Acoustic-based Machine Learning Diagnostic Tool for Voice Disorders

Huiyi Wu

Department of Electronic and Electrical Engineering

University of Strathclyde

Glasgow, United Kingdom

This dissertation is submitted for the degree of

Doctor of Philosophy

2020

Blessings for all of us:

Imagine there's no countries
It isn't hard to do
Nothing to kill or die for
And no religion, too
Imagine all the people
Living life in peace
You may say I'm a dreamer
But I'm not the only one
I hope someday you will join us
And the world will be as one
Imagine no possessions
I wonder if you can
No need for greed or hunger
A brotherhood of man
Imagine all the people
Sharing all the world
You may say I'm a dreamer
But I'm not the only one
I hope someday you will join us
And the world will live as one

"*Imagine*" by *John Lennon*

Huiyi Wu
University of Strathclyde, 2020

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Huiyi Wu

2020

Acknowledgements

The research in this thesis has been carried out at the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom.

I would like to appreciate my supervisor Professor John Soraghan, Professor Anja Lowit, and Dr Gaetano Di Caterina for their excellent and invaluable support, encouragement, guidance and contribution towards the developments of successful completion of the research work presented in this thesis. I sincerely appreciate the trust and confidence they give me in conducting research work. Under their exceptional supervision, the PhD journey helped me to learn critical and analytical thinking, creative idea generation, resource searching, and collaborating research.

I would like to thank my colleagues at the Centre for Signal and Image Processing and the Speech and Language Therapy Research groups for their peer reviews, support and collaboration. I also appreciate my friends in Glasgow: Shenchi Zhang, Xiao Liu, Hong Yang, Keerati Kaewrak, Yilan Xiao, Ping Ma, Lei Luo, Yura Zaiki, Weijie Ke, Baixiang Zhao, Yannan Xing, Ming Gong, Guoliang Xie, Kletia Vassiliou, and Xrusa Fwth; my ex-boyfriends: Lin Zhang and Cong Niu; my rock band friends: Leon Hui, Hewei Li, watermelon, and Zhe Wang. In addition, I really appreciate the friends in Edinburgh, Christopher Galazis, Emile Mackute, and Ming Tang. I appreciate the seaside breeze near Edinburgh Leith walk and the industrial atmosphere in Glasgow. Finally, I would like to appreciate my family who stand by me and support me throughout the years.

This research would have been impossible without the research grants from the University of Strathclyde and Capita Fund.

Huiyi Wu
University of Strathclyde, 2020

Abstract

The research presented in this thesis addresses the application of deep neural networks and digital signal processing algorithms in the pathological voice detection. In this thesis, the novel methods are presented, including deep acoustic recurrent model that combines frame-based cepstral and spectral features and Bi-directional Long short-term memory (Bi-LSTM) network, a 10-layer convolutional neural network (CNN) model with spectrogram of the speech as input, transfer learning from image recognition applications to pathological voice detection field using time-frequency representation as input, and a novel CNN model using data augmentation idea with scalogram of the speech as input.

The deep acoustic recurrent model explores the relationship of frame-based cepstral features with RNN model. Two novel cepstral features based on cepstrum are proposed: Second Peak Perturbation (SPP) and standard deviation of cepstrum (CepStd). These novels cepstral features are validated to improve the classification performance on three databases. In addition, traditional acoustic analysis is compared with the proposed deep acoustic recurrent model. It is shown that frame-based cepstral features shows overall better performance on deep recurrent model than traditional classifiers.

A 10-layer convolutional neural network is proposed with spectrogram of the speech as input. This is the first model that applies time-frequency representation in deep learning for pathological voice detection. The experimental results have shown that it is an effective and efficient model for detecting pathological speech data. However, it shows overfitting problem to some extent. This is a commonly seen problem due to the small data size. In order to address this issue, transfer learning with state-of-the-art CNN networks from image recognition field is applied in the pathological voice detection field. The results shows that transfer learning improves the testing data accuracy. However, the overfitting problem is still severe.

Finally, the concept of data augmentation is explored and a novel CNN model called the R-Net is proposed. This method uses continuous wavelet transform to obtain the scalograms of the speech onset, and data augmentation within a CNN environment. This model significantly reduces the overfitting problems, and improves the testing

performance between 15% to 20% on the most challenging SVD database. It validates the efficiency of data augmentation on small-data-size problems.

List of Figures

List of Tables

Abbreviations

ADSV	Analysis of Dysphonia in Speech and Voice software
AE	Auto-Encoder
ANN	Artificial Neural Network
APQ	Amplitude Perturbation Quotient
ATRI	Amplitude Tremor Intensity Index
AUC	Area Under the Curve
AVPD	Arabic Voice Pathological Database
BERT	Bidirectional Encoder Representations from Transformers
CAPE	Consensus Auditory Perceptual Evaluation
CDBN	Convolutional Deep Belief Network
CNN	Convolutional Neural Network
CPP	Cepstral Peak Prominence
CRBM	Convolutional Restricted Boltzmann Machine
CSL	Computerized Speech Laboratory
CWT	Continuous Wavelet Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DL	Deep Learning
DNN	Deep Neural Network
DSH	Degree of Subharmonics
DUV	Degree of Voiceless
DVB	Degree of Voice Breaks
DWT	Discrete Wavelet Transform
EEG	Electroglottography
ENT	Ear, Nose and Throat
FDR	Fisher Discriminant Ratio
FN	False Negative
FP	False Positive
FPR	False Positive Rate

FTRI	Frequency Tremor Intensity Index
GMM	Gaussian Mixture Model
GNE	Glottal to Noise Excitation Ratio
GRBAS	Grade, Roughness, Breathiness, Asthenia, and Strain.
HAN	Hierarchical Attention Network
HMM	Hidden Markov Model
HNR	Harmonic-to-Noise Ratio
HPV	Human Papilloma Virus
KNN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LHR	Low-to-High Harmonic Ratio
LPCC	Linear Prediction Coding Coefficients
LRT	Likelihood Radio Test
LSTM	Long Short-Term Memory
LTAS	Long-Term Average Spectrum
MDVP	Multi-Dimensional Voice Program
MEEI	Massachusetts Eye and Ear Infirmary
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
NHR	Noise Harmonic Ratio
NLP	Natural Language Processing
NNE	Normalized Noise Energy
NSH	Number of Subharmonics
NUV	Number of Unvoiced Segments
PCA	Principle Component Analysis
PDA	Principe de Asturias Database
PLP	Perceptual Linear Prediction
PPQ	Period Perturbation Quotient
PVD	Pathological Voice Detection
RAP	Relative Average Perturbation
RBM	Restricted Boltzmann Machine
RNN	Recurrent Neural Network
ROC	Receiver Operating Curve

RRP	Recurrent Respiratory Papillomatosis
SLT	Speech and Language Therapists
SNR	Signal-to-Noise Ratio
SPL	Speech Pressure Level
SPP	Second Peak Perturbation
STFT	Short-Time Fourier Transform
SVD	Saarbrucken Voice database
SVM	Support Vector Machine
TPR	True Positive Rate
VAD	Voice Activity Detection
VCN	Vocal Cord Nodule
VTI	Turbulent Noise Index

Table of Contents

Chapter 1

1 Introduction

1.1 Preface

Dysphonia refers to disorder of voice which is caused by changes in the phonatory organs such as tissue infection, systemic changes, mechanical stress, surface irritation, tissue changes, neurological and muscular changes, abnormal muscle tension and other factors [1]. It is characterized by prolonged hoarseness and loss of voice. According to statistics, it is estimated that over 50000 patients are referred by their General Practitioner (GP) to otolaryngology/voice clinics in UK per year[2]. Due to the complexity of the causes of dysphonia, the diagnosis requires Speech and Language Therapists (SLT), Ear, Nose and Throat (ENT) clinicians, and it might also require some other professionals including neurologists, and psychologists [3]. This has a significant impact on the UK economy considering the cost of lost productivity, workforce replacement and medical treatment. Studies have shown that there is a significant statistical correlation between dysphonia and the quality of life [4, 5]. It is essential that dysphonia is treated effectively in order to allow the affected individual to function normally within society and to ensure their wellbeing.

Dysphonia can be classified in different categories by the severity, which often can be distinguished depending on whether there are organic changes on the larynx. Around 15% of dysphonia are organic dysphonia, which would display organic changes and disease in the larynx that require surgery, including cancer, ulcers or vocal nodules. Organic dysphonia are physiological voice disorders in nature and result from alterations in respiratory, laryngeal, or vocal tract mechanisms. The remaining type of dysphonia are functional dysphonia, which results from improper or inefficient use of the vocal mechanism when the physical structure is normal. (E.g. vocal fatigue; muscle tension dysphonia or aphonia; diplophonia; ventricular phonation).

Patients presenting with prolonged voice problems will be referred to the Ear Nose and Throat (ENT) consultant by their GP. ENT consultation will exclude the possibility of cancer and determine the requirement for surgery if it is organic dysphonia, or apply a range of behavioural voice therapy techniques from Speech and Language Therapy if patients are diagnosed with functional dysphonia.

Organic dysphonia can lead to serious consequences if the referral system does not meet the requirements, while functional dysphonia may be developed to organic type if it is remained untreated which will lead to even more costly intervention. In this case, GP routinely refer patients with prolonged hoarseness to ENT, which consumes significant costs and resources in NHS, as patients will require to be seen by a consultant and undergo expensive laryngeal examination (endoscopic imaging of their larynx). Due to only 15% of the patients presenting with the organic dysphonia, 85% of the examination are therefore unnecessary so that the alternative reliable method of diagnosing dysphonia need to be developed.

Perceptive measurement are the main diagnostic methods used in clinicians. However, as signal processing tools are becoming more mature and computers are introduced in clinics, acoustic measurement has becoming more common these days[6-8]. The speech of the patients would be imported to the diagnostic system and acoustic features can be automatically extracted for classification. Novel classification methods such as deep learning have appeared in recent years and shown to have more significant impact in speech processing field [9, 10]. However, deep learning has not been explored deeply in pathological voice detection field yet.

1.2 Motivation

Patients are referred to General Practitioners (GP) first to distinguish whether it is organic dysphonia (requires surgery) or functional dysphonia (do not require surgery). The GP performs perceptual analysis to assess the severity and individual characteristics of the dysphonia. However, these procedures require medical experts to make diagnostic decisions, and hence suffer from a certain level of subjectivity. If the patients are diagnosed to be severe organic dysphonia, they will be referred to ENT clinicians for further treatment. These patients are assessed using expensive instrumental examination using fibre-optic endoscopy, and laryngoscopy. The

endoscopy requires to be put in patients nose, which will lead to discomfort and invasiveness. Perceptual analysis suffers from subjectivity, while the instrumental examination is expensive and invasive. Therefore, non-invasive acoustic analysis assistant tool becomes a popular research topic.

1.3 Research Aim & Objectives

The aim of this project is to develop a non-invasive acoustic analysis tool for dysphonia, using state-of-the-art deep neural network techniques and signal processing methods. This will help to improve the referral process, eliminating unnecessary examination of many patients by ENT consultants. The objectives of the research include the following:

- Extracting proper features from acoustic speech, which is capable of representing dysphonic characteristics
- Exploring deep learning architectures for detecting dysphonic speech
- Develop the system for pathological voice detection

1.4 Contributions

There are three main novel contributions in this research which include:

A. Development of a deep acoustic recurrent model for pathological voice detection. Two novel cepstral features are proposed: second peak perturbation (SPP) and standard deviation of cepstrum (CepStd), and combine the traditional cepstral and spectral features for a Bi-LSTM model for training. This model achieves 71.60%, 82.58%, and 78.43% accuracy respectively on pathological voice detection with Saarbrucken Voice Database (SVD), Principe de Asturias Database (PdA) and Arabic Voice Pathological Database (AVPD). This is the first time cepstral features that has been explored in RNN models, and the proposed two novel cepstral features shows strong correlation with dysphonia too. (Chapter 4)

B. A novel 10-layer CNN model is proposed for pathological voice detection. The proposed model is the first one exploring spectrogram as input in pathological voice detection using deep learning techniques. This model works effectively with smaller network architecture compared to [11], which using raw speech data as input to a DNN. However, the overfitting problem appears significant due to the small-data-

size issue. In order to solve the small-data-size issue, transfer learning with state-of-the-art CNN models from image recognition field has been applied into this field. Different types of 2D representations (spectrogram, zoomed spectrogram, scalogram) of pathological and normal speech are explored. ResNet achieves the best testing accuracy using scalogram as input, with 70.75%, 80.30%, and 80.39% testing accuracy on SVD, PdA, and AVPD database.

C. A novel CNN model R-Net is proposed for pathological voice detection. This model uses the idea of data augmentation, and it significantly reduces overfitting problem and improves the performance by around 15% to 20% on the most challenging SVD database compared to the benchmark. It achieves 89.27%, 75.62%, and 85.05% accuracy for pathological voice detection on SVD, PdA and AVPD databases respectively. R-Net helps to reduce the overfitting problems greatly and improves the overall performance. (Chapter 6.3)

In addition, transfer learning from pathological voice detection model R-Net to paediatric Recurrent Respiratory Papilloma (RRP) data detection project also shows successful performance with very-small-sized data. (Chapter 6.4)

1.5 Outline of the Thesis

This thesis consists of seven chapters. The first chapter is an introductory chapter which includes the background of dysphonia, motivation, research aim and objectives, novel contributions, and academic outputs of this research. Chapter 2 reviews causes, characteristics and symptoms of dysphonia, and both perceptual-based and acoustic-based analysis in pathological voice detection. Chapter 3 reviews the feature extraction and classification techniques used in this field, with sufficient literature on the state-of-the-art deep learning architectures. The novel contributions of this research are presented in Chapters 4, 5 and 6. The first novel contribution is combining traditional cepstral features and proposed novel cepstral features with deep recurrent model for pathological voice detection. This is the first time frame-based cepstral features applied on deep learning models. In chapter 5, a novel deep convolutional spectrogram model is proposed, and compared with state-of-the-art CNN model transfer learning performance on detecting dysphonia. Before our work, the work in [11] is the first and only work successfully using deep learning in

pathological voice detection. It applies RNN model on raw speech data. However, our work is the first time transforming the speech into time-frequency 2D representations on deep learning models for pathological voice detection, which achieves better performance compared to the previous work. Both deep recurrent acoustic model and convolutional spectrogram model are shown to provide successful performance compared to the bench-mark. However, due to the data limitation, over-fitting problems are hard to be avoided. There are generally two ways of solving the problem. One is transfer learning shown in Chapter 5.3, another is data augmentation methods, which is applied in R-Net in Chapter 6. This uses the noise-robustness idea, and focuses on addressing the over-fitting problems. It improves the result by 15% to 20% on the most challenging SVD database, and it performs successfully on detecting paediatric RRP data with transfer learning. The last chapter of this thesis, Chapter 7, gives a summary of contributions in this research work as well as the future works relevant to this study.

Chapter 2

2 Review of pathological voice detection and acoustic feature extraction techniques

2.1 Introduction

In this chapter, a review of pathological voice detection and the acoustic analytical technologies for feature extraction is discussed. In order to understand dysphonia, the causes, characteristics and symptoms are described in the beginning. This provides the etiological view for different types of dysphonia. There are two main ways of assisting diagnostic methods for dysphonia [8, 12-14]. One is perceptual analysis, which can be subjective due to the dependency on experiences of practitioners [15-17]. Another one is acoustic analysis, which is relying on the latent pathological features in the voice recordings, especially sustained vowels [6, 18, 19]. Acoustic analysis has been the most popular assistant tool for pathological voice detection, and our research will focus on acoustic analysis too. The characterisation process extract acoustic features that are able to describe the underlying attributes of different pathologies [6]. The literature review of several perspective view of characterisation will be discussed, which is essential for pathological classification process in the next chapter.

2.2 Causes, Characteristics, and Symptoms of Dysphonia

There are several ways of organizing and classifying dysphonia. Above all, etiologic method reveal the underlying causes for signs and symptoms. In addition, dysphonia describes abnormal voice in its psychoacoustic parameters of pitch, loudness, quality, and variability[20]. These acoustic perceptual features can be used to classify dysphonia in a perceptual way. Another approach is to classify disorders with regards to malignant or benign pathologies.

According to etiological view, over 120 different laryngeal voice states are categorized into 8 major groups in [21]:

1. Structural pathologies
2. Inflammatory conditions
3. Trauma
4. Systemic conditions
5. Non-laryngeal autodigestive disorders
6. Psychiatric and psychological disorders
7. Neurological voice disorders
8. Other disorders

This section organizes the major disorder groups to provide a better comprehensive structure of the causes of dysphonia, by discussing organic dysphonia (caused by changes in the larynx with neurological or structural abnormality reasons), psychogenic dysphonia (functional dysphonia caused by psychological reasons), and mixed dysphonia (complex characteristics with unknown etiological reasons).

2.2.1 Organic Dysphonia

Organic Dysphonia is a disease with structural abnormalities in the larynx itself or a laryngeal structure and function alteration due to neurological diseases.

A. Dysphonia caused by structural abnormalities

1) Vocal Fold Nodules

Vocal fold nodules are caused by phono traumatic behaviours such as over high intensity and repetitive overuse in talking, coughing, screaming, singing with over high or over low pitch [22]. It commonly appears in young male children and female adults, with higher pitch and similar vibratory patterns [23]. Overuse of vocal fold lead to irritation in some areas. This area can develop into a callous texture region with continued unrest or misuse of the vocal folds [24]. Figure 2.1 shows an example of vocal fold nodule. Perceptually, voice quality has only subtle minor changes in the beginning stage of the development. With the nodule getting firmer and harder, the voice quality appears hoarseness and slight breathiness due to the fatigue of vocal fold. This lead to some perceptual conditions such as overpressure of glottal airflow and respiratory system, and asymmetric vibration of vocal cord [25]. Singers are

special groups of patients diagnosed with this pathology, and it is reported that they experience “loss of vocal range”, fatigue, and less endurance in vocal folds physiologically.

Figure 2.1 Vocal fold nodule [26]

2) Vocal Fold Polyps

Similar to vocal cord nodules, acute vocal injury or different forms of phono traumatic behaviours including singing, overuse and excessive talking will lead to vocal fold polyps. One sudden individual trauma will also cause a type of polyp called haemorrhagic polyp, with an abrupt start of hoarseness [27]. However, compared to vocal fold nodules, it forms with more blood vessels and the size will increase quickly. One example of vocal fold polyps are shown in Figure 2.2. The occurrence in different positions in larynx and different size of the polyp will cause various degree of effects to the voice quality [28]. Hoarseness, roughness and breathiness is still the main physiological perceptual signs. In some occasions, no specific voice quality changes appear when the pedunculated polyp grows below the vocal cord edge, while this will lead to overpressure of the inspiratory system and challenge of breathing.

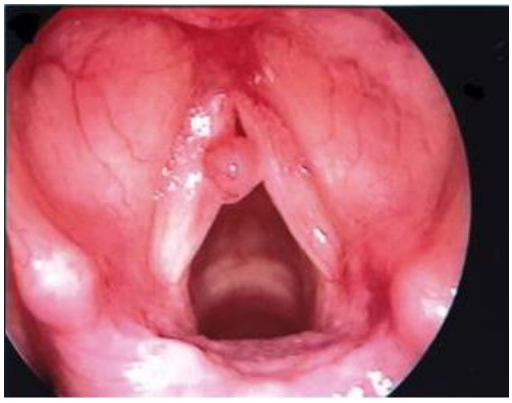


Figure 2.2 Vocal fold polyp [29]

3) Cysts

Vocal cord cyst grows with a sac around a benign fluid-filled lesion. Similar as vocal fold nodules and polyps, it is caused by phono traumatic behaviours or instant trauma[30]. Cysts commonly exist in the midmembranous position of the vocal cord[31]. One example of vocal cord cysts are shown in Figure 2.3. Perceptually, vocal cord asymmetry and abnormal vocal cord closure will lead to different degrees of hoarseness. Different size and shape of the cyst will cause variation in voice quality and the firmness. Globus sensation may appear too with the increasing size of cyst.

Figure 2.3 Vocal Cord Cysts [32]

4) Laryngeal Papilloma

Vocal cord nodule, polyps and cyst are all caused by phono traumatic behaviours. Different from that, laryngeal papilloma is caused by a type of virus called human papilloma virus (HPV). One example are shown in Figure 2.4. It also called recurrent respiratory papillomatosis (RRP) or glottal papillomatosis. There are two forms of RRP, with adult-form RRP and paediatric-form RRP. It is a rare disease and more seen in children than in adults are (ratio: 4.5 to 2 in 100000 individuals). The

common sign and symptom of RRP is hoarseness. With the gradual development, it might lead to fatal symptom of airway obstruction.

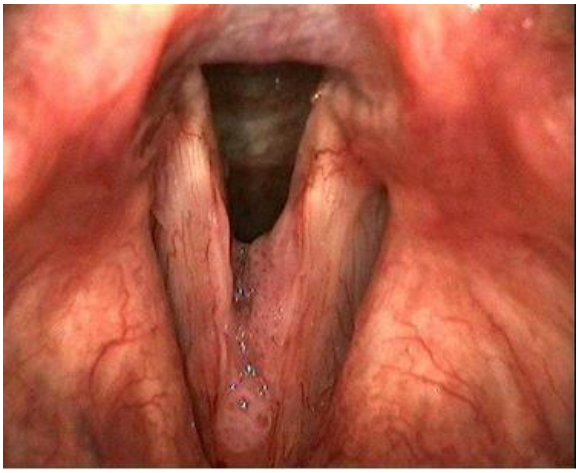


Figure 2.4 Laryngeal Papilloma [32]

5) Reinke's edema

Most edema grow on the superficial layer of the vocal cord. It appears with the swelling of the vocal cords due to the accumulating fluid [33]. One example of this pathology are shown in Figure 2.5. It was first recognized by Reinke et al [34] in 1895, so it is called as Reinke's edema. The main symptom of Reinke's edema is hoarseness, similar to laryngitis. Research found that smoking is the main cause of Reinke's edema, with 97% of the patients have the smoking habits. Other reasons such as overuse of the vocal fold, gastroesophageal reflux, and hypothyroidism are also risking factors. It is a benign (non-cancerous) polyp takes up to 10% of all the benign vocal cord changes.

Figure 2.5 Reinke's edema [32]

6) Laryngitis

Laryngitis describes the vocal cord inflammation phenomenon. It can be regarded in two types, chronic laryngitis or acute laryngitis [20]. Acute laryngitis is normally

one of the symptom caused by upper respiratory tract infection [35]. There are other factors including vocal cord trauma (caused by coughing) or non-virus infection by pathogens. Chronic laryngitis is generally caused by factors such as smoking, tuberculosis, allergy, gastroesophageal reflux, rheumatoid arthritis, or sarcoidosis [36, 37]. The commonly seen symptom of laryngitis is hoarseness, difficulty of swallowing, and fore-neck pain. Acute laryngitis patients generally recover by themselves with reduced usage of vocal cord. Chronic laryngitis patients are commonly seen in adults, with more male patients than female ones.

7) Granular Cell Tumour

Similar to vocal cord nodules, acute vocal injury or different forms of phono traumatic behaviours including singing, overuse and excessive talking will lead to vocal fold polyps. Granular cell tumour is a type of tumour that can develop on any part of skin or mucosal surface [38]. Tyrosine-protein phosphatase non-receptor type 11 (PTPN11) gene mutation is the cause of LEOPARD syndrome, where multiple granular cell tumours appears. It will affect every part of the body, while head and neck part will be affected 45% to 65% of the time. Vocal cord Granular Cell Tumour takes up to 10% in head-and-neck cases [39].

8) Laryngeal cancer

Laryngeal cancer is formed on the squamous epithelial cells of the laryngeal epithelium. An example of this laryngeal cancer is shown in Figure 2.6. The main risk factor that will lead to laryngeal cancer is smoking [40]. The death rate from this disease on smokers is 20 times more than it on non-smokers [41]. Abuse consumption of alcohol is another main factor, which has synergistic effect with smoking [40]. The symptoms of laryngeal cancer can be different with different locations and size of the cancerous area. The main symptoms would be hoarseness or other voice quality changes, sore throat, continuous cough, wheezing (due to the airway blockage by a tumour) or earache. Male over 55 years old tend to have higher risk of diagnosed with this disease [42].

Figure 2.6 laryngeal cancer [43]

9) Leucoplakia

Leucoplakia refers to a symptom of white plaques that adhere to the oral mucosa, and is generally related to laryngeal cancer. An example of Leucoplakia is shown in Figure 2.7. The cause of leucoplakia has not been fully determined yet[44]. The possible factors include smoking, abuse assumption of alcohol, or chewing betal nut [45, 46]. Leucoplakia is seen as a tissue mutation that may develop to cancer. The risk of leucoplakia relates to age too[47]. In most cases, patients diagnosed with leucoplakia is over 30. Male patients over 70 years old have 8% risk of developing leucoplakia.

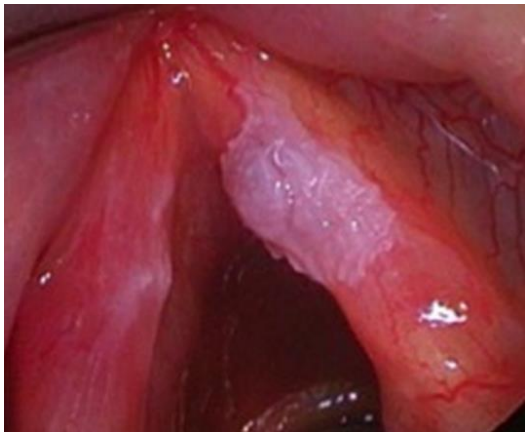


Figure 2.7 Leucoplakia [48]

B. Neurological dysphonia

1) Vocal fold paralysis

Vocal fold paralysis is also called recurrent laryngeal nerve paralysis. It is caused by lesion of recurrent laryngeal nerve, which control the muscles of vocal cord. The

unilateral injury of the nerve will lead to hoarseness due to the irregular vibration of the vocal cord. Reduced mobility will cause the weakness of the voice. The bilateral injury will impair air-flow and lead to breathing difficulties. The cause of vocal fold paralysis varies. It includes congenital causes, infectious reasons, tumours, trauma, endocrinologic diseases, and systematic neurological diseases. Study found that around 26% of the patients in the voice disorder clinics are diagnosed with vocal fold paralysis[49]. While in some cases, the risk of vocal fold paralysis takes up to 80% of all the patients with voice disorders[50].

2) Spasmodic dysphonia

When the muscle around vocal cord develop into a spasm condition, we call it spasmodic dysphonia [51]. The mains symptom is discontinuous speech with breaks between sentences, which makes it hard for people to understand [52]. The voice quality is affected with hoarseness and breathiness in the beginning [53]. In addition, the voice sounds tense and strained, or the patient might even unable to speak [52]. Voice tremor may also appears. The reason that cause spasmodic dysphonia is not determined yet. In this case, spasmodic dysphonia can be regarded as mixed dysphonia sometimes. This also increase the difficulty in diagnosis [20, 54]. The possible main factor might be the family history. Other causes are infected upper respiratory tract, trauma, overuse or heavy use of the vocal cord, and psychological reasons. This disease is generally related with central nervous system, and is a type of focal dystonia. It generally appears in mid-age people, but it can also be seen in people in twenties [51].

2.2.2 Psychogenic dysphonia

Other than dysphonia with organic changes in the vocal fold, there is another type of dysphonia that with poor voice quality without any anatomical changes in the vocal fold, which is called psychogenic dysphonia, also called functional dysphonia. There are two types of functional dysphonia. One is hypo-functional dysphonia, which shows incomplete closure of the vocal cords. Another type is hyper-functional dysphonia, which shows the adduction of the vocal cord with overuse of the muscles. The sign of voice quality changes would be hoarseness and roughness, instability and weakness. The cause of functional dysphonia is not determined yet. It is reported to be related with stress and virus diseases such as measles and mumps[55]. Survey

conducted in Bedford College with a group of dysphonic patients. Evidence shows that patients group report 54% of patients have “conflicts of speaking out” recently compared to the female control group with only 16% reported[56]. One study which focus on structured psychiatric interviews reported that there are one third of the functional dysphonia patients are diagnosed with adjustment disorders[57]. This corrects the former research suggesting that functional dysphonia patients are all tend to have hysterical condition[55].

2.3 Perceptual Techniques used in Pathological Voice Detection

Perceptual analysis for speech is one of the most traditional method in analysing voice quality. Speech language therapists evaluates patient speech based on auditory impression. Then speech language therapists will give diagnostic results for dysphonia patients. In this way, it will help to make treatment plans for improving the quality of life for patients.

Auditory perceptual analysis is recognized as a golden standard method for dysphonia worldwide[16, 58]. However, evaluators have different experiences and knowledge levels of perceptual analysis[59]. In this case, the evaluating tools for this method are investigated to reduce latent variability and inconsistency. Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) [30, 60] and Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) [61] are two popular perceptual analysis tools for clinicians to make decisions. GRBAS refers to Grade, Roughness, Breathiness, Asthenia, and Strain. These speech characteristics will be introduced in detail in Chapter 2.4. It is the most common method for clinicians and researchers to evaluate voice disorders. However, the reliability of GRBAS is doubted in general. It is reported that it will take eight hours training for four unexperienced evaluators to get 80% similarity in evaluating results[12]. In [13], it claims that grade(G) parameter ranging from fair to good, while the other four parameters ranging from moderate to fair. It also proved the significant effect of experiences to evaluation results.

Consensus Auditory Perceptual Evaluation (CAPE-V) is capable of analysing similar metrics without Asthenia. Apart from this, CAPE-V uses visual analogue scale, which allows practitioners to evaluate pitch and loudness. It also allows two

parameters that are not settled, and it contains other functions such as resonance classification.

GRBAS classify patient voice by severity of dysphonia, while CAPE-V contains scales of mild, moderate, and severe dysphonia with asymmetric distribution. Kreiman et al [62] suggests that the visual analogue scaling process in CAPE-V would help to solve the limitations in GRBAS method. However, there are still discussions in GRBAS and CAPE-V advantages. In [63], researchers compare the original GRBAS and visual analogue scale edition of GRBAS that they developed. They claim that clinicians shows better agreement with the original GRBAS than the visual analogue scale edition of GRBAS does.

2.4 Acoustic-based Techniques used in Pathological Voice Detection

In section 2.3, we introduced the perceptual method in diagnosing dysphonia. However, perceptual analysis have limitations that the results rely on clinician's experiences. This places significant pressure on the clinicians. Another evaluating method for diagnosis of dysphonia is based on instrumental examination, such as electroglottography, endoscopy, aerodynamic analysis etc [8, 64]. While this method requires expensive high quality instruments, acoustic analysis has become the alternative assistant tool for diagnosis of dysphonia. Acoustic analysis automatically measure the voice quality based on the acoustic features that extracted from the voice corpus. This is a non-invasive method that reduces the expenses for diagnosis and treatment.

There are four perceptual traits defining voice conditions, including *loudness*, *pitch*, *quality* and *variability*. *Loudness* refers to the intensity of the airflow from the vocal cord. Apart from the distance of the mouth to the microphone, the psychological reasons of the patients will affect loudness too. Loudness impairments are often caused by Parkinson disease that patients react with personality disorders, such as Hyperphonia and Hypophonia. *Pitch* refers to vibrating frequency of the vocal cord, which is the fundamental frequency. Pitch impairments are often due to the abnormal voice and tremor of the vocal cord. It appears with too high or too low pitch, sometimes with pitch breaks. *Quality* is constructed with various of acoustic features, which describes different psychological appearance[20, 65].

1) Breathiness: Breathiness appears with audible airflow due to the failure of vocal cord closure.

2) Roughness: Roughness shows the aberrant vocal fold vibration, which can be described as harshness too.

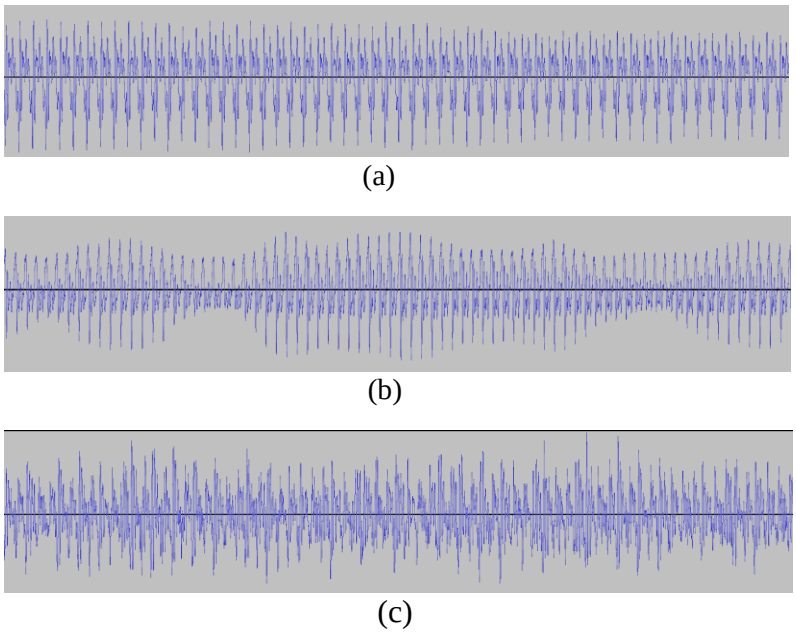


Figure 2.8 (a) Type I signal (periodic) (b) Type II signal (modulation) (c) Type III signal (aperiodicity)

3) Strain: It is caused by over adduction of the vocal cord, which is usually due to the neurological changes or psychological reasons. For example, muscle tension dysphonia and hyper-functional dysphonia all appears with strained voice quality.

4) Resonance: This shows with too much or too little airflow through the nose. Vocal tract (throat, mouth and nose) performs like filters to the airflow. Resonance impairment can be shown as Hypernasaliy and Hyponasality.

Strain, breathiness and roughness are voice aperiodicity metrics used to describe the degree of abnormality in the vocal folds. Therefore, these are also part of GRBAS metrics for perceptual analysis described in section 2.3. There are three main types of voice signals, as illustrated in Figure 2.8. Type I shows a periodic behaviour, which can be regarded as normophonic voice most of the times; Type II signal contains modulating frequencies and sub-harmonics owing to additional perturbations, which is often related with roughness; Type III signal represents aperiodic behaviour with additive noise which is often linked with breathiness [66-69]. Aperiodicity reveals that the vocal folds are dysfunctional and vibrating with non-constant airflow[70, 71].

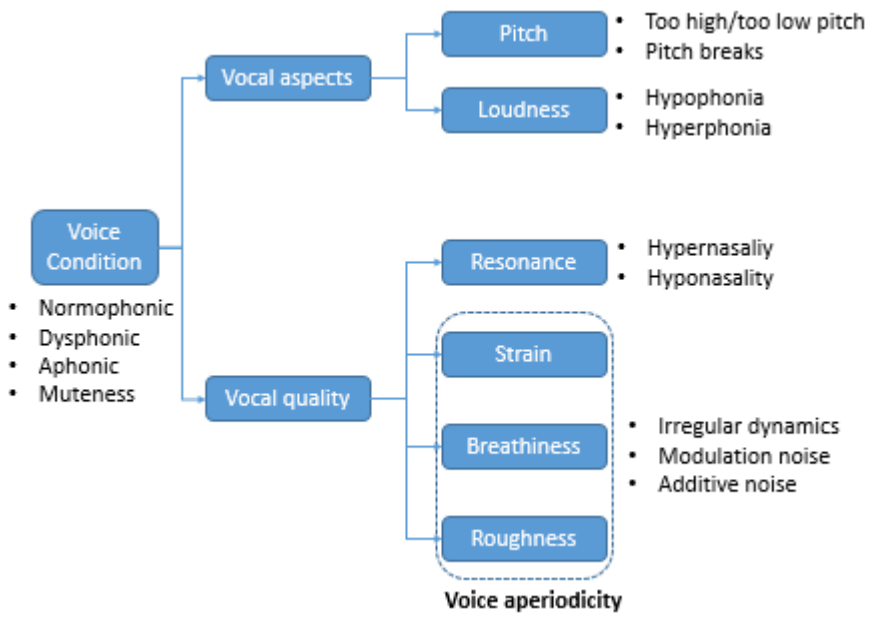


Figure 2.9 Perceptual traits of voice conditions

In conclusion, there are few levels of voice conditions: normophonic, dysphonic, aphonic and muteness. There are several perceptual traits to describe the voice conditions which are shown in Figure 2.9. This includes voice aspects *pitch, loudness,* and few characteristics that tells the vocal quality: resonance, strain, breathiness, and roughness. Among the characteristics, breathiness, roughness and strain show voice aperiodicity, which can be seen from the waveform.

In this section, we first introduce the signal pre-processing "framing" method for feature extraction, and then provide a critical literature review on acoustic feature extraction.

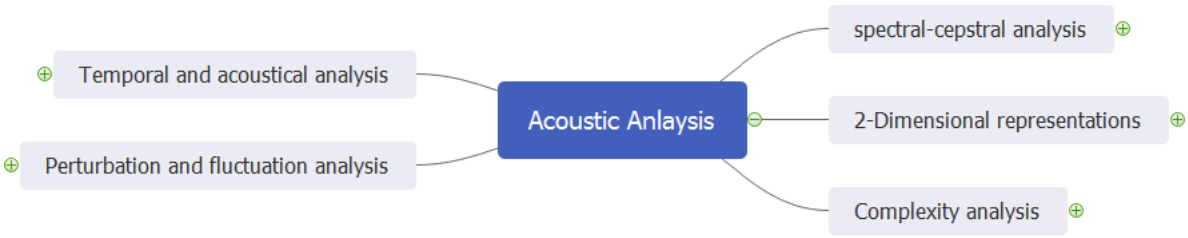
2.4.1 Framing before feature extraction

Speech signals have non-stationary characteristic, so that signal processing techniques are developed to conduct pre-processing steps on speech signals. Short-time signal analysis is a commonly seen technique to analyse speech signals in stationary status. First *framing* the signals into short-time small pieces; then *windowing* the frames to improve spectral properties. In this case, the frames are assumed to be stationary. The most popular choices of windows are *Hamming* [72-84] and *Hanning* [69, 85-94] windows. The length of the window is commonly set to 20ms [95-102] or 40ms [88, 103-110]. There are some other popular window

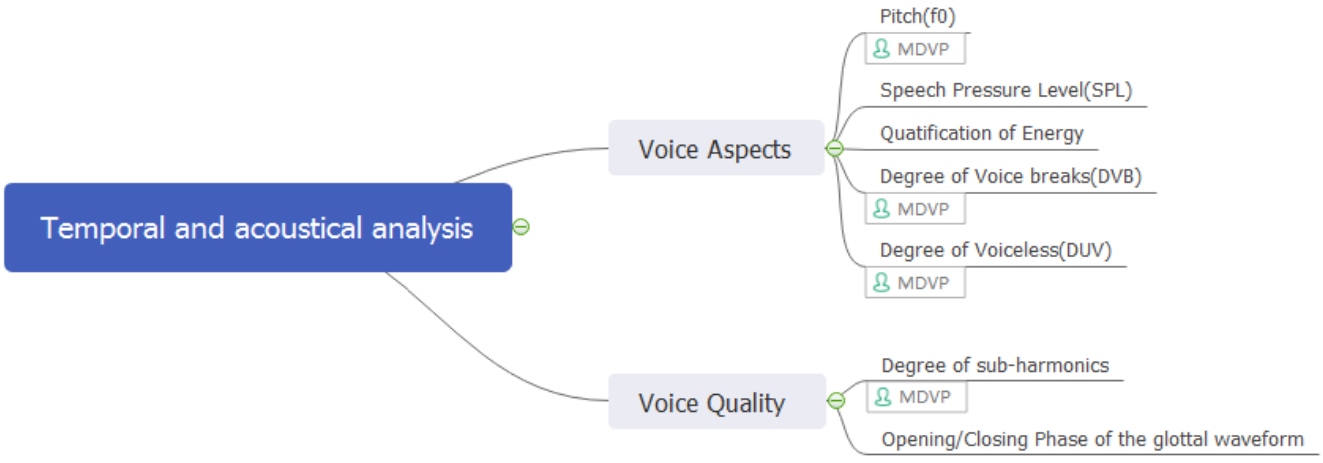
lengths in literatures, such as 10ms[85, 89, 111], 16ms[112], 25ms[113], 30ms[72, 78, 114-116], 50ms[76, 86, 92, 94, 117, 118], 55ms[70, 103] etc.

2.4.2 Acoustic feature extraction Mind Map

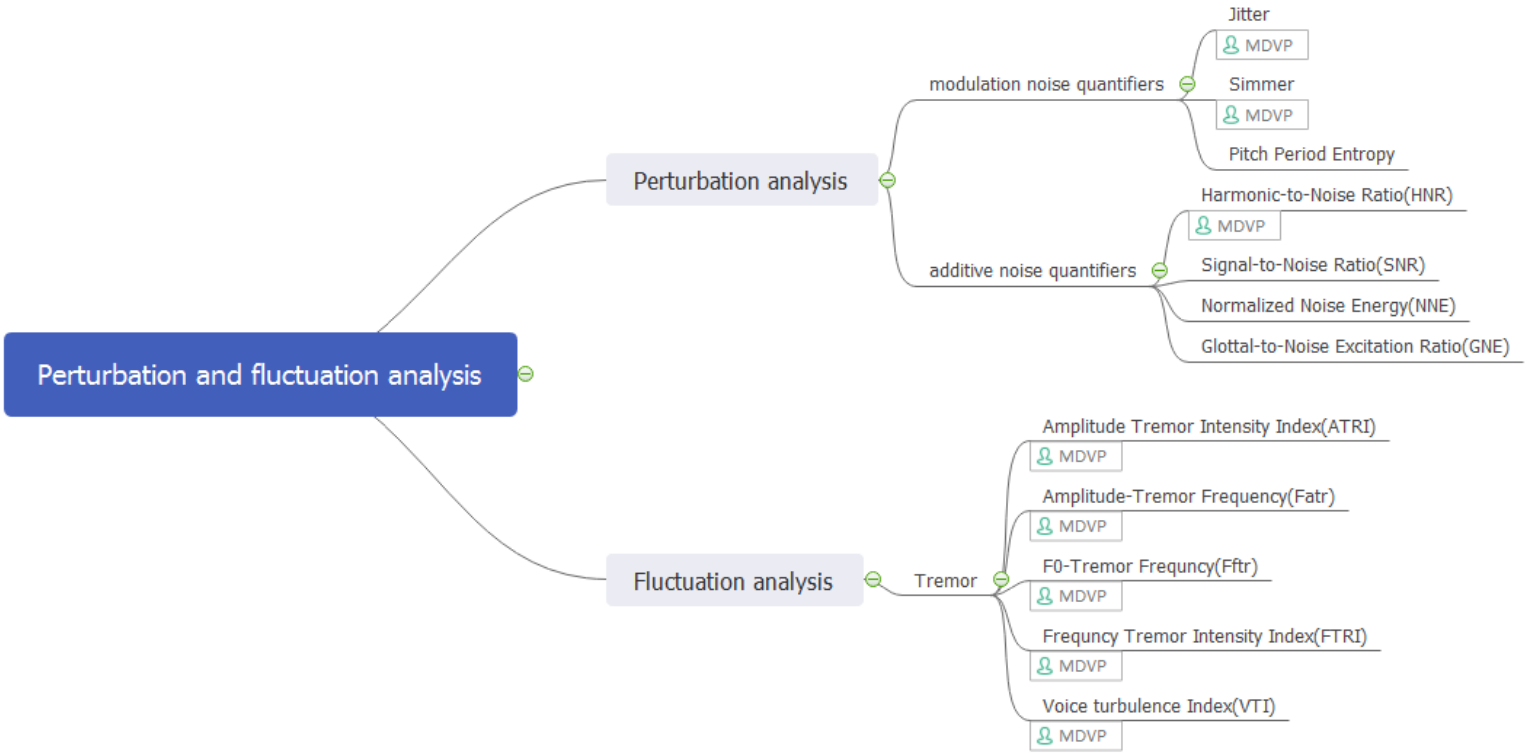
After pre-processing steps for transforming the speech signal into short-time frames, acoustic features representations are extracted from each frame. These features are short-time based features, while there exist other features that are extracted in long-term basis from the whole audio signal. With one or two related acoustic features, it is hard to distinguish pathological and normal speech[119]. This is because aperiodicity might appear in some non-pathological status with other disturbing reasons[120]. This is so called "weak label" issue which will be discussed in Section 2.5. Therefore, the most common way for pathological voice detection is to combine related significant features as a fusion to form a feature set. In Figure 2.10, the commonly used Acoustic features are listed in a Mind Map.



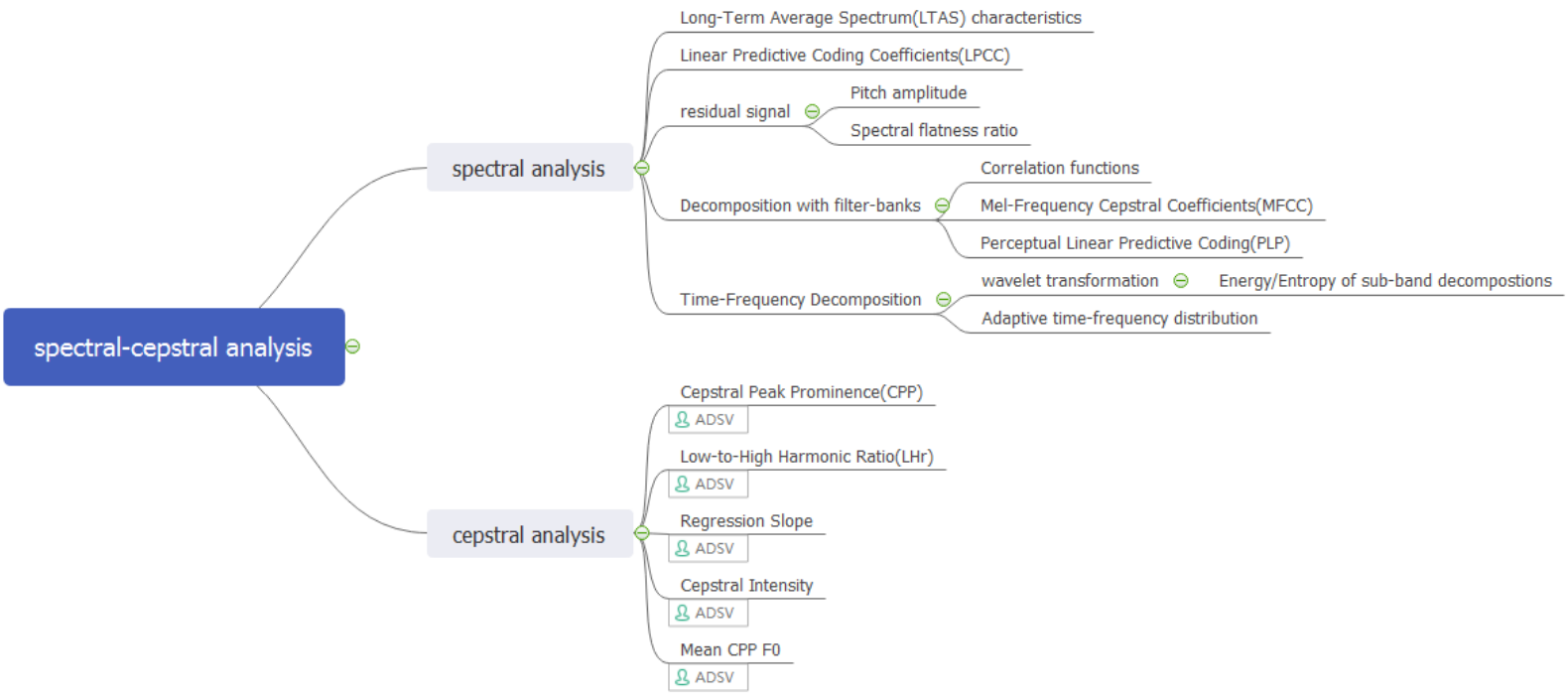
(a)



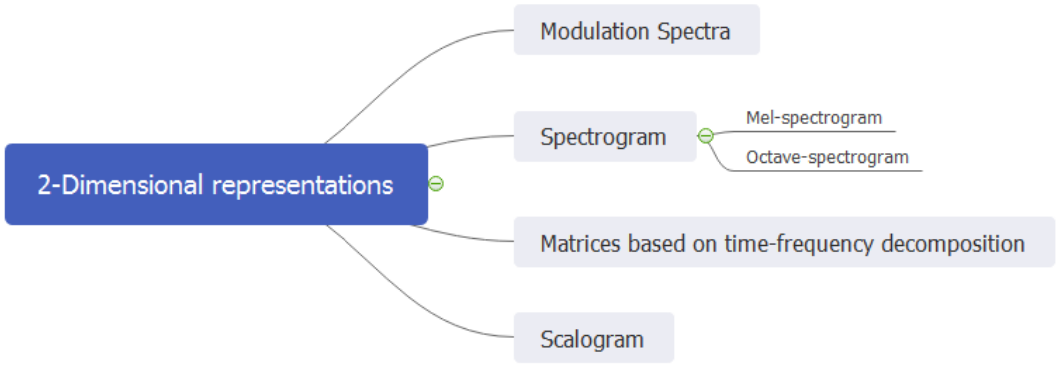
(b)


Perturbation and fluctuation analysis
Perturbation analysis
modulation noise quantifiers
Jitter
MDVP
Simmer
MDVP
Pitch Period Entropy
additive noise quantifiers
Harmonic-to-Noise Ratio(HNR)
MDVP
Signal-to-Noise Ratio(SNR)
Normalized Noise Energy(NNE)
Glottal-to-Noise Excitation Ratio(GNE)
Fluctuation analysis
Tremor
Amplitude Tremor Intensity Index(ATRI)
MDVP
Amplitude-Tremor Frequency(Fatr)
MDVP
F0-Tremor Frequncy(Fftr)
MDVP
Frequncy Tremor Intensity Index(FTRI)
MDVP
Voice turbulence Index(VTI)
MDVP

(c)


spectral-cepstral analysis
spectral analysis
Long-Term Average Spectrum(LTAS) characteristics
Linear Predictive Coding Coefficients(LPCC)
residual signal
Pitch amplitude
Spectral flatness ratio
Decomposition with filter-banks
Correlation functions
Mel-Frequency Cepstral Coefficients(MFCC)
Perceptual Linear Predictive Coding(PLP)
Time-Frequency Decomposition
wavelet transformation
Energy/Entropy of sub-band decompostions
Adaptive time-frequency distribution
cepstral analysis
Cepstral Peak Prominence(CPP)
ADSV
Low-to-High Harmonic Ratio(LHr)
ADSV
Regression Slope
ADSV
Cepstral Intensity
ADSV
Mean CPP F0
ADSV

(d)



(e)

Figure 2.10 Acoustic features Mind Map (a) overall acoustic analysis (b) Temporal and acoustic analysis (c) Perturbation and fluctuation analysis (c) Spectral-cepstral analysis (e) 2-Dimensional representations (MDVP note refers to the feature in the manual MDVP, ADSV note refers to the feature in the manual ADSV)

These acoustic features can be regarded into five types, including temporal acoustic analysis, perturbation and fluctuation analysis, spectral and cepstral analysis, 2D representations, and complexity analysis. Complexity analysis is also called non-linear analysis, which is another field in this research, without relationship with the voice perceptual traits. In this work, we do not expand research in this field. The other types of acoustic features will be reviewed in the following sections.

2.4.3 Temporal acoustic analysis

In Figure 2.9, voice conditions can be evaluated from both voice aspects and voice quality. In voice aspects, Speech Pressure Level (SPL) are used to represent loudness. Pitch and SPL are two common metrics [117, 121]. In order to identify modulated and additive noise in type II and type III signal, some research developed a method to track the vocal cord vibration based on pitch tracking and the related low-order statistics[14, 64, 84, 95, 122-126]. In [127], it proposed a system to distinguish normal voice and pathological voice by pitch tracking with respect to sex and age. SPL requires a sound pressure level meter(accelerometer), such as a pulsar nova[117]. Photograms are also used in voice pathology detection, which is the measurement of the sound pressure level at different frequency ranges [128, 129]. Some variation of photograms, based on the characterization of speech level range, were investigated in[130, 131].

Apart from pitch and loudness, voice breaks and unvoiced segments are also important acoustic features[132]. Degree of Voice Breaks (DVB), Number of Voice

Breaks (NVB), Degree of Voiceless (DUV), and Number of Unvoiced Segments (NUV) are all acoustic features that is recorded in Multi-dimensional Voice program (MDVP) manual. The threshold of DVB and DUV in normal voice is set to zero in sustained vowels. Experimental observation shows that functional dysphonia patients and patients with neurogenic voice disorders tend to have higher Degree of Subharmonics(DSH) and Number of Subharmonics(NSH)[133, 134].

