การรู้จำเสียงที่ไม่ขึ้นกับเสียงรบกวนพื้นหลังโดยใช้แบบรูปสเปกโทรแกรมเชิงภาพ

นายพีระพล ขุนอาสา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2555

BACKGROUND-NOISE INDEPENDENT SOUND RECOGNITION USING IMAGERIAL SPECTROGRAM PATTERNS

Mr Peerapol Khunarsa

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Science
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic year 2011

Thesis Title	BACKGROUND-NOISE INDEPENDENT SOUND RECOGNITION USING IMAGERIAL SPECTROGRAM PATTERNS
By	Mr. Peerapol Khunarsa
Field of Study	Computer Science
Thesis Advisor	Professor Chidchanok Lursinsap, Ph.D.
Thesis Co-Advisor	Thanapant Raicharoen, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

……..…….. Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE

……Chairman
(Suphakant Phimoltares, Ph.D.)

…… Thesis Advisor
(Professor Chidchanok Lursinsap, Ph.D.)

……..……Thesis Co-Advisor
(Thanapant Raicharoen, Ph.D.)

……. Examiner
(Assistant Professor Rajalida Lipikorn, Ph.D.)

……..….Examiner
(Assistant Professor Thanarat Chalidabhongse, Ph.D.)

……..…..External Examiner
(Chularat Tanprasert, Ph.D.)

……..…..External Examiner
(Chai Wutiwiwatchai, Ph.D.)

พีระพล ขุนอาสา : การรู้จำเสียงที่ไม่ขึ้นกับเสียงรบกวนพื้นหลังโดยใช้แบบรูปสเปกโทรแกรมเชิงภาพ (BACKGROUND-NOISE INDEPENDENT SOUND RECOGNITION USING IMAGERIAL SPECTROGRAM PATTERNS) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ศ.ดร. ชิดชนก เหลือสินทรัพย์, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม น.อ. ดร. ธนพันธุ์ หร่ายเจริญ จำนวนหน้า 89 หน้า.

ในการรู้จำเสียงนั้นหมายถึงการรู้จำและจำแนกเสียงประเภทต่างๆ รวมไปถึงเสียงเพลง เสียงร้องนองนักร้องหรือเสียงเรียกสายเป็นต้น โดยในการรู้จำเสียงนั้นมีงานวิจัยที่เกี่ยวข้องอยู่ด้วยกัน 2 ส่วน คือการประมวลผลสัญญาณและการรู้จำ ในปัจจุบันได้มีการประยุกต์ใช้เทคนิคหลายๆ แบบเข้ามาช่วยในการแก้ปัญหาการรู้จำเสียงประเภทต่างๆมากมาย โดยทั่วไปนั้นประกอบไปด้วย 2 ขั้นตอนวิธีคือการดึงแยกคุณลักษณะและการจำแนกประเภท ในงานวิจัยนี้เราได้นำเสนอกระบวนการรู้จำเสียงที่ไม่ขึ้นกับเสียงรบกวนพื้นหลังโดยใช้การเทียบแบบรูปสเปกโทรแกรมเชิงโครงข่ายประสาทเทียม โดยเราได้นำเอาระเบียบขั้นตอนวิธีดังกล่าวมาแก้ปัญหาการรู้จำเสียงประเภทต่างๆ ที่มีเสียงรบกวนพื้นหลังสูง ในขั้นตอนแรกนั้นเราทำการแปลงสัญญาณเสียงให้รู้ในรูปแบบสเปกโทรแกรมในขั้นตอนของการดึงแยกคุณลักษณะและในขั้นตอนของการจำแนกประเภท เราได้ใช้เครือข่ายประสาทเทียมและวิธีค้นหาสมาชิกที่ใกล้ที่สุด ในงานวิจัยนี้เราได้ประยุกต์เทคนิคดังกล่าวกับปัญหา การรู้จำคำร้องในเพลงที่มีเสียงดนตรีเป็นพื้นหลังและปัญหาการรู้จำเสียงสภาวะแวดล้อมในสิ่งแวดล้อมประเภทต่างๆ

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต.....

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา 2554 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5073856323 : MAJOR COMPUTER SCIENCE
KEYWORDS : Spectrogram / Singing voice recognition / Environmental sound recognition / Automatic speech recognition / Spectrogram pattern matching / feed-forward neural network (ANN) / k-nearest neighbour (KNN)

PEERAPOL KHUNARSA : BACKGROUND-NOISE INDEPENDENT SOUND RECOGNITION USING IMAGERIAL SPECTROGRAM PATTERNS.

ADVISOR : PROF. CHIDCHANOK LURSINSAP, Ph.D.,

CO-ADVISOR : Gp.Capt. Thanapant Raicharoen, Ph.D., 89 pp.

Audio recognition is defined as the task of recognizing a particular piece of audio (could be music, ring-tone, speech and singing as well, from a given sample set of audio tracks. The field of audio recognition tries to emulate this behavior by using concepts from Biological modeling, signal processing theory and pattern recognition theory. Several techniques have been proposed to solve the problem of audio recognition. Most of the proposed methods are divided into two processing steps: feature extraction and classification. This research proposes a Background Noise Independence Sound Recognition algorithm that is able to automatically recognize a piece of audio with background by using the concept of spectrogram pattern matching. Each signal is analyzed and generated to its spectrogram that is used to train data for the classifier. Several classification functions are used, such as feed-forward neural network and k-Nearest Neighbor. This research applies a concept of matching of spectrogram pattern with various audio problem singing voice recognition and the environment sound recognition.

Department : Mathematics and Computer Science Student's Signature .....
Field of Study : Computer Science Advisor's Signature .....
Academic Year : 2011 Co-advisor's Signature .....

Acknowledgments

During my years as a Ph.D. student, I have received a lot of tuition, care and friendship from several people, some of which I wish to thank here.

- First of all I would like to thank the Uttaradit Rajabhat University who sponsor the Ph.D. scholarships.

- During my time as a Ph.D.s student, I am grateful to my supervisor, Prof.Dr. Chidchanok Lursinsap, to whom with his advice, guidance and care, help me to overcome the necessary difficulties of the process of research and make this dissertation possible.

- I would also like to thank my co-supervisor, Dr.Thanapant Raicharoen, who gives me a wonderful suggestion in Ph.D. research methodologies.

Finally, I would like to dedicate my Ph.D. to my family who has given me the opportunity of an education from the best institutions and support throughout my life.

CONTENTS

List of Figures

List of Tables

Table Page

CHAPTER I
INTRODUCTION

1.1 Music Information Retrieval (MIR)

Music Information Retrieval (MIR) is an interdisciplinary research area which appeared in recently. It converge Computer Science, Information Retrieval, Engineering, Signal Processing, Musicology and Music Theory. The term MIR encompasses a number of different research that have the common denominator of being related to music access.

Despite its name, MIR is not only about retrieving information from music but to full users' music information, amusement or training needs. And as these needs are more aimed at music retrieval that music information retrieval, so are the consequent approaches. Also, the term "retrieval" has a broader sense since it encompasses tasks such as filtering, classification, identification, indexing and visualization that become increasingly useful for the final users [1].

Most of the research works on MIR, of the proposed techniques, and of the developed systems are content-based. The main idea underlying content-based approaches is that a document can be described by a set of features that are directly computed from its content [1]. In the case of MIR, the content is the implicit and explicit information related to a sound or a piece of music and that is embedded in the signal itself. The methodologies of MIR are based on Infomation Retrieval, Thus techniques of statistics and probability theory are used to describe the underlying models. Some of the topics MIR include are:

- Computational methods for classification, clustering, and modeling
- Musical feature extraction for monophonic and polyphonic audio
- Similarity and pattern matching
- Music identification and recognition
- Filtering for music and music queries, query languages, standards and other metadata or protocols for music information handling and retrieval
- Software for music information retrieval, human-computer interaction and interfaces, mobile applications, user behavior
- Music perception, cognition, an affect and emotions

- Music similarity metrics, syntactical parameters, semantic parameters, musical forms, structures, styles and genres,
- Music annotation methodologies,
- Music automatic summarization, analysis and knowledge representation, downgrading, transformation, formal models of music, digital scores and representations,
- Music indexing and metadata
- Music archives and digital collections
- Intellectual property rights, national and international intellectual property right issues, digital rights management, identification and traceability,
- Sociology and economy of music,

1.2 Singing Voice vs. Speech

Although singing voice and speech sounds have many properties because they originate from the same apparatus, there are several differences [2] , [3]:

- Duration of voiced sounds
- Loudness
- Pitch
- Vibrato
- Formants
- Rhythm
- Rhyme

1.3 Related Works

In addition to vision, sound is one of human being important sense. It is the sense most used to gather information about the environment. Despite this, comparatively little research has been done the field of environmental sound classification. The research that has been done mainly centered on the recognition of speech and music.

The voice recognition has been most popular as the other is to recognize the human voice. There are many problems related to the management of musical data that have not yet been solved. They are now being extensively considered in the field of Music Information Retrieval (MIR) [4]. In this research, we are interested singing voice recognition in polyphonic recordings of popular music. Our assumption is that, for any song, it is unnecessary to filter the instrumental background from the singing voice to recognize the singing words. By following this direction, we expected to achieve high recognition accuracy. Singing voice recognition is very different from Automatic Speech Recognition (ASR) because of the differences between speech and singing voice such as duration of voice sound, loudness, pitch, vibrato, formant, rhythm and rhyme [5] [6] [7] [8]. To make the problem realistic and feasible, we considered singing voices in polyphonic audio signal sampled from commercial compact disc (CD) recordings of popular music. In addition, various music genres for popular music, such as Rock, hard rock, soft rock, dance, hip-pop, soul, R&B, folk, and acoustic were concerned. All music genres have a man and woman singers. The type of songs emphasized in this study is Thai songs. The study of Thai singing word recognition is rather few and Thai words have special characteristics due to their intonation patterns. Different intonations have different meanings. The intonation of a Thai word may be changed during the singing, depending on the rhythm.

Several techniques concerning the English words was proposed for to solve the problem of audio recognition [9] [10] [11] [12] [13] [14] [15] [16] [17]. Most of the proposed methods were divided into two processing steps: feature extraction and classification. In the first step, feature exaction, the redundant information contained in the signal was transformed into descriptors used as the input of a classifier. In the second step, classification, the singing voice was recognized. Shenoy [18] used the amplitude variation over time in each sub-band and a threshold method on the energy function such as the proportion of frames classified as vocals to be equivalent to the proportion of the singing in the entire song. Nwe [19] used Harmonic Attenuated LFPCs with Hidden Markov Model (HMM) models based on three parameters, e.g. section type (intro, verse, chorus, bridge and outro), tempo, and loudness. Tsai [20] used Mel-Frequency Cepstral Coefficient s (MFCCs) andGMM models to classify vocal from non-vocal signals. Berenzweig and Ellis [21] used vector of posterior probability as a feature and HMM framework with two states, "singing" and "non-singing". Chou and Gu [22] used 4 Hz modulation energy, harmonic coefficient, 4Hz harmonic coefficient, delta MFCC and delta log energy as features and GMM model to detect singing voice. Berenzweig [23] applied 13 PLPCs and MLP. Maddage [24] considered LPC, LPC derived cepstrums (LPCC), MFCC, spectral power (SP), short time energy (STE), and ZCR as feaures and a multi-layer neural network, SVM and GMM for classification. SVM was found to outperform the other classifiers. Maddage [25] latter tried Twice Iterated Composite Fourier Transform (TICFT) to each audio frame. Rocamora and Herrera [26] used different sets of features such as MFCCs and their deltas, LFPC their deltas and double deltas, PLPCs and their deltas, HC and pitch and different classifiers such as a SVM, a back propagation NN, a decision tree classifier, and two different K-nearest neighbors. Tzanetakis [27] used spectral shape feature, MFCCs, mean and deviation of pitch , centroid and LPCs for feature extraction and a naive bayes network, nearest neighbor

algorithms, back-propagation ANN, a decision tree classifier based on the C4.5 algorithm, a SVM classifiers. Kim [28] used a harmonic measure, defined as the ratio of the total signal energy to the maximally harmonically attenuated signal and threshold method on the harmonic measure to classify the segment.

As compared to other areas in audio such as speech or music, research on general unstructured audio-based scene recognition has received little attention. To the best of our knowledge, only a few systems (and frameworks) have been proposed to investigate of singing voice recognition with raw audio. Most of investigations of singing voice recognition deal with recognition phoneme first and use a speech recognizer for lyrics recognition. Sasou [29] tested an Auto Regressive HMM with pure singing voice signals from the RWC database. These studies presumed pure monophonic singing voices without accompaniment, posing additional difficulties for practicable use with musical audio signals like CD recordings. Suzuki [30] combined both the melody and the lyrics of the user's singing voice to retrieve a song from a database. The authors used a large vocabulary speech recognition system with a HMM as the acoustic model adapted to the singing voice using the speaker adaptation technology.

Wong [31] proposed a system for real-time alignment of Cantonese music, which is a particular tone language. The meaning of a word changes when pronounced with a different pitch. A MLP was used to segregate the vocal from the non-vocal segments taking as input the spectral flux, the HC, the ZCR, the MFCCs, the amplitude level and the 4Hz modulation energy. DTW algorithm was used to align the two sequences. However, this method is not consistently effective because the durations of uttered phonemes are based on location, even though they are the same phonemes.

Kan [32] is probably the first English lyrics sentence level alignment system for aligning the lyrics to the music signals for a specific structure of songs. Gruhne [33] implemented a system that performed automatic classification of 15 voiced sung phonemes in polyphonic audio. Their procedure was based on harmonics extraction and re-synthesis of a number of partials as a preprocessing step, in order to reduce influences from accompanying sounds. Then, low-level features were extracted from the audio and classified using different classification techniques like SVM, GMM and MLP. Fujihara Gruhne [34] performed automatic synchronization between lyrics and polyphonic music signals for Japan CD recordings. Their proposed system included detection of vocal segments, segregation of vocals and adaptation of a speech recognizer to the segregated vocal signals. During the first step, harmonics extraction and re-synthesis was performed as in Gruhne [33]. A simple HMM was used in order to keep only the vocal regions and remove the non-vocal sections. Last, features were extracted from the audio (MFCCs, delta MFCCs, and delta power) and the Viterbi algorithm was used to align the segmented vocal parts with corresponding lyrics. Pawel [35] presented an automatic singing voice recognition using neural network and rough sets. The method also required and combined many type of feature vector for classification method. Annamaria [36] studied the use of n-gram language models in recognizing phonemes and words in monophonic and polyphonic music. They considered uni-, bi-, and tri-gram language

models for phonemes and bi- and tri-grams for words. In the recognition, a Hidden Markov Model based phonetic recognizer was adapted to singing voice. The word recognition system achieved only 24% correct recognition rate, where the first retrieved in Figure 1.1 was an approach used in previous research.

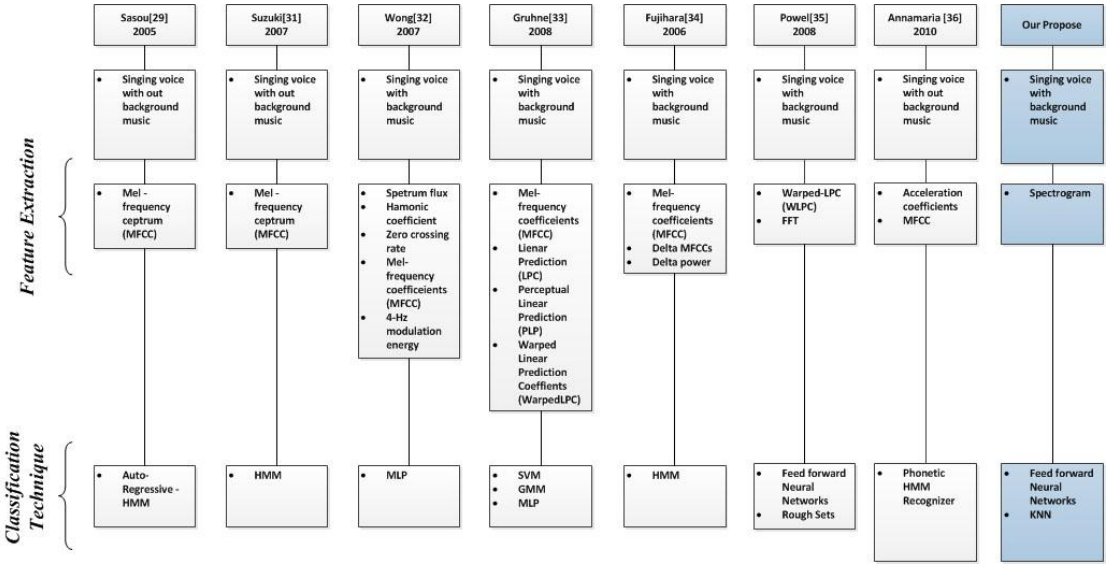

Figure 1.1 Comparison of our propose and another work.

An algorithm for audio recognition can be applied to new problems, such as other environment sound recognition. We considered the task of classifying the environment sounds to understand the scene surrounding the audio sensor. By auditory scenes, we referred to a location with different acoustic characteristics such as a streets, restaurants, offices, homes, and cars. In this research, we proposed a system of classifying a unstructured environmental sound in polyphonic audio signal sampled from commercial compact-disc (CD) recordings of popular database including various types of environmental sounds in this research such as car engine, construction, crowd applause, crow clamor, fire, helicopter, office, outdoor sounds-forest, outdoor sounds-road, restaurant stores, transportation-motorcycle, transportation-train, water, weather - rain, weather-thunder, household, air plane, water(ocean), chicken farm, and auto racing.

Similar to the above problems, Most of the proposed methods were divided into two processing steps [37], [38], [39], [40], [41] [42] are feature extraction and classification. In feature exaction step, the redundant information contained in the signal was transformed into descriptors used as the inputs of a classifier. Malkin and Waibel [43] extracted sixty-four dimensional MFCC and the spectral centroid, at a rate of 100 frames per second. They introduced linear auto encoding neural networks

for classifying the environment. A hybrid auto-encoder and GMM were used in their experiments and 80.05% average accuracy was obtained. However, they selected only those segments that were quieter than the average power in an audio file for the experiments.

Wang et al. [44] used three MPEG-7 audio low-level descriptors spectrum centroid, spectrum spread, and spectrum flatness are used as features in their study on environmental sound classification. They proposed a hybrid SVM and k-NN classifier in their study. For SVM, they used three different types of kernel functions: linear kernel, polynomial kernel and radial basis kernel. The system with 3 MPEG-7 features achieved 85.1% accuracy averaged over 12 classes.

Kraetzer et al. [45] developed a method to detect the used microphone and the background environments of audio recordings. Kraetzer extracted 63 statistical features from audio signals. Seven of the features were in time domain, i.e. empirical variance, covariance, entropy, LSB ratio, LSB flipping rate, mean of samples and median of samples. Besides these temporal features, they used 28 mel-cepstral features and 18 filtered mel-cepstral features. For classification, the data mining tool WEKA with K-means as a clustering and Naive Bayes as a classification technique were applied with the goal to evaluate their classification in regard to the classification accuracy on known audio features. For the evaluation of hypothesis I, i.e. the classification of the microphones for all rooms and a fixed number of vectors per file, the best results for the Bayesian classification achieved 75.99% and K-means clustering achieved 41.57%. For the evaluation of hypothesis II, i.e. the room classification, the results showed less impressive accuracy than the microphone classification evaluated in hypothesis I. The best result here was found with 41.54% accuracy in the case of Bayesian classification, the Headset and 100 vectors computed per file. The clustering with K-means resulted generally in worse accuracies than Bayes classification (about 15% worse in the maximum case).

Ntalampiras et al. [46] used MFCC along with MPEG-7 features to classify urban environments. They exploited a full use of MPEG-7 low level descriptors, namely audio waveform, audio power, audio spectrum centroid, audio spectrum spread, audio spectrum flatness, harmonic ration, upper limit of harmonics, and audio fundamental frequency. This work was based on a Hidden Markov Model (HMM) classification frameork.

Toyoda [47] used a multi-layered perception neural system for environmental sound recognition. The input data were the combination of instantaneous spectrum at power peak and the power pattern in time domain. Since for almost environmental sounds, their spectrum changes were not remarked when being compared with speech or voice, the combination of power and frequency pattern would reserve the major features of environmental sounds but with drastically reduced data. The recognition rate for 45 data types kinds of environmental sound was about 90%.

Eronen et al. [48], identified time and frequency domain features, as well as stochastic features, to classify various everyday outdoor and indoor scenes.

Eronen used Zero-crossing rate (ZCR), Mel- Frequency Cepstral Coefficients (MFCC), Mel-Frequency Delta Cepstral Coefficients (MFCCs), Band-energy, Spectral roll-off, Linear Prediction Coefficients (LPCs)and Linear Prediction Cepstral Coefficients(LPCC) for features. They employed k-nearest neighbor (k-NN) and the one-state Hidden Markov Model (HMM) as classifiers, and applied Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for feature transformation. They reported that, by using Mel-Frequency Cepstral Coefficients (MFCCs) and Hidden Markov Models (HMMs), they were able to achieve a recognition accuracy of up to 88%. The recognition accuracy as a function of the length of testing sequence converged after about 30-60 s. interestingly, they reported that human's recognition accuracy of the same data set was 82% with an average reaction time of 14 s.

Wang et al. [49] applied signal enhancement prior to recognition, and divided the recognition procedure into environmental sound classification and speech recognition. For signal enhancement, they used the perceptual wavelet analysis filterbank and the Karhunen-Loeve Transform (KLT). These approaches achieved satisfactory results, when combined with traditional features and classification methodologies.

Byeong-jun Han and Eenjun Hwang [50] considered three types of features, i.e. Traditional Features (TFs), Change Detection Features (CDFs), and Acoustic Texture Features (ATFs). To mitigate this problem of high dimension of the feature data, they used no-negative matrix factorization (NMF) and employed Support Vector Machine (SVM) as a classifier. Experimental results showed that the combination of these features with traditional features can achieve 86.09% of the maximum accuracy in environmental sound classification when compared with 74.35% of the maximum accuracy under traditional features.

Lozano [51] et. al. presented a short paper on a method for classifying audio sounds. The presented techniques in this research can be used as input to parse audio to make sure the alternative text is correctly describing the audio content. The following acoustic parameters were extracted from the data: Mel Frequency Cepstrum Coefficients, Zero Crossing Rate, Centroid and Roll- Off Pint. The feature extraction included a multi-resolution analysis technique with multiple windows of different sizes, instead of the traditional fixed length window. This gave a more number of parameters which would worsen the performance of the classifier. However, this was compensated by a heuristic selection of parameters to reduce the size of the feature array. The classification algorithm applied was the Gaussian Mixture Model. The experimental results were based on 60% training data and 40% testing data. The windows sizes ranged from 20 to 80 milliseconds. With the most preferable configuration, the classified reached an accuracy of 92.44%. The results also showed that using multi-resolution analysis outperformed the use of single-windows.

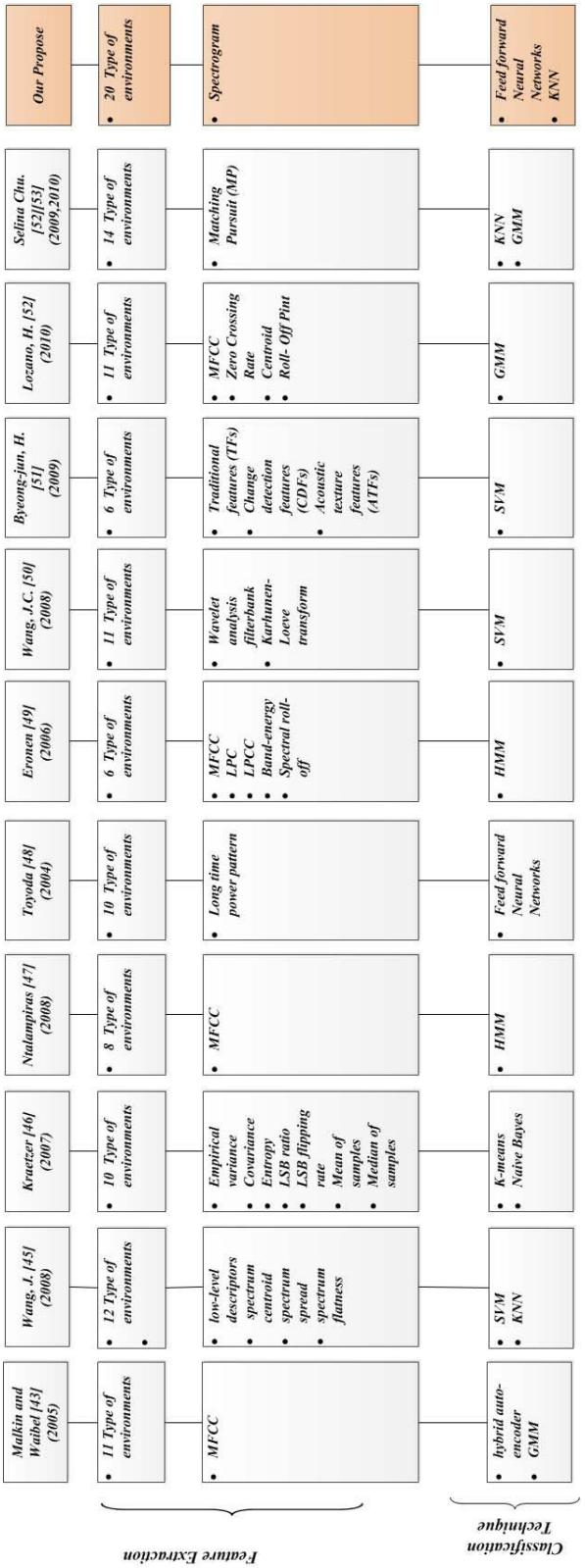

Figure 1.2 Comparison of our propose and another work.

