On the Use of Audio Fingerprinting Features for Speech Enhancement with Generative Adversarial Network

Farnood Faraji

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

November 2020

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

Abstract

Recently, the advent of learning-based methods in speech enhancement has revived the need for robust and reliable training features that can compactly represent speech signals while preserving their vital information. Time-frequency domain features, such as the Short-Term Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCC), are preferred in many approaches. They represent the speech signal in a more compact format and contain both temporal and frequency information. Compared to STFT, MFCC requires less memory and drastically reduces the learning time and complexity by removing the redundancies in the input. The MFCC are a powerful Audio FingerPrinting (AFP) technique among others which provides for a compact representation, yet they ignore the dynamics and distribution of energy in each mel-scale subband. In this work, a state-of-art speech enhancement system based on Generative Adversarial Network (GAN) is implemented and tested with a new combination of two types of AFP features obtained from the MFCC and Normalized Spectral Subband Centroid (NSSC). The NSSC capture the locations of speech formants and complement the MFCC in a crucial way. In experiments with diverse speakers and noise types, GAN-based speech enhancement with the proposed AFP feature combination achieves the best objective performance in terms of objective measures, i.e., PESQ, STOI and SDR, while reducing implementation complexity, memory requirements and training time.

Sommaire

Récemment, avec l'avènement de méthodes basées sur l'apprentissage dans l'amélioration de la parole, le besoin de de caractéristiques d'apprentissage robustes et fiables qui peuvent représenter de manière compacte les signaux vocaux tout en préservant leurs informations vitales a été ravivé. Les caractéristiques du domaine temps-fréquence, telles que la transformée de Fourier à court terme (STFT) et les coefficients cepstraux Mel-Frequency (MFCC), sont préférées dans de nombreuses approches. Ils représentent le signal vocal dans un format plus compact et contiennent à la fois des informations temporelles et fréquentielles. Par rapport à STFT, les MFCC nécessitent moins de mémoire et réduisent considérablement le temps d'apprentissage et la complexité en supprimant les redondances dans les données d'entrée. Les MFCC, qui font partie de la famille de caractéristiques de type Audio FingerPrinting (AFP), offrent une représentation compacte, mais ils ignorent la dynamique et la distribution de l'énergie dans chaque sous-bande à l'échelle mel. Dans cette thèse, un système d'amélioration de la parole à la pointe de la technologie basé sur le Generative Adversarial Network (GAN) est mis en œuvre et testé avec une nouvelle combinaison de deux types de caractéristiques AFP, soit les MFCC et les centroides normalisés de sous-bandes spectrales (NSSC). Les NSSC capturent les emplacements des formants de parole et complémentent ainsi MFCC d'une manière cruciale. Dans des expériences avec divers locuteurs et types de bruit, l'amélioration de la parole basée sur GAN avec la combinaison de caractéristiques AFP proposée atteint les meilleures performances en termes de mesures objectives, à savoir PESQ, STOI et SDR, tout en réduisant la complexité de la mise en œuvre, les besoins en mémoire et le temps d'apprentissage.

Acknowledgments

First and foremost, I wish to express my deepest appreciation to my supervisor, Prof. Benoit Champagne for the patient guidance, encouragement and advice he has provided throughout my time as his student. I would also like to thank Dr. Yazid Attabi (post-doctoral fellow) for his help and constructive comments over the course of my thesis work. I was honoured to receive the McGill Engineering Undergraduate Student Masters Award (MEUSMA) which helped me to undertake the graduate degree at McGill University. I am also grateful for the financial support provided by Prof. Champagne via his research grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and Microsemi Canada Ltd, without which the realization of this thesis would not have been possible. I must express my gratitude to my family for their continued support and encouragement throughout the years. I will be forever indebted to them, without whose unconditional support, love, encouragement, I would have never made it this far. Special appreciation goes out to my friends and fellows in the Telecommunications and Signal Processing (TSP) laboratory for their moral support and inspiring discussions.