2.4.4 Perturbation and fluctuation analysis

From a voice quality point of view, perturbation and fluctuation analysis is able to describe aperiodicity. Perturbation refers to temporal changes or unexpected disturbances. The most commonly seen perturbation parameters related to modulation noise are jitter and shimmer.

Jitter and shimmer have its variations. Jitter is perturbation of fundamental frequency, which is an evaluation of the period-to-period variability of the pitch period in speech signals. Period-to-period variability is caused by impairment of vocal cord which support periodic vibration. Generally, this type of variation is random, which relates with hoarseness. It has been generally used in acoustic analysis[14, 89, 95, 123, 125, 132, 135-145]. It is reported that normal voice exhibit jitter too, while pathological voice exhibit increased jitter[146, 147]. Mathematical representation of absolute jitter is defined as,

$$Jitt\_abs = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_{0i} - T_{o(i+1)}\right| \tag{2.1}$$

T_{0i} $(i = 1,2, \dots N)$ is the pitch period of the i-th frame and N is the number of extracted pitch periods(frames). As the fundamental frequency varies from frame to frame, the jitter value increases. This parameter is directly related with pitch period, so that correct detection of pitch is very essential for calculating this parameter. There are few variations of jitter as follows:

1) Jitter relative (jitt_rel): Relative evaluation of the period-to-period (very short-term) variability of the pitch within the analysed voice sample. Relative jitter is more suitable as a parameter than absolute jitter. However, relative jitter is affected by pitch extraction algorithm to a large extent. The mathematical representation is,

$$Jitt\_rel = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_{0i} - T_{o(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_{0i}} \tag{2.2}$$

2) Jitter Relative Average Perturbation (RAP): Relative evaluation of the period-to-period variability of the pitch within the analysed voice sample with smoothing factor of 3 periods. RAP reduces the sensitivity to the error of pitch detection compared to absolute jitter and jitter relative. Hoarse or breathy voices may have an increased RAP. The mathematical representation is,

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|\frac{T_{0(i-1)} + T_{0i} + T_{0(i+1)}}{3} - T_{0i}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_{0i}} \tag{2.3}$$

3) Jitter period perturbation quotient (PPQ5): Relative evaluation of the period-to-period variability of the pitch within the analyzed voice sample with a smoothing factor of 5 periods. The difference between PPQ and RAP is that PPQ uses 5 pitch period and RAP uses 3 pitch period.

$$PPQ5 = \frac{\frac{1}{N-4}\sum_{i=1}^{N-4}\left|\frac{\sum_{r=0}^{4} T_{0(i+r)} - T_{0(i+2)}}{5}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_{0i}} \tag{2.4}$$

Shimmer is evaluation of the period-to-period (very short-term) variability of the peak-to-peak amplitude within the analysed voice sample. It has been successfully applied in acoustic analysis[14, 89, 95, 123, 125, 132, 135, 137, 139-141, 143, 145]. Similarly, pitch extraction errors affect shimmer percent significantly. The absolute shimmer is computed from the extracted peak-to-peak amplitude data as:

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(A_{i+1}/A_i\right)\right| \tag{2.5}$$

A_i $(i = 1,2, \dots, N)$ is the extracted peak-to-peak amplitude data, and N is the number of extracted impulses. There are few variations of shimmer as follows:

1) Shimmer relative (shim_rel): It is relative evaluation of the period-to-period (very short-term) variability of the peak-to-peak amplitude within the analysed voice sample. It is computed from the extracted peak-to-peak amplitude data as:

$$Shim\_rel = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{2.6}$$

2) Amplitude perturbation quotient (APQ): It is relative evaluation of the period-period variability of the peak-to-peak amplitude within the analysed voice sample at smoothing of 11 periods. The mathematical algorithm is:

$$APQ = \frac{\frac{1}{N-10}\sum_{i=1}^{N-10}\left|\frac{\sum_{r=0}^{10} A_{(i+r)} - A_{(i+5)}}{11}\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \tag{2.7}$$

It is shown in [148] that although jitter and shimmer related features shows great power in analyzing type I signal, it is not as uesful in analysing type II and type III signal due to the requirement of accurately detecting pitch. Some methods using spectral techniques to calculate jitter and shimmer, in order to avoid pitch tracking algorithms [69, 91].

Other popular acoustic parameters discover the relationship of harmonics with noises in the speech corpus. For example, Harmonic-to-Noise Ratio (HNR) [89, 95, 139, 141, 149] is referred as the average ratio of the inharmonic spectral energy to the harmonic spectral engergy in the frequency range 70 to 4200 Hz. Increasing value of HNR reveals the increase of the spectral noise, which is caused by amplitude and frequency variations. Apart from HNR, there are some variations of it to describe the additive noise, such as Signal-to-Noise Ratio (SNR) [140, 141, 144, 150], Normalized Noise Energy (NNE) [14, 141, 151, 152], Glottal-to-Noise Excitation Ratio (GNE) [87, 90, 153] etc.

Fluctuation analysis is the research that focuses on the dynamics of the vocal cord tract, showing the instability of the tract. Tremor involves with rhythmic muscle movements, which causes the quavering of the voice. In the Multi-dimensional Voice Program (MDVP) manual, there are few parameters listed to represent tremor related features, including Frequency Tremor Intensity Index (FTRI), Amplitude Tremor Intensity Index (ATRI), Amplitude-Tremor Frequency (Fatr), F0-Tremor frequency (Fftr) [125, 154, 155], and Turbulent Noise Index (VTI) [79, 156]. The tremor analysis illustrates the amplitude modulation of the voice and the strongest periodic frequency. Fftr describes the rate of periodic tremor of the frequency, and Fatr describes the rate of change of the amplitude. When Fftr and Fatr surpass the low threshold of detection, it shows the potential of voice disorders. Magnitude of FTRI and ATRI interprets the rate of the amplitude and frequency tremors.

2.4.6 Spectral or Cepstral Analysis

Spectral and cepstral measurement methods have been generally employed in pathological voice detection. They show high correlation with dysphonia in classification tasks, with good sensitivity [80, 157]. Compared to temporal and perturbation features, spectral and cepstral analysis do not rely on pitch detection, and are also applicable on both sustained vowels and continuous speeches [80, 82].

There are spectral features that are directly extracted from the spectrum of the speech, such as Long-term Average Spectrum (LTAS) characteristics [68, 139, 158, 159]. Other features focus on the spectral envelope, such as Linear prediction coding coefficients (LPC) [157, 160-164]. They can be applied for decomposing the signal into residual signal part and a vocal tract signal part[164]. From the residual signal, studies explored characteristics such as pitch amplitude [68, 152, 158], which shows the main peak amplitudes in autocorrelation of the residual signal), and spectral flatness ratio[136, 152, 158], which shows the flatness of the residual signal part.

There are several ways of decomposing the speech signal into sub-bands using filter-banks. In [165], octave filter-banks are applied for pathological voice detection. Other than octave filter banks, filter banks with psychoacoustic features are employed to emphasise information in frequency bands which are in the range of human hearing system. For example, Perceptual Linear Predictive Coding (PLPC) [96, 166, 167], RASTA-PLP [70, 96, 162, 166], Mel-Frequency Cepstral Coefficients (MFCC) [70, 88, 97, 99, 103, 115, 162, 168-175], and its variations[74, 88, 93, 98, 102, 104, 167] have been investigated in many studies. Among them, MFCC is the most popular feature in speech field. The Mel-frequency band is shown in Figure 2.11. Although MFCC has been the most popular acoustic feature used, it has its own limitations. One is that additive noise affects the robust computation of MFCC[176], another one is that it is dependent on the range of the filter banks and the number of filters[177]. In [178], it was shown that when the number of filters are increased above certain threshold, it will affect the performance of the MFCC.

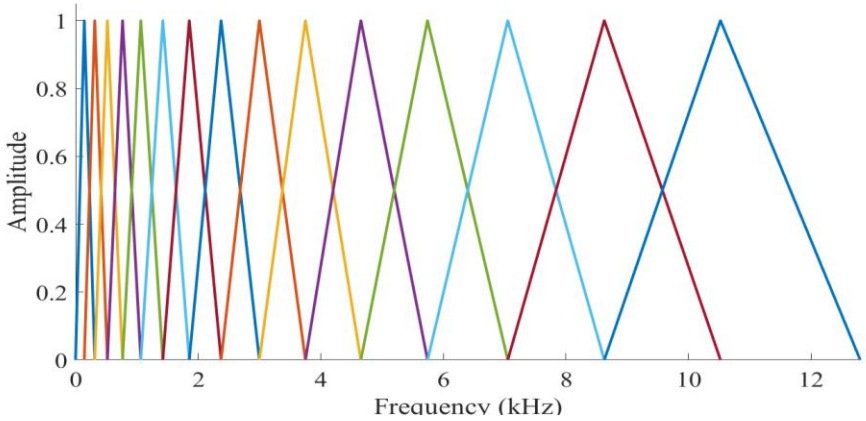


Figure 2.11 Mel-scale filter banks

Except for spectral features, cepstral features [80, 139, 140, 159, 179-181] based on cepstrum has also been explored in dysphonia detection, such as cepstral peak prominence (CPP). Cepstrum is the Fourier transform of the logarithm power spectrum of speech signal. In cepstrum, the dominant peak is called cepstral peak. It is shown that this is the most correlated acoustic feature with dysphonia in hoarseness characteristic compared to NHR or other perturbation features [80, 82, 182]. The cepstrum is computed by the following algorithm:

$$x_C(n) = Real(IDFT[\log|X(k)|]) \tag{2.11}$$

$Real$ refers to the real part of IDFT. $|X(k)|$ refers to the magnitude of the frequency domain of the Fourier transform of $x(n)$. $x_C(n)$ is notated as cepstrum. In another aspect, cepstrum refers to inverse Fourier transform of the Fourier transform of a signal, which is "double spectrum" of the signal. Cepstrum is an anagram of spectrum, and quefrency is an anagram of frequency. A cepstrum of guitar sound example sampled at 48 kHz is shown in Figure 2.12. The highest peak is at 109 on the quefrency band, which corresponds to frequency of $\frac{fs}{109}$, which is 440.367 Hz, representing the perceived pitch of note A4. The second largest peak is at quefrency 215, which corresponds to frequency of $\frac{fs}{215}$, which is 223.2558 Hz, representing the perceived pitch of note A3. The maximum peak is at 0, and the cepstrum near it corresponds to very high frequency noise, which near or beyond the edge of the audible spectrum.

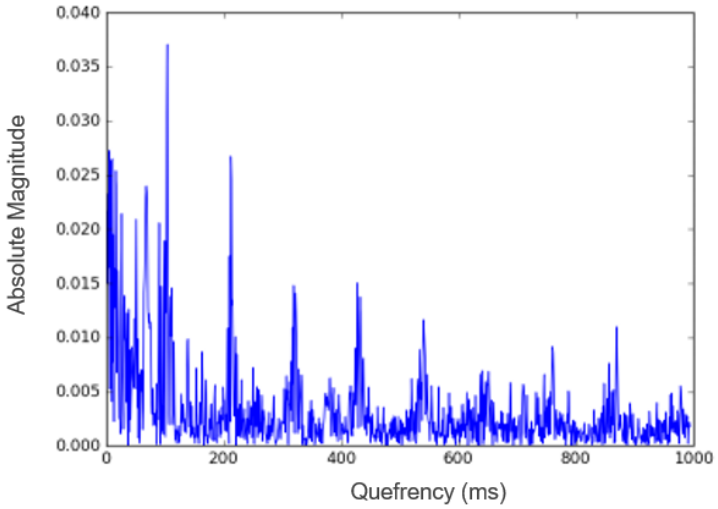


Figure 2.12 Cepstrum of a guitar sound

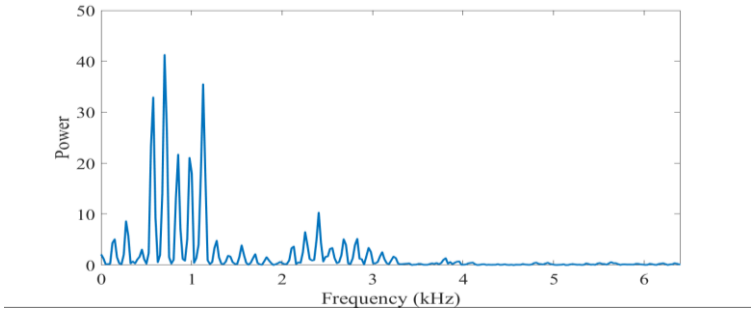
The pitch and formants can be easily shown in frequency band of spectrum. However, these components is hard to be distinguished in cepstrum quefrency band. The cepstrum quefrency band is index of time. For example, note A3 in Figure 2.12 correspond to $\frac{1}{f_{A3}}$ (f_{A3} is 223.2588Hz), which is 0.00448 seconds. The quefrency band is also different from time-domain. The quefrency gap between two samples is calculated as $\frac{1}{f_s}$, which is $\frac{1}{48000}$=0.00002083 seconds. In this case, the quefrency bin q of $\frac{1}{f_{A3}}$ is calculated as,

$$q = \frac{\frac{1}{f_{A3}}}{\frac{1}{f_s}} = \frac{f_s}{f_{A3}} \tag{2.12}$$

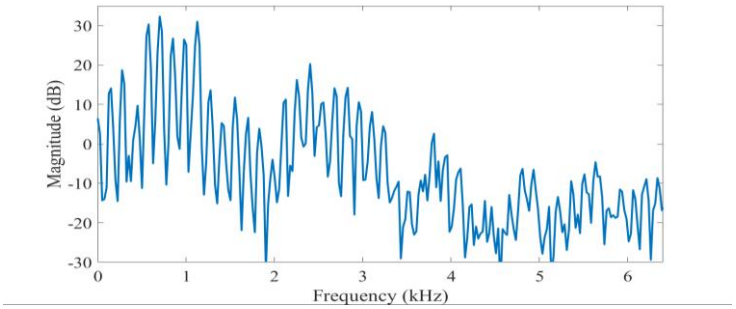
It represents that note A3 locates in $\frac{48000}{223.2588}$ = 215 quefrency bin.

Although MFCC is an important spectral feature in many speech applications, it also have strong relationship with cepstrum. Therefore, we regard it as a cepstral feature for analysis in Chapter 4. The logarithm of the spectrum approximates roughly the sensitivity of the ear. Therefore, logarithmic spectrum are used to evaluate the auditory features. The power spectrum and a logarithm-spectrum of a speech segment are shown in (a) and (b) of Figure 2.13. From the logarithm-spectrum, the periodic structure reveals the harmonic components (formants) caused by the fundamental frequency. This represents the resonances of the vocal tract. One way to assessing this periodic structures is to use the Fourier Transform. In this case,

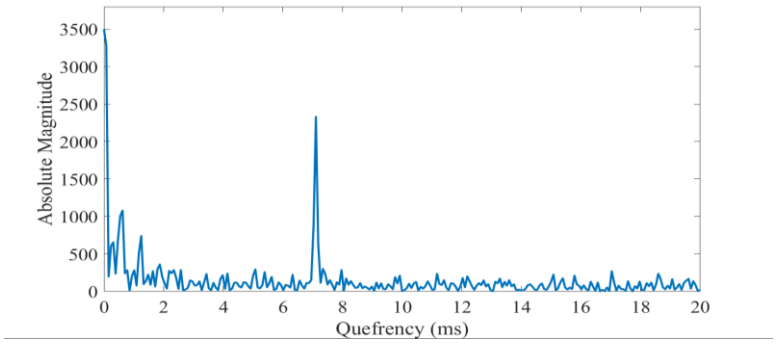
Discrete Fourier Transform (DFT) or Discrete Cosine Transform (DCT) is conducted on the logarithm-spectrum for the evaluation, which is the cepstrum. The cepstrum of the speech segment example is shown in Figure 2.13 (c). In the cepstrum, the low-quefrency end of the cepstrum shows the formant information.



(a)



(b)



(c)

Figure 2.13 (a) power spectrum (b) logarithm spectrum (c) cepstrum of a speech segment

As indicated in Figure 2.11, Mel-scale are introduced for the speech signal with regards to the auditory perception sensitivity to difference frequency scales. For the specific frequencies, it models the human auditory system. To avoid losing all other information except for these specific frequencies, the weighted sum of energies near the target frequency in the filter banks are calculated. Finally, the discrete cosine

transform is conducted on this robust estimate corresponding for human auditory system, which produce the representation known as mel-frequency cepstral coefficients (MFCC). Therefore, although there are variations including logarithm, mel-scale filter, and DCT, MFCC is still a cepstral feature essentially.

In Chapter 4, we explore MFCC and cepstral features based on the cepstrum for pathological voice detection, using deep recurrent models to process these short-time-frame based features.

2.4.7 2D representations

2D representations are a type of acoustic features, which have been explored in recent years. One method is based on modulation spectra (MS), which shows the modulation and frequency component of the speech signal. It is a 2D representation that has been investigated to employ in pathological voice detection [77, 169, 183]. Spectrogram and the variations has also been explored in this field [107]. Spectrogram is a "counter map" which is formed by stacking a series of spectrum amplitude in time sequence [184]. Time sequence is the horizontal axis, and frequency is on the vertical axis. Generally, it is a grey-scale time-frequency representation. While in some research, it is transformed into RGB 3-channel images.

Different time frame length will lead to different kinds of spectrogram [185]. If the time section is short (about 3ms), the spectrogram emphasize on the temporal changes in the signal, so the time-resolution is high while the frequency resolution is low. This is wide-band spectral analysis [186], which will benefit the analysis of characteristics of the source (vocal cord), such as pitch detection. Otherwise, when time section is long (about 20ms), the spectrogram emphasize on the frequency changes in the signal. This is narrow-band spectral analysis [186], and is convenient for analysing characteristics of the vocal tract filter. It highlights the vocal tract resonances (formants) and shows how it continues to vibrate after the source (vocal cord) vibration passed through. In our work, spectral and cepstral characteristics of the speech is emphasized, and wide-band spectrogram is applied.

There are other 2D representations that are formed with time-frequency decomposition techniques, and features are extracted using multi-directional regression[101] or interlaced derivative patterns of the speech signals[78]. Energy

and entropy of the wavelets from Discrete Wavelet Transform (DWT) are generally used in analysing aperiodic signals and signal with additive noises for pathological voice detection [83, 187]. The processing flow of 3-level DWT is shown in Figure 2.14. The output of DWT are coefficients, which are function of scale/frequency and time. It can be seen that DWT process is equivalent to comparing a signal with discrete multi-rate filter banks, similar to the decomposition techniques in MFCC in Section 2.4.6.

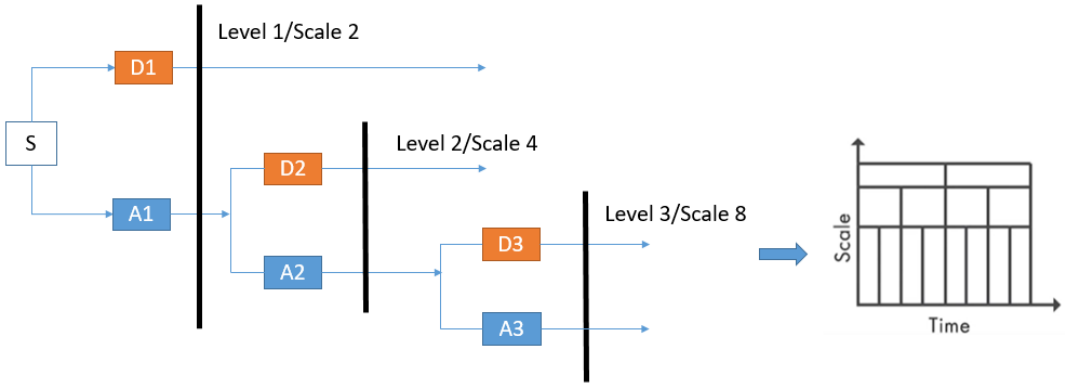


Figure 2.14 The processing flow of 3 – level DWT

DWT is a sampled version of the CWT and is used extensively in signal denoising, compression and enhancement [83, 187, 188]. Compared to DWT, Continuous Wavelet Transform (CWT) can obtain a simultaneous time-frequency analysis of a signal, which is applied in our work in Chapter 5 and Chapter 6. CWT can be used to study frequency breaks, time discontinuity, signal bursts, signal's damping, vibration pattern etc., and it is suitable to be used as feature map input to a DNN for signal classifications. DWT scale a wavelet by a factor of 2, and the corresponding equivalent frequency is shown in Table 2.1. It can be seen that the wavelet scale reduces the equivalent frequency by an octave. CWT have the added flexibility to analyse the signal at intermediary scales within each octave, which allows for fine-scale analysis. The parameters in Table 2.2 is referred as the number of scales per octave. The higher the number of scales per octave, the finer the scale discretization. Typical values for this parameter are 10, 12, 16, or 32. In Chapter 6, we use 16 per octave.

Table 2.1 Equivalent frequency with wavelet scale on DWT

Wavelet Scale	2^1	2^2	2^3	2^4
Equivalent Frequency(F_{EQ})	$\frac{F_{EQ}}{2^1}$	$\frac{F_{EQ}}{2^2}$	$\frac{F_{EQ}}{2^3}$	$\frac{F_{EQ}}{2^4}$

Table 2.2 An example of CWT scale transformed from DWT octave scale

Scale: 2^1	$2^{\frac{11}{10}}$	$2^{\frac{12}{10}}$	$2^{\frac{13}{10}}$	$2^{\frac{14}{10}}$	$2^{\frac{15}{10}}$	$2^{\frac{16}{10}}$	$2^{\frac{17}{10}}$	$2^{\frac{18}{10}}$	$2^{\frac{19}{10}}$	Scale:2^2

A CWT of a signal $f(t)$ is given by the equation as it shown as

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) \cdot \Psi^* \left(\frac{t-b}{a}\right) dt \tag{2.13}$$

,where a is the scale (dilation), and b is the position (translation parameter) of the wavelet. In our work in Chapter 6, we use default analytic Morse wavelet [189-192] in Matlab, The Fourier transform of the generalized Morse wavelet is

$$\Psi_{P,\gamma}(\omega) = U(\omega) a_{P,\gamma} \omega^{\frac{P^2}{\gamma}} e^{-\omega^{\gamma}} \tag{2.14}$$

Where $U(\omega)$ is the unit step, $a_{P,\gamma}$ is a normalizing constant, P^2 is the time-bandwidth product, and γ controls the symmetry of the Morse wavelet in time. By adjusting the square root of the time-bandwidth product P (proportional to the wavelet duration in time) or symmetry controller γ, the wavelet obtains different characteristics and behaviour. We use the default symmetry parameter (gamma) as 3, and time-bandwidth product equal as 60. The frequency response decays to 50% of the peak magnitude at the Nyquist.

Each scaled wavelet is shifted in time along the entire length of the signal, and it forms the corresponding wavelet coefficients called “Scalogram”. Spectrogram uses a constant window so that the time resolution and frequency resolution is fixed. Compared to that, CWT uses “wavelets” to window the signal, which scale through the time. An example of scalogram in the onset of speech are shown in Figure 2.15. The scaling operation shrinks and stretches the prototype wavelet. When scale a is low, the wavelet is compressed and the frequency is high, so it gets short-duration with fast variations in the coefficients. It can be seen in Figure 2.15 that it captures

the fast changing details of the signal. When scale a is high, the wavelet is stretched and the frequency is low, and it gets long-duration, so it is capable of capturing the low varying detail of the signal. Compared to DWT, CWT is shift-invariant. A simple shift in a signal will not cause realignment of signal energy in CWT.

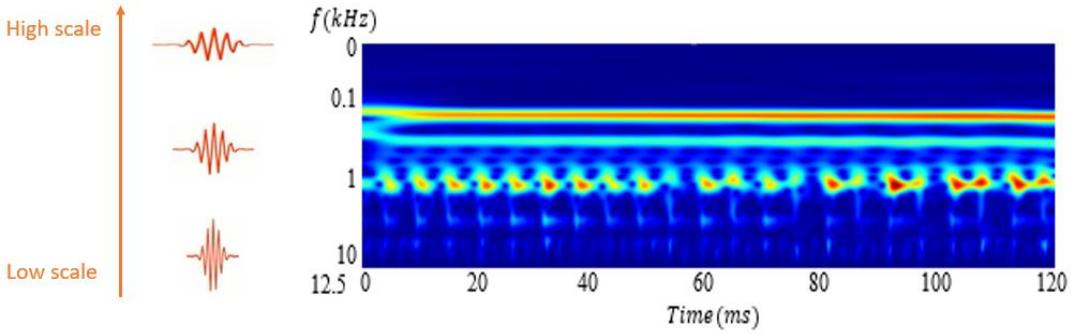


Figure 2.15 Example of scalogram from 0 to 0.12s (Normal voice sample in PdA database: Mtaaa2.wav)

Compared to spectrogram, scalogram is superior in detecting abrupt changes and localize the frequency information in a signal. Thus it is helpful for detecting the potential pathological characteristics in the onset of the speech.

2.5 Datasets

In order to reduce background noise sources, specific principles and policy are set for data collection process, such as keeping the similar acoustic recording instruments conditions (microphone standards etc.) and data storage method (sampling frequency etc.). In [193], it listed several acoustic speech analysis data collection suggestions. It suggests a professional condenser microphone with the minimum sensitivity of -60dB should be used. The distance of mouth to microphone would be less than 10 centimetres, and the sampling frequency is around 10 kHz to 100 kHz. The recording environment should be sound-proof room, with environmental noise lower than 50dB. The recorded corpuses should be able to contain all the possible variables in different classes, including different age range or balanced gender distribution.

There are two types of speech corpuses for acoustic analysis: sustained vowel and continuous speech. Sustained vowels contains more advantages[68, 158] as follows:

1) It is easier to record sustained vowels.

2) It is much less affected by dialects, tones, speed, or speech characteristics by speakers from different countries or regions.

3) The structure of sustained vowels are simple, which leads to better consistency in voice quality measurement.

In this case, sustained vowels are preferred to be used in acoustic analysis research. Open vowels such as /a/ are more popular than other vowels such as /i/ or /u/, because open vowels requires the usage of entire vocal tract[194].

However, there are some research argues that continuous speech have advantages than sustained vowels in some other aspects[179, 182]:

1) It helps the exploration of dynamic acoustic characteristics.

2) Continuous speech is commonly seen in realistic environment.

3) Some acoustic characteristics appears more frequently in continuous speech than it in sustained vowels in some speeches such as spasmodic dysphonia.

In this section, we introduce four databases commonly used in this field: MEEI database [195], SVD database [196], PdA database [119], and AVPD database [96].

2.5.1 Massachusetts Eye & Ear Infirmary (MEEI) Database

The popular MEEI database [195] comprises a number of public corpuses. It is a widely used resource for speech and language therapy studies. It contains over 1400 sustained vowel /a/ and the first sentence of "rainbow passage" text. This database was recorded using different sampling frequencies (10 kHz, 25 kHz or 50 kHz). The study in [152] with MEEI database has been the standard partition that many other works compares onto. Although this database is popular, it has a major problem that limits it use. The healthy speech and pathological speech were recorded in different environments [172, 197]. Consequently, it is very hard to tell whether the classifier is classifying pathological features or environmental information. In addition, the database is gender unbalanced, and the two classes are unbalanced too. The significant pathologies in this database is shown in Table 2.3. There are 657 pathological speakers with different types of pathologies and 53 healthy speakers, so that unbalanced classification problem may appear when doing pathological voice detection.

Table 2.3 Significant pathologies in MEEI database

Pathology Type	Amount
Healthy voice (Control)	53
Hyper-functional dysphonia	267

A-P squeezing	167
Ventricular compression	110
Paralysis	67
Gastric reflux	48

2.5.2 Saarbrucken Voice Database (SVD)

Saarbrucken Voice Database (SVD) [196] database is a free downloadable database online, with voice recordings and electroglottography (EGG) from more than 2000 persons. This database contains 71 different pathologies with recordings from 687 healthy people (428 females and 259 males) and 1356 patients (727 females and 629 males). Each individuals' recording session contains the following recordings:

- Sustained vowels /a/, /i/, and /u/ with different intonations including normal, low, high and low-high-low
- Sentence "Guten Morgen, wie geht es linen?" ("Good morning, how are you?")

All the voice recordings are sampled at 50 kHz with 16-bit resolution, and all of them are recorded in the same environment. The significant pathologies are listed in Table 2.4.

Table 2.4 Significant pathologies in SVD database

Pathology Type	Amount
Healthy voice (Control)	687
Functional dysphonia	112
Hyper-functional dysphonia	187
Hypo-functional dysphonia	16
Psychogenic dysphonia	91
Bulbar paralysis	2
Contact pachyderm	71
Cyst	6
Fibro sarcoma	2
Granuloma	2
Laryngeal cancer	1
Laryngitis	140
Larynx tumour	5

Leucoplakia	41
Median neck cyst	1
Nasosinusitis	18
Papilloma	1
Phonation nodule	17
Recurrent laryngeal nerve paralysis	213
Reinke's edema	68
Spasmodic dysphonia	64
Vocal fold carcinoma	22
Vocal fold polyps	45

Since all data are recorded in the same environment, SVD database is a much more reliable database compared to MEEI database. In addition, the healthy data and pathological data have similar amount, resulting in no unbalanced classification problems.

2.5.3 Principe de Asturias Database (PdA)

Principe de Asturias Database (PdA) [119] is a database recorded by the Unversidad Politecnica de Madrid (UPM). This database contains sustained vowel /a/ from 200 pathological speakers covering a variety of organic pathologies (eg: nodules, polyps, edemas, and carcinomas etc.), and 239 healthy speakers. All voice recordings in this database are sampled at 25 kHz with 16 bit resolution, and all of them are recorded in the same environment. This database also contains the perceptual analysis metric "GRBAS" (Grade, roughness, asthenia, breathiness, strain) for each individual speaker, which might be helpful to guide the severity of dysphonia for future research. The significant pathologies are listed in Table 2.5.

Table 2.5 Significant pathologies listed in PdA database

Pathology Type	Amount
Healthy voice (Control)	239
Bilateral vocal cord nodule	29
Bilateral Reinke's edema	29
Pediculate vocal cord polyp	28
Sulcus	22
Epidermoide cyst	20

2.5.4 Arabic Voice Pathology Database (AVPD)

AVPD database [96] was recorded at the Communication and Swallowing Disorders Unit of King Abdul Aziz University Hospital, Riyadh, Saudi Arabia in a sound-treated room using a Kay C (CSL) utilizing MDVP software. All the normal and pathological voice samples were diagnosed by laryngeal stroboscope as a tool for clinical assessment. This database also contains the information of severity of dysphonia for each sample based on consensus of a panel of three experts. This perceptual score was rated on a scale of 1 to 3, where 1 represents mild, 2 represents moderate, and 3 represents severe voice quality disorder. This database includes the following recording task:

- Three sustained vowels /a/, /e/ and /o/ with onset and offset information
- Arabic reading of counting from 0 to 10, and three common words
- Standardized Arabic passage

All the voice recordings were sampled at 48 kHz with a bit depth of 16 bits, and recorded in the same environment. There are overall 5 types of pathologies in this database, and the healthy-pathological data amount is balanced. The data distribution is shown in Table 2.6.

Table 2.6 Data distribution in AVPD database

Pathology Type	Amount
Healthy voice (Control)	175
Vocal cord paralysis	51
Sulcus	43
Vocal cord polyp	39
Vocal cord cyst	25
Vocal cord nodules	20

It can be seen that compared to MEEI database, SVD, PdA and AVPD databases are more reliable in healthy-pathological data balance and consistent recording environment. In this case, SVD, PdA and AVPD databases are adopted in this work.

2.6 Conclusion

In this chapter, we introduced the causes, characteristics and symptoms of various types of dysphonia. A review of both perceptual and acoustic analysis of dysphonia were also been illustrated with current state-of-the-art techniques. From perceptual point of view, voice quality metrics such as breathiness, roughness, strain, are represented by GRBAS and CAPE-V. In acoustic point of view, these conditions can be evaluated from acoustic features based on signal processing techniques. These short-time acoustic features has been generally employed in pathological voice detection, including temporal acoustic analysis, perturbation and fluctuation analysis, spectral or cepstral analysis and 2D representations. Cepstral and spectral analysis will be explored in Chapter 4, and 2D time-frequency representations will be explored in Chapter 5 and Chapter 6.

Chapter 3

3 Machine Learning based classification techniques for pathological voice detection

3.1 Introduction

In Chapter 2, acoustic analysis and feature extraction of speech signals in pathological voice detection were reviewed. In this chapter, the main machine learning based classification techniques are reviewed. An overview of processing flow of the pathological voice detection is described first. Then a literature review table presents the state-of-the-art studies using the three databases described in Chapter 2. After the overview of techniques, a review of supervised learning and unsupervised learning techniques for pathological voice detection is presented in section 3.3 and section 3.4 respectively. In supervised classification problem, traditional machine learning techniques such as Support vector machine, Gaussian mixture model, k-Nearest Neighbours are described. Deep learning in the field of Pathological Voice Detection (PVD) has been increasing in popularity in recent years. In section 3.3.2, the basic structure of classical DNN models will be described, and the review of its variations with its applications in speech processing field is presented.

3.2 Review of techniques in pathological voice detection

In general, acoustic analysis based pathological voice detection requires manual feature extraction, pattern recognition and classification. The basic processing structure is shown in Figure 3.1. After pre-processing and feature extraction with acoustic analysis, dimensionality reduction is conducted to select essential relevant features for classification. Then the feature sets are fed into machine learning system to make classification decisions.

In this section, an overview of the processing flow of pattern recognition with state-of-the-art technologies is presented in detail in the first part. In the second part, the overview of literature review based on these technologies are presented as a table.

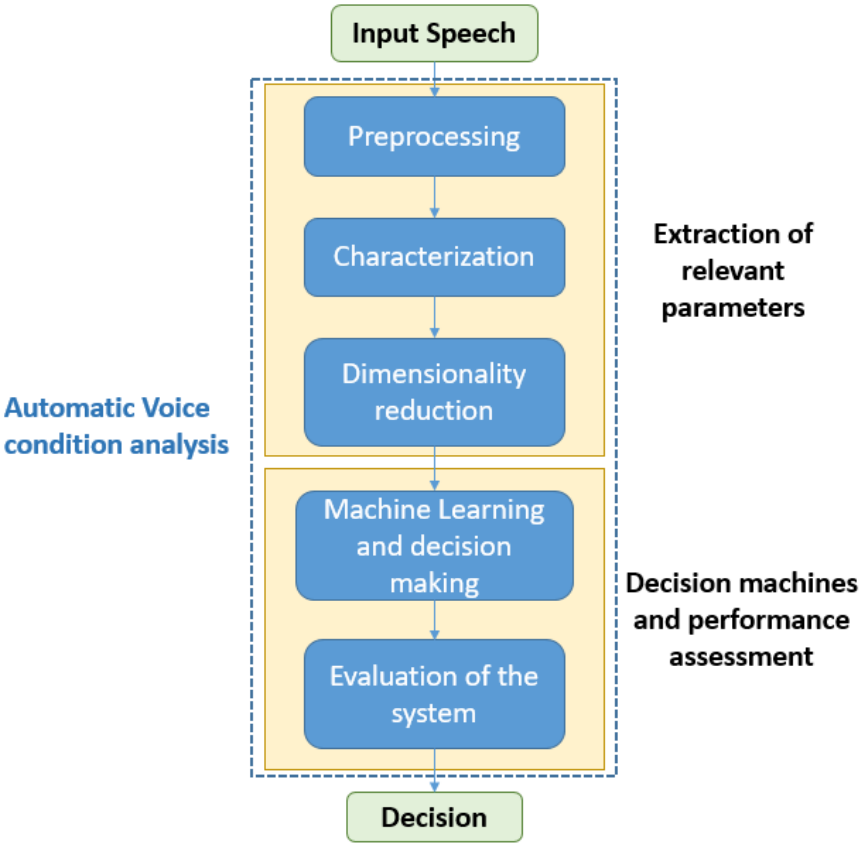


Figure 3.1 Pattern recognition processing flow

3.2.1 Overview of processing flow in pathological voice detection

The pathological voice detection process flow is similar to the basic pattern recognition processing flow as shown in Figure 3.1. In this section, the overview of processing flow in pathological voice detection is presented in Figure 3.2. It can be seen from Figure 3.2 that there are two main branches of processing flow in pathological voice detection. One processing flow is shown in red line. This traditional processing flow is to extract traditional acoustic features such as temporal features, perturbation, fluctuation features, and spectral/cepstral features. In Chapter 2, pre-processing and feature characterization were reviewed. Then unsupervised dimensionality reduction techniques are applied to select the most correlated features.

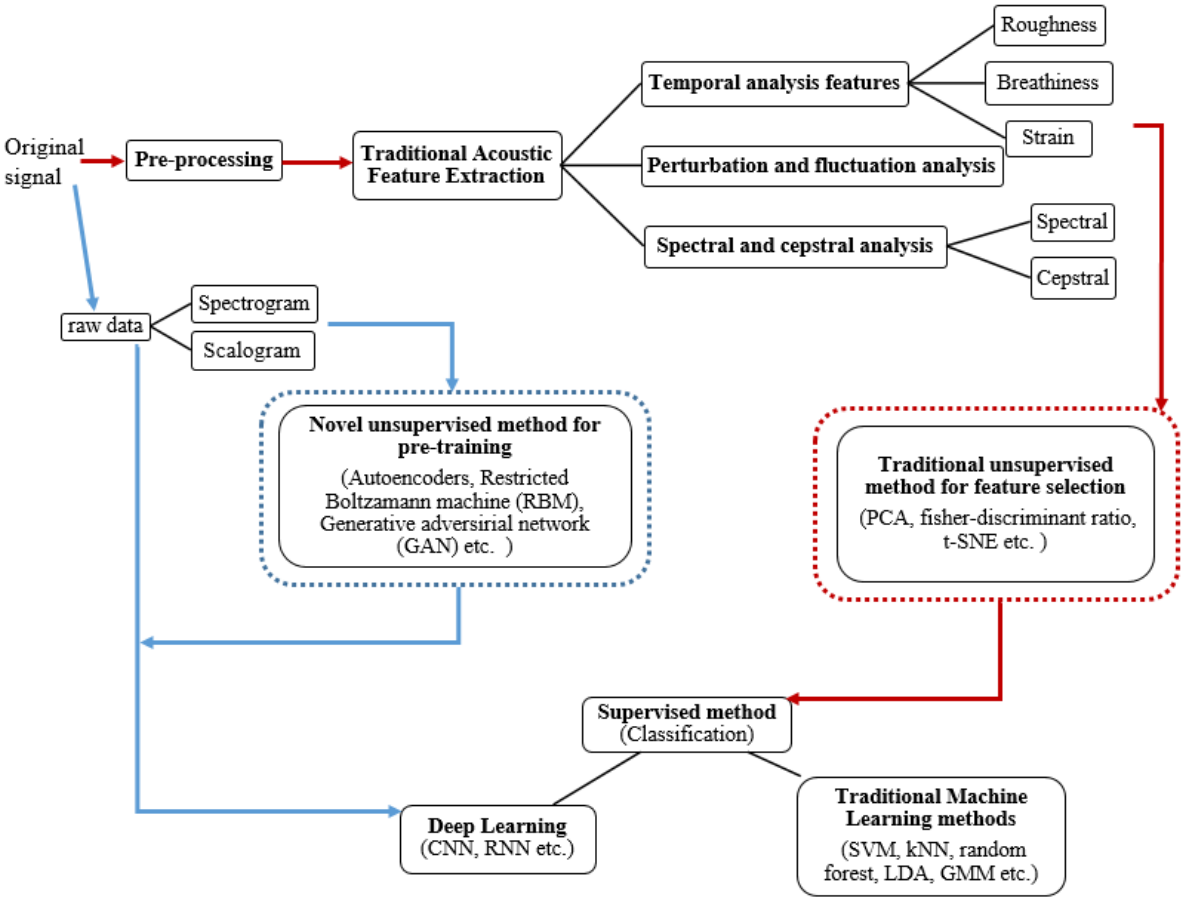


Figure 3.2 Overview of processing flow in pathological voice detection

Dimensionality reduction focuses on reducing the feature space, to reduce unrelated redundant features in the feature set map. It transform the original parameter set to another high-dimensionality space, which is a statistical way to discover the inner rule of data. Some dimensionality reduction technique transforms the original feature space to a new space. These axes in the new feature space are non-correlated. This method can be more efficient, while it is hard to interpret the parameters in the new feature space. This technique include Linear Discriminant Analysis (LDA) [83, 109, 160, 187, 198-200], Singular Value Decomposition (SVD) [110], principle component analysis (PCA)[83, 84, 87, 109, 164, 171, 179, 198, 200-205], and variations of PCA (kernel PCA[202], neural-networks PCA[204]). PCA is an unsupervised learning linear method. The main idea is to make the variance of the data on each coordinate axis to be the largest in the new feature space, thus it orthogonalize and de-correlates the former feature space. There are also some methods for transformation based on Hidden Markov Model (HMM) [205] and clustering thoughts[109, 110].

Other works uses feature selection to reduce the complexity of the feature space, which remove the features with high-correlation/similarity. There are three methods in feature selection. First is to use wrapper feature selection with regards to the performance of the classifier/regressor. There are some examples using this method to select features such as genetic algorithm [171, 199, 206-208], and binary logistic regression analysis[130]. The second method is to use filtering feature selection such as Fisher discriminant ratio (FDR) [100, 104, 126, 200, 209, 210], Fisher discriminant analysis[85], correlation analysis[87] and mutual information analysis[211]. The last method is to using embedded feature selection, to conduct feature selection in classification process.

After feature extraction and dimensionality reduction process, machine learning techniques are applied to make system decisions. Traditional machine learning methods have been generally applied in pathological voice detection, such as Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). SVM is built by a variety of kernel functions, and has been the most popular classification method in application [1, 70, 74, 77, 78, 83, 89, 93, 94, 100, 124, 126, 137, 160, 164-166, 169, 170, 198, 199, 206-210, 212-218]. GMM is based on probabilistic models, which is a commonly used method in speech processing applications. In this field, it also

achieves good performance in classification [72, 93, 99, 101, 103, 104, 111, 157, 161, 162, 168-170]. Hidden Markov model (HMM) is also a basic probabilistic graphical model, which has been applied in [73, 99, 111, 113, 137, 162, 205, 217, 219]. Among them, SVM is the most popular model for small-data classification, and it works well on subsets of a database. However, GMM works better on large-data classification applications compared to SVM.

Other traditional machine learning techniques have also been explored, such as k-Nearest Neighbours (KNN) [75, 94, 175, 207], Artificial neural networks (ANN)[71, 83, 97, 114, 116, 140, 171, 187, 198, 199, 220], random forests[1, 141, 218, 221], LDA[68, 79, 95, 111, 125, 152, 170, 175, 179, 222-224], and its variation quadratic-LDA[225].

The second processing flow branch is shown as the blue line. Deep learning is a branch of machine learning technique which is originated from ANN. It is a revolutionary End-to-End method, which extracts and selects features automatically. It emerged in 1990s, while explodes after 2010, and has just been applied in pathological voice detection field in recent years. The raw speech data is transformed into 2D representations such as spectrogram or scalogram. Unsupervised generative models such as Restricted Boltzmann Machine (RBM)/convolutional RBM is used as a pre-training technique. Convolutional neural networks (CNN) is employed as an End-to-End system for analysing 2D representations in our work and the work [107, 226, 227]. In the work of Harar et al. [11, 113], CNN and Long short-term memory (LSTM) are also applied on raw speech data, this is also the first time deep learning has been applied in this field. Auto-encoder is an unsupervised reconstruction method for extracting features in the bottleneck. We explored it too, while the result is not to our satisfaction. It is suggested that auto-encoder only uses raw speech data as training information.

3.2.2 Related work on PVD

There are literatures employing a variety of public and private datasets for pathological voice detection. As shown in section 2.5, MEEI database record pathological and normal voice data in two different environments, so that it was not considered in this study. We adopt three accessible public databases on hand: SVD,

Table 3.1 List of related works on pathological voice detection

Refs[1].	Research Group Institute	Year	Database	Data amount	Features	Classifier	Accuracy
[228]	University of Crete	2009	PdA	100N – 100P	MS	SVDM	82.70%
[169]	Polytechnic university of Madrid	2011	PdA	199N – 200P	MFCC, MS	GMM/SVM	81.70%
[106]	University of Zaragoza	2012	SVD	650N – 1320P	MFCC, HNR, NNE, GNE	GMM	67.00%
[229]	King Saud University	2014	SVD	100N – 100P	Peak value and lag for every frequency band	GMM, SVM	72%
[1]	Brno University of Technology	2015	PdA	239N – 200P	92 speech features2 with feature selection method Mann-Whitney U test	SVM/RF	82.1 ± 3.3%
[70]	University of Antioquia	2015	PdA	199N – 200P	Noise measures (HNR, VTI, SPI, NNE, GNE) / periodicity of voice (shimmer, jitter, PPQ, APQ) / spectral-cepstral features (MFCC, LPC etc.) / non-linear features	SVM	80 ± 8% / 82 ± 11% / 78 ± 7% / 82 ± 6%
[230]	University of Tunis El-Manar	2015	SVD	50N – 70P	MFCC, 1st and 2nd derivative	LDA, SVM	86.44%
[231]	University of Tunis El-Manar	2016	SVD	50N – 70P	MFCC, 1st and 2nd derivative	LDA, ANN	87.82%
[102]	Badji Mokhtar University	2016	SVD	60N – 40S[3]	MFCC	GMM-SVM	96.5%
[168]	King Saud University	2016	SVD / AVPD	60N – 101P / 87N – 78P	MFCC	GMM	80.20% / 83.65%
[126]	Kind Saud University	2016	SVD / AVPD	262N – 244P / 118N – 75P	MDVP features	FDR	99.68% / 72.53%
[165]	King Saud University	2016	SVD / AVPD	266N – 263P / 169N – 127P	cross-correlation	SVM	90.98% / 91.17%
[11]	Brno University of Technology	2017	SVD	687N – 1356P	raw waveform	LSTM, DNN	68.08%
[232]	Ecole Nationale Polytechnique d'Alger	2017	SVD	40N – 40S – 40P[4]	MFCC, jitter, shimmer, pitch	NBN	80% / 90%
[233]	Tsinghua University	2018	SVD	686N – 40C / 686N – 29F[5]	jitter, shimmer and 22 variation of them	proposed linear model + KNN	85.86 ± 6.5% / 94.46 ± 5.9%
[234]	Polytechnic University of Madrid	2018	PdA / SVD	197N – 169P / 568N – 970P	perturbation features, spectral/cepstral features, non-linear features, MS	GMM	76.61± 4.3% / 75.42 ± 2.19%
[113]	Brno University of Technology	2018	MEEI+SVD +PdA+AVPD	8042 chunks[6]	MFCC, MDVP features	XGBoost (CNN)	73.3%
[227]	University of Strathclyde	2018	SVD	482N – 482P	spectrogram	CNN / CDBN+CNN	77% / 71%
[107]	King Saud University	2018	SVD	686N – 743P	octave-spectrogram	AlexNet (CNN)	98.5%
[235]	Georgia Institute of Technology	2019	PdA	239N – 200P	MFCC	CNN + SVM / AE + SVM	70.1% / 71.8%

[1] Refs.: reference, PdA: Principe de Asturias Database, N – P: Normal voice recordings to pathological voice recordings, MS: Modulation spectra, SVDM: singular value decomposition method, MFCC: Mel-Frequency cepstral coefficients, GMM: Gaussian mixture model, SVM: Support vector machine, SVD: Saarbrucken voice database, HNR: harmonic-to-noise ratio, NNE: normalized noise energy, GNE: glottal-to-noise excitation ratio, RF: random forest, VTI: voice turbulence index, SPI: soft phonation index, PPQ: pitch perturbation quotient, APQ: amplitude perturbation quotient, LPC: linear prediction coefficients, LDA: linear discriminative analysis, AVPD: Arabic voice pathology database, FDR: fish discriminative ratio, LSTM: long short-term memory, NBN: Naïve Bayes networks, KNN: k-nearest neighbor, MEEI: Massachusetts eye and ear infirmary, CNN: convolutional neural network, CDBN: convolutional deep belief network, AE: autoencoder

[2] This work reveals sufficient pathological speech parameterization including MDVP features, spectral-cepstral features, wavelets, non-linear dynamic features and some new proposed features

[3] 60 normophonic voice – 40 spasmodic dysphonia voice

[4] 40 normophonic voice – 40 spasmodic dysphonia voice – 40 vocal fold paralysis

[5] 686 normophonic voice – 40 cordectomy voice / 686 normophonic voice – 29 front-lateral voice

[6] These chunks are 0.750s length splits of voice recordings from fusion of 4 databases

PdA, and AVPD database. Relevant research with these public databases is summarized in Table 3-1. These three databases are widely used as benchmark databases with good quality data. From the literature review, we conclude the following three phenomenon.

A. Small Subset VS. Large Fusion Database

As shown in Table 3.1, a number of works select only small subset of databases as reported in [102, 168, 230-233]. These works used around 50 to 80 speech data, and achieved high accuracy using traditional methods such as SVM, GMM or ANN.

However, when applying the similar techniques (SVM, GMM, DNN) on comparatively larger database [11, 106, 113, 227, 229, 234], it is obvious that fusion of large datasets appears much more challenging. This shows the difficulty of generalizing the model on large database.

Apart from this, most of experiments use a balance quantity of the normal voice data and pathological voice data. However, one study [233] included a significantly higher number of normal voices (n=686) than pathological samples (n=29). Therefore, although this work achieves high accuracy, it contains the imbalance classification problem which is a high risk to the application.

Small data size is the most common challenge in the bio-medical research. There are generally two ways of reducing small data size problems, one is transfer learning, and another is data augmentation.

Transfer learning[236] is developing to a research problem that focuses on storing knowledge obtained from solving one specific application and applying it to another application. In other words, we can train model A on dataset A and expect model A as a pre-trained model for unseen data in dataset B. In this case, non-sufficient labelled data in domain B is able to train a more robust and reliable model B. From [237], a sufficient transfer learning survey gives detailed definition of transfer learning with binary document classification example. Given domain, $A = \{X, P(X)\}$ (X is the feature space; $P(X)$ is the marginal probability distribution on the feature space), task $B = \{Y, P(Y|X)\}$ (Y is the label space; $P(Y|X)$ is the

conditional probability distribution) will learn from training data from feature space X and Y.

Best way of dealing small-data-size problem is to get more data. Therefore, data augmentation is capable of addressing it by creating fake data for training. In biomedical image field such as pulmonary nodule detection or brain tumour detection, translation and rotation are commonly used techniques for data augmentation [238-240]. Other operations such as cropping, zooming, and changing the brightness or contrast are also proved to be effective [241-243]. There are also some transformation which is too complex to be implemented such as out-of-plane rotation.

Noise injection in the input to a neural network have also been applied and seen as an operation of data augmentation [244]. Some research uses this in unsupervised learning applications such as de-noising autoencoder [245]. In [246], noise injection is applied in the hidden layers, and it verifies that this can work effectively when the magnitude of the noise is carefully tuned. Dropout [247] can also be seen as one of this technique because it is equivalent to multiplying noise to hidden layer for creating new input to the next layer.

It is said in [248] that hand-designed data augmentation methods usually dramatically reduces the generalization error of a machine learning/deep learning model. In chapter 6, we show the strong power of data augmentation in solving the dataset size problem.

B. Weak Labels

There are many challenges in the pathological voice detection field. The most significant one is the documented overlap between normophonic voice and pathological voice samples [234]. This is of particular importance in the absence of physiological changes, classifying a sample as pathological is based on subjective views of expert and can thus lead to inaccurate subjective labels. In this case, these "weak" labels can confuse DNN classifiers during training process and limit the performance. This issue will be described in more detail later in the thesis.

C. Black-box Dilemma

Compared to traditional acoustic features, the DNN system extracts the features automatically, so that interpretability is limited. The extracted features of DNN models are largely dependent on given knowledge – labels and the characteristics of speech data. In this case, the database affects the results heavily. Success on a specific database might not guarantee the same on another. The Deep Learning (DL) approach increases the robustness of the classification system and is good for generalization on large datasets. It has been shown to be very effective in both image classification and audio classification in recent years, while it was introduced in pathological voice detection in 2018 by Harar et.al [47], with 68% testing accuracy on Saarbrucken voice database.