The analysis of sound environments in Selina Chu, [52] [53], which is closest to our work, presented Matching Pursuit (MP) as features. Chu introduced the Matching Pursuit (MP) technique in environmental sounds recognition. MP provides a way to extract features that can describe sounds where other audio feature such as MFCC fails. In their MP technique, they used Gabor function based time-frequency dictionaries. It was claimed that features with Gabor properties could provide a flexible representation of time and frequency localization of unstructured sounds in the background environment. They applied KNN (k = 1) and GMM with 5 mixtures to recognize fourteen types of environmental noise events.

Jonathan [54] presented a novel feature extraction method for sound event classification, based on the visual signature extracted from the sound's time-frequency representation. The motivation stems from the fact that spectrograms form recognizable images, which can be identified by a human reader, with perception enhanced by pseudo-coloration of the image. All the four step process as follows. 1) The spectrogram is normalized into greyscale with a fixed range. 2) The dynamic range is quantized into regions, each of which is then mapped to form a monochrome image. 3) The monochrome images are partitioned into blocks, and the distribution statistics in each block are extracted to form the feature. The proposed method comes from the fact that the noise is normally more diffuse than the signal and therefore the effect of the noise is limited to a particular quantization region, leaving the other regions less changed. The method is tested on a database of 60 sound classes containing a mixture of collision, action and characteristic sounds and shows a significant improvement over other methods in mismatched conditions, without the need for noise reduction. In Figure 1.2 is an approached used in previous research.

1.4 Objective

- To classify a singing word in a singing voice signal with background music especially a singing word pronounced similarly.
- To classify a type of environment sound.
- To find the optimal windows size used to create a spectrogram for classification.
- To compare performance of the proposed method with Mel-Frequency Cepstral Coecients (MFCCs), Linear Predictive Coding (LPC) and other Features extraction method in optimal parameters.

1.5 Scope and Limitations

In this dissertation, the scope of work is constrained as follows:

- This proposed algorithm is tested with the Thai and English music 5000 albums. Sample files were coded in stereo of frequency 44.2 kHz with 128/s bit rate manually captured.

- The BBC Sound Effects Library, The Warner Bros Sound Effects Library, 56 TV-series, 356 DVD Movie. Sample files were coded in stereo of frequency 44.2 kHz with 128/s bit rate manually captured.

- The result of this approach is compared with the other Features extraction method such as Linear Predictive Coding(LPC), Mel-Frequency Cepstral Coecients (MFCC))

1.6 Dissertation Organization

This thesis is organized as follows. The next chapter, this thesis provides theoretical preliminary on grasping which is used subsequently in the remaining of the dissertation. The remaining chapters describe algorithms to solve the problem in each setting. Chapter 2 gives brief introduction and literature reviews. Chapter 3 describes algorithms and data collection. The results and discussion are given in Chapter 4. Finally, Chapter 5 concludes our work and describes future extension.

CHAPTER II

RELATED BACKGROUND

This chapter provides the necessary background for audio classification. We first discuss methods for feature extraction and classification, and conclude with a section summarizing relevant research on these fields.

2.1 Audio Classification.

Audio signal processing, sometimes referred to as audio processing, is the intentional alteration of auditory signals, or sound. As audio signals may be electronically represented in either digital or analog format, signal processing may occur in either domain. Analog processors operate directly on the electrical signal, while digital processors operate mathematically on the digital representation of that signal. Most audio recognition and classification problems are implemented using the following two-stage process.

- Feature Extraction.
- Classification.

The sequence of recognition and classification problems is shown in Figure 2. Generally, a computer represents sounds in a digital format. First, an audio signal is analyzed and calculated to generate a feature. After that, a classifier such as Feed-Forward neural network and k Nearest Neighbors (k-NN) are used for classification.

Figure 2.1 Traditional Classification Sequence.

2.2 Audio Feature Extraction

Feature extraction is the process of computing a compact numerical representation that characterizes a segment of audio. The design of descriptive feature

for a specific application is the main challenge in building pattern recognition systems. Here, we examine some of the commonly used audio signal features.

2.2.1 Fast Fourier Transform

A fast Fourier transform (FFT) is an algorithm to compute the discrete Fourier transform (DFT) and its inverse. There are many distinct FFT algorithms involving a wide range of mathematics, from simple complex-number arithmetic to group theory and number theory; this article gives an overview of the available techniques and some of their general properties, while the specific algorithms are described in subsidiary articles linked below.

A DFT decomposes a sequence of values into components of different frequencies. This operation is useful in many fields but computing it directly from the definition is often too slow to be practical. An FFT is a way to compute the same result more quickly: computing a DFT of N points in the naive way, using the definition, takes $O(N)^2$ arithmetical operations, while an FFT can compute the same result in only *O(NlogN)* operations. The difference in speed can be substantial, especially for long data sets where N may be in the thousands or millions in practice, the computation time can be reduced by several orders of magnitude in such cases, and the improvement is roughly proportional to *N/log(N).* This huge improvement made many DFT-based algorithms practical; FFTs are of great importance to a wide variety of applications, from digital signal processing and solving partial differential equations to algorithms for quick multiplication of large integers.

The most well-known FFT algorithms depend upon the factorization of N, but there are FFTs with *O(NlogN)* complexity for all N, even for prime N. Many FFT algorithms only depend on the fact that $e^{-\frac{2\pi i}{N}}$ is an *N th* primitive root of unity, and thus can be applied to analogous transforms over any finite field, such as number-theoretic transforms. Since the inverse DFT is the same as the DFT, but with the opposite sign in the exponent and a 1/N factor, any FFT algorithm can easily be adapted for it.

An FFT computes the DFT and produces exactly the same result as evaluating the DFT definition directly. The only difference is that an FFT is much faster. In the presence of round-off error, many FFT algorithms are also much more accurate than evaluating the DFT definition directly, as discussed below.

Let $x_0, \ldots, x_{N-1}$ be complex numbers. The DFT is defined by the formula

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k \frac{n}{N}} \qquad k = 0, \ldots, N-1. \tag{2.1}$$

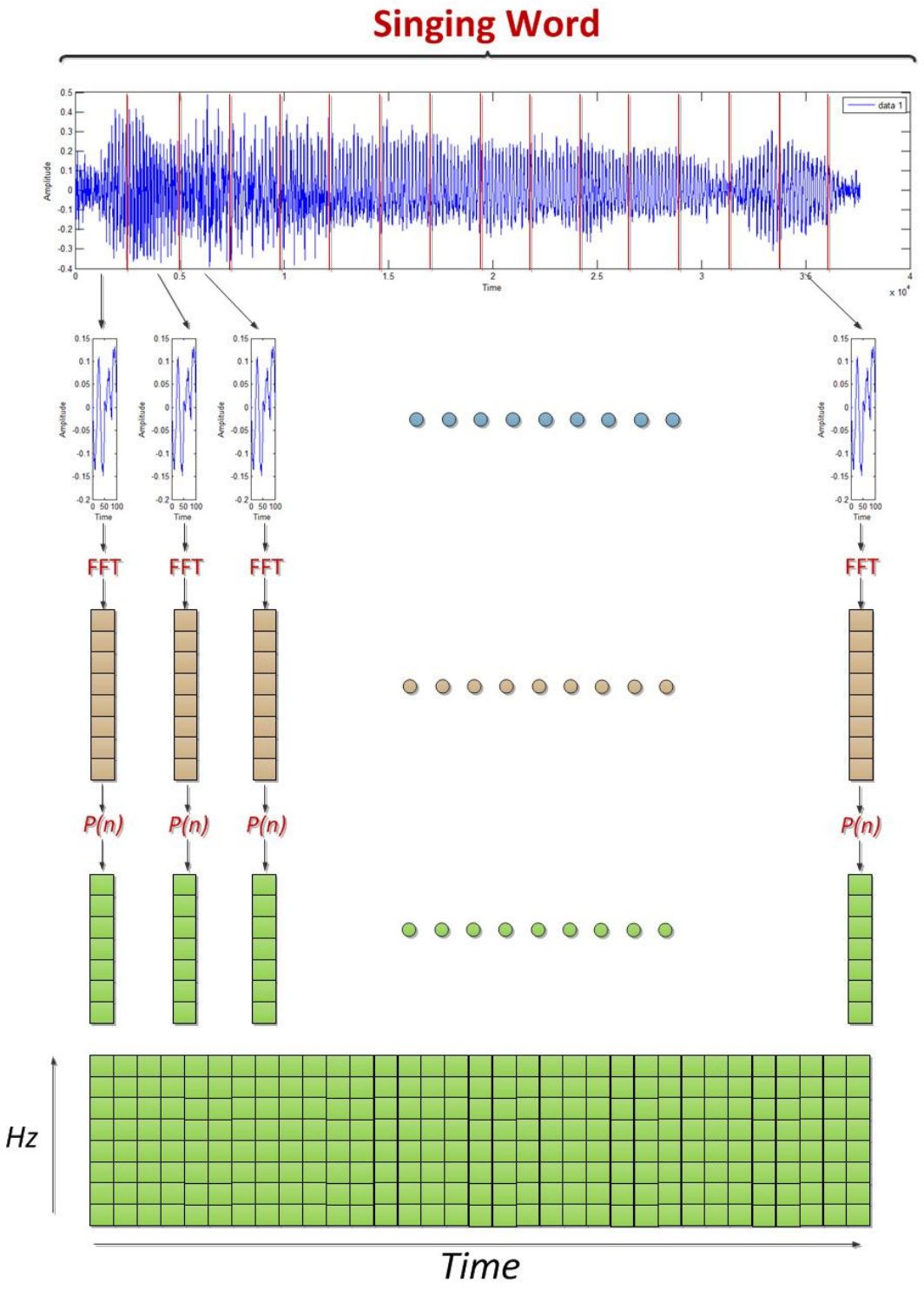

Figure 2.2 Diagram illustration of spectrogram.

2.2.2 Spectrogram (Power Spectrum)

A spectrogram is a visual representation of the distribution of acoustic energy across frequencies and over time. The horizontal axis of a spectrogram typically represents time. The vertical axis represents the discrete frequency steps. The strength of power detected is represented as the intensity at each time-frequency point.

First, the input audio signal *x(n)* of each singing word is sliced into a number of small windows or frames equal to a power of two. Each signal window is calculated by using the short-time Fourier transform (STFT) defined as follows.

$$X(k) = \sum_{n=0}^{N-1} w(n)x(n)exp(-\frac{2\pi kn}{N}) \tag{2.2}$$

For $k = 0, 1, \dots, N-1$ where k corresponds to the frequency $f(k) = \left(\frac{kf_s}{N}\right)$; f_s is the sampling frequency in Hertz and *w(n)* is Hamming time-window given by

$$w(n) = 0.54 - 0.46\cos\left(\frac{\pi n}{N}\right) \tag{2.3}$$

The power of each *X(k)*, denoted by *P(k)*, is computed by following equation.

$$P(k) = 10log_{10}(X(k)) \tag{2.4}$$

Each *P(k)* is plotted against time step to form a power spectrogram of each singing word. Figure 2.2 shows an example of how a power spectrogram is created.

2.3 Classification Basics

Classification is the task of assigning objects to one of several predefined categories. Especially, a classifier takes as input data a collection of records that are characterized by a tuple (x,y), where x is the attribute set and y is the class label. The attribute set includes several features or properties of the instance and can be either discrete or continuous. On the other hand, the class label must be a discrete attribute and this distinguishes classification from regression. Thus, classification is the task of learning a target function f that maps each attribute set x to

one of the predefined class labels y. This target function is also known as classification model. A classification technique (or classifier) is a systematic approach to build classification models from a given data set.

Examples of classifiers are decision trees, neural networks, support vector machines, logistic models etc. Each technique employs a learning algorithm in order to build a model that best fits the relationship between the attribute set and the class label of the data. This model should apart from fit well input data, correctly predict the class labels of instances it has never seen before. The input data consist the training set, while the unknown records consist the testing set. In order to measure the performance of a model, the number of correctly and incorrectly predicted test records is measured. These measures are usually presented in a tabular form, known as a confusion matrix:

Actural Class	Predicted Class	
	Class1	Class2
Class1	True Positive	False Positive
Class2	False Positive	True Positive

In order to compare the performance of different models metrics such as accuracy and error rate are widely used:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{2.5}$$

$$Error\ rate = \frac{Total\ number\ of\ predictions}{Number\ of\ correct\ predictions} \tag{2.6}$$

2.3.1 K-nearest neighbor method

KNN method is simplest methods for general, non-parametric classification and based on supervised learning [55]. The aim is to find nearest k sample from the existing training data when a new sample appears and classify the appeared sample according to most similar class [56]. Generally closeness is defined with Euclidean distance. Mitchell (1997) had explained Euclidean distance precisely with a formula. An arbitrary instance x be described by the feature vector

$$\{a_1(x), a_2(x), ..., a_n(x)\} \tag{2.7}$$

Where $a_n(x)$ denotes the value of nth attribute of instance *x*. Then the distance between two instances x_j and x_i is defined to be $d(x_i, x_j)$ as follows

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{n} (x_i - x_j)^2} \tag{2.8}$$

In general the following steps are performed for KNN algorithm:

1. Choose of k value: *k* value is completely up to user. Generally after some trials a k value is chosen according to results.

2. Distance calculation: Any distance measurement can be used for this step. Generally most known distance measurements like Euclidean and Manhattan distances are preferred.

3. Distance sort in ascending order: Chosen *k* value is also important in this step. Found distances are sorted in ascending order and k of minimum distances are taken.

4. Classification of nearest neighbors: Classes of *k* nearest neighbor are identified.

5. Finding dominant class: In the last step, queried data is classified according to class of identified *k* nearest neighbor by utilizing maximum ratio. This ratio is calculated for each class of *k* nearest neighbor with the number of data owned by that class over *k*. Let $= (p_1, p_2, \dots, p_n)$ is the set of k nearest neighbor probabilities for each class where *n* is number of class. Maximum ratio is calculated as in Eq.

2.3.2 Artificial Neural Networks

Artificial neural networks (ANN) has originated from the studies on how animal brains work, hence one has to study brain to understand fundamentals of ANNs. The brain is an extremely complex, nonlinear and parallel computer, which has the ability of organizing neurons to perform certain computations like pattern recognition, perception. Artificial neural networks born after McCulloc and Pitts introduced a set of simplified neurons in 1943. These neurons were represented as models of biological networks into conceptual components for circuits that could perform computational tasks. The basic model of the artificial neuron is founded upon

the functionality of the biological neuron. By definition, "Neurons are basic signaling units of the nervous system of a living being in which each neuron is a discrete cell whose several processes are from its cell body"

The biological neuron has four main regions to its structure. The cell body, or soma, has two offshoots from it. The dendrites and the axon end in pre-synaptic terminals. The cell body is the heart of the cell. It contains the nucleolus and maintains protein synthesis. A neuron has many dendrites, which look like a tree structure, receives signals from other neurons. The electrical signals are generated by the membrane potential which is based on differences in concentration of sodium and potassium ions and outside the cell membrane.

Consequently, a crude analogy between an artificial and a biological neuron can be made: dendrites of other neurons refer to input signals, synapses are the connection weights and activity in the cell body is represented by an activation function [46]. Illustration of such an artificial neuron is displayed in Figure 2.3.

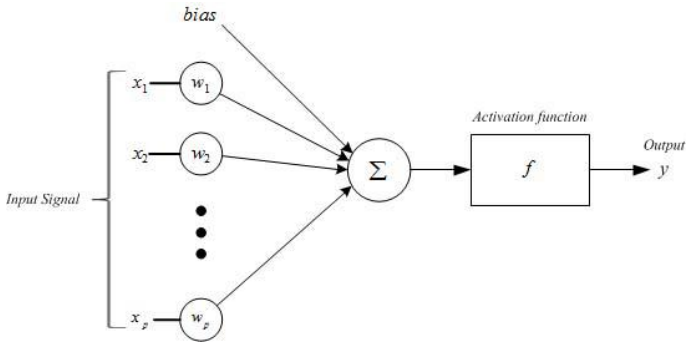

Figure 2.3 an illustration of an artificial neuron

From this model the interval activity of the neuron can be shown to be:

$$v_k = \sum_{j=1}^{m} w_{kj} x_j \tag{2.10}$$

The output of the neuron, y_k, would therefore be the outcome of some activation function on the value of v_k.

The activation function defines a mapping between the output of a neuron and its inputs. Figure 2.4 demonstrates three basic types of activation functions.

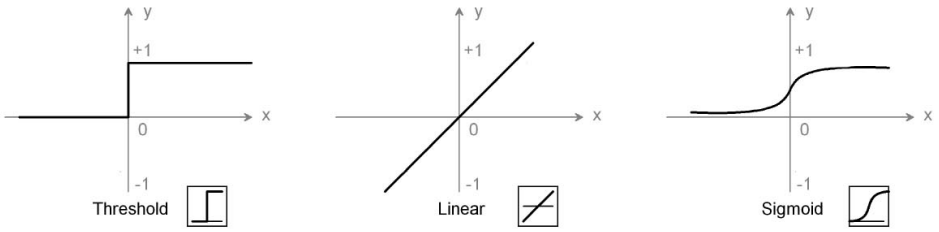

Figure 2.4 Different activation functions used in neural networks.

Emerging from the studies on how animal brains work, ANNs are composed of layers of neurons gathered in a parallel architecture with a high degree of interconnection between them. Although each neuron performs linear discrimination, ANNs can solve most of the real-world (often non-linear) problems thanks to their parallel structure. Researchers have provided numerous ANN algorithms with different architectures, learning paradigms or parameters in the literature. It is extremely arduous to cover all available ANNs due to this diversity and numerousness. Consequently, we selected several ANNs from different architectures to be used in this work. The following subsections present these ANNs with basic explanations. Please note that these explanations do not cover aspects of learning process; like learning methods, learning rate adaptation, weights update, and convergence.

2.3.3 Feed-forward Networks

In a feed-forward network data propagates in the forward (from input layer to output layer) direction, thus its neurons has only unidirectional connections (no feedback or same layer neuron to- neuron connections). Figure 2.5 displays an example of such a network.

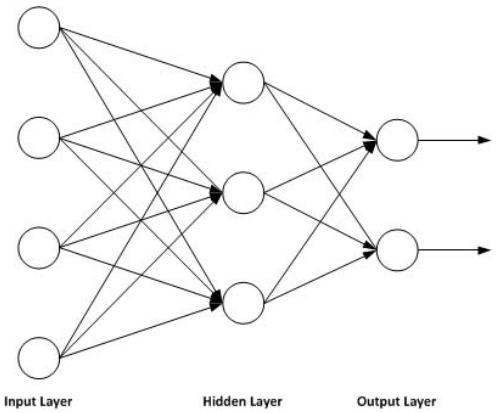

Figure 2.5 Architecture of a feed-forward neural network.

2.3.4 Multi-Layer Perceptron (MLP)

MLPs are also known as feed-forward networks, because input signals propagate layer-by-layer through the network in forward direction. MLP performs back propagation learning, where two passes of signals through the network are employed. Forward pass: Input signals are propagated in forward direction, while weights at each layer are fixed and actual output of the network is produced. Error between the actual output and the desired output (label) is calculated. In backward pass, this error signal is propagated backward and weights are adapted a second time. Therefore, this algorithm is also known as *errorback - propagation*.

2.4 Audio Time Scale Modification (TSM)

Since each sound was captured from different songs. Therefore, the time interval of each sound is not equal, depending on a singer. For this reason we apply time-scale modification algorithm (TSM) to scale the time interval of each sound become equal. Time Scale Modification (TSM) refers to the process of speeding up or slowing down a sound without changing the pitch of any tonal components. For example, TSM of speech should sound like the speaker is talking at a slower or faster rate. The idea of time-scale modification of a audio signal is used not to change the speaking rate of a signal, but to reconstruct the signal segment which is lost or delayed.

Waveform Similarity Overlap-and-Add (WSOLA) technique were used in the time-scale modification of audio signals. The WSOLA time-scale modification (TSM) technique is capable of generating an output signal with the same pitch period from the signal provided to the algorithm. This technique also minimizes discontinuities at the boundaries between good packets and reconstructed packets. It is possible to use the WSOLA time-scale modification technique on the residual signal in the same way it is used on the original signal.

2.4.1 Waveform Similarity Based Overlap-Add (WSOLA)

The Waveform Similarity Overlap-Add (WSOLA) algorithm proposed in [57] is a robust and computationally efficient algorithm used for high quality time-scale modification of speech. Timescale modification techniques aim to change only the apparent speaking rate, while preserving other perceived aspects of speech such as timbre, voice quality, and pitch. The basic idea of WSOLA is to decompose the input into overlapping segments of equal length, which are then realigned and superimposed with fixed overlap to form the output. The realignment leads to an increase or decrease in the output length. Specifically, WSOLA produces a synthetic waveform, *y(k)*, that maintains maximal local similarity to the original waveform, *x(n)*, in the neighborhoods of all sample indices given by the mapping, $n = \tau(n)$, where $\tau(n)$ is the transformation function defined as $\tau(t) = \alpha t$, being the time-scaling factor. If $\alpha > 1$, the output speech is stretched, and if $\alpha < 1$, the output speech is compressed.

The WSOLA algorithm operates entirely in the time domain. The algorithm works by segmenting the input audio waveform into blocks of equal length. Audio blocks in the input waveform are selected and overlap-added to produce the output audio. If the source blocks were taken at regular intervals in the original waveform, the output file would be of poor quality as the pitch pulses are not equally spaced. Thus, the selection of similar source blocks in the input to use for overlap-add is critical to achieving high output quality.

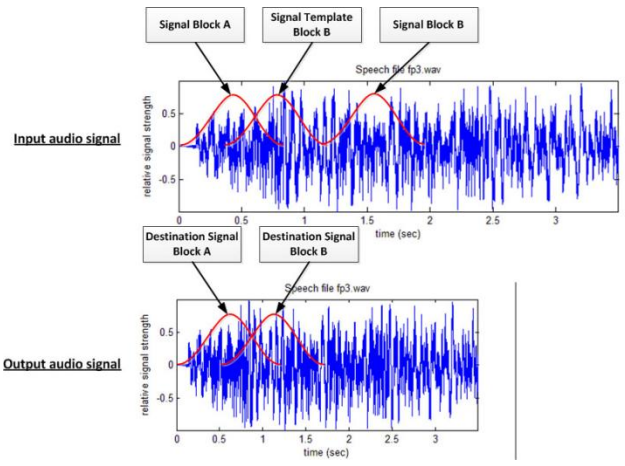

Figure 2.6 WSOLA compression at α = 0.6

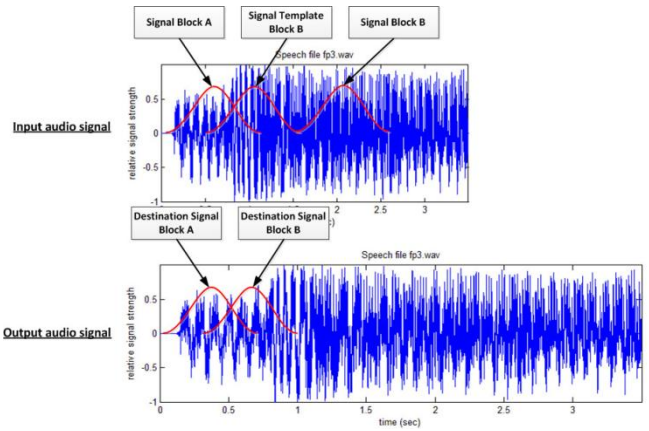

Figure 2.7 WSOLA compression at α = 1.6

Figure 2.6 and 2.7 illustrates the basic operation of the WSOLA algorithm. The algorithm iteratively constructs the output waveform, block by block. In Figure 2.6, source block A is copied to the destination block A. Template block B is the block following source block A with 50% overlap. WSOLA now needs to find a block to copy to destination block B to overlap-add with destination block A. Therefore, source block B is desired to closely resemble template block B.

The reverse transformation, $\tau^{-1}(t) = \frac{1}{\alpha}t$ gives the center of the search region in which to look for source block B. A measure of waveform similarity is computed between template block B and blocks in the search region. The source block with the greatest similarity is then copied to destination block B. The template block for the next iteration will be the block right after source block B with 50% overlap. For a given iteration, the source block follows the template block in WSOLA compression, and it precedes the template block in WSOLA expansion, as shown in Figure 2.6 and 2.7, respectively.

Once the positions of the template block and the search region are known, a series of correlations is computed between the template block and blocks in the search region. Each source block in the search region is shifted by δ, where $\frac{-L}{2} \leq \delta \leq \frac{L}{2} - 1$, and L is the length of the search region. The similarity measure used in this work is the cross-correlation coefficient,

$$\rho = \sum_{N}^{k=0} TemplateBlock(k) \times SourceBlock(k+\delta), \tag{2.11}$$

where N is the length of a block. The weighting window used in this work for the overlapadd operation is the Hamming time-window *w(n)*, $w(n) =$

$0.54 - 0.46\cos(\frac{2\pi n}{N})$, with 50% overlap. The window size is set to be the length of a block, N. Based on our experimental setup, we use a window of 512 points for length of a block *N*.

The speech quality and algorithm computation time are affected by the block size and the length of the search region for the source block. Larger blocks contain more pitch periods so the correlations will give a better measure of the waveform similarity between template and source blocks. However, if the block size is too large, artifacts such as echoes and tinny sounds will be introduced into the output. A larger search region results in more correlations being computed, thus it is computationally more expensive. Nevertheless, a better match with higher correlation between template and source block may be found within a larger search region.

CHAPTER III

PROPOSED METHODOLOGY

3.1 Concept of Proposed Solution

The singing style and duration of sing voice make it difficult to define effectively the number of states in Hidden Markov Model to recognize those singing words. Moreover, separately eliminating the instrumental background signal under uncontrollable loudness, pitch, vibrato, formant, and rhythm is not simple. In our solution, the instrumental background signal will not be filtered from the singing word signal.

Both signals are considered as one entity and this 2-dimensional signal is transformed into a 2-dimensional in spectrum domain to magnify the features of singing words. Then, the image of spectrogram is used as an input for a classifier to recognize the singing words.

