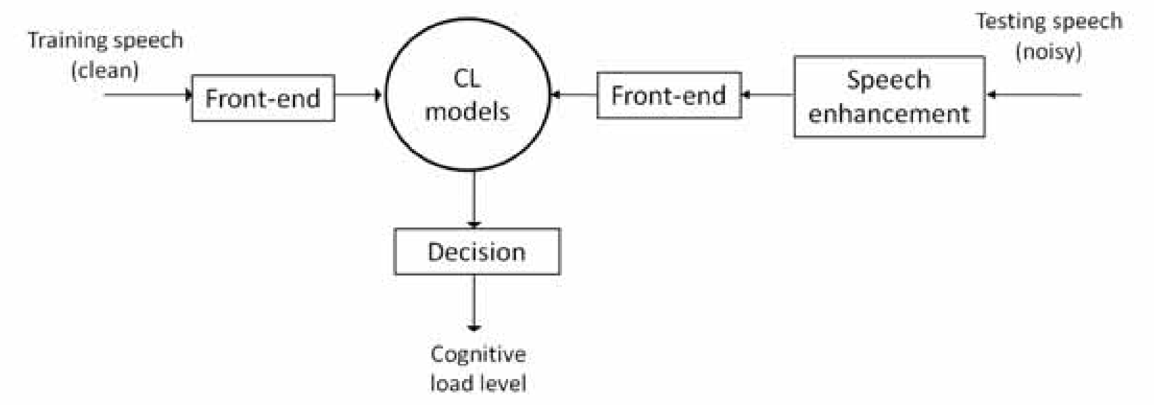


Figure 6.13: Diagram of the system incorporating speech enhancement.

The accuracy of the classification system averaged across all SNRs, under the effect of individual noise types, with and without incorporating the speech enhancement module using the improved soft thresholding DCT method is presented in Table 6.3. The average relative reduction of error rate when using the proposed thresholding DCT speech enhancement method is also presented in this table.

Table 6.3: Average accuracy of the system in noisy conditions over all SNRs.

Noise type	Average accuracy (%)		Average relative reduction of error rate by using DCT speech enhancement (%)
	Without speech enhancement	With DCT speech enhancement	
Pink	58.0	59.0	2.5
White	54.3	55.7	3.1
Leopard	70.9	71.8	3.1
Factory	55.8	56.7	2.3
F16	57.8	59.4	4.3
Buccaneer	53.9	54.9	2.6
Babble	60.1	61.5	3.3
Average	58.7	59.9	3.0

It can be seen from Table 6.3 that the proposed DCT speech enhancement method improves the performance of the classification system under the effect of all noise types tested. It provides an average relative error rate reduction of 3.0% over all tested noise and a maximum relative error rate reduction of 4.3 % for F16 noise.

The average accuracy of the system in noisy conditions across all noise types tested at an individual SNR, with and without using the proposed DCT speech enhancement method, is presented in Table 6.4. The average relative error rate reduction using the improved soft thresholding DCT method is also shown in this table.

Table 6.4: Average accuracy of the system in noisy conditions across all noise types tested.

SNR (dB)	Average accuracy (%)		Average relative reduction of error rate using DCT speech enhancement (%)
	Without speech enhancement	With DCT speech enhancement	
0	43.9	45.3	2.6
5	51.4	52.7	2.8
10	60.9	61.8	2.5
15	67.1	68.2	3.3
20	70.2	71.3	3.9
Average	58.6	59.9	3.0

It can be seen from Table 6.4 that the use of the proposed improved soft thresholding DCT speech enhancement method provides higher accuracy for the classification system in noisy conditions at all levels of SNR tested. The relative error rate reduction when using the DCT speech enhancement method is quite consistent for various SNRs and the maximum relative error rate reduction is 3.9% for a signal to noise ratio of 20 dB.

In addition, the accuracies of the system obtained for each signal to noise ratio and each noise type (appendix A) indicate that the proposed DCT method provides a relative error rate reduction of up to 7.5% with F16 noise at 20 dB SNR.

6.4 Summary

This chapter has proposed three speech enhancement methods. Two of them are novel methods namely the non-uniform subband Kalman filtering and empirical mode decomposition based methods. The third method improves the existing soft thresholding for discrete cosine transform (DCT) speech enhancement method. It was found that the proposed subband Kalman filtering method provides a 11.4% increase of δ, i.e. the relative improvement of PESQ compared to noisy speech, compared to the traditional full-band Kalman filtering method. The proposed empirical mode decomposition based method yields a 7.2% of δ. The proposed improved thresholding DCT method provides a 1.6% increase in δ when compared to the traditional thresholding method.