3.3 Supervised learning techniques for pathological voice detection

In this section, we review the supervised learning techniques for pathological voice detection. In the first part of this section, we review some traditional machine learning techniques that have been generally applied in this field, including support vector machine, hidden Markov model, Gaussian mixture model, k-Nearest Neighbours, and shallow artificial neural network (multi-layer perceptron [249]). These methods are generally employed in "red" processing flow in Figure 3.2, with traditional acoustic features extracted for classification. However, this method is only showing good performance on small subset of database as described in Section 3.2.2.A. When processing large datasets (even entire database), end-to-end deep learning method (the blue processing flow in Figure 3.2) have better generalization ability on it. In this second part of this section, a review of related deep learning techniques is described with three topics in detail. CNN shows superiority in image detection field, while RNN has the ability of processing sequence data such as speech. Attention mechanism is rising to be the focus of deep learning research field in recent years. It performs as a "feature selection" tool which automatically learns the attention of weights in the parameters.

3.3.1 Review of Traditional Machine Learning methods for pathological voice detection

Over the years, traditional supervised learning techniques such as support vector machine and probabilistic graphical models such as hidden Markov models are the most popular techniques being applied in speech field.

A. Support Vector Machines

Support vector machine (SVM) is a supervised learning method introduced by Boser et.al [250]. SVM classifier uses the idea of kernel methods. Kernel method uses kernel functions which transforms the input data into high-dimensional feature space [251]. In this case, the classifier is capable to conduct algorithm in high-dimensional feature space. The main idea behind SVM is to generate linear or non-linear hyperplane for classification. Linear hyperplane is easier to train while non-linear hyperplane lead to more accurate training performance. Nonlinear SVM kernel contains polynomial kernels (quadratic or cubic kernel) and Gaussian kernels (fine Gaussian, medium Gaussian, coarse Gaussian) [252]. These nonlinear kernels are capable of dealing with more complex classification problems.

Due to the flexibility, robustness for handling data, SVM has been applied in a variety of fields [251]. SVM is capable of handling high complex feature space and obtaining non-linear kernel boundaries [252], which makes it a popular method in speech processing field and many other areas [253].

B. Hidden Markov Model

Hidden Markov Model (HMM) is the most classical probabilistic model, which predict output by analysing input characteristics in past states [253]. HMM is a type of variation form Markov chains, of which the current state only relates to the adjacent last state, and is independent from all the other past states [253]. It has been generally explored in speech related applications such as speech recognition [254, 255], speech synthesis [256, 257], speech classification [111], speech emotion recognition [258], and other speech fields. It has been the most popular model for speech research before the deep learning era.

This is an approach with stochastic characteristic so that it is capable of dealing with variability in pathological voice detection robustly [259]. There are also limitations when applying HMM method [260]. First of all, HMM models requires several observation sequence. It is hard to estimate parameters of the model with only one observation sequence. However, the system will become more complex when processing multiple sequences [260]. Another issue when applying HMM is in model selection process. Researchers may find it hard to select the most suitable type of model for a specific problem [260]. Finally, HMM requires large amount of training data due to the complexity of transition matrices and the large amount of parameters. In this case, small data will limit the usage of HMM models, and the balance of interpolation and clustering techniques. Due to the limited bio-medical data, so that HMM model is hard to be applied to medical usage in most cases. In this work, we do not explore with this model.

C. k-Nearest Neighbours

A traditional classification method k-Nearest Neighbour (kNN) is generally applied in speech processing applications such as speech emotion recognition [261, 262], speech synthesis and speech recognition [263]. It identifies the closest k elements near the testing element in the training dataset [253]. These closest k elements and the testing element are regarded as one class. KNN is a supervised learning method, and label the training data is the first essential step. The Euclidean distance function is the most common function to measure the distance between the data elements. Multidimensional distance function can also be applied to represent the element distance. Then by weights or majority voting mechanism, the model selects the most suitable class [264].

The limitation of KNN is similar to all other traditional machine learning methods. It requires large amount of data for training which is time consuming. In addition, it is a supervised learning method so that it requires pre-labelled data. Another difficulty is the most appropriate choice of distance function for specific classification problems.

D. Artificial Neural Networks

Artificial Neural Networks (ANN) is a model that mimics the brain neural system. The data processing of ANN borrows the process of neurotransmitter conduction[265]. It is a model connected by neurons, which is mainly for achieving optimization goals. ANN is generally applied in pattern recognition [266], object detection [267], classification [268], and anomaly detection [269]. ANN computes the cost function by interconnected neurons, so that non-linearity characteristic makes it a complex system for learning high-dimensional information [270]. Shallow ANN with only one hidden layer is capable of recognizing patterns from non-linear input data and achieve good performance [266, 271].

Traditional shallow ANNs such as the multi-layer perceptron (MLP) have been generally applied in speech processing field such as speech recognition [266, 270, 271], speech emotion recognition [272, 273], speaker identification [267, 274], and speech classification [266, 275]. It has also been explored in pathological voice detection [270] [268].

3.3.2 Deep Learning techniques for pathological voice detection

Deep learning was called as cybernetics which originates from 1940s, and went through three main periods. The first period was from 1940s to 1960s. With the development of biological learning [276, 277], the original model was called the perceptron proposed in [278]. This model was capable of training a single neuron. The second period was from 1980s to 1990s, it was known as connectionism (Traditional ANN). Backpropagation [279] was proposed to train a neural network with one or two hidden layers. The third wave of it was from 2006, which is called deep learning [280-282]. From the earliest period, researchers were developing ANN to understand the working mode and function of brain [283].

In the first period of time, McCulloch-Pitts neuron [276] were the earliest model to mimic biological neurons. This is a linear model to distinguish binary inputs by the output (negative or positive) of optimization function $f(x, w)$. However, the weight of this model requires to be set beforehand. By 1950s, perceptron [278, 284] was proposed to become the first model to learn weights. At the same period of time, [285] proposed adaptive linear element (ADALINE) to return scalar for prediction.

Stochastic gradient descent was a variation developed from the ADALINE, which is the main algorithm generally being used in deep learning field.

In the second wave of time from 1980s, connectionism and parallel distributed processing arise with the research interest in neural networks [286, 287]. Connectionism originates from cognitive science. Most researchers studied symbolic modelling in this field at the early stage. In [288], researchers started to build cognitive models based on neural implementations. The main idea behind connectionism is that large amount of neurons connected for calculation is capable of building high-intelligent system.

Backpropagation [279, 289] was proposed and successfully applied in neural netowrks in the second wave. Although it has been ignored for a period in the meantime, it arise in the third wave and has become the baseline algorithm for most of deep learning models today.

The theory behind traditional ANN has been built in 1980s, and it has achieved impressive success in some tasks[290, 291]. However, in some situations, the computation ability and storage memory were still not enough for training. Therefore, this faded in the middle of 1990s. Kernel SVM based methods [250, 292, 293] and probabilistic graphic models [294] became alternative substitute for shallow neural networks.

The third wave starts from 2006 with deep belief network proposed by Geoffrey Hinton [280]. This method has also been applied to achieve impressive results on other CIFAR databases. With the development of advanced computational technologies, researchers have explored various structures of deep neural networks [295-297], and it has achieved impressive results on image and speech applications.

A. Convolutional Neural Network

Convolutional Neural Network (CNN) is an important example of deep learning technologies. With the development in understanding of brain, it achieved significant success in a varieties of machine learning applications. It was the first neural network that was applied in industry and been employed for commercial use. Up to the present, CNN is still the front edge popular algorithm used in most applications. In 1990s, LeCun et al. [290] developed the basic CNN model for reading checks. By the end of 1990s, NEC applied this system for reading over 10% of all the checks in the US.

By 2003, CNN based hand-written recognition system has been developed by [298]. From then on, CNN was the winning technology in many key computer vision challenges. In 2012, ImageNet proposed by Krizhevesky et al. [299] achieved successful performance on ImageNet object recognition challenge. This brought further attention from the commercial and interest of industrial investments. The popularity of deep learning systems has led to successful applications into many different fields.

Compared to fully-connected neural networks, CNN tends to have higher efficiency on training, with easier hyper-parameter tuning process. Therefore, when fully-connected neural networks and backpropagation were seen as problematic in 1990s, CNN succeed in applications and brought the attention of researchers. This motivated more researchers to carry on the neural network investigation.

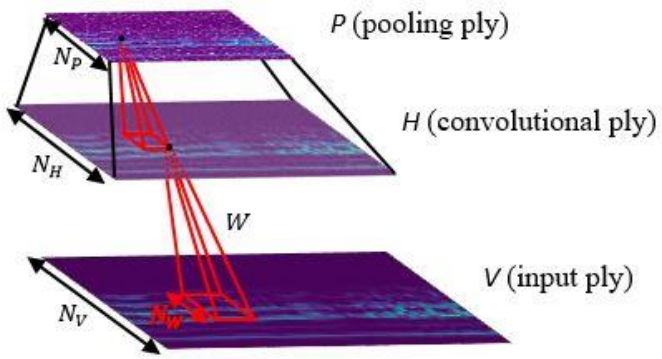


Figure 3.3 CNN structure in one layer

CNN is built by an input layer and several hidden layers. Each individual layer consists of convolutional ply *H* and pooling ply *V*, which is shown in Figure 3.3. For notations, the input feature map is set as $V_l(l = 1, \dots, L)$, and convolutional feature map is set as $H_k(k = 1, \dots, K)$. Weights (filters) are shared among all the units on convolutional ply, on which each unit is computed as,

$$h_m^k = \sigma\left(\sum_{l=1}^{I}\sum_{n=1}^{N_W} v_{l,n+m-1} w_{l,n}^k + w_0^k\right) \tag{3.1}$$

where $v_{l,m}$ means the *m*-th unit of the *l*-th input ply V, and h_m^k means the *m*-th unit of the *k*-th convolutional ply *H*. N_w is set as the size of filters (weights), $w_{l,n}^k$ is the *n*-th unit of the weight (kernel). In this procedure, features are detected locally and automatically by shared-weights throughout the feature map.

In order to reduce the resolution in the convolutional ply and reduce the computational complexity, pooling from convolutional map is essential. Maximization or averaging function are commonly applied to build pooling ply. We set *G* as the size of pooling window, using maximization function, and unit on pooling ply p_m^k is defined as,

$$p_m^k = max_{n=1}^{G} h_{l,(m-1)\times s+n} \tag{3.2}$$

where *s* is stride when the pooling windows shifting among the convolutional ply, *n* is gliding through the filter.

The essence of CNN model is the parameter-sharing characteristic, which refers to the tied weights of the model. This idea requires less storage memory, so that it improves the computation speed and efficiency.

The signal input to CNN contains different dimensionalities. Time series input can be seen as one-dimensional topology data, and image input can be seen as two-dimensional topology data. Different data types has been used in CNN are shown in Table 3.2.

Table 3.2 Different types of data input in CNN

	Single-channel	Multi-channel
1D	Audio waveform: The audio signal changes through time, and it is convolved though time axis.	Skeleton animation data: The angle of each joint in the skeleton of the amination are seen as a signal changing through time. Each joint represent one channel.
2D	2D representations transformed from audio data: Audio data can be transformed to 2D representations such as modulation spectra, spectrogram, scalogram etc.	RGB data images: Red, blue, green are three channels of the image data.
3D	Volumetric data: This type of data generally sourced from biomedical imaging, such as CT scans.	Color video data

In our work, we used biomedical audio signal, and transform it into 2D representations for CNN training.

After forward-propagating the 2D inputs to the CNN layers, the last layer uses Softmax function to convert a vector of numbers into a vector of probabilities as the estimated output $\hat{y}^{(t)}$. The probabilities of each value are proportional to the relative scale of each value in the vector. The loss function are obtained by calculating the cross entropy of target $y^{(t)}$ and estimated output $\hat{y}^{(t)}$. Finally, the weights are updated from the last layer to the first layer to minimize the loss. This process is called back-propagation. Hundreds of epochs of this process leads to a trained CNN network, which is used for predicting the testing data.

B. Recurrent Neural Network

1) RNN

Speech data is a typical sequential time-series data. In the speech recognition field, generalization ability is essential when the machine is fed with variable-length speech data, especially when the information appears in different position of the speech. In this case, "parameter sharing" on different time steps is important. 1D convolution on the speech data is proposed in the late 1980s [300-302]. The output of the 1D convolution is a sequence where each node is calculated by a small number of neighbouring nodes of the input. This is the base of RNN model with weight sharing on different kernels in each time steps. RNN has a different parameter sharing technique. Similar to graphical model HMM, it contains the output of function with the former inputs.

The RNN processing model is shown in Figure 3.4. For notations, at time step *t*, the model are fed into input vector $x^{(t)}$, and the hidden node is notated as $h^{(t)}$, the output is shown as $o^{(t)}$. Loss function is noted as $L^{(t)}$. The shaded box in the left side denotes the time delay of one step. The loop on the shaded box denotes the recurrent connection. In order to describe the feedforward process explicitly, the unfolded RNN model is shown in the right side.

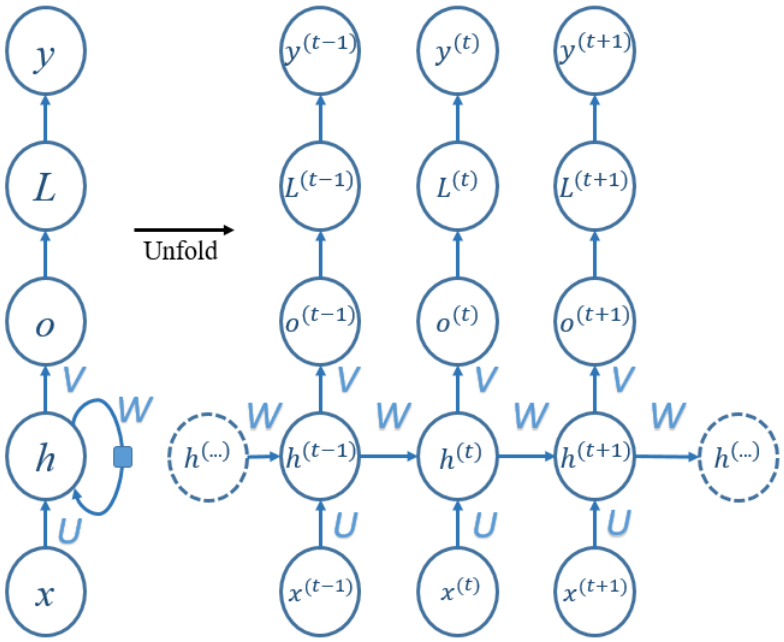


Figure 3.4 *RNN computational graph*

Assume there is tangent activation function for the model in Figure 3.4. From time step $t = 1$ to $t = \tau$, the feedforward updating equations are shown as follow,

$$a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)} \tag{3.3}$$

$$h^{(t)} = \tanh\left(a^{(t)}\right) \tag{3.4}$$

$$o^{(t)} = c + Vh^{(t)} \tag{3.5}$$

$$\hat{y}^{(t)} = softmax(o^{(t)}) \tag{3.6}$$

The loss function at time step t $L^{(t)}$ is calculated by comparing target $y^{(t)}$ with estimated target $\hat{y}^{(t)}$.

$$L^{(t)}\left(\hat{y}^{(t)}, y^{(t)}\right) = -y^{(t)} log\hat{y}^{(t)} - (1 - y^{(t)})\log\left(1 - \hat{y}^{(t)}\right) \tag{3.7}$$

The overall loss is the sum of loss from each time step, which is used for backpropagation to train the model.

$$L(\hat{y}, y) = \sum_{t=1}^{\tau} L^{(t)}\left(\hat{y}^{(t)}, y^{(t)}\right) \tag{3.8}$$

In many applications, the estimated target $\hat{y}^{(t)}$ depends on the whole input sequence data. For example, in speech recognition field, it requires information from both the

past state and the future state. Therefore, bi-directional RNN is proposed [303]. It has been successful in hand-writing recognition [304] and speech recognition[305, 306], and bioinformatics [307].

When training very deep neural networks, gradient vanishing/exploding problem appears in many occasions. Similarly, when RNN process long-sequence input, it is hard for the outputs of the errors associated with the later time steps to affect the computations that are earlier. This is a challenge of learning long-term dependencies that the gradients propagated over a long sequence will vanish (most of the time) or explode. In this case, gated RNN are introduced.

2) Long Short-Time Memory (LSTM)

Gated RNNs is a variation from traditional RNN model, Long-short term memory (LSTM) is a typical gated RNN. Traditional RNN is capable of accumulating information from past time steps. In LSTM, memory gate decides whether the past information is useful in the current time step.

The original LSTM model is proposed in [308], who introduced the recurrent idea (self-loop) in the model. Then an essential development on this is the flexible weight on the self-loop depending on the length of the context[309]. At this stage, the time scale is variable with the changing of the input sequence length. This achieves successful performance on a variety of fields, including hand-written recognition[304] or hand-written generation[305], speech recognition[9, 306], automatic translation[310], image captioning[311-313] etc.

The LSTM cell structure is shown in Figure 3.5. The connected LSTM computational graph is shown in Figure 3.6. Deeper LSTM has shown great advantage in [9, 297]. The feed forward propagation in one cell are shown as below. The current memory states is notated as $c^{(t)}$; current hidden layer output is shown as $h^{(t)}$; the estimated target is $\hat{y}^{(t)}$, and the current input is $x^{(t)}$. The output of forget gate is notated as $f^{(t)}$; the output of update gate is notated as $g^{(t)}$; the output of output gate is notated as $o^{(t)}$. The output of forget gate, update gate, and output gate are calculated as,

$$f^{(t)} = \sigma(b^f + W^f h^{(t-1)} + U^f x^{(t)}) \tag{3.9}$$

$$g^{(t)} = \sigma(b^g + W^g h^{(t-1)} + U^g x^{(t)}) \quad (3.10)$$

$$o^{(t)} = \sigma(b^o + W^o h^{(t-1)} + U^o x^{(t)}) \quad (3.11)$$

The memory state in next time step is calculated as,

$$c^{(t)} = c^{(t-1)} * f^{(t)} + g(t)\tanh(b + Wh^{(t-1)} + Ux^{(t)}) \quad (3.12)$$

This memory state is also been used for calculating the output of this hidden layer. When $c^{(t)}$ is set to 0, it refers that the old state requires to be forgotten. Similarly, the estimated output $\hat{y}^{(t)}$ is the Softmax of the hidden layer $h^{(t)}$.

$$h^{(t)} = \tanh(c^{(t)}) * o^{(t)} \quad (3.13)$$

$$\hat{y}^{(t)} = softmax(h^{(t)}) \quad (3.14)$$

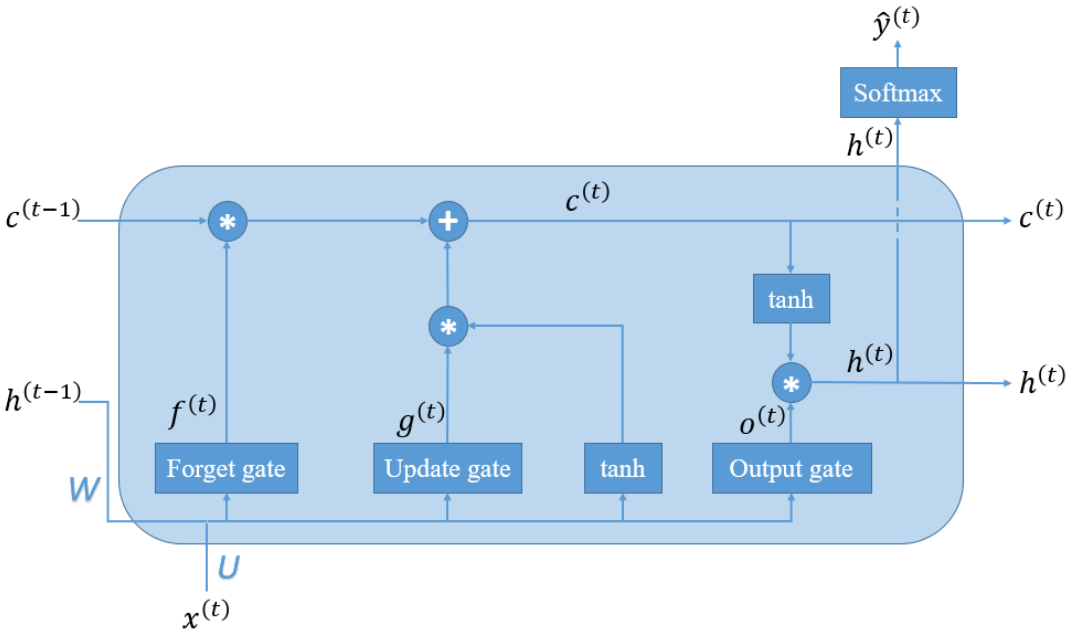


Figure 3.5 One LSTM cell structure

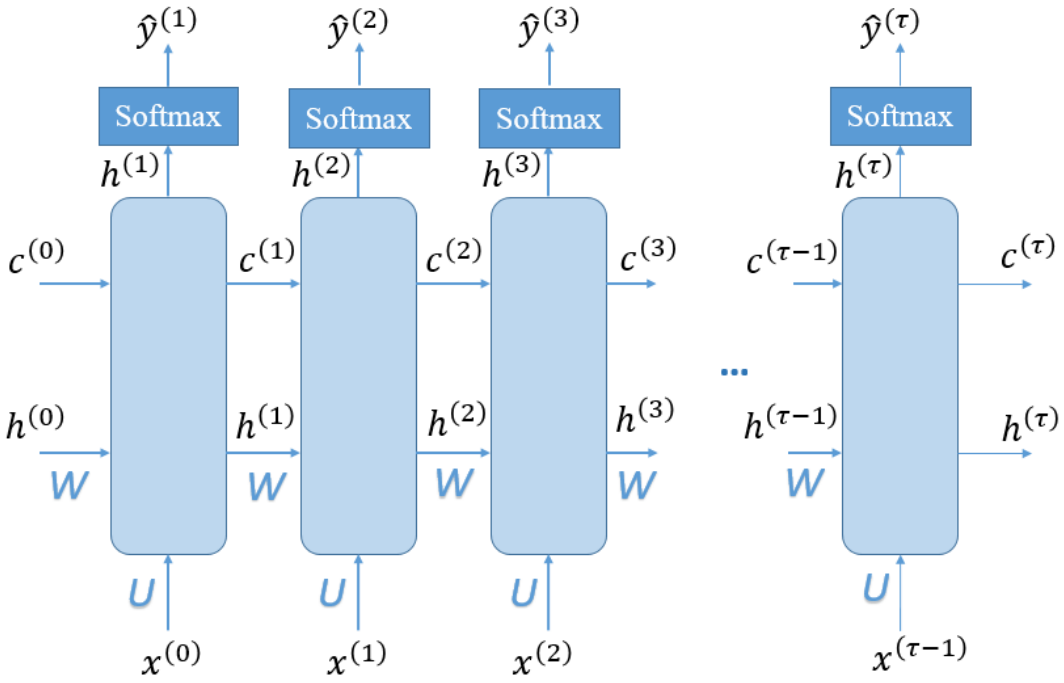


Figure 3.6 LSTM computational graph

Compared to traditional RNN model, it is proved that LSTM shows more advantage in learning long sequence data with the "memory decision" characteristic. LSTM eliminates the vanishing gradient problem in traditional RNN. It shows superior not only on applications that based on artificial datasets [308, 314, 315], but also on up-to-date popular data challenges [305, 310, 316].

3) Attention Mechanism

Attention mechanism was introduced into ML for hand-written generation applications in 2013 [305]. Then it achieved great success in natural language processing (NLP) field and it has become a popular research interest in deep learning applications. Cho et al.[317] and Sutskever et al.[310] introduced similar sequence to sequence (Seq2Seq) architecture respectively for machine translation in 2014. However, this model only processes fixed-length vectors. In order to obtain semantic information from flexible positions in the whole sentence/paragraph, Bahdanau et al. [318] proposed to align the source and the target. They introduced the context vector c, which is weighted average by the hidden layer's nodes. In 2016, Yang et al. [319] proposed hierarchical attention network (HAN) and achieved successful performance on classification tasks compared to the benchmark. In 2017, the

transformer neural network was proposed in NLP field, which improved the efficiency while reducing the computational requirements. This was a breakthrough in NLP field that gave rise to more research interest. Devlin et al [320] introduced the Bidirectional Encoder representations from transformers (BERT) in 2019, and demonstrated its superiority in NLP challenges.

3.4 Review of unsupervised learning for pathological voice detection

In the processing flow in Figure 3.2, generative models represent state-of-the-art unsupervised learning techniques, which shows the probabilistic distribution over the data itself.

3.4.1 Restricted Boltzmann Machine

Initially, Boltzmann machine was introduced for learning random probabilistic distribution on binary variables [321-324]. Suppose there is a set of training samples, the joint probabilistic distribution can be deduced from the observable variable. The probability of specific unit is deduced by logistic regression of other units. However, when there exists latent variable, Boltzmann machine is capable of showing superiority and the latent variables acts as hidden layer in ANN/MLP[325]. Hinton et al. [326] and Bengio et al. [327] proposed feasible backpropagation method for Boltzmann machine. In 1986, restricted Boltzmann machine (RBM) is introduced by [328], of which the name is inspired by harmonium. It contains a visible layer and a hidden layer, with no connections in visible layer or hidden layer itself. Stacked RBM forms deep belief network[280, 329], which is a deep neural network structure. This also provokes the research interest to deep learning field since 2006.

Convolutional Restricted Boltzmann Machine (CRBM) is a typical generative model, and is an extension to the RBM with visible ply and hidden ply as images, which is generally used for pre-training in CNN models. The model is trained to reach thermal equilibrium state, which is the deepest energy minimum state. In this state, hidden ply is able to model the structure of the input data.

CRBM consists of two plies, the visible (input) ply V, and a hidden (convolutional) ply H. Similar to CNN setting as shown in Figure 3.3, weights W^k between input

ply and convolutional ply are shared among all locations in the hidden ply. Hidden units are binary-valued while visible units can be real-valued or binary-valued.

As the CNN structure shown in Figure 3.3, assume the size of visible ply is N_V, and the size of hidden ply is N_H. There are *K* filters (weights) and each weight W^k is convolved with visible ply, and there are bias b_k for each weights and bias c for visible ply. The energy function with binary input is defined as,

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{k=1}^{K}\sum_{j=1}^{N_H}\sum_{r=1}^{N_W} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{N_H} h_j^k - c\sum_{i=1}^{N_V} v_i \tag{3.15}$$

The energy function with real-value data input is defined as,

$$E(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{2}\sum_{i}^{N_V} v_i^2 - \sum_{k=1}^{K}\sum_{j=1}^{N_H}\sum_{r=1}^{N_W} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{N_H} h_j^k - c\sum_{i=1}^{N_V} v_i \tag{3.16}$$

The joint distribution is defined as,

$$P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z}\exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right) \tag{3.17}$$

Similarly, CRBM is trained using block Gibbs Sampling as an extension to Gibbs Sampling in RBM, to maximize the similarity of distribution between construction visible ply and input visible ply, in which case reach the equilibrium state.

3.4.2 Auto-Encoder

Auto-encoder has been explored in this field since 1980s [249, 330, 331]. It was successfully employed in dimensionality reduction, feature extraction, and generative models. The most classical dimensionality reduction application is in Hinton et al.[332] work, which trained the weights using stacked RBM and obtained a bottleneck parameter set with 30 nodes. This bottleneck parameter set can be seen as a features extracted from the original data input without target knowledge. This bottleneck parameter set has been shown to provide better clustering performance than other dimensionality reduction techniques such as PCA [332]. This dimensionality reduction tool also provide convenience in information retrieval. For example, semantic hashing [333, 334] are utilized in textual input and image processing [335-337].

3.5 Conclusion

In this chapter, machine learning techniques are reviewed for pathological voice detection. An overview of the processing flow in this field and sufficient literature review was presented. It was seen that there were three essential challenges: the amount of biomedical data is limited; subjective “weak labels” leads to poor classification; “black box” dilemma still exists.

Both supervised learning and unsupervised learning methods were reviewed. Unsupervised learning method explores the inner rule of the data itself, without learning from the targets. However, Pathological Voice Detection (PVD) is a typical supervised classification problem. In this work, we place emphasis on exploring supervised deep learning methods on pathological voice detection. In the next chapter, we will combine the frame-based traditional acoustic analysis with LSTM model for training.

Chapter 4

4 Deep recurrent acoustic analysis model for pathological voice detection

4.1 Introduction

In chapter 3, machine learning techniques were reviewed for pathological voice detection. Compared to traditional machine learning techniques, deep learning methods have stronger generalization ability on training larger dataset. RNN is a deep learning method suitable for dealing with time-series sequence data. Acoustic analysis extracts stationary features based on short-time frames, which is suitable for RNN model training. In this chapter, a novel deep recurrent acoustic model is proposed. This model applies the frame-based acoustic features in the state-of-the-art deep recurrent model. To our knowledge, this is the first time cepstral features have been used in RNN for pathological voice detection. In Section 4.2, an overview of the proposed deep recurrent acoustic model will be presented. In Section 4.3, experiments on individual acoustic features will be conducted using the model. The relative performance comparison between traditional machine learning methods and the proposed deep recurrent model are given in Section 4.4, which shows the generalization superiority of the deep learning model on the datasets tested.

4.2 Overview of Novel Deep Recurrent Acoustic Analysis model

A block diagram of the proposed model is illustrated in Figure 4.1. In order to extract short-frame features, the signal is pre-processed into time-series data blocks. First, the pre-processing steps are conducted, including DC removal, resampling, frame-blocking and windowing. Then perturbation features, cepstral and spectral features are extracted from each short frame. Afterwards, the selected feature sets are fed into the deep recurrent model.

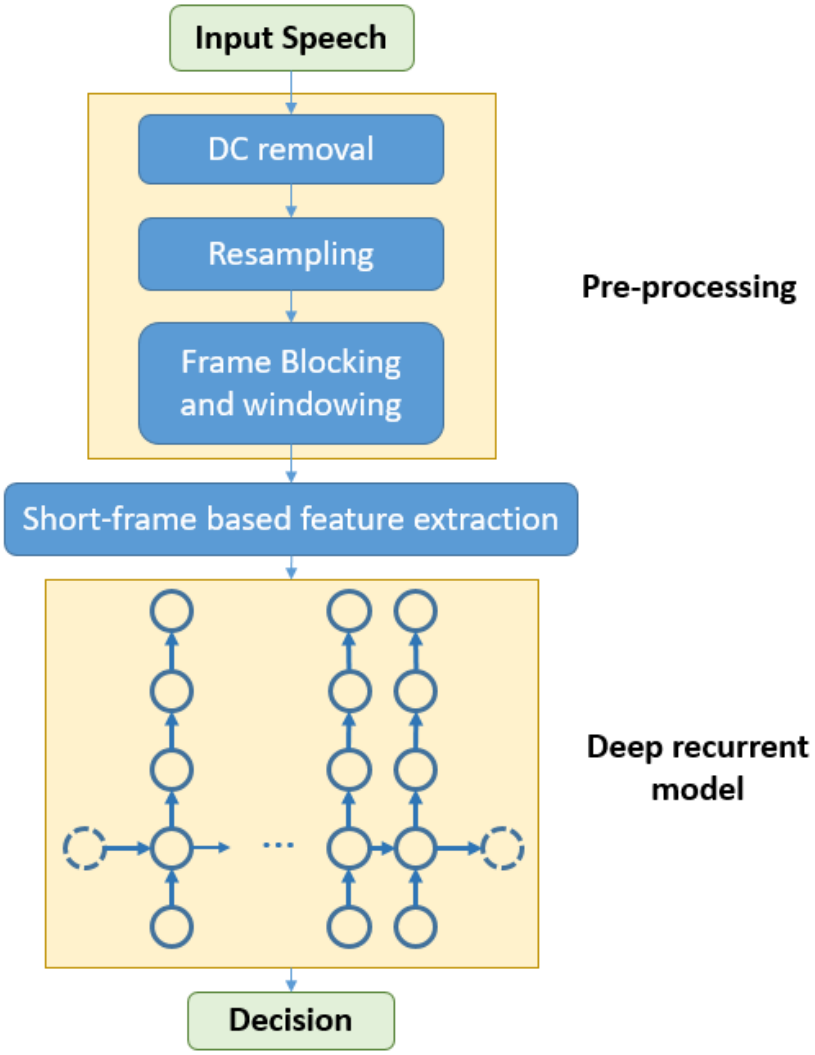


Figure 4.1 Block diagram of the proposed Deep Recurrent Acoustic model

4.3 Pre-processing to get frame-based acoustic features

In this section, the pre-processing steps for obtaining frame-based stationery acoustic features are described in detail.

4.3.1 DC removal

The speech input might contain DC component, a constant on the mean value [338]. The DC component might be caused by recording devices or storage devices. The DC bias components might lead to miscalculation of energy of the signal, and affect some other acoustic features. The main idea behind DC removal is to subtract the mean value from the speech input data. The mathematical equation is shown as,

$$s'[n] = s[n] - \mu_s \tag{4.1}$$

μ_s refers to the mean value of the speech signal, which show as,

$$\mu_s = \frac{1}{N}\sum_{n=0}^{N-1} s[n] \tag{4.2}$$

,where N is the length of the signal.

4.3.2 Resampling

From section 2.5, we can see that the four databases adopted different sampling frequency, with different recording devices. Therefore, there is a requirement for resampling the data to keep it in standard sampling frequency. With regard to the Nyquist sampling theory [339], the sampling frequency must be more than twice of the maximum frequency of the speech input.

$$f_s \geq f_{Ny} = 2f_{max} \tag{4.3}$$

The sampling frequency is notated as f_s. f_{max} is the maximum frequency. f_{Ny} is Nyquist frequency.

Human speech frequency have components ranging from 20 Hz to 20 kHz [339]. In this case, Nyquist sampling frequency is 40 kHz. However, most essential frequency components and formants are distributed in the 0 to 12500 Hz range. In order to reduce the computation burden, we resample all the speech data to 25 kHz using Matlab.

4.3.3 Frame Blocking and Windowing in Speech Processing

In order to process the time-varying features on the non-linear speech data, the signal is generally framed into small segments. In each small signal frame, the short-time segment can be seen as stationary and linear data. Each frame has overlaps with the adjacent frames to avoid the effect of discontinuity of consecutive frames. The frame size and the overlapping percentage are seen as parameters to set. When using smaller frame sizes, the signal has higher time resolution but lower frequency resolution, and is called wideband analysis [340]. In contrast, when using larger

frame sizes, the signal has lower time resolution but higher frequency resolution, and is called narrowband analysis.

According to different age, gender, and voice quality, speech signal ranges in different frequency band [340]. Generally, the maximum fundamental frequency of human speech signal is 500Hz. Researchers often use 20ms frame size for keeping time and frequency resolution in balance [11, 104-106]. In this work, we use 20ms as the frame size, and 75% overlapping percentage. Figure 4.2 shows the block framing process on the speech waveform. After block framing process, windowing process is conducted on each frame. This reduces the effect of discontinuity of the consecutive frames and spectral leakage phenomenon [341].

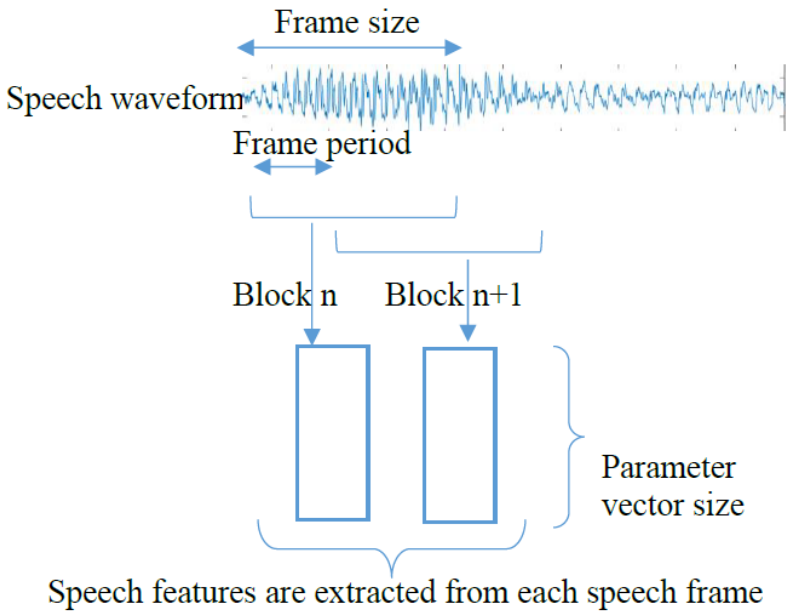


Figure 4.2 Block framing and overlapping process

4.4 Short-frame based feature extraction, feature selection and classification

In this section, we first conduct classification experiments on the extracted short-frame based acoustic features individually, and then select the correlated features as a feature set for training. The performance with traditional machine learning methods are compared with the performance on the proposed deep recurrent acoustic model.

4.4.1 Deep recurrent acoustic model for classification

The architecture of classification model is shown in Figure 4.3. In this model, there are two hidden bi-directional LSTM layer with 512 nodes connected on each layer, a fully connected hidden layer with 512 nodes, and a Softmax layer for classification outputs. The mini-batch size is set 30, the maximum epoch is set 50, and the training-validation split is 70% to 30%.

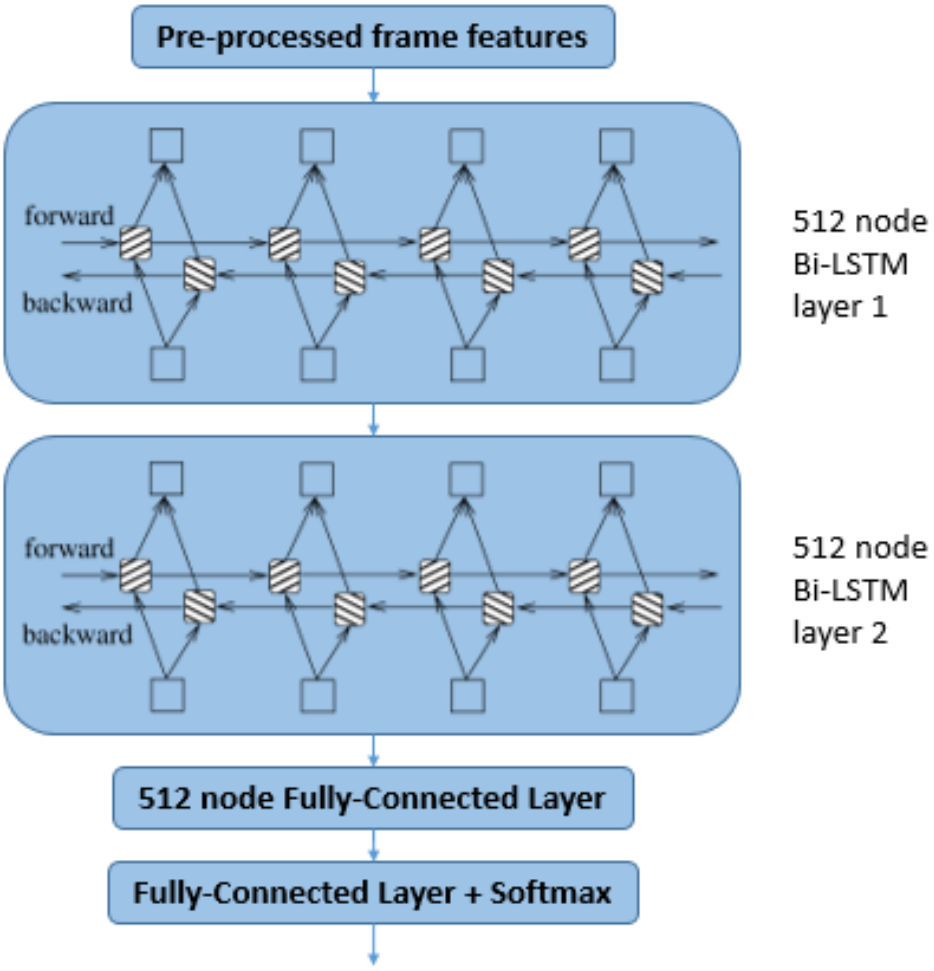


Figure 4.3 The deep recurrent model architecture

4.4.2 Experiments on individual acoustic features for feature selection

In this section, we explore the performance on a variety of features based on the short-time frames for feature selection. Since the SVD database is a comparatively reliable database, we employ this database for experiments on individual acoustic features. There are overall 1000 pathological patients and 687 healthy patients. For each patient, one sustained vowel /a/ is used. Therefore, there are overall 1000

pathological speech data versus 687 healthy speech data, which contains certain level of unbalance. In order to keep training data in balance, we select 70% of the healthy speech data (480) and the same amount of pathological speech data (480) as the training data. In this case, there are overall 960 training data and 727 testing data.

Studies on voice disorders demonstrate that cepstral features are more reflective of perceptual impression as discussed in Section 2.4.6. CPP and CPPs contain great levels of correlation with overall dysphonia, breathiness, and asthenia ratings than the other acoustic measures [342]. In this section, we explored performance of cepstral related features on the recurrent models.

In the Analysis of Dysphonia in Speech and Voice software manual (ADSV), cepstral features are the main statistics for analyzing the speech signal, including Cepstral Peak Prominence (CPP), Low-to-High Harmonic Ratio (LHR), Peak-position frequency of cepstrum, CPP/AVG statistics, Regression Slope and Cepstral Intensity [343, 344]. In this section, these features are individually explored. In addition, we proposed two self-developed cepstral features. ADSV statistics are achieved in a cumulative manner, with some statistics of parameters used in the generation of another.

Similarly, when analysing the cepstrum of human speech, the pitch is shown in the first cepstral peak, which corresponds to the fundamental frequency of the speech signal. The first peak of the cepstrum appears to be useful in pathological voice detection too. In Figure 4.4 and Figure 4.5, the cepstrum of control voice and pathological voice are compared.

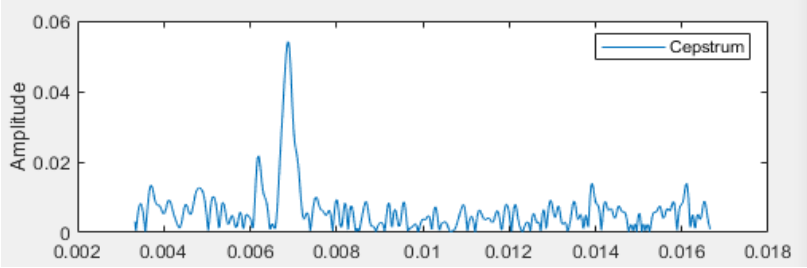


Figure 4.4 Cepstrums of a normal voice example on quefrency band 60Hz to 300Hz

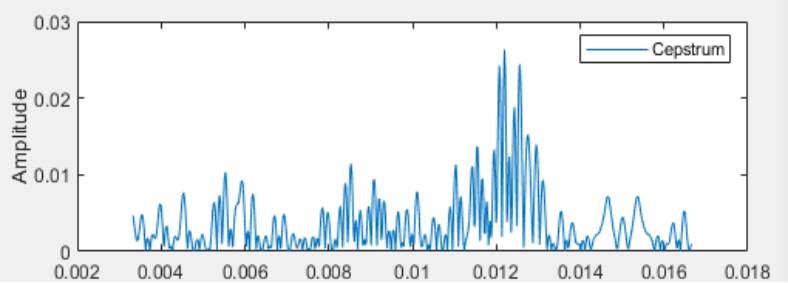


Figure 4.5 Cepstrums of a pathological voice example on quefrency band 60Hz to 300Hz

The main peak – maximum cepstral peak of the speech signal can be extracted from the quefrency band 60Hz to 300Hz. We call the corresponding position MaxPos. From Figure 4.4 and Figure 4.5, we can see that the cepstrum of pathological voice samples have more irregular side lopes near the main peak, while the cepstrum of the normal voice samples shows smoother status. The irregular side lopes in Figure 4.5 can be seen as noisy components in speech signal, which correlates to hoarseness in GRBAS scale. With regards to this characteristic, we proposed two novel cepstral features: Standard deviation of Cepstrum (CepStd) and Second Peak Perturbation (SPP).

In this section, the ADSV manual cepstral features and our proposed two cepstral features are extracted from the cepstrum for classification evaluation. Feature details and the related performance on each feature are discussed as follows.

1) Cepstral Peak Prominence (CPP)

Cepstral peak prominence has been shown to be one of the strongest correlations of breathiness. For each cepstrum that based on the short frame, a linear regression is fit to the array of cepstral magnitude in the quefrency band 0 to 10kHz. The difference between the estimated linear regression value on the cepstral peak position (MaxPos) and the real cepstral peak value (MaxCepstrum) are termed as the cepstral peak prominence. Traditional features calculates the statistics of CPP on all frames such as mean value, maximum value, minimum value or standard deviation of the CPP value on all frames. Comparatively, deep recurrent model learns from CPP information on all the frames. The Training progress and the loss is shown in Figure 4.6. It can be seen that the accuracy is slightly improving during the training process. The performance is 64.69% accuracy, which proves CPP correlating well on detecting dysphonia.

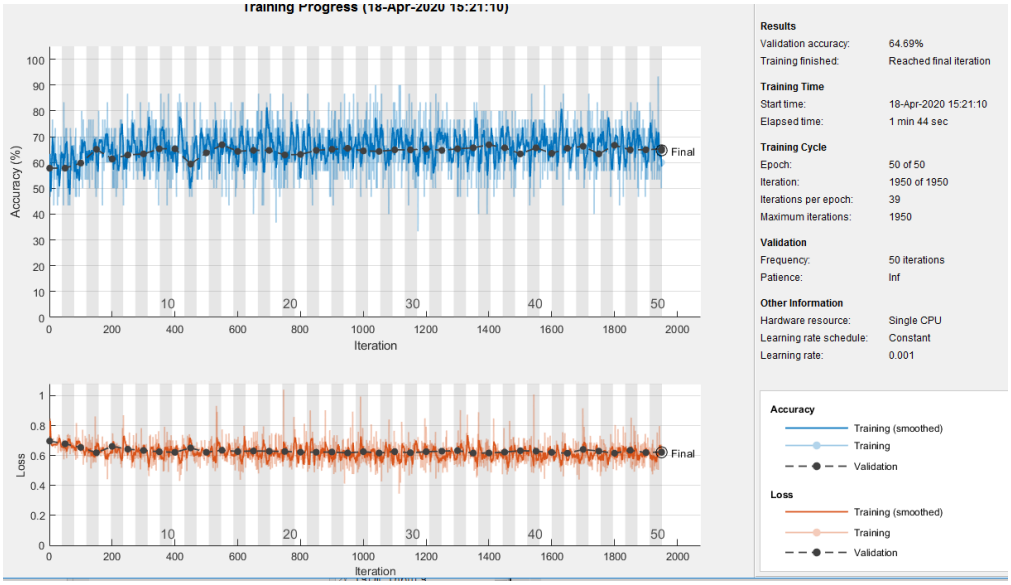


Figure 4.6 Training progress and loss with CPP

2) CPPf0 (Peak frequency position)

CPPf0 refers to the peak frequency postion (MaxPos) ranging from 60Hz to 300Hz on the quefrency band. It is often recognized as pitch information. MaxPos in each frame is sent as frame-based features to the deep recurrent classifier. The result is shown in Figure 4.7. It can be seen from the training progress that the network is not training at all. This means that pitch information might not be a good metric for evaluating the pathological information, compared to CPP itself. The accuracy is 56.8%, which is very poor. Therefore, we abandon this feature.

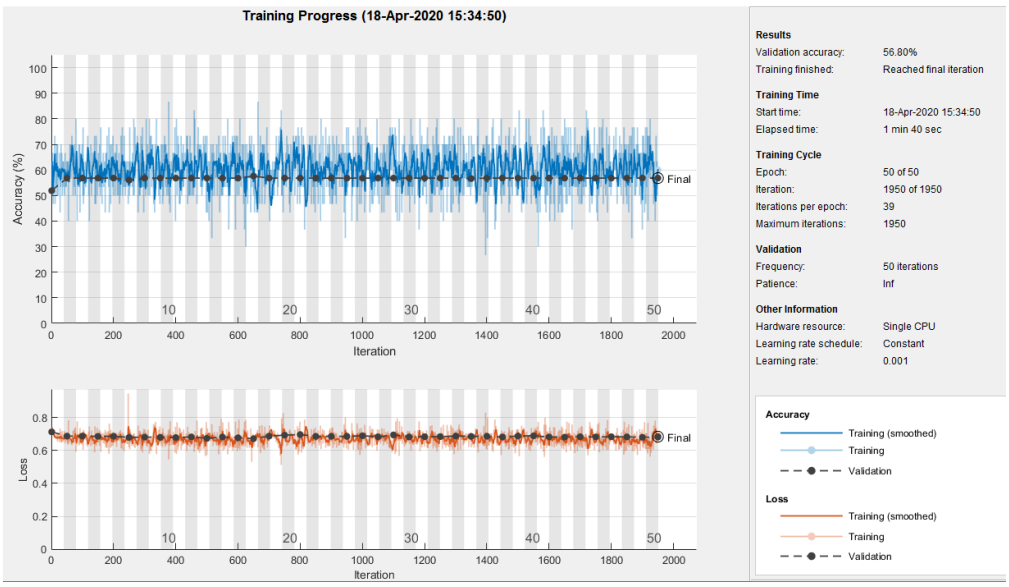


Figure 4.7 The training progress and loss with CPPf0

3) CepAvg

This feature is also estimated on the frequency range from 60Hz to 300Hz. The ADSV definition for this feature is that this is the difference between the largest cepstral peak value and the mean value of the cesptrum ranging in 60Hz to 300Hz for the selected voiced frames. It can be seen from the training progress in Figure 4.8 that it is slowly training with 66.27% accuracy at end. This proves that CepAvg is a comparatively good feature.

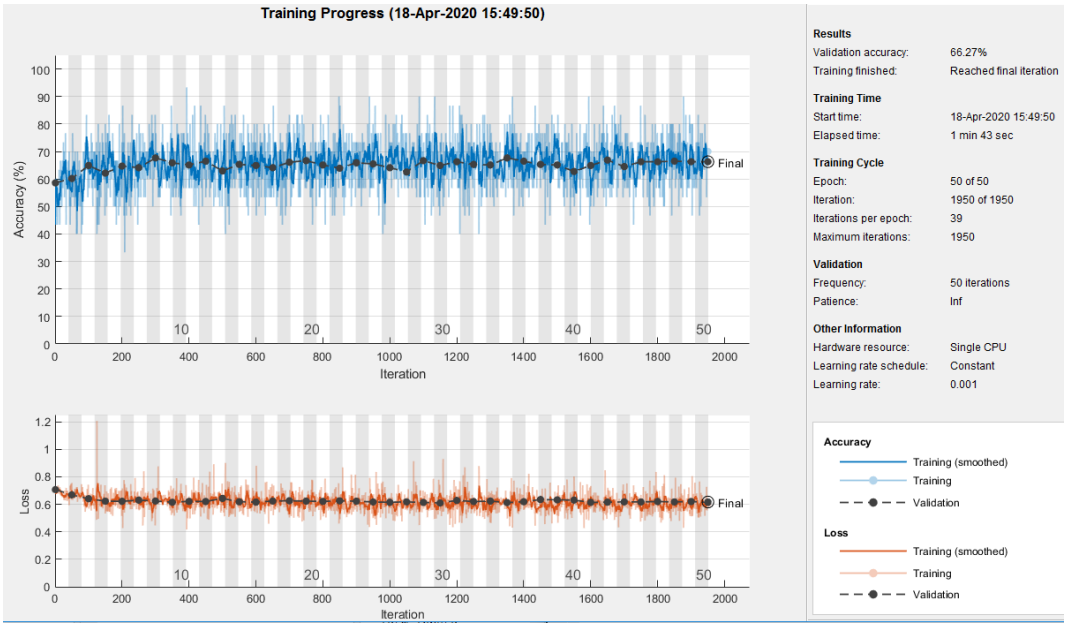


Figure 4.8 The training progress and loss with CepAvg

4) Cepstral Intensity

Cepstral intensity refers to the mean averaged value of the cepstral magnitudes. Traditional cepstral intensity features such as mean or standard deviation are statistics that derived from this array of cepstral averages. For deep recurrent model, cepstral intensities in the array are seen as a sequence. However, from the performance in Figure 4.9, we can see the training progress is not changing through time, with 59.17% accuracy at the end. Therefore, we abandon this feature.