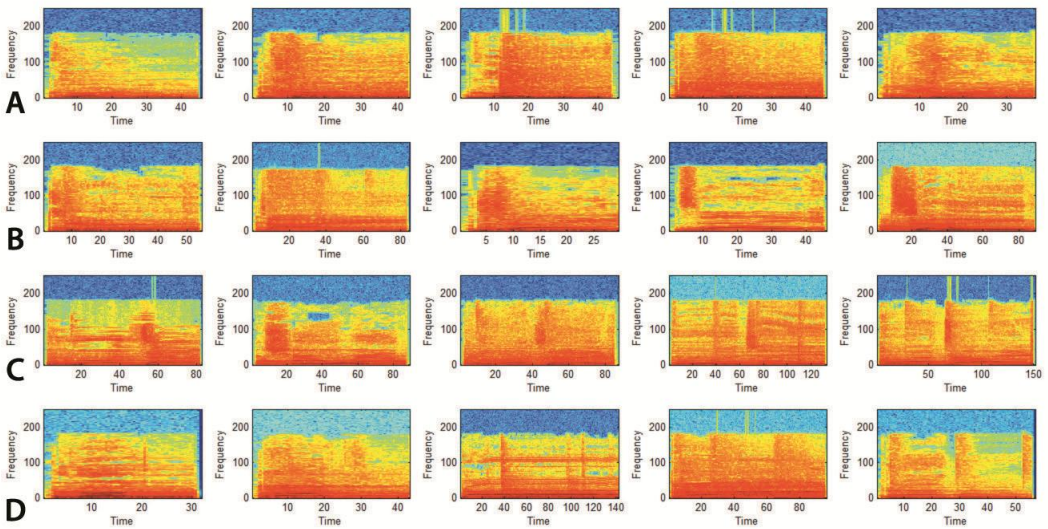

Figure 3.1: Examples of four singing words represented in forms of spectrograms. (A) Word A. (B) Word B. (C) Word C. (D) Word D.

The examples of spectrogram for four singing words are presented in Figure 3.1. Each spectrogram is shown in Figure 3.1. It can be seen that each vertical band of frequency magnitude in the spectrogram can be viewed as a vertical image whose color of each pixel implies the magnitude of the corresponding frequency. This

vertical band is called a spectrogram image. Each row of Figure 3.1 presents the same word sang from different people and time. The characteristics of a spectrogram of same word obtained from different people are very similar. Then, this thesis apples the concept of recognition like image recognition, such as hand-written digit recognition and fingerprint identification.

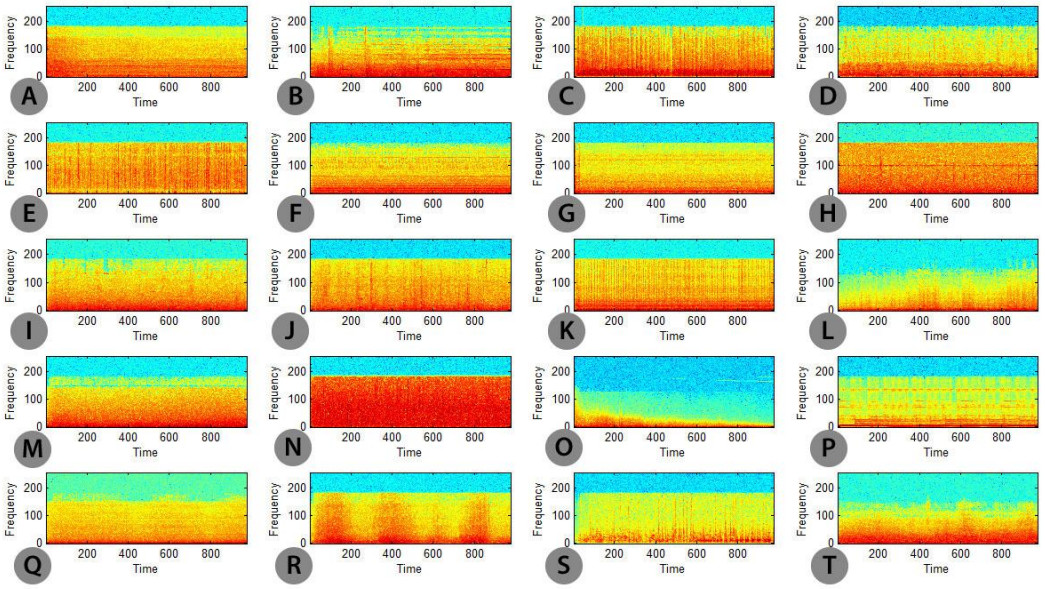

Figure 3.2: The spectrogram of each sound type. (A) car engine. (B) construction. (C) crowd Applause. (D) crow clamor. (E) cire. (F) celicopter. (G) office. (H) outdoor sounds - forest. (I) outdoor sounds - road. (J) restaurant stores. (K) transportation-motorcycle. (L) transportation-train. (M) water. (N) weather-rain. (O) weather-thunder. (P) household. (Q) airplane. (R) water(Ocean). (S) chicken farm. (T) auto racing.

With techniques discussed above, it is possible to apply the concept to other problems. The examples of spectrogram of each environmental sound type are presented in Figure 3.2. It can be seen that each spectrogram clearly displayed different characteristics. Based on different characteristics of the spectrogram displayed, this thesis can use a different characteristic of spectrogram for classification.

3.2 Singing Word Recognition Problem

The purpose of our research is to recognize a singing voice with instrumental interference. Our system take polyphonic music audio signal as input, which was sampled from music CD recording and different music genres are included in the experiments such as rock ,hard rock, soft rock , dance , hip-pop, soul, r&b ,folk and acoustic. The files are all from different artists. the following two issues are :

- Singing voice different from speech because of the differences between speech and singing voice such as duration of voice sound, loudness, pitch, vibrato, formant, rhythm and rhyme [58]. It is difficult to use algorithm of speech recognition to solve this problem.
- In polyphonic music recordings, the instrumental interference is treated as the noise source that causes degradation to the intelligibility of the singing voice signal.

The goal of this research is to solve singing voice recognition without using any method to separate music in environment. Especially, the audio music with instrumental interference is treated as the noise source degraded the performance of recognition system.

3.2.1 Methodology

Figure 3.3 shows a diagram of singing voice recognition algorithm. Start by reading the audio file. Since each sound was captured from different songs. Therefore, the time interval of each sound is not equal, depending on a singer.

For this reason this research applies time-scale modification algorithm to make the time interval of each sound to be equal. An audio signal is analyzed and calculated with the short-time Fourier transform (STFT), to generate a spectrogram. After that a classifier, such as Feed-Forward neural network and K nearest neighbors (KNN) are experimentally chosen.

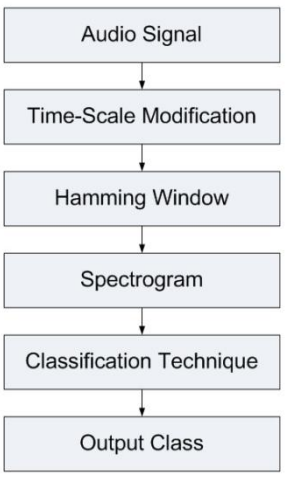

Figure 3.3 Flowchart of the singing voice recognition algorithm.

A spectrogram of each word is viewed as a matrix that describes the time waveform energy distribution in the joint time-frequency domain. Because a

spectrogram of each singing word is a 2D array, a spectrogram of each singing word is arranged to a feature vector before classification method , as shown in Figure 3.4.

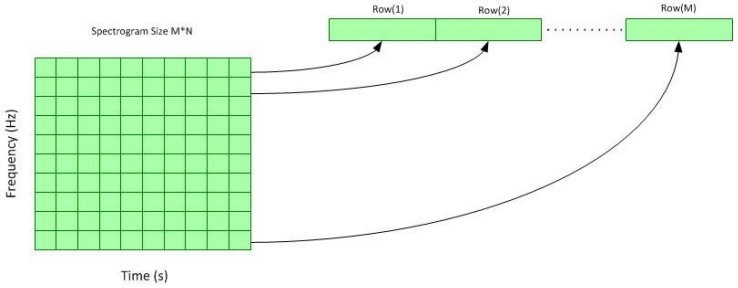

Figure 3.4: The method to convert a spectrogram images matrix to a spectrogram images vector by using row data spectrogram.

3.2.2 Data Collection

First, we investigated the performance of a spectrogram of audio features to solve the problem of singing voice recognition and provide an empirical evaluation on two data sets. The first database, denoted as DB-THS, is a collection of songs randomly chosen from Thai popular music CDs. It contains over 1500 Albums. DB-THS consists of 12 Thai One syllable singing word, 7200 sound samples, and 600 for each words. The singing words were selected from the most frequently in the all song. The considered singing words are shown in Table 3.1.

Table 3.1 DATABASES DB-THS USED IN EXPERIMENTS

Class.	Singing word	Time duration(min-max)	Pronunciation (in Thai)
1	"คน"	0.65s-2.95s	"kon"
2	"ความ"	0.26s-0.60s	"kwarm"
3	"เคย"	0.33s-0.62s	"koey"
4	"ใคร"	0.33s-0.70s	"krai"
5	"ใจ"	0.44s-1.38s	"jai"
6	"ฉัน"	0.26s-1.23s	"chan"
7	"ที่"	0.26s-0.54s	"tee"
8	"เธอ"	0.23s-0.78s	"ther"
9	"มี"	0.28s-0.86s	"mee"
10	"รัก"	0.18s-1.48s	"rug"
11	"รู้"	0.28s-0.47s	"roo"
12	"เรา"	0.26s-0.73s	"rao"

The second database, denoted as DB-TH-ENG. DB-TH-ENG, is a collection of songs randomly chosen from English and Thai popular music CDs. For the second dataset that consisting of two or more words. DB-TH-ENG consists 12 singing word. We used five words in English and seven words in Thai. DB-TH-ENG that contains 7200 sound samples, 600 for each word. The singing word was selected from the most frequently in the all song. The 12 considered singing words are showing in Table 3.2.

Table 3.2 DATABASES DB-THS USED IN EXPERIMENTS

Class.	Singing word	Time duration(min-)	Pronunciation (in Thai)
1	I love you	0.65s-2.95s	
2	Love you	0.57s-2.92s	
3	Together	1.04s-2.11s	
4	Tomorrow	1.07s-6.63s	
5	Yesterday	0.81s-5.90s	
6	"ความรัก"	0.52s-3.65s	"kwarm-luck"
7	"คิดถึง"	0.88s-1.11s	"kit-thun"
8	"ใครสักคน"	0.99s-4.62s	"krai-sak-kon"
9	"ไม่เคย"	0.41s-1.99s	"mai-koey "
10	"ไม่มี"	0.57s-1.17s	"mai-mee"
11	"รักเธอ"	0.47s-1.93s	"ruk-ther"
12	"หัวใจ"	0.73s-1.46s	"hua-jai"

All sample files in DB-THS and DB-TH-ENG were coded in stereo of frequency 44.2 kHz with 128/s bit rate. All audio signals were converted to mono and down-sampling types at rate of 8,000 Hz.

The following comparisons are conducted. The objective of the experiments is to investigate which features, i.e. (1) Mel Frequency Delta Cepstral Coefficients (MFCC); and (2) Linear Prediction Coefficients (LPC); and classifier, i.e. feed forward neural network and k-nearest neighbor network (KNN).

- The average accuracy based on spectrogram, MFCC, and LPC features versus a feed forward neural network.
- The average accuracy based on spectrogram, MFCC, and LPC features versus a KNN.

- The average accuracy between the feed forward neural network and the KNN with spectrogram, MFCC, LPC, and MP features.
- The average accuracy based on spectrogram, MFCC, and LPC features versus a feed forward neural network with different window sizes.
- The average accuracy based on spectrogram, MFCC, and LPC features versus a KNN with different window sizes.
- Computational Speed Tests on Spectrogram Features.
- Experiment dimension reduction on Spectrogram Features.

3.3 Environmental sound Recognition Problem

This research considers the task of classifying the environment sounds to understand the scene surrounding the audio.

3.3.1 Methodology

Fig 3.5 shows a diagram of Environment sound recognition algorithm. An audio signal is analyzed and calculated with the short-time Fourier transform (STFT) to generate a spectrogram. In Environment sound recognition, we send each column of spectrogram to classifier. After that a classifier, such as Feed-Forward neural network and k Nearest Neighbors (k-NN) are experimentally chosen.

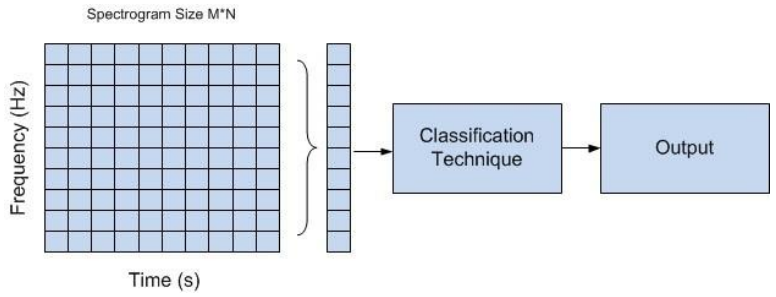

Figure 3.5 Flowchart of the Environment sound recognition algorithm.

3.3.2 Data Collection

The third database denoted as DB-ENG was a collection of 20 different types of environmental sounds. As summarized in Table 4.1, the natural sound clips are obtained from famous sound database such as the BBC [23] and Sound Ideas - The General Series 6000. The sounds were recorded in wav format to avoid introducing artifacts in our data. All audio signals were converted to mono and down- sampled from the CD sampling rate of 44.1 kHz to 16 kHz. The environment sound types were chosen so that they are made up of non-speech and non-music sounds.

The experiment consisted of the test on 20 different types of unstructured environmental sound which are: Car engine, Construction, Crowd Applause, Crow Cheering, Fire, Helicopter, Office, Out- door Sounds - Forest, Outdoor Sounds - Road, Restaurant Stores, Transportation - Motorcycle (start and idle), Transportation- Train, Water, Weather - Wind, Weather - Rain and Thunder, Household, Airplane, Water(Ocean), Chicken Farm, Auto Racing. The length of each sound is listed in the last column of Table 3.3.

Table 3.3 The length and class of 20 different types of unstructured environmental sound.

Class	Type of Environmental Sound	Time (Minutes)
1	Car engine	14.5
2	Construction	12.1
3	Crowd Applause	13.8
4	Crowd Clamor	15.7
5	Fire	13.5
6	Helicopter	14.2
7	Office	15.2
8	Outdoor Sounds - Forest	15.8
9	Outdoor Sounds - Road	15.9
10	Restaurant Stores	15.9
11	Transportation - Motorcycle	13.8
12	Transportation - Train	14.1
13	Water	15.8
14	Weather - Wind	15.7
15	Weather - Rain and Thunder	12.8
16	Household	16.6
17	Airplane	12.2
18	Water(Ocean)	20.0
19	Chicken Farm	22.3
20	Auto Racing	23.2

The following comparisons are conducted. The objective of the experiments is to investigate which features, i.e. (1) Mel Frequency Delta Cepstral Coefficients (MFCC); (2) Linear Prediction Co-efficients (LPC); and (3) Matching Pursuit (MP), and classifier, i.e. feed forward neural network and k-nearest neighbor network (KNN), are suitable for recognizing the environmental sounds. Since the

details of extracted features depend upon the sampling rate, following comparison of different sampling is also investigated.

- The average accuracy based on spectrogram features versus a feed forward neural network.
- The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a feed forward neural network.
- The average accuracy based on spectrogram features versus a feed forward neural network.
- The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a feed forward neural network.
- The average accuracy spectrogram features versus a KNN.
- The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a KNN.
- The average accuracy between the feed forward neural network and the KNN with spectrogram, MFCC, LPC, and MP features.
- The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a feed forward neural network with different window sizes. The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a KNN with different window sizes.
- The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a feed forward neural network with different sampling rates.
- The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a KNN with different sampling rates.

CHAPTER IV

RESULTS AND DISCUSSION

The experimental environment is Dell OptiPlex 755 Desktop, Intel Core 2 Duo E6750 Processor operating at 2.66-GHz and 6 GB total memory, running Microsoft Windows 7 64bit. All data sets were divided into four groups of equal sizes. Then, arbitrarily selected three groups were used for training and the rest is used for testing. For cross-validation procedure, the same process was repeated 50 times with the different training and test sets, to ensure that all samples are included at least once in the test set. The experimental results are shown in the following sections.

4.1 Singing Word Recognition Problem

For Singing Word Recognition Problem, Each environmental in DB-THS and DB-TH-ENG in table 3.1 and 3.2 was segmented into several sub-signals. These sub-signals were randomly divided into four groups of equal sizes. Then, arbitrarily selected three groups were used for training and the rest is used for testing. In each experiment, we performed 50 runs on each classifier to obtain statistically reliable results. The mean recognition rate was calculated based on the error average for one run on test set. The following classification techniques are commonly used for speech/speaker recognition or have, in the past, been used for this application domain. They are:

- K nearest neighbor method
- Artificial Neural Networks
- Minimum least square linear
- Normal densities based linear
- Naive Bayes
- Parzen
- Radial basis neural network
- Decision tree

4.1.1 Experimental on DB-THS Dataset

Based on our experimental setup, we use a window of 512 points with a 25% overlap. This corresponds to the window size used for all feature extractions. First, we applied Waveform Similarity Based Overlap-Add (WSOLA) for time-scale modification in each signing word audio data was applied to equalize to the lengths of all samples. A time of interval of each singing word equal 0.5 seconds.

The overall recognition accuracy from K nearest neighbor method, Artificial Neural Networks, Minimum least square linear, Normal densities based linear, Naive Bayes, Parzen, Radial basis neural network and Decision tree are summarized in figure 4.1.

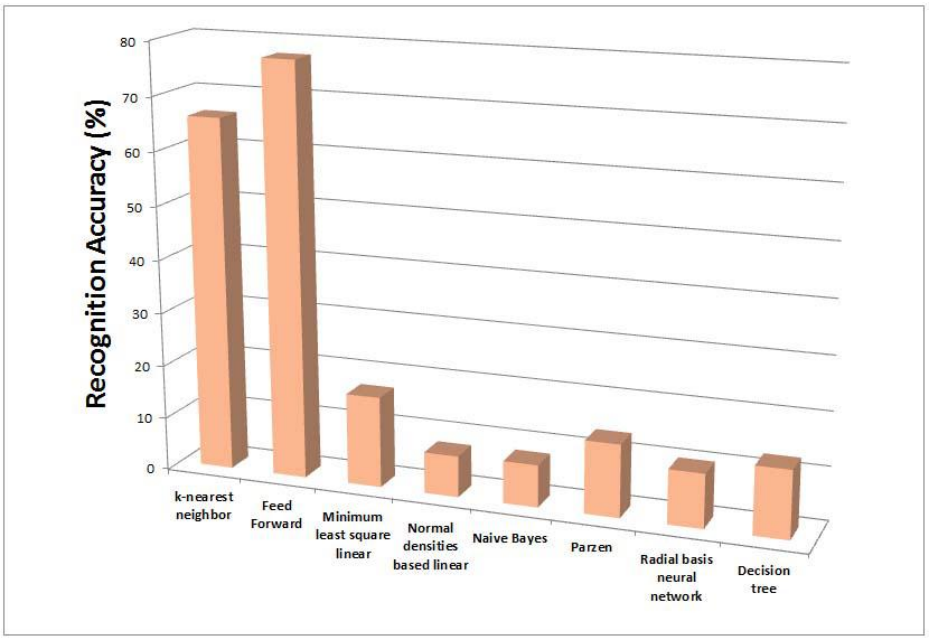

Figure 4.1: Overall recognition rate comparing 8 classifier using spectrogram as features on DB-THS dataset.

As shown in this figure, K nearest neighbor and Artificial Neural Networks performed better than another classification technique. As a result of this experiment, we will consider the use of K nearest neighbor method and Artificial Neural Networks as the most powerful. These researches examine the results from varying the number of neighbors and using the same for each environment type. The

By using feed forward neural network (ANN) with spectrogram feature, this research examines the results from verity number of hidden neural unit and using the same for each singing word. The overall recognition rates by varying are given in Fig 4.2. The highest recognition rates were obtained using 20 hidden neural units, with the average accuracy of 78.60%.

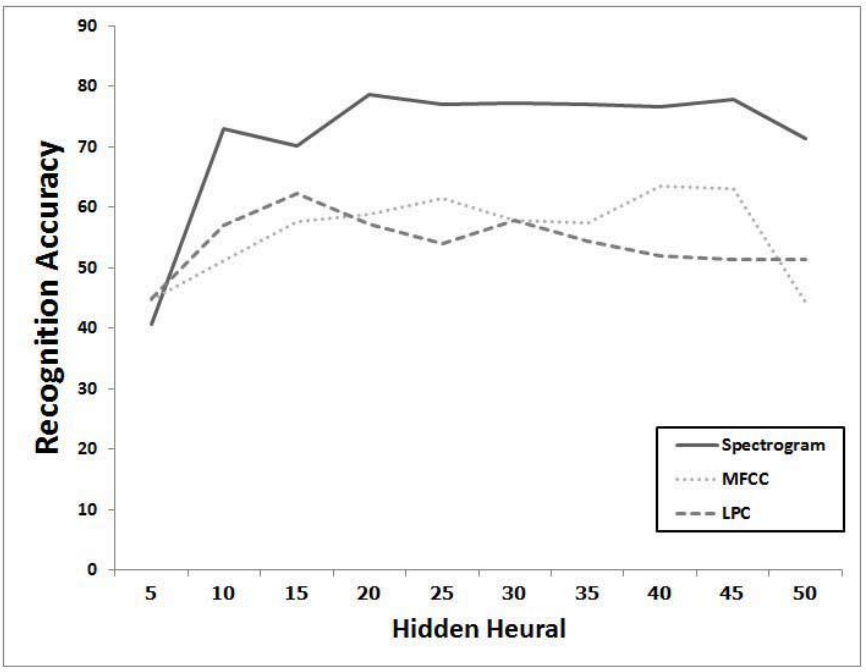

Figure 4.2: Preliminary experiments to obtain the candidate number of hidden neurons based on the features of Spectrogram, MFCC, and lpc on DB-THS of hidden neural = 20.

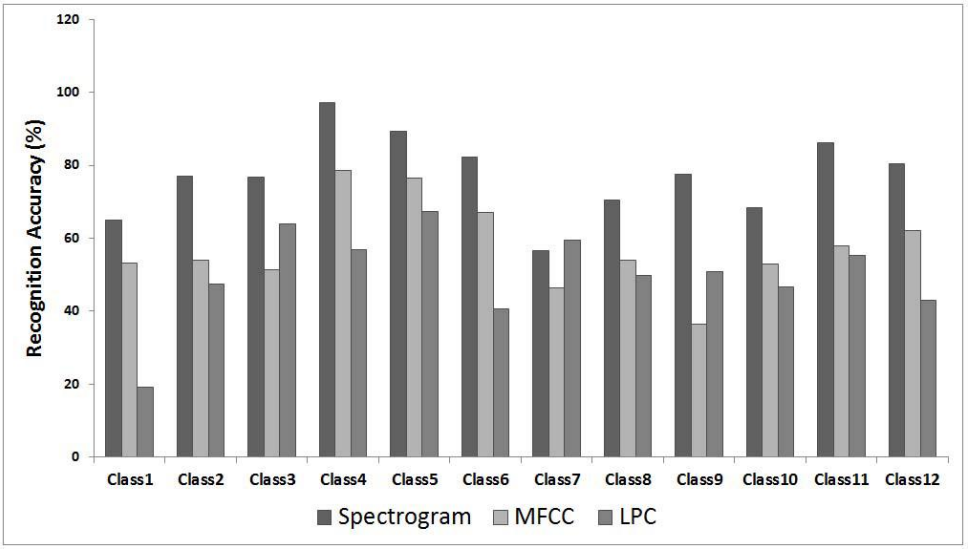

Figure 4.3: Overall recognition rate (ANN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with DB-THS data set.

An interesting benchmark is showing in figure 4.3 when ran the same experiments using all feature, including MFCC and lpc. In lpc feature, this research

use 13 order of the prediction filter polynomial. MFCC feature, this research use 13 number of cepstra to return, 0.6 for exponent for liftering,0.97 for apply pre-emphasis filter, lowest band edge of mel filters 133.3 Hz, highest band edge of mel filters 8000Hz, 40 numbers of warped spectral bands to use. The frequency warping scale used for filter spacing in MFCC is the Mel (Melody) scale.

This research compares the overall recognition accuracy using Spectrogram, MFCC, and LPC for 12 classes of sounds using feed forward neural network (ANN) with 20 hidden neural units in Fig. 4.3. As shown in this figure, Spectrogram features demonstrate the ability to better. They perform better than MFCC features in 11 of the examined classes while producing poor results in the case of 1 other class. Compared with the LPC features, Spectrogram feature were better in every class. The having the highest recognition rate at 97.08% in class 4 (Pronounced "krai").

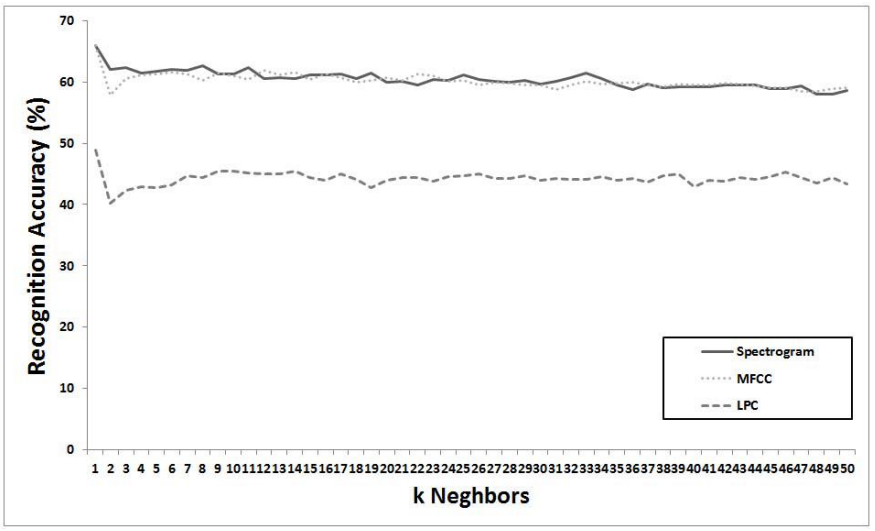

Figure 4.4: Preliminary experiments to obtain the candidate number of K nearest neighbors based on the features of Spectrogram, MFCC, and lpc. on DB-THS.