Contents

List of Figures

List of Tables

List of Acronyms

AFP	audio fingerprinting
AFPC	audio fingerprinting combination
AI	artificial intelligence
AMS	amplitude modulation spectrum
ANN	artificial neural network
ART	adaptive resonance theory
ASR	automatic speech recognition
BNN	biological neural network
CDAE	convolutional denoising auto-encoder
CGAN	conditional generative adversarial network
CNN	convolutional neural network
CRNN	convolutional recurrent neural network
DCT	discrete cosine transform
DNN	deep neural network
DRNN	deep recurrent neural network
DWT	discrete wavelet transform
FFT	fast Fourier transform
GAN	generative adversarial network
HF	high frequency
IFFT	inverse fast Foutier transform
IRM	ideal ratio mask
LSGAN	least-square generative adversarial netwrok
LSTM	long short-term memory
MFCC	mel-frequency cepstral coefficients
MMSE	minimum mean-square error
MSE	mean square error
NMF	non-negative matrix factorization
NSSC	normalized spectral subband centroid

PCA	principal component analysis
PESQ	perceptual evaluation of speech quality
PLP	perceptual linear prediction
POLQA	perceptual objective listening quality analysis
PSD	power spectral density
ReLU	rectified linear unit
RL	reinforcement learning
RNN	recurrent neural network
SAR	signal-to-artifacts ratio
SBE	spectral band energy
SDR	signal-to-distortion ratio
SE	speech enhancement
SEGAN	speech enhancement generative adversarial network
SEP	spectral energy peak
SFM	spectral flatness measure
SIR	signal-to-interference ratio
SNR	signal-to-noise ratio
SOM	self-organizing map
SSC	spectral subband centroid
SSE	spectral subband energy
SSNR	segmental signal-to-noise ratio
STFT	short-time Fourier transform
STOI	short-time objective intelligibility
SVD	singular value decomposition

Chapter 1

Introduction

This chapter provides a general introduction to the thesis. It begins with a high-level overview of the single-channel speech enhancement problem under study is given. Then existing literature aimed at solving the denoising problem is surveyed. Next, the main technical contributions made by this thesis are summarized. Finally, the thesis organization is explained and key notations are defined for reference.

1.1 The Speech Enhancement Problem

Speech communication is an integral part of every human interaction. Over time, humans have developed a very distinct way of sound production compared to other animals. Humans learned how to produce distinctive sounds by combining different frequencies, formants and consonants produced from different vocal chord vibrations, air flow constrictions and mouth shapes. These complex and intricate mechanisms allowed us to form words, sentences and express concepts to transfer knowledge through our children and advance civilization. Human development throughout centuries required us to preserve and transfer

knowledge through different means other than mere speech, especially hand writing and printing. Up until the last century, most of this knowledge transfer was carried out using conventional oral and written forms.

In the 20th century, humans found ways to communicate with their voice over long distances through radio transmissions as well as to record their voices by means of electronic devices. In the last few decades up until now, computer networks emerged as a mainstream technology for transferring knowledge throughout the world. This has led to the escalation of human-human voice communications and human-computer voice enabled interactions. Nowadays, either type of interactions is extensively dependent on the voice acquisition, recording and transmission technologies, which are available through our devices such as personal computers, and smartphones. But there is a problem with this method of communication which ultimately relies on the use of microphones, since the latter are not as sophisticated and evolved as human ears in discerning desired speech from the acoustic background. This has caused problems for speech-based human-human communications as well as the more recent human-computer interactions.

From the early days of electronic voice communications, additive noise has been a recurrent problem for a variety of speech processing devices and applications. Different types of noise corrupt the speech in hearing aids, mobile devices, airplane communications and Automatic Speech Recognition (ASR) systems. These noise types present large variability in terms of their temporal, spectral and other fundamental characteristics, e.g., stationary versus non-stationary, spectrally white versus colored, etc. Speech enhancement aims to isolate a desired speech signal from the additive background noise, and increase the quality or intelligibility of the processed speech for storage, transmission or reproduction [1]. Hence, it is very important for the speech enhancement to work with different noise types, noise levels and speakers. In general, speech enhancement has two separate goals: (1)

improve quality/intelligibility signals to be consumed by humans and (2) improve system performance for signals to be consumed by machines.