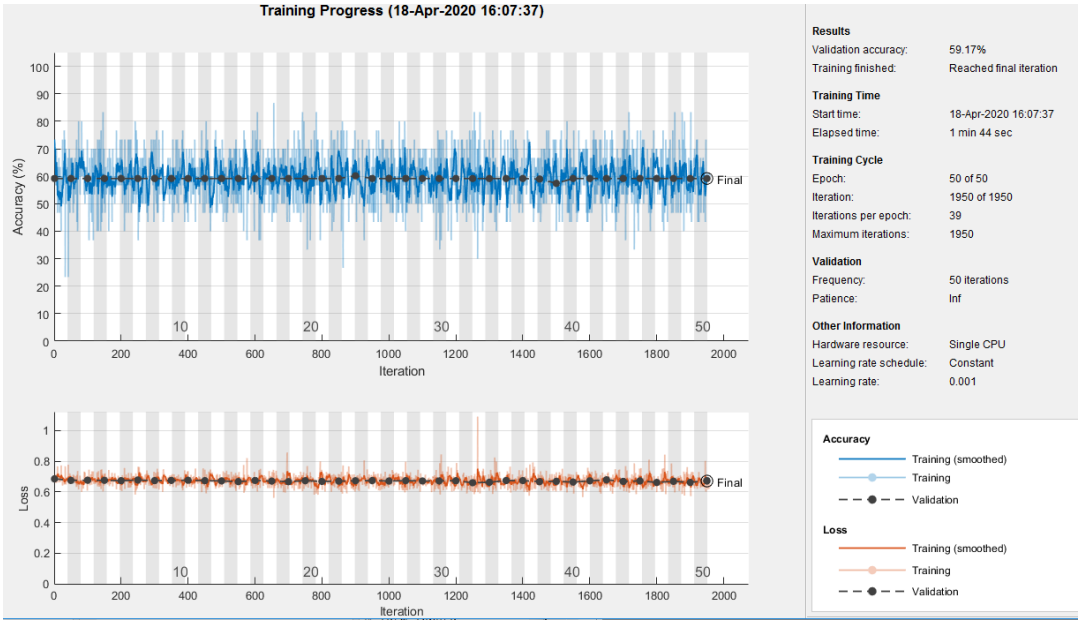


Figure 4.9 The training progress and loss with cesptral intensity

5) Low-to-High Harmonic Ratio

In each frame, Low-to-High ratio (L/H ratio) is the ratio of energy in the frequency band ranging from 0 to 4000Hz and the energy above 4000Hz. Similarly, statistics such as mean value, standard deviation, maximum and minimum value of L/H ratio of the frames are calculated as traditional acoustic features. In this model, the L/H ratio in each frame is seen as a sequence input to be fed into the deep recurrent network. In Figure 4.10, the training progress and loss with L/H ratio are shown. It can be seen that L/H ratio achieves 65.09% on detecting pathological voice, and the validation loss drops quickly in the start of the training process. It can be seen from the training progress that the validation accuracy is steadily increasing. In this case, we decide to adopt this feature.

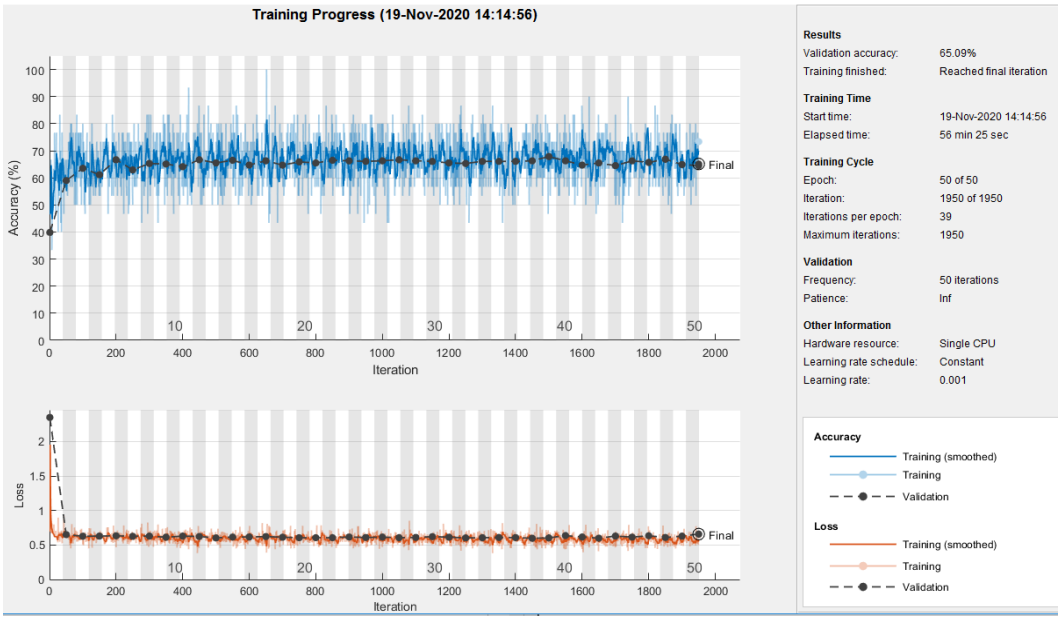


Figure 4.10 The training progress and loss with L/H ratio

6) Regression Slope

Regression slope refers to the normalized regression slope values of the estimated linear regression of the cesptrum on specific frame. In Figure 4.11, the training progress and loss with regression slope are shown. It can be seen that the model is not training at all, with 54.83% training accuracy at the end. In this case, we abandon this feature.

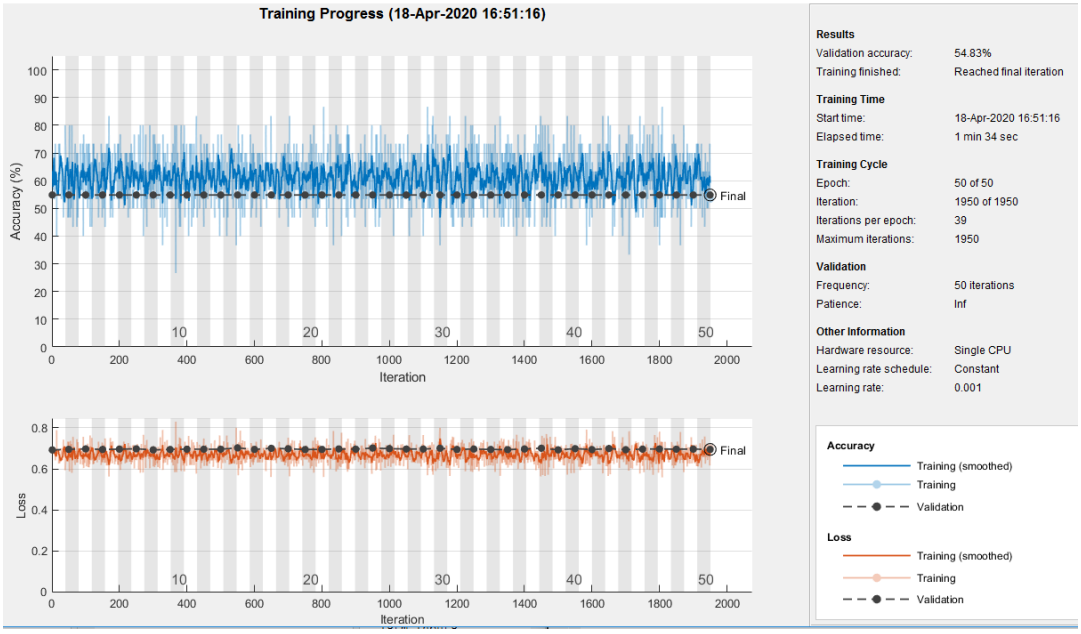


Figure 4.11 The training progress and loss with regression slope

6) Energy

Energy on each frame is also been evaluated on the model, with 59.17% accuracy shown in Figure 4.12. We abandon this feature.

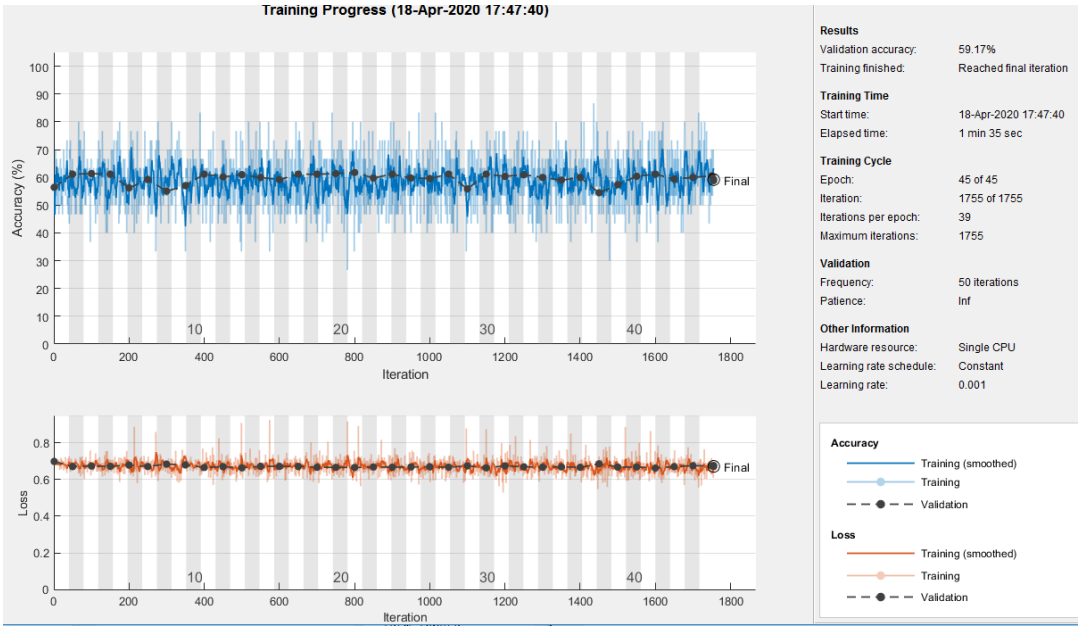


Figure 4.12 Training progress and loss with energy

7) Standard deviation of Cepstrum (CepStd)

CepStd is standard deviation of cepstrum on the specific frame, which is a new

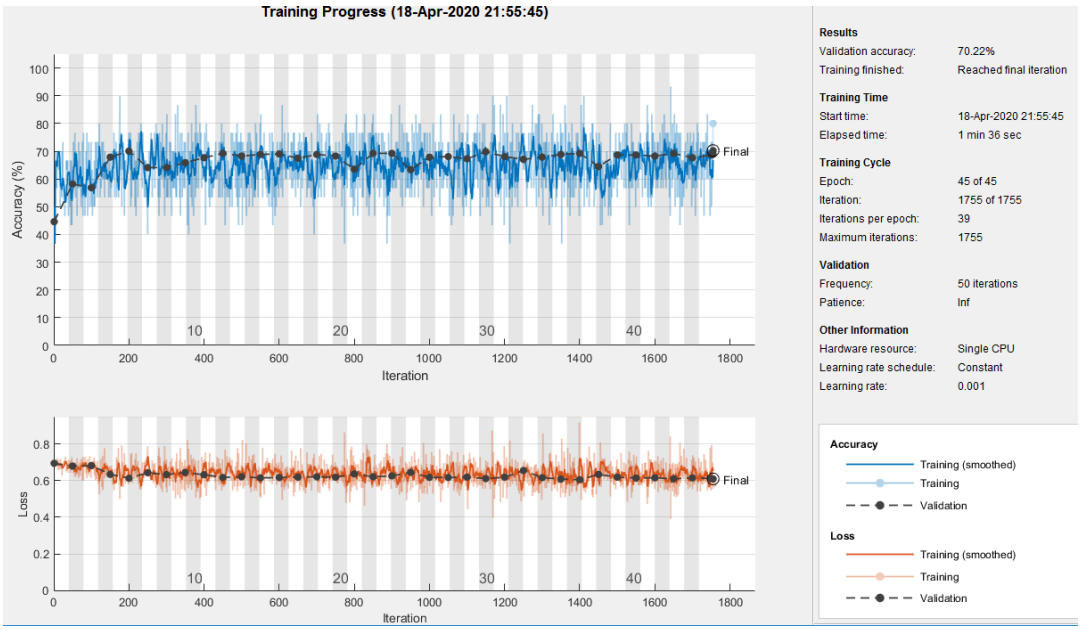
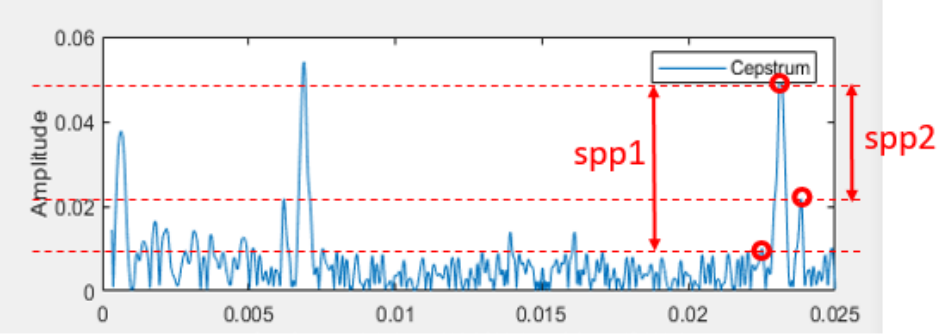


Figure 4.13 Training progress and loss with CepStd

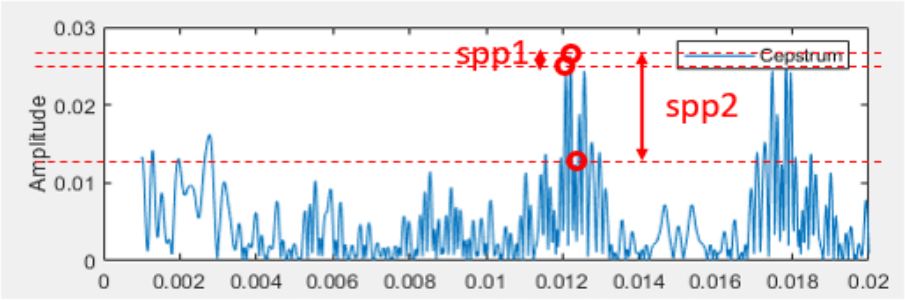
feature we proposed. Since it can be seen from Figure 4.5 and Figure 4.6 that the cepstrum of pathological data shows more irregularities than the cepstrum of normal data, standard deviation of the cepstrum on each frame is calculated as a feature sequence for classification. In Figure 4.13, we can see that this feature achieved 70.22% accuracy, and the validation loss is reducing through time, with training accuracy increasing during the epochs. This shows that CepStd is a comparatively good frame-feature for classifying pathological voice and normal voice.

8) Second Peak Perturbation (SPP)

When comparing the cepstrum of normal voice and pathological voice, the perturbation near the second largest peak contains a serious amount of information too. Not only the irregularities appearing at this area, we can also observe that the difference of the second largest peak and the peaks adjacent with it is distinguishable. We proposed this feature as second peak perturbation (SPP). From Figure 4.14, SPP1(left) and SPP2(right) of normal voice cepstrum and pathological voice cepstrum are shown.



(a)



(b)

Figure 4.14 spp1 and spp2 shown in cepstrum in (a) a normal voice sample (b) a pathological voice sample

Since there are more intense perturbation of peaks in the neighbouring area of second peak in pathological voice, the SPP are much smaller than it in the normal voice. We feed this feature into the deep recurrent network, and it achieved 67.46% accuracy. This shows that the proposed feature is comparatively reliable. The performance of the training progress and loss are shown in Figure 4.15.

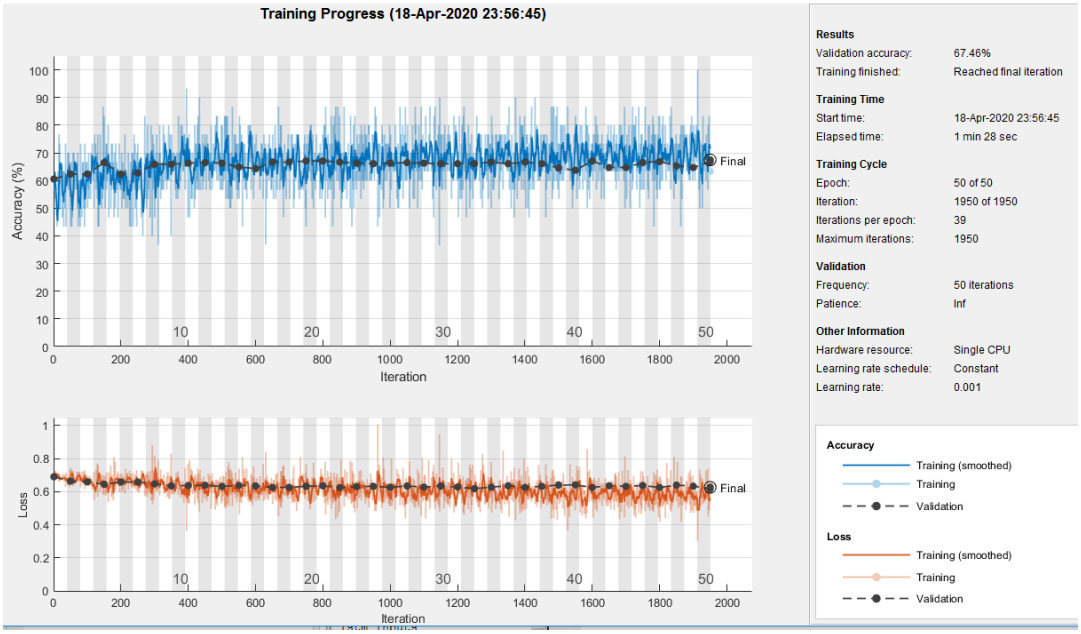


Figure 4.15 Training progress and loss with SPP1 and SPP2

9) Mel-frequency Cepstral Coefficients (MFCC)

The vocal cords (source) produces low-frequency excitation, while the vocal tract (filter) is shown as the formant. Mel-frequency cepstrum seperates the impact of the source and the filter clearly, and this is the reason why cepstral features performs better than traditional spectral features. The lower order coefficients contains most information of overall spectral shape of the source-filter components, while the higher order coefficients represents the spectral details. Generally, 12 to 20 cepstral coefficients are applied for speech applications. In our work, we choose 16 cepstral coefficients. The performance of the 16-MFCC on the deep recurrent model is shown in Figure 4.16.

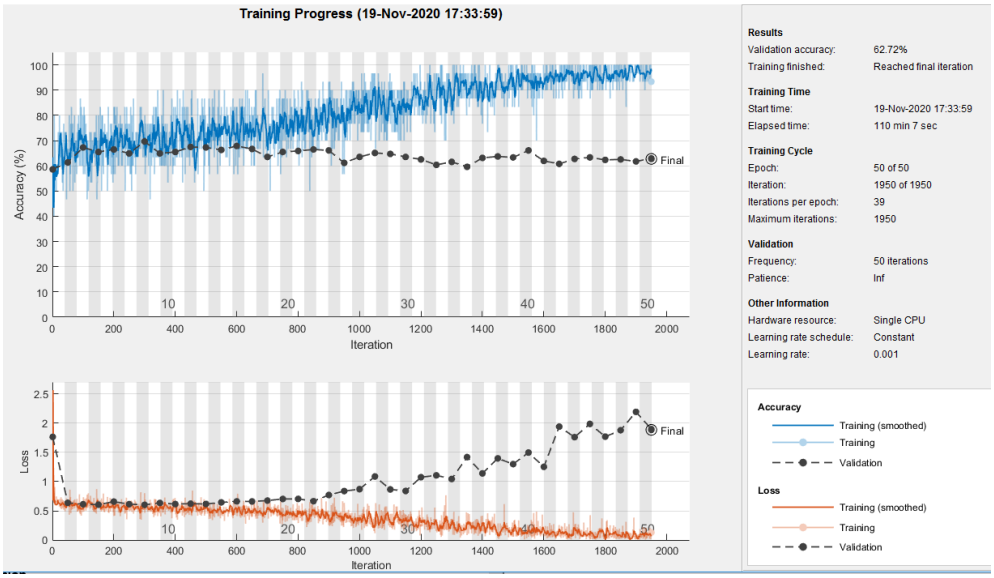


Figure 4.16 Training progress and loss with MFCC (training epoch:50)

The validation accuracy is 62.72%. However, it can be seen that the validation loss starts to increase at the end of the training process, and the training accuracy approaching to 100% accuracy. This overfitting phenomenon might be due to the number of MFCC coefficients, and MFCC characteristic itself. Therefore, MFCC is applicable for representing the dysphonic feature, but the training epochs needs to be stopped before 12 (early stopping). When set the number of epochs to 12, the performance is shown in Figure 4.17. The accuracy achieves 66.47%, and the overfitting problem is eliminated. This Early Stopping technique is also applied in the experiments in next section.

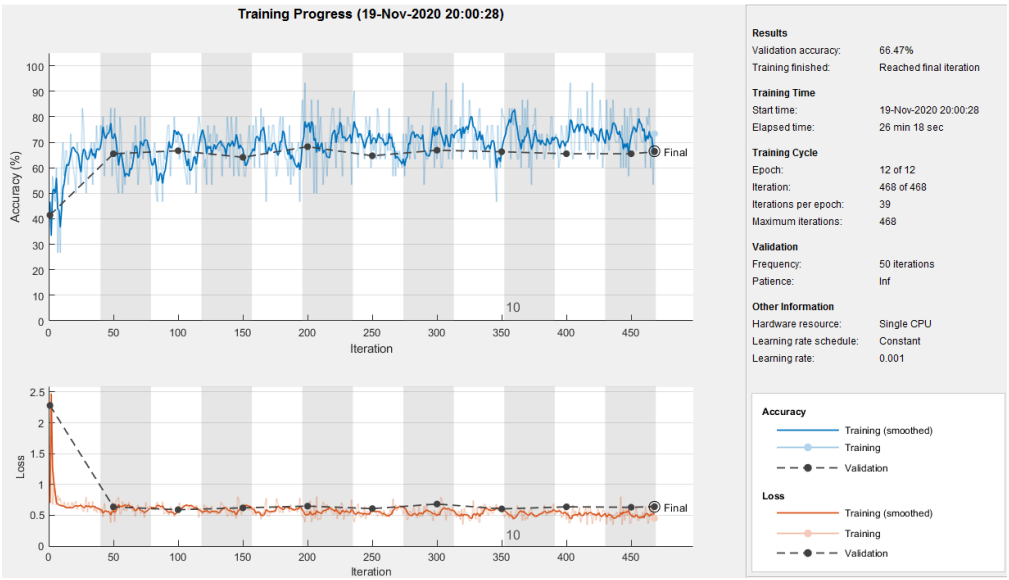


Figure 4.17 training progress and loss with MFCC (training epoch:12)

In conclusion, the overall classification performance on each individual feature are shown in Table 4.1. Some correlated features are adopted for feature selection in the next section, due to the comparatively good accuracy and the performance during the training process shown in the figures above.

Table 4.1 Overview of performance on individual cepstral feature

Cepstral Feature	Performance
CPP (dB)	**64.69%**
CPPf0 (Hz)	56.80%
CepAvg (dB)	**66.27%**
Cepstral Intensity (dB)	59.17%
Low-High harmonic ratio	**65.09%**
Regression Slope	54.83%
Energy	59.17%
CepStd (dB)	**70.22%**
SPP	**67.46%**
16-MFCC	**66.47%**

4.3.3 Experiments on selected feature sets

We also explored perturbation features such as jitter, shimmer. However, these features shown poor results. Cepstral features and spectral features tend to performs much better, since this does not require pitch detection. Pitch detection is not robust in some situations, which leads to error in the later classification process. In addition, CNN extracted features perform poorly, which is due to the complexity and redundancy of the process of data.

From Table 4.1, we can see that CPP, CepAvg, L/H harmonic ratio, CepStd, SPP and MFCC shows good performance. In this section, we exploited the fusion of the selected cepstral features to feed into the deep recurrent model for pathological voice detection. Apart from SVD database, we also validates the performance on PdA and AVPD databases. Similarly, SVD contains 70% of the normal data (480) and the same amount of pathological data (480) as the training data, and rest are testing data

(207 healthy data and 520 pathological data). There are overall 280 training data in PdA database, with 140 healthy and 140 pathological data, and testing data contains 99 healthy data and 61 pathological data, overall 160 data samples. AVPD database contains 232 training data (116 healthy and 116 pathological) and 107 testing data (51 healthy and 56 pathological). The overview summary table of the experiments' performance are shown in Table 4.2.

Table 4.2 Overview performance summary on deep acoustic recurrent model with different feature set

Experiment number: Feature set	Performance		
	SVD	PdA	AVPD
1: CPP, CepAvg	66.07%	69.70%	70.59%
2: CPP, LHR	64.69%	71.97%	68.63%
3: CPP, CepAvg, LHR	67.26%	72.73%	64.71%
4: CPP, CepAvg, LHR, 16-MFCC	70.02%	80.30%	73.53%
5: CPP, CepAvg, LHR, SPP	65.29%	72.73%	74.51%
6: CPP, CepAvg, LHR, CepStd	67.65%	80.30%	70.59%
7: CPP, CepAvg, LHR, SPP, CepStd	67.06%	81.06%	66.67%
8: CPP, CepAvg, LHR, SPP, CepStd, 16-MFCC	71.60%	82.58%	78.43%

Compared to experiment 1,2, and 3 using the traditional ADSV cepstral features, experiment 5, 6 and 7 achieved better performance with proposed new feature SPP and CepStd. It shows the validity of the proposed novel features on the traditional cepstral features. In this table, we can also see that MFCC increase the robustness of the model. In experiment 4 and experiment 8, the feature sets predict successfully on all three databases. Especially in experiment 8, the traditional cepstral features CPP, CepAvg, LHR and 16-coefficient MFCC achieved 71.60%, 82.58%, and 78.43% accuracy respectively on SVD, PdA, and AVPD databases, which is the best overall performance in these experiments. Comparing experiment 5,6,7 with experiment 4, we can see that CepStd contribute to the better performance on PdA database, while SPP contribute to a better performance on AVPD database. In addition, compared to experiment 4, the overall accuracy of experiment 8 in PdA and SVD databases

improves. This gives us evidence that the two novel proposed cepstral features SPP and CepStd works well in presenting dysphonic characteristics.

The proposed model achieved 71.60% on the most challenging SVD database. Compared to the bench mark, which surpasses Pavol's work (68.08%) by 3.52% [11]. Pavol's work is the only work which adopted whole SVD database. This is the first work exploring deep learning in pathological voice detection field. In addition, the performance on PdA database achieved 82.58%, it surpasses the work using traditional features and GMM [234] (76.71%) by 5.87%, and it surpasses the work using traditional features (noise perturbation features, MFCC, non-linear features) and SVM by 3% to 5% [70]. For AVPD database, this model achieved 78.43%, which is also comparable with the work using MDVP features and FDR (72.53%) [126].

In order to analyse more details in the performance of the proposed novel features, we introduce the following metrics. Sensitivity (SN) reveals how good the classifier is at detecting the pathological voice files, which has the same meaning as "recall" and is calculated as in (4.4). Specificity (SP) calculated as in (4.5) reveals the proportion of normal voice files that are correctly identified. Precision (P) in (4.6) shows how many of the pathological voice files classified are relevant, and F1-score (F1) has also been taken into account, calculated as in (4.7).

$$SN = \frac{TP}{TP + FN} \tag{4.4}$$

$$SP = \frac{TN}{FP + TN} \tag{4.5}$$

$$P = \frac{TP}{TP + FP} \tag{4.6}$$

$$F1 = 2\frac{P \cdot SN}{P + SN} \tag{4.7}$$

True Negative (TN) represent normal voice recordings that are correctly detected as "normal voice"; True Positive (TP) represent pathological voice recordings that are correctly detected as "pathological voice"; False Negative (FN) represent pathological voice recordings that are detected as "normal voice", False Positive (FP) represent normal voice recordings that are classified as "pathological voice".

We use the ROC curve for analysing the performance of the models. In binary classification problem, the model will not predict the result with simply 0 or 1, but a probability for each sample. When applying Softmax for obtaining probabilistic forecast in each class, it will allow us to set a threshold to decide positive and negative sample. ROC curve) has False Positive Rate (FPR) as x-axis, and True Positive Rate (TPR) as y-axis. TPR refers to the ratio of positive samples that is predicted correctly, and FPR refers to the ratio of negative samples that is predicted as positive sample wrongly. The definition of TPR and FPR are shown in equation (4.8) and (4.9).

$$TPR = \frac{TP}{TP + FN} = SN \tag{4.8}$$

$$FPR = \frac{FP}{FP + TN} = 1 - SP \tag{4.9}$$

Each point in the curve represent different FPR and TPR with different threshold. When sensitivity (SN) equals to 1 and specificity (SP) equals to 1, it refers to the point (0, 1). This means all the predictions are correct. When sensitivity (SN) equals to 0 and specificity (SP) equals to 0, it refers to the point (1, 0). This means all the predictions are wrong. When point of the curves falls on the diagonal line, it means that it is a random prediction with 50% accuracy. Therefore, the best condition is that the curve is closing to the point (0, 1).

The area under the curve (AUC) is a metric for analysing the performance of the classification model. When AUC equals to 1, the curve falls on the point (0, 1), it means that this is a perfect classifier. However, perfect classifier does not exist in most cases. The higher the AUC area, the better the classifier is.

We will now investigate the confusion matrix and performance of experiment 4 and 8 for the three databases. The confusion matrix on SVD, PdA, and AVPD databases with feature set in experiment 4 (CPP, CepAvg, LHR, 16-MFCC) are shown in Table 4.3, Table 4.4 and Table 4.5 respectively, and the related ROC curves on three databases are presented in Figure 4.18. The related metrics for evaluating model 4 are shown in table 4.6.

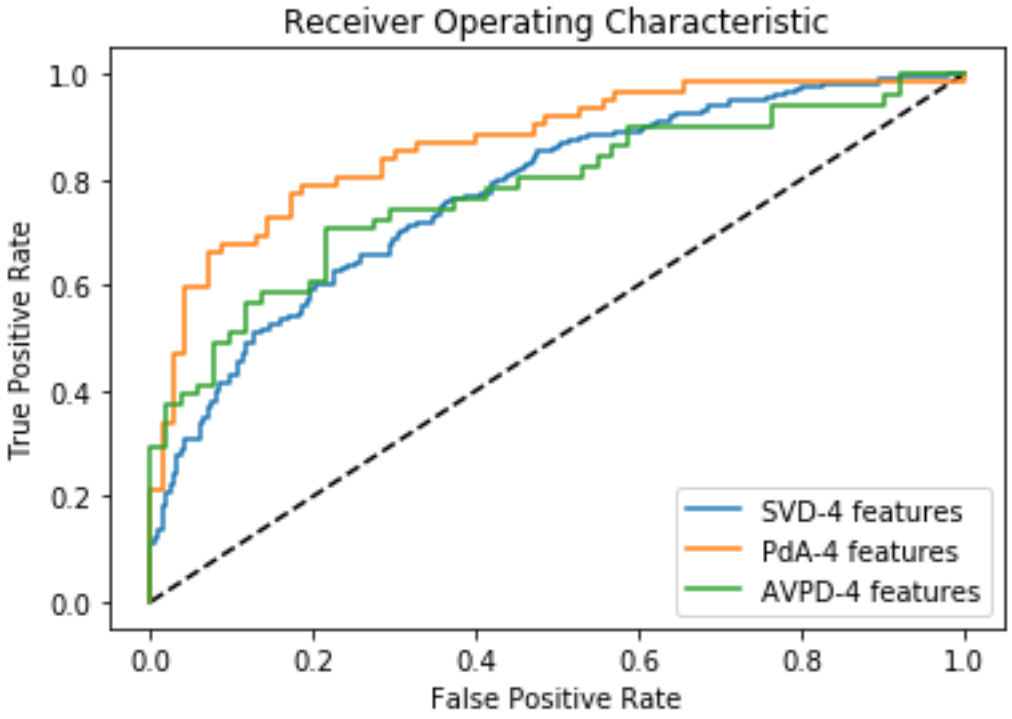


Figure 4.18 ROC curve on three databases with four features (experiment 4: CPP, CepAvg, LHR, 16-MFCC)

Table 4.3 Confusion matrix on SVD database with feature set of experiment 4 (CPP, CepAvg, LHR, 16-MFCC)

	True: heathy	**False: Pathological**
Prediction: healthy	147	85
Prediction: Pathological	67	208

Table 4.4 Confusion matrix on PdA database with feature set of experiment 4 (CPP, CepAvg, LHR, 16-MFCC)

	True: healthy	**False: pathological**
Prediction: healthy	57	13
Prediction: pathological	13	49

Table 4.5 Confusion matrix on AVPD database with feature set of experiment 4 (CPP, CepAvg, LHR, 16-MFCC)

	True: healthy	**False: pathological**
Prediction: healthy	39	12
Prediction: pathological	15	36

Table 4.6 Classification performance with selected feature set on experiment 4 (CPP, CepAvg, LHR, 16-MFCC)

	TPR(SN)(r)	SP(1-FPR)	p	F1	Accuracy
SVD	68.69%	70.99%	63.36%	65.92%	70.02%
PDA	81.43%	79.03%	81.43%	81.43%	80.30%
AVPD	76.47%	70.59%	72.22%	74.29%	73.53%

It can be seen that the PdA database achieves best performance with 80.30% accuracy. The left top point of ROC curve in PdA database is closest to (0,1). Compared to experiment 1, 2 and 3, the overall performance improved greatly. It validates the usefulness of MFCC coefficients.

The confusion matrix on SVD, PdA, and AVPD databases with feature set in experiment 8 (CPP, CepAvg, LHR, SPP, CepStd, 16-MFCC) are shown in Table 4.7, Table 4.8 and Table 4.9 respectively, and the related ROC curves on three databases are presented in Figure 4.19. The overall metrics evaluating the performance of experiment 8 is listed in Table 4.10.

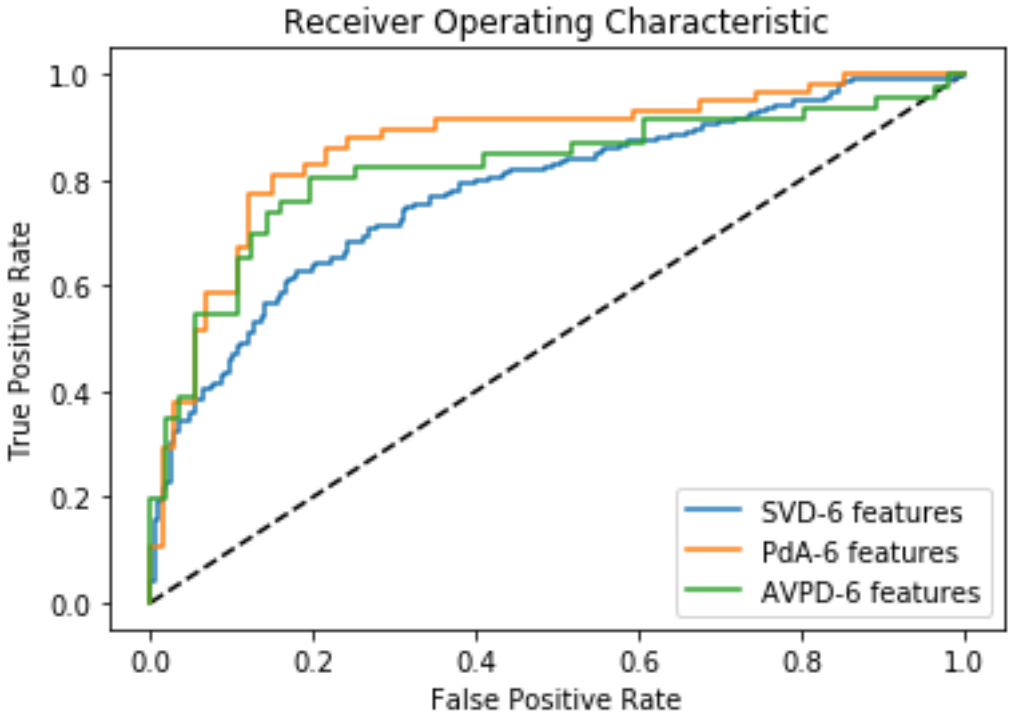


Figure 4.19 ROC curve on three databases with six features (experiment 8: CPP, CepAvg, LHR, SPP, CepStd, 16-MFCC)

Table 4.7 Confusion matrix on SVD database with feature set of experiment 8 (CPP, CepAvg, LHR, SPP, CepStd, 16-MFCC)

	True: healthy	**False: pathological**
Prediction: healthy	156	101
Prediction: pathological	50	200

Table 4.8 Confusion matrix on PdA database with feature set of experiment 8 (CPP, CepAvg, LHR, SPP, CepStd, 16-MFCC)

	True: healthy	**False: pathological**
Prediction: healthy	62	11
Prediction: Pathological	12	47

Table 4.9 Confusion matrix on AVPD database with feature set of experiment 8 (CPP, CepAvg, LHR, SPP, CepStd, 16-MFCC)

	True: healthy	**False: pathological**
Prediction: healthy	42	8
Prediction: pathological	14	38

Table 4.10 Classification performance with selected feature set on experiment 8 (CPP, CepAvg, LHR, SPP, Cepstd, 16-MFCC)

	TPR(SN)(r)	**SP(1-FPR)**	**p**	**F1**	**Accuracy**
SVD	72.33%	71.10%	63.14%	67.42%	71.60%
PDA	83.78%	81.03%	84.93%	84.35%	82.58%
AVPD	75.00%	82.61%	84.00%	79.25%	78.43%

It can be seen from the ROC curves that PdA database still achieves the best performance with this model, with overall accuracy (82.58%) and highest F1-score (84.35%), and the ROC curves closest to the left-top point. Compared to experimental 5, 6 and 7, it can be seen that the feature CepStd also contributes to the improved performance in PdA databases. In Figure 4.20, the overall ROC performance on the two models are shown. The ROC curves of three databases in experiment 8 gets closer to the left-top point (0,1) than it in experiment 4. It gives evidence that the proposed feature SPP and CepStd help to detect dysphonic characteristics.

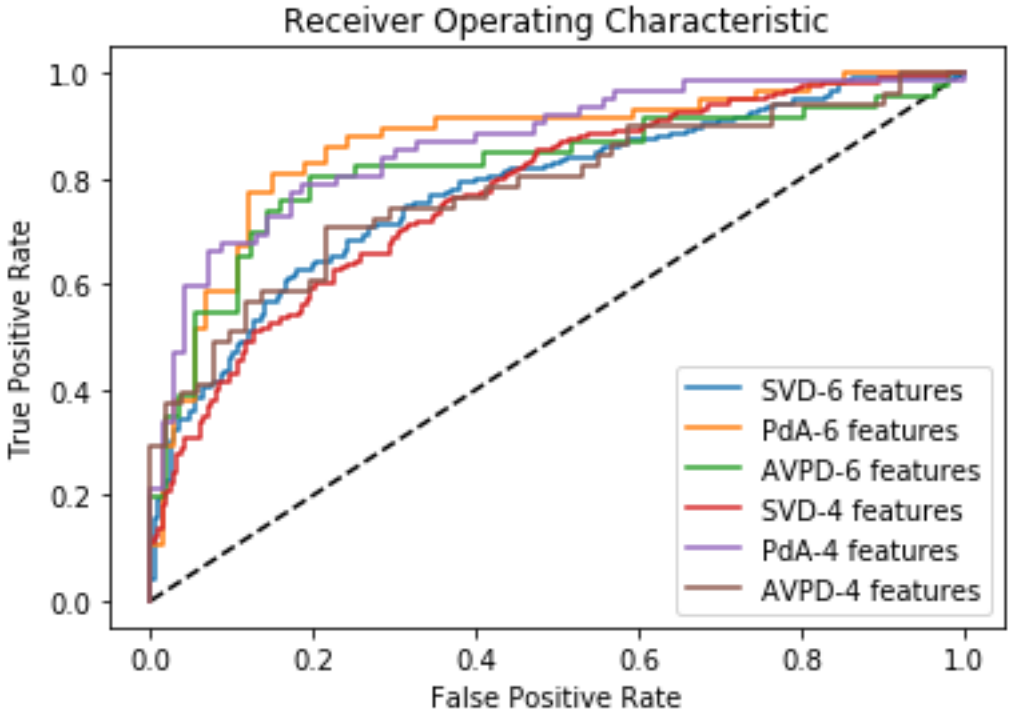


Figure 4.20 Overall ROC performance comparison on deep acoustic recurrent model

4.3.4 Comparison with traditional machine learning methods

The last section validated the usefulness of specific traditional cepstral features including CPP, CepAvg, LHR, and MFCC. It also demonstrated the power of our proposed novel cepstral features: SPP and CepStd. The proposed model combines time-series features very well with the deep recurrent model. In this section, in order to evaluate deeper with the deep recurrent model without regard to the proposed features, we compare the performance on traditional machine learning classifiers with the same parameter sets on the deep recurrent model.

In order to fit the features set into traditional classifiers, the statistics of cepstral features are calculated. We use the fusion of features from the ADSV manual and spectral features MFCC and wavelet packet transform features. Experiments shows that the best parameter set is with: mean value and standard deviation value of pitch, MFCC, CPP, CPPf0, CepInt, CepAvg, regression slope, energy, L/H ratio, and maximum and minimum value of CPP, L/H ratio, and the energy and entropy of 6 level wavelet packets. (Overall 174 coefficients) Similarly, the data is split into 70%

for training and 30% for testing. The model is evaluated with 5-fold validation on training data, and the testing data performance is shown in Table 4.11.

It can be seen that Linear SVM achieved overall best performance on three databases, with 71.15%, 76.52% and 78.43% on SVD, PDA and AVPD database respectively. The best performance with SVD and PDA databases on the traditional classification model are not as good as the proposed deep recurrent acoustic model. The performance on AVPD database is same on traditional methods and proposed model. This is due to the AVPD database characteristic itself. AVPD database contains a large amount of speech corpuses with long silence before speech onset. This affect less on statistics of features. However, it affect seriously on recurrent model since it uses the series data information.

Table 4.11 Performance on traditional classifiers with selected feature sets (including cepstral features, MFCC, and energy and entropy of wavelet packet transform)

Classifier	Setting	Performance/%		
		SVD	PDA	AVPD
Fine Tree	Maximum number of splits:100 Split criterion: Gini's diversity index	62.25	63.64	75.49
Kernel Naïve bayes	Kernel: Gaussian	61.26	74.24	70.59
Linear SVM	Kernel function: Linear	71.15	76.52	78.43
weighted KNN	Number of neigbors:10 Distance metric: Euclidean Distance weight: squared inverse	67.00	80.30	74.51
Boosted Trees Ensemble	Ensemble method: AdaBoost Maximum number of splits: 20 Number of learners: 30 learning rate: 0.1	69.57	78.03	77.45

4.4 Conclusion

In this chapter, we proposed a deep recurrent acoustic model for pathological voice detection. Cepstral features were investigated with two novel proposed features: SPP

and CepStd. Compared to the benchmark, the overall performance shows that Bi-LSTM is successful in detecting frame-based dysphonic features. This is the first time cepstral features has been explored with deep recurrent models. It also demonstrates that cepstral features, especially Cepstral peak prominence (CPP) and MFCC correlates well with dysphonia.

In addition, we compared the deep recurrent model with traditional classification models. It was shown that deep recurrent model achieves higher accuracy with less features than traditional models. In the next chapter, a CNN-based model is proposed for pathological voice detection, which explores 2D representation of speech data.

Chapter 5

5 Deep convolutional model for pathological voice detection

5.1 Introduction

In this chapter, we propose a novel automatic pathological voice detection model based on deep convolutional neural networks (CNN). The proposed algorithm uses spectrograms of voice samples. To the best of our knowledge, our model is the first one applying spectrogram in deep neural network for pathological voice detection. The performance of the proposed algorithm is validated on SVD database. In order to reduce the overfitting problem, transfer learning is explored in the following section with varieties of 2D time-frequency representations (spectrogram, zoomed spectrogram, scalogram) of the speech. Four state-of-the-art CNN models (AlexNet, ResNet, GoogleNet, and XceptionNet) are used for transfer learning in pathological voice detection, and the performance are compared on three databases.

5.2 Deep convolutional model for pathological voice detection

It was shown in Chapter 2 that temporal perturbation amplifiers jitter, shimmer rely heavily on steady status of the sustained vowels, and the pitch estimation algorithms are sensitive to pathological unstable voice conditions. In Chapter 4, we showed the advantage of cepstral features and spectral features in deep recurrent models, which achieved successful performance. Nevertheless, compared to these traditional parametric features, time-frequency 2D representations can also show hoarseness, roughness, and breathiness, with complementary dynamic pathological information[169] as reviewed previously in Section 2.4.7. In addition, there is no requirements for pitch detection.

Speech data are time-series data which can be thought of 1D data [248]. In Section 3.3.2, we introduced Convolutional Neural Networks (CNN) [345], which is a

particular type of neural network to process grid-like topology data, especially 2D grid data input. Time-frequency representation is a 2D grid data input transformed from speech. In this work, we transformed the 1D speech data into 2D time-frequency representations, and explored their use in CNN end-to-end training system.

5.2.1 Overall processing flow of CNN model for PVD

The overall processing flow block diagram is shown in Figure 5.1. Python programming language has been used with the signal processing package *scipy.signal*. Firstly, a pre-processing step is carried out, and the original speech is resampled at 25 kHz to keep consistency. Then a Short-Time Fourier Transform (STFT) is applied to the resampled data for transforming the time-domain signal into spectral-domain signal. In STFT, in order to achieve the balance of time resolution and frequency resolution, each file use 20 ms Hamming window segments, with 75% overlap between consecutive windows. Overall 1000 pathological and 687 normal data in the SVD database is used, and the training testing split is 70% to 30%. In order to keep training data in balance, we select 70% of the normal data (480) and the same amount of pathological data (480) as the training data. In this case, there are overall 960 training data (480 healthy and 480 pathological) and 727 testing data (207 healthy and 520 pathological). Finally, the spectrogram is cropped to the same size of 84 * 257 points to keep all the data in the same length. 0.435s (84) from the onset of speech, this is the length of the shortest speech corpus. 257 is the number of frequency width of the spectrogram.

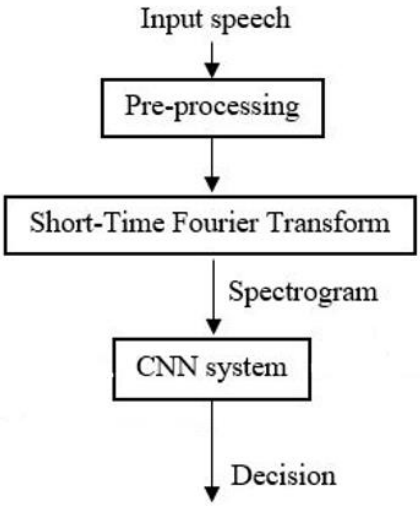
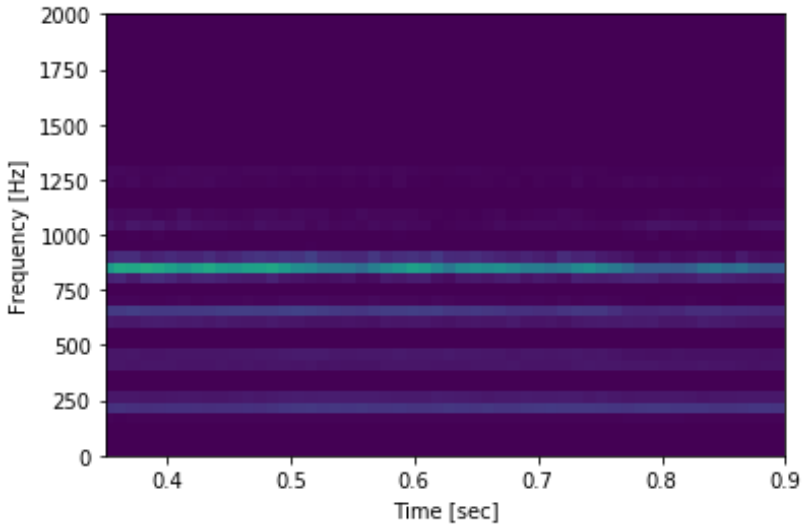
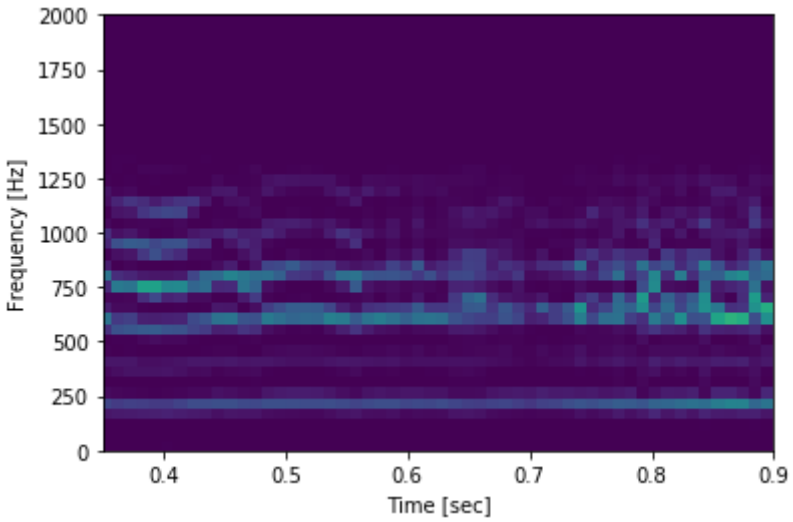


Figure 5.1 Block diagram flow of CNN model for pathological voice detection

An example of the comparison of input spectrogram between normal voice and pathological voice is shown in Figure 5.2. From Figure 5.2, we can see that the spectrogram contains more irregularities in pathological speech data than healthy speech. Not all the healthy speech spectrogram is as smooth as it is in Figure 5.2 (a), while this is representative in most examples.



(a)



(b)

Figure 5.2 Example of spectrogram representation of (a) a healthy speech (b) a pathological speech

Afterwards, the 2D representations are fed into a CNN system for training, the CNN architecture details is described in the next section 5.2.2.

5.2.2 Proposed CNN architecture

Pathological voice contains subtle differences that can be seen on the spectrogram compared to normal voice, which are difficult to be manually defined using particular criteria. An End-to-End CNN system is capable of extracting features automatically and use it for classification. The size of the 2D representation input is 84*257*1. In order to find the most appropriate parameters adapted to the input, we conducted a large amount of experiments to search it. A 5-layer, 10-layer, and 15 layer architecture was investigated and the resulting performance on the SVD database are in Figure 5.3. In addition, we compare the performance of the model with small number of nodes and the model with increasing number of nodes by layers (5-layer: 8, 10, 20, 50, 100; 10-layer: 8, 10, 20, 50, 80, 100, 150, 200, 250, 300; 15 layer: 8, 10, 20, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550). The training progress and loss function detail is shown in Appendix A.

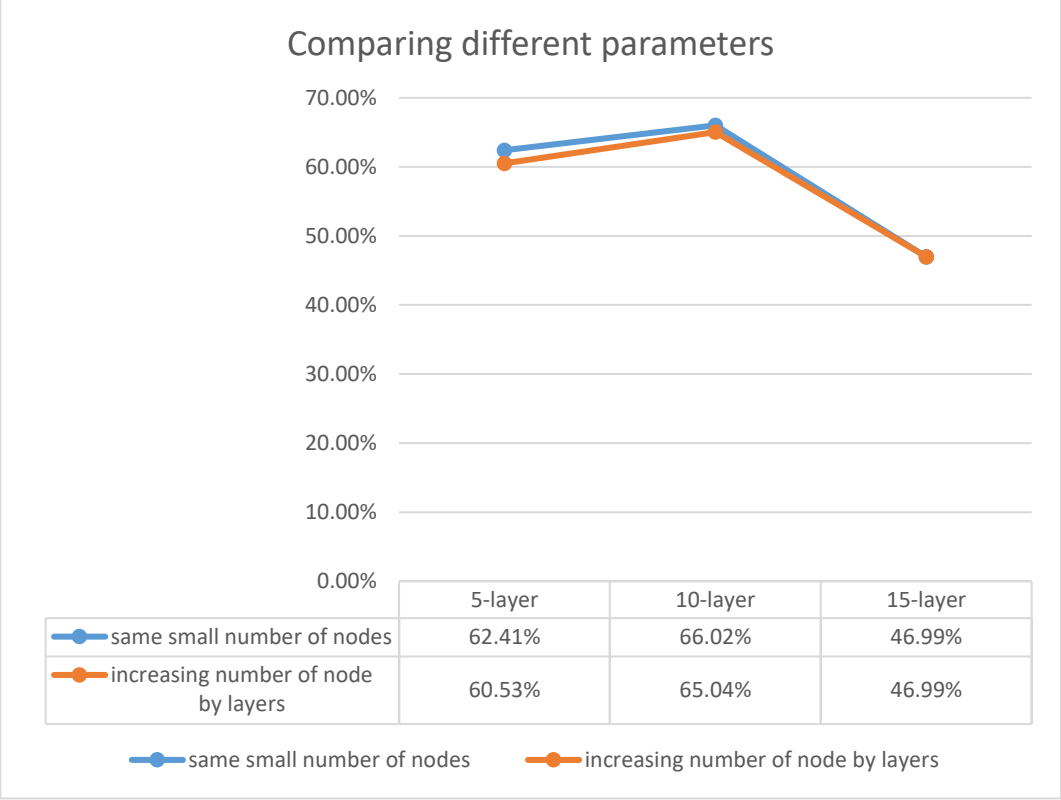


Figure 5.3 comparing the performance with different parameters

In Figure 5.3, we can see that 15-layer CNN performs very poor. The training progress are shown in Appendix A4 and A5, and the 15-layer CNN did not shows training loss change with the epochs. This means that complex model such as very deep CNN is not suitable for this application. 10-layer CNN system is the most

appropriate model and achieves the best performance compared to 5-layer CNN and 15-layer CNN system. In addition, 8 nodes in each layer is enough for training compared to the increasing number of node by layers. It achieves better performance while it reduces the space and computational complexity. Therefore, we choose the 10-layer CNN structure with 8 nodes in each layer.