For completeness, this research examines the classification by using K-nearest neighbors (KNN) with spectrogram. These researches examine the result from verity number of K by using same windows size 512 and same data in Feed Forward Neural Network. Figure 4.4 was showing overall recognition accuracy using K-nearest neighbors (KNN) with a verity number of K for each singing word. The overall recognition rate was obtained using K=1 with average accuracy of 66.0%. The performance is not high compared with Feed Forward Neural Network.

This research compares the overall recognition accuracy using Spectrogram, MFCC, and LPC for 12 classes of sounds using K-nearest neighbors (KNN) using K=1 in Fig 4.5. As shown in figure 4.4 and Fig. 4.5, Spectrogram a feature in recognition performance is not much higher than MFCC. They perform better than MFCC features in 6 of the examined classes while producing poor results for 6 other classes. Compare with the LPC features, Spectrogram feature were better

in every class. By using K-nearest neighbors (KNN) with spectrogram feature, the highest recognition rate was 78.77% in class 6 (Pronounced "chan").

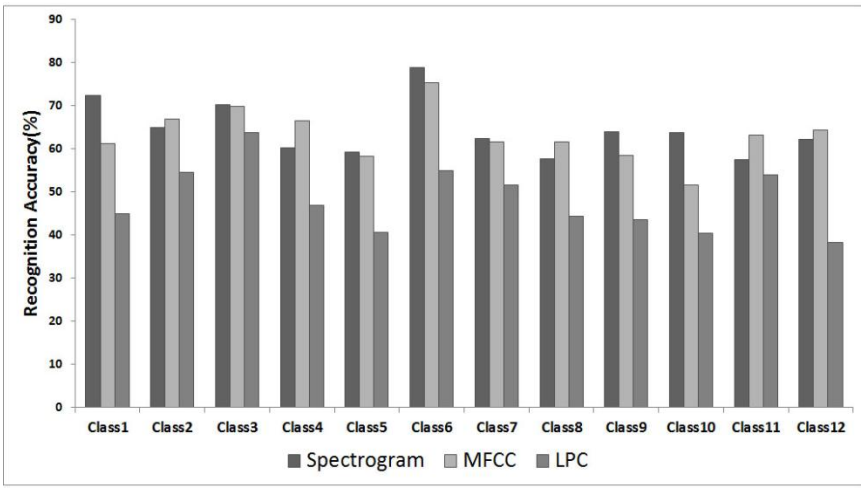

Figure 4.5: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-THS.

4.1.2 Experimental on DB-TH-ENG Dataset

Singing voice recognition experiments were conducted using Longer term and cross language. The data in this section can be used to handle English and Thai music data. Using the same experiment settings associated with section 4.1.1. This research obtained the accuracies of this section.

This research compares the overall recognition accuracy K nearest neighbor method, Artificial Neural Networks, Minimum least square linear, Normal densities based linear, Naive Bayes, Parzen, Radial basis neural network and Decision tree in figure 4.1.

As shown in this figure, K nearest neighbor and Artificial Neural Networks perform better than classification techniques. As a result of this experiment, we will consider the use of K nearest neighbor method and Artificial Neural Networks as the most powerful.

Using the same experiment settings associated with 4.1.1 this research examines the classification by using feed forward neural network (ANN) with spectrogram. This research examines the results from verity number of hidden neural unit and using the same for each environment type. The overall recognition rates by varying are given in Fig. 4.7. The highest recognition rate was obtained using 45 hidden neural units, with an average accuracy of 92.93%. In this Experiment, spectrogram feature show a higher performance when used longer term data than the

others. An interesting benchmark is show in Figure 4.8 when ran the same experiments using all feature, including MFCC and lpc. This research compares the overall recognition accuracy using Spectrogram, MFCC, and LPC for 12 classes of sounds using feed forward neural network (ANN) with 45 hidden neural units in Figure 4.8. As shown in this figure, Spectrogram features demonstrate the ability to better. They perform better than MFCC features in 7 of the examined classes while producing poor results in the case of 5 other classes. Compared with the LPC features, Spectrogram feature were better in every class. The having highest recognition rate at 96.12% in class 4 (Pronounced ""mai-mee""). Especially, in figure 4.8spectrogram feature can recognize Cross-Language Music Data.

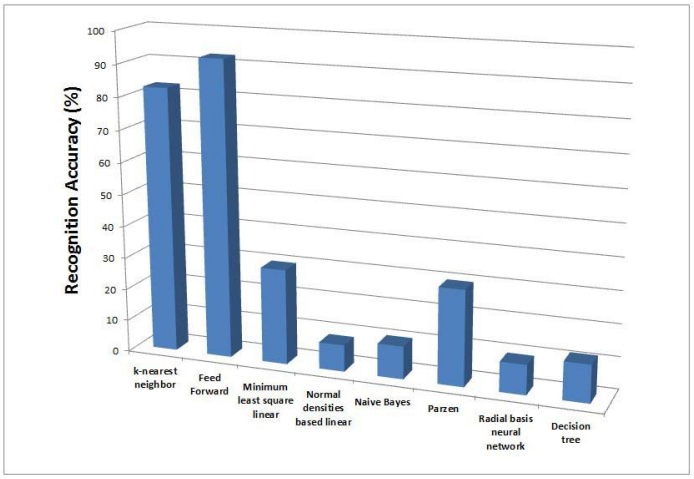

Figure 4.6: Overall recognition rate comparing 8 classifier using spectrogram as features on DBTHS-ENG dataset.

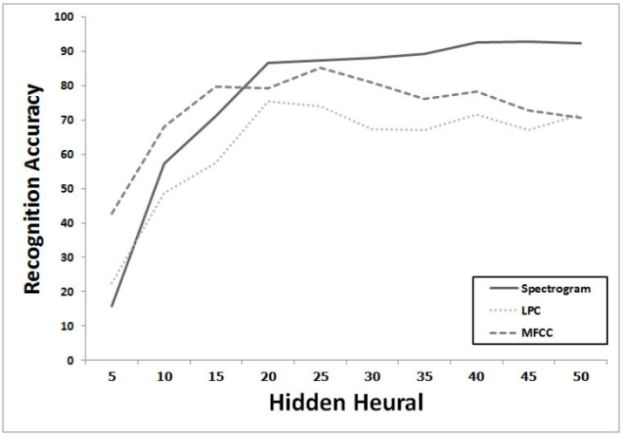

Figure 4.7: Preliminary experiments to obtain the candidate number of hidden neurons based on the features of Spectrogram, MFCC, and lpc on DB-TH-ENG dataset of hidden neural = 45.

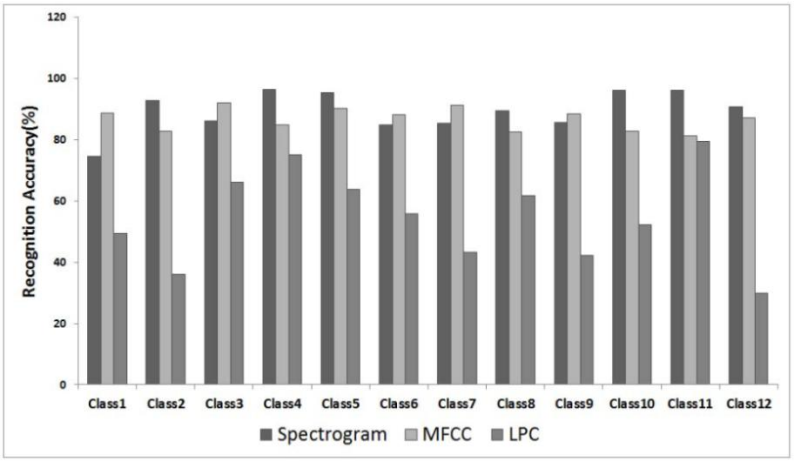

Figure 4.8: Overall recognition rate (ANN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-TH-ENG dataset.

For completeness, this research examines the classification by using K-nearest neighbors (KNN) with spectrogram. This research examines the result from verity number of K by using same windows size 512 and same data in Feed Forward Neural Network. Figure 4.9 was showing overall recognition accuracy using K-nearest neighbors (KNN) with a verity number of K for each singing word. The overall recognition rate was obtained using K=1 with accuracy of 82.67%. The performance is not high compared with Feed Forward Neural Network.

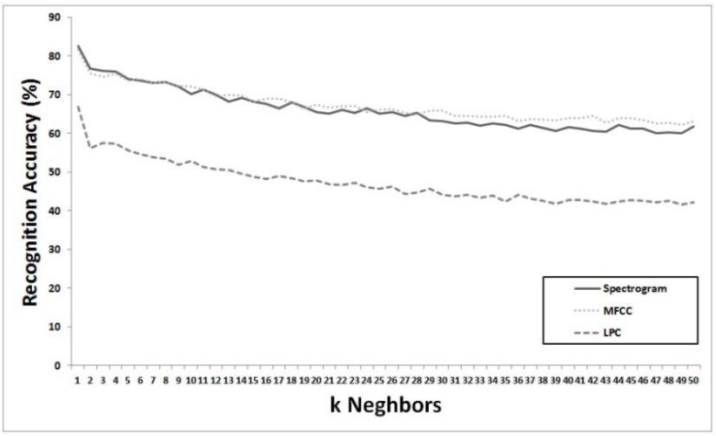

Figure 4.9: Preliminary experiments to obtain the candidate number of K nearest neighbors based on the features of Spectrogram, MFCC, and lpc. on DB-TH-ENG dataset

This research compares the overall recognition accuracy using Spectrogram, MFCC, and LPC for 12 classes of sounds using K-nearest neighbors (KNN) using K=1 in Fig. 4.10. As shown in figure 4.9 and Fig 4.10, Spectrogram features in recognition performance are not much higher than MFCC. They perform better than MFCC features in 7 of the examined classes while producing poor results

in the case of 5 other classes. Compared with the LPC features, Spectrogram feature were better in every class. The having the highest recognition rate was 95.09% in class 3 (Pronounced "Together").

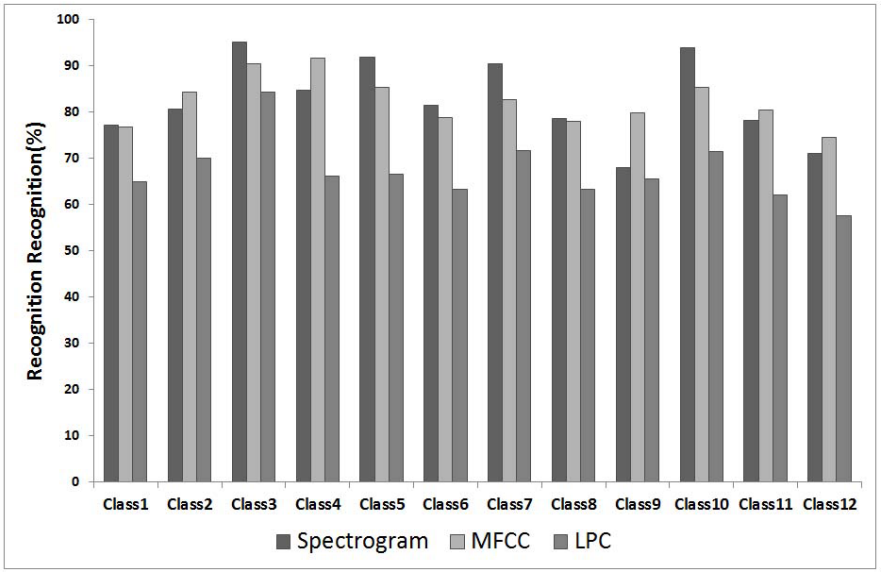

Figure 4.10: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-TH-ENG dataset.

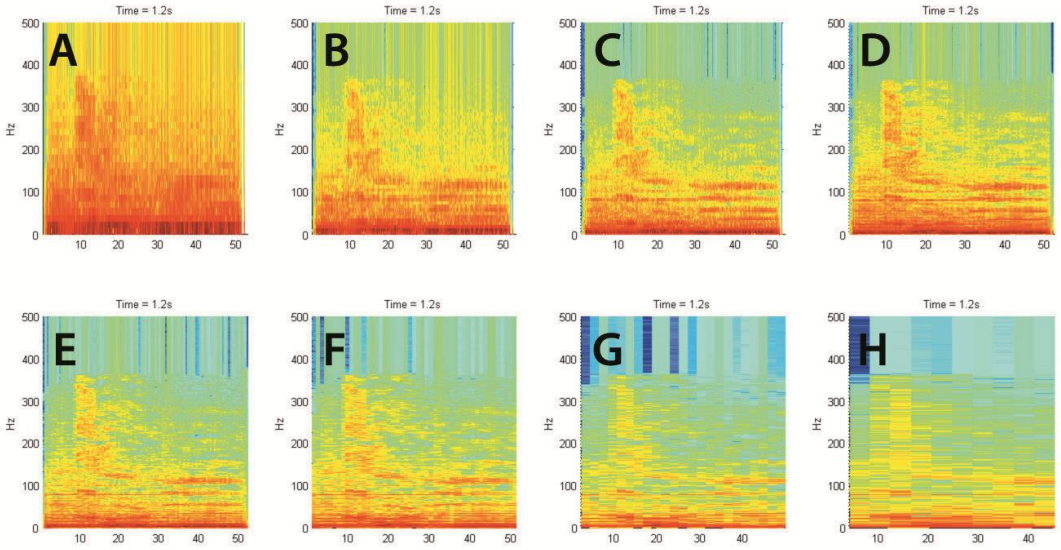

Figure 4.11: Example of spectrogram obtained from different sizes of windowed segment a) 64, b) 128, c) 256, d) 512, e) 1024, f) 2048, g) 4096, h) 8192.

4.1.2 Experiment on different sizes of windowed segment on DB-TH-ENG and DB-THS

A spectrogram can be obtained from different sizes of windowed segment. Figure 4.11 show a spectrogram obtained from different sizes of windowed segment. From the figure this research can see a different characteristic of a spectrogram obtained from different sizes of windowed segment. This research wanted to find out the effect in classification performance of different sizes of windowed segment and size of windowed segment that gives the best accuracy rate in classification for the data set.

This experiment used the same data in Table 3.1 and Table 3.2. The following window sizes were experimented: 4096, 2048, 1024, 512 and 256 . All window sizes were used overlapping at 25%.

Figure 4.12 and 4.13, show the average accuracy obtained from different classifier with a different size of windows segment by using k-nearest neighbors (KNN) with K=1 and feed forward neural network with 20 hidden neural units on DB-THS. As showing Figure 4.12 and 4.13 and, a spectrogram that created from a large size of windows segment gives better classification accuracy than a spectrogram that created from a small size of windows segment. For another feature, by using feed forward a spectrogram perform better than MFCC and lpc in all sizes of windows segment. Comparing to the results from Figs 6 and 4.12, these results are interesting. When increasing a window size of 512 to 4096 in the spectrogram feature. The Average recognition accuracy increased from 78.60% to 82.17%.

Figure 4.14 show the details of each group. As shown in this figure, Spectrogram features demonstrate the ability to better. They perform better than MFCC features in 12 of the examined classes. Compared with the LPC features, Spectrogram feature were better in every class. Compared to the results from Figure 7. When using a window size of 512 for spectrogram feature, the recognition performance than MFCC only 11 class. For K -nearest neighbor network, when window size is 256, MFCC provides higher accuracy than spectrogram. However, in other sizes, the results of spectrogram are better than MFCC. When this research used the same data in Table 3.2. The following window sizes were experimented: 4096, 2048, 1024, 512 and 256. All window sizes were used overlapping at 25%. Figure 4.16 and 4.17, show the average accuracy obtained from different classifier with a different size of windows segment by using feed forward neural network with 45 hidden neuron units and k-nearest neighbors (KNN) with K=1.

In case of feed forward network, similar results in DB-THS dataset a large window size achieves higher accuracy than a small window size for spectrogram features. The DB-TH-ENG dataset is the same effect. By using windows size 512, the highest average recognition rate was 89.43%. When the window size is change to 4096. The highest average recognition rate was 90.30%. Compared with other feature in the spectrogram, it also provides higher performance anyway. Figure 4.17 show the details of each group. As shown in this figure, Spectrogram features demonstrate the ability to better. They perform better than MFCC features in 12 of the examined

classes. Compared with the LPC features, Spectrogram feature were better in every class. Compared to the results from Figure 4.5. When using a window size of 512 for spectrogram feature, the recognition performance than MFCC only 7 class.

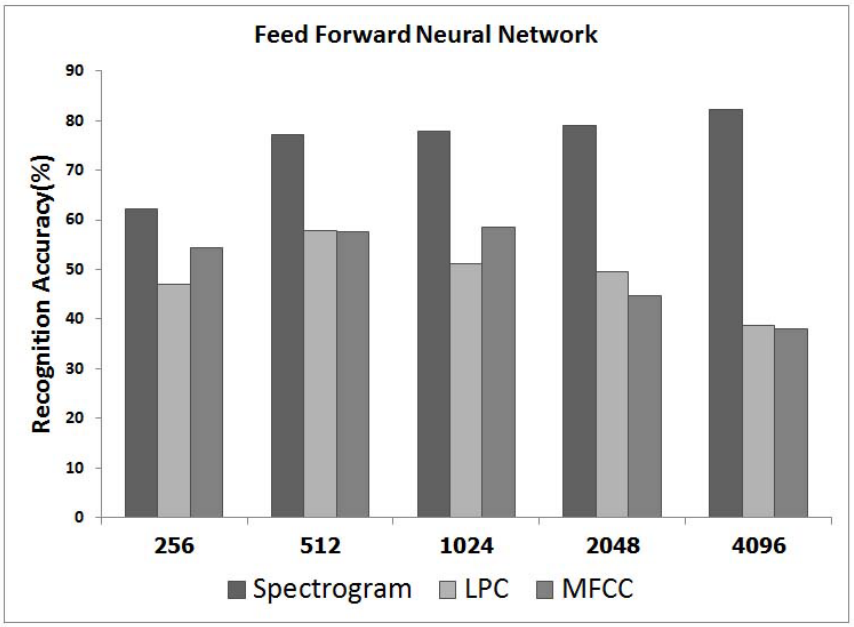

Figure 4.12: Average recognition performance of Feed-Forward Neural Networks on a spectrogram MFCC and lpc obtained from different sizes of windowed segment on DBTHS data set.

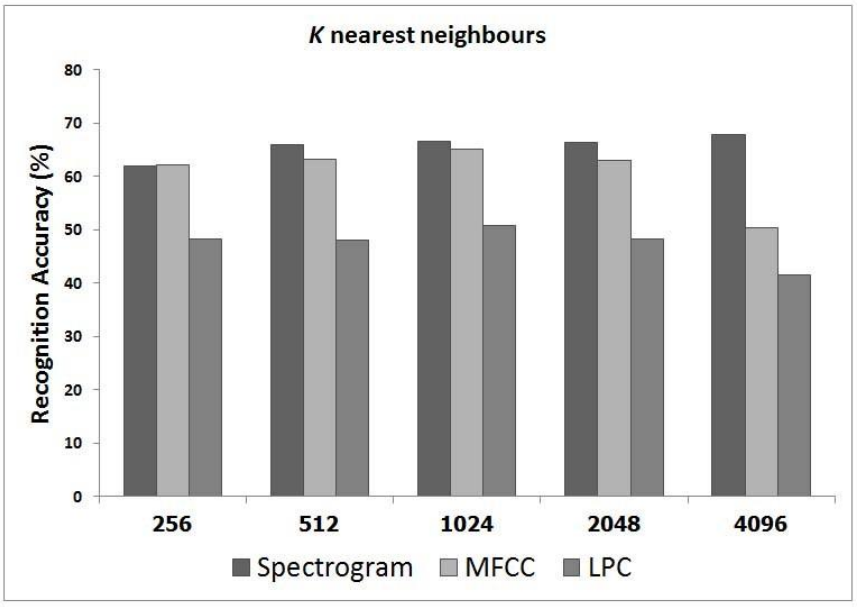

Figure 4.13: The comparison of recognition accuracy for different window sizes based on K = 1 nearest neighbor network and different features on DBTHS data set.

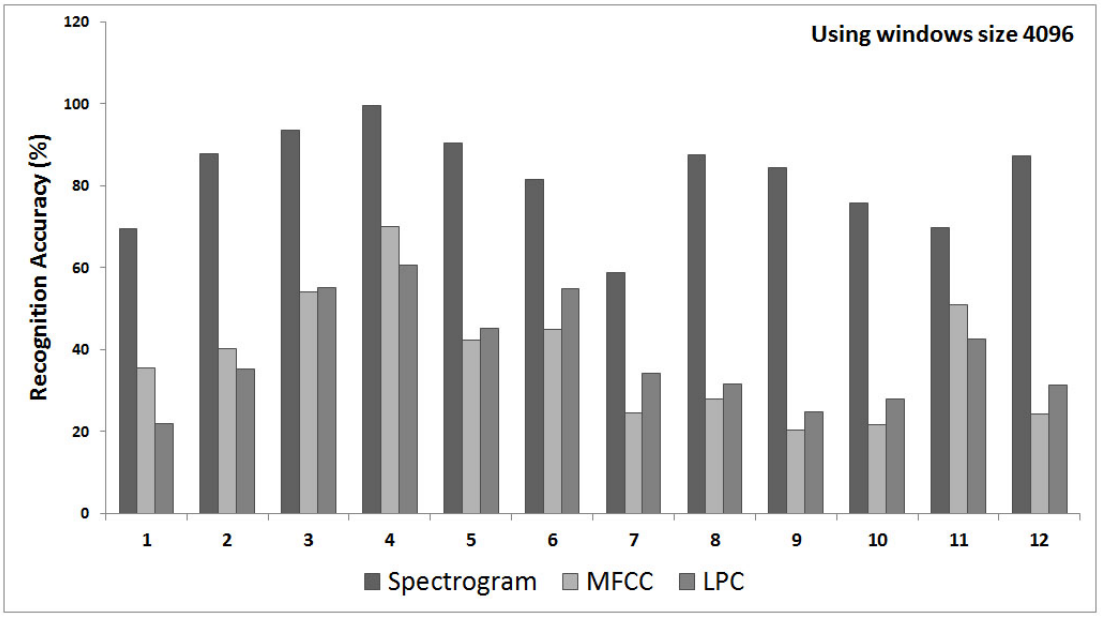

Figure 4.14: Overall recognition rate (ANN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DBTHS data set on Windows Size 4096.

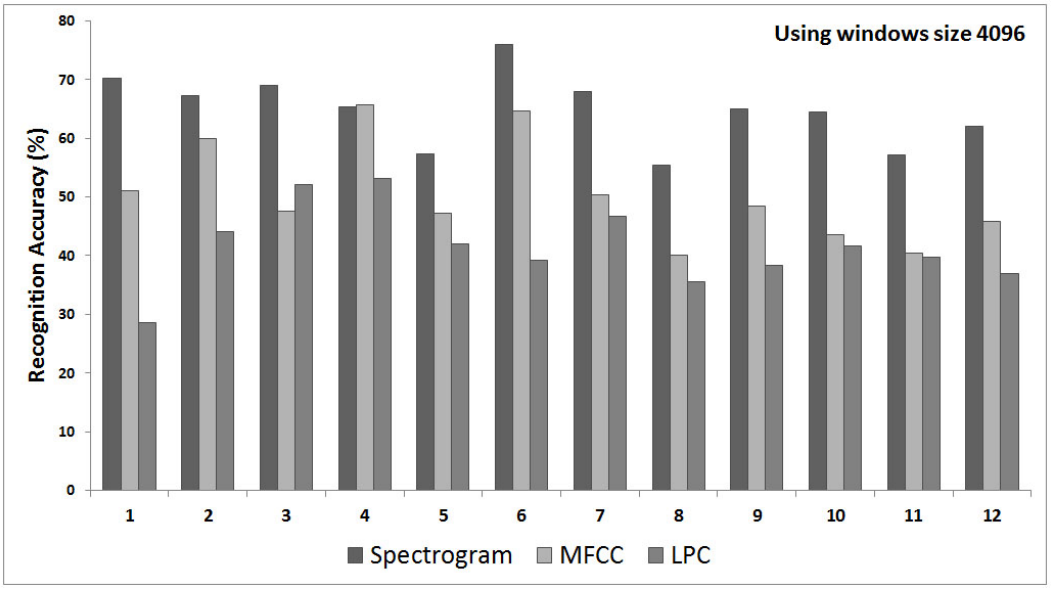

Figure 4.15: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-THS dataset by using windows size 4096.

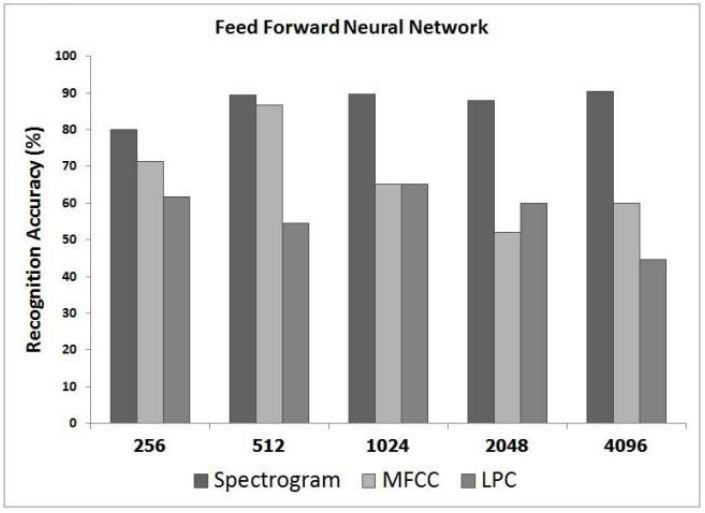

Figure 4.16: Average recognition performance of Feed-Forward Neural Networks on a spectrogram MFCC and lpc obtained from different sizes of windowed segment on DB-TH-ENG data set.