1.2 Literature Review

In this section, we provide an overview of the different digital processing methods used for single-channel speech enhancement, which can be broadly classified as statistical and machine learning based methods. We also briefly discuss the main audio features used for the digital representation of speech signals in these methods.

1.2.1 Statistical Methods

The early works in the field of speech enhancement were based on conventional signal processing techniques. These approaches mainly rely on the time-frequency decomposition of the speech signal as obtained from the Short-Time Fourier Transform (STFT). Typically, they exploit *a priori* knowledge of the statistical distribution of the noise and signal power during each time-frequency bin to enhance the degraded speech signal [2]. These approaches make use of various probability models and distributions as well as statistical filtering techniques.

Below, we provide a brief overview of the main approaches or algorithms within this category of methods along with representative references; for more details, the reader is referred to [2, 3].

Spectral subtractive algorithms: The basic principles of these methods, which rely on the assumption of additive uncorrelated background noise, is to subtract an estimation of the noise power spectrum from the instantaneous noisy speech power spectrum, in order to recover the power spectrum of the clean speech. Commonly, in the statistical methods,

the phase of the noisy speech is not processed and used directly to synthesize the enhanced speech. However, some methods address the importance of phase in speech enhancement [4]. Spectral subtractive algorithms were initially proposed by Boll *et al.* [5], while further relevant extensions can be found in [6].

Spectral subtraction algorithms, however, suffer from perceptually annoying spectral artifacts. An improvement over these algorithms is based on modulation domain processing [7]. These methods use the information in the speech modulation spectrum which represents how the vocal tract changes as a function of time. More recent works in this field can be found in [8–10].

Minimum Mean Square Error (MMSE) algorithms: These approaches seek to achieve a better estimation of the clean speech STFT magnitude (also known as spectral magnitude) by exploiting available *a priori* knowledge of signal and noise distribution within established statistical estimation framework, including MMSE estimation and closely related maximum likelihood and maximum *a posteriori* estimation. This type of approach was initially proposed in [11, 12], while some more recent works along this avenue include [13, 14].

Kalman filter-based algorithms: In these methods, the linear predictive model for speech generation is recast as a state space model, allowing the use of a discrete-time Kalman filter to estimate the clean speech from the noisy speech signal. These methods, originally proposed in [15], operate directly on the time-domain signal samples. However, in past several works, various extensions to subband processing have been investigated [16, 17]

Subspace algorithms: Unlike the previous algorithms, the subspace algorithms use concepts of linear algebra to decompose the vector of noisy speech samples into orthogonal signal and noise components. These methods employ well-known orthogonal matrix decomposition techniques such as the Singular Value Decomposition (SVD) [18] or the eigenvalue decomposition [19]. Most recently in [20], the authors developed a subspace algorithm

based on human hearing model.

Over the years, several objective metrics have been developed to characterize the quality and intelligibility of the enhanced speech [3]. This includes Segmental Signal-to-Noise Ratio (SSNR), Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), short-time objective intelligibility measure (STOI) [21], Perceptual Evaluation of Speech Quality (PESQ) [22] and Perceptual Objective Listening Quality Analysis (POLQA) [23]. Statistical methods in general tend to improve the quality of the speech with regards to these metrics without training or *a priori* knowledge. Thus, compared to the learning-based algorithms discussed below, these methods require less computational resources and are less time-consuming. However, one of their main disadvantages is the creation of so-called *musical noise* in the processed speech. Musical noise is generated partly by non-linearities in the spectral processing stage of these classical algorithms, which results isolated peaks in the time-frequency representation of the enhanced speech signal [3]. All the above methods require the estimation of the noise power spectrum as well as other parameters that may be needed to characterize particular statistical distributions. In this regard, a well-known method for the estimation of the noise power spectrum is presented in [24]. In general, the above statistical methods can adapt to the noise level with quasi-stationary noises, but exhibit limited performance when used for impulse non-speech noise types [25].