The proposed CNN architecture is shown in Figure 5.4. Since it is the spectrogram of the speech file, the depth of this input layer is 1, and it has the same meaning as "colour channels", i.e. Red-Green-Blue (RGB) in computer vision field; in other words, the spectrogram can be seen as a grey scale image input.

The input feature map is then convolved with a set of 8 filters. Each filter has the shape of 8*3*1 and stride of 1. We use the rectangular filters in this work due to the spectrogram characteristics. Unlike image input, spectrogram is time-frames-based representation. In order to explore the pathological information between the adjacent time frames, the width is chosen as 3, representing the time frame before and after the current time frame. The length is 8, corresponding to the frequency axis. This is explored by experiments. Furthermore, max-pooling filters with the shape 4*4 and stride of 1 are applied to pool the significant values out and reduce the computational complexity. Then the activation function RELU [346] is applied to make the neural network non-linear and fit for classification.

After the first hidden layer, each layer was convolved with 8 filters with the shape 8*3*8 and stride of 1. The depth is 8 due to the number of filters in the first layer. Max-pooling filters and activation function is the same as for the first hidden layer. After 10 hidden layers to extract the features from the spectrogram, the feature map is formed into a Dense Layer, which is a fully-connected layer, to train the model for classification. L2-regularization is used in the layer to avoid overfitting problems.

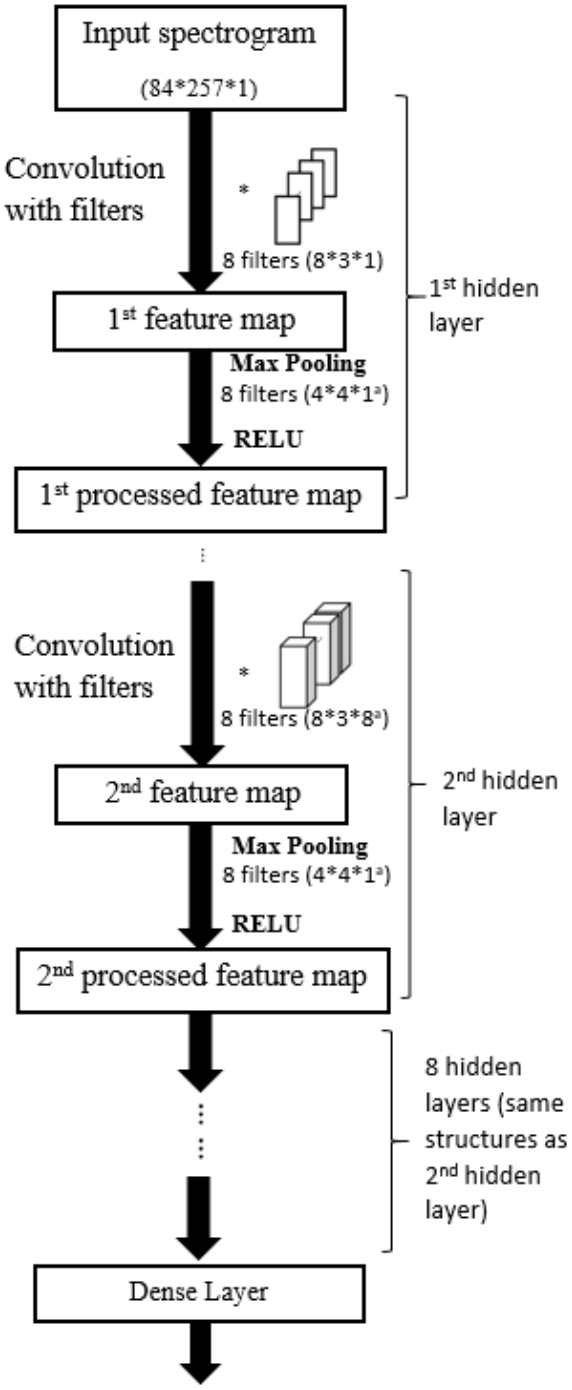


Figure 5.4 CNN architecture (a. length*width*depth)

5.2.3 Experimental Results and Discussion

Python based Tensorflow is used as a framework for the training process. Because the mini-batch gradient descent [347] use GPU for matrix computation and will lead to high speed, the training samples are divided into 256 samples in each mini-batch to be trained on GPU NVidia GTX1070 in this work. Adam Optimizer [348] was applied in the experiment with initial learning rate 0.0003 so that the training process

becomes more robust. Delta value of the L2 regularization is set to 0.0001 and the maximum epochs of training is 100. The loss with the training process is shown in Figure 5.5. The confusion matrix of testing dataset is shown in Table 5.1, and the metrics to measure the performance is shown in Table 5.2.

```
Cost after epoch 0: 0.692995
Cost after epoch 10: 0.660903
Cost after epoch 20: 0.626552
Cost after epoch 30: 0.574588
Cost after epoch 40: 0.506523
Cost after epoch 50: 0.466331
Cost after epoch 60: 0.427205
Cost after epoch 70: 0.429317
Cost after epoch 80: 0.459113
Cost after epoch 90: 0.388603
Confusion Matrix:
[[113  94]
 [139 381]]
Train Accuracy: 0.8833333
Test Accuracy: 0.6795048
C:\Users\npb16184\anaconda3\envs\tensorflow-gpu\lib\site-p
Function plot_roc_curve is deprecated; This will be remove
  warnings.warn(msg, category=DeprecationWarning)
```

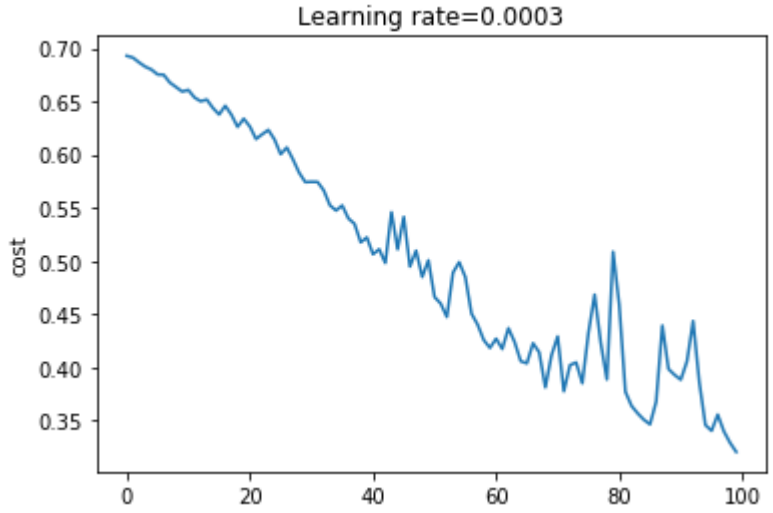


Figure 5.5 The loss with training process

Table 5.1 Confusion Matrix of testing dataset

	True: healthy	**False: pathological**
Prediction: healthy	113	139
Prediction: pathological	94	381

Table 5.2 The performance of 10-layer CNN network for PVD

SN(r)	***SP***	***p***	***F1***	***ACC***
54.59 %	73.27%	80.21%	76.58%	67.95%

It can be seen from Table 5.2 that the classifier achieved overall accuracy (ACC) is 67.95%. Compared to Pavol's work [11], the spectrogram features show comparable performance to it on raw speech data. Nevertheless, it uses a smaller network architecture which reduces the computations and memory required for the training. The ROC curve of the 10-layer CNN model are shown in Figure 5.6.

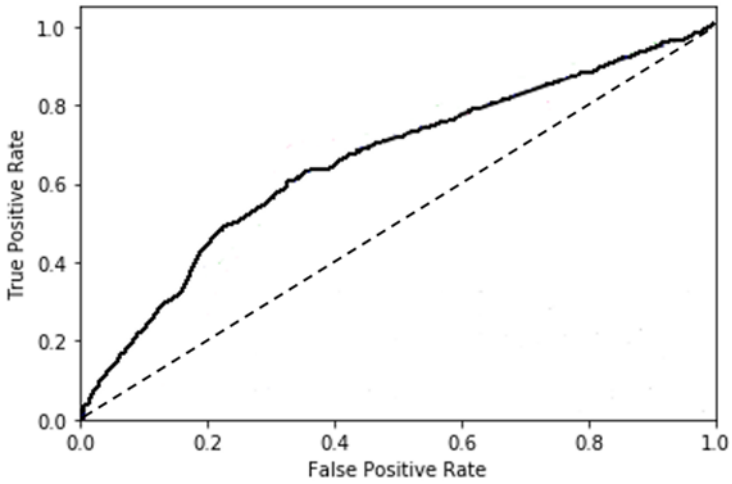


Figure 5.6 ROC curve of the 10-layer CNN model

The 10-layer CNN model trained from scratch is the first exploration of 2D representations in pathological voice detection field using deep learning techniques. This result shows that the spectrogram can be used effectively as the input to classify pathological voice and normal voice, without necessity to extract features manually. However, the training data accuracy is still much higher than testing data accuracy, which reveals a degree of overfitting.

Overfitting problems is commonly seen in small-data-size problems. In Section 3.2.2, we reviewed that there are generally two methods of addressing small-data-size problem. One way is using transfer learning to transfer one model from other applications into a different application. Another is to use data augmentation to generate more data for training. In the following section, we explore the transfer learning using state-of-the-art CNN models from image recognition field to pathological voice detection field.

5.3 Transfer Learning exploration

The model we proposed in section 5.2 is a CNN model that trained from scratch. However, deep learning requires a large amount of data for training to obtain good generalization ability. Since the amount of pathological data is limited, it may not be sufficient for training a robust generalized model for deep learning. In this work, we transferred state-of-the-art CNN models from image recognition field to pathological voice detection field. In this section, we will first give the overview of transfer learning block diagram flow, then describe the time-frequency representation inputs in detail. Moreover, the architectures of four CNN models trained from image classification field and altered for pathological voice detection will be illustrated.

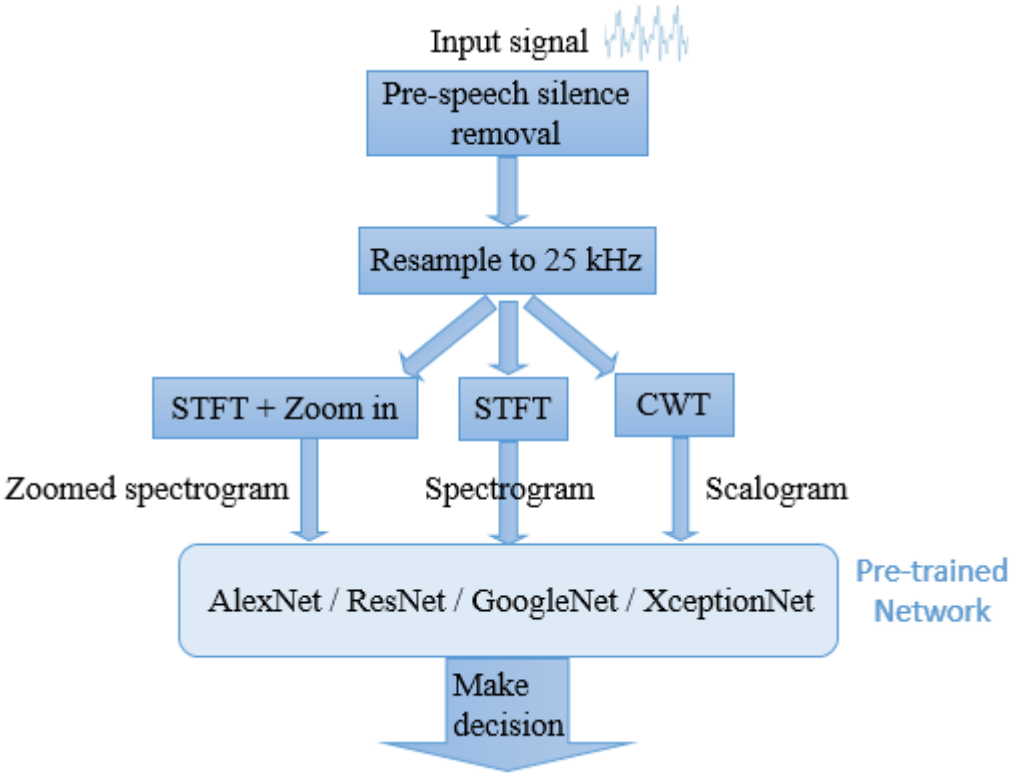


Figure 5.7 Transfer Learning architecture for pathological voice detection

The transfer learning architecture is shown in Figure 5.7. Firstly, the signal is resampled to 25 kHz, then the signal is transformed into three types of 2D representation input: spectrogram, zoomed spectrogram, and scalogram. These 2D representations are respectively fed into the pre-trained CNN networks for transfer learning. We explored four state-of-the-art CNN networks: AlexNet, GoogleNet, ResNet and XceptionNet.

5.3.1 Time-frequency representation inputs

A. Spectrogram

After pre-speech silence removal, the speech corpuses are resampled 25 kHz. Afterwards, the signal is decomposed with short-time Fourier transform by applying 20ms windows every 15 ms (75% overlapping). Hamming window is used. This window size and overlapping percentage is different from the one that proposed in section 5.2, because the 2D input to AlexNet is set to 227*227*3. In this case, the pre-processing parameters are adjusted for transfer learning.

In chapter 2.4, we reviewed perceptual techniques generally used in pathological voice detection. GRBAS is a classical tool for measuring the severity of dysphonia by perceptual analysis. The PdA database provides GRBAS ratings that describe the severity of dysphonia and specific pathological features, so that it will give us a more reliable overview of pathological features on 2D representations with different state of severities. Some examples of spectrogram standard input in two different categories are shown in Figure 5.8. In order to illustrate the formants clearly, we show the spectrogram from 0 to 3kHz. It can be seen the severe pathological voice with GRBAS grade 11 in Figure 5.8 (d) shows more irregular patterns. The normal voice with GRBAS grade 0 in Figure 5.8 (a) shows significant and continuous patterns. However, we can still see that mild pathological voice with GRBAS grade 2 in Figure 5.8 (c) diagnosed of vocal cord polyp shows smoother and less irregular patterns than normal voice with GRBAS grade 2 in Figure 5.8 (b). This is the challenge that we described as “weak labels”. As mentioned before, these weak labels reveal that the dataset has overlap between normophonic voice and pathological voice. This will lead to confusion when training classifiers.

An example of spectrogram and zoomed spectrogram are shown in Figure 5.9. The onset of the speech in Figure 5.9 (b) was shown to be important in detecting pathological features [101]. It can be seen that Figure 5.9 (b) contains some voice breaks shown in the spectrogram, and the formants exhibits discontinuous patterns compared to Figure 5.9 (a). The resulting spectrograms and zoomed spectrograms are all resized to 227*227*3 for transfer learning.

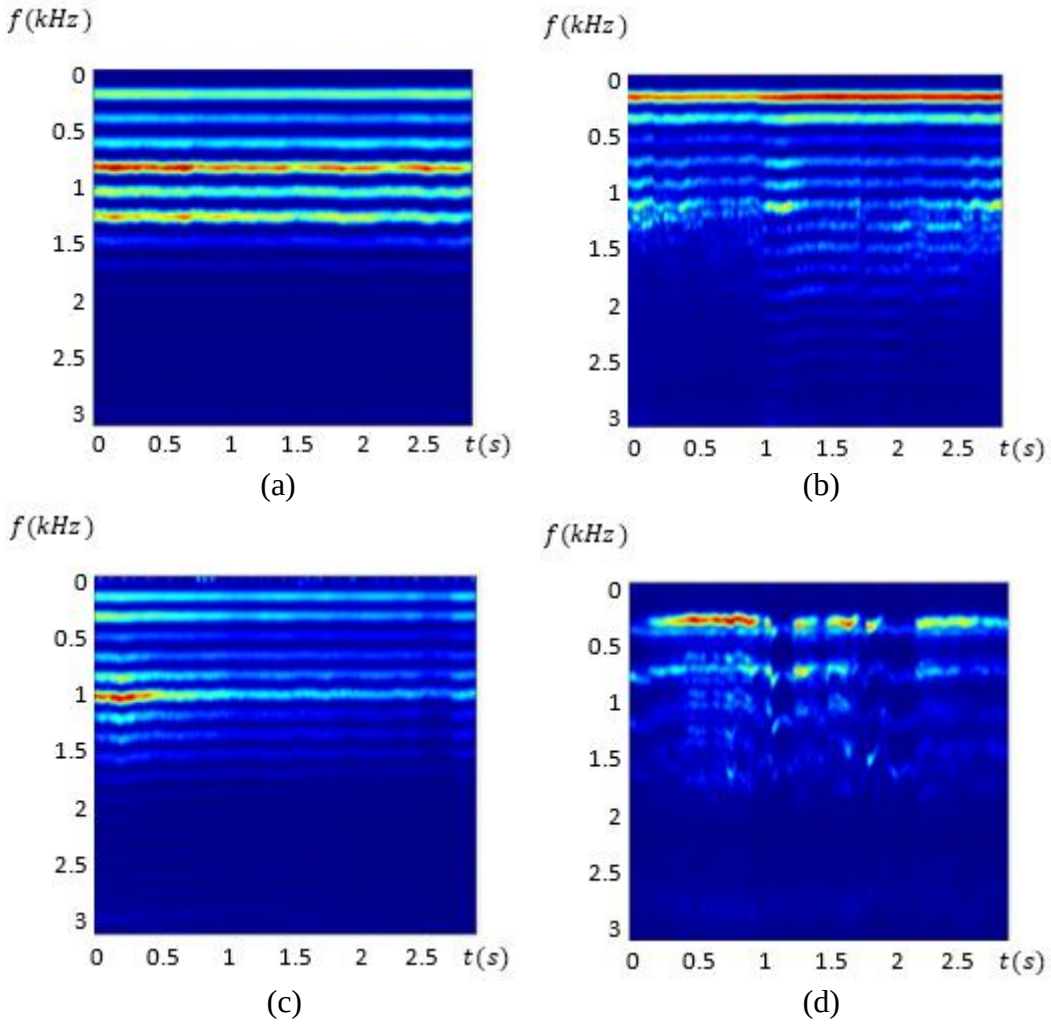


Figure 5.8 Spectrogram of (a) Normal voice (asra.wav) (GRBAS grade: 0) (b) Normal voice (Mtaaa2.wav) (G:0, R:0, A:1, B:1, S:0, Grade:2) (c) Pathological voice (Polipo11.wav) (padiculated polyp) (G:0, R:0, A:1, B:1, S:0, Grade:2) (d) Pathological voice (rpjbis.wav)

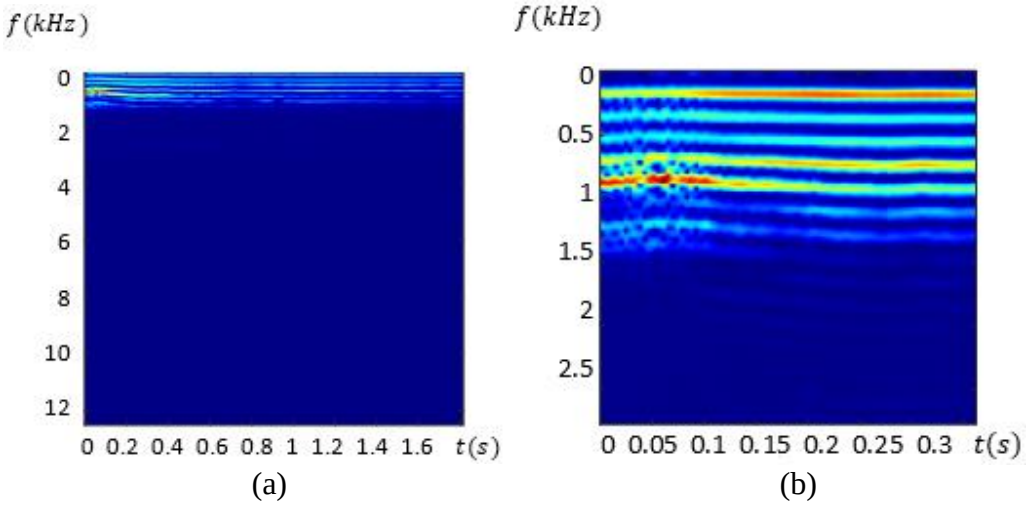


Figure 5.9 (a) spectrogram (b) zoomed spectrogram

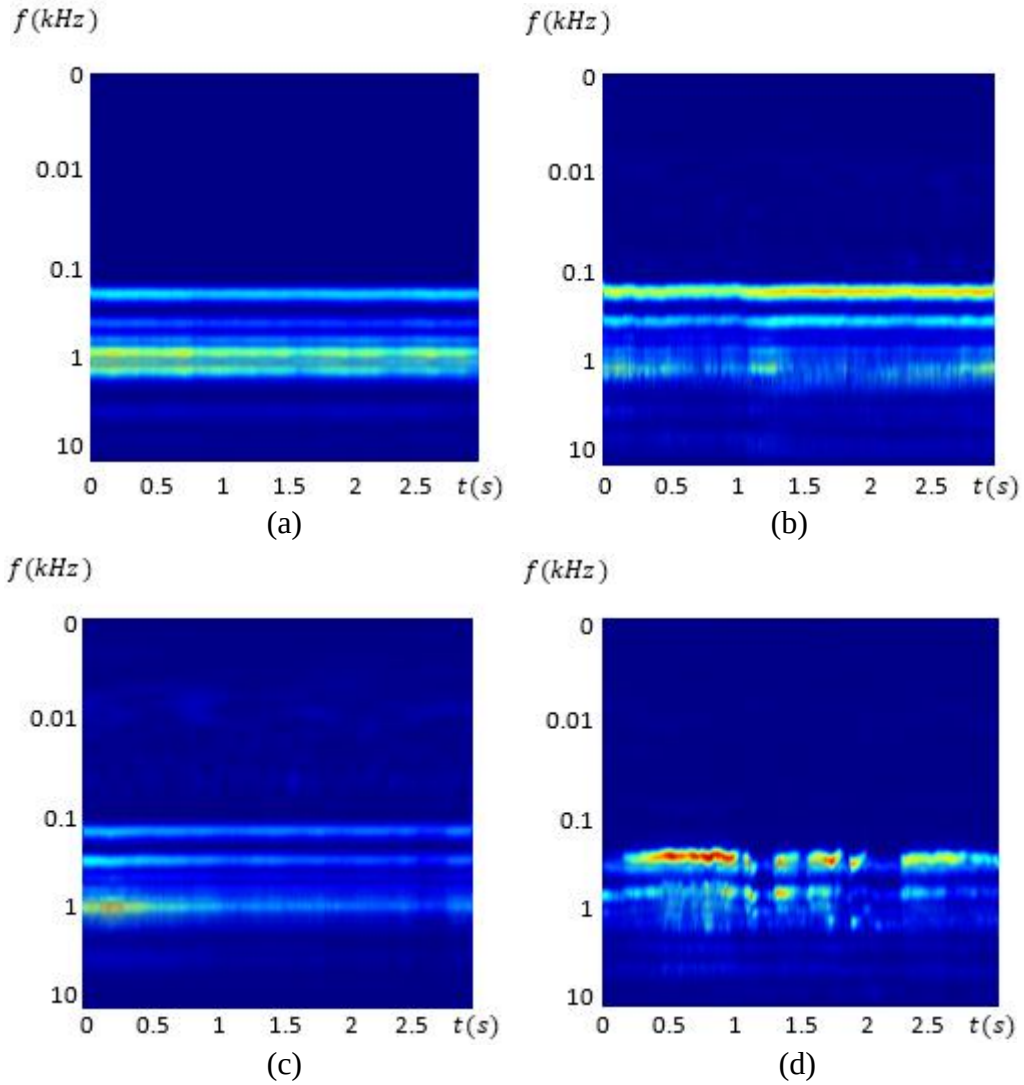


Figure 5.10 Scalogram of (a) Normal voice (asra.wav) (GRBAS grade: 0) (b) Normal voice (Mtaaa2.wav) (G:0, R:0, A:1, B:1, S:0, Grade:2) (c) Pathological voice (Polipo11.wav) (padiculated polyp) (G:0, R:0, A:1, B:1, S:0, Grade:2) (d) Pathological voice (rpjbis.wav)

B. Scalogram

We use the continuous wavelet transform to obtain the scalogram as another type of 2D representation for transfer learning. Firstly, the pre-speech silence part is removed and the signal is resampled to 25 kHz. A continuous wavelet transform of the whole signal is computed obtain the scalogram. The wavelet is Morse wavelet, with the default symmetry parameter (gamma) as 3, and time-bandwidth product equal as 60. The frequency response decays to 50% of the peak magnitude at the Nyquist. Finally, the scalogram is resized to 227 by 227 by 3. Typical scalograms of the same voice examples are shown in Figure 5.10. It can be seen that the pathological voice with GRBAS grade 11 in Figure 5.10 (d) shows more irregular

voice break patterns than normal voice with GRBAS grade 0 in Figure 5.10 (a). While scalogram of pathological voice with GRBAS grade 2 in Figure 5.10 (c) shows more irregular patterns than spectrogram in Figure 5.10 (c).

5.3.2 Transfer learning experiments

A. AlexNet

The altered AlexNet Structure is shown in Figure 5.11. The original AlexNet [299] is a baseline CNN network designed with 227*227*3 input size, so that there is no augmentation required to be done to the inputs. The last three layers are changed to a fully-connected set of 3 layers. The first two layers contain 4096 nodes each and the last layer has two nodes, representing pathological voice and healthy voice. Initial learning rate is set as 10^{-5}, and the mini batch size is 40, maximum epochs is set to 20.

B. GoogleNet

GoogleNet [349] is a 22 layers deep network with repeated inception blocks. The altered GoogleNet Structure is shown in Figure 5.12, there are nine similar inception blocks followed by three baseline convolutional layers. Each inception block contains four branches with different type of filters. The input size of this architecture is 224 by 224 by 3, so that we conduct augmentation to the inputs first. The last four layers are removed, and change them with a new dropout layer, a fully connected layer, a softmax layer, and an output layer.

C. ResNet-101

Original Residual network [350] is a structure designed with 152 layers, which is far more deeper than traditional CNN networks. It contains same first convolutional layer architecture as GoogleNet, then followed with two types of residual learning blocks. The essential idea behind residual block is skip connections. It makes neural network easier to train when going deep and make the neural network more dynamic. In this work, we use ResNet-101, which is a variation of ResNet with 101 layers. There are overall 33 residual blocks. Similarly, we remove the last three layers, and adding a fully connected layer, a softmax layer, and the classification output layer at the end of the network. The altered ResNet-101 Structure is shown in Figure 5.13.

D. Xception Network

Xception network [351] is 71 layers deep and is a variation on GoogleNet and ResNet. It outperforms inception V3 on the ImageNet dataset with smaller number of parameters. The altered Xception Network Structure is shown in Figure 5.14. There are three types of Xception block with skip connections in this architecture. Similarly, we augment the input size to 299*299*3, and remove the last three layers, then add a new fully connected layer, a Softmax layer and a output layer at the end of the network.

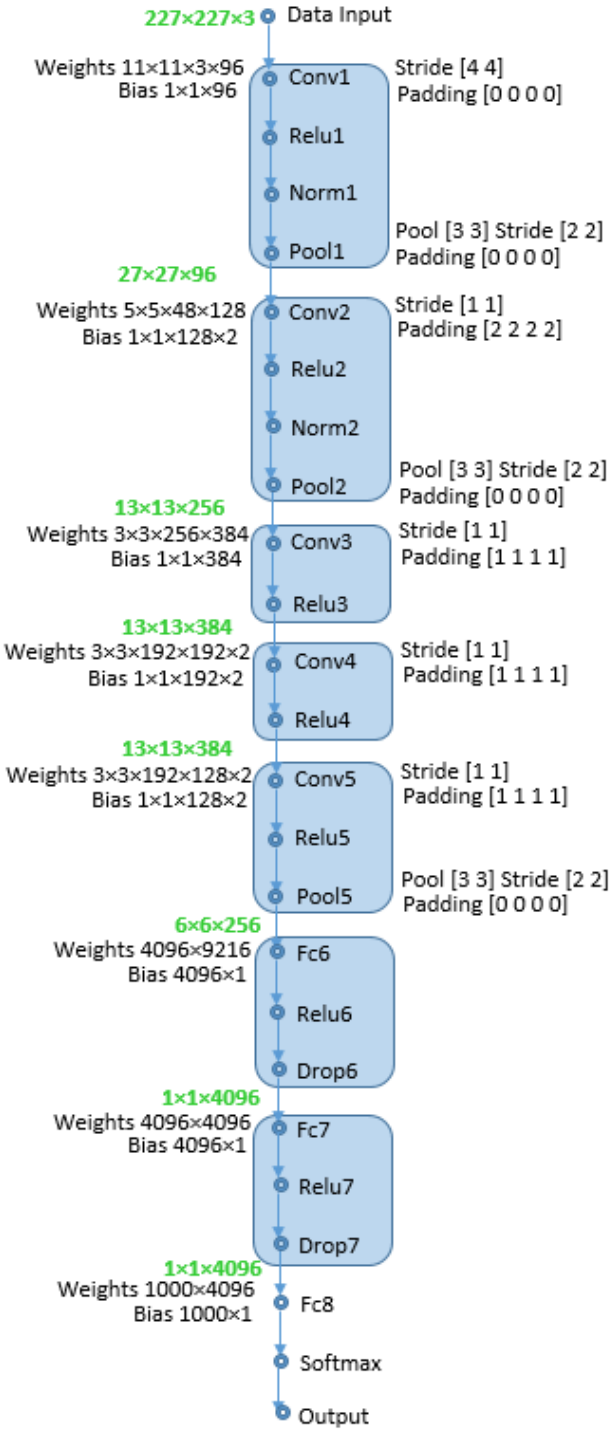


Figure 5.11 Altered AlexNet architecture for pathological voice detection transfer learning

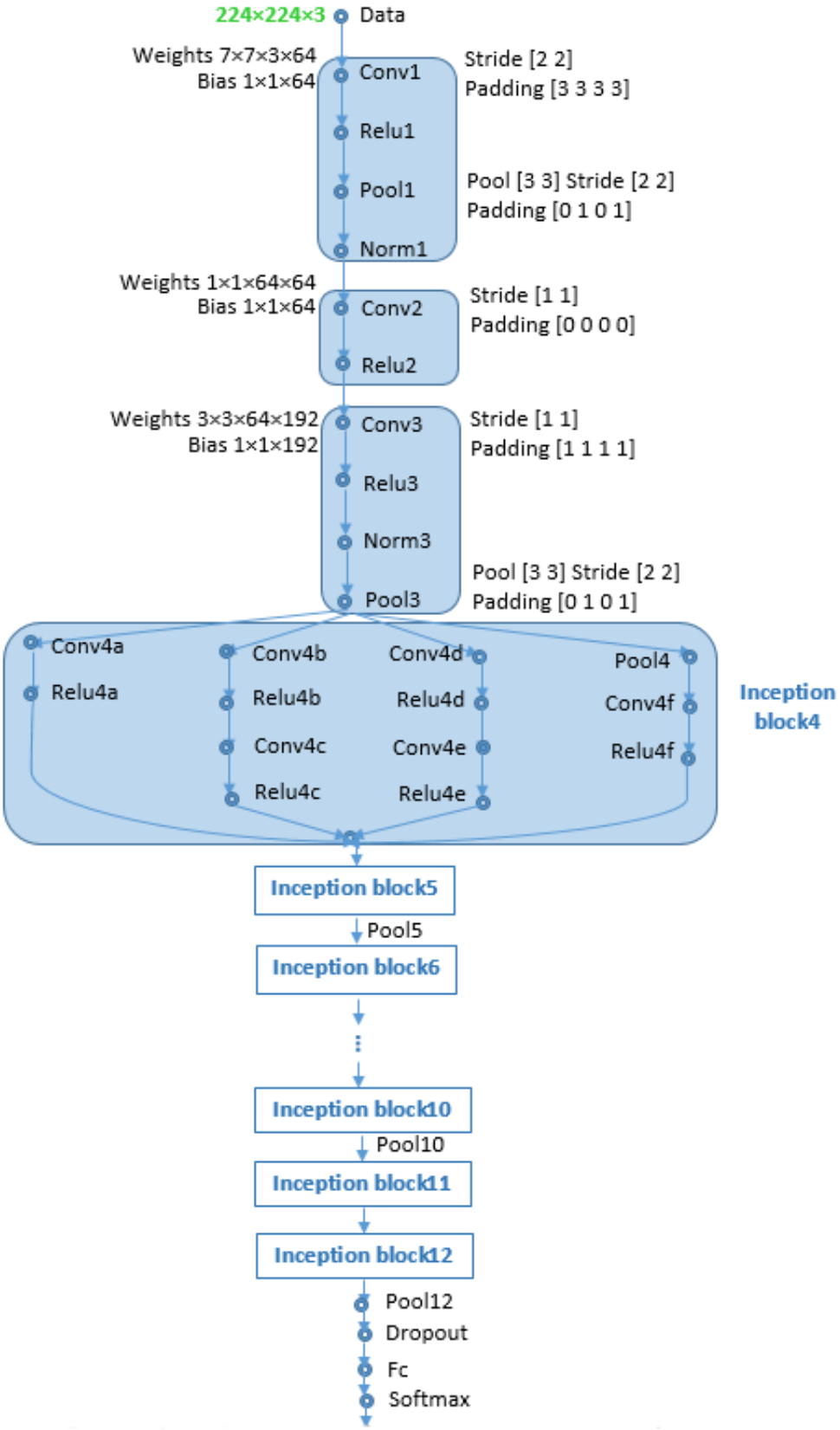


Figure 5.12 Altered GoogleNet architecture for pathological voice detection transfer learning

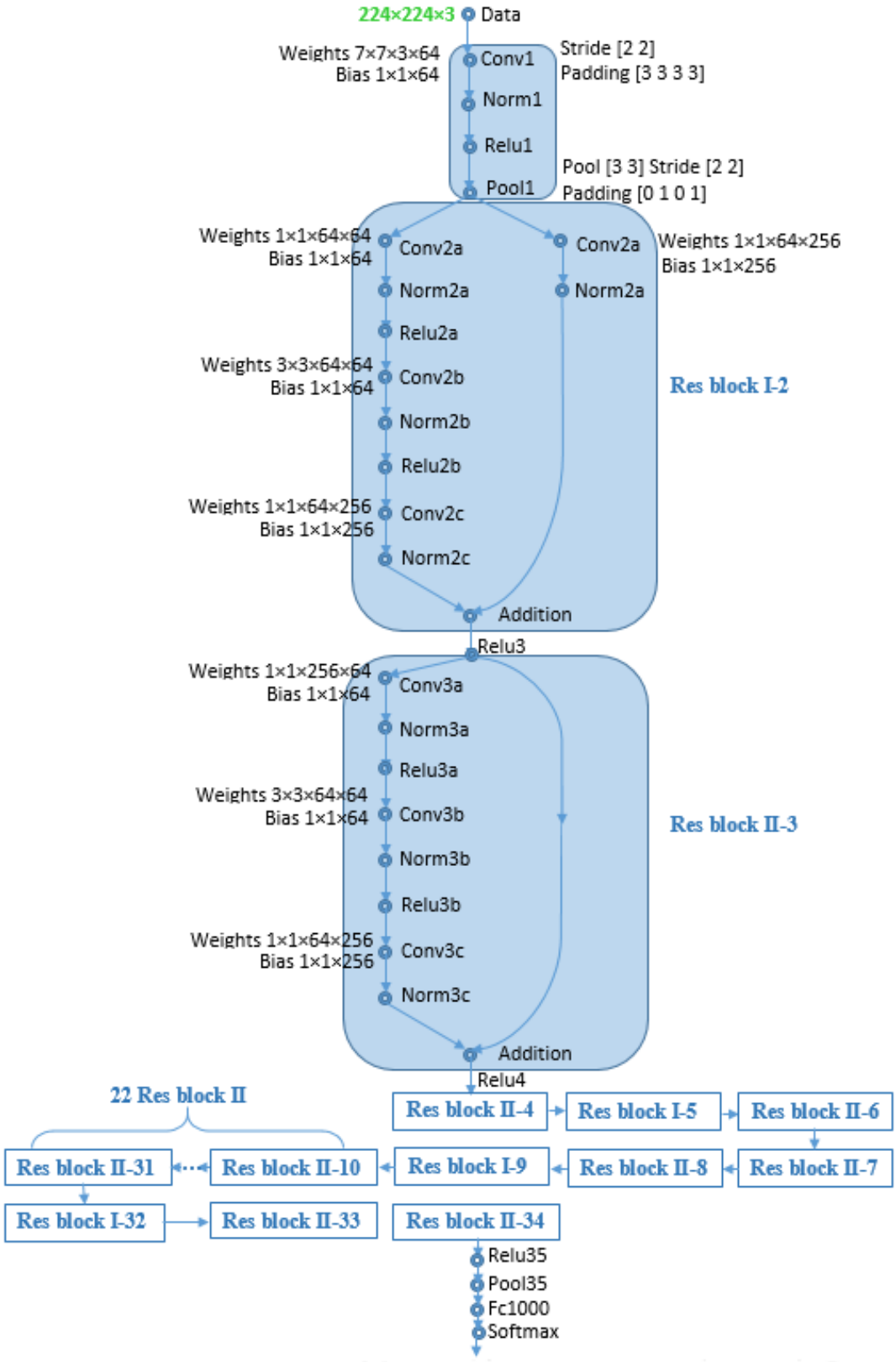


Figure 5.13 Altered ResNet architecture for pathological voice detection transfer learning

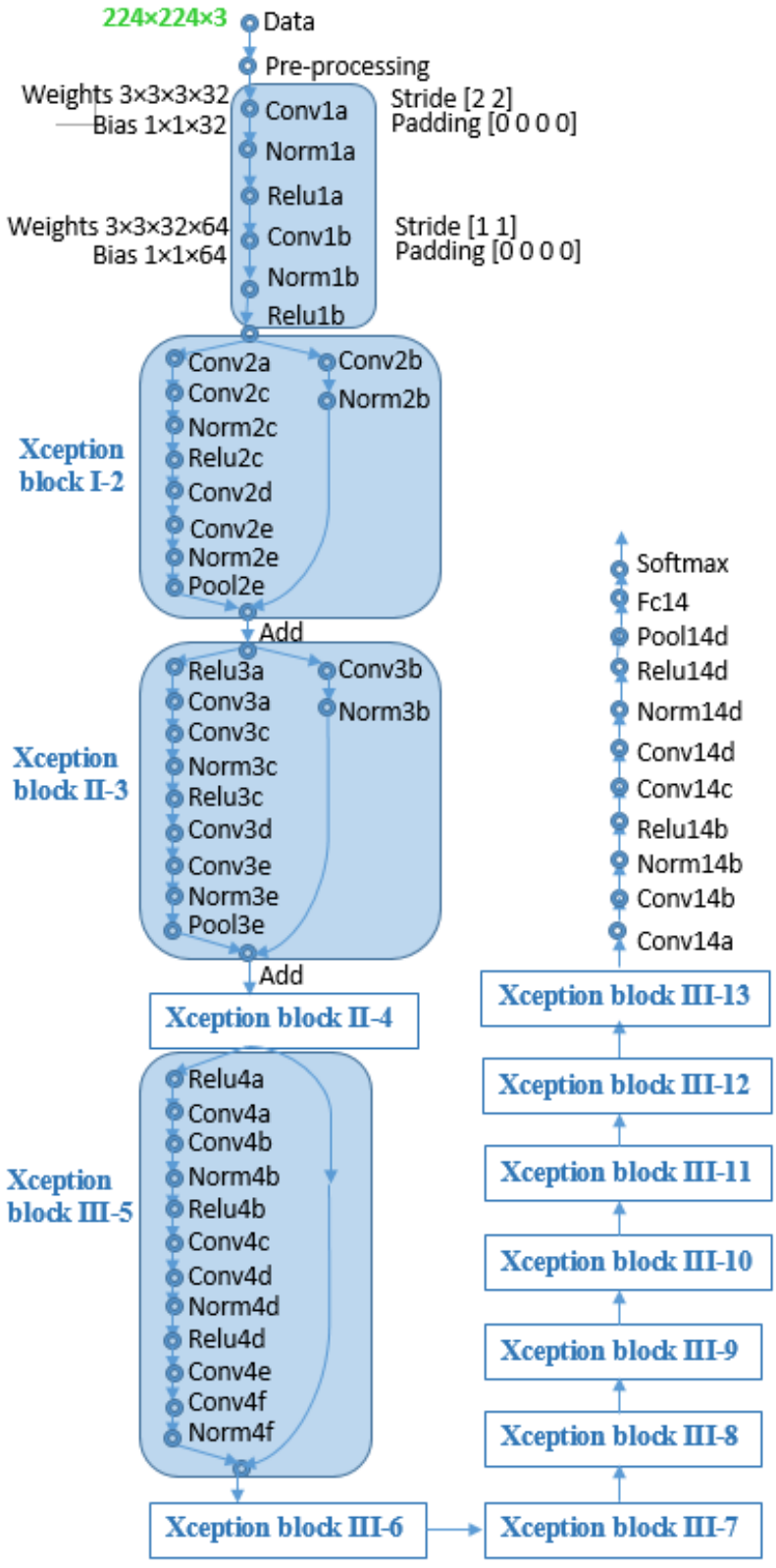


Figure 5.14 Altered XceptionNet architecture for pathological voice detection transfer learning

5.3.3 Results and discussion

The experiments are conducted on three databases for pathological voice detection and the whole databases are used. The train-test split is 70% to 30%. SVD database: there are 960 training data (480 healthy and 480 pathological), and 727 testing data (207 healthy and 520 pathological). PdA database: there are 280 training data (140 healthy and 140 pathological), and testing data amount is 160 (99 healthy and 61 pathological). AVPD database: there are 232 training data (116 healthy and 116 pathological), and 107 testing data (51 healthy and 56 pathological).

Spectrogram, zoomed spectrogram and scalogram are all resized to match the input size for transfer learning. The AlexNet classification results are shown in Table 5.3. The severity of overfitting problems can be evaluated by the difference of training accuracy and testing accuracy. It can be seen that the training accuracy is around 10% higher than the testing accuracy, which means a certain degree of overfitting problem exists. The best testing accuracy achieved is applying scalogram for transfer learning in AVPD database, with 81.37%. However, it performs relatively poor on SVD database, with the highest accuracy 69.57% using zoomed spectrogram as input. AlexNet transfer learning result did not show better advantage than training a CNN model from scratch.

Table 5.3 Classification result report of AlexNet transfer learning

Accuracy		AlexNet		
Type of input representation	Train vs. Test	SVD	PdA	AVPD
Spectrogram	Train	77.50%	80.00%	80.00%
	Test	68.58%	70.45%	76.47%
Zoomed spectrogram	Train	75.00%	80.00%	77.50%
	Test	69.57%	72.73%	73.53%
Scalogram	Train	87.50%	77.50%	90.00%
	Test	67.80%	74.24%	81.37%

The GoogleNet classification results are shown in Table 5.4. Since GoogleNet architecture contains inception blocks, which is a more complex topology of the network, the overfitting problem becomes even more severe than AlexNet. The best

performance is the scalogram applied on PdA database, with 81.06% accuracy. SVD database remains to be the most challenging database with highest accuracy of 71.74% using zoomed spectrogram input.

Table 5.4 Classification result report of GoogleNet transfer learning

Accuracy		GoogleNet		
Type of input representation	Train vs. Test	SVD	PdA	AVPD
Spectrogram	Train	90.00%	100.00%	97.50%
	Test	68.97%	69.70%	70.59%
Zoomed spectrogram	Train	92.50%	100.00%	95.00%
	Test	71.74%	80.30%	79.41%
Scalogram	Train	100.00%	95.00%	100.00%
	Test	71.54%	81.06%	79.41%

The ResNet classification results are shown in Table 5.5. The overfitting problem is more severe than GoogleNet, while the performance achieves significantly better than the former two networks. It achieves 84.31% accuracy in classifying AVPD database with zoomed spectrogram as input. It also achieves better performance in the most challenging SVD database, with 73.12% accuracy using spectrogram as input. This shows the superiority of skip connection structure in residual block compared to AlexNet and GoogleNet.

Table 5.5 Classification result report of ResNet transfer learning

Accuracy		ResNet-101		
Type of input representation	Train vs. Test	SVD	PdA	AVPD
Spectrogram	Train	100.00%	100.00%	100.00%
	Test	73.12%	75.00%	76.47%
Zoomed spectrogram	Train	100.00%	97.50%	100.00%
	Test	69.37%	78.03%	84.31%
Scalogram	Train	100.00%	100.00%	100.00%
	Test	70.75%	80.30%	80.39%

The Xception Net classification results are shown in Table 5.6. Xception Net is the variation of ResNet, with similar residual block but lighter network structure. It shows 84.09% accuracy in PdA database with spectrogram as input, while the performance on SVD database is not as good as ResNet.

Table 5.6 Classification result report of XceptionNet transfer learning

Accuracy		XceptionNet		
Type of input representation	Train vs. Test	SVD	PdA	AVPD
Spectrogram	Train	100.00%	100.00%	100.00%
	Test	69.57%	84.09%	79.41%
Zoomed spectrogram	Train	97.50%	100.00%	100.00%
	Test	69.57%	75.76%	82.35%
Scalogram	Train	100.00%	100.00%	100.00%
	Test	69.76%	80.30%	73.53%

It can be seen from these results that training accuracy appears to be higher than testing accuracy in all experiments. Furthermore, overfitting problems appears much more severe with GoogleNet, Xception Net, and ResNet, rather than AlexNet. AlexNet is 8-layers deep with a baseline CNN structure; GoogleNet is a 22-layer deep network that contains inception layers; ResNet is a 152-layer network with residual learning blocks; Xception Net is a 71-layer network based on variations on GoogleNet and ResNet. AlexNet represents a much simpler architecture and less number of layers, with less overfitting issues. However, GoogleNet, Xception Net and ResNet performed with higher accuracy on testing data on at least two databases. For example, GoogleNet achieved 81.06% and 79.41% accuracy using scalogram on the PdA database and AVPD database respectively; Xception Net achieved 80.30% accuracy using a scalogram on the PdA database, and 82.35% accuracy using a zoomed spectrogram on AVPD database; ResNet-101 achieved 80.30% accuracy using scalograms on the PdA database and 84.31% accuracy using zoomed spectrograms on AVPD database. Scalograms and zoomed spectrograms are shown to be working better than spectrograms in most cases. It can also be seen that the scalogram achieved better results on the PdA database most of the times, and the

zoomed spectrogram performed better on the AVPD database in general. This might be due to the characteristics of different databases.

Among these networks, it can be seen that the ResNet-101 achieved the overall best performance on the three databases. This is due to the high efficiency of the residual blocks. However, the SVD database appears to be much more challenging compared to other databases. The highest accuracy on SVD database with transfer learning is by ResNet, which achieved 73.12% accuracy with spectrogram representation. This is because that SVD contains much more data amount than other two databases, and it is a fusion of 71 pathological types of data. Therefore, it is suggested that SVD is a much more challenging database that requires the network for higher generalization ability.

When comparing the transfer learning performance with the 10-layer CNN model proposed in Section 5.2, it did not solve the overfitting problem although the testing data performance is improved. When applying more complex networks in pathological voice detection (such as ResNet and XceptionNet), the overfitting problem is even more severe.

5.4 Conclusion

In this chapter, we proposed two type of ways in applying CNN in pathological voice detection field with time-frequency representations, and both achieved successful results. One is a novel 10-layer CNN model designed for training from scratch, another one is the transfer learning with four state-of-the-art CNN networks from image classification field. Compared to the CNN network trained from scratch, transfer learning shows superiority in testing accuracy. Although transfer learning experiments improves the overall performance, the over-fitting problem becomes even more severe with very-sophisticated CNN architecture.

It was seen that it is hard for transfer learning to address over-fitting problems. In addition, the balance of good performance and reducing overfitting problem is a dilemma. In the next Chapter, we propose a model based on data augmentation idea, which addresses the over-fitting problem and improve the overall performance.

Chapter 6

6 R-Net

6.1 Introduction

Chapter 4 and Chapter 5 explored different approaches of deep learning techniques for pathological voice detection. However, over-fitting problem were evident in both which is due mainly to the data limitation in pathological voice detection. In this chapter, a novel method R-Net, for reducing the overfitting problem while using deep learning training with limited amount of data is presented. Following a description of the R-Net system, the importance of speech onset and pre-speech onset silence is described. Then we describe the pre-processing step with pre-speech silence removal algorithm. Experimental results of R-Net are presented in Section 6.3. These include a comparison to traditional CNN models. In addition, the trained R-Net from SVD database is applied for transfer learning in detecting paediatric recurrent respiratory papillomatosis (RRP). RRP is also a type of laryngeal papilloma as illustrated in Section 2.2.1. It validates its superiority for the diagnosis of RRP.

6.2 The R-Net system

The most useful way of generalizing a deep learning model is to train it on more data. Since the number of pathological data is limited (due to expensive cost and time in data collection), a way to create “fake data” is to train the model using data augmentation. Data augmentation has been demonstrated to be useful in classification applications such as object recognition. For example, affine transformations such as rotation or translation, even clipping, zooming, or contrast or illumination changes are all common data augmentation techniques. Injecting noise in the input to a neural network [352] can also be a form of data augmentation. De-noising auto encoder model [245] is a successful example of unsupervised learning algorithm with injected input noise. Adding noise with a small variance at the input of the model has a similar effect of regularization [353]. The regularization

effect is much more powerful than simply reducing the size of the network. There are some other ways of adding noise such as adding it to the weights or to the output targets.

The principle of the novel R-Net is that it combines the idea of translation and characteristics on the onset of the speech for a data augmentation method. The aim of it is to improve the generalization ability of the model, and thus successfully reducing the overfitting problem.

6.2.1 Speech onset definition

The composition of a typical speech signal corpus $s[i]$ is shown in Figure 6.1. The pre-speech silence time T_{sil} is defined as the time before the speech onset point. It normally contains background noise during the speech recording process. The speech onset time T_{aper} is the period after the speech onset point to a point where the vocal cords are functioning pseudo periodically. During the speech onset time, the voice signal will change from aperiodic status to periodic status. Some research works report that speech onset time is important for detecting pathological features[101]. The duration of periodic time T_{per} in Figure 6-1 shows periodic patterns, indicating vocal cords vibrating pseudo periodically.