Among the three proposed methods, the subband Kalman filtering method was found to provide the highest relative improvement of PESQ and the thresholding DCT method was found to provide 7.9% lower δ than the Kalman filtering method. However, in terms of processing time, the use of the thresholding method saves 93.5% processing time compared to the Kalman filtering method. Therefore, the improved soft thresholding discrete cosine transform method was chosen to increase the performance of the cognitive load classification system under noisy conditions. It was indicated that the use of the DCT method provides an average relative error rate reduction of 3.0% for the system under the effect of seven noise types and five levels of SNR tested. In particular, it provided a maximum relative error rate reduction of 7.5% under the effect of F16 noise at 20 dB SNR.

Chapter 7: Conclusion and Future work

7.1 Conclusion

This thesis presents research in automatic cognitive load classification based on speech with the aim of proposing techniques to improve its performance in both clean and noisy conditions. This research provides a number of novel contributions including:

(i) the proposal of the use of new features to complement features in the existing classification systems;

(ii) the investigation of the distribution of cognitive load information across different frequency bands;

(iii) the proposal of the use of the multi-band approach for cognitive load classification and the investigation of different weighting schemes for a multi-band classification system;

(iv) the design of filterbanks specifically for classifying cognitive load;

(v) the proposal of speech enhancement methods to improve the performance of the classification system in noisy conditions.

7.1.1 Implementation of human listening test

In order to validate the supposition that cognitive load information is conveyed in speech and to study what sort of speech cues human use to identify the cognitive load levels, a human listening test was implemented in which the participants were asked to detect the cognitive load levels of speakers by listening their speech. The ability of subjects to accurately classify the cognitive load levels implied that there are cognitive load cues contained in speech. Furthermore, the feedback from the participants in the test indicated that breath pattern, speech rate, the use of filler sounds such as ‘*uh*’ and ’*ah*’, and the intonation of the utterance are the most important cues used to identify the cognitive load levels by humans. The usefulness of the intonation in this test supported the effectiveness of the shifted delta feature of the pitch for the classification system, presented in this thesis as well as in the previous study [3, 5].

7.1.2 The use of spectral based speech features

The study of the effectiveness of different speech features for CL classification was presented in Chapter 3. The speech features were categorized as the source-based features such as pitch, intensity, and SMFCC (source MFCC); filter-based features such as formant frequencies, FMFCC (filter MFCC); and combined features i.e. features relating to both the voice source and the vocal tract filter of human speech production system such as MFCC, group delay (GD) and frequency modulation (FM). It was found that although filter-based features are somewhat better at classifying than source-based features, both of these features are effective for the classification. Therefore an effective classification system needs to utilize both types of features. Furthermore, this suggested that in the source-filter model of human speech production, the filter component is more important than the source component in characterizing the variation of cognitive load.

The use of the spectral centroid features, namely spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) for cognitive load classification was proposed in Section 3.5. It was found that both of these features produced accuracies comparable to the traditional MFCC features. In addition, fusion of either the spectral centroid frequency based system or the spectral centroid amplitude based system with the MFCC-based system consistently outperformed a solely MFCC-based system. In particular, fusion of the spectral centroid frequency based system and the spectral centroid amplitude based system with the MFCC-based system produced a relative reduction in error rate of 8.9% and 31.5% respectively, compared to the MFCC-based system performed on the Stroop test corpus. These results indicated that cognitive load information contained in the spectral centroid features are complementary to that contained in the MFCC features. Therefore the spectral centroid features can be used to provide additional cognitive load information to improve the performance of the traditional MFCC-based system. Among all the features used in Chapter 3, the spectral features namely MFCC, spectral centroid frequency, and spectral centroid amplitude were found to be the most effective and hence were chosen for all further studies reported in this thesis.

7.1.3 Analysis of the distribution of cognitive load information

In order to capture cognitive load information effectively for the purpose of improving the performance of the classification system, it is crucial to know how cognitive load information is distributed across the different frequency bands. The preliminary investigation of cognitive load information distribution in a small number of mel subbands i.e. two bands or three bands, whose bandwidths are equal in the mel frequency scale, indicated that the low frequency mel subband contains significantly more CL information than the high frequency mel subband. Furthermore, the rigorous and systematic investigation of the cognitive load information distribution in a large number of uniform subbands i.e. 20 bands or 32 bands, whose bandwidths are equal in the linear frequency scale presented in Section 5.3 showed that cognitive load information is mainly concentrated in the region approximately from 0 Hz to 1.5 kHz, reaching a peak in (400-1000) Hz. Beyond 1 kHz, the amount of cognitive load information contained in individual subbands decreases with respect to frequency. The results found in this study strongly suggest that spectral information in the frequency region of (0-1.5) kHz needs to be emphasized to improve cognitive load classification results.