1.2.2 Machine Learning Methods

In the past decade, due to important theoretical advances, faster and cheaper computational resources, and the availability of large recorded data set for training, neural networks have been applied successfully to a variety of non-linear mapping problems, including speech enhancement. Multiple research studies has been conducted with different neural network architectures and training features. Here, we provide a brief summary of these approaches

from the perspective of speech processing, and especially enhancement.

Non-negative Matrix Factorization (NMF): It is a popular dictionary-based approach which has been successfully applied to speech enhancement [26], speech separation [27], and speech recognition [28]. The NMF can be categorized as a learning-based approach since it requires training based on the observed data to build dictionaries. In this approach, a given non-negative matrix of signal descriptors is decomposed into the product of a non-negative basis matrix (also known as dictionary) and activation matrix. NMF is a dimensionality reduction tool which, as apposed to principal components analysis (PCA) and vector quantization (VQ), only allows additive (and not subtractive) combinations of the basis vectors [29]. In speech enhancement, the non-negative input matrix is usually the Short-Time power or magnitude spectrum of the speech signal. Some recent works in this area include [30, 31].

Deep Neural Network (DNN): These types of networks consist of fully connected multi-layer perceptrons designed to learn non-linear patterns in data. One of the early works on the application of DNN to speech enhancement is by Narayanan and Wang [32], which uses Relative Spectral filtered Perceptual Linear Prediction cepstral coefficients (RASTA-PLP), Mel Frequency Cepstral Coefficients (MFCC) and Amplitude Modulation Spectrum (AMS) as input features. Xu *et al.*, [33] propose a supervised speech enhancement system based on DNN that can outperform the conventional statistical methods. Finally, [34, 35] uses a perceptually modified loss function to train a DNN model using logarithmic magnitudes of the STFT features.

Recurrent Neural Network (RNN): RNNs are a class of neural networks which exhibit temporal dynamic behaviour. Hence, due to the sequential nature of speech, RNN provides a powerful tool in speech enhancement. Some notable works along this avenue include [36–38] which report an improvement in objective measures over statistical and DNN methods.

Convolutional Neural Network (CNN): In recent years, Covolutional Neural Networks (CNN) have achieved notable performance improvements in the context of speech recognition and image processing [39], while requiring much smaller number of model parameters than DNN and RNN. Because of their ability to learn and extract robust structures of inherent within clean speech and noise signals in the time-frequency domain CNN has found successful application in speech processing. In [40], the authors improve the CNN architecture by using it as an auto-encoder. Their approach, called Convolutional Denoising Auto-Encoder (CDAE), uses a fully connected CNN which takes two-dimensional (2D) time-frequency domain inputs and encodes and decodes them into a corresponding 2D output. This technique noticeably reduces the number of parameters present in the architecture. In these works, CNN delivers a considerable improvement in terms of objective measures such as SDR and SIR, while significantly reducing the number of trained and stored model parameters, when compared to DNN and RNN [41].

Generative Adversarial Network (GAN): The GAN aims to generate more realistic output patterns that exhibit characteristics closer to the real data [42]. Adversarial training can also be employed in the field of speech enhancement. Proposed by [43, 44], Speech Enhancement GAN (SEGAN) operates in the time-domain and employs a one dimensional Convolutional Neural Network (CNN). A similar neural network architecture is investigated in [45] but using STFT features. In [46, 47], the authors use Gammatone and STFT features, respectively, along with a GAN architecture for speech enhancement, and propose modified network training targets.

The above neural network based methods require substantial training data to give the best performance. Thus, having a reliable feature set which reduces memory requirements and training time is an important asset, especially for embedded systems and real-time applications. Speech enhancement based on neural networks can work with both time [43,

48] and frequency domain data [45–47]. However, it appears that frequency-domain features have a clear advantage over the former, especially in terms of speech quality measures like PESQ [48].