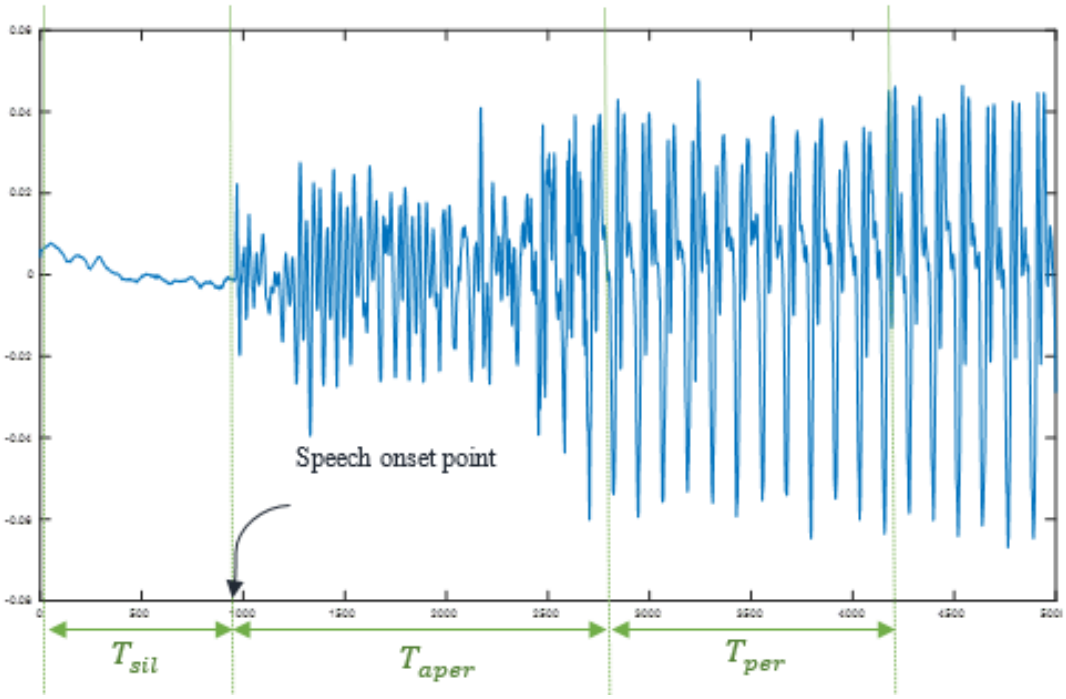


Figure 6.1 Speech onset silence and speech onset (voice sample called “A1-cyt-f-52-no-2-3119kac-p” from AVPD database)

6.2.2 Pre-speech silence removal

A "pre-speech silence removal" algorithm to eliminate pre-speech silence was developed and illustrated in Figure 6.2. First, the speech signal with sampling frequency 25kHz is segmented into frames. In this work, the frame size is set 50 samples. A 'Speech Ratio' is defined as the proportion of real speech in $s[i]$ compared to noise in each frame and can be expressed as:

$$M[i] = \begin{cases} 1 & if\ s[i] > 0.2 \times \max(s[i]) \\ 0 & otherwise \end{cases} \quad (6.1)$$

$$Speech\ ratio = \frac{\sum_{i=1}^{N} M[i]}{N}, i = 0,1, \dots N-1 \quad (6.2)$$

where N represents a whole number of data in each frame. In this case, N equals to the length of frame size 50. $M[i]$ is a measure of "valid speech frame". It equals to 1 when $s[i]$ has values larger than $0.2 \times \max(s[i])$, which is classified as real speech.

The speech ratio in each frame is computed to form the "speech ratio" distributions graph. Using a threshold as 0.1, the first frame which contains the speech ratio larger than 0.1 will be defined as the speech onset frame. The frames that occur before this speech onset frame are silent frames and are removed.

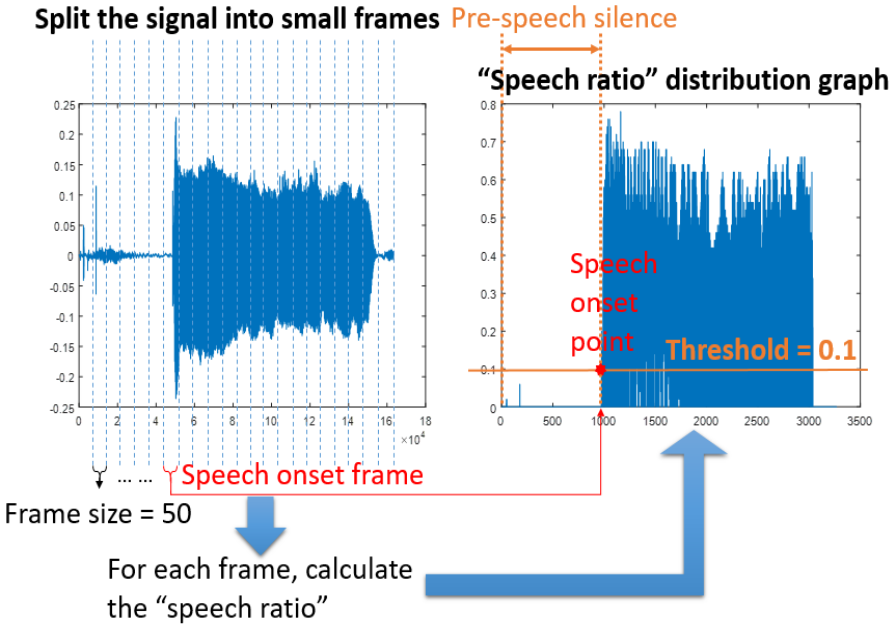


Figure 6.2 Speech onset silence cutting process

Voice activity detection (VAD) technique are generally applied for activating noise classification or estimation and switching on/off for smartphones [354-356].

Traditional method of VAD is statistical modelling. G.729 Annex B (G729.B) [357] uses a fixed threshold for features such as line spectral frequencies, full-band energy, low-band energy, and zero crossing difference etc. It is seen as a standard for silence compression that is generally used in Voice over Internet Protocal (VoIP). Other VAD applies likelihood radio test (LRT) on the speech, and sees the DFT coefficients of the noise and speech as random variables [358, 359]. In [360], it models the DFT coefficients as generalised Gamma distribution. Apart from these statistical modelling methods, machine learning methods are also been employed. In [361], it uses SVM to decide the threshold on the features same as it in G729.B. Some combines SVM with statistical models such as LRT [362]. In addition, deep learning methods have also been explored in recent years. RNN is applied with 13-dimensional perceptual linear prediction (PLP) features of speech in [363], and CNN is explored with log-mel spectrogram and its delta and acceleration coefficients [364].

Although there are many VAD techniques exist in the literature, realistic scenarios is complex to specific applications. Our proposed VAD method and the parameters are designed specifically for the three databases, so that it achieved successful performance on almost all the databases, which is suitable for this pathological voice detection scenario.

6.2.3 Architecture of R-Net

The architecture of the novel R-Net is shown in Figure 6.3. The input speech is fed into two channels, a training channel and a validation channel. Both channels conduct resampling procedure and converts the speech in scalogram using the Continuous Wavelet Transform (CWT) as described in Chapter 5.3.1. In the transfer learning in Chapter 5.3, the scalogram are resized to 227 by 227 by 3 to fit the input of AlexNet. In R-Net, there is no requirements to this size of the representation.

The only difference between the two channels is that one channel obtains pre-speech onset silence removal at start while the other one does not. As indicated in Figure 6.3, the first channel converts the speech into scalogram without pre-speech onset silence removal, and the representations obtained from this approach is defined as “raw input”. The CNN is trained with these raw representations. Then the second channel converts the speech into scalogram with speech-onset-silence removed, and

the scalogram in second channel is fed into the trained network as the "Real input" to help the CNN system to make decisions.

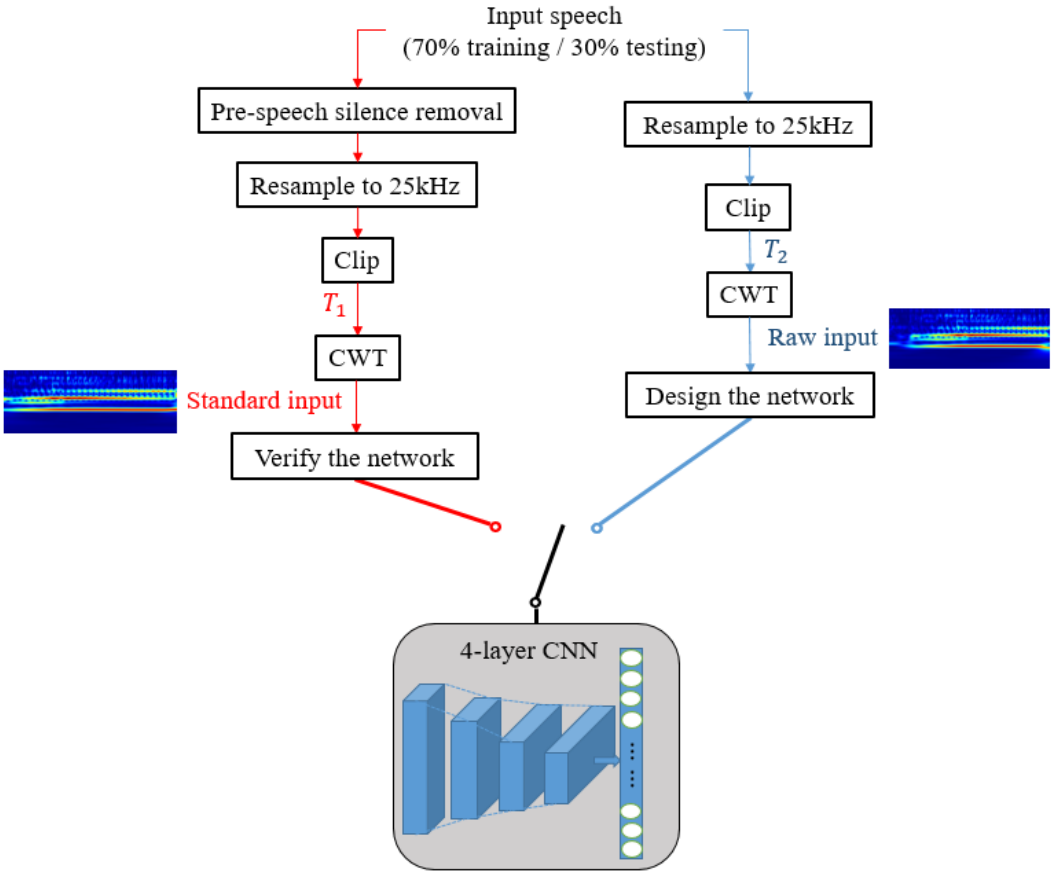


Figure 6.3 The architecture of "R-Net"

An illustration of the standard input representation and a raw input representation are shown in examples in Figure 6.4. Since it is reported that speech onset time is important for detecting pathological features [101], the speech onset are emphasized in this model. The resampling frequency is 25 kHz. The time section of standard input and raw input are annotated as T_1 and T_2. The speech onset 0.12s is cropped for analysis. (0 to 3000 point). It can be seen that 0.12s contains all the information including T_{sil} and T_{aper} and T_{per} in Figure 6.4 (a).

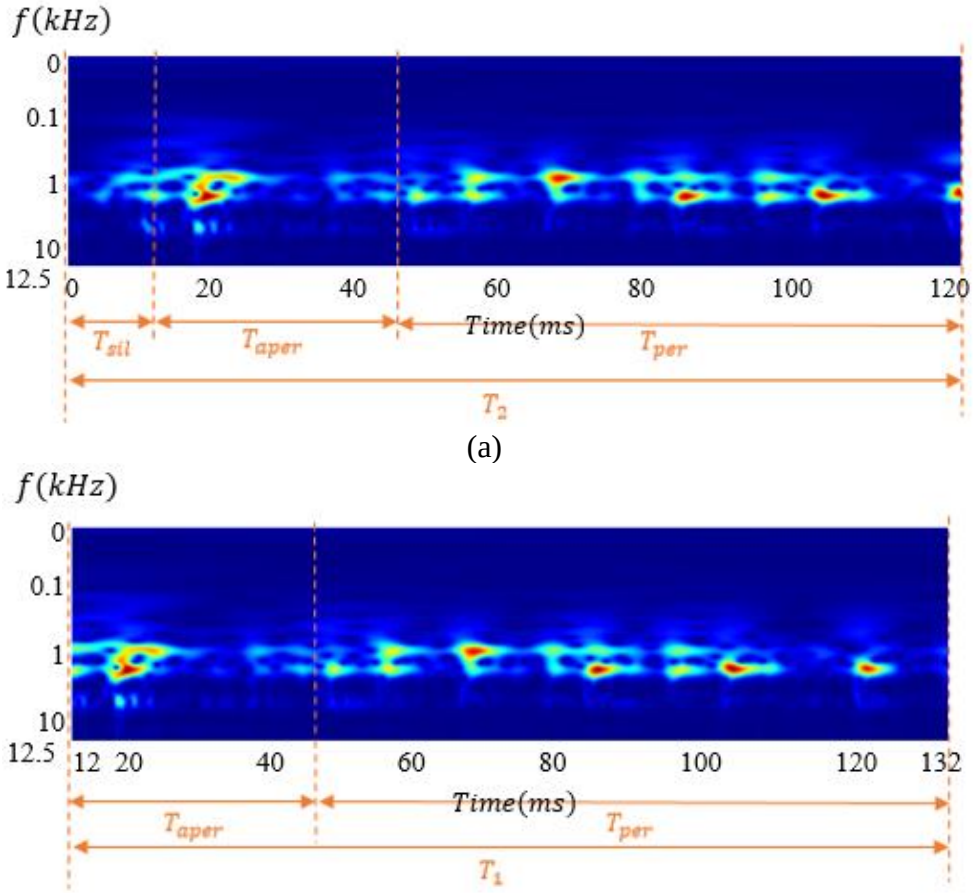


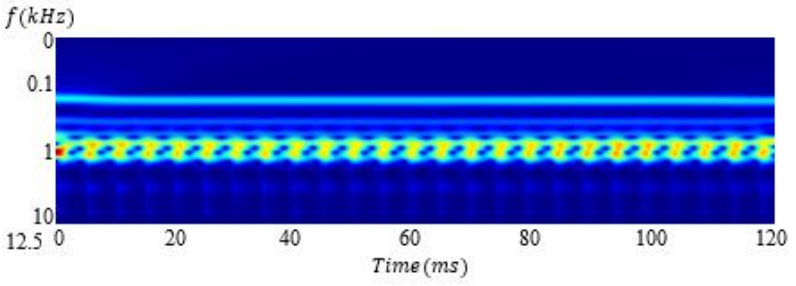
Figure 6.4 (a) raw input (901-a_n.wav) (b) standard input (901-a_n.wav)

The essence of R-Net is that the network is trained with the "augmented data" - raw input that includes the pre-speech silence. In this case, the network is designed with appropriate architecture and robust parameters. Afterwards, the network verify the decisions on standard input representations without the pre-speech silence.

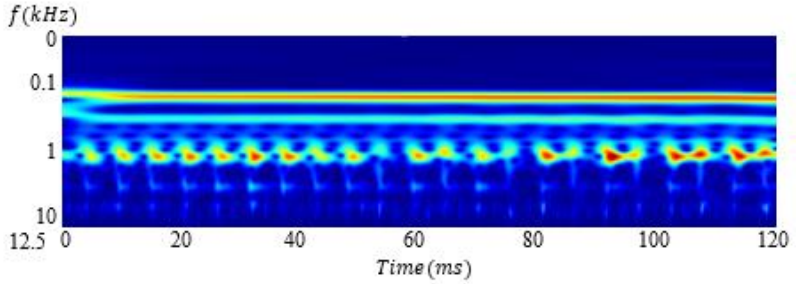
This mimics the noise-injection model which adds the randomness to the system. That makes the model more robust and accurate on new data. It was noted in [365] that adding random noise to the training data can improve generalization and fault tolerance ability. Jim et.al [366] demonstrated the effectiveness of noise in convergence and generalization of recurrent neural networks (RNN); Graves et.al [353] applied weight noise (the noise on the parameters) per training sequence in long short-term memory (LSTM), and claimed that weight noise 'simplify' neural networks which helps to improve the generalization. One reason behind these observations is that the addition of noise is shown to have the similar effect of regularization. Furthermore it is noted that adding noise to the input increases the

sample in the domain of input space, which can be seen as a form of data augmentation [6].

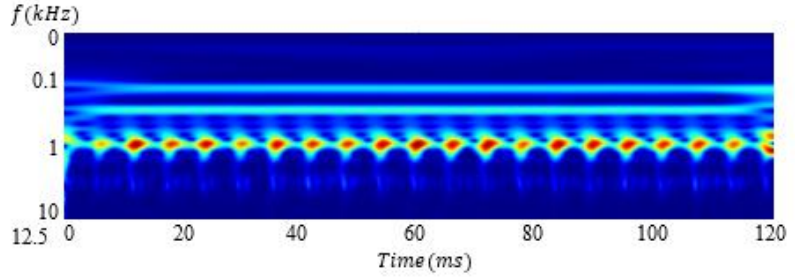
Some examples of scalograms of the standard input in two different categories are shown in Figure 6.5. The normal voice with GRBAS grade 0 in Figure 6.5 (a) shows significant and continuous patterns. It can be seen that the severe pathological voice with GRBAS grade 11 in Figure 6.5 (d) shows more irregular patterns.



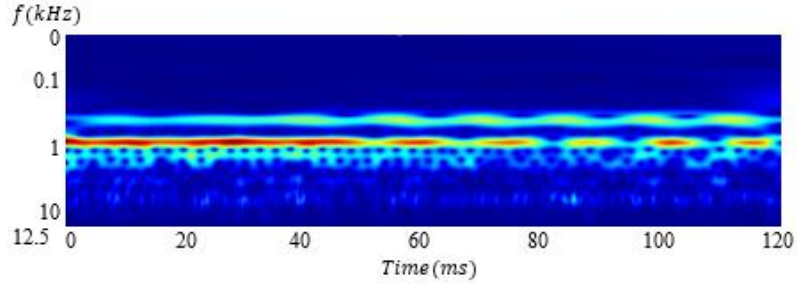
(a)



(b)



(c)



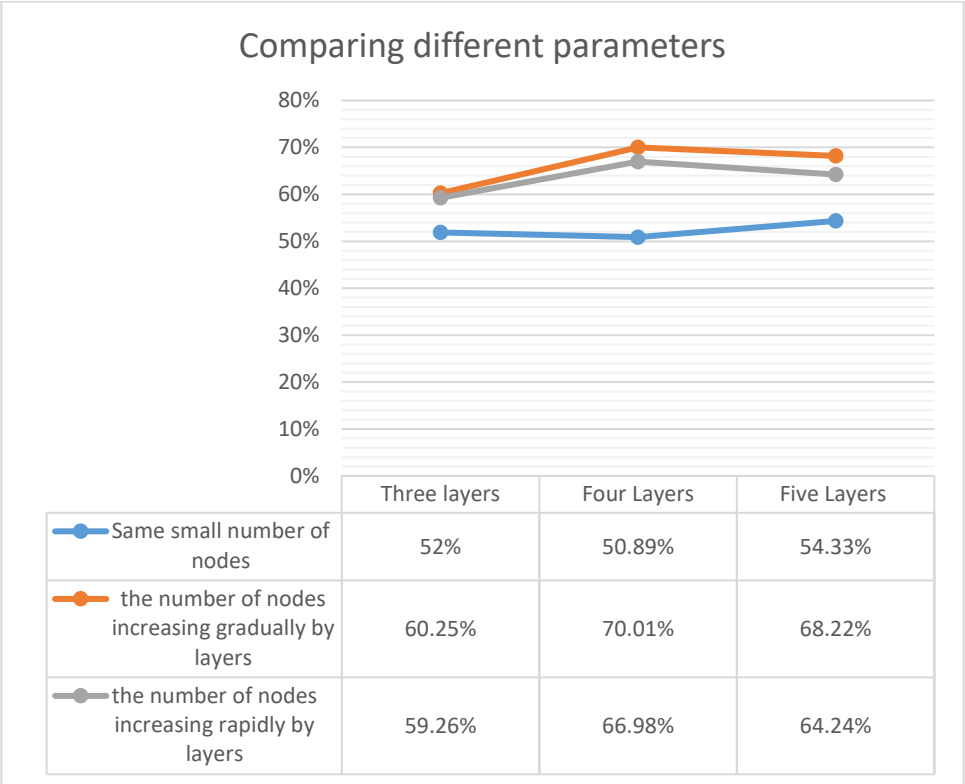
(d)

Figure 6.5 Standard input of (a) Normal voice (asra.wav) (GRBAS grade: 0) (b) Normal voice (Mtaaa2.wav) (G:0, R:0, A:1, B:1, S:0, Grade:2) (c) Pathological voice (Polipo11.wav) (pediculate polyp) (G:0, R:0, A:1, B:1, S:0, Grade:2) (d) Pathological voice (rpjbis.wav) (G:2, R:2, A:2, B:2, S:3, Grade:11)

However, we can still see that mild pathological voice with GRBAS grade 2 in Figure 6.5 (c) diagnosed of vocal cord polyp shows smoother and less irregular patterns than normal voice with GRBAS grade 2 in Figure 6.5 (b). This is the challenge that we described as “weak labels”. As mentioned before, these weak labels reveal that the dataset has overlap between normophonic voice and pathological voice. This will lead to confusion when training classifiers.

6.2.4 CNN architecture

In order to search for the most appropriate parameters, we explored the testing performance on three layers, fours layers, and five layers with the SVD database. The comparison of the performance with different kind of parameters are shown in Figure 6.6. The first type is using the same amount small number of nodes (8) on each layer. The number of nodes in the second type is increasing gradually by layers (3 layers: 10, 30, 50; 4 layers: 30, 50, 100, 150; 5 layers: 30, 50, 100, 150, 200). The number of nodes in the third type is increasing with larger number of nodes by layers (3 layers: 50, 100, 200; 4 layers: 50, 100, 150, 200; 5 layers: 50, 100, 150, 200, 300). The training progress and performance are listed in Appendix B.



	Three layers	Four Layers	Five Layers
Same small number of nodes	52%	50.89%	54.33%
the number of nodes increasing gradually by layers	60.25%	70.01%	68.22%
the number of nodes increasing rapidly by layers	59.26%	66.98%	64.24%

Figure 6.6 Searching for the most appropriate parameters for CNN model

From Figure 6.6, we can see that small number of nodes in each layer performs worst on different-number-of-layer CNN models. Scalogram input in this model (127 by 3000 by 3) is very large compared to the spectrogram input shown in Chapter 5.2 (84 by 257 by 1). Therefore, 8 nodes of each layer adopted in the model in Chapter 5.2 is not suitable for this situation. The number of nodes increasing gradually by layers performs the best, and four-layer model achieves the best performance among these experiments. In this case, we choose the four-layer CNN system, and the number of filters from first layer to the fourth layer is 30, 50, 100, and 150 respectively as illustrated in Figure 6.7.

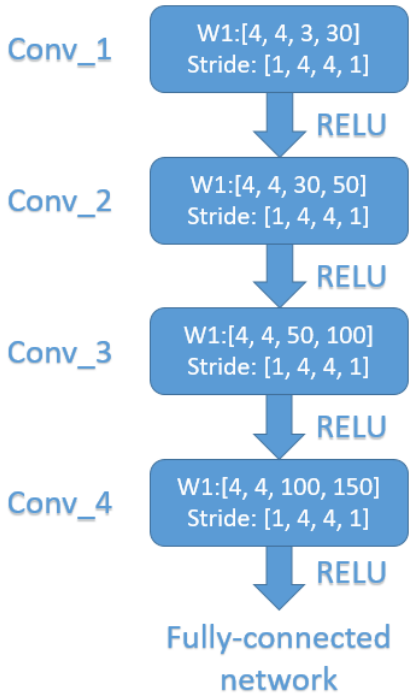


Figure 6.7 4-layer CNN system architecture

Each layer contains a convolutional layer and an activation layer. It is shown in research[367] that max-pooling can simply be replaced by a convolutional layer with increased stride with no significant effect to the accuracy on a number of image recognition application benchmarks. In addition, the CNN model reduces computational complexity without pooling layer. The stride of the filter in each convolutional layer is 4 by 4, the size of filters are 4 by 4, and the activation function is RELU. The input 2D representation contains three channels, so the input filter size is 4 by 4 by 3. In Section 5.2, we use 4 by 4 pooling layer for the 10-layer CNN model. In R-Net, the selected filter size and the large stride replaced the pooling layer, and this does not affect the performance at the same time (with the pooling layer, the

testing accuracy is 70.15% as shown in Appendix B.9). Python based Tensorflow is used as a framework for the training process. The training samples are divided into 64 samples in each mini-batch to be trained on GPU NVidia GTX1070 in this work. Adam Optimizer is applied in the experiment with initial learning rate 0.00003. Delta value of the L2 regularization in dense layer is set to 0.00005 and the maximum epochs of training is 80.

The experiments are conducted on three databases for pathological voice detection and the whole databases are used. The train-test split is 70% to 30%. SVD database: there are 960 training data (480 healthy and 480 pathological), and 727 testing data (207 healthy and 520 pathological). PdA database: there are 280 training data (140 healthy and 140 pathological), and testing data amount is 160 (99 healthy and 61 pathological). AVPD database: there are 232 training data (116 healthy and 116 pathological), and 107 testing data (51 healthy and 56 pathological).

6.3 Experimental Results

In order to validate the data augmentation idea in R-Net, we compare the R-Net performance with two comparison experiments. One uses the same CNN architecture as R-Net, but trained and tested with only the raw input representation, another one also uses the same CNN architecture as R-Net, but trained and tested with only the standard input representation.

6.3.1 Performance comparison

The overall performance using R-Net on three databases is shown in Table 6.1. It also contains the performance of one comparison experiments, which is training and testing the network with only raw input representation. This comparison experiment is also used for designing R-Net (architecture, parameters). The training accuracy achieved 86.77%, 96.79% and 99.14% accuracy on SVD, PDA and AVPD database respectively, while the testing accuracy achieves 64.10%, 71.88% and 65.42% accuracy on the three databases respectively. It can be seen that the overfitting problem is significant. After using R-Net, with standard input representation for verification, we can see that the performance on the training dataset achieved 84.27%, 88.21%, and 78.02% accuracy on SVD, PDA, and AVPD database respectively, and the performance on testing dataset achieved 89.27%, 75.62% and 85.05% accuracy

on the three databases. It shows that the training dataset accuracy drops but the testing dataset accuracy increases. Compared to traditional CNN models on the other two comparison experiments, R-Net improves the overall performance on all three databases, especially on the most challenging SVD database. It achieved 89.27% testing accuracy on SVD database, which improves the benchmark performance by around 15% to 18%.

In Table 6.2, the performance of another comparison experiment is shown. It is with the same architecture as R-Net, but trained and tested with standard input representation. This experiment is to verify whether it is the standard data input representation that plays the important role in R-Net or the data augmentation idea that leads to the success in R-Net. It is shown in Table 6.2 that the training accuracy achieves 91.04%, 97.86%, and 99.57% on SVD, PDA, and AVPD databases respectively, while the testing accuracy achieves 59.15%, 70.63%, and 67.29% accuracy on these three databases. Similar to the first comparison experiment which trained and tested with raw input representation, the overfitting problem is severe, and it shows that the standard input representation is not the decisive factor that contributes to the success of R-Net. Area under the Curve (AUC) of the ROC curves are also an important evaluating metric. From Table 6.3, it can be seen that R-Net achieves 94.01%, 78.95%, and 89.15% AUC on SVD, PDA, and AVPD database, which surpasses the other two comparison experiments too.

Table 6.1 Overall performance using R-Net

	Input	SVD	PDA	AVPD
Design (70% training - 30% testing)	Training raw input	86.77	96.79	99.14
	Testing raw input	64.10	71.88	65.42
Verification	Training standard input	84.27	88.21	78.02
	Testing standard input	**89.27**	**75.62**	**85.05**

Table 6.2 Overall performance using only standard input

	Input	SVD	PDA	AVPD
Design (70% training – 30% testing)	Training standard input	91.04	97.86	99.57
	Testing standard input	59.15	70.63	67.29

Table 6.3 Comparison of Area under the Curve (AUC) with R-Net

Comparison/%	SVD	PDA	AVPD
only raw input	66.05	76.68	71.04
R-Net	**94.01**	**78.95**	**89.15**
only standard input	64.37	75.51	73.77

In order to analyse the performance in detail, the classification results report with different metrics on three databases are explored. The classification result report on SVD are shown in Table 6.4. For SVD database, the confusion matrix of testing data on R-Net, on the comparison experiment with only raw input representation, and on the comparison experiment with only standard input representation are shown in Table 6.5, Table 6.6, and Table 6.7 respectively. The related ROC curves on SVD database are compared in Figure 6.8.

Table 6.4 Classification result report on SVD

	TPR(SN)(r)	SP(1-FPR)	p	F1	AUC	Accuracy
R-Net	91.79	88.27	75.70	82.97	94.01	89.27
only raw input	57.49	66.73	40.75	47.70	66.05	64.10
only standard input	55.56	60.58	35.94	43.64	64.37	59.15

Table 6.5 Confusion matrix of experiments of R-Net on SVD database

Testing data	**True: healthy**	**False: pathological**
Prediction: healthy	190	61
Prediction: pathological	17	459

Table 6.6 Confusion matrix of comparison experiments with only raw input representation on SVD database

Testing data	**True: healthy**	**False: pathological**
Prediction: healthy	119	173
Prediction: pathological	88	347

Table 6.7 Confusion matrix of comparison experiments with only standard input representation on SVD database

Testing data	**True: healthy**	**False: pathological**
Prediction: healthy	115	205
Prediction: pathological	92	315

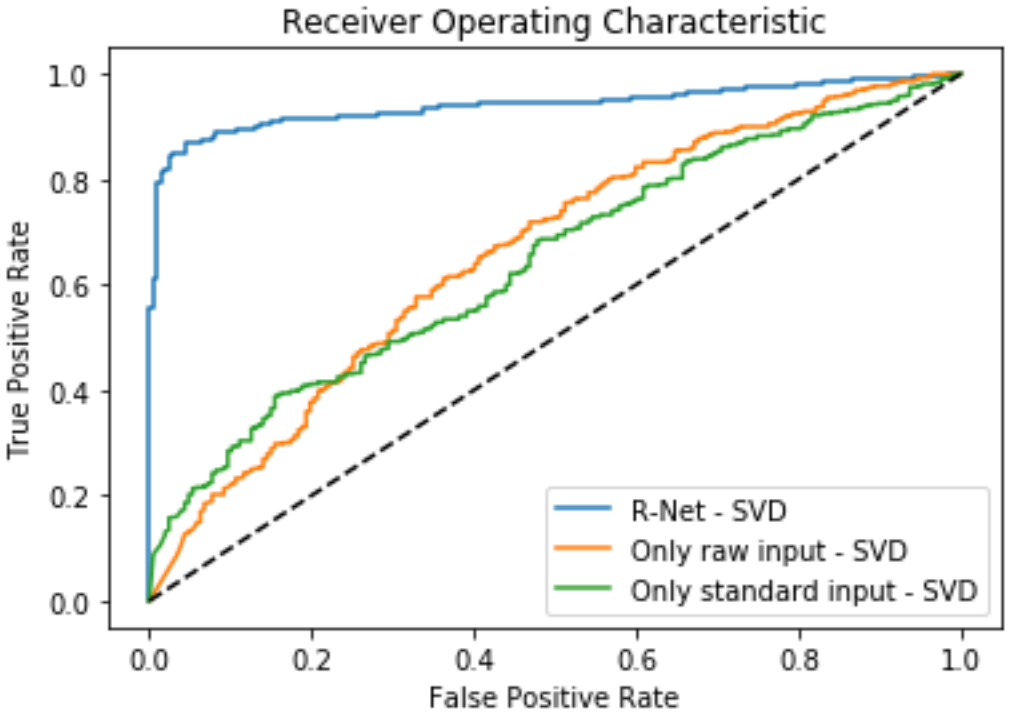


Figure 6.8 Performance comparison on SVD database

It can be seen in Table 6.4 that R-Net performs significantly superior on all the metrics for evaluation, including sensitivity, specificity, precision, F1-score, AUC, and overall accuracy. From Figure 6.8, the ROC curve can also shows the strong potential of R-Net compared to other two traditional CNN models' experiments. The superiority of R-Net is very significant in SVD database.

The classification result report on PdA database are shown in Table 6.8. For PdA database, the confusion matrix of testing data on R-Net, on the comparison experiment with only raw input representation, and on the comparison experiment with only standard input representation are shown in Table 6.9, Table 6.10, and Table 6.11 respectively. The related ROC curves on PdA database are compared in Figure 6.9.

Table 6.8 Classification result report on PdA

	TPR(SN)(r)	SP(1-FPR)	p	F1	AUC	Accuracy
R-Net	76.77	73.77	82.61	79.58	78.95	75.62
only raw input	71.72	72.13	80.68	75.94	76.68	71.88
only standard input	71.72	68.85	78.89	75.13	75.51	70.63

Table 6.9 Confusion matrix of experiments of R-Net on PdA database

Testing data	True: healthy	False: pathological
Prediction: healthy	76	16
Prediction: pathological	23	45

Table 6.10 Confusion matrix of comparison experiments with only raw input representation on PdA database

Testing data	True: healthy	False: pathological
Prediction: healthy	71	17
Prediction: pathological	28	44

Table 6.11 Confusion matrix of comparison experiments with only standard input representation on PdA database

Testing data	True: healthy	False: pathological
Prediction: healthy	71	19
Prediction: pathological	28	42

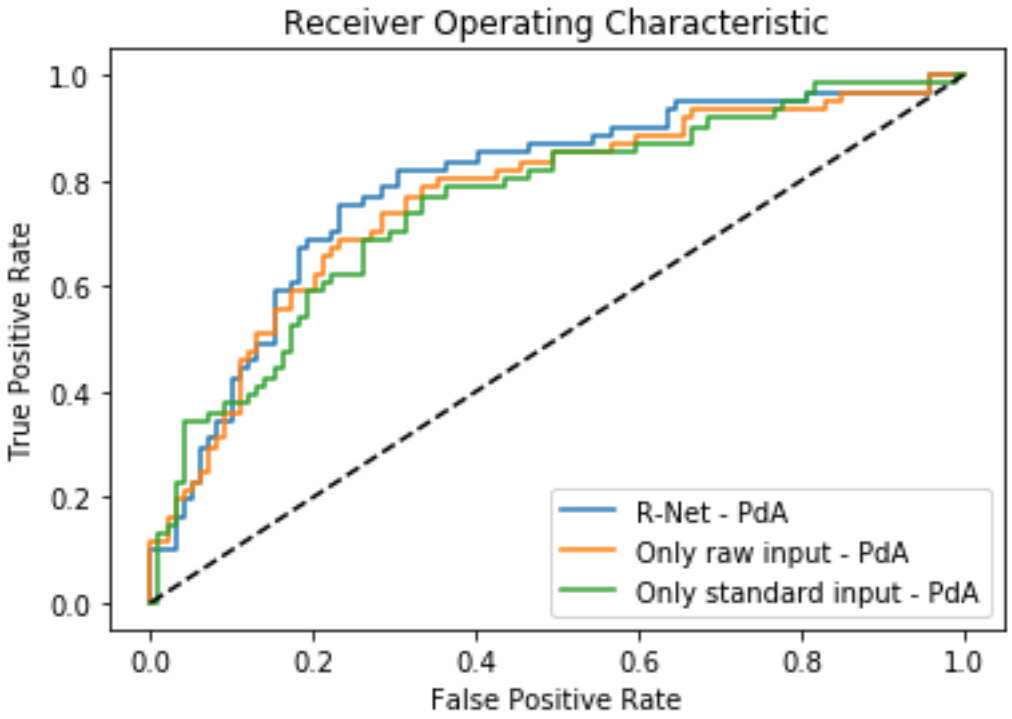


Figure 6.9 Performance comparison on PdA database

It can be seen in Table 6.8 that R-Net have slight improvement on all the metrics for evaluation. From Figure 6.9, the ROC curve can also show slight improvement of R-

Net from the other two traditional CNN models' experiments. However, the difference is not as significant as it is in SVD database.

The classification result report on AVPD database are shown in Table 6.12. For AVPD database, the confusion matrix of testing data on R-Net, on the comparison experiment with only raw input representation, and on the comparison experiment with only standard input representation are shown in Table 6.13, Table 6.14, and Table 6.15 respectively. The related ROC curves on AVPD database are compared in Figure 6.10.

Table 6.12 Classification result report on AVPD

	TPR(SN)(r)	SP(1-FPR)	p	F1	AUC	Accuracy
R-Net	84.31	85.71	84.31	84.31	89.15	85.05
only raw input	62.75	67.86	64.00	63.37	71.04	65.42
only standard input	66.67	67.86	65.38	66.02	73.77	67.29

Table 6.13 Confusion matrix of experiments of R-Net on AVPD database

Testing data	**True: healthy**	**False: pathological**
Prediction: healthy	43	8
Prediction: pathological	8	48

Table 6.14 Confusion matrix of comparison experiments with only raw input representation on AVPD database

Testing data	**True: healthy**	**False: pathological**
Prediction: healthy	32	18
Prediction: pathological	19	38

Table 6.15 Confusion matrix of comparison experiments with only standard input representation on AVPD database

Testing data	**True: healthy**	**False: pathological**
Prediction: healthy	34	18
Prediction: pathological	17	38

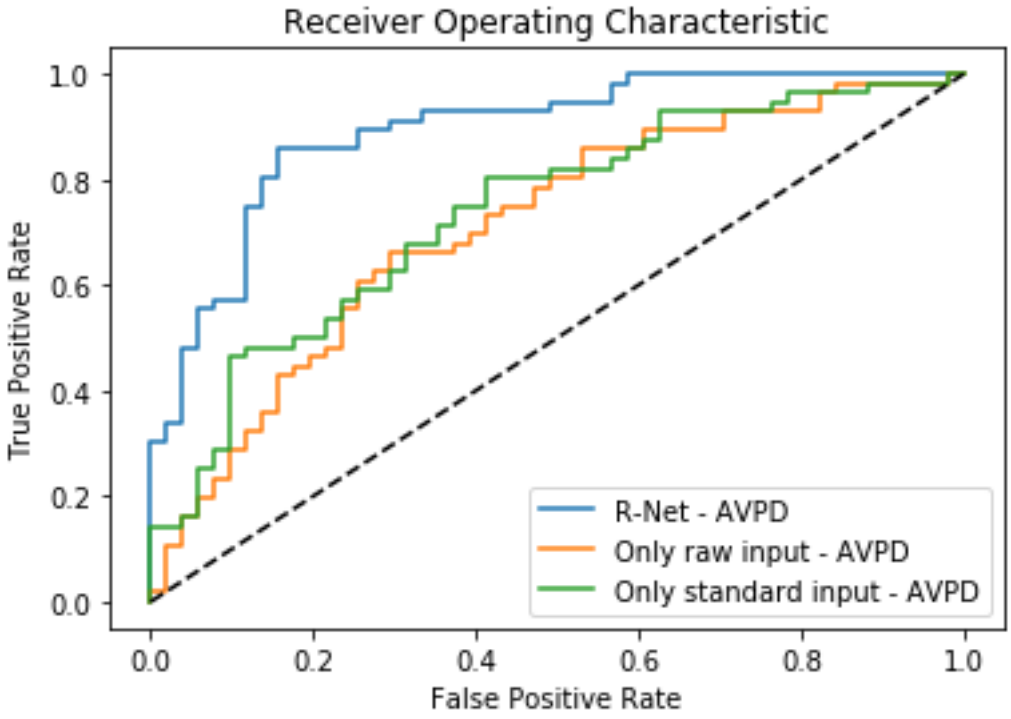


Figure 6.10 Performance comparison on AVPD database

It can be seen in Table 6.12 that R-Net performs significantly better on all the metrics for evaluation, including sensitivity, specificity, precision, F1-score, AUC, and overall accuracy. From Figure 6.10, the ROC curve can also shows the strong potential of R-Net compared to other two traditional CNN models' experiments on AVPD database. The superiority of R-Net is also significant in AVPD database.

It can be seen that ROC curve in SVD database achieves best performance, with the curve closest to the ideal curve and the top-left point closest to (0,1) point. The performance on AVPD database also achieves excellent result. However, the PdA database performance not as good as other two databases. This is because the PdA database does not contain enough data with pre-speech silence. This can also been seen in the RNN model in Chapter 4. The deep acoustic recurrent model achieves best performance on PdA database compared to SVD and AVPD database, because PdA database contains less data with pre-speech onset. Recurrent models requires good consistent sequence data, and PdA contains best consistent sequence data that satisfy the recurrent model.

6.3.2 Discussion

Table 6.1 and Table 6.2 compares the overall training and testing accuracy over R-Net model and another two comparison experiments on three databases. It shows that R-Net significantly improves the testing data performance and balances the training data performance. From the ROC performance comparisons on the three databases, R-Net shows its significant superiority over the other two comparison experiments. It shows that the success of R-Net on reducing the overfitting problems and improving the overall performance relies on the data augmentation ideas, not the CNN architectures or other pre-processing techniques.

The main reason of the over-fitting problem is that the network can easily learn the training data's characteristics, so that it prevent the network from training further for detecting real pathological characteristics. In order to reduce over-training for training data's characteristics, data augmentation idea is applied and cropping/translation operations are conducted to some data. In this work, R-Net shows its great potential in eliminating the overfitting problems in small-data problems.

6.4 R-Net on RRP detection application

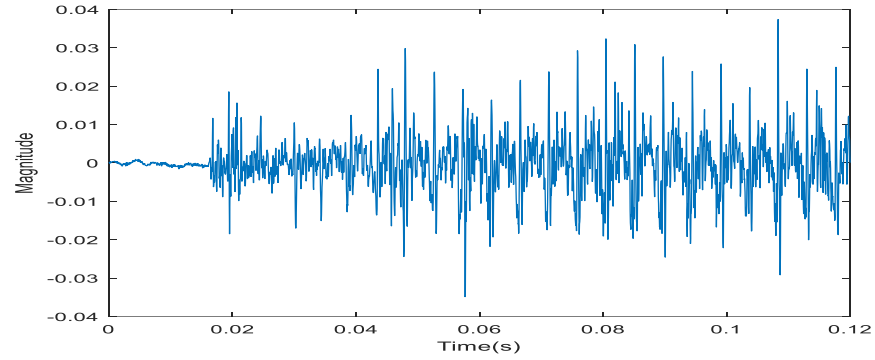
R-Net achieves significant improvement in pathological voice detection, with overfitting problems significantly reduced. Specifically, it achieved 89.27% testing accuracy on the most challenging SVD database. The trained model demonstrated good generalization ability with comparatively large amount of data in SVD database. In order to evaluate the transfer learning ability of the trained R-Net model, and expand the application field, we conduct transfer learning experiments on Recurrent Respiratory Papillomatosis (RRP) detection. RRP is a disease that is caused by the human papilloma virus (HPV), normally affecting young children. This disease commonly appears at 3-4 years old, and it will cause airway blockage when the lesions becomes larger. Paediatric RRP is important to be diagnosed at the early stage, which will reduce the risk of deterioration. Acoustic analysis and the related assistant diagnostic tools are investigated to help clinicians to detect RRP. In this work, we apply the R-Net model trained by SVD database for transfer learning, and fine-tune the model with RRP and vocal cord nodule (VCN) data.

6.4.1 Background

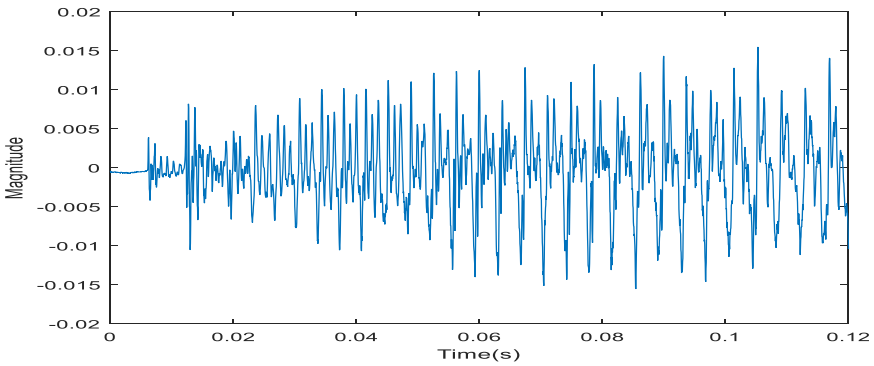
RRP is the most common benign (non-cancerous) tumour that appears in the upper airway. It will lead to voice quality changes with symptom like hoarseness or loss of voice.

Paediatric RRP develop before the age of 12, which is more severe and recurring than adult-onset form of RRP. In Figure 2.4 in Section 2.2.1, an example of vocal cord papilloma was presented. The development of the tumour is gradual and slow, so that it is rare in children with sudden breathing problem. However, after a period of time, this may lead to airway obstruction, which is a life-threatening problem. Generally, children with RRP requires multiple surgical treatment.

Apart from laryngoscopy for diagnosing paediatric RRP, acoustic diagnostic techniques are considered to minimize discomfort for children. Expert clinicians is able to tell the subtle difference between paediatric RRP and vocal cord nodules (VCN) (Another commonly seen voice disorder in children). In this work, deep learning techniques are investigated to distinguish paediatric RRP and VCN. The examples of these two types of voice disorders are shown in Figure 6.11.



(a)



(b)

Figure 6.11 Voice example of RRP and VCN (a) RRP (24.wav) (b) VCN (29.wav)

6.4.2 Experiments

We collaborated with clinicians in Glasgow Royal Infirmary and speech and language therapists in University of Strathclyde with this feasibility study. Clinicians collected acoustic samples from 8 RRP patients and 8 VCN patients. Sustained vowel /a/ is applied in this research. In Chapter 2.2.1, the causing reasons and symptoms of VCN and RRP were introduced in detail. VCN is often caused by overuse of vocal folds while RRP is caused by HPV virus. RRP appears mostly in children, which is a serious disease that leads to airway blockage. Diagnosing RRP is challenging suitable to only experienced professional clinicians. In this project, we explore the assistant diagnostic tool for clinicians to use for detecting RRP.

There are overall 8 RRP and 8 VCN patients in the dataset for classification. It can be seen that it is a small dataset classification problem. Since machine learning models trained by small dataset often lack generalization ability in most cases, transfer learning becomes an essential method for this problem.

A pre-trained R-Net model for pathological voice detection using the SVD database was the model used in detecting pathological voice. In this work, RRP and VCN data is fine-tuned using the pre-trained model. There are 8 RRP and 8 VCN samples, each sample is sustained vowel /a/ from one patient. The training-testing split is 70% - 30%, with 6 testing data and 10 training data. We conduct the experiments 20 times, with the input data shuffled with different random seed. In this case, the results are more reliable on small data classification problem.

In order to validate the noise robustness method in R-Net, we use both raw input and standard input to evaluate R-Net. The transfer learning results on RRP to VCN classification with R-Net that pre-trained on SVD database are shown in Table 6.16. In the table, 83.34% testing accuracy means 1 error prediction, and 66.67% means 2 error prediction, 50% testing accuracy means 3 error prediction. The left part is the performance of the model that fine-tuned and tested with raw input and the right part is the performance of the model that fined-tuned and tested with standard input. From top to bottom, different random seeds (2 to 42) are used to shuffle the training and testing data.

We can see from Table 6.16 that the result are not robust when using raw input for fine-tuning the R-Net model, with 50% testing accuracy appears three times in 20

experimental trials. However, the performance with standard input shows no less than 66.67% testing accuracy, with 100% testing accuracy appears five times in 20 experimental trials. It shows that R-Net works on classifying RRP data and VCN data, with better robustness and higher accuracy. Furthermore, this work shows that transfer learning from pathological voice detection to pediatric RRP detection achieves excellent performance using limited small amount of data. As a feasibility study, pediatric RRP detection deserves to be investigated to the next level, with larger amount of data and further experiments with R-Net.

Table 6.16 Transfer learning on RRP to VCN classification with R-Net that pre-trained on SVD database

random shuffle seed	Evaluate with raw input		Evaluate with standard input	
	training accuracy/%	testing accuracy/%	training accuracy/%	testing accuracy/%
2	100.00	66.67	100.00	**83.34**
3	90.00	**83.34**	90.00	**100.00**
4	90.00	**83.34**	90.00	**83.34**
5	90.00	**50.00**	90.00	**83.34**
6	90.00	**83.34**	90.00	66.67
7	90.00	**100.00**	90.00	**83.34**
8	100.00	66.67	100.00	66.67
9	100.00	66.67	100.00	66.67
10	90.00	**83.34**	90.00	**100.00**
11	100.00	66.67	100.00	**83.34**
12	100.00	66.67	100.00	66.67
13	90.00	**50.00**	90.00	**83.34**
14	90.00	**100.00**	90.00	**83.34**
15	90.00	**100.00**	90.00	**100.00**
16	90.00	**83.34**	90.00	**100.00**
17	100.00	**50.00**	100.00	66.67
18	90.00	**83.34**	90.00	**83.34**
19	90.00	**83.34**	90.00	**100.00**

20	100.00	66.67	100.00	66.67
42	90.00	**100.00**	90.00	**83.34**

6.5 Conclusion

This issue of small-data problem is generally addressed using two methods. One way is to use transfer learning, which was explored in Section 5.3. Another way is data augmentation, which is the essential idea in R-Net proposed in this Chapter. In R-Net, the raw input representation can be seen as augmented data for designing the network, with pre-speech silence kept. Then the standard input representation which pre-speech silence removed is used for making decisions. This data augmentation skill reduces the over-fitting problem and improves the performance. The proposed R-Net shows with strong performance on the most challenging SVD database, with 89.27% testing accuracy. It improves around 15% to 20% accuracy compared to the benchmark results. Finally, we validated the trained R-Net on detecting paediatric RRP data, and it achieved successful result on a limited RRP data. This gives more evidence to noise robustness theory, which is an area that requires to be investigated further in the future, especially in biomedical field with limited data.

Chapter 7

7 Conclusion and Future Works

7.1 Conclusion

In this thesis, a variety of signal processing techniques and deep learning methods have been explored in different aspects for developing assistant diagnostic model for dysphonia. Signal processing techniques include Short-Time Fourier Transform, Wavelet transform, Mel Frequency Cepstral Analysis, and Cepstral analysis etc. Deep learning methods include RNN, CNN, are explored to design the model for pathological voice detection.

In Chapter 2, an extensive literature review was presented, focusing on introducing the causes, characteristics and symptoms of various types of dysphonia, and the techniques proposed for diagnosis over the years. Both perceptual and acoustic analysis of dysphonia are also been illustrated with current state-of-the-art techniques. From perceptual point of view, GRBAS and CAPE-V are two common tools for evaluating voice quality in a variety of aspects, such as breathiness, roughness, and strain. From acoustic point of view, these voice quality metrics are evaluated from acoustic features based on signal processing techniques. These short-time acoustic features have been generally employed in pathological voice detection, including temporal acoustic analysis, perturbation and fluctuation analysis, spectral or cepstral analysis and 2D representations. Among them, cepstral and spectral analysis are explored in Chapter 4, and 2D time-frequency representations are explored in Chapter 5 and Chapter 6. As final part, four databases commonly used in this field are introduced (MEEI, SVD, PDA and AVPD). The healthy voice and pathological voice in MEEI database are recorded in different environments, so that it is not a reliable database for usage. Only SVD, PDA and AVPD databases were used in this work.

In Chapter 3, the overview of the processing flow in pathological voice detection are presented with a sufficient literature review. In Section 3.2,2, we presented the

overall three main challenges in this field: "*weak labels*", "black-box" dilemma, and limited data.

"Weak Label" challenge was due to the subjectivity of diagnosis from clinicians with different background and level of experiences. This leads to inaccurate labels. These weak labels can confuse deep learning classifiers during training process and limit the performance. "Black-box" dilemma refers to the unclear interpretability of deep learning system. For example, the working mechanism of CNN is still a black-box, so that it is still hard to interpret dysphonic characteristics from CNN end-to-end system.

Furthermore, since the cost of data collection is high, limited data problem exists in many applications in the biomedical field. The overall amount of annotated data in the used three databases are around 2000. However, state-of-the-art deep learning models are designed for other applications such as speech recognition or object detection, which contains millions of data samples for deep learning models for training. Generally, there are two methods to solve the problem: transfer learning and data augmentation. Transfer learning methods are explored in Chapter 5.3.2, and data augmentation ideas are used in Chapter 6 for designing CNN models.

In Section 3.4, a variety of machine learning techniques are reviewed in detail, including both supervised learning and unsupervised learning methods. Unsupervised learning method explores the inner rule of the data itself, without learning from the targets. Since pathological voice detection is a classification problem, unsupervised learning methods tends to be weak compared to supervised learning methods, especially state-of-the-art deep learning techniques. In this work, we place emphasis on exploring supervised deep learning methods on pathological voice detection.

The first novel contribution of a deep recurrent acoustic model for pathological voice detection are proposed in Chapter 4. Since cepstral features are proved to contain strong correlation with dysphonia, cepstral/spectral features are explored in this model as short-time-frame based features. Apart from traditional cepstral features in ADSV manual, two novel features based on cepstrum are proposed (SPP and CepStd). It is shown in Chapter 3 that RNN model contains the ability of analysing sequence data. In this model, cepstral feature sets are fed into a two-layer Bi-LSTM

model, and it achieved successful performance on three databases. It shows 71.60%, 82.58%, and 78.43% accuracy respectively on SVD, PDA and AVPD database. This is the first time cepstral features has been explored with deep recurrent models. It shows that cepstral features show the ability of detecting dysphonic characteristics with the deep recurrent model. The two novel cepstral features also show their superiority in improving the classification performance. In addition, we compared the proposed model with traditional acoustic classification models. Traditional classifiers such as SVM, random forests are explored with the selected best feature set. It was shown that deep recurrent model achieves higher accuracy with less features than traditional acoustic models.