7.1.4 Multi-band approach and the effectiveness of weighting schemes

Chapter 4 investigated the effectiveness of the multi-band approach and compared it with that of the traditional full-band approach for cognitive load classification. It was found that the multi-band approaches (both feature combination and likelihood combination) were more effective than the full-band approach in both clean and noisy conditions. In particular when performed on the Stroop test corpus, the 2-band multi-band systems based on likelihood combination with an accuracy weighting scheme and the feature combination method reduced the relative error rate by 9.5% and 17% respectively, compared to the traditional full-band system in clean conditions. The corresponding relative error rate reductions in noisy conditions are 3.9% and 9.9%.

Furthermore, this chapter investigated the effectiveness of different weighting schemes, namely accuracy and SNR weighting schemes, for a multi-band classification system based on likelihood combination. It was found that the accuracy weighting scheme was more effective than the non-weighting scheme in clean conditions and both SNR and non-weighting schemes in noisy conditions. This indicated that the performance of the likelihood combination multi-band classification system can be improved by assigning a larger weight to emphasize the speech features in the low frequency band where cognitive load information is mainly concentrated.

7.1.5 Designing effective filterbanks to extract spectral features

As cognitive load information was found to be mainly concentrated in the low frequency region, a novel filterbank design was proposed specifically for cognitive load classification which emphasizing the spectral information in the low frequency region.

This filterbank was designed to have a high frequency resolution in the low frequency region by allocating a large number of filters in this region. It was shown that the designed filterbank consistently performs better than the existing perceptual filterbanks (mel, Bark and ERB) and the Hertz filterbank for the system based on cepstral coefficients in both clean and noisy conditions.

Furthermore, for the system based on the fusion of the classification results of the spectral centroid frequency based and spectral centroid amplitude based systems, the designed filterbank performed better than the mel, Bark, ERB and Hertz filterbanks in clean conditions. In noisy conditions, the fusion system based on the designed filterbank performed better than those based on the mel and Hertz filterbanks but worse than those based on the Bark and ERB filterbanks.

It was found that spectral features with six dimensions produced the highest performance for the system irrespective of the feature type, and both very high and very low dimensional features degraded the system. Hence, six dimensional spectral features were used in the study of filterbank design.

7.1.6 Proposed speech enhancement methods

Section 6.2 proposed two novel speech enhancement methods, namely the non-uniform subband Kalman filtering and empirical mode decomposition based and one separate approach to improve the existing soft thresholding for discrete cosine transform speech enhancement method. This was done with the aim of reducing the effect of noise in order to improve the robustness of the CL classification system under noisy conditions. It was found that the proposed non-uniform subband method provided on average an improvement of δ, i.e. the relative PESQ improvement, of 11.4% compared to the traditional full-band Kalman filtering method. Furthermore, the proposed empirical mode decomposition based method produced an average relative PESQ improvement of 7.2% compared to the noisy speech. The proposed improved soft thresholding method provided on average an improvement higher relative improvement of PESQ than the traditional soft thresholding method.

Among the three proposed speech enhancement methods, the non-uniform subband Kalman filtering method was found to provide the largest relative improvement of PESQ. The improved soft thresholding method was found to provide slightly less relative PESQ improvement than the non-uniform subband method. However, it saved 93.5% processing time compared to the subband Kalman filtering method. As such, the improved soft thresholding method was chosen to improve the quality of speech and increase the performance of the cognitive load classification system under noisy conditions. It was indicated in Section 6.3 that the use of the proposed improved soft thresholding method based on the discrete cosine transform reduced the relative error rate by 3.0% when averaged over the seven noise types and five SNRs tested. In particular, it reduced the relative error rate by a maximum of 7.5% for the system under the effect of the F16 noise at 20 dB SNR.

7.2 Future work

- This thesis has found that cognitive load (CL) information is mainly concentrated in low frequency region. Two methods are proposed to emphasize the contribution of speech features in this region in order to improve the performance of the system. A related investigation suggested by this is to examine the temporal dependence of cognitive load specific information. An appropriate approach to emphasize the contribution of speech features in segments that is more important for CL classification is expected to further improve the performance of the system.
- The usefulness of the shifted delta features in this study indicates that temporal variation of speech features is very important for classifying cognitive load. However the shifted delta features can only capture the temporal variation to some extent. By developing other techniques to describe the temporal variation of speech feature more effectively, the performance of the cognitive load classification system can be further improved.
- The filterbanks designed in this study perform well with speech sampled at 16 kHz where the signal bandwidth ranges from 0 Hz to 8 kHz. These filterbanks may not be very effective for speech collected through a telephone channel as the bandwidth of this medium is approximately 300 Hz to 3400 Hz. Designing filterbanks for CL classification based on telephone speech is an interesting area for future research as one of the common methods to capture speech for CL classification is from a telephone. Furthermore, the designed filterbanks in this thesis were optimized separately for each of the two databases. An interesting extended study would be to develop a common filterbank for two databases. This may be done by designing a filterbank based on a development dataset which is independent to the traning and test datasets.