1.2.3 Speech Features

Multiple possibilities exist regarding the choice of audio features to be used as input in neural network based speech processing systems. These features can be classified into two main categories, i.e., time-domain versus transform domain features. Time-domain features are directly obtained from the discrete-time samples of the audio signals under consideration, with a minimum amount or processing (or none). In contrast, transform-domain features are obtained by applying a linear transformation on the audio signal samples, possibly followed by further processing. Typical transformations include the STFT, which allows a representation of the audio signal as a temporal sequence of complex spectral values [3], and the Discrete Wavelet Transform (DWT), whose representation coefficients allow a balance between temporal and frequency resolution [49–51].

Frequency-domain features, such as the unprocessed STFT, the Gammatone spectrum and the Mel-Frequency Cepstral Coefficients (MFCC) have been used frequently in the literature. In addition, a combination of STFT with MFCC is employed in [52] for training wide residual networks for speech enhancement. Compared to STFT, filter-based features derived from the latter, such as MFCC, exhibit reduced dimensionality and are more suitable for learning algorithms, as they can reduce memory and computational requirements while maintaining comparable level of performance [47, 53–55]. MFCC belong to a larger family of so-called Audio Fingerprinting (AFP)[1] features, which include the Spectral Sub-

[1]We understand the AFP terminology is more commonly used in the music information retrieval literature. However, in this thesis, we apply the same concept to speech and prefer to use the same terminology for the sake of simplicity.

band Centroids (SSC) and Spectral Energy Peaks (SEP). These features have been used effectively in the implementation of various audio processing tasks, including data compression and pattern extraction [56].

The MFCC are computed by applying the Discrete Cosine Transform (DCT) to a set of weighted subband energies obtained from the application of a Mel-spaced filterbank to the STFT magnitude coefficients. However, the filter-based energy computation of this process ignores important information about the audio signal in each subband, such as the locations of energy peaks corresponding to speech formants. The SSC introduced by Paliwal [57], provides crucial information about the centroid frequency in each subband, which has proven to be of great value in several applications. The SSC have been successfully employed in speech recognition, speaker identification and music classification, with non-learning or dictionary-based systems [58–60]. Besides a combination of MFCC and SSC was proposed for speaker authentication with non-learning methods in [61].

1.3 Thesis Objectives and Contributions

To the best of our knowledge, in the field of speech enhancement using neural networks and machine learning, there has been minimal effort to incorporate *multiple* AFP features in the training and processing phases of the network operation. In this thesis, we propose to use and investigate the performance of a combination of AFP features in speech enhancement applications of neural networks. To achieve this, we choose a state-of-the-art network architecture model, namely GAN, which we adapt for training and testing with different combinations of feature sets including two prominent AFP ones, i.e. MFCC and SSC.

Indeed, while the MFCC lead to a significant reduction of the processing complexity, they do not perform as well as the STFT. We believe that by adding the SSC to the MFCC,

we can palliate to their intrinsic limitation (i.e. lack of frequency resolution) without significantly increasing the complexity, when compared to a neural network system based on STFT features.

The main contributions can be a summarized as follows:

- We implement a state-of-art speech enhancement system based on GAN to predict the Ideal Ratio Mask (IRM) of the noisy speech.
- We propose using a compact set of features obtained from the combination of MFCC, Normalized SSC (NSSC) and their respective time differences (i.e. delta versions) for training the GAN.
- We evaluate the performance of the resulting system by means of standard objective measures, and compared the results to that of other possible combinations of features, including the STFT coefficients.
- Our results show that the proposed combination of AFP features based on MFCC and NSSC can achieve best (or near best) performance under a wide range of SNR and noise type, while significantly reducing memory requirements and training time.

The above contributions have led to a publication in a peer-reviewed conference:

- F. Faraji, Y. Attabi, B. Champagne, and W.-P. Zhu, "On the use of audio fingerprinting features for speech enhancement with generative adversarial network," in *Proc. IEEE Int. Workshop on Signal Processing Systems (SiPS)*, Coimbra, Portugal, pp. 77-82, Oct. 2020.