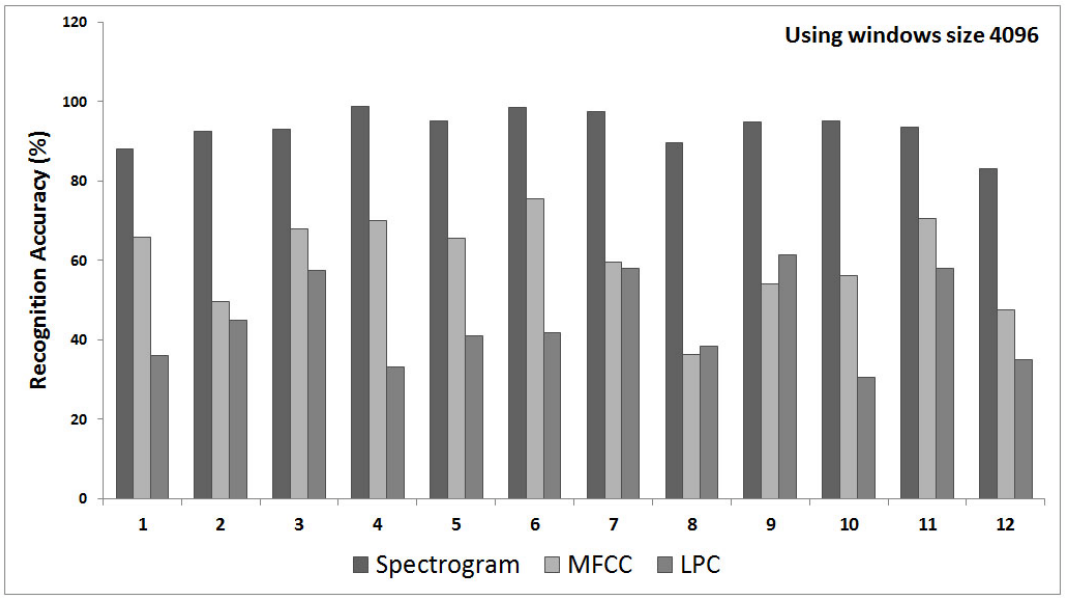

Figure 4.17: Overall recognition rate (ANN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-TH-ENG dataset by using windows size 4096.

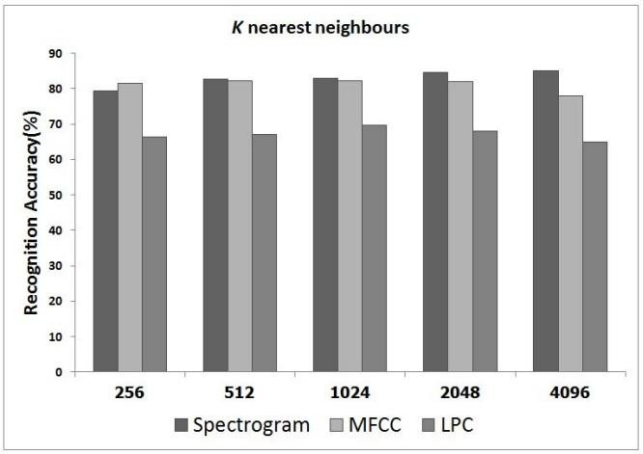

Figure 4.18: The comparison of recognition accuracy for different window sizes based on K = 1 nearest neighbour network and different features on DB-TH-ENG data set.

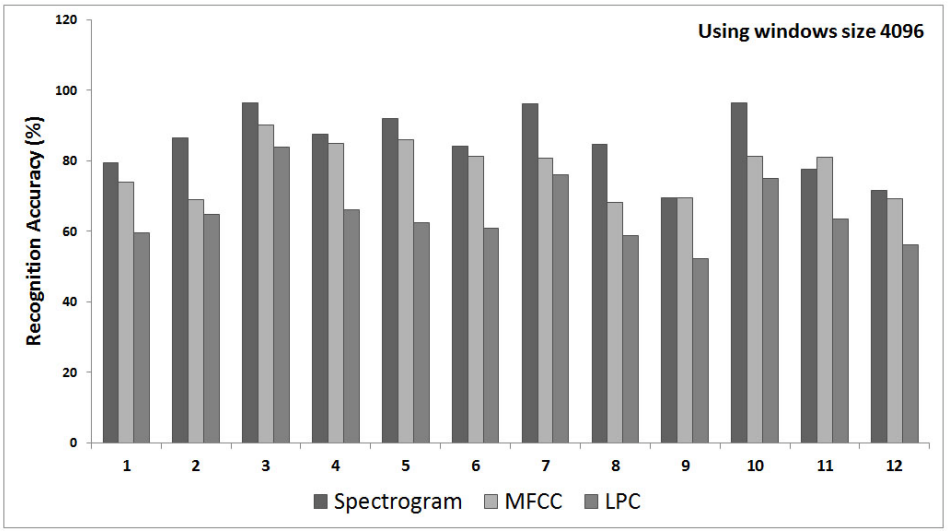

Figure 4.19: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-TH-ENG dataset by using windows size 4096.

Figure 4.18, For K -nearest neighbor network, when window size is 256, MFCC provides higher accuracy than spectrogram. However, in other sizes, the results of spectrogram are better than MFCC. Similar to neural feed forward network, a large window is better than a small window size. At maximum window size of 4096, the recognition rate is up to 85.16% for DB-TH-ENG dataset.

4.1.4 Experiment Dimension Reduction on Spectrogram Features.

However, the use of Spectrogram is also limited. When converting from Audio Signal to a Spectrogram. A dimension of Spectrogram is higher than other feature. Therefore, Spectrogram was using more memory more than the other types feature vector in same windows size. As compared to using the MFCC feature performance close to the Spectrogram. When used a same size of Windows to create spectrogram and MFCC feature. MFCC was used 13 numbers.

The spectrogram feature usually has a high dimensionality. The size of a spectrogram of each singing word can be calculated from:

$$SpectrogramSize = \left\lfloor \frac{N}{M-Q} \right\rfloor \times \frac{M}{2} \tag{4.1}$$

Where N is the length of a input signal x(n), M is length of windows w(n) and Q is hop size. For example, this research created a spectrogram from short input signal. Time duration of input signal is 0.5s with 8000 H z sample rate. The sampling rate defines the number of samples per unit of time (usually seconds). This signal sampling rates are 8000 Hz and time duration is 0.5s. Therefore, the length of input signal is 8000/2 = 4000. If spectrogram used windows size 256 and overlap 25%. This research can be calculated from the sample below. A size of spectrogram equal 2560. A dimension of spectrogram feature is very high; its dimensionality needs to be reduced. Dimensionality reduction is the transformation of high-dimensional data into a meaningful representation of reduced dimensionality.

This research apply 6 dimension reduction technique in this section for reduce spectrogram feature dimensions.

- Diffusion maps ('DiffusionMaps').
- Linear Local Tangent Space Alignment ('LLTSA')
- Principal Component Analysis('PCA').
- Stochastic Neighbor Embedding ('SNE')
- Symmetric Stochastic Neighbor Embedding ('SymSNE')
- t-Distributed Stochastic Neighbor Embedding ('tSNE')

First, the research performs an estimation of the intrinsic dimensionality of both dataset based on the method specified by method. Possible values for method are maximum likelihood estimator (MLE). In our experiments, this research set neighborhood range k1 and k2 in maximum likelihood estimator (MLE) to 6 and 20. After that, this research run dimension reduction technique on same data in Table 3.1 and Table 3.2.

Two classifiers, i.e. feed forward neural network and K-nearest neighbour (kNN), are deployed in this experiment. In section 4.1.3, a spectrogram that create from window of 4096 points with a 25% overlap show the highest average recognition rate. Then, this section used that window of 4096 points for spectrogram

feature. The research compare the overall recognition accuracy using Spectrogram and their combination for 6 dimension reduction technique in Fig. 4.20 , 4.21 , 4.22 and 4.23, After this research apply dimension reduction technique for reduce spectrogram feature dimension.

In Figure 4.20, this research listed the accuracies achieved with spectrogram, MFCC, LPC and spectrogram with dimension reduction technique for the testing data. This research can see from Figure 4.20 that spectrogram feature with dimension reduction technique recede the accuracies noticeably, compared to the results obtained from spectrogram feature without dimension reduction technique.

By using Feed forward neural network (ANN) with spectrogram feature with dimension reduction technique. The recognition performance is greatly reduced. Especially when singing word is short as show in figure 4.20 and 4.23.

However, we compared with K-nearest neighbour (kNN) with spectrogram feature with dimension reduction technique. Recognition performance was not reduced as much as the using Feed forward neural network (ANN)as show in figure 4.21 and 4.23. In particular, t-Distributed Stochastic Neighbor Embedding ('tSNE') techniques can provide performance equivalent to the spectrogram feature without dimension reduction technique.

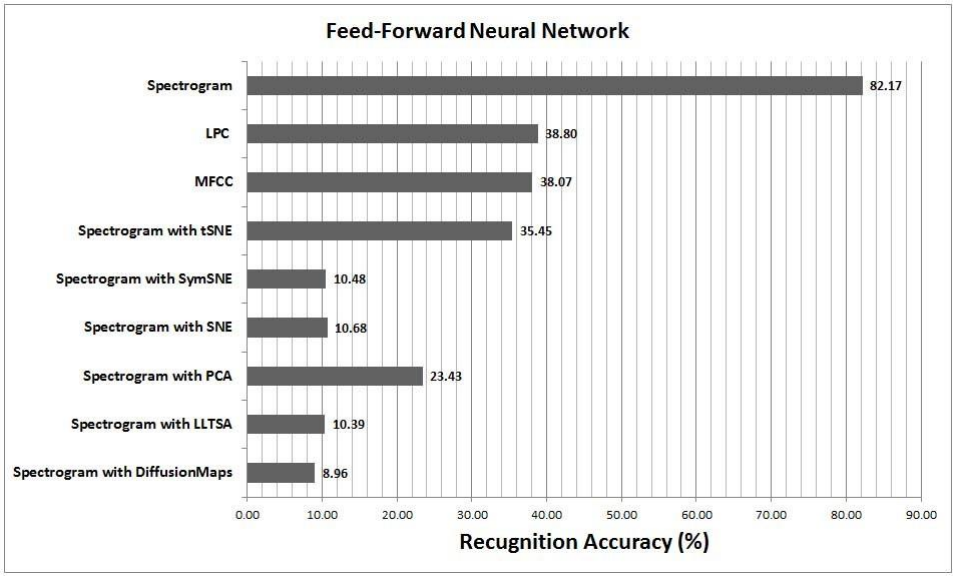

Figure 4.20: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-THS dataset.

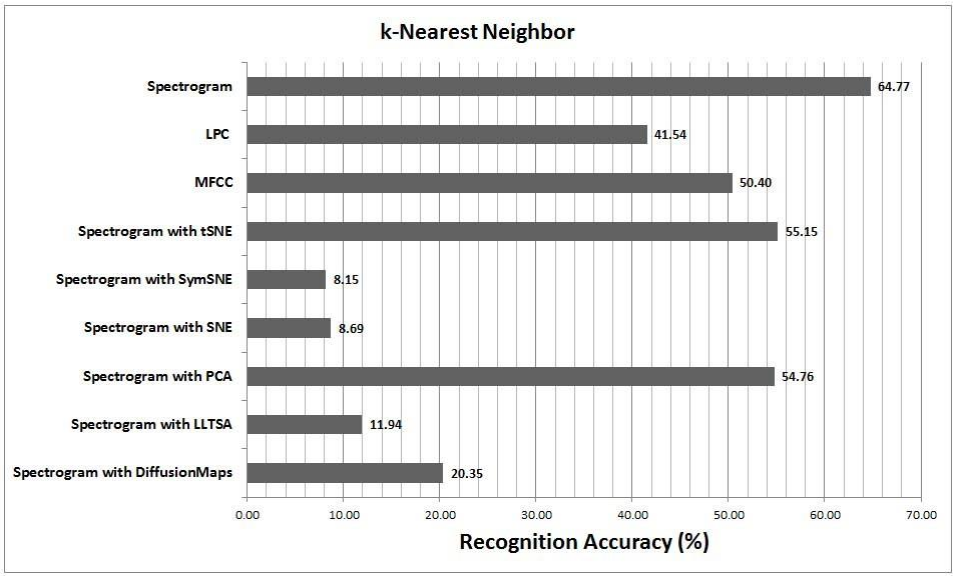

Figure 4.21: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC, and MFCC as features with a DB-THS dataset.

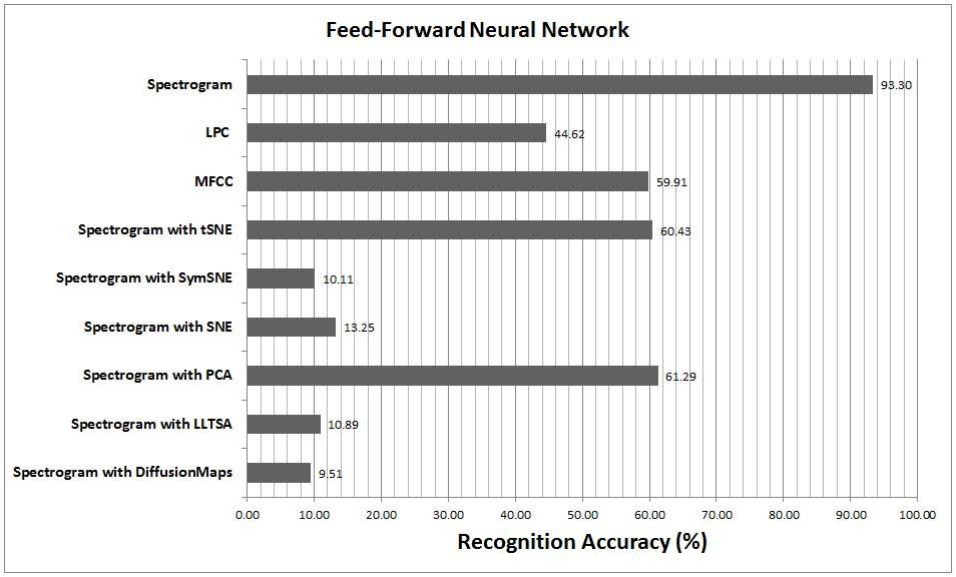

Figure 4.22: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC , and MFCC as features with a DB-TH-ENG dataset.

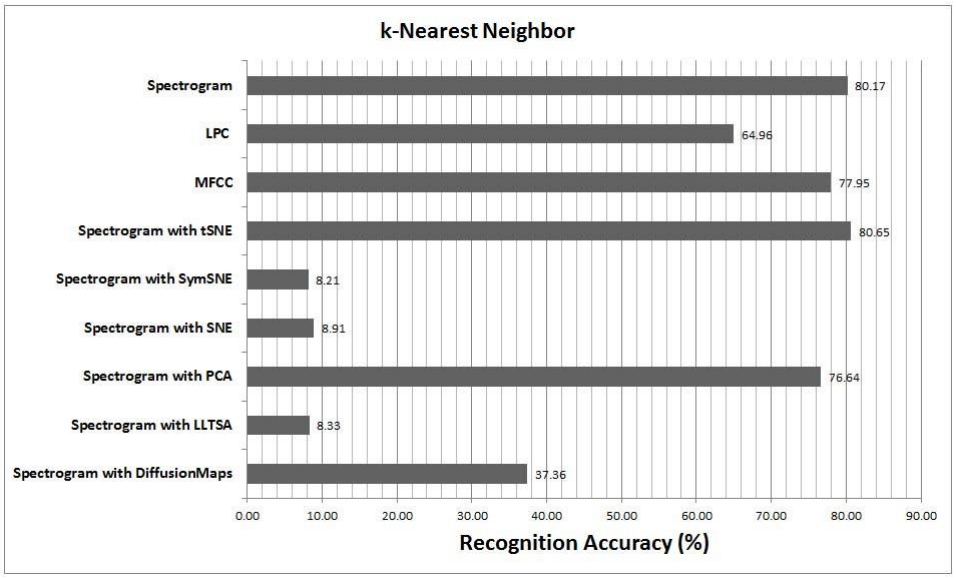

Figure 4.23: Overall recognition rate (KNN) comparing 12 classes using Spectrogram, LPC , and MFCC as features with a DB-TH-ENG dataset.

4.1.5 Computational Speed Tests

In this section this research show the execution time of the algorithm. The experimental en- vironment is Dell OptiPlex 755 Desktop , Intel Core 2 Duo E6750 Processor operating at 2.66-GHz and 6 GB total memory, running Microsoft Windows 7 64bit with Matlab 2011b 64bit. The programs were tested on 2 data sets acquired from Table 3.1 and Table 3.2.

The execution time report in this section, Our timing of the experiment in section The 12 considered singing word that contains 7200 sound samples, 600 for each word. Each singing words are randomly divided into four groups of equal sizes. Each group contains 150 sounds for each word. Then, arbitrarily selected three groups are used for training and the rest is used for testing. For cross-validation procedure, the same process is repeated 50 times. The average retrieval time with 50 times of a test set report in Table 4.1. By using feed forward n which has a computational time between 0.376s 0.526s for each singing word. However, when using knn computation time increases slightly. By using KNN which has a computational time between 0.485s 0.576s for each singing word.

Table 4.1 COMPUTATIONAL TIMES (Second) with spectrogram feature

	FFN		KNN	
Windows Size	DB-THS	DB-TH-ENG	DB-THS	DB-TH-ENG
256	0.52s	0.40s	0.57s	0.51s
512	0.48s	0.40s	0.57s	0.50s
1024	0.47s	0.39s	0.54s	0.48s
2048	0.44s	0.39s	0.57s	0.49s
4096	0.45s	0.37s	0.55s	0.49s

Table 4.2 COMPUTATIONAL TIMES (Second) with MFCC feature

	FFN		KNN	
Windows Size	DB-THS	DB-TH-ENG	DB-THS	DB-TH-ENG
256	0.65s	0.50s	0.72s	0.64s
512	0.60s	0.50s	0.72s	0.63s
1024	0.59s	0.49s	0.68s	0.60s
2048	0.55s	0.49s	0.72s	0.62s
4096	0.57s	0.47s	0.69s	0.62s

Table 4.3 COMPUTATIONAL TIMES (Second) with LPC feature

	FFN		KNN	
Windows Size	DB-THS	DB-TH-ENG	DB-THS	DB-TH-ENG
256	0.25s	0.19s	0.27s	0.24s
512	0.23s	0.19s	0.27s	0.24s
1024	0.22s	0.18s	0.25s	0.23s
2048	0.21s	0.18s	0.27s	0.23s
4096	0.21s	0.17s	0.26s	0.23s

Table 4.2 and 4.3 show a computational time by using MFCC and LPC feature. By using MFCC feature will take to process an average of 25% slower than the spectrogram feature. However, when compared with LPC feature. The time it takes to process up to 60% faster than spectrogram feature.

4.2 Environmental sound Recognition Problem

For Singing Word Recognition Problem, Each environmental in DB-ENG in table 3.3 were segmented into several sub-signals. These sub-signals were randomly divided into five groups of equal sizes. Then, arbitrarily selected four groups were used for training and the rest is used for testing. In each experiment, we performed 50 runs on each classifier to obtain statistically reliable results. The mean recognition rate was calculated based on the error average for one run on test set. The following comparisons are conducted. The objective of the experiments is to investigate which features, i.e. (1) Mel Frequency Delta Cepstral Coefficients (MFCC); (2) Linear Prediction Coefficients (LPC); and (3) Matching Pursuit (MP), and classifier, i.e. feed forward neural network and k-nearest neighbour network (KNN), are suitable for recognizing the environmental sounds. Since the details of extracted features depend upon the sampling rate, the comparison of different sampling is also investigated.

1.The average accuracy based on spectrogram features versus a feed forward neural network.

2.The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a feed forward neural network.

3.The average accuracy spectrogram features versus a KNN.

4.The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a KNN.

5.The average accuracy between the feed forward neural network and the KNN with spectrogram, MFCC, LPC, and MP features.

6.The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a feed forward neural network with different window sizes.

7.The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a KNN with different window sizes.

8.The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a feed forward neural network with different sampling rates.

9.The average accuracy based on spectrogram, MFCC, LPC, and MP features versus a KNN with different sampling rates.

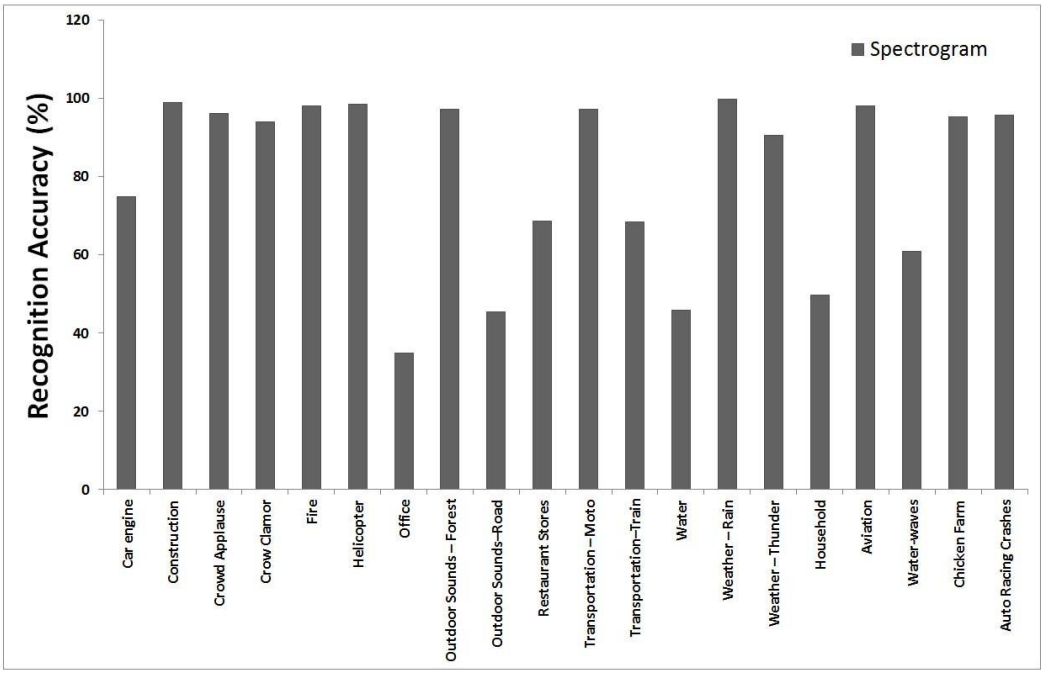

Figure 4.24: Classification accuracy obtained with Spectrogram features and Feed-Forward Neural Network.

4.2.1 Various Features with Feed-Forward Neural Network

This experiment is to test the feasibility of spectrogram feature on the accuracy of classification realized by a neural network. Each spectrogram is computed from a window of size 256 sampling points. There are 30 hidden neurons used in this experiment. Figure 4.24 summarizes the accuracy of each sound type. It can be seen that 12 sound types, i.e. Construction, Crowd Applause, Crow Clamor, Fire, Helicopter, Outdoor Forest, Transportation-Moto, Weather-Wind, Weather-Rain and Thunder, Aviation, Chicken Farm, and Auto Racing, achieve more than 90% accuracy rate. Another four sound types, i.e. Car Engine, Restaurant Stores, Transportation Train , and Water(Ocean) achieve between 75-60% accuracy rate. But the rest of sound types, i.e. Office, Outdoor-Road, Water, and Household, achieve rather poor accuracy.

Although the accuracy based on spectrogram feature is acceptable up to some certain degree, it is not conclusive that spectrogram feature is the most suitable feature for this recognition. Three other features, namely MFCC, LPC, and MP, are tested against spectrogram feature. For MFCC, the parameters are the following: number of cestrum's is 13; exponent for littering is 0.6; highest band edge of Mel filters is 4000 Hz; number of warped spectral bands is 40. The frequency warping scale used for filter spacing in MFCC is the Mel (Melody) scale. For MP, the signal is decomposed by using Gabor dictionary of 1120 atoms with dyadic scales

from 2 to 256 samples and translations in 0, 64, 128, and 192. In each atom, 35 different exponentially distributed modulation frequencies are considered.

Anyhow, from the MP decomposition of the segment, only the first 5 atoms are concerned. From these atoms, a four-dimensional feature vector from the mean, standard deviations of the modulation frequencies, and scales of the 5 atoms is formed. The classification results for the spectrogram, MP, MFCC, and LPC features by using feed forward neural network with 30 hidden neural are shown in Figure 4.25.

The number of hidden neurons is also another relevant factor affecting the accuracy. However, theoretically estimating this number is rather difficult. Several numbers of hidden neurons are tested. The accuracy based on the number of hidden neurons for each feature type is summarized in Figure 4.26. The highest accuracy is achieved when the number of hidden neurons is set to 30. But when this number is increased, the accuracy is gradually decreased. This may be due to the over fitting effect during the neural training process.