- The two databases used in this thesis were collected in a laboratory through a microphone and each of them contains the speech of fifteen speakers. The factors unrelated to cognitive load such as speaker variation and channel mismatch therefore have been minimized or eliminated. In practical scenarios, speech data are usually collected from a very large number of speakers and through many different channel types such as a telephone and microphone and hence the effect of the above-mentioned factors would be much more serious. In other words, there is a gap between the databases used in this study and those collected in more realistic scenarios. Hence, although the techniques proposed in this thesis are promising in terms of improving the performance of the cognitive load classification system, they should be validated by using other databases that contain speech from a larger number of speakers and are collected from different channels.
- This thesis focused only on a single back-end classifier. Alternative classification techniques including Hidden Markov Model (HMM), simple linear kernel Support Vector Machine (SVM), hybrid SVM-GMM which accepts the likelihood scores from GMM as inputs for SVM, and a fusion approach which integrates GMM, SVM and SVM-GMM systems together should hence be explored.

Appendix A

Relative error rate reduction of the cognitive load classification system using the improved soft thresholding discrete cosine transform speech enhancement method

Noise \ SNR	0 dB	5 dB	10 dB	15 dB	20 dB	Average
Pink	3.0	2.0	2.7	3.2	1.6	2.5
White	5.5	1.0	1.1	1.3	6.7	3.1
Leopard	3.0	3.3	1.7	4.4	3.3	3.1
Factory	1.8	1.8	1.2	3.6	3.3	2.3
F16	2.1	1.9	5.4	4.7	7.5	4.3
Buccaneer	1.2	3.2	2.3	2.5	3.6	2.6
Babble	1.8	6.1	3.4	3.7	1.7	3.3
Average	2.6	2.8	2.5	3.3	4.0	3.0

Bibbliography

[1] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. V. Gerven, "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory," *Education Psychologist*, vol. 38, pp. 63-71, 2003.

[2] A. Berthold and A. Jameson, "Interpreting Symptons of Cognitive Load in Speech Input," *in Proc of International Conference on User Modeling*, 1999, pp. 235-244.

[3] H. Bo il, S. O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and Detection of Cognitive Load and Frustration in Drivers' Speech," *in Proc. of Interspeech* Makuhari, Chiba, Japan, 2010, pp. 502-505.

[4] T. F. Yap, J. Epps, E. H. C. Choi, and E. Ambikairajah, "Glottal feature for speech-based cognitive load classification," *in Proc. of ICASSP*, pp. 5234-5237, 2010a.

[5] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," *in Proc. of ICASSP*, 2008, pp. 2041-2044.

[6] H. J. M. Steeneken and J. H. L. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," *in Proc of ICASSP*, pp. 2079-2082, 1999.

[7] J. Sweller, J. J. G. Van Merrienboer, and F. G. W. C. Paas, "Cognitive architecture and instructional design," *Educational psychology review*, vol. 10, pp. 251-296, 1998.

[8] S. Miyake, "Multivariate workload evaluation combining physiological and subjective measures," *International journal of psychophysiology*, vol. 40, pp. 233-238, 2001.

[9] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," vol. 63, pp. 81-97, 1956.

[10] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behavioral and brain sciences*, vol. 24, pp. 87-114, 2001.

[11] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive science*, vol. 12, pp. 257-285, 1988.

[12] R. C. Anderson, "The notion of schemata and the educational enterprise," *Schooling and the acquisition of knowledge*, pp. 415-431, 1977.

[13] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture," *Instructional Science*, vol. 32, pp. 1-8, 2004.

[14] P. Chandler and J. Sweller, "Cognitive load theory and the format of instruction," *Cognition and instruction,* vol. 8, pp. 293-332, 1991.

[15] J. L. Plass, R. Moreno, and R. Brunken, *Cognitive load theory*: Cambridge university press, 2010.

[16] R. Brunken, J. L. Plass, and D. Leutner, "Direct measurement of cognitive load in multimedia learning," *Educational Psychologist,* vol. 38, pp. 53-61, 2003.

[17] F. G. Paas, "Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach," *Journal of educational psychology,* vol. 84, p. 429, 1992.