Regarding the contributions of the authors to the paper above, the first author, Mr. Farnood Faraji, developed the ideas, implemented the algorithms, conducted the experiments and wrote the first draft of the manuscript. The co-authors, Dr. Y. Attabi, Prof. B.

Champagne and Prof. W.-P. Zhu provided guidance and advice throughout the research by suggesting further ideas, validating the theoretical developments, suggesting refinements to the experimental methodology, and contributing to the writing and editing of the final manuscript.

1.4 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 outlines the motivation behind deep learning and provides a brief overview on how artificial neural networks have inspired researchers to derive successful algorithms to solve a wide variety of problems. In particular, the chapter further discusses the fundamentals of neural network, including GAN, along with their training by providing mathematical descriptions. In Chapter 3, different audio features are presented and briefly discussed using mathematical representations. On the basis of the feature equations provided in Chapters 2 and 3, a new feature combination of AFP features, i.e. consisting of MMCF, NSSC, and their deltas, is proposed and its incorporation into GAN is explained. The experimental evaluation and performance results for the proposed method for various noise types are presented in Chapter 5. Finally, we summarize the findings of our work in Chapter 6, where we also briefly discuss potential research directions for future research.

Throughout the thesis, vectors are denoted as bold or uppercase letters while scalars are shown as lowercase. $\mathbb{Z}$ and and $\mathbb{R}$ denote the set of signed integer and real numbers, respectively. $\|.\|$ represents the Euclidean norm of its vector argument. Finally, $\mathbb{E}$ denotes the expected value of a random quantity.

Chapter 2

Machine Learning in Speech Enhancement

In this chapter, we will review some underlying concepts of Machine Learning (ML) and their application to speech enhancement. First, an overview of neural networks and deep learning is presented. Next, fundamental concepts of artificial neural network architecture and training are reviewed. Finally, the principles of Generative Adversarial Networks (GANs) are briefly discussed along with their application to speech enhancement.

2.1 Deep Learning

With the recent innovations and advancements in computer and digital processing technologies, we are able to run more computations in less time and with relatively physically smaller devices. This technological leap has spurred the application of machine learning methods which, although they had been around for decades, could not previously be implemented in real world scenarios. Machine learning is concerned with the development of data

processing algorithms, typically in the form of computer programs, which can learn fundamental laws that governing various learning processes through experience [62]. Nowadays, Artificial Intelligence (AI) and machine learning have infiltrated several aspects of our lives, as evidenced when we shop online, when we search on our frequently used search engines or when we seek to find new friends through social media. AI is now deeply implanted in our lives and it seems that this trend will continue to expand.

Deep learning is an important branch of machine learning methods based on Artificial Neural Networks (ANN). This branch includes a variety of methods such as DNN, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), etc. which aim to learn complex and non-linear patterns and relationships in large amounts of data. These methods are very flexible and come in various forms, namely: unsupervised, supervised and reinforcement learning, as further explained below.

The goal of unsupervised methods is to find the underlying pattern in data as well as a corresponding mapping function, without any target values. This form of learning only employs the unlabelled data. In neural networks the Adaptive Self-Organizing Map (SOM) [63] and Resonance Theory (ART) [64] are commonly used for unsupervised learning.

The most commonly used machine learning methods belong to the category of supervised learning which is the approach used in this thesis. In this type of method, the input values have corresponding output targets and the underlying neural network model is trained in a way to learn this mapping. A well-defined and trained model is capable of extending the learned patterns to unseen examples. Supervised learning is widely used in the field of speech enhancement, while unsupervised learning has been studied as well [65].

The third machine learning method is Reinforcement Learning (RL), in which the software agent is trained in a way to maximize its cumulative reward in a given environment. The difference between RL and supervised learning is the lack of labelled target outputs.

In RL, unlike supervised learning, sub-optimal actions of the model are not explicitly corrected. Thus, the model is trained in a way to find a balance between exploration and exploitation that is, exploring new knowledge territories while retaining the attained ones. This type of problem is studied in many disciplines, such as information theory, genetic algorithm, game theory, etc. and the realm of applications is still growing [66].