In Chapter 5, we explored CNN models in pathological voice detection field with 2D time-frequency representations of the speech data. Firstly, a novel 10-layer CNN model designed for training from scratch. This model achieves 67.95% accuracy on SVD database, which is comparable to the benchmarks. However, the overfitting problem due to the limited data shows. In order to reduce the overfitting problem, transfer learning are explored with four state-of-the-art CNN networks from image classification field. The testing data accuracy improves while the overfitting problem is even more serious with complex CNN models. Among them, GoogleNet achieves 71.74%, 80.30%, and 79.41% accuracy with zoomed spectrogram on SVD, PDA, and AVPD database respectively; ResNet achieves 70.75%, 80.30%, and 80.39% accuracy with scalogram on SVD, PDA, and AVPD database respectively. These two networks contain the highest number of layers. They achieves the best testing data accuracy, while the training data accuracy is around 100%. It can be concluded that, with limited data, simpler architecture will benefit for reducing overfitting problems, while the testing data accuracy and the training-testing-performance-balance is a trade-off.

The main reason of the over-fitting problem is that the network can easily learn the training data's characteristics, so that it prevent the network from training further for detecting real pathological characteristics. In Chapter 6, data augmentation idea is applied to the traditional CNN model, and we propose the R-Net model for pathological voice detection. In R-Net, pre-speech silence is kept as "noisy" raw input representation for training, and pre-speech silence is removed as standard input representation for evaluating the trained model. The proposed R-Net shows with

strong performance on the most challenging SVD database, with 89.27% testing accuracy. It improves around 15% to 20% accuracy compared to the benchmark results. It also achieved 75.62% and 85.05% on PdA and SVD database. The training accuracy of R-Net is 84.27%, 88.21%, and 78.02% on these databases respectively. It can be seen that the overfitting problem is largely reduced and the testing performance is greatly improved. In order to validate the reason why the model is working, comparison experiments with only raw data input and only standard data input are conducted on the same CNN architecture as R-Net. It can be seen that R-Net improves the AUC and testing accuracy greatly due to the data augmentation idea, not the CNN architectures or other pre-processing settings.

Finally, we validated the trained R-Net on very-small-data paediatric RRP detection, and it achieved successful results. This gives more evidence to noise robustness theory and usefulness of data augmentation idea in limited data problems, which is an area that requires to be investigated further in the future.

7.2 Future Works

In this work, various research topics can be explored for further research. First of all, limited-data problem in pathological voice detection increases the difficulty on deep learning models for applications. In this thesis, the most essential contribution R-Net addressed this problem, with improved testing performance (89.27%, 75.62%, and 85.05% accuracy on SVD, PdA, and AVPD database respectively). The data augmentation idea is the essential idea that improves the performance and reduces the overfitting problems. Therefore, this idea is not only for pathological voice detection, but also for other limited-data-deep-learning applications, especially in bio-medical applications. More exploration on R-Net for transfer learning in other medical field can be explored in the future. Except for R-Net, more data augmentation techniques will be investigated for eliminating the overfitting problem, to improve the performance.

Moreover, apart from CNN models, RNN model with short-time-frame-based cepstral features also shows potential in pathological voice detection. More cepstral features that represents pathological information can be designed, and attention-mechanism can be explored with LSTM models for classification.

Furthermore, unsupervised learning such as Auto-encoders can be further investigated for extracting the dysphonic features from speech. Although it is shown from experiments that unsupervised learning methods does not show good performance compared to supervised deep learning techniques, it can still be employed when the labelled data is limited, or when it came across with “weak label” problems.

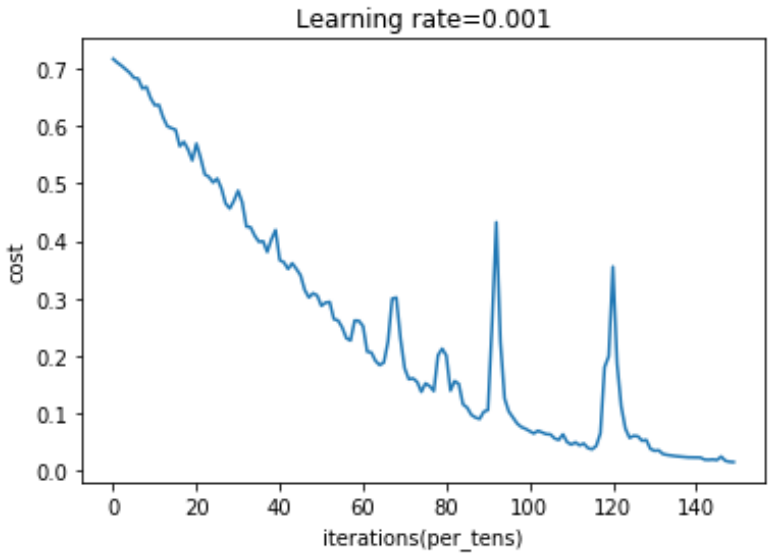
Finally, the principle of the deep learning field is “more data, better performance”. It is suitable to be applied in pathological voice detection too. More data will increase the generalization ability of the model. Therefore, a further dysphonia data collection will be conducted to add on the training process of R-Net or other deep recurrent models.

Appendix

Appendix A – Training progress and loss performance comparison on CNN model (Grid search for most appropriate parameters) (Referred to Chapter 5.2.1)

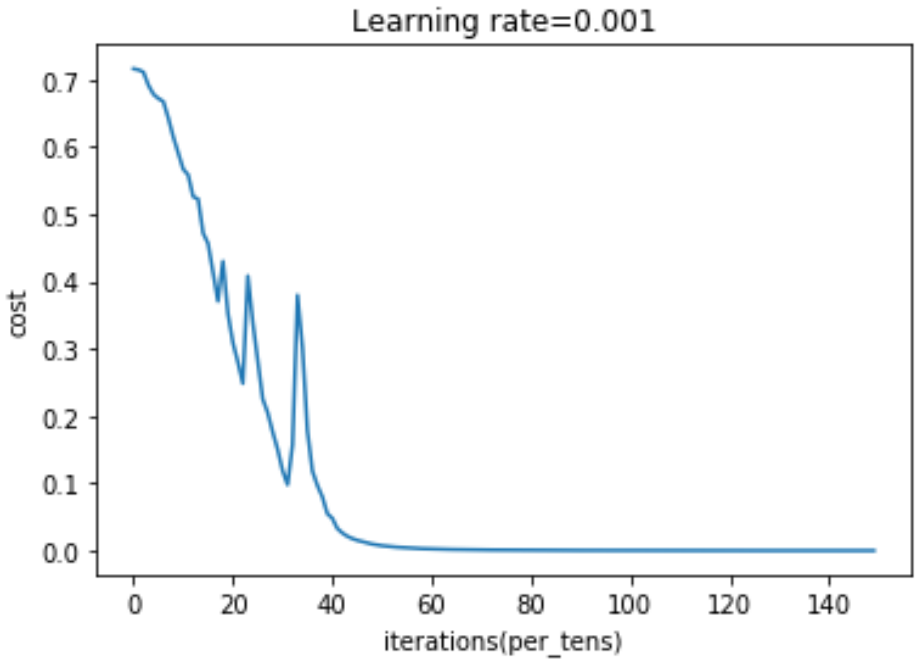
A1. 5-layer (8, 8, 8, 8, 8)

```
Cost after epoch 0: 0.716977
Cost after epoch 10: 0.636445
Cost after epoch 20: 0.569510
Cost after epoch 30: 0.487614
Cost after epoch 40: 0.366710
Cost after epoch 50: 0.287221
Cost after epoch 60: 0.251900
Cost after epoch 70: 0.178451
Cost after epoch 80: 0.200698
Cost after epoch 90: 0.105825
Cost after epoch 100: 0.069090
Cost after epoch 110: 0.045557
Cost after epoch 120: 0.355411
Cost after epoch 130: 0.034850
Cost after epoch 140: 0.022845
Confusion Matrix:
[[99 42]
 [58 67]]
Train Accuracy: 0.9987469
Test Accuracy: 0.62406015
```



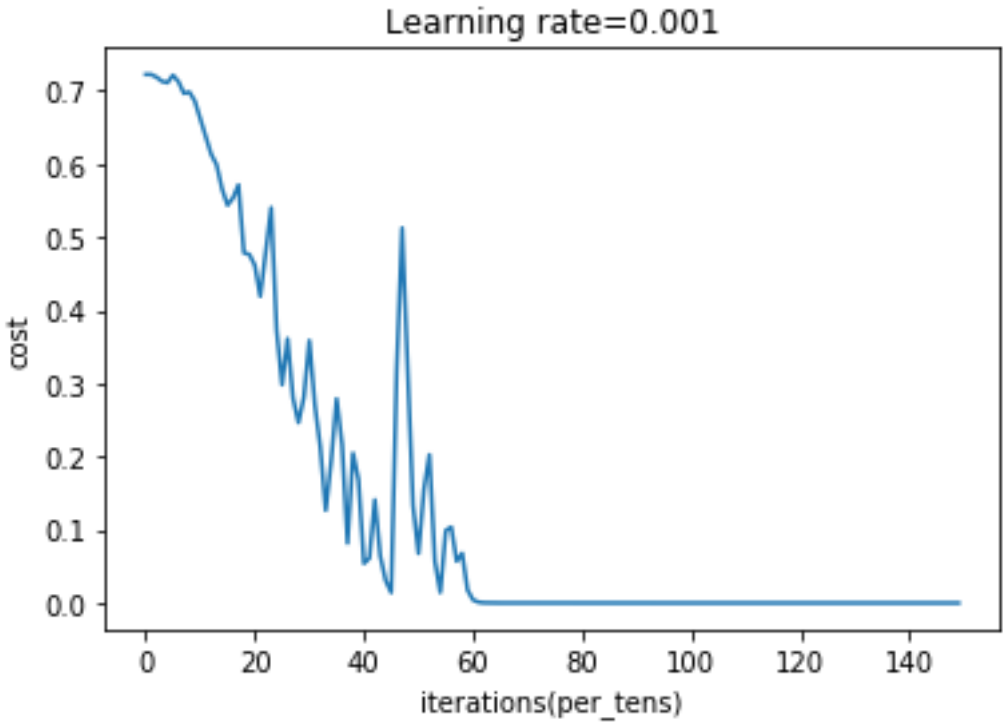
A2. 5-layer (8, 10, 20, 50, 100)

```
Cost after epoch 0: 0.715996
Cost after epoch 10: 0.566918
Cost after epoch 20: 0.309159
Cost after epoch 30: 0.119150
Cost after epoch 40: 0.048372
Cost after epoch 50: 0.007631
Cost after epoch 60: 0.002799
Cost after epoch 70: 0.001447
Cost after epoch 80: 0.000930
Cost after epoch 90: 0.000694
Cost after epoch 100: 0.000565
Cost after epoch 110: 0.000493
Cost after epoch 120: 0.000441
Cost after epoch 130: 0.000411
Cost after epoch 140: 0.000387
Confusion Matrix:
[[91 50]
 [55 70]]
Train Accuracy: 1.0
Test Accuracy: 0.6052632
```



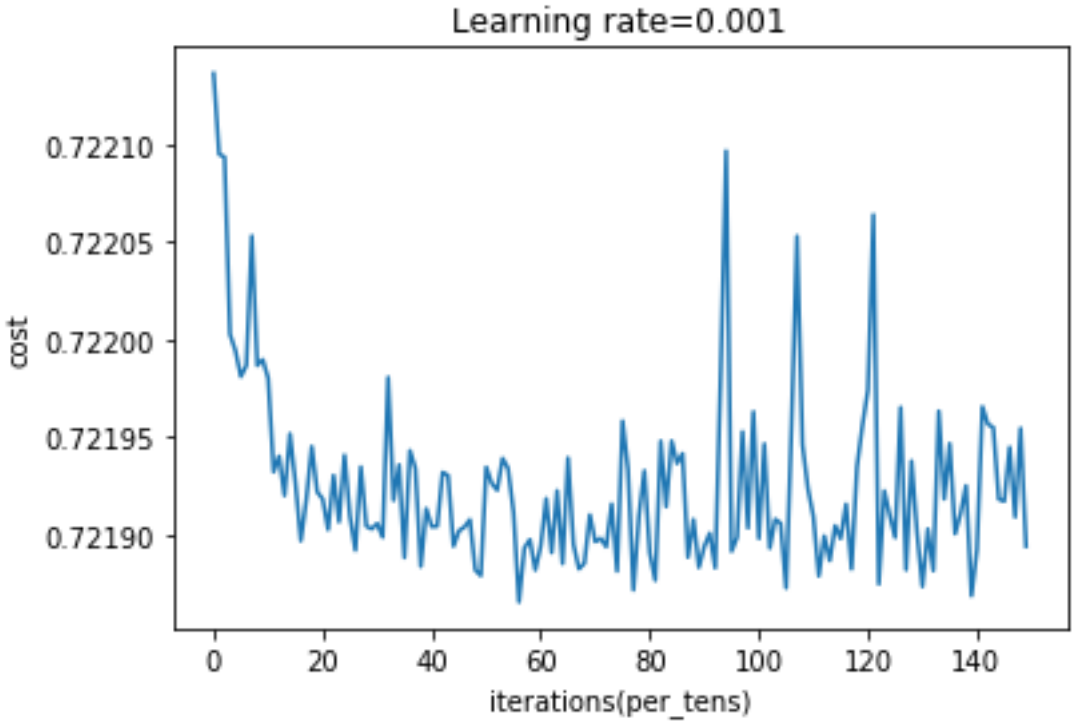
A3. 10-layer (8, 10, 20, 50, 80, 100, 150, 200, 250, 300)

```
Cost after epoch 0: 0.722187
Cost after epoch 10: 0.662010
Cost after epoch 20: 0.461497
Cost after epoch 30: 0.358743
Cost after epoch 40: 0.054127
Cost after epoch 50: 0.067735
Cost after epoch 60: 0.004249
Cost after epoch 70: 0.000061
Cost after epoch 80: 0.000052
Cost after epoch 90: 0.000051
Cost after epoch 100: 0.000050
Cost after epoch 110: 0.000049
Cost after epoch 120: 0.000049
Cost after epoch 130: 0.000048
Cost after epoch 140: 0.000048
Confusion Matrix:
[[89 52]
 [41 84]]
Train Accuracy: 1.0
Test Accuracy: 0.65037596
```



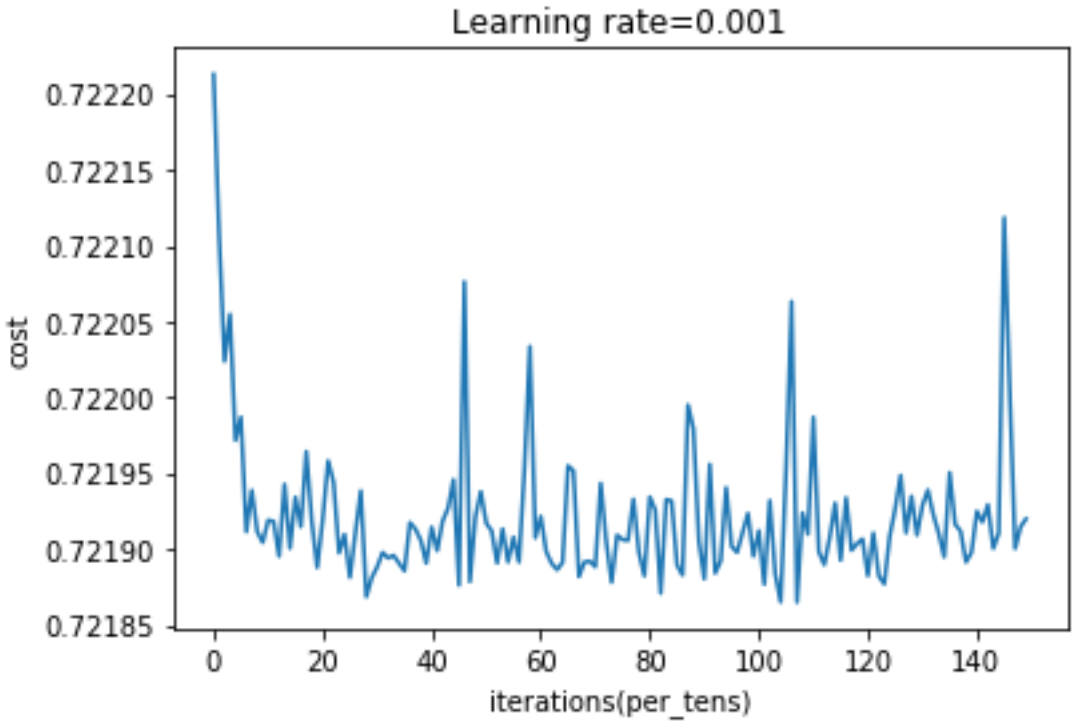
A4. 15-layer (8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8)

```
Cost after epoch 0: 0.722136
Cost after epoch 10: 0.721980
Cost after epoch 20: 0.721918
Cost after epoch 30: 0.721906
Cost after epoch 40: 0.721904
Cost after epoch 50: 0.721935
Cost after epoch 60: 0.721894
Cost after epoch 70: 0.721897
Cost after epoch 80: 0.721891
Cost after epoch 90: 0.721894
Cost after epoch 100: 0.721898
Cost after epoch 110: 0.721909
Cost after epoch 120: 0.721974
Cost after epoch 130: 0.721874
Cost after epoch 140: 0.721893
Confusion Matrix:
[[  0 141]
 [  0 125]]
Train Accuracy: 0.50877196
Test Accuracy: 0.4699248
```



A5. 15-layer (8, 10, 20, 50, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550)

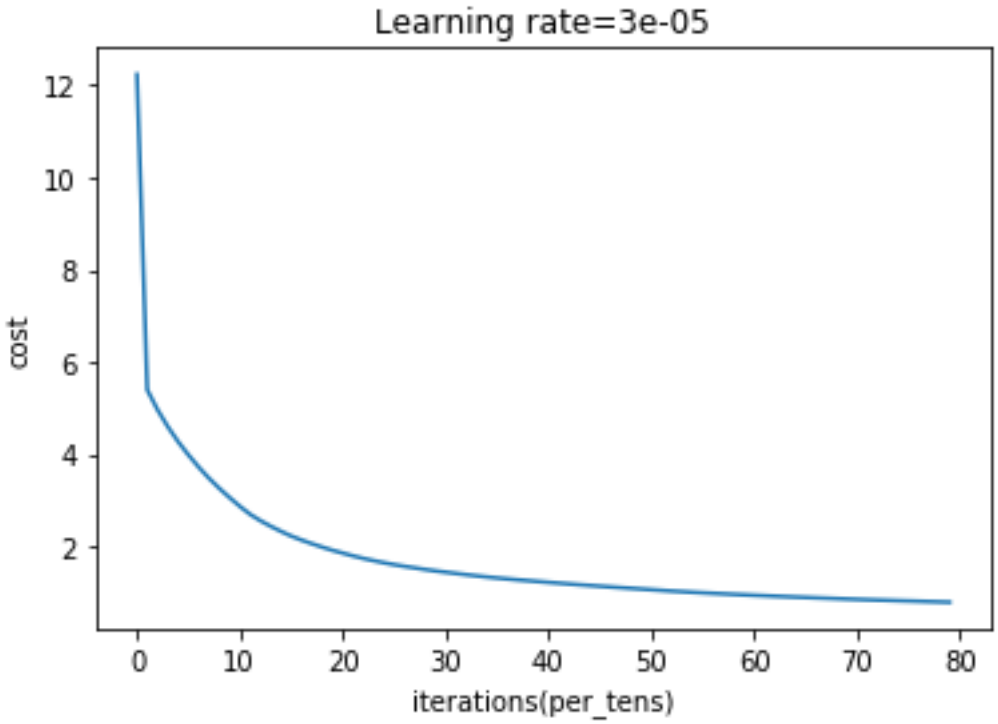
```
Cost after epoch 0: 0.722213
Cost after epoch 10: 0.721919
Cost after epoch 20: 0.721919
Cost after epoch 30: 0.721888
Cost after epoch 40: 0.721914
Cost after epoch 50: 0.721917
Cost after epoch 60: 0.721922
Cost after epoch 70: 0.721888
Cost after epoch 80: 0.721934
Cost after epoch 90: 0.721880
Cost after epoch 100: 0.721912
Cost after epoch 110: 0.721987
Cost after epoch 120: 0.721882
Cost after epoch 130: 0.721929
Cost after epoch 140: 0.721925
Confusion Matrix:
[[  0 141]
 [  0 125]]
Train Accuracy: 0.50877196
Test Accuracy: 0.4699248
```



Appendix B – Training progress and loss performance comparison on R-Net (Grid search for most appropriate parameters) (Referred to Chapter 6.2.4)

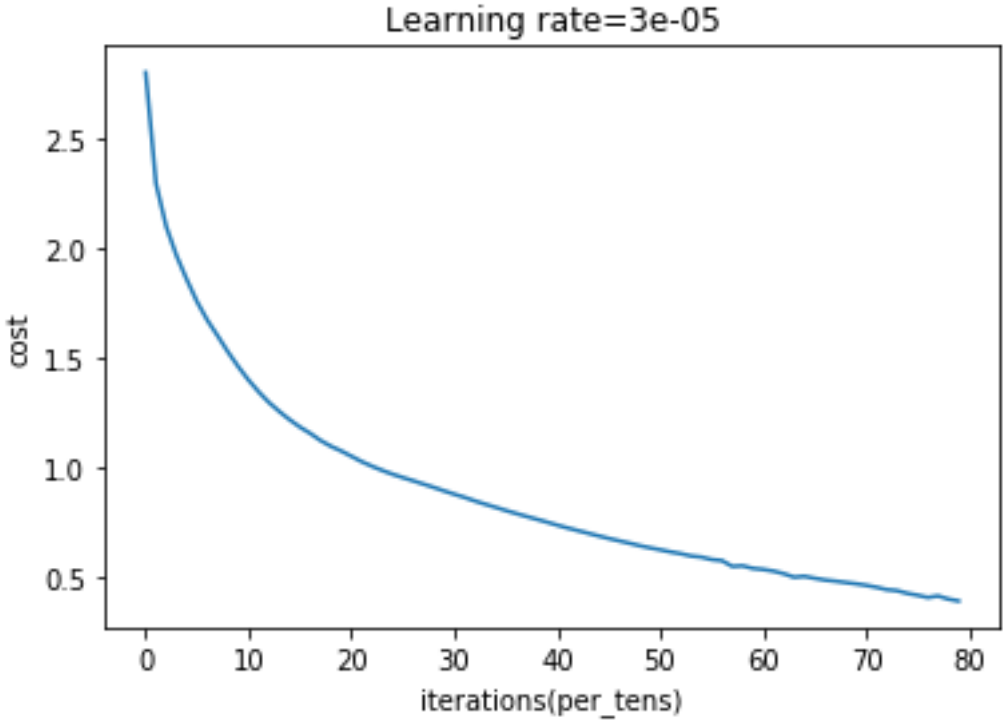
B1. 3-layer(8, 8, 8)

```
Cost after epoch 0: 12.240294
Cost after epoch 5: 3.989891
Cost after epoch 10: 2.875034
Cost after epoch 15: 2.229549
Cost after epoch 20: 1.860988
Cost after epoch 25: 1.610134
Cost after epoch 30: 1.446843
Cost after epoch 35: 1.321505
Cost after epoch 40: 1.226707
Cost after epoch 45: 1.142981
Cost after epoch 50: 1.067557
Cost after epoch 55: 0.999465
Cost after epoch 60: 0.945827
Cost after epoch 65: 0.899440
Cost after epoch 70: 0.857467
Cost after epoch 75: 0.822945
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/
DNN_Python/CNN/model225423_12102020\\saved_model.pb'
accuracy_train: 0.653125
accuracy_test: 0.5185694635488308
```



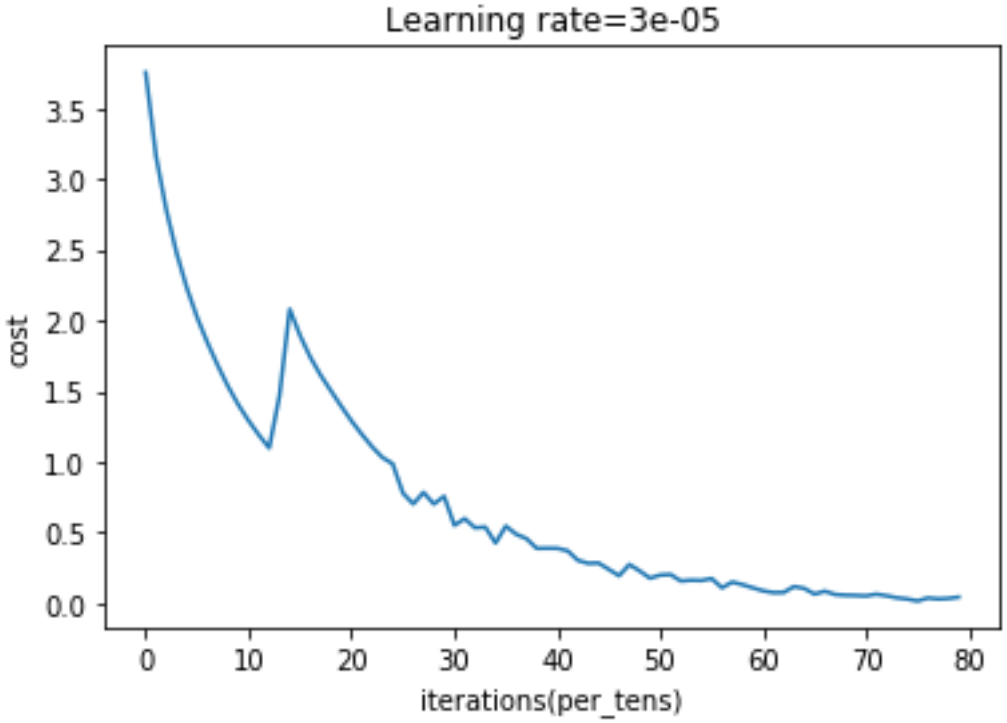
B2. 3-layer (10, 30, 50)

```
Cost after epoch 0: 2.801133
Cost after epoch 5: 1.755722
Cost after epoch 10: 1.399386
Cost after epoch 15: 1.185202
Cost after epoch 20: 1.052267
Cost after epoch 25: 0.955273
Cost after epoch 30: 0.879368
Cost after epoch 35: 0.805339
Cost after epoch 40: 0.738509
Cost after epoch 45: 0.677971
Cost after epoch 50: 0.626344
Cost after epoch 55: 0.581886
Cost after epoch 60: 0.536840
Cost after epoch 65: 0.496698
Cost after epoch 70: 0.464551
Cost after epoch 75: 0.419312
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/DNN_Python/CNN/
model202902_12122020\\saved_model.pb'
accuracy_train: 0.8697916666666666
accuracy_test: 0.6024759284731774
```



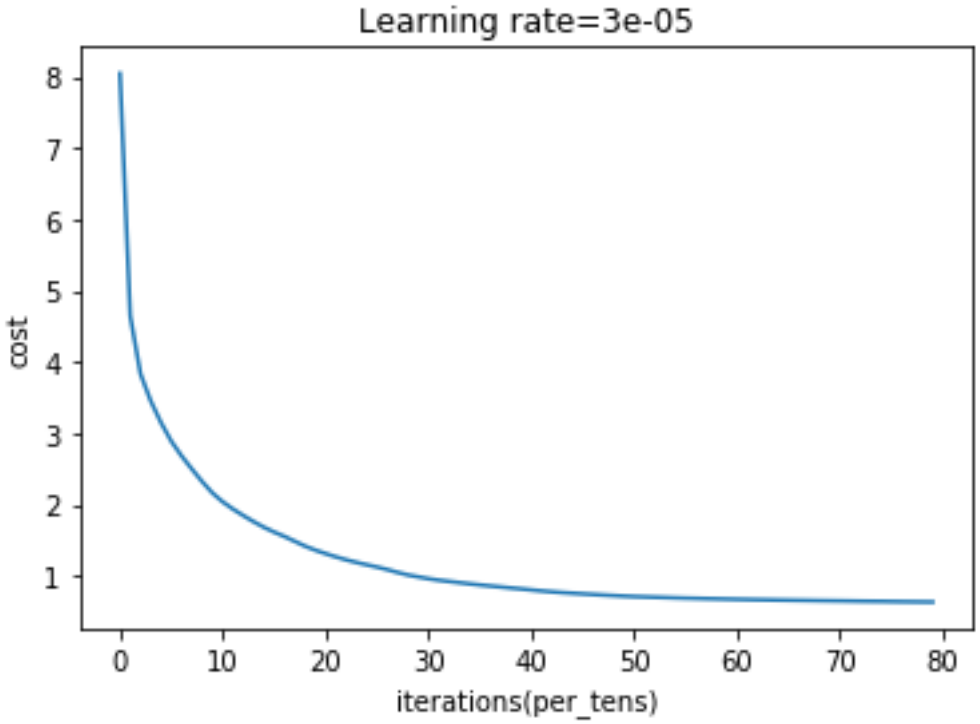
B3. 3-layer (50, 100, 200)

```
Cost after epoch 0: 3.755422
Cost after epoch 5: 2.022261
Cost after epoch 10: 1.291173
Cost after epoch 15: 1.893831
Cost after epoch 20: 1.289227
Cost after epoch 25: 0.778316
Cost after epoch 30: 0.550460
Cost after epoch 35: 0.545353
Cost after epoch 40: 0.387886
Cost after epoch 45: 0.239262
Cost after epoch 50: 0.198527
Cost after epoch 55: 0.174107
Cost after epoch 60: 0.086858
Cost after epoch 65: 0.063805
Cost after epoch 70: 0.051059
Cost after epoch 75: 0.015155
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/DNN_Python/CNN/
model040304_12132020\\saved_model.pb'
accuracy_train: 1.0
accuracy_test: 0.5928473177441541
```



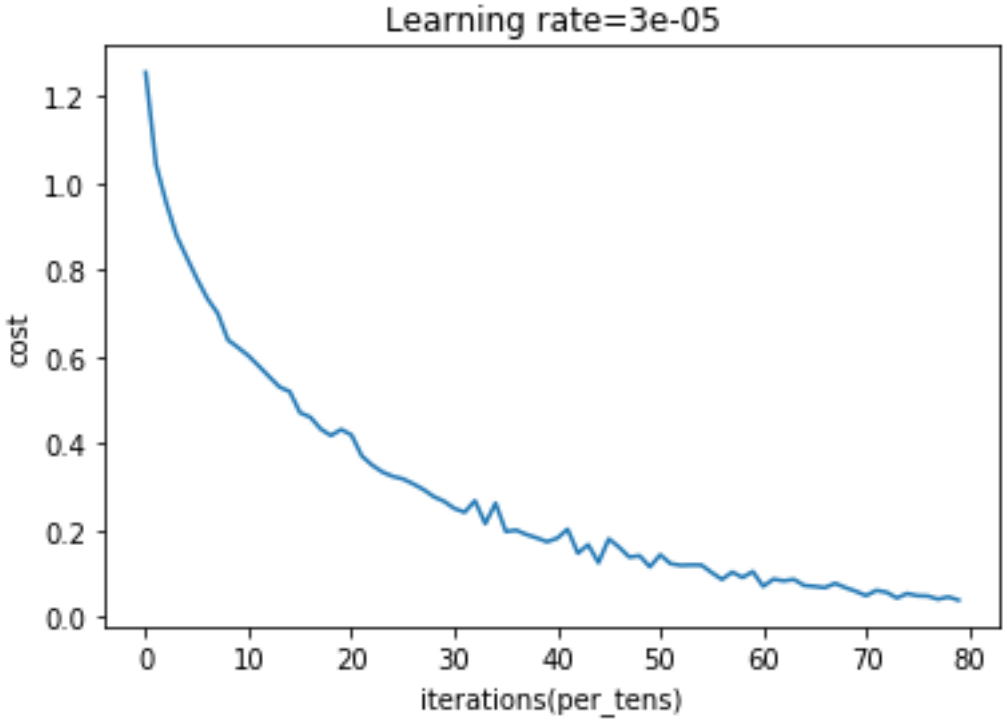
B4. 4-layer (8, 8, 8, 8)

```
Cost after epoch 0: 8.051433
Cost after epoch 5: 2.886188
Cost after epoch 10: 2.036345
Cost after epoch 15: 1.612049
Cost after epoch 20: 1.310194
Cost after epoch 25: 1.121641
Cost after epoch 30: 0.958376
Cost after epoch 35: 0.870526
Cost after epoch 40: 0.800991
Cost after epoch 45: 0.745649
Cost after epoch 50: 0.707450
Cost after epoch 55: 0.685552
Cost after epoch 60: 0.669314
Cost after epoch 65: 0.656802
Cost after epoch 70: 0.647251
Cost after epoch 75: 0.637706
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/
DNN_Python/CNN/model051129_12122020\\saved_model.pb'
accuracy_train: 0.6520833333333333
accuracy_test: 0.5089408528198074
```



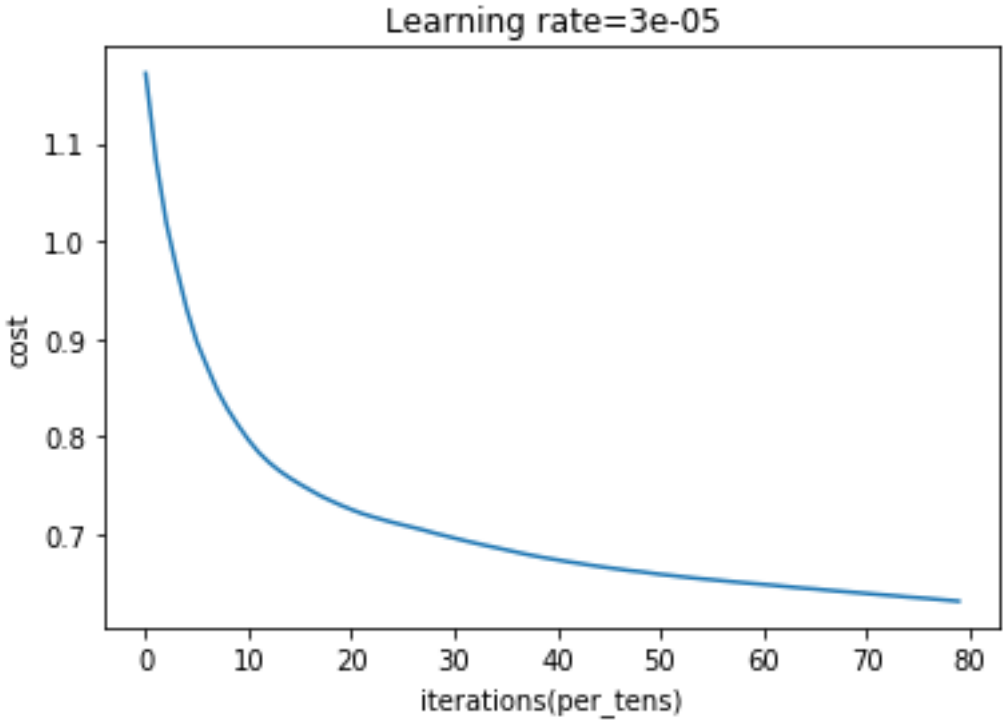
B5. 4-layer (50,100,150,200)

```
Cost after epoch 0: 1.255686
Cost after epoch 5: 0.779096
Cost after epoch 10: 0.602278
Cost after epoch 15: 0.471433
Cost after epoch 20: 0.420044
Cost after epoch 25: 0.318833
Cost after epoch 30: 0.250645
Cost after epoch 35: 0.197540
Cost after epoch 40: 0.182159
Cost after epoch 45: 0.180554
Cost after epoch 50: 0.144011
Cost after epoch 55: 0.103078
Cost after epoch 60: 0.071254
Cost after epoch 65: 0.071104
Cost after epoch 70: 0.049826
Cost after epoch 75: 0.049600
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/DNN_Python/CNN/
model160704_12132020\\saved_model.pb'
accuracy_train: 0.9989583333333333
accuracy_test: 0.6698762035763411
```



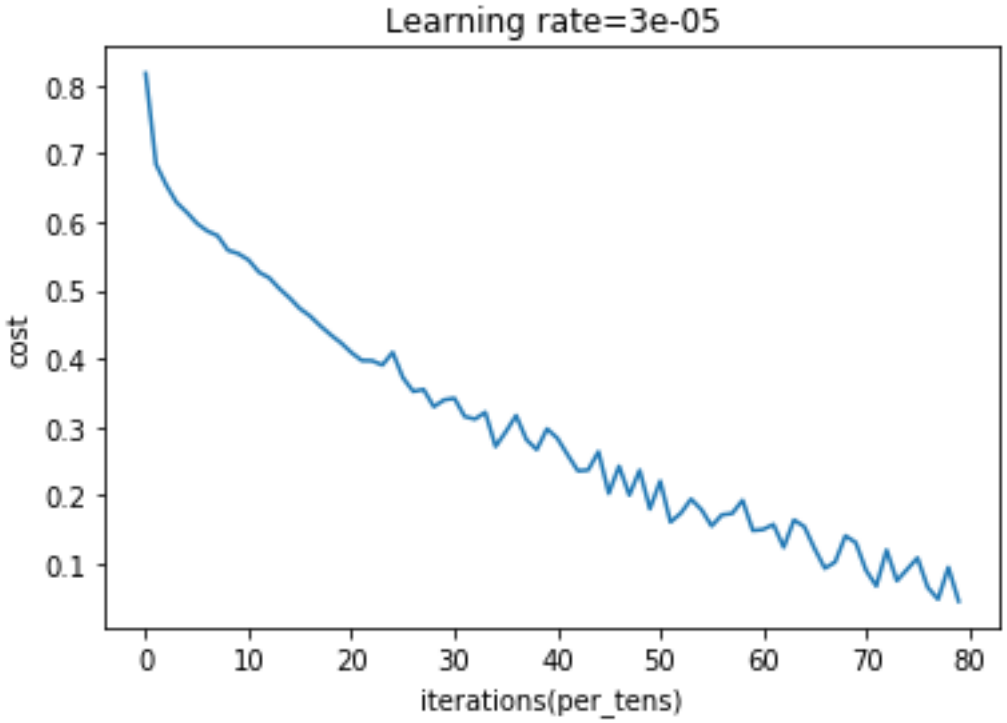
B6. 5-layer (8, 8, 8, 8, 8)

```
Cost after epoch 0: 1.171506
Cost after epoch 5: 0.896089
Cost after epoch 10: 0.796622
Cost after epoch 15: 0.751423
Cost after epoch 20: 0.725552
Cost after epoch 25: 0.709807
Cost after epoch 30: 0.696365
Cost after epoch 35: 0.684414
Cost after epoch 40: 0.674124
Cost after epoch 45: 0.666026
Cost after epoch 50: 0.659363
Cost after epoch 55: 0.653871
Cost after epoch 60: 0.649062
Cost after epoch 65: 0.644423
Cost after epoch 70: 0.639834
Cost after epoch 75: 0.635469
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/
DNN_Python/CNN/model212850_12112020\\saved_model.pb'
accuracy_train: 0.6395833333333333
accuracy_test: 0.5433287482806052
```



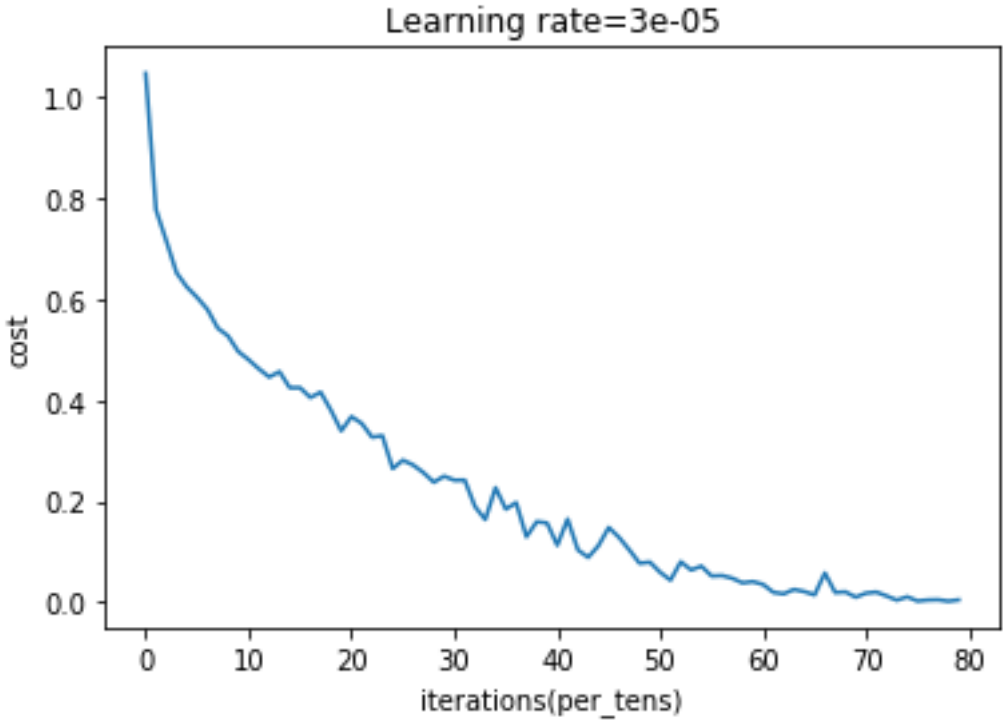
B7. 5-layer (30, 50, 100, 150, 200)

```
Cost after epoch 0: 0.817543
Cost after epoch 5: 0.597816
Cost after epoch 10: 0.544595
Cost after epoch 15: 0.473395
Cost after epoch 20: 0.409074
Cost after epoch 25: 0.372369
Cost after epoch 30: 0.342367
Cost after epoch 35: 0.293551
Cost after epoch 40: 0.284003
Cost after epoch 45: 0.203254
Cost after epoch 50: 0.221166
Cost after epoch 55: 0.155597
Cost after epoch 60: 0.150524
Cost after epoch 65: 0.122498
Cost after epoch 70: 0.090731
Cost after epoch 75: 0.109012
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/
DNN_Python/CNN/model034919_12112020\\saved_model.pb'
accuracy_train: 0.9958333333333333
accuracy_test: 0.6822558459422283
```



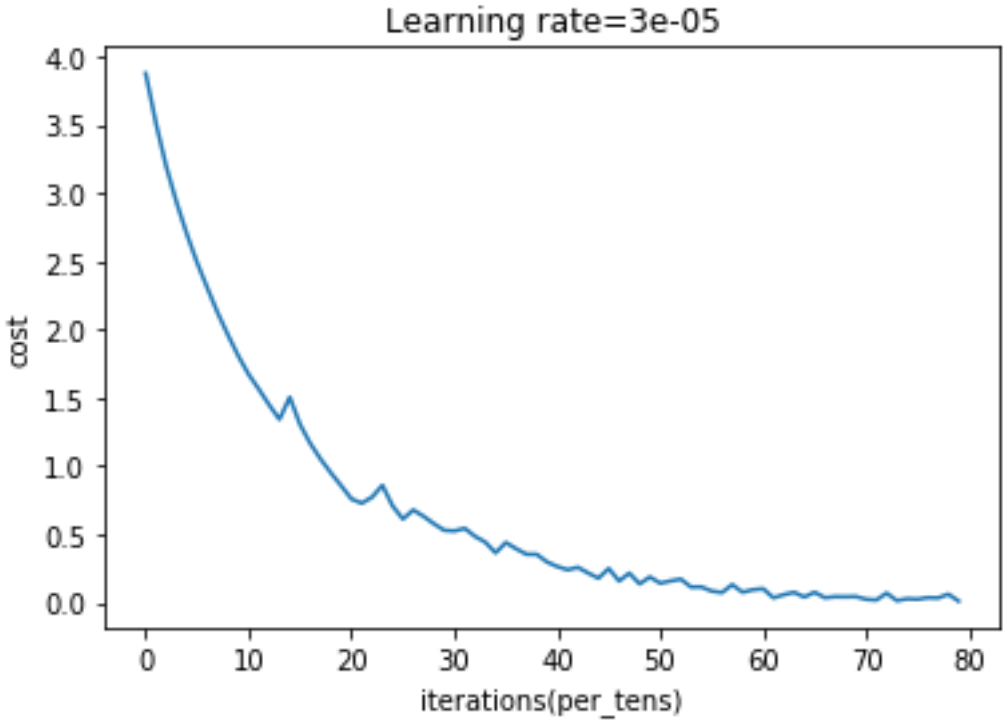
B8. 5-layer (50, 100, 150, 200, 300)

```
Cost after epoch 0: 1.048318
Cost after epoch 5: 0.604159
Cost after epoch 10: 0.480808
Cost after epoch 15: 0.425339
Cost after epoch 20: 0.367992
Cost after epoch 25: 0.281662
Cost after epoch 30: 0.241985
Cost after epoch 35: 0.184703
Cost after epoch 40: 0.113197
Cost after epoch 45: 0.148396
Cost after epoch 50: 0.059500
Cost after epoch 55: 0.052123
Cost after epoch 60: 0.035491
Cost after epoch 65: 0.014770
Cost after epoch 70: 0.018231
Cost after epoch 75: 0.002697
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/
DNN_Python/CNN/model162940_12112020\\saved_model.pb'
accuracy_train: 1.0
accuracy_test: 0.6423658872077029
```



B9. Four-layer CNN system with pooling layer

```
Cost after epoch 0: 3.877169
Cost after epoch 5: 2.490328
Cost after epoch 10: 1.673113
Cost after epoch 15: 1.305637
Cost after epoch 20: 0.759807
Cost after epoch 25: 0.613402
Cost after epoch 30: 0.526004
Cost after epoch 35: 0.441059
Cost after epoch 40: 0.265136
Cost after epoch 45: 0.251151
Cost after epoch 50: 0.143207
Cost after epoch 55: 0.085440
Cost after epoch 60: 0.102889
Cost after epoch 65: 0.077196
Cost after epoch 70: 0.025675
Cost after epoch 75: 0.025930
INFO:tensorflow:Assets added to graph.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: b'H:/Documents/DNN_Python/CNN/
model001904_12132020\\saved_model.pb'
accuracy_train: 0.996875
accuracy_test: 0.7015130674002751
```



Appendix C – Publications

- Wu, H., Soraghan, J., Lowit, A. and Di Caterina, G., 2018, July. Convolutional Neural Networks for Pathological Voice Detection. In 40th International Conference of the IEEE Engineering in Medicine and Biology Society. [12] (Conference paper contribution – Chapter 5)

- Wu, H., Soraghan, J., Lowit, A. and Di Caterina, G., 2018. A deep learning method for pathological voice detection using convolutional deep belief networks. The 19th Conference of the International Speech Communication Association - INTERSPEECH 2018. [13] (Conference paper contribution – Chapter 5)

- Wu, H., Soraghan, J., Lowit, A. and Di Caterina, G., 2020. R-Net: Robust Deep Learning for Pathological Voice Detection. Physics in medicine and biology. Submitted (Journal paper contribution – Chapter 6.3)

- Wu, H., Soraghan, J., Lowit, A. and Di Caterina, G., 2017. Development of Dysphonia Diagnostic Tools with Acoustic Analysis. In the Royal College of Speech and Language Therapists Conference 2017. (Conference poster presentation – Chapter 4)

- Wu, H., Soraghan, J., Lowit, A. and Di Caterina, G., 2021. Novel cepstral features with Bi-LSTM on pathological voice detection. IEEE Access. In preparation (Journal paper contribution – Chapter 4)

- Wu, H., Soraghan, J., Lowit, A. and Di Caterina, G., 2021. Pediatric RRP detection using transfer learning from pathological voice detection. The 28th European Signal Processing Conference (EUSIPCO 2021). In preparation (Conference paper contribution – Chapter 6.4)

References

[1] J. Mekyska, E. Janousova, P. Gomez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova*, et al.*, "Robust and complex approach of pathological speech signal analysis," *Neurocomputing,* vol. 167, pp. 94-111, 2015/11/01/ 2015.

[2] K. MacKenzie, A. Millar, J. A. Wilson, C. Sellars, and I. J. Deary, "Is voice therapy an effective treatment for dysphonia? A randomised controlled trial," *Bmj,* vol. 323, p. 658, 2001.

[3] R. T. Sataloff, "Professional singers: the science and art of clinical care," *American journal of otolaryngology,* vol. 2, pp. 251-266, 1981.

[4] A. L. Spina, R. Maunsell, K. Sandalo, R. Gusmão, and A. Crespo, "Correlation between voice and life quality and occupation," *Revista Brasileira de Otorrinolaringologia,* vol. 75, pp. 275-279, 2009.

[5] J. A. Wilson, I. J. Deary, A. Millar, and K. Mackenzie, "The quality of life impact of dysphonia," *Clinical Otolaryngology & Allied Sciences,* vol. 27, pp. 179-182, 2002.

[6] J. P. Teixeira and P. O. Fernandes, "Acoustic Analysis of Vocal Dysphonia," *Procedia Computer Science,* vol. 64, pp. 466-473, 2015.

[7] I. Smits, P. Ceuppens, and M. S. De Bodt, "A comparative study of acoustic voice measurements by means of Dr. Speech and Computerized Speech Lab," *J Voice,* vol. 19, pp. 187-96, Jun 2005.

[8] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: a meta-analysis," *J Acoust Soc Am,* vol. 126, pp. 2619-34, Nov 2009.

[9] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645-6649.

[10] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8599-8603.

[11] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice Pathology Detection Using Deep Learning: a Preliminary Study," in *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, 2017, pp. 1-4.

[12] C. J. Bassich and C. L. Ludlow, "The use of perceptual methods by new clinicians for assessing voice quality," *Journal of Speech and Hearing Disorders,* vol. 51, pp. 125-133, 1986.

[13] M. S. De Bodt, F. L. Wuyts, P. H. Van de Heyning, and C. Croux, "Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality," *Journal of Voice,* vol. 11, pp. 74-80, 1997.

[14] V. Uloza, V. Saferis, and I. Uloziene, "Perceptual and acoustic assessment of voice pathology and the efficacy of endolaryngeal phonomicrosurgery," *Journal of Voice,* vol. 19, pp. 138-145, 2005.

[15] K. Nemr, M. Simões-Zenari, G. F. Cordeiro, D. Tsuji, A. I. Ogawa, M. T. Ubrig*, et al.*, "GRBAS and Cape-V Scales: High Reliability and Consensus When Applied at Different Times," *Journal of Voice,* vol. 26, pp. 812.e17-812.e22, 2012.

[16] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS

scale," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 2478-2481.

[17] A. Stráník, R. Čmejla, and J. Vokřál, "Acoustic Parameters for Classification of Breathiness in Continuous Speech According to the GRBAS Scale," *Journal of Voice,* vol. 28, pp. 653.e9-653.e17, 2014/09/01/ 2014.

[18] G. Castellanos-Dominguez, G. Daza, L. Sanchez Giraldo, O. Castrillón, and J. Suárez Cifuentes, *Acoustic Speech Analysis for Hypernasality Detection in Children* vol. 1, 2006.

[19] G. Niedzielska, "Acoustic analysis in the diagnosis of voice disorders in children," *International Journal of Pediatric Otorhinolaryngology,* vol. 57, pp. 189-193, 3/1/ 2001.

[20] A. E. Aronson and D. M. Bless, *Clinical Voice Disorders*: Thieme, 2009.

[21] K. Verdolini, C. A. Rosen, and R. C. Branski, *Classification manual for voice disorders-I*: Psychology Press, 2014.

[22] N. Roy, K. I. Holt, S. Redmond, and H. Muntz, "Behavioral characteristics of children with vocal fold nodules," *Journal of Voice,* vol. 21, pp. 157-168, 2007.

[23] E. T. Stathopoulos and C. M. Sapienza, "Developmental changes in laryngeal and respiratory function with variations in sound pressure level," *Journal of Speech, Language, and Hearing Research,* vol. 40, pp. 595-614, 1997.

[24] D. M. Bless, L. E. Glaze, D. B. Lowery, G. Campos, and R. C. Peppard, "Stroboscopic, acoustic, aerodynamic, and perceptual analysis of voice production in normal speaking adults," *NCVS Status and Progress Report,* vol. 4, pp. 121-134, 1993.

[25] C. M. Sapienza and E. T. Stathopoulos, "Respiratory and laryngeal measures of children and women with bilateral vocal fold nodules," *Journal of Speech, Language, and Hearing Research,* vol. 37, pp. 1229-1243, 1994.

[26] M. Eiji yanagisawa, "Vocal Fold Nodule ".

[27] R. Colton, J. Casper, and R. Leonard, "Differential diagnosis of voice problems," *Understanding voice problems. A physiological perspective for diagnosis and treatment. 3rd ed. Lippincott Williams & Wilkins,* pp. 12-63, 2006.