[18] F. G. W. C. Pass and J. J. G. V. Merrienboer, "Instructional Control of Cognitive Load in the Training of Complex Cognitive Tasks," *Educational Psychology Review,* vol. 6, pp. 351-371, 1994.

[19] S. Kalyuga, P. Chandler, and J. Sweller, "Levels of expertise and instructional design," *Human factors,* vol. 40, 1998.

[20] S. Tindall-Ford, P. Chandler, and J. Sweller, "When Two Sensory Modes Are Better Than One," *Journal of experimental psychology: Applied,* vol. 3, pp. 257-287, 1997.

[21] K. C. Hendy, K. M. Hamilton, and L. N. Landry, "Measuring subjective workload: when is one scale better than many?," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 35, pp. 579-601, 1993.

[22] S. G. Hill, H. P. Iavecchia, J. C. Byers, A. C. Bittner, A. L. Zaklad, and R. E. Christ, "Comparison of four subjective workload rating scales," *Human Factors: The Journal of the Human Factors and Ergonomics Society,* vol. 34, pp. 429-439, 1992.

[23] F. Zijlstra, "Efficiency in work behaviour: A design approach for modern tools," 1993.

[24] D. De Waard, R. te Groningen, and V. Studiecentrum, *The measurement of drivers' mental workload*: Groningen University, Traffic Research Center, 1996.

[25] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," *Human mental workload,* vol. 1, pp. 139–183, 1988.

[26] V. J. Gawron, "Human performance measures handbook," 2000.

[27] R. ODONNELL and F. EGGEMEIER, "Workload assessment methodology," *Handbook of perception and human performance.,* vol. 2, pp. 42-1, 1986.

[28] E. Leyman, G. Mirka, D. Kaber, and C. Sommerich, "Cervicobrachial muscle response to cognitive load in a dual-task scenario," *Ergonomics,* vol. 47, pp. 625-645, 2004.

[29] F. G. W. C. Paas and J. J. G. Van Merriënboer, "Instructional control of cognitive load in the training of complex cognitive tasks," *Educational psychology review*, vol. 6, pp. 351-371, 1994.

[30] F. Wada, M. Iwata, and S. Tano, "Information presentation based on estimation of human multimodal cognitive load," *Joint 9th IFSA World Congress and 20th NAFIPS Int. Conference*, pp. 2924-2929 vol. 5, 2001.

[31] B. Kerr, "Processing demands during mental operations," *Memory & Cognition*, vol. 1, pp. 401-412, 1973.

[32] W. Knowles, "Operator loading tasks " *Human factors*, vol. 5, p. 155, 1963.

[33] Y. Shi, N. Ruiz, R. Taib, E. H. C. Choi, and F. Chen, "Galvanic Skin Response (GSR) as an Index of Cognitive Load" *in Proc. of* CHI, San Jose, California, U.S.A., 2007.

[34] F. G. W. C. Pass and J. J. G. V. Merrienboer, "Variability of Worked Examples and Transfer of Geometrical Problem-Solving Skills: A Cognitive-Load Approach," *Jounal of Educational Psychology*, vol. 86, pp. 122-133, 1994.

[35] P. W. M. V. Gerven, F. Pass, and J. J. G. V. Merrienboer, "Memory load and the cognitive pupillary response in aging," *Psychophysiology*, vol. 41, pp. 167-174, 2004.

[36] C. Gütl, M. Pivec, C. Trummer, V. M. García-Barrios, F. Mödritscher, J. Pripfl, and M. Umgeher, "Adele (adaptive e-learning with eye-tracking): Theoretical background, system architecture and application scenarios," *European Journal of open, Distance and E-learning (EURODL)*, pp. 2005-12, 2005.

[37] W. S. Ark, D. C. Dryer, and D. J. Lu, "The emotion mouse," *in Proc. of the 8th International Conference on Human-Computer Interaction*, 1999, pp. 818-823.

[38] J. Liu, C. K. Wong, and K. K. Hui, "An adaptive user interface based on personalized learning," *IEEE Intelligent Systems*, pp. 52-57, 2003.

[39] S. Oviatt, "Human-centered design meets cognitive load theory: designing interfaces that help people think," *in Proc. of The 14th annual ACM international conference on Multimedia*, 2006, pp. 871-880.

[40] M. Khawaja, F. Chen, C. Owen, and G. Hickey, "Cognitive Load Measurement from User's Linguistic Speech Features for Adaptive Interaction Design," *Human-Computer Interaction–INTERACT 2009*, pp. 485-489, 2009.

[41] J. B. Sexton and R. L. Helmreich, "Analyzing cockpit communication: the links between language, performance, error, and workload," in *The Tenth International Symposium on Aviation Psychology*, 1999, pp. 689-695.