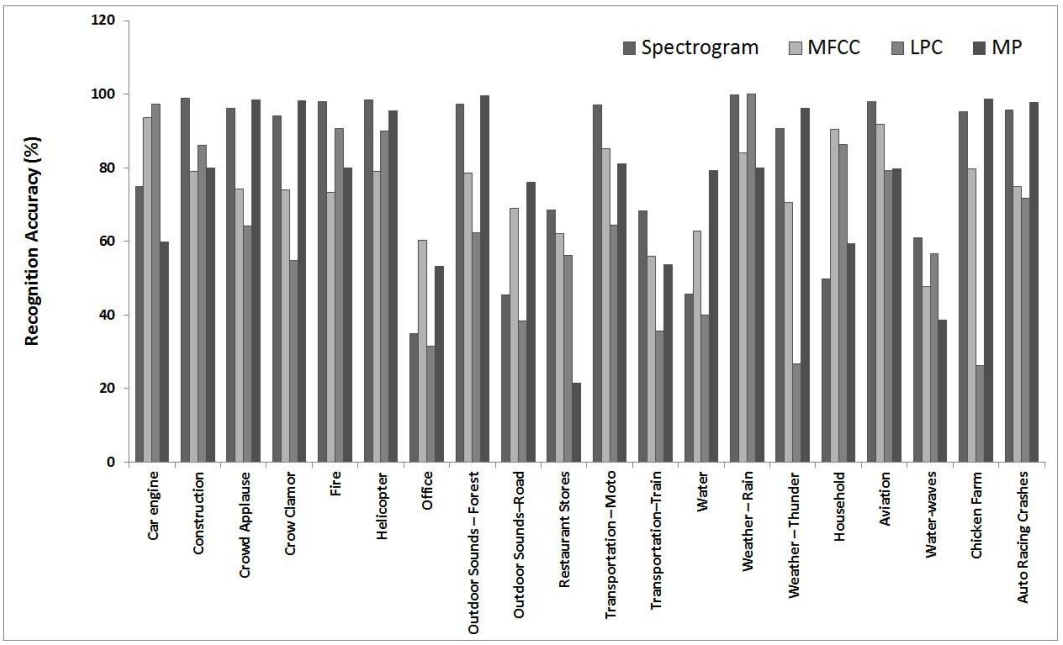

Figure 4.25: Classification accuracy obtained with different features and Feed-Forward Neural Network.

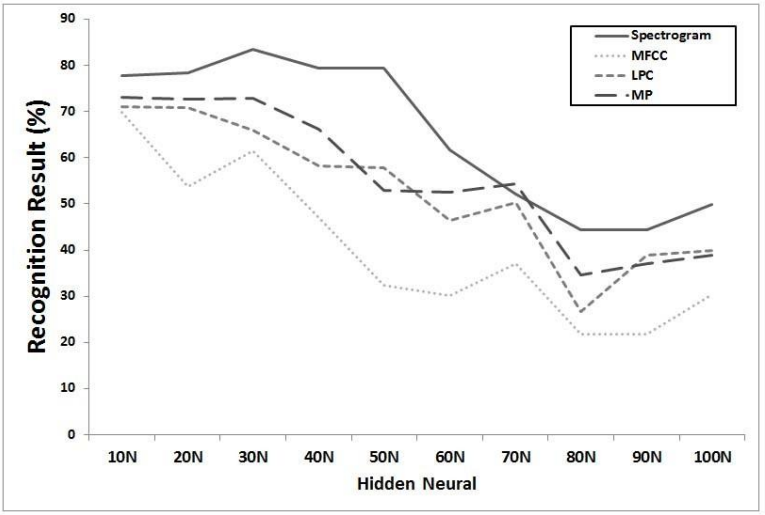

Figure 4.26: Classification performance of Feed-Forward Neural Network with varying number of hidden neural unit.

4.2.2 Various Features with K-Nearest Neighbours (KNN)

A similar experiment is conducted with a k-nearest neighbour classifier (KNN). The number of nearest neighbours is set to 10. The rationale of setting this number will be discussed later. Firstly, the spectrogram feature of each sound type is extracted and trained with a KNN. The testing result of each sound type is summarized in Figure 4.27. It can be seen that 13 sound types, i.e. Car Engine, Construction, Crowd Applause, Fire, Helicopter, Office, Outdoor-Forest, Transportation-Moto, Wa- ter, Weather-Wind, Weather-Rain and Thunder, Household, and Auto Racing, achieve the accuracy range of 92.66%-99.88%. Five sound types, i.e. Crow Clamor, Outdoor-Road, Restaurant Stores, Transportation-Train, and Chicken Farm, are in the accuracy range of 76.12%-89.37%. The other two sound types, i.e. Aviation and Water (Ocean), have the accuracy in between 56.85%-66.24%. By average, the KNN classifier with spectrogram feature performs much better than the neural classifier.

The effectiveness of different features, i.e. spectrogram, MFCC, LPC, and MP, with a KNN classifier is also investigated in this experiment. Figure 4.28 shows the result of this experiment. Obviously, over all classification accuracy using KNN classifier is better than a neural network. The spectrogram feature indicates the best classification accuracy of more than 90%. This accuracy is higher than the results based on MFCC, LPC, and MP features in 12 classes, i.e. Car engine, Construction, Construction, Office, Restaurant Stores, Transportation-Train, Weather Wind, Weather-Rain and Thunder, Household, Water(Ocean), Chicken Farm and Auto Racing. But spectrogram feature results less accuracy than MFCC feature for the sound of Outdoor-Forest and MP for the sounds of Construction, Crow Clamor, Outdoor Sounds-Road, Transportation-Moto, Water, and Aviation.

Different numbers of nearest neighbours are tested by varying the values of K from 10 to 20 to achieve the maximum accuracy. Figure 4.29 summarizes the accuracy under different numbers of nearest neighbours. The maximum accuracy of94.43% occurs when K is equal to 15.

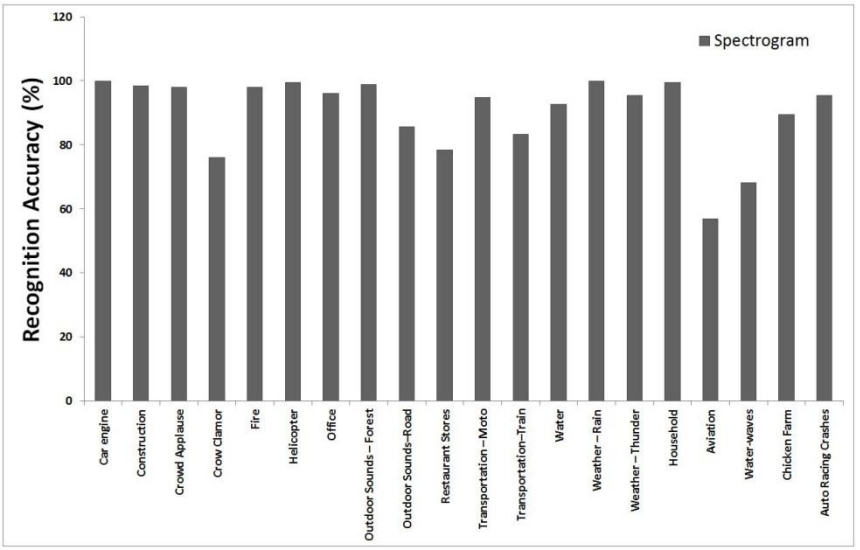

Figure 4.27 Classification accuracy obtained with spectrogram features using the KNN classifier.

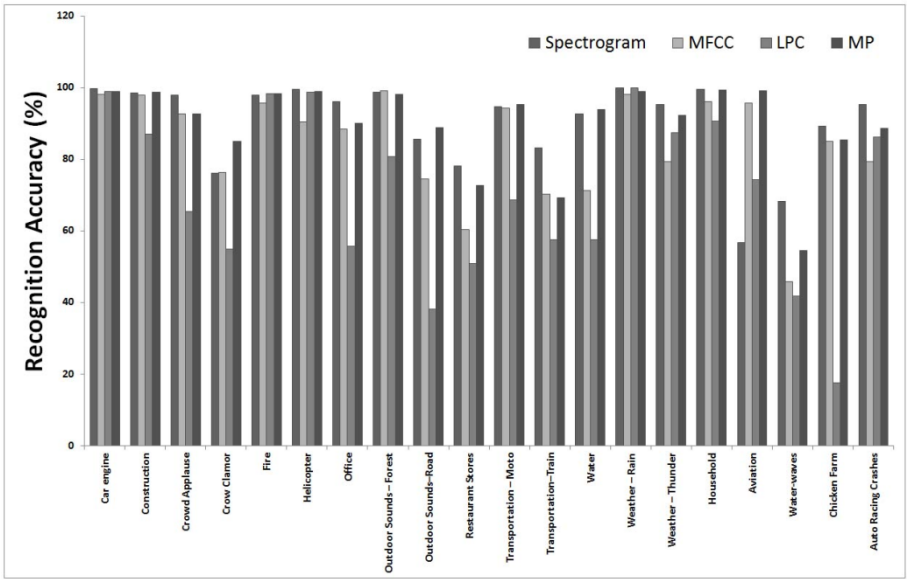

Figure 4.28: Average classification performance of KNN with k = 10 on spectrogram, MP, MFCC and LPC features.

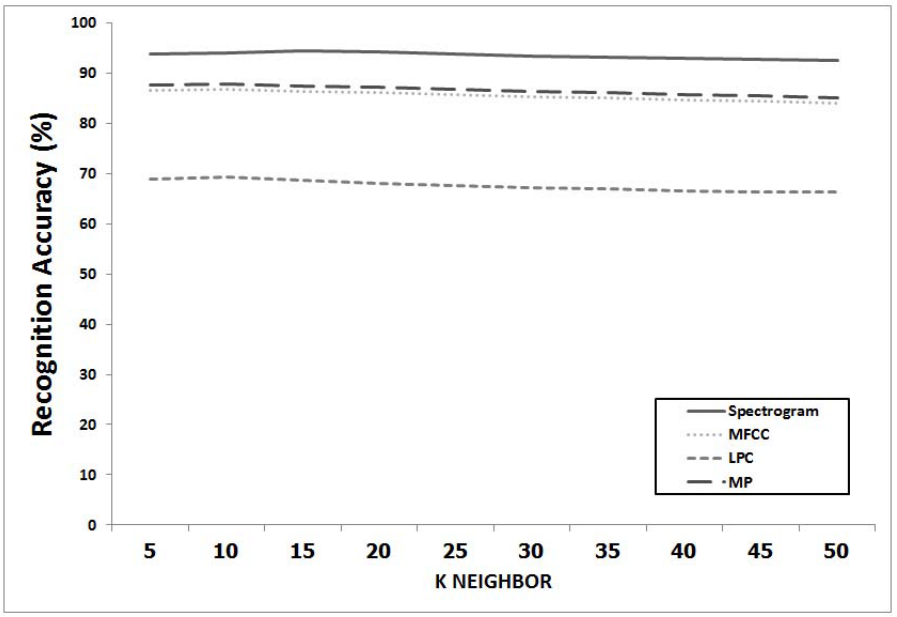

Figure 4.29: Average classification performance of KNN with different numbers of nearest neighbours, spectrogram, MP, MFCC, and LPC features.

4.2.3 Comparison of Neural Network and KNN Performances

To conclude which classifiers, namely neural network and KNN classifiers performs best on environmental sound recognition, the average accuracy of all sound types based each feature and each classifier is computed. Figure 4.30 summarizes the comparison. The first two vertical bars are the performances of neural network and KNN classifiers with spectrogram feature. The second vertical bars are of MPCC feature. The third vertical bars are of LPC feature and the last ones are of MP feature. Obviously, KNN classifier performs significantly better than neural network classifier in all features.

4.2.4 Effect of Different Window Sizes

A spectrogram can be obtained from different sizes of windowed segment. Note that the amount information of any sound wave represented in forms of spectrogram depends upon the window size. However, predicting an appropriate window size for each sound type is not simple. What should be a feasible window size that can be applied to every sound type to possibly achieve the maximum classification accuracy? To answer this problem, the following set of window sizes {64, 128,256, 512, 1024, 2048, 4096, and 8192} with neural and KNN classifiers are experimented. Based on spectrogram, MFCC, LPC, and MP with different window sizes, Figures 4.31 illustrates the average neural classification accuracy and Figure 4.32 illustrates the average KNN classification accuracy. The same configurations of neural classifier and KNN classifier discussed in the previous section are deployed in this experiment.

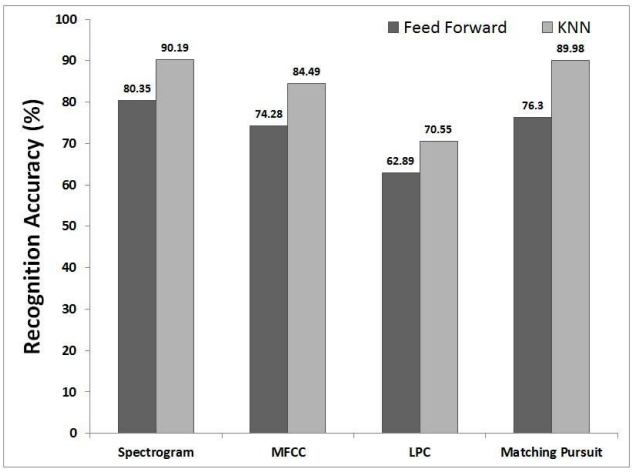

Figure 4.30: Average Classification accuracy obtained with KNN and Feed-forward Neural Network on a variety features.

It is interesting to note that, regardless of neural as well as KNN classifiers and window sizes, spectrogram feature provides the highest accuracy among the other features, MFCC, LPC, and MP. The maximum average accuracy occurs when the window size is equal to 8192. This is because a large window contains more classifiable feature information than a small window size. However, for MFCC, LPC, and MP features, the average accuracy with different window sizes is not conclusive. For example, with window size of 512, LPC performs better than MP by neural classifier but MP performs better than LPC by KNN classifier.

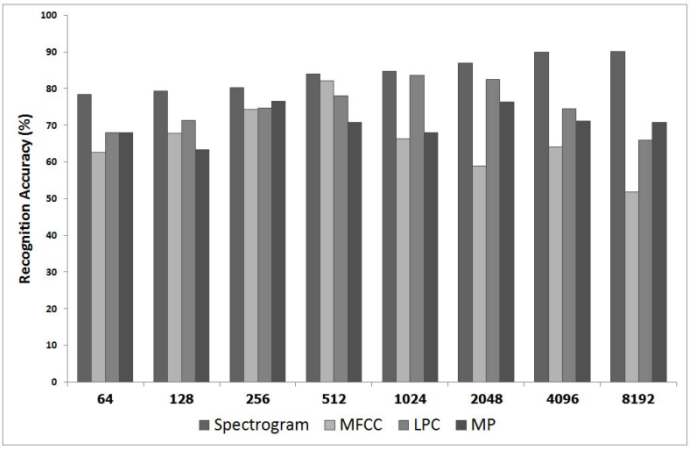

Figure 4.31: Average classification performance of a feed forward neural network with spectrogram, MP, MFCC, LPC features and different window sizes.

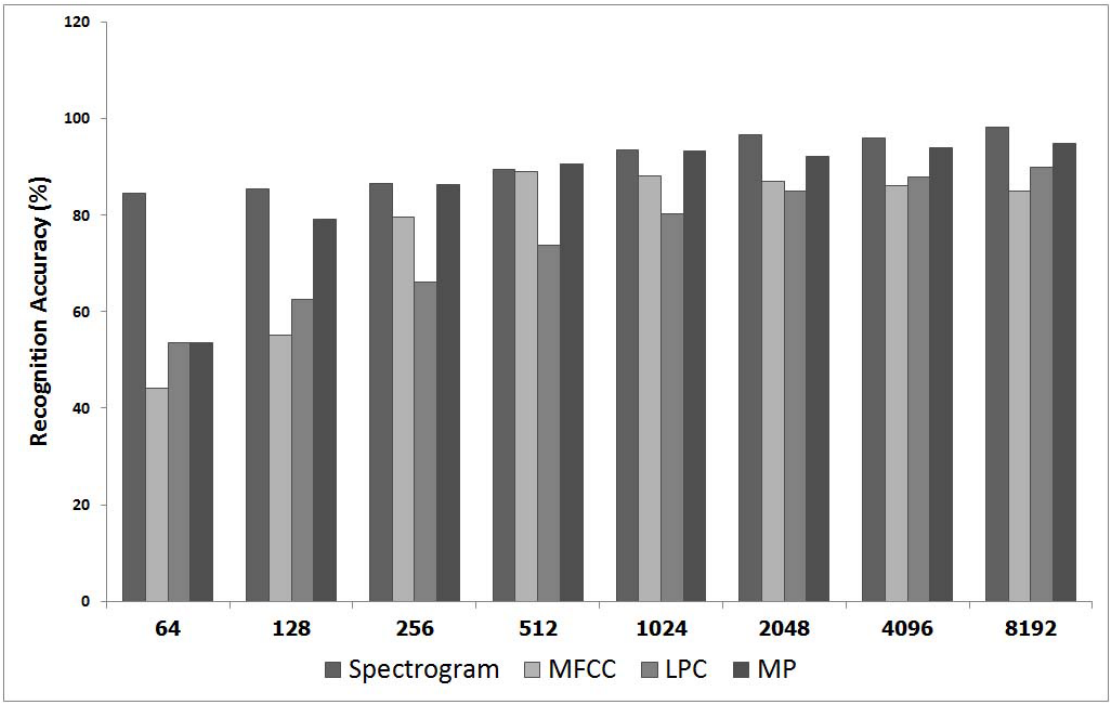

Figure 4.32: Average classification performance of KNN with spectrogram, MP, MFCC and LPC feature using different window sizes.

4.2.5 Effect of Different Sampling Rates

A digital audio signal can be collected from different sampling rates. Many different sampling rates are used in current digital signal processing applications. Telephone systems sample speech at 8 kHz, 11.025 kHz is used for AM-radio quality audio, and 44.1 kHz is the standard for CD quality digital music. A spectrogram feature computed from a digital audio signal by different sampling rates contains different amounts of information which obviously affect the accuracy. The interesting problem is which sampling rate is most suitable for classifying environmental sound types. Since it is very difficult to reach the theoretical conclusion on the best sampling rate for this problem, an empirical study on different sampling rates versus classification accuracy is conducted. The following sampling rates are tested against the classification accuracy, i.e. 5500 Hz, 6000Hz, 7333 Hz, 8000 Hz, 11025 Hz, 16000 Hz, 22050 Hz, 32000 Hz, and 44100 Hz.

In this experiment, the sampling rate of 44100 Hz with 128 KB/s bit-rates is originally applied to the audio signals. Then, the obtained signals are down sampled to 5500Hz, 6000Hz, 7333Hz, 8000 Hz, 11025 Hz, 16000 Hz, 22050 Hz and 32000 Hz with mono channel, respectively. In addition to different sampling rates, different window sizes in each sampling rate are also involved. The following window sizes are concerned for each sampling rate, i.e. 128, 256, 512, 1024, 2048, 4096 and 8192 with 25% overlap.

Figure 4.33, 4.34, 4.35, 4.36, 4.37,4.38, and 4.39 show the average classification rates by using feed-forward neural network with 30 hidden neural on spectrogram, MFCC, LPC and MP features. It can be seen that, in case of the feed forward neural network, spectrogram feature provides better performances than those from MFCC, LPC and MP features in all sampling rates. The performance of spectrogram feature does not change much when changing the sampling rate but MFCC, LPC and MP features significantly change a lot. In all sampling rates, it was found that, regardless of the feature types used, a large window size will provide better performance than a small window size.

Figure 4.40, 4.41, 4.42, 4.43, 4.44, 4.45, and 4.46 show the performances of spectrogram, MFCC, LPC, and MP features with a KNN for 10 nearest neighbours under different sampling rates. Similar to the previous experiment, the same conclusion on the performance of each feature type as well as the window sizes can be drawn. Remarkably, the performance of spectrogram feature, however, remains to be relatively high and even very stable for different sampling rates. This verifies that spectrogram is very robust to use for environmental sound classification.

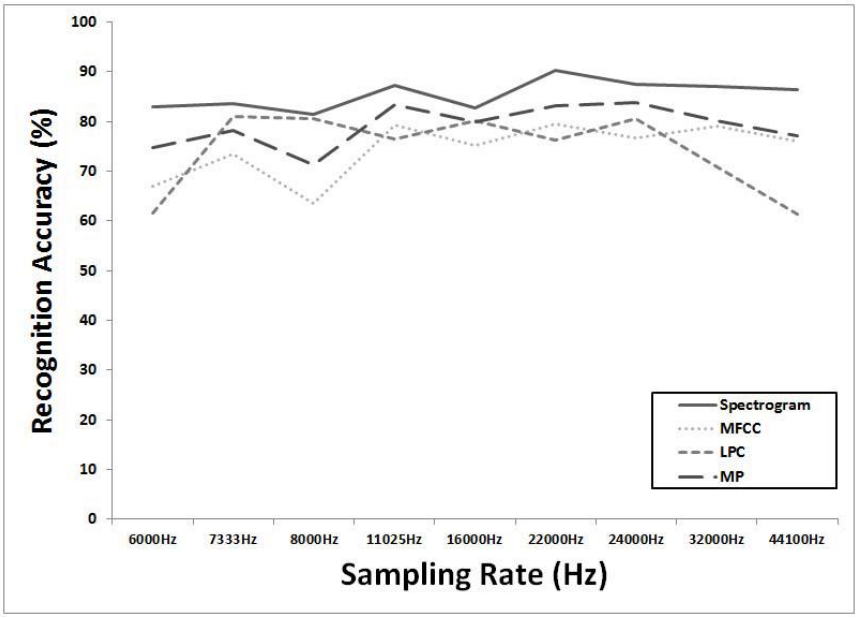

Figure 4.33: Average classification performance of feed forward neural network having 30 hidden neurons with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 8192.

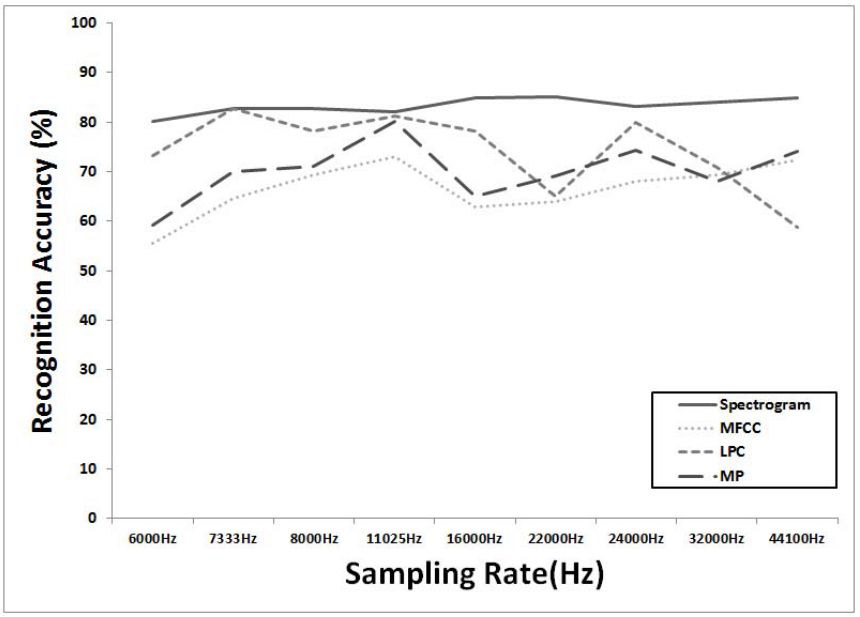

Figure 4.34: Average classification performance of feed forward neural network having 30 hidden neurons with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 4096.

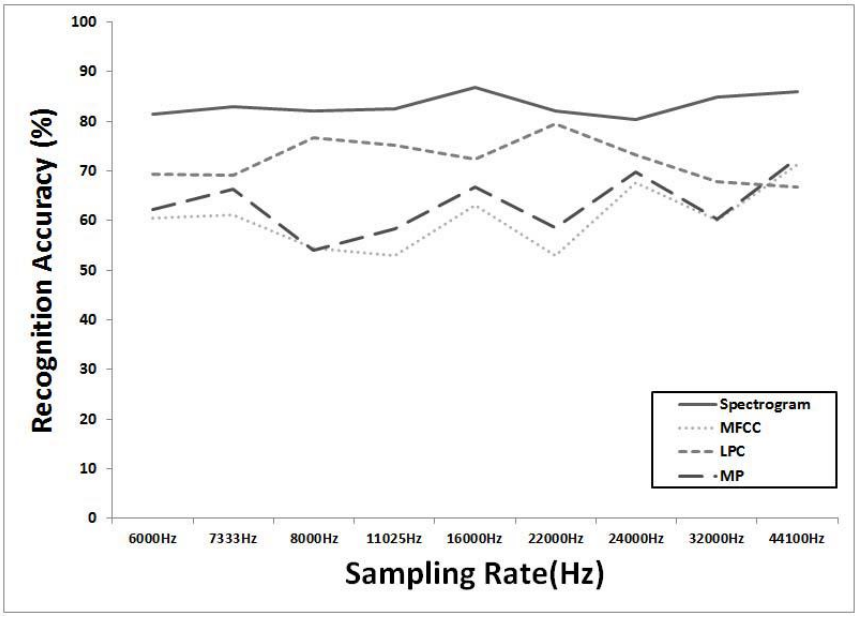

Figure 4.35: Average classification performance of feed forward neural network having 30 hidden neurons with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 2048.

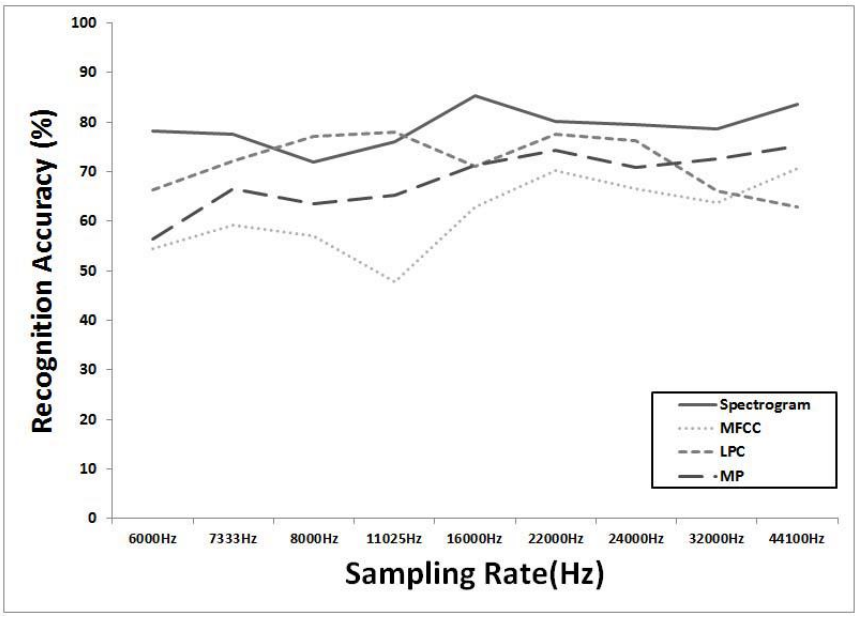

Figure 4.36: Average classification performance of feed forward neural network having 30 hidden neurons with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 1024.