[28] K. J. Cho, I. C. Nam, Y. S. Hwang, M. R. Shim, J. O. Park, J. H. Cho*, et al.*, "Analysis of factors influencing voice quality and therapeutic approaches in vocal polyp patients," *European archives of oto-rhino-laryngology,* vol. 268, pp. 1321-1327, 2011.

[29] M. eiji yanagisawa, "Vocal fold polyp ".

[30] M. Hirano and K. R. McCormick, "Clinical examination of voice by Minoru Hirano," ed: Acoustical Society of America, 1986.

[31] O. D. Smith, V. Callanan, J. Harcourt, and D. M. Albert, "Intracordal cyst in a neonate," *International journal of pediatric otorhinolaryngology,* vol. 52, pp. 277-281, 2000.

[32] J. Bohlender, "Diagnostic and therapeutic pitfalls in benign vocal fold diseases," *GMS current topics in otorhinolaryngology, head and neck surgery,* vol. 12, pp. Doc01-Doc01, 2013.

[33] R. H. Colton, P. Woo, D. W. Brewer, B. Griffin, and J. Casper, "Stroboscopic signs associated with benign lesions of the vocal folds," *Journal of Voice,* vol. 9, pp. 312-325, 1995.

[34] F. Reinke, "Untersuchungen uber das menschiche Stimmband," *Fortschritte Med,* vol. 13, pp. 469-478, 1895.

[35] J. M. Wood, T. Athanasiadis, and J. Allen, "Laryngitis," *Bmj,* vol. 349, p. g5827, 2014.

[36] J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, S. Aguilera-Navarro, and P. Gómez-Vilda, "An integrated tool for the diagnosis of voice disorders," *Medical Engineering & Physics,* vol. 28, pp. 276-289, 4// 2006.

[37] J. P. Dworkin, "Laryngitis: types, causes, and treatments," *Otolaryngologic clinics of North America,* vol. 41, pp. 419-436, 2008.

[38] N. G. Ordóñez, "Granular cell tumor: a review and update," *Advances in Anatomic Pathology,* vol. 6, pp. 186-203, 1999.

[39] N. G. Ordonez and B. Mackay, "Granular cell tumor: a review of the pathology and histogenesis," *Ultrastructural pathology,* vol. 23, pp. 207-222, 1999.

[40] M. Cattaruzza, P. Maisonneuve, and P. Boyle, "Epidemiology of laryngeal cancer," *European Journal of Cancer Part B: Oral Oncology,* vol. 32, pp. 293-305, 1996.

[41] E. De Stefani, P. Correa, F. Oreggia, J. Leiva, S. Rivero, G. Fernandez*, et al.*, "Risk factors for laryngeal cancer," *Cancer,* vol. 60, pp. 3087-3091, 1987.

[42] H. T. Hoffman, K. Porter, L. H. Karnell, J. S. Cooper, R. S. Weber, C. J. Langer*, et al.*, "Laryngeal cancer in the United States: changes in demographics, patterns of care, and survival," *The Laryngoscope,* vol. 116, pp. 1-13, 2006.

[43] C. E. Steuer, M. El‐Deiry, J. R. Parks, K. A. Higgins, and N. F. Saba, "An update on larynx cancer," *CA: a cancer journal for clinicians,* vol. 67, pp. 31-50, 2017.

[44] B. W. Neville, D. D. Damm, C. M. Allen, and A. C. Chi, *Oral and maxillofacial pathology*: Elsevier Health Sciences, 2015.

[45] A. Villa and S. B. Woo, "Leukoplakia—a diagnostic and management algorithm," *Journal of oral and maxillofacial surgery,* vol. 75, pp. 723-734, 2017.

[46] M. Underner, J. Perriot, and G. Peiffer, "Smokeless tobacco," *Presse medicale (Paris, France: 1983),* vol. 41, p. 3, 2012.

[47] C. A. Waldron and W. G. Shafer, "Leukoplakia revisited. A clinicopathologic study 3256 oral leukoplakias," *Cancer,* vol. 36, pp. 1386-1392, 1975.

[48] H. Klimza, J. Jackowska, M. Tokarski, K. Piersiala, and M. Wierzbicka, "Narrow-band imaging (NBI) for improving the assessment of vocal fold leukoplakia and overcoming the umbrella effect," *PLoS One,* vol. 12, p. e0180590, 2017.

[49] S. V. Stager, "Vocal fold paresis: etiology, clinical diagnosis and clinical management," *Current opinion in otolaryngology & head and neck surgery,* vol. 22, pp. 444-449, 2014.

[50] L. Sulica, "Vocal fold paresis: an evolving clinical concept," *Current Otorhinolaryngology Reports,* vol. 1, pp. 158-162, 2013.

[51] M. Aminoff, H. Dedo, and K. Izdebski, "Clinical aspects of spasmodic dysphonia," *Journal of Neurology, Neurosurgery & Psychiatry,* vol. 41, pp. 361-365, 1978.

[52] J. M. Schweinfurth, M. Billante, and M. S. Courey, "Risk factors and demographics in patients with spasmodic dysphonia," *The Laryngoscope,* vol. 112, pp. 220-223, 2002.

[53] C. L. Ludlow, C. H. Adler, G. S. Berke, S. A. Bielamowicz, A. Blitzer, S. B. Bressman*, et al.*, "Research priorities in spasmodic dysphonia," *Otolaryngology—Head and Neck Surgery,* vol. 139, pp. 495-505, 2008.

[54] G. Schlotthauer, M. E. Torres, and M. C. Jackson-Menaldi, "A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification," *J Voice,* vol. 24, pp. 346-53, May 2010.

[55] A. E. Aronson, H. W. Peterson Jr, and E. M. Litin, "Psychiatric symptomatology in functional dysphonia and aphonia," *Journal of Speech and Hearing Disorders,* vol. 31, pp. 115-127, 1966.

[56] A. O. House and H. B. Andrews, "Life events and difficulties preceding the onset of functional dysphonia," *Journal of psychosomatic research,* vol. 32, pp. 311-319, 1988.

[57] A. House and H. B. Andrews, "The psychiatric and social characteristics of patients with functional dysphonia," *Journal of Psychosomatic Research,* vol. 31, pp. 483-490, 1987.
[58] T. Eadie, A. Sroka, D. R. Wright, and A. Merati, "Does knowledge of medical diagnosis bias auditory-perceptual judgments of dysphonia?," *Journal of Voice,* vol. 25, pp. 420-429, 2011.
[59] T. L. Eadie and C. R. Baylor, "The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice," *Journal of voice,* vol. 20, pp. 527-544, 2006.
[60] N. Isshiki, H. Okamura, M. Tanabe, and M. Morimoto, "Differential diagnosis of hoarseness," *Folia Phoniatrica et Logopaedica,* vol. 21, pp. 9-19, 1969.
[61] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," *American Journal of Speech-Language Pathology,* vol. 18, pp. 124-132, 2009.
[62] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech, Language, and Hearing Research,* vol. 36, pp. 21-40, 1993.
[63] F. L. Wuyts, M. S. De Bodt, and P. H. Van de Heyning, "Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia," *Journal of Voice,* vol. 13, pp. 508-517, 1999.
[64] M. Pützer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements," *Clinical linguistics & phonetics,* vol. 22, pp. 407-420, 2008.
[65] J. J. Ballenger and J. B. Snow, *Ballenger's otorhinolaryngology: head and neck surgery*: Pmph-usa, 2003.
[66] I. R. Titze, N. C. f. Voice, and Speech, *Workshop on Acoustic Voice Analysis: Summary Statement*: National Center for Voice and Speech, 1995.
[67] J. Schoentgen, "Spectral models of additive and modulation noise in speech and phonatory excitation signals," *The Journal of the Acoustical Society of America,* vol. 113, pp. 553-562, 2003.
[68] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice," *Journal of Speech, Language, and Hearing Research,* 2001.
[69] M. Vasilakis and Y. Stylianou, "Spectral jitter modeling and estimation," *Biomedical Signal Processing and Control,* vol. 4, pp. 183-193, 2009.
[70] J. R. Orozco-Arroyave, E. A. Belalcazar-Bolanos, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz*, et al.*, "Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases," *IEEE journal of biomedical and health informatics,* vol. 19, pp. 1820-1828, 2015.
[71] J. B. Alonso, J. De Leon, I. Alonso, and M. A. Ferrer, "Automatic detection of pathologies in the voice by HOS based parameters," *EURASIP Journal on Applied Signal Processing,* vol. 4, pp. 275-284, 2001.
[72] Eskidere, #xd6, mer, #xfc, rhanl, #x131*, et al.*, "Voice Disorder Classification Based on Multitaper Mel Frequency Cepstral Coefficients Features," *Computational and Mathematical Methods in Medicine,* vol. 2015, p. 12, 2015.
[73] A. A. Dibazar, T. W. Berger, and S. S. Narayanan, "Pathological voice assessment," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 1669-1673.

[74] A. Alpan, J. Schoentgen, Y. Maryn, and F. Grenez, "Automatic perceptual categorization of disordered connected speech," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[75] K. Shama, A. Krishna, and N. U. Cholayya, "Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology," *EURASIP Journal on Advances in Signal Processing,* vol. 2007, pp. 1-9, 2006.

[76] C. Manfredi, M. d'Aniello, P. Bruscaglioni, and A. Ismaelli, "A comparative analysis of fundamental frequency estimation methods with application to pathological voices," *Medical engineering & physics,* vol. 22, pp. 135-147, 2000.

[77] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions on audio, speech, and language processing,* vol. 19, pp. 1938-1948, 2011.

[78] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki*, et al.*, "Voice pathology detection using interlaced derivative pattern on glottal source excitation," *Biomedical Signal Processing and Control,* vol. 31, pp. 156-164, 2017/01/01/ 2017.

[79] S. Hadjitodorov and P. Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening," *Medical engineering & physics,* vol. 24, pp. 419-429, 2002.

[80] C. R. Watts and S. N. Awan, "Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts," *J Speech Lang Hear Res,* vol. 54, pp. 1525-37, Dec 2011.

[81] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, J. Minelga, M. Hållander*, et al.*, "Fusing voice and query data for non-invasive detection of laryngeal disorders," *Expert Systems With Applications,* vol. 42, pp. 8445-8453, 2015.

[82] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and J. M. Gutiérrez-Arriola, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: sustained vowels and running speech," *Journal of Voice,* vol. 27, pp. 11-23, 2013.

[83] M. K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," *Biomedical Signal Processing and Control,* vol. 7, pp. 3-19, 2012/01/01/ 2012.

[84] D. Panek, A. Skalski, and J. Gajda, "Quantification of linear and non-linear acoustic analysis applied to voice pathology detection," in *Information Technologies in Biomedicine, Volume 4*, ed: Springer, 2014, pp. 355-364.

[85] T. Dubuisson, T. Dutoit, B. Gosselin, and M. Remacle, "On the use of the correlation between acoustic descriptors for the normal/pathological voices discrimination," *EURASIP Journal on advances in signal processing,* vol. 2009, p. 173967, 2009.

[86] C. Moore, K. Manickam, T. Willard, S. Jones, N. Slevin, and S. Shalet, "Spectral pattern complexity analysis and the quantification of voice normality in healthy and radiotherapy patient groups," *Medical engineering & physics,* vol. 26, pp. 291-301, 2004.

[87] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *The Journal of the Acoustical Society of America,* vol. 103, pp. 1628-1639, 1998.

[88] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network

based detectors," *IEEE Transactions on Biomedical Engineering,* vol. 51, pp. 380-384, 2004.

[89] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, "An investigation of vocal tract characteristics for acoustic discrimination of pathological voices," *BioMed research international,* vol. 2013, 2013.

[90] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *Journal of Voice,* vol. 24, pp. 47-56, 2010.

[91] M. Vasilakis and Y. Stylianou, "Voice pathology detection based eon short-term jitter estimations in running speech," *Folia Phoniatrica et Logopaedica,* vol. 61, pp. 153-170, 2009.

[92] C. J. Moore, K. Manickam, and N. Slevin, "Collective spectral pattern complexity analysis of voicing in normal males and larynx cancer patients following radiotherapy," *Biomedical Signal Processing and Control,* vol. 1, pp. 113-119, 2006.

[93] G. Daza-Santacoloma, J. D. Arias-Londono, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruíz, and G. Castellanos-Dominguez, "Dynamic feature extraction: an application to voice pathology detection," *Intelligent Automation & Soft Computing,* vol. 15, pp. 667-682, 2009.

[94] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Computer Methods and Programs in Biomedicine,* vol. 91, pp. 36-47, 2008.

[95] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering,* vol. 53, pp. 468-477, 2006.

[96] T. A. Mesallam, M. Farahat, K. H. Malki, M. Alsulaiman, Z. Ali, A. Al-Nasheri*, et al.*, "Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms," *Journal of healthcare engineering,* vol. 2017, 2017.

[97] R. Fraile, N. Saenz-Lechon, J. Godino-Llorente, V. Osma-Ruiz, and C. Fredouille, "Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex," *Folia phoniatrica et logopaedica,* vol. 61, pp. 146-152, 2009.

[98] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, "Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia)," 2005.

[99] S. Jothilakshmi, "Automatic system to detect the type of voice pathology," *Applied Soft Computing,* vol. 21, pp. 244-249, 2014.

[100] G. Muhammad and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomedical Signal Processing and Control,* vol. 11, pp. 1-9, 2014/05/01/ 2014.

[101] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, A. Mahmood, and M. Alsulaiman, "Multidirectional Regression (MDR)-Based Features for Automatic Voice Disorder Detection," *Journal of Voice,* vol. 26, pp. 817.e19-817.e27, 2012/11/01/ 2012.

[102] F. Amara, M. Fezari, and H. Bourouba, "An improved GMM-SVM system based on distance metric for voice pathology detection," *Appl. Math,* vol. 10, pp. 1061-1070, 2016.

[103] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using

complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering,* vol. 58, pp. 370-379, 2010.

[104] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters," *IEEE Transactions on Biomedical Engineering,* vol. 53, pp. 1943-1953, 2006.

[105] D. Martínez, E. Lleida, A. Ortega, and A. Miguel, "Score Level versus Audio Level Fusion for Voice Pathology Detection on the Saarbrücken Voice Database," in *Advances in Speech and Language Technologies for Iberian Languages: IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández*, et al.*, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 110-120.

[106] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit," in *Advances in Speech and Language Technologies for Iberian Languages: IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández*, et al.*, Eds., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 99-109.

[107] G. Muhammad, M. F. Alhamid, M. Alsulaiman, and B. Gupta, "Edge Computing with Cloud for Voice Disorder Assessment and Treatment," *IEEE Communications Magazine,* vol. 56, pp. 60-65, 2018.

[108] G. Muhammad, M. Alhamid, M. Hossain, A. Almogren, and A. Vasilakos, "Enhanced Living by Assessing Voice Pathology Using a Co-Occurrence Matrix," *Sensors,* vol. 17, p. 267, 2017.

[109] M. Hariharan, K. Polat, R. Sindhu, and S. Yaacob, "A hybrid expert system approach for telemonitoring of vocal fold pathology," *Applied Soft Computing,* vol. 13, pp. 4148-4161, 2013/10/01/ 2013.

[110] M. Hariharan, K. Polat, and S. Yaacob, "A new feature constituting approach to detection of vocal fold pathology," *International Journal of Systems Science,* vol. 45, pp. 1622-1634, 2014/08/03 2014.

[111] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, 2002, pp. 182-183 vol.1.

[112] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin*, et al.*, "Detection of Pathological Voice Using Cepstrum Vectors: A Deep Learning Approach," *Journal of Voice,* 2018/03/19/ 2018.

[113] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and Z. Smekal, "Towards robust voice pathology detection," *Neural Computing and Applications,* pp. 1-11, 2018.

[114] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Díaz-de-María, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE transactions on audio, speech, and language processing,* vol. 17, pp. 1186-1195, 2009.

[115] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, and V. Uloza, "Exploring similarity-based classification of larynx disorders from human voice," *Speech Communication,* vol. 54, pp. 601-610, 2012.

[116] G. Muhammad, S. M. M. Rahman, A. Alelaiwi, and A. Alamri, "Smart health solution integrating IoT and cloud: A case study of voice pathology monitoring," *IEEE Communications Magazine,* vol. 55, pp. 69-73, 2017.

[117] D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza*, et al.*, "Using ambulatory voice monitoring to investigate common voice disorders: Research update," *Frontiers in bioengineering and biotechnology,* vol. 3, p. 155, 2015.

[118] P. R. Scalassara, M. E. Dajer, C. D. Maciel, R. C. Guido, and J. C. Pereira, "Relative entropy measures applied to healthy and pathological voice characterization," *Applied Mathematics and Computation,* vol. 207, pp. 95-108, 2009.

[119] J. I. Godino-Llorente, P. Gómez-Vilda, F. Cruz-Roldán, M. Blanco-Velasco, and R. Fraile, "Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness," *Journal of Voice,* vol. 24, pp. 667-677, 2010.

[120] R. F. Leonarduzzi, G. A. Alzamendi, G. Schlotthauer, and M. E. Torres, "Wavelet leader multifractal analysis of period and amplitude sequences from sustained vowels," *Speech Communication,* vol. 72, pp. 1-12, 2015.

[121] D. D. Mehta, M. Zanartu, S. W. Feng, H. A. Cheyne II, and R. E. Hillman, "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform," *IEEE Transactions on Biomedical Engineering,* vol. 59, pp. 3090-3096, 2012.

[122] M. Petrović-Lazić, S. Babac, M. Vuković, R. Kosanović, and Z. Ivanković, "Acoustic voice analysis of patients with vocal fold polyp," *journal of Voice,* vol. 25, pp. 94-97, 2011.

[123] A. Rovirosa, E. Martínez-Celdrán, A. Ortega, C. Ascaso, R. Abellana, M. Velasco*, et al.*, "Acoustic analysis after radiotherapy in T1 vocal cord carcinoma: a new approach to the analysis of voice quality," *International Journal of Radiation Oncology* Biology* Physics,* vol. 47, pp. 73-79, 2000.

[124] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta*, et al.*, "Categorizing normal and pathological voices: automated and perceptual categorization," *Journal of Voice,* vol. 25, pp. 700-708, 2011.

[125] M. Döllinger, M. Kunduk, M. Kaltenbacher, S. Vondenhoff, A. Ziethe, U. Eysholdt*, et al.*, "Analysis of vocal fold function from acoustic data simultaneously recorded with high-speed endoscopy," *Journal of Voice,* vol. 26, pp. 726-733, 2012.

[126] A. Al-nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat*, et al.*, "An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification," *Journal of Voice,* vol. 31, pp. 113.e9-113.e18.

[127] L. Verde, G. De Pietro, and G. Sannino, "A methodology for voice classification based on the personalized fundamental frequency estimation," *Biomedical Signal Processing and Control,* vol. 42, pp. 134-144, 2018.

[128] A. M. Sulter, H. P. Wit, H. K. Schutte, and D. G. Miller, "A structured approach to voice range profile (phonetogram) analysis," *Journal of Speech, Language, and Hearing Research,* vol. 37, pp. 1076-1085, 1994.

[129] A. Behrman, C. J. Agresti, E. Blumstein, and G. Sharma, "Meaningful features of voice range profiles from patients with organic vocal fold pathology: a preliminary study," *Journal of Voice,* vol. 10, pp. 269-283, 1996.

[130] E. Ma, J. Robertson, C. Radford, S. Vagne, R. El-Halabi, and E. Yiu, "Reliability of speaking and maximum voice range measures in screening for dysphonia," *Journal of Voice,* vol. 21, pp. 397-406, 2007.

[131] A. E. Hallin, K. Fröst, E. B. Holmberg, and M. Södersten, "Voice and speech range profiles and Voice Handicap Index for males—methodological issues and data," *Logopedics Phoniatrics Vocology,* vol. 37, pp. 47-61, 2012.
[132] J. Goddard, G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "Dimensionality reduction for visualization of normal and pathological speech data," *Biomedical Signal Processing and Control,* vol. 4, pp. 194-201, 2009/07/01/ 2009.
[133] P. Aichinger, I. Roesner, B. Schneider-Stickler, M. Leonhard, D.-M. Denk-Linnert, W. Bigenzahn*, et al.*, "Towards objective voice assessment: the diplophonia diagram," *Journal of voice,* vol. 31, pp. 253. e17-253. e26, 2017.
[134] P. Aichinger, I. Roesner, M. Leonhard, B. Schneider-Stickler, D.-M. Denk-Linnert, W. Bigenzahn*, et al.*, "Comparison of an audio-based and a video-based approach for detecting diplophonia," *Biomedical Signal Processing and Control,* vol. 31, pp. 576-585, 2017.
[135] M. Farrus and J. Hernando, "Using jitter and shimmer in speaker verification," *IET Signal Processing,* vol. 3, pp. 247-257, 2009.
[136] M. de Oliveira Rosa, J. C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Transactions on Biomedical Engineering,* vol. 47, pp. 96-104, 2000.
[137] M. Kohler, M. M. Vellasco, and E. Cataldo, "Analysis and classification of voice pathologies using glottal signal parameters," *Journal of Voice,* vol. 30, pp. 549-556, 2016.
[138] A. Giovanni, M. Ouaknine, and J.-M. Triglia, "Determination of largest Lyapunov exponents of vocal signal: application to unilateral laryngeal paralysis," *Journal of Voice,* vol. 13, pp. 341-354, 1999.
[139] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. De Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels," *J Voice,* vol. 24, pp. 540-55, Sep 2010.
[140] M. Shu, J. J. Jiang, and M. Willey, "The effect of moving window on acoustic analysis," *Journal of Voice,* vol. 30, pp. 5-10, 2016.
[141] V. Uloza, E. Padervinskis, A. Vegiene, R. Pribuisiene, V. Saferis, E. Vaiciukynas*, et al.*, "Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening," *European Archives of Oto-rhino-laryngology,* vol. 272, pp. 3391-3399, 2015.
[142] M. l. N. Vieira, F. R. McInnes, and M. A. Jack, "On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures," *The Journal of the Acoustical Society of America,* vol. 111, pp. 1045-1055, 2002.
[143] Y. Zhang, J. J. Jiang, L. Biazzo, and M. Jorgensen, "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis," *Journal of Voice,* vol. 19, pp. 519-528, 2005.
[144] Y. Zhang and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies," *Journal of Voice,* vol. 22, pp. 1-9, 2008.
[145] Y. Zhang, C. McGilligan, L. Zhou, M. Vig, and J. J. Jiang, "Nonlinear dynamic analysis of voices before and after surgical excision of vocal polyps," *The Journal of the Acoustical Society of America,* vol. 115, pp. 2270-2277, 2004.
[146] P. Lieberman, "Perturbations in vocal pitch," *The Journal of the Acoustical Society of America,* vol. 33, pp. 597-603, 1961.
[147] H. Hollien, J. Michel, and E. T. Doherty, "A method for analyzing vocal jitter in sustained phonation," *Journal of phonetics,* vol. 1, pp. 85-91, 1973.

[148] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke, "Comparison of voice analysis systems for perturbation measurement," *Journal of Speech, Language, and Hearing Research,* vol. 39, pp. 126-134, 1996.

[149] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics‐to‐noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America,* vol. 71, pp. 1544-1550, 1982.

[150] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," *The Journal of the Acoustical Society of America,* vol. 87, pp. 2218-2224, 1990.

[151] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *The Journal of the Acoustical Society of America,* vol. 80, pp. 1329-1334, 1986.

[152] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of speech, language, and hearing research,* vol. 43, pp. 469-485, 2000.

[153] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio–a new measure for describing pathological voices," *Acta Acustica united with Acustica,* vol. 83, pp. 700-706, 1997.

[154] C. Peng, W. Chen, X. Zhu, B. Wan, and D. Wei, "Pathological voice classification based on a single Vowel's acoustic features," in *7th IEEE International Conference on Computer and Information Technology (CIT 2007)*, 2007, pp. 1106-1110.

[155] W. S. Winholtz and L. O. Ramig, "Vocal tremor analysis with the vocal demodulator," *Journal of speech, language, and hearing research,* vol. 35, pp. 562-573, 1992.

[156] P. Mitev and S. Hadjitodorov, "A method for turbulent noise estimation in voiced signals," *Medical and Biological Engineering and Computing,* vol. 38, pp. 625-631, 2000.

[157] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Automatic Voice Pathology Detection With Running Speech by Using Estimation of Auditory Spectrum and Cepstral Coefficients Based on the All-Pole Model," *Journal of Voice,* vol. 30, pp. 757.e7-757.e19, 2016/11/01/ 2016.

[158] T. L. Eadie and P. C. Doyle, "Classification of dysphonic voice: acoustic and auditory-perceptual measures," *Journal of Voice,* vol. 19, pp. 1-14, 2005.

[159] S. Y. Lowell, R. H. Colton, R. T. Kelley, and Y. C. Hahn, "Spectral-and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker," *Journal of Voice,* vol. 25, pp. e223-e232, 2011.

[160] A. Akbari and M. K. Arjmandi, "Employing linear prediction residual signal of wavelet sub-bands in automatic detection of laryngeal pathology," *Biomedical Signal Processing and Control,* vol. 18, pp. 293-302, 2015.

[161] Z. Ali, G. Muhammad, and M. F. Alhamid, "An automatic health monitoring system for patients suffering from voice complications in smart cities," *IEEE Access,* vol. 5, pp. 3900-3908, 2017.

[162] M. Alsulaiman, "Voice pathology assessment systems for dysphonic patients: detection, classification, and speech recognition," *IETE Journal of Research,* vol. 60, pp. 156-167, 2014.

[163] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. Gómez-Vilda, "Lpc, LPCC and MFCC parameterisation applied to the detection of voice impairments," in *Sixth International Conference on Spoken Language Processing*, 2000.

[164] G. Muhammad, G. Altuwaijri, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat*, et al.*, "Automatic voice pathology detection and classification using vocal tract area

irregularity," *Biocybernetics and Biomedical Engineering,* vol. 36, pp. 309-317, 2016.

[165] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z. Ali, "Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions," *Journal of Voice,* vol. 31, pp. 3-15.

[166] J. R. Orozco‐Arroyave, F. Hönig, J. D. Arias‐Londoño, J. F. Vargas‐Bonilla, and E. Nöth, "Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words," *Expert Systems,* vol. 32, pp. 688-697, 2015.

[167] L. Moro-Velazquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease," *Applied Soft Computing,* vol. 62, pp. 649-666, 2018.

[168] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, A. Al-nasheri, T. A. Mesallam*, et al.*, "Intra-and inter-database study for arabic, english, and german databases: do conventional speech features detect voice pathology?," *Journal of Voice,* vol. 31, pp. 386. e1-386. e8, 2017.

[169] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology,* vol. 36, pp. 60-69, 2011.

[170] H. Cordeiro, J. Fonseca, I. Guimarães, and C. Meneses, "Hierarchical classification and system combination for automatically identifying physiological and neuromuscular laryngeal pathologies," *Journal of voice,* vol. 31, pp. 384. e9-384. e14, 2017.

[171] V. Majidnezhad, "A novel hybrid of genetic algorithm and ANN for developing a high efficient method for vocal fold pathology diagnosis," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2015, p. 3, 2015.

[172] N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control,* vol. 1, pp. 120-128, 2006.

[173] X. Wang, J. Zhang, and Y. Yan, "Discrimination between pathological and normal voices using GMM-SVM approach," *Journal of Voice,* vol. 25, pp. 38-43, 2011.

[174] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," *Multimedia Systems,* vol. 25, pp. 565-575, 2019.

[175] J. C. Saldanha, T. Ananthakrishna, and R. Pinto, "Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features," *Journal of medical imaging and health informatics,* vol. 4, pp. 168-173, 2014.

[176] S. Magre and R. Deshmukh, "A review on feature extraction and noise reduction technique," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 4, pp. 352-356, 2014.

[177] N. Moritz, J. Anemüller, and B. Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, pp. 1926-1937, 2015.

[178] U. Sharma, S. Maheshkar, and A. Mishra, "Study of robust feature extraction techniques for speech recognition system," in *2015 International conference on futuristic trends on computational analysis and knowledge management (ABLAZE)*, 2015, pp. 654-658.

[179] A. Alpan, Y. Maryn, A. Kacha, F. Grenez, and J. Schoentgen, "Multi-band dysperiodicity analyses of disordered connected speech," *Speech Communication,* vol. 53, pp. 131-141, 2011.
[180] L. F. Brinca, A. P. F. Batista, A. I. Tavares, I. C. Gonçalves, and M. L. Moreno, "Use of cepstral analyses for differentiating normal from dysphonic voices: A comparative study of connected speech versus sustained vowel in European Portuguese female speakers," *Journal of Voice,* vol. 28, pp. 282-286, 2014.
[181] B. R. Kumar, J. S. Bhat, and N. Prasad, "Cepstral analysis of voice in persons with vocal nodules," *Journal of Voice,* vol. 24, pp. 651-653, 2010.
[182] S. N. Awan, N. Roy, and C. Dromey, "Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectral/cepstral model," *Clinical Linguistics & Phonetics,* vol. 23, pp. 825-841, 2009/11/01 2009.
[183] L. Moro-Velázquez, J. A. Gómez-García, and J. I. Godino-Llorente, "Voice pathology detection using modulation spectrum-optimized metrics," *Frontiers in bioengineering and biotechnology,* vol. 4, p. 1, 2016.
[184] A. V. Oppenheim, "Speech spectrograms using the fast Fourier transform," *IEEE spectrum,* vol. 7, pp. 57-62, 1970.
[185] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *arXiv preprint arXiv:1706.09559,* 2017.
[186] S. Cheung and J. S. Lim, "Combined multi-resolution (wideband/narrowband) spectrogram," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1991, pp. 457-460.
[187] A. Akbari and M. K. Arjmandi, "An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features," *Biomedical Signal Processing and Control,* vol. 10, pp. 209-223, 2014/03/01/ 2014.
[188] R. T. S. Carvalho, C. C. Cavalcante, and P. C. Cortez, "Wavelet transform and artificial neural networks applied to voice disorders identification," in *2011 Third World Congress on Nature and Biologically Inspired Computing*, 2011, pp. 371-376.
[189] S. C. Olhede and A. T. Walden, "Generalized morse wavelets," *IEEE Transactions on Signal Processing,* vol. 50, pp. 2661-2670, 2002.
[190] J. M. Lilly and S. C. Olhede, "Higher-order properties of analytic wavelets," *IEEE Transactions on Signal Processing,* vol. 57, pp. 146-160, 2008.
[191] J. M. Lilly and S. C. Olhede, "On the analytic wavelet transform," *IEEE Transactions on Information Theory,* vol. 56, pp. 4135-4156, 2010.
[192] J. M. Lilly and S. C. Olhede, "Generalized Morse wavelets as a superfamily of analytic wavelets," *IEEE Transactions on Signal Processing,* vol. 60, pp. 6036-6041, 2012.
[193] I. R. Titze and W. S. Winholtz, "Effect of microphone type and placement on voice perturbation measurements," *Journal of Speech, Language, and Hearing Research,* vol. 36, pp. 1177-1190, 1993.
[194] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of children, women, and men: The effects of vocal intensity variation," *The Journal of the Acoustical Society of America,* vol. 106, pp. 1532-1542, 1999.
[195] M. Eye and E. Infirmary, "Voice disorders database, version. 1.03 (cd-rom)," *Lincoln Park, NJ: Kay Elemetrics Corporation,* 1994.
[196] B. Woldert-Jokisz, "Saarbruecken voice database," 2007.
[197] N. Malyska, T. F. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically-inspired amplitude-modulation features," in *Proceedings. (ICASSP*

'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., 2005, pp. I/873-I/876 Vol. 1.

[198] M. K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, and A. Moqarehzadeh, "Identification of voice disorders using long-time features and support vector machine with different feature reduction methods," *Journal of Voice,* vol. 25, pp. e275-e289, 2011.

[199] H. Ghasemzadeh, M. T. Khass, M. K. Arjmandi, and M. Pooyan, "Detection of vocal disorders based on phase space parameters and Lyapunov spectrum," *biomedical signal processing and control,* vol. 22, pp. 135-145, 2015.

[200] P. Gómez, E. San Segundo, L. M. Mazaira, A. Álvarez, and V. Rodellar, "Using Dysphonic Voice to Characterize Speaker's Biometry," *Language and Law= Linguagem e Direito,* vol. 1, 2017.

[201] P. Gómez-Vilda, R. Fernández-Baillo, A. Nieto, F. Díaz, F. Fernández-Camacho, V. Rodellar*, et al.*, "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters," *Journal of Voice,* vol. 21, pp. 450-476, 2007.

[202] D. Hemmerling, A. Skalski, and J. Gajda, "Voice data mining for laryngeal pathology assessment," *Computers in Biology and Medicine,* vol. 69, pp. 270-276, 2016/02/01/ 2016.

[203] L. Matassini, R. Hegger, H. Kantz, and C. Manfredi, "Analysis of vocal disorders in a feature space," *Medical engineering & physics,* vol. 22, pp. 413-418, 2000.

[204] D. Panek, A. Skalski, J. Gajda, and R. Tadeusiewicz, "Acoustic analysis assessment in speech pathology detection," *International Journal of Applied Mathematics and Computer Science,* vol. 25, pp. 631-643, 2015.

[205] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "An improved method for voice pathology detection by means of a HMM-based feature space transformation," *Pattern recognition,* vol. 43, pp. 3100-3112, 2010.

[206] R. Behroozmand and F. Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients' speech signal with unilateral vocal fold paralysis," *Computers in Biology and Medicine,* vol. 37, pp. 474-485, 2007.

[207] H. K. Heris, B. S. Aghazadeh, and M. Nikkhah-Bahrami, "Optimal feature selection for the assessment of vocal fold disorders," *Computers in Biology and Medicine,* vol. 39, pp. 860-868, 2009.

[208] A. Verikas, A. Gelzinis, M. Bacauskiene, M. Hållander, V. Uloza, and M. Kaseta, "Combining image, voice, and the patient's questionnaire data to categorize laryngeal disorders," *Artificial intelligence in medicine,* vol. 49, pp. 43-50, 2010.

[209] Z. Ali, M. Talha, and M. Alsulaiman, "A Practical Approach: Design and Implementation of a Healthcare Software for Screening of Dysphonic Patients," *IEEE Access,* vol. 5, pp. 5844-5857, 2017.

[210] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients' speech signals using nonlinear dynamical analysis," *Computers in biology and medicine,* vol. 40, pp. 54-63, 2010.

[211] T. Drugman, T. Dubuisson, and T. Dutoit, "On the mutual information between source and filter contributions for voice pathology detection," *arXiv preprint arXiv:2001.00583,* 2020.

[212] Z. Ali, M. Alsulaiman, I. Elamvazuthi, G. Muhammad, T. A. Mesallam, M. Farahat*, et al.*, "Voice pathology detection based on the modified voice contour and SVM," *Biologically Inspired Cognitive Architectures,* vol. 15, pp. 10-18, 2016.

[213] Z. Ali, I. Elamvazuthi, M. Alsulaiman, and G. Muhammad, "Detection of voice pathology using fractal dimension in a multiresolution analysis of normal and disordered speech signals," *Journal of medical systems,* vol. 40, p. 20, 2016.
[214] C. Middag, Y. Saeys, and J.-P. Martens, "Towards an ASR-free objective analysis of pathological speech," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
[215] P. Saidi and F. Almasganj, "Voice disorder signal classification using m-band wavelets and support vector machine," *Circuits, Systems, and Signal Processing,* vol. 34, pp. 2727-2738, 2015.
[216] S. Shilaskar, A. Ghatol, and P. Chatur, "Medical decision support system for extremely imbalanced datasets," *Information Sciences,* vol. 384, pp. 205-219, 2017.
[217] C. M. Travieso, J. B. Alonso, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, E. Nöth, and A. G. Ravelo-García, "Detection of different voice diseases based on the nonlinear characterization of speech signals," *Expert Systems with Applications,* vol. 82, pp. 184-195, 2017.
[218] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Z. Kons, A. Satt*, et al.*, "Fusion of voice signal information for detection of mild laryngeal pathology," *Applied Soft Computing,* vol. 18, pp. 91-103, 2014.
[219] V. Majidnezhad, "A HTK-based method for detecting vocal fold pathology," *Acta Informatica Medica,* vol. 22, p. 246, 2014.
[220] C. D. P. Crovato and A. Schuck, "The Use of Wavelet Packet Transform and Artificial Neural Networks in Analysis and Classification of Dysphonic Voices," *IEEE Transactions on Biomedical Engineering,* vol. 54, pp. 1898-1900, 2007.
[221] A. Verikas, A. Gelzinis, E. Vaiciukynas, M. Bacauskiene, J. Minelga, M. Hållander*, et al.*, "Data dependent random forest applied to screening for laryngeal disorders through analysis of sustained phonation: acoustic versus contact microphone," *Medical engineering & physics,* vol. 37, pp. 210-218, 2015.
[222] M. Kaleem, B. Ghoraani, A. Guergachi, and S. Krishnan, "Pathological speech signal analysis and classification using empirical mode decomposition," *Medical & biological engineering & computing,* vol. 51, pp. 811-821, 2013.
[223] K. Umapathy and S. Krishnan, "Feature analysis of pathological speech signals using local discriminant bases technique," *Medical and Biological Engineering and Computing,* vol. 43, pp. 457-464, 2005.
[224] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Transactions on Biomedical Engineering,* vol. 52, pp. 421-430, 2005.
[225] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomedical engineering online,* vol. 6, p. 23, 2007.
[226] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "Convolutional neural networks for pathological voice detection," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1-4.
[227] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," *Interspeech 2018,* 2018.
[228] M. Markaki and Y. Stylianou, "Normalized modulation spectral features for cross-database voice pathology detection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.
[229] A. Al-nasheri, Z. Ali, G. Muhammad, and M. Alsulaiman, "Voice pathology detection using auto-correlation of different filters bank," in *2014 IEEE/ACS 11th*

International Conference on Computer Systems and Applications (AICCSA), 2014, pp. 50-55.

[230] N. Souissi and A. Cherif, "Dimensionality reduction for voice disorders identification system based on Mel Frequency Cepstral Coefficients and Support Vector Machine," in *2015 7th International Conference on Modelling, Identification and Control (ICMIC)*, 2015, pp. 1-6.

[231] N. Souissi and A. Cherif, "Speech recognition system based on short-term cepstral parameters, feature reduction method and Artificial Neural Networks," in *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 2016, pp. 667-671.

[232] M. Dahmani and M. Guerti, "Vocal folds pathologies classification using Naïve Bayes Networks," in *2017 6th international conference on systems and control (ICSC)*, 2017, pp. 426-432.

[233] K. Wu, D. Zhang, G. Lu, and Z. Guo, "Joint learning for voice based disease detection," *Pattern Recognition,* vol. 87, pp. 130-139, 2019.

[234] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," *Biomedical Signal Processing and Control,* vol. 51, pp. 181-199, 2019/05/01/ 2019.

[235] H. Guan and A. Lerch, "Learning Strategies for Voice Disorder Detection," in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 295-301.

[236] L. Y. Pratt, "Discriminability-based transfer between neural networks," in *Advances in neural information processing systems*, 1993, pp. 204-211.

[237] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering,* vol. 22, pp. 1345-1359, 2009.

[238] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel*, et al.*, "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging,* vol. 35, pp. 1160-1169, 2016.

[239] M. Winkels and T. S. Cohen, "3D G-CNNs for pulmonary nodule detection," *arXiv preprint arXiv:1804.04656,* 2018.

[240] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in *International MICCAI Brainlesion Workshop*, 2018, pp. 61-72.

[241] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing,* vol. 321, pp. 321-331, 2018.

[242] J. Nalepa, M. Marcinkiewicz, and M. Kawulok, "Data augmentation for brain-tumor segmentation: A review," *Frontiers in Computational Neuroscience,* vol. 13, 2019.

[243] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE transactions on medical imaging,* vol. 35, pp. 1240-1251, 2016.

[244] J. Sietsma and R. J. Dow, "Creating artificial neural networks that generalize," *Neural networks,* vol. 4, pp. 67-79, 1991.

[245] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research,* vol. 11, 2010.

[246] B. Poole, J. Sohl-Dickstein, and S. Ganguli, "Analyzing noise in autoencoders and deep networks," *arXiv preprint arXiv:1406.1831,* 2014.
[247] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research,* vol. 15, pp. 1929-1958, 2014.
[248] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*: MIT press, 2016.
[249] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological cybernetics,* vol. 59, pp. 291-294, 1988.
[250] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144-152.
[251] L. Wang, Z. Zhang, and C. X. R. C. Design, "Theory and applications," *Support Vector Machines, Springer-Verlag, Berlin Heidelber,* 2005.
[252] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data mining techniques for the life sciences*, ed: Springer, 2010, pp. 223-239.
[253] M. Awad and R. Khanna, *Efficient learning machines: theories, concepts, and applications for engineers and system designers*: Springer Nature, 2015.
[254] S. L. Christina, P. Vijayalakshmi, and T. Nagarajan, "HMM-based speech recognition system for the dysarthric speech evaluation of articulatory subsystem," in *2012 International Conference on Recent Trends in Information Technology*, 2012, pp. 54-59.
[255] P. Sujatha and M. R. Krishnan, "Lip feature extraction for visual speech recognition using Hidden Markov Model," in *2012 International Conference on Computing, Communication and Applications*, 2012, pp. 1-5.
[256] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, 2000, pp. 1315-1318.
[257] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227-230.
[258] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 216-221.
[259] P. D. Polur and G. E. Miller, "Experiments with fast Fourier transform, linear predictive and cepstral coefficients in dysarthric speech recognition algorithms using hidden Markov model," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 13, pp. 558-561, 2005.
[260] C. Chakraborty and P. H. Talukdar, "Issues and Limitations of HMM in Speech Processing: A Survey," *International Journal of Computer Applications,* vol. 141, pp. 13-17, 05/17 2016.
[261] Y. Qian, L. Ying, and J. Pingping, "Speech emotion recognition using supervised manifold learning based on all-class and pairwise-class feature extraction," in *IEEE Conference Anthology*, 2013, pp. 1-5.
[262] S. A. Rieger, R. Muraleedharan, and R. P. Ramachandran, "Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers," in *The 9th International Symposium on Chinese Spoken Language Processing*, 2014, pp. 589-593.
[263] T.-L. Pao, W.-Y. Liao, T.-N. Wu, and C.-Y. Lin, "Automatic visual feature extraction for Mandarin audio-visual speech recognition," in *2009 IEEE International Conference on Systems, Man and Cybernetics*, 2009, pp. 2936-2940.

[264] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers," *arXiv preprint arXiv:2004.04523,* 2020.
[265] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in neural information processing systems*, 1995, pp. 231-238.
[266] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4605-4608.
[267] K. Saeed and M. K. Nammous, "A speech-and-speaker identification system: Feature extraction, description, and classification of speech-signal image," *IEEE transactions on industrial electronics,* vol. 54, pp. 887-897, 2007.
[268] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5070-5074.
[269] M. S. Holi, "Automatic detection of neurological disordered voices using mel cepstral coefficients and neural networks," in *2013 IEEE Point-of-Care Healthcare Technologies (PHT)*, 2013, pp. 76-79.
[270] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly*, et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine,* vol. 29, pp. 82-97, 2012.
[271] T. Ghiselli-Crippa and A. El-Jaroudi, "Voiced-unvoiced-silence classification of speech using neural nets," in *IJCNN-91-Seattle International Joint Conference on Neural Networks*, 1991, pp. 851-856.
[272] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on HMM and ANN," in *2009 WRI World congress on computer science and information engineering*, 2009, pp. 225-229.
[273] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital signal processing,* vol. 22, pp. 1154-1160, 2012.
[274] E. Chandra and C. Sunitha, "A review on Speech and Speaker Authentication System using Voice Signal feature selection and extraction," in *2009 IEEE International Advance Computing Conference*, 2009, pp. 1341-1346.
[275] E. C. Guerra and D. F. Lovey, "A modern approach to dysarthria classification," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, 2003, pp. 2257-2260.
[276] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, pp. 115-133, 1943.
[277] D. O. Hebb, *The organization of behavior: a neuropsychological theory*: J. Wiley; Chapman & Hall, 1949.
[278] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological review,* vol. 65, p. 386, 1958.
[279] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature,* vol. 323, pp. 533-536, 1986.
[280] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation,* vol. 18, pp. 1527-1554, 2006.
[281] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153-160.
[282] M. A. Ranzato, C. Poultney, S. Chopra, and Y. L. Cun, "Efficient learning of sparse representations with an energy-based model," in *Advances in neural information processing systems*, 2007, pp. 1137-1144.

[283] G. E. Hinton and T. Shallice, "Lesioning an attractor network: Investigations of acquired dyslexia," *Psychological review,* vol. 98, p. 74, 1991.
[284] R. A. Davis, K.-S. Lii, and D. N. Politis, "Remarks on some nonparametric estimates of a density function," in *Selected Works of Murray Rosenblatt*, ed: Springer, 2011, pp. 95-100.
[285] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Stanford Univ Ca Stanford Electronics Labs1960.
[286] J. L. McClelland, D. E. Rumelhart, and P. R. Group, "Parallel distributed processing," *Explorations in the Microstructure of Cognition,* vol. 2, pp. 216-271, 1986.
[287] J. L. McClelland, D. E. Rumelhart, and G. E. Hinton, "The appeal of parallel distributed processing," *MIT Press, Cambridge MA,* pp. 3-44, 1986.
[288] D. S. Touretzky and G. E. Hinton, "Symbols among the neurons: Details of a connectionist inference architecture," in *IJCAI*, 1985, pp. 238-243.
[289] L. Yann, "Modeles connexionnistes de lapprentissage," *Ph. D. dissertation, PhD thesis,* vol. 6, 1987.
[290] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, pp. 2278-2324, 1998.
[291] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research,* vol. 3, pp. 1137-1155, 2003.
[292] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.
[293] A. Smola, "Advances in kernel methods: support vector learning," 1999.
[294] M. I. Jordan, *Learning in graphical models* vol. 89: Springer Science & Business Media, 1998.
[295] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," *Large-scale kernel machines,* vol. 34, pp. 1-41, 2007.
[296] O. Delalleau and Y. Bengio, "Shallow vs. deep sum-product networks," in *Advances in neural information processing systems*, 2011, pp. 666-674.
[297] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026,* 2013.
[298] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Icdar*, 2003.
[299] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," presented at the Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Lake Tahoe, Nevada, 2012.
[300] K. J. Lang, "A time delay neural network architecture for speech recognition," 1990.
[301] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing,* vol. 37, pp. 328-339, 1989.
[302] A. Waibel, T. Hanazawa, K. Shikano, G. Hinton, and K. Lang, "Speech recognition using time‐delay neural networks," *The Journal of the Acoustical Society of America,* vol. 83, pp. S45-S46, 1988.
[303] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing,* vol. 45, pp. 2673-2681, 1997.
[304] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, pp. 855-868, 2008.

[305] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850,* 2013.
[306] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764-1772.
[307] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," *Bioinformatics,* vol. 15, pp. 937-946, 1999.
[308] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, pp. 1735-1780, 1997.
[309] A. G. Felix, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural computation,* vol. 12, pp. 2451-2471, 2000.
[310] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104-3112.
[311] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539,* 2014.
[312] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156-3164.
[313] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov*, et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048-2057.
[314] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks,* vol. 5, pp. 157-166, 1994.
[315] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," ed: A field guide to dynamical recurrent neural networks. IEEE Press, 2001.
[316] A. Graves, "Supervised sequence labelling," in *Supervised sequence labelling with recurrent neural networks*, ed: Springer, 2012, pp. 5-13.
[317] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk*, et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078,* 2014.
[318] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.
[319] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480-1489.
[320] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805,* 2018.
[321] S. E. Fahlman, G. E. Hinton, and T. J. Sejnowski, "Massively parallel architectures for Al: NETL, Thistle, and Boltzmann machines," in *National Conference on Artificial Intelligence, AAAI*, 1983.
[322] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive science,* vol. 9, pp. 147-169, 1985.
[323] G. E. Hinton, T. J. Sejnowski, and D. H. Ackley, *Boltzmann machines: Constraint satisfaction networks that learn*: Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, 1984.

[324] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in Boltzmann machines," *Parallel distributed processing: Explorations in the microstructure of cognition,* vol. 1, p. 2, 1986.
[325] N. Le Roux and Y. Bengio, "Representational power of restricted Boltzmann machines and deep belief networks," *Neural computation,* vol. 20, pp. 1631-1649, 2008.
[326] G. Hinton, "How to do backpropagation in a brain," in *Invited talk at the NIPS'2007 deep learning workshop*, 2007.
[327] Y. Bengio and A. Fischer, "Early inference in energy-based models approximates back-propagation," *arXiv preprint arXiv:1510.02777,* 2015.
[328] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," Colorado Univ at Boulder Dept of Computer Science1986.
[329] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences,* vol. 11, pp. 428-434, 2007.
[330] Y. Le Cun and F. Fogelman-Soulié, "Modèles connexionnistes de l'apprentissage," *Intellectica,* vol. 2, pp. 114-143, 1987.
[331] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Advances in neural information processing systems*, 1994, pp. 3-10.
[332] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science,* vol. 313, pp. 504-507, 2006.
[333] R. Salakhutdinov and G. Hinton, "Semantic hashing," *RBM,* vol. 500, p. 500, 2007.
[334] R. Salakhutdinov and G. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning,* vol. 50, pp. 969-978, 2009.
[335] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.
[336] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in neural information processing systems*, 2009, pp. 1753-1760.
[337] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *ESANN*, 2011, p. 2.
[338] X. Li, "SPEech Feature Toolbox (SPEFT) design and emotional speech feature extraction," 2007.
[339] Y. C. Eldar, *Sampling theory: Beyond bandlimited systems*: Cambridge University Press, 2015.
[340] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*: Springer, 2007.
[341] N. Alcaraz Meseguer, "Speech analysis for automatic speech recognition," Institutt for elektronikk og telekommunikasjon, 2009.
[342] S. Jannetts and A. Lowit, "Cepstral analysis of hypokinetic and ataxic voices: correlations with perceptual and other acoustic measures," *J Voice,* vol. 28, pp. 673-80, Nov 2014.
[343] C. R. Watts, S. N. Awan, and Y. Maryn, "A Comparison of Cepstral Peak Prominence Measures From Two Acoustic Analysis Programs," *J Voice,* Oct 14 2016.
[344] C. Sauder, M. Bretl, and T. Eadie, "Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and Analysis of Dysphonia in Speech and Voice (ADSV)," *Journal of Voice,* vol. 31, pp. 557-566, 2017.
[345] Y. LeCun, "Generalization and network design strategies," *Connectionism in perspective,* vol. 19, pp. 143-155, 1989.

[346] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, ed: Springer, 1982, pp. 267-285.
[347] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747,* 2016.
[348] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980,* 2014.
[349] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov*, et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.
[350] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[351] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251-1258.
[352] J. Sietsma and R. J. F. Dow, "Creating artificial neural networks that generalize," *Neural Networks,* vol. 4, pp. 67-79, 1991/01/01/ 1991.
[353] C. M. Bishop, "Training with Noise is Equivalent to Tikhonov Regularization," *Neural Computation,* vol. 7, pp. 108-116, 1995.
[354] A. Sehgal, F. Saki, and N. Kehtarnavaz, "Real-time implementation of voice activity detector on ARM embedded processor of smartphones," in *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*, 2017, pp. 1285-1290.
[355] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, "Smartphone-based real-time classification of noise signals using subband features and random forest classifier," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2016, pp. 2204-2208.
[356] A. Bhattacharya, A. Sehgal, and N. Kehtarnavaz, "Low-latency smartphone app for real-time noise reduction of noisy speech signals," in *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*, 2017, pp. 1280-1284.
[357] S. SECTOR and O. ITU, "ITU-Ty. 4465."
[358] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters,* vol. 6, pp. 1-3, 1999.
[359] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transactions on Speech and Audio Processing,* vol. 11, pp. 498-505, 2003.
[360] J. W. Shin, J.-H. Chang, H. S. Yun, and N. S. Kim, "Voice activity detection based on generalized gamma distribution," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 2005, pp. I/781-I/784 Vol. 1.
[361] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *6th International Conference on Signal Processing, 2002.*, 2002, pp. 1124-1127.
[362] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing,* vol. 3, pp. 205-210, 2009.
[363] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7378-7382.
[364] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in

2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 2519-2523.

[365] R. Reed and R. J. MarksII, *Neural smithing: supervised learning in feedforward artificial neural networks*: Mit Press, 1999.

[366] J. Kam-Chuen, C. L. Giles, and B. G. Horne, "An analysis of noise in recurrent neural networks: convergence and generalization," *IEEE Transactions on Neural Networks,* vol. 7, pp. 1424-1438, 1996.

[367] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806,* 2014.