[42] J. H. L. Hansen and S. Patil, *Speech Under Stress: analysis, modeling and recognition*: Springer-Verlag Berlin Heidelberg, 2007.

[43] A. Jameson, J. Kiefer, C. Müller, B. Großmann-Hutter, F. Wittig, and R. Rummer, "Assessment of a user’s time pressure and cognitive load on the basis of features of speech," *Resource-Adaptive Cognitive Processes,* pp. 171-204, 2009.

[44] M. A. Khawaja, N. Ruiz, and F. Cheng, "Potentical Speech Features for Cognitive Measurement," *in Proc. of the19th Australian conference on Computer-Human Interaction: Entertaining User Interfaces,* pp. 57-60, 2007.

[45] S. Oviatt, R. Coulston, and R. Lunsford, "When do we interact multimodally?: cognitive load and multimodal communication patterns," *in Proc. of the 6th international conference on Multimodal interfaces*, 2004, pp. 129-136.

[46] J. SCfflLPEROORD, "On the cognitive status of pauses in discourse production," *Contemporary tools and techniques for studying writing,* vol. 10, 2001.

[47] L. Stirling, G. Barrington, and S. Douglas, "Two times three little pigs: dysfluency, cognitive complexity and autism," Annual Meeting of the Australian Linguistic Society, 2006.

[48] M. Honda, "Human Speech Production Mechanisms," *NTT Technical Review,* vol. 1, 2003.

[49] P. K. Rajasekaran, Doddington, G.R., Picone, J.W., "Recognition of speech under stress and in noise," *in Proc. of the 11th IEEE international conference on acoustics, speech, and signal processing (ICASSP '86)*, Tokyo, 1986, pp. 733-736.

[50] K. R. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Bänziger, "Acoustic correlates of task load and stress," *in Proc. of ICSLP*, Denver, Colorado, U.S.A., 2002, pp. 2017-2020.

[51] E. Mendoza and G. Carballo, "Acoustic analysis of induced vocal stress by means of cognitive workload tasks," *Journal of Voice,* vol. 12, pp. 263-273, 1998.

[52] G. Griffin and C. Williams, "The effects of different levels of task complexity on three vocal measures," *Aviation, space, and environmental medicine,* vol. 58, p. 1165, 1987.

[53] H. Boril, O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and detection of cognitive load and frustration in drivers' speech," *in Proc. of Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 502-505.

[54] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. Choi, "An investigation of formant frequencies for cognitive load classification," *in Proc. of InterSpeech*, 2010b, pp. 2022-2025.

[55] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Formant Frequencies Under Cognitive Load: Effects and Classification," *EURASIP journal on advances in signal processing*, 2011.

[56] S. Lively, D. Pisoni, W. Van Summers, and R. Bernacki, "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences," *The Journal of the Acoustical Society of America*, vol. 93, pp. 2962-2973, 1993.

[57] B. Yin, N. Ruiz, F. Chen, and M. A. Khawaja, "Automatic cognitive load detection from speech features," *in Proc. of CHISIG,* pp. 249-255, 2007.

[58] B. Bielefeld, "Language identification using shifted delta cepstrum," *Fourteenth Annual Speech Research Symposium,* 1994.

[59] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *in Proc. of International of Conference on Spoken Language Processing*, 2002, pp. 89-92.

[60] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language Identification: A Tutorial," *Circuits and Systems Magazine, IEEE,* vol. 11, pp. 82-108, 2011.

[61] T. F. Yap, E. Ambikairajah, E. Choi, and F. Chen, "Phase-based features for cognitive load measurement system," *in Proc. of ICASSP,* pp. 4825-4828, 2009.

[62] T. F. Yap, E. Ambikairajah, J. Epps, and E. H. C. Choi, "Cognitive load classification using formant features," *in Proc. of ISSPA*, pp. 221-224, 2010.

[63] R. Fernandez and R. W. Picard, "Modeling drivers' speech under stress," *Speech Communication*, vol. 40, pp. 145-159, 2003.

[64] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *in Proc. of ISCA Workshop on Speaker Recognition: A speaker Oddyssey* 2001.

[65] K. Abou-Moustafa, C. Suen, and M. Cheriet, "A generative-discriminative hybrid for sequential data classification," *in Proc. of* ICASSP 2004, pp. 805–808.

[66] B. Yin, N. Ruiz, F. Chen, and E. Ambikairajah, "Exploring classification techniques in speech based cognitive load monitoring," *in Proc. of Interspeech*, 2008, pp. 2478-2481.

[67] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 39, pp. 1-38, 1977.