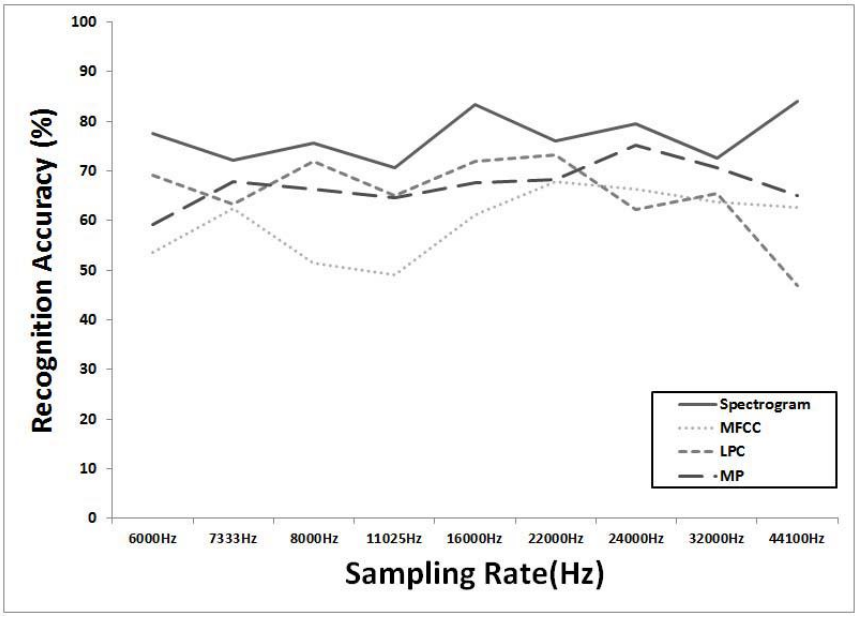

Figure 4.37: Average classification performance of feed forward neural network having 30 hidden neurons with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 512.

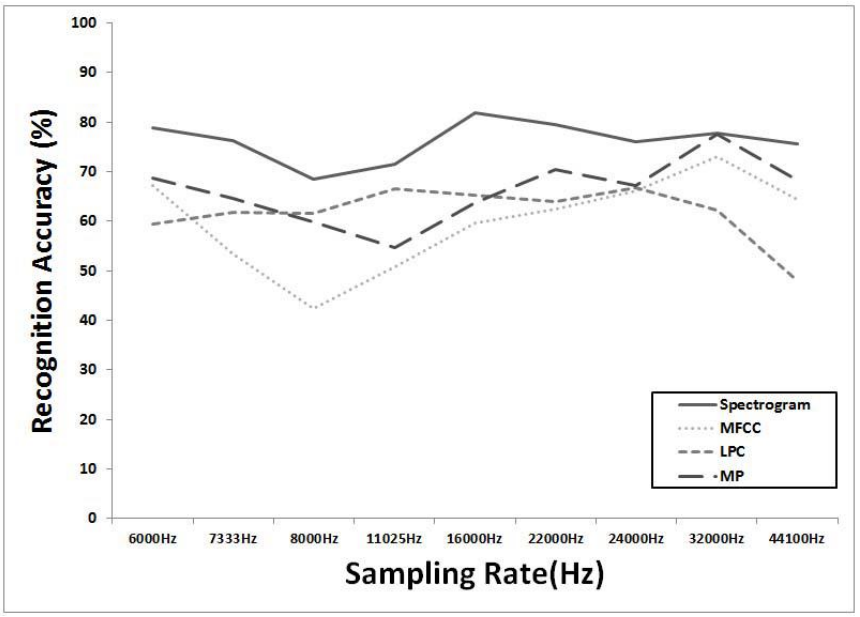

Figure 4.38: Average classification performance of feed forward neural network having 30 hidden neurons with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 256.

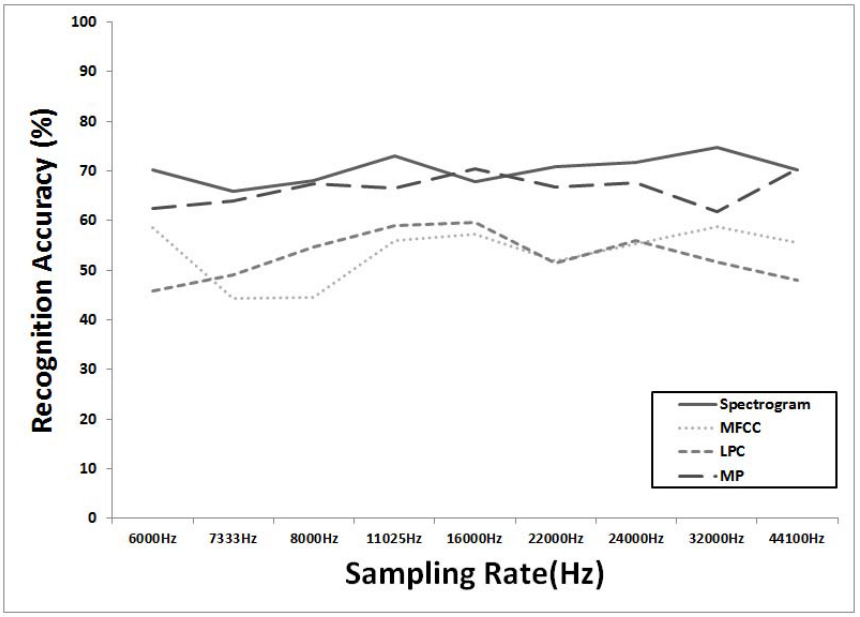

Figure 4.39: Average classification performance of feed forward neural network having 30 hidden neurons with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 128.

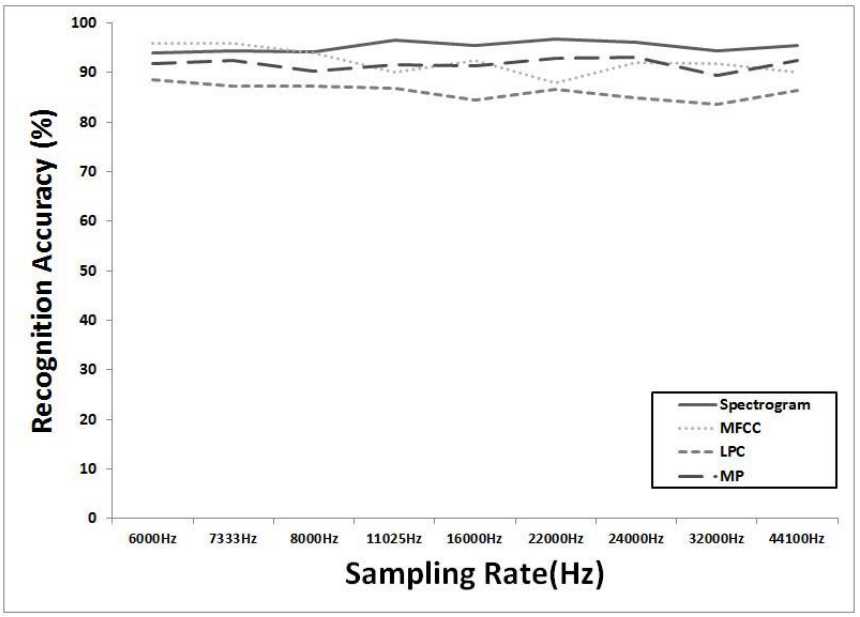

Figure 4.40: Average classification performance of K-nearest neighbour forK10 with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 8192.

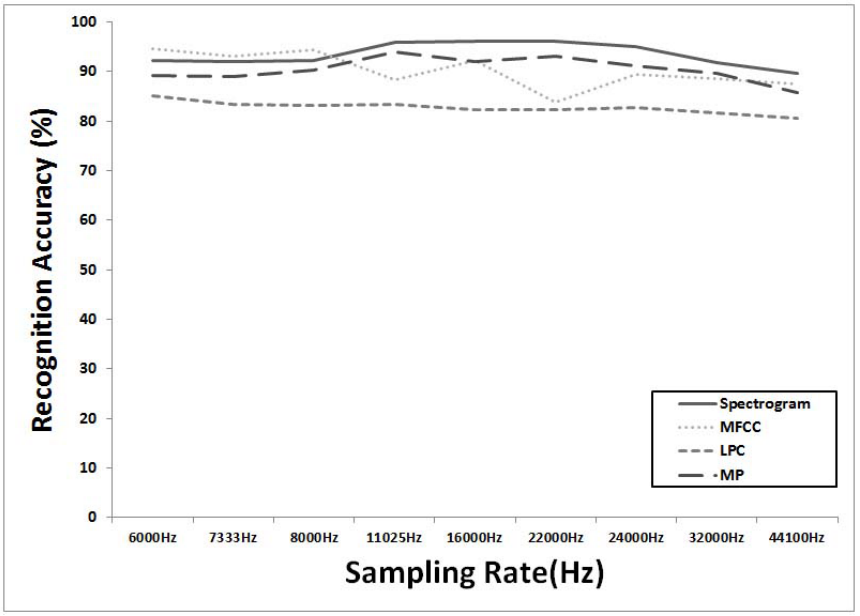

Figure 4.41: Average classification performance of K-nearest neighbour forK10 with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 4096.

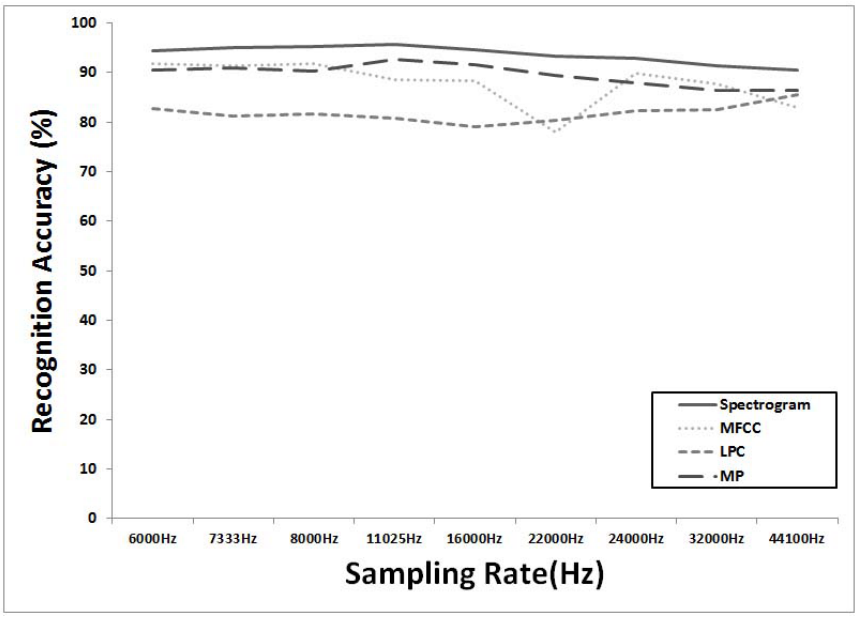

Figure 4.42: Average classification performance of K-nearest neighbour forK10 with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 2048.

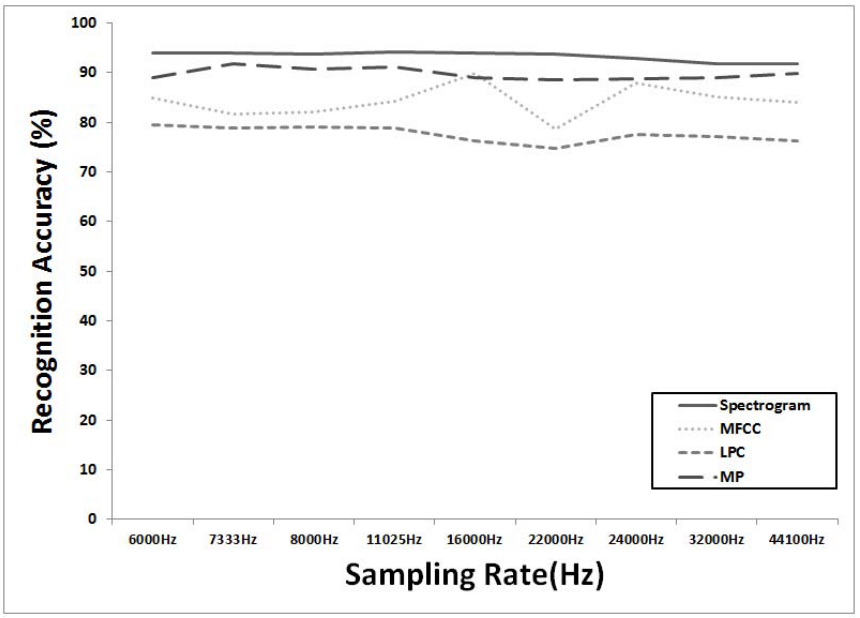

Figure 4.43: Average classification performance of K-nearest neighbour forK10 with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 1024.

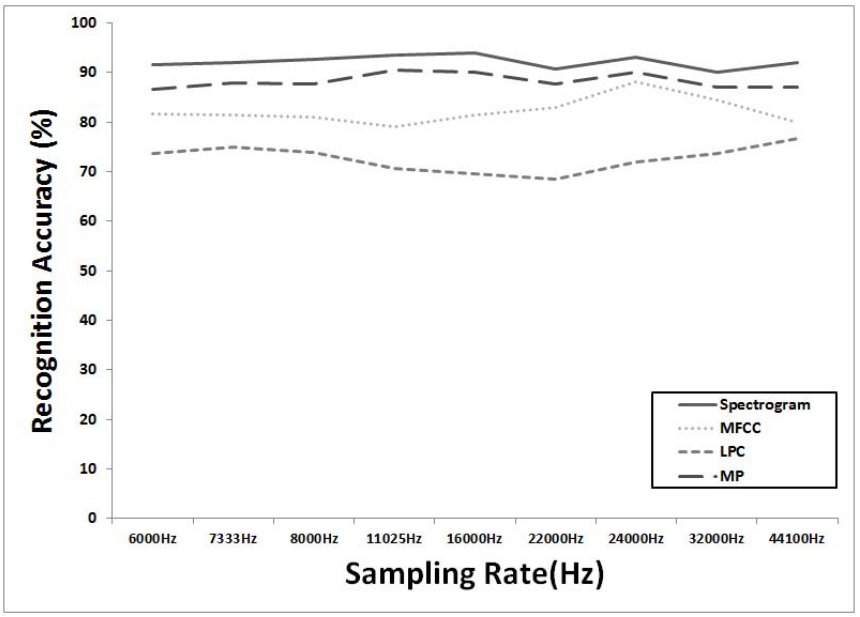

Figure 4.44: Average classification performance of K-nearest neighbour forK10 with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 512.

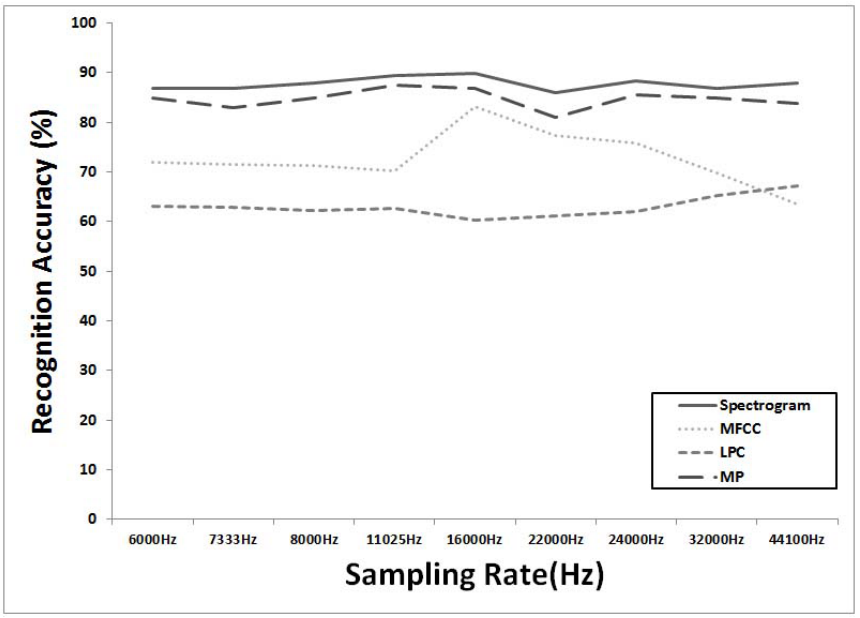

Figure 4.45: Average classification performance of K-nearest neighbour forK10 with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 256.

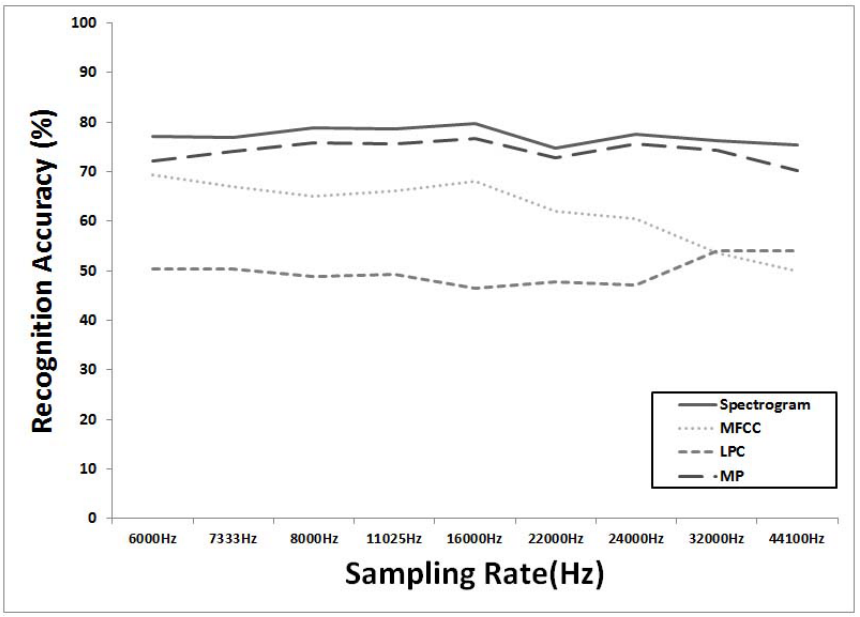

Figure 4.46: Average classification performance of K-nearest neighbour forK10 with spectrogram, MP, MFCC and LPC feature using different sampling rates on Window Sizes 128.

4.2.6 Classifying Short Duration Sounds

The work by Selina Chu [22] performed a listening test to study human recognition capability on these environmental sounds. The duration of the studied audio clips varied between 2, 4, and 6 seconds. The test consisted of 140 audio clips from 14 categories, with ten clips in each class. The study concluded that a longer duration increases the chance that a listener can correctly identify the sounds in each clip. However, this conclusion may not be true if an interested sound is recognized by a machine. To confirm this hypothesis, our technique is deployed to different duration sounds as follows.

Our algorithms are tested with the audio clips summarized in Table 4.4 for various time periods between 1 to 6 seconds. Some these periods are shorter than those of Selina Chu's experimental times [22]. Our experiments consist of 5-fold cross validation and the average accuracy is measured. In each fold, the data in each of the audio clip are partitioned into five sections. Four out of five sections are for training and the rest is for testing. The duration of each clip in minutes is given in Table 4.4. The audio data in the training set are converted into a set of spectrograms by using a window of size 8196 and 4096 with 25% overlapping sampled points for feature extraction. A feed- forward neuron network with 30 hidden neurons and 20 outputs is applied for classification. For KNN, the value of K is set to 10 for classification.

The audio data in test set are chopped into a set of short audio clips. A duration of each clip varied between 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5 and 6 seconds.

This research used each window in each audio clip to test the accuracy. The accuracy rate is computed by analyzing the maximum output value observed from cumulative frequency.

Fig 4.47 shows the results of the experiments. By using a feed-forward neural network with spectrogram for the duration of one second, the classification accuracy ranges from 85.66% to 90.57% for the window size of 8196 and 81.19% to 84.93% for the window size of 4096. Obviously, the accuracy increases when the duration is lengthened. The accuracy of each duration is summarized in Tables 4.5 and 4.6. Nine best classification scenes are Car engine, Crowd Applause, Crowd Clamor, Fire, Outdoor Sounds - Forest, Outdoor Sounds - Road, Water, Household, and Chicken Farm. All nine classes achieve more than 90% of classification rate.

Table 4.4 Duration of training and testing sets.

Audio Clips	Training (min)	Testing (min)
Car engine	11.6	3
Construction	9.68	2.5
Crowd Applause	11	3
Crow Cheering	12	4
Fire	10	3.7
Helicopter	11.36	3.08
Office	12.16	3.24
Forest	12.64	3.6
Road, Restaurant Stores	12.02	3.8
Transportation - Motorcycle	12.72	3.58
Transportation-Train	11.04	3.2
Water, Weather - Wind	12.56	3.54
Weather - Rain and Thunder	10.24	3.56
Household	13.28	3.32
Airplane	9.76	2.44
Water(Ocean)	16	4.8
Chicken Farm	17.84	4.45
and Auto Racing	18.56	4.64

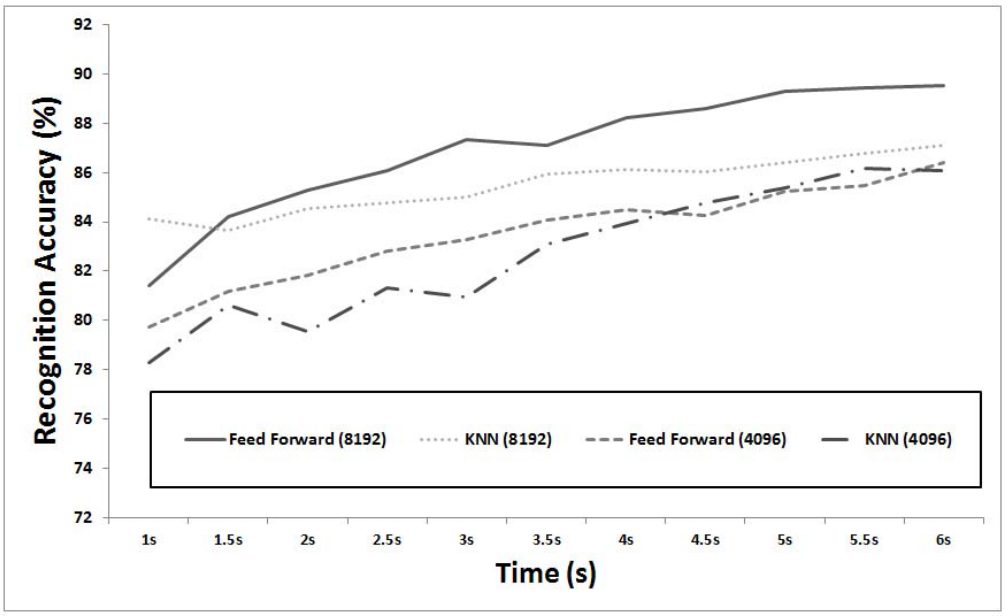

Figure 4.47: Classification accuracy obtained from different time duration.

By KNN using with spectrogram windows of sizes 8192 and 4096, the classification accuracy ranges from 84.11% to 87.09% for the window of size 8196, and 78.29% to 86.10% for the window size of 4096.

Tables 4.7 and 4.8 show the overall classification rates by using KNN with K 10 on spectrogram with different time duration. For window size of 8192, the following eight classes, i.e. Car engine, Crowd Applause, Fire, Outdoor Sounds - Forest, Outdoor Sounds - Road, Transportation - Train, Household, and Auto Racing, achieve the accuracy rate more than 90%. From these experiments, when the duration of audio clip is short, the results from the feed-forward neural network outperform the results from KNN. In addition, the accuracy rate depends upon the length of the audio clips.

Table 4.5 The average accuracy of all classes from feed-forward 30 hidden neurons with spectrogram using window size of 4096 in different time duration.

Window Sizes 4096	1s	1.5s	2s	2.5s	3s	3.5s	4s	4.5s	5s	5.5s	6s
Class1	94.63	96.09	97.56	99.39	100.00	100.00	100.00	99.63	100.00	100.00	100.00
Class2	68.60	68.97	70.54	70.68	68.92	70.09	68.72	71.57	72.48	76.23	74.33
Class3	88.58	87.87	88.27	89.79	93.23	93.09	92.11	89.80	90.00	91.86	91.06
Class4	95.45	95.99	95.74	97.92	96.58	98.00	98.67	97.34	96.67	96.65	98.88
Class5	90.89	90.13	92.05	90.63	92.91	90.76	92.35	92.26	91.43	93.95	94.59
Class6	75.41	79.22	79.38	82.25	82.95	82.98	83.00	81.22	84.13	85.21	85.10
Class7	68.91	70.71	72.55	73.72	73.83	72.43	74.91	76.40	76.65	75.58	77.53
Class8	96.75	96.27	98.45	100.00	98.19	98.85	100.00	99.68	99.55	99.44	100.00
Class9	77.00	85.62	81.09	81.47	86.00	86.08	89.15	87.83	92.65	88.00	91.80
Class10	76.00	77.46	79.04	80.26	78.23	79.65	79.75	80.08	80.79	80.69	82.36
Class11	78.17	82.75	81.49	84.46	82.90	84.50	86.78	85.33	89.01	90.57	92.77
Class12	72.00	75.62	75.09	76.47	77.00	79.08	82.15	80.83	82.65	82.00	84.80
Class13	97.98	98.65	100.66	99.43	100.00	100.00	99.86	100.00	100.00	100.00	100.00
Class14	70.20	71.06	71.03	70.70	70.62	72.43	71.53	70.68	74.09	70.81	72.75
Class15	67.25	68.27	70.04	69.88	69.57	73.19	70.64	72.71	73.66	75.22	76.79
Class16	96.57	98.03	97.19	98.75	99.44	98.10	99.39	99.21	100.00	100.00	100.00
Class17	60.16	61.27	61.39	61.45	62.04	63.04	64.60	63.61	64.73	64.77	65.22
Class18	87.05	86.33	88.17	89.50	89.49	89.10	90.29	89.12	88.96	90.41	90.61
Class19	93.71	96.42	98.06	98.33	99.89	99.81	99.67	100.00	100.00	100.00	100.00
Class20	68.54	68.67	68.80	70.54	71.45	74.50	73.38	75.68	75.56	74.75	76.68
Average	81.19	82.77	83.33	84.28	84.66	85.28	85.85	85.65	86.65	86.81	87.76

Table 4.6 The average accuracy of all classes from feed-forward 30 hidden neurons with spectrogram using window size of 8192 in different time duration.