2.2 Artificial Neural Networks

2.2.1 Fundamental Concepts

The basic ideas behind ANN and machine learning first originated from analogies with Biological Neural Networks (BNN). The human brain or BNN consists of approximately 10^{11} interconnected brain cells, also called neurons. The structure of a typical biological neuron is illustrated in Fig. 2.1. A neuron consists of three main parts, namely: dendrites, cell body and axon. Electrical or chemical signals are captured by the dendrites, processed by the cell body, and then carried away through the axon to connect to thousands of other neurons via their dendrites, forming a BNN. The juncture between the axon terminal and the dendrites is referred to as a synapse which serves like a gate, regulating the flow of information within the brain.

This conceptually simple interconnected web of neurons comprising billions of interconnected neurons, can learn and execute very complex processes. The ANN aims to imitate the concept in BNN ability to solve complex problems and use it in engineering and scientific applications. Fig. 2.2 illustrates the model of an artificial neuron, which consists of several components, namely: inputs, weights, bias, accumulator, activation function and output. The input feature vector is defined as $\boldsymbol{x} = [x_1, x_2, \ldots, x_I]$, where I is the number of inputs. TWhile the features can be in any form, in the case speech signal processing, they

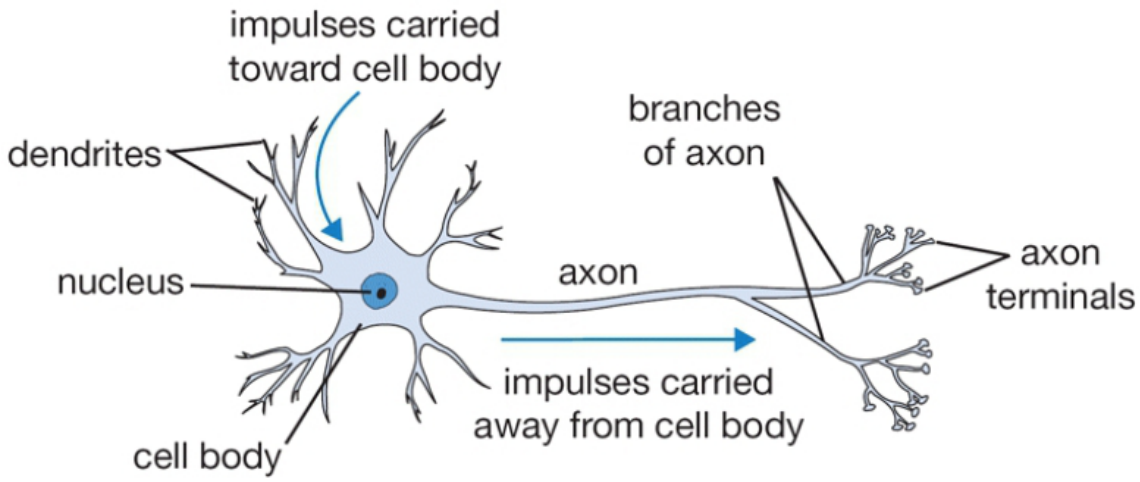

Fig. 2.1 Structure of a biological neuron

often consist of time-frequency features such as the STFT, MFCC, etc. The weight vector is defined as $\boldsymbol{W} = [w_1, w_2, \ldots, w_I]$, where the individual weight w_i is associated with input x_i. The operation of an artificial neuron, which aim to imitate the behavior of a biological cell, and especially the synaptic connection, can be mathematically expressed as,

$$z = \sum_{i=1}^{I} w_i x_i + w_0, \tag{2.1}$$

$$y = f(z), \tag{2.2}$$

where w_0 is a bias, z is the output after applying the weights and bias to the input and y is the output after applying an activation function $f(.)$ to z. Activation functions are used to add non-linearity to the network, thereby providing the capability to learn very complex and nontrivial tasks.

In BNNs, the activation function usually an abstraction representing the rate of action potential firing in the cell. The role of the activation functions is to map or compress the permissible amplitude range of the output signal to some other more appropriate range.

Therefore, depending on the application, the desired activation function can vary.