[68] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *in Proc. of Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[69] T. Thiruvaran, E. Ambikairajah, and J. Epps, "FM features for automatic forensic speaker recognition," *in Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 1497–1500.

[70] E. Wong and S. Sridharan, "Fusion of output scores on language identification system," *Multilingual Speech and Language Processing*, p. 7, 2003.

[71] J. Villing, "Dialogue behaviour under high cognitive load," presented at the in Proc. of the 10th annual meeting of the special interest group in discourse and dialoge London, UK, 2009.

[72] C. Muller, B. Grossmann-Hutter, A. Jameson, R. Jummer, and F. Wittig, "Recognizing time pressure and cognitive load on the basis of speech: An experimental study," *Lecture Note in Computer Science*, pp. 24-33, 2001.

[73] A. Jameson, J. Kiefer, C. Muller, B. G. Hutter, F. Wittig, and R. Rummer, "Assessment of a user's time pressure and cognitive laod on the basis of features of speech " *Journal of computer science and technology* 2006.

[74] B. Yin, N. Ruiz, F. Chen, and E. Ambikairajah, "Investigating speech features and automatic measurement of cognitive load," in *in Proc. of IEEE 10th Workshop on Multimedia Signal Processing*, Cairns, Queensland, 2008, pp. 988-993.

[75] J. Stroop, "Studies of interference in serial verbal reactions," *Journal of Experimental Psychology: General*, vol. 18, pp. 643-662, 1935.

[76] Metametrics, "The Lexile Framework for Reading," 2007

[77] J. John R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*: Macmillan Publishing Company, 1993.

[78] S. Koolagudi, R. Reddy, and K. Rao, "Emotion recognition from speech signal using epoch parameters," in *IEEE International Conference on Signal Processing and Communication (SPCOM)*, IISC Bangalore, India, 2010, pp. 1-5.

[79] T. Kinnunen and P. Alku, "On separating glottal source and vocal tract information in telephony speaker verification," in *ICASSP*, Taipei, 2009, pp. 4545-4548.

[80] A. Acero, "Source-filter models for time-scale pitch-scale modification of speech," in *ICASSP*, Seattle, WA , USA 1998, pp. 881-884.

[81] J. Harrington and S. Cassidy, "Techniques in speech acoustic," *Kluwer Academic Publishers*, 1999.

[82] L. Rabiner and R. Schafer, *Digital processing of speech signals*: Pearson Education India, 1978.

[83] P. Alku, "Glottal Wave Analysis With Pitch Synchronous Iterative Adaptive Inverse Filtering," *In Proc. of EuroSpeech*, pp. 1081-1084, 1991.

[84] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *in Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science B. V., Amsterdam*, pp. 495-518, 1995.

[85] P. Boersma and D. Weenink. *Praat: doing phonetics by computer* Available: http://www.fon.hum.uva.nl/praat/

[86] K. Sjolander and J. Beskow, "Wavesurfer - an open source speech tool," presented at the In Proc. of international conference on spoken language processing, Beijing, China, 2000.

[87] T. Thiruvaran, E. Ambikairajah, and M. Epps, "Group delay features for speaker recognition," 2008, pp. 1-5.

[88] H. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Processing,* vol. 22, pp. 259-267, 1991.

[89] V. Sethu, E. Ambikairajah, and J. Epps, "Group delay features for emotion detection," presented at the in Proc. of interspeech, Antwerp, Belgium, 2007.

[90] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing,* vol. 41, pp. 3024-3051, 1993.

[91] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of FM components from speech signals using all-pole model," *Electronics Letters,* vol. 44, p. 449, 2008.

[92] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Acustica united with acta acustica,* vol. 88, pp. 416-422, 2002.

[93] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *ICASSP*, 1998, pp. 617-620.

[94] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition," *in Proc. of Odyssey, The Speaker and Language Recognition Workshop,* pp. 34-39, 2010.

[95] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Speaker identification using FM features," *in Proc. of the 11th Australia International Conference on Speech Science & Technology,* pp. 148-152, 2006.

[96] S. Kim, M. Ji, Y. Suh, and H. Kim, "Noise Robust Speaker Identification Using Sub-Band Weighting in Multi-Band Approach," *IEICE Transactions on Information and Systems,* vol. 90, pp. 2110-2114, 2007.

[97] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," *Proc. ICASSP,* pp. 641-644, 1998.

[98] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker dependency of spectral features and speech production cues for automatic emotion classification," in *ICASSP*, 2009, pp. 4693-4696.

[99] B. J. Shannon and K. K. Paliwal, "A comparative study of filter bank spacing for speech recognition," *Microelectronic engineering research conference,* pp. 1-3, 2003.

[100] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication*, vol. 50, pp. 312-322, 2008.