Window Sizes 8192	1s	1.5s	2s	2.5s	3s	3.5s	4s	4.5s	5s	5.5s	6s
Class1	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class2	81.10	80.00	81.11	83.33	82.61	83.87	82.14	86.27	85.11	86.09	85.37
Class3	98.90	96.00	98.44	100.00	97.96	97.73	100.00	100.00	100.00	100.00	100.00
Class4	93.33	93.33	92.19	94.55	93.88	93.18	97.50	100.00	96.97	96.77	96.43
Class5	95.61	95.74	95.06	95.71	96.77	98.21	98.00	95.65	97.62	94.87	97.22
Class6	82.00	81.93	84.51	83.87	83.64	85.71	86.36	90.00	86.49	88.24	87.50
Class7	74.64	74.78	75.51	77.65	76.32	79.41	78.69	78.57	78.85	79.17	81.82
Class8	96.43	97.10	100.00	98.04	97.78	100.00	100.00	96.97	100.00	100.00	100.00
Class9	99.01	100.00	98.59	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class10	77.50	77.27	77.19	81.63	79.07	79.49	80.00	81.25	82.76	81.48	84.00
Class11	82.61	80.52	83.08	87.72	86.00	84.44	85.37	91.89	97.06	97.55	97.10
Class12	71.01	73.84	75.51	72.87	72.81	74.51	74.19	75.29	75.64	76.39	73.13
Class13	98.86	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class14	71.01	73.84	75.51	72.87	72.81	74.51	74.19	75.29	75.64	76.39	73.13
Class15	79.76	83.45	86.55	85.58	86.96	85.54	85.33	86.76	88.89	86.21	92.59
Class16	98.39	98.04	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class17	73.98	77.45	75.86	76.32	77.61	78.33	80.00	76.00	73.91	76.19	76.92
Class18	83.78	83.70	82.28	82.35	85.25	87.04	83.67	86.67	90.24	90.84	90.71
Class19	89.47	93.65	94.34	93.48	92.68	91.89	90.91	93.33	96.43	96.00	96.67
Class20	65.71	68.97	68.92	69.23	73.68	74.51	73.91	76.19	79.49	75.00	78.79
Average	85.66	86.48	87.23	87.76	87.79	88.42	88.51	89.51	90.25	90.06	90.57

Table 4.7 The average accuracy of all classes from 10-nearest neighbour network with spectrogram using window size of 4096 in Different time duration.

Window Sizes 4096	1s	1.5s	2s	2.5s	3s	3.5s	4s	4.5s	5s	5.5s	6s
Class1	78.00	82.92	69.29	77.38	71.48	79.58	82.68	74.78	84.88	88.98	84.08
Class2	71.27	76.39	68.00	79.19	78.80	76.13	79.03	74.93	85.83	72.73	78.63
Class3	87.25	92.15	88.04	94.55	89.59	91.94	99.72	92.49	101.27	96.04	99.82
Class4	73.53	74.49	80.73	75.19	73.93	77.82	75.92	77.02	75.12	84.22	71.32
Class5	71.97	79.70	83.13	81.15	78.71	76.92	76.21	85.50	81.80	82.09	77.38
Class6	95.21	96.04	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class7	70.86	70.65	67.09	68.11	65.37	72.81	74.95	73.09	73.24	79.38	79.52
Class8	92.11	87.39	85.59	92.41	88.15	93.61	97.89	101.17	99.45	100.00	100.00
Class9	86.35	85.48	85.72	83.79	92.56	95.52	94.94	96.36	90.78	91.20	88.61
Class10	53.84	61.07	65.00	66.16	66.84	68.51	61.93	73.35	64.77	70.19	77.61
Class11	75.00	77.92	74.29	75.38	76.48	78.58	81.68	79.78	81.88	83.98	84.08
Class12	72.27	73.39	72.00	74.19	73.80	74.13	74.03	75.93	80.83	77.73	82.63
Class13	91.25	93.15	91.04	93.55	92.59	94.94	97.72	96.49	98.27	98.04	98.82
Class14	75.53	78.49	76.73	76.19	76.93	78.82	77.92	80.02	80.12	79.22	76.32
Class15	76.97	77.70	78.13	78.15	78.71	80.92	77.21	80.50	80.80	82.09	78.38
Class16	98.21	98.04	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class17	67.86	67.65	68.09	70.11	70.37	70.81	72.95	74.09	74.24	75.38	79.52
Class18	89.11	92.39	90.59	92.41	93.15	95.61	97.89	98.17	96.45	99.74	100.00
Class19	82.35	85.48	87.72	86.79	88.56	91.52	89.94	91.36	87.78	89.20	88.61
Class20	56.84	62.07	60.00	62.16	62.84	63.51	65.93	70.35	69.77	69.19	73.61
Average	78.00	82.92	69.29	77.38	71.48	79.58	82.68	74.78	84.88	88.98	84.08

Table 4.8 The average accuracy of all classes from 10-nearest neighbour network with spectrogram using window size of 8192 in Different time duration.

Window Sizes 4096	1s	1.5s	2s	2.5s	3s	3.5s	4s	4.5s	5s	5.5s	6s
Class1	99.02	95.55	99.00	99.01	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class2	74.21	73.23	77.14	71.11	75.64	75.36	79.03	75.00	78.43	78.72	75.00
Class3	96.54	92.81	97.37	96.70	96.00	96.88	98.21	93.88	97.73	97.50	97.22
Class4	85.59	78.29	87.61	81.11	85.33	84.38	85.45	83.67	86.36	85.00	86.11
Class5	96.88	95.29	95.80	95.61	94.68	95.06	95.71	95.16	92.86	94.00	95.65
Class6	73.23	70.41	72.22	71.00	73.49	71.83	72.58	69.09	73.47	72.73	75.00
Class7	68.53	71.68	70.29	72.17	71.43	76.47	73.68	76.47	73.77	75.00	75.00
Class8	88.65	90.48	90.48	89.86	89.83	92.16	91.11	90.24	91.89	93.94	93.55
Class9	92.94	94.49	93.07	95.24	97.18	93.55	96.36	95.92	95.56	95.12	97.30
Class10	58.52	70.30	60.00	68.18	61.40	65.31	65.12	69.23	68.57	68.75	65.52
Class11	68.59	74.14	72.83	72.73	69.23	71.93	74.00	73.33	73.17	72.97	76.47
Class12	92.94	94.49	93.07	93.63	93.69	93.76	93.82	93.88	93.95	94.01	94.08
Class13	88.59	90.09	88.64	87.67	88.71	87.04	87.50	88.37	87.18	88.57	87.50
Class14	78.39	83.08	80.68	83.14	78.23	81.40	78.95	83.33	81.72	80.00	79.49
Class15	82.38	79.17	82.73	80.67	83.65	84.78	83.13	84.00	83.82	82.54	84.48
Class16	97.44	96.77	98.04	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Class17	67.53	67.48	69.61	71.26	72.30	73.63	74.97	76.30	77.63	78.96	80.30
Class18	88.49	82.88	86.96	83.54	85.12	85.45	85.78	86.11	86.45	86.78	87.11
Class19	86.32	82.89	83.02	88.89	87.24	88.02	88.81	89.59	90.38	91.16	91.95
Class20	97.42	89.81	92.11	93.92	96.77	101.75	97.93	97.11	95.29	100.00	100.00
Average	99.02	95.55	99.00	99.01	100.00	100.00	100.00	100.00	100.00	100.00	100.00

CHAPTER V

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this dissertation propose an algorithm for Thai singing voice recognition in monaural poly- phonic music based on the time-frequency domain feature technique of power spectrogram with neural classifier and classification algorithms. The advantage of spectrogram is its ability to magnify and represent the relevant information of an audio signal captured under a defined window segment. This approach is simpler than the existing methods which used MFCC and LPC as features. Our approach achieved higher accuracy than the other techniques. However, the performance depends on several factors such as window size. The experiment showed that feed-forward network performed better than accuracy rate more K-nearest neighbor network than 94%. . Especially, this algorithm can recognize Cross-Language Music Data.

We also presented the result of time-frequency domain feature technique for unstructured environmental sound classification by using spectrogram pattern. All possible relevant factors such as sampling rates, window sizes, and different features are thoroughly studied to conclude which factors actually define the acceptable classification performance. The experimental results show a promising performance in classifying 20 different audio environmental. Both KNN and feed-forward neural network can effectively classify unstructured environmental sound. In particular, feed-forward neural network gives the best result in this experiment. A longer duration was increase classification accuracy within each sound clip.

However, the use of spectrogram is also limited. When converting from audio signal to a spectrogram. A dimension of Spectrogram is higher than other feature. Therefore, Spectrogram was using more memory than the other types feature vector in same windows size. As compared to using the MFCC feature performance close to the Spectrogram. When used a same size of Windows to create spectrogram and MFCC feature. This research apply dimension reduction technique for reduce spectrogram feature dimension. This research found, by using feed forward neural network (ANN) with spectrogram feature with dimension reduction technique. The recognition performance is greatly reduced. However, when compared with K-nearest neighbour (kNN) and spectrogram feature with dimension reduction technique. Recognition performance was not reduced as much as the using feed forward neural network (ANN).

REFERENCES

[1] Orio, N. Music Retrieval: A Tutorial and Review. Foundations and Trends in Information Retrieval 1 , 1 (2006): 1-90.

[2] Gerhard, D., Computationally Measurable Differences between Speech and Song. Ph.D. dissertation Simon Fraser University, 2003.

[3] Loscos, A., Cano, P., and Bonada, J., Low-Delay Singing Voice Alignment to Text. International Computer Music Conference.

[4] Cullity, B. D. Music information retrieval, vol. 35. Information Today Books, 2003.

[5] Gerhard, D. B., Computationally measurable differences between speech and song. Ph.D. dissertation, 2003. AAINQ81587.

[6] Sasou, A., Goto, M., Hayamizu, S., and Tanaka, K., An Auto-Regressive, Non-Stationary Excited Signal Parameter Estimation Method and an Evaluation of a Singing-Voice Recognition. Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on 1 (18-23, 2005): 237 - 240.

[7] Yaguchi, Y. and Oka, R., Song Wave Retrieval Based on Frame-Wise Phoneme Recognition. Information Retrieval Technology (Lee, G., Yamada, A., Meng, H., and Myaeng, S., eds.) 3689 (2005): 503-509.

[8] Gruhne, M., Schmidt, K., and Dittmar, C., Phoneme recognition in pop-pular music. 8th International Conference on Music Information Retrieval, Vienna, Austria, Sep 23-27 2007.

[9] Makeyev, O., Sazonov, E., Schuckers, S., Melanson, E., and Neuman, M., Limited receptive area neural classifier for recognition of swallowing sounds using short-time Fourier transform. Neural Networks, 2007, IJCNN 2007. International Joint Conference on (aug. 2007): 1601 -1606.

[10] Lin, C.-C., Chen, S.-H., Truong, T.-K., and Chang, Y. Audio Classification and Categorization Based on Wavelets and Support Vector Machine. Speech and Audio Processing, IEEE Transactions on 13 (sept. 2005): 644 - 651.

[11] Esmaili, S., Krishnan, S., and Raahemifar, K., Content based audio classification and retrieval using joint time-frequency analysis. Acoustics, Speech, and Signal Processing, 2004, Proceedings (ICASSP '04). IEEE International Conference on 5 (may 2004): V - 665-8 vol.5.

[12] Wang, J.-C., Lee, H.-P., Wang, J.-F., and Lin, C.-B. Robust Environmental Sound Recognition for Home Automation. Automation Science and Engineering, IEEE Transactions on 5 (jan. 2008): 25 -31.

[13] Yoshii, K., Goto, M., and Okuno, H. G. Drum Sound Recognition for Polyphonic Audio Signals by Adaptation and Matching of Spectrogram Templates With Harmonic Structure Suppression. Audio, Speech, and Language Processing, IEEE Transactions on 15 (jan. 2007): 333 -345.

[14] Toyoda, Y., Huang, J., Ding, S., and Liu, Y. Environmental sound recognition by the instantaneous spectrum combined with the time pattern of power. (2004): 169-172.

[15] Makeyev, O., Sazonov, E., Schuckers, S., Lopez-Meyer, P., Melanson, E., and Neuman, M., Limited receptive area neural classifier for recognition of swallowing sounds using continuous wavelet transform. Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE (aug. 2007): 3128 -3131.

[16] Ajmera, J., McCowan, I., and Bourlard, H. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Speech Commun. 40 (May 2003): 351- 363.

[17] Toyoda, Y., Huang, J., Ding, S., and Liu, Y., Environmental sound recognition by multilayered neural networks. Computer and Information Technology, 2004. CIT '04. The Fourth International Conference on (sept. 2004): 123 - 127.

[18] Shenoy, A. Singing voice detection for karaoke application. Proceedings of SPIE 5960 (2005): 752-762.

[19] Nwe, T. L., Shenoy, A., and Wang, Y., Singing voice detection in popular music. Proceedings of the 12th annual ACM international conference on Multimedia, MULTIMEDIA '04, New York, NY, USA, ACM, (2004): 324--327.

[20] Tsai, W.-H., Wang, H.-M., Rodgers, D., Cheng, S.-S., and Yu, H.-M., Blind clustering of popular music recordings based on singer voice characteristics. ISMIR.

[21] Berenzweig, A. and Ellis, D., Locating singing voice segments within music signals. Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the (2001): 119-122.

[22] Chou, W. and Gu, L., Robust singing detection in speech/music discriminator design. Proceedings of the Acoustics, Speech, and Signal Processing, 200. on IEEE International Conference - Volume 02, Washington, DC, USA, IEEE Computer Society, (2001): 865-868.

[23] Berenzweig, D. P. W. L. S., Using Voice Segments to Improve Artist Classification of Music. Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio, 6 2002.

[24] Maddage, N. C., Xu, C., and Wang, Y., An SVM-based classification approach to musical audio. ISMIR.

[25] Maddage, N., Wan, K., Xu, C., and Wang, Y., Singing voice detection using twice-iterated composite Fourier transform. Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on 2 (june 2004): 1347 -1350 Vol.2.

[26] Rocamora, M. and Herrera, P., Comparing audio descriptors for singing voice detection in music audio files. Brazilian Symposium on Computer Music, 11th. San Pablo, Brazil, sep 2007.

[27] Tzanetakis, G., Song-specific bootstrapping of singing voice structure. Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on 3 (june 2004): 2027 - 2030 Vol.3.

[28] Kim, Y. E., Singer identification in popular music recordings using voice coding features. In Proceedings of the 3rd International Conference on Music Information Retrieval, (2002): 164-169.

[29] Sasou, A., Goto, M., Hayamizu, S., and Tanaka, K., An Auto-Regressive, Non-Stationary Excited Signal Parameter Estimation Method and an Evaluation of a Singing-Voice Recognition. Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on 1 (18-23, 2005): 237 - 240.

[30] Suzuki, M., Hosoya, T., Ito, A., and Makino, S. Music information retrieval from a singing voice using lyrics and melody information. EURASIP J. Appl. Signal Process. 2007 (January 2007): 151--151.

[31] Wong, C., Szeto, W., and Wong, K. Automatic lyrics alignment for Cantonese popular music.. Multimedia Systems 12 (Mar. 2007): 307--323.

[32] Kan, M.-Y., Wang, Y., Iskandar, D., Nwe, T. L., and Shenoy, A. LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals. IEEE Transactions on Audio, Speech & Language Processing 16 , 2 (2008): 338-349.

[33] M. Gruhne, K. S. and Dittmar, C., Phoneme recognition in pop-pular music. 8th International Conference on Music Information Retrieval, Vienna, Austria., Sep 23-27 2007.

[34] Fujihara, H., Goto, M., Ogata, J., Komatani, K., Ogata, T., and Okuno, H. G., Automatic Synchronization between Lyrics and Music CD Recordings Based on Viterbi Alignment of Segregated Vocal Signals. Proceedings of the Eighth IEEE International Symposium on Multimedia, ISM '06, Washington, DC, USA, IEEE Computer Society, (2006): 257- 264.

[35] Zwan, P., Szczuko, P., Kostek, B., and Czyzewski, A., Transactions on Rough Sets IX. ch. Automatic Singing Voice Recognition Employing Neural Networks and Rough Sets, pp. 455-473, Berlin, Heidelberg : Springer-Verlag, 2008.

[36] Mesaros, A. and Virtanen, T., Recognition of phonemes and words in singing. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (march 2010): 2146 -2149.

[37] Makeyev, O., Sazonov, E., Schuckers, S., Melanson, E., and Neuman, M., Limited receptive area neural classifier for recognition of swallowing sounds using short-time Fourier transform. Proc. International Joint Conference on Neural Networks IJCNN 2007, Orlando, USA.

[38] Lin, C.-C., S.-H.Chen, Truong, T.-K., and Chang, Y. Audio Classification and Categorization Based on Wavelets and Support Vector Machine. 13 (2005): 644-651.

[39] Esmaili, S., Krishnan, S., and Raahemifar, K., Content Based Audio Classification and Retrieval Using Joint Time-Frequency Analysis. International Conference on Acoustics, Speech and Signal Processing 2004, May 2004.

[40] Wang, J.-C., Lee, H.-P., Wang, J.-F., and Lin, C.-B. Robust environmental sound recognition for home automation. 5 (Jan. 2008): 25-31.

[41] Yoshii, K., Goto, M., and Okuno, H. Drum Sound Recognition for Polyphonic Audio Signals by Adaptation and Matching of Spectrogram Templates With Harmonic Structure Suppression. 15 (Jan. 2007): 333 - 345.

[42] Marzano, F., Scaranari, D., and Vulpiani, G. Supervised Fuzzy-Logic Classification of Hydrometeors Using C-Band Weather Radars. 45 , 11 (2007): 3784 - 3799.

[43] Malkin, R. and Waibel, A., Classifying user environment for mobile applications using linear autoencoding of ambient audio. Proc. IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), USA, Mar.18-23 2005.

[44] Wang, J., Wang, J., He, K., and Hsu, C., Environmental sound classification using hybrid SVM/KNN Classifier and MPEG-7 audio lowlevel descriptor. Proc. IEEE International Joint Conferenceon Neural Networks, 26-29 2006.

[45] Kraetzer, C., Oermann, A., Dittmann, J., and Lang, A., Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification. Proc. ACM Multi Media and Security, 20-21 2007.

[46] Ntalampiras, S., Potamitis, I., and Fakotakis, N., Automatic recognition of urban environmental sounds events. Proc. International Association for Pattern Recognition Workshop on Cognitive Information Processing, (2008): 110-113.

[47] Toyoda, Y., Huang, J., Ding, S., and Liu, Y., Environmental Sound Recognition by Multilayered Neural Networks. Proceedings of the The Fourth International Conference on Computer and Information Technology, CIT '04, Washington, DC, USA, IEEE Computer Society, (2004): 123--127.

[48] Eronen, A. J., Peltonen, V. T., Tuomi, J. T., Klapuri, A. P., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J. Audio-based context recognition. IEEE Transactions on Audio, Speech and Language Processing 14 , 1 (2006): 321--329.

[49] Wang, J. C., Lee, H. P., Wang, J. F., and Lin, C. B. Robust Environmental Sound Recognition for Home Automation. Automation Science and Engineering, IEEE Transactions on Robotics and Automation, IEEE Transactions 5 , 1 (2008): 25--31.

[50] Byeong-jun, H. and Eenjun, H., Environmental sound classification based on feature collaboration. Proc. IEEE International Conference on Multimedia and Expo, New York, USA, (2009): 542-545.

[51] Lozano, H., Hernáez, I., Picón, A., Camarena, J., and Navas, E., Audio classification techniques in home environments for elderly/dependant people. Proceedings of the 12th international conference on Computers helping people with special needs: Part I, ICCHP'10, Berlin, Heidelberg, Springer-Verlag, (2010): 320--323.

[52] Chu, S., Narayanan, S., and Kuo, C.-C. J., Environmental sound recognition using mp-based features. Proc. 2008 IEEE international conference on acoustics, speech, and signal processing, Las Vegas, USA, (2008): 1-4.

[53] Chu, S., Narayanan, S., and Kuo, C.-C. Environmental sound recognition with time-frequency audio features. Trans. Audio, Speech and Lang. Proc. 17 (August 2009): 1142--1158.

[54] Dennis, J., Dat, T. H., and Li, H. Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions. IEEE Signal Process. Lett. 18 , 2 (2011): 130-133.

[55] Bay, S. D. Nearest neighbor classification from multiple feature subsets. Intelligent Data Analysis 3 (1999): 191--209.

[56] Mitchell, T. M. Machine Learning. New York : McGraw-Hill, 1997.
[57] Verhelst, W. and Roelands, M., An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. Proceedings of the 1993 IEEE international conference on Acoustics, speech, and signal processing: speech processing - Volume II, ICASSP'93, Washington, DC, USA, IEEE Computer Society, (1993): 554-557.
[58] Gerhard, D., Computationally Measurable Differences Between Speech and Song. Ph.D. dissertation Simon Fraser University, 2003.

Biography

Name: Ms. Peerapol Khunarsa.

Date of Birth: 27th July, 1975.

Educations:

• M.Sc. Program in Information Technology, Faculty of Information Technology, King Mongkut's Institute of Technology North Bangkok, Bangkok, Thailand .

• B.Sc. Program in Computer Science, Faculty of Science, Rajabhat Uttaradit University, Uttaradit, Thailand.

Publication papers:

• Khunarsa, P., Lursinsap, C., and Raicharoen, T., Impulsive Environment Sound Detection by Neural Classification of Spectrogram and Mel-Frequency Coefficient Images. in Advances in Neural Network Research and Applications (Zeng, Z. and Wang, J., eds.), vol. 67 of Lecture Notes in Electrical Engineering, pp. 337346, Springer Berlin Heidelberg, 2010.

• P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Singing voice recognition based on matching of spectrogram pattern", in Proc. IJCNN, 2009, pp.1595-1599.

Thesis

by Peerapol Khunarsa

Thesis

ORIGINALITY REPORT

18 %	9 %	13 %	5 %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Peerapol Khunarsal. "Singing voice recognition based on matching of spectr… *Publication*	3%
2	Submitted to Chulalongkorn University *Student Paper*	1%
3	Al-Zhrani, Saleh AlQahtani, Mubarak. "Audio environment recognition using z.. *Publication*	1%
4	www.learnartificialneuralnetworks.com *Internet Source*	1%
5	Byeong-jun Han. "Environmental sound classification based on feature collab.. *Publication*	1%
6	Pafan Doungpaisan. "Singer Identification Using Time-Frequency Audio Fea… *Publication*	1%
7	Aci, M.. "K nearest neighbor reinforced expectation maximization method", E… *Publication*	1%
8	www.algorithm.randki-na-seks.waw.pl *Internet Source*	1%
9	wwwiti.cs.uni-magdeburg.de *Internet Source*	1%
10	mtg.upf.edu *Internet Source*	1%
11	www.tcts.fpms.ac.be *Internet Source*	< 1%
12	Peerapol Khunarsa. "Impulsive Environment Sound Detection by Neural Cl… *Publication*	< 1%
13	music.kosmix.com *Internet Source*	< 1%
14	Submitted to KTH - The Royal Institute of Technology *Student Paper*	< 1%
15	wwwpub.zih.tu-dresden.de *Internet Source*	< 1%
16	www.actapress.com *Internet Source*	< 1%

75	www2.austin.cc.tx.us *Internet Source*	< 1%
76	scholar.lib.vt.edu *Internet Source*	< 1%
77	www.ims.tuwien.ac.at *Internet Source*	< 1%
78	Schumacher, M.. "Neural networks and logistic regression: Part I", Comput... *Publication*	< 1%
79	Roberto Iglesias Rodríguez. "Comparison of several chemometric techniqu... *Publication*	< 1%
80	www.rle.mit.edu *Internet Source*	< 1%
81	www.jdl.ac.cn *Internet Source*	< 1%
82	Amir, A.. "Automatic generation of conference video proceedings", Journal ... *Publication*	< 1%
83	www.gits.kmutnb.ac.th *Internet Source*	< 1%
84	Ilk, H.G.. "Adaptive time scale modification of speech for graceful degrading... *Publication*	< 1%
85	Hui Li Tan. "Rhythm analysis for personal and social music applications usin... *Publication*	< 1%
86	www.idiap.ch *Internet Source*	< 1%
87	Santosh Gaikwad. "Feature extraction using fusion MFCC for continuous m... *Publication*	< 1%
88	student.grm.hia.no *Internet Source*	< 1%
89	www.coursehero.com *Internet Source*	< 1%
90	Yee Leung. "Algorithmic Approach to the Identification of Classification Rule... *Publication*	< 1%
91	Dhanalakshmi, P.. "Pattern classification models for classifying and indexing... *Publication*	< 1%
92	Hiroshi Okuno. "Automatic Synchronization between Lyrics and Music CD R... *Publication*	< 1%
93	Bunte, K.. "Stochastic neighbor embedding (SNE) for dimension reduction ... *Publication*	< 1%
94	M. Roelands. "An overlap-add technique based on waveform similarity (WS... 	< 1%

	Publication	< 1%
95	library.utcc.ac.th *Internet Source*	< 1%

EXCLUDE QUOTES AN

EXCLUDE BIBLIOGRAPHY AN

EXCLUDE MATCHES OFF