[101] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," 2002, pp. 1255-1258.

[102] S. r. g. a. C. M. University. (1995, *Noisex-92 database*. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[103] A. Webb, *Statistical pattern recognition*: A Hodder Arnold Publication, 1999.

[104] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unscented transform and its application to speaker recognition," in *Interspeech*, Lisboa, Potugal, 2005, pp. 1985-1988.

[105] J. Jensen, D. Ellis, M. Christensen, and S. Jensen, "Evaluation of distance measures between Gaussian mixture models of MFCCs," 2007, pp. 107–108.

[106] T. Stadelmann and B. Freisleben, "Fast and robust speaker clustering using the earth mover's distance and Mixmax models," in *ICASSP*, 2006, pp. 989-992.

[107] T. Thiruvaran, "Automatic speaker recognition using phase based features," PhD, Awarded By: University of New South Wales. Electrical Engineering & Telecommunications, 2009.

[108] A. El-Solh, A. Cuhadar, and R. Goubran, "Evaluation of Speech Enhancement Techniques for Speaker Identification in Noisy Environments," in *Ninth IEEE International symposium on Multimedia Workshops*, 2008, pp. 235-239.

[109] C. H. You, S. Rahardja, and H. Li, "Speech enhancement for telephony name speech recognition," in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 973-976.

[110] L. M. Arslan and J. H. L. Hansen, "Speech enhancement for crosstalk interference," *Signal Processing Letters, IEEE*, vol. 4, pp. 92-95, 1997.

[111] R. Martin and G. Enzner, "Speech enhancement in hearing aids-from noise suppression to rendering of auditory scenes," in *IEEE 25th Convention of Electrical and Electronics Engineers in Israel*, 2008, pp. 363-367.

[112] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113-120, 1979.

[113] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 197-210, 2003.

[114] T. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Transactions on Speech and Audio Processing,* vol. 4, pp. 383-389, 2002.

[115] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109-1121, 2003.

[116] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing, ,* vol. 13, pp. 857-869, 2005.

[117] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *ICASSP*, 2003, pp. 177-180.

[118] C. H. You, S. N. Koh, and S. Rahardja, "Kalman filtering speech enhancement incorporating masking properties for mobile communication in a car environment," 2005, pp. 1343-1346.

[119] C. H. You, S. Rahardja, and S. N. Koh, "Perceptual Kalman filtering speech enhancement," in *ICASSP*, 2006, pp. 461-464.

[120] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings: Mathematical, Physical and Engineering Sciences,* vol. 454, pp. 903-995, 1998.

[121] X. Zou, X. Li, and R. Zhang, "Speech enhancement based on Hilbert-Huang transform theory," 2008, pp. 208-213.

[122] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech communication,* vol. 24, pp. 249-257, 1998.

[123] E. Deger, M. K. I. Molla, K. Hirose, N. Minematsu, and M. K. Hasan, "EMD based soft-thresholding for speech enhancement," in *InterSpeech*, 2007, pp. 810-813.

[124] S. Salahuddin, S. Al Islam, M. K. Hasan, and M. Khan, "Soft thresholding for DCT speech enhancement," *Electronics Letters,* vol. 38, pp. 1605-1607, 2002.

[125] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*: Englewood Cliffs: Prentice Hall, 1998.

[126] T. S. Gunawan, E. Ambikairajah, and J. Epps, "Perceptual speech enhancement exploiting temporal masking properties of human auditory system," *Speech communication,* vol. 52, pp. 381-393, 2010.

[127] EBU. (1988, *Sound Quality Assessment Material Recordings for Subjective Tests. European Broadcasting Union.*

[128] Y. Gui and H. Kwan, "Adaptive subband Wiener filtering for speech enhancement using critical-band gammatone filterbank," in *48th Midwest Symposium on Circuits and Systems*, 2005, pp. 732-735 Vol. 1.

[129] T. Irino, "Noise suppression using a time-varying, analysis/synthesis gamma chirp filterbank," in *ICASSP*, 1999, pp. 97-100.

[130] L. Lin, E. Ambikairajah, and W. Holmes, "Speech enhancement for nonstationary noise environment," in *Asia-Pacific Conference on Circuits and Systems*, 2002, pp. 177-180 vol. 1.

[131] T. Irino, "Noise suppression using a time-varying, analysis/synthesis gamma chirp filterbank," 1999, pp. 97-100.

[132] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and Trends in Signal Processing*, vol. 1, pp. 1-194, 2007.

[133] L. Lin, W. Holmes, and E. Ambikairajah, "Speech enhancement based on a perceptual modification of Wiener filtering," *Electronics Letters*, pp. 1486 - 1487, 2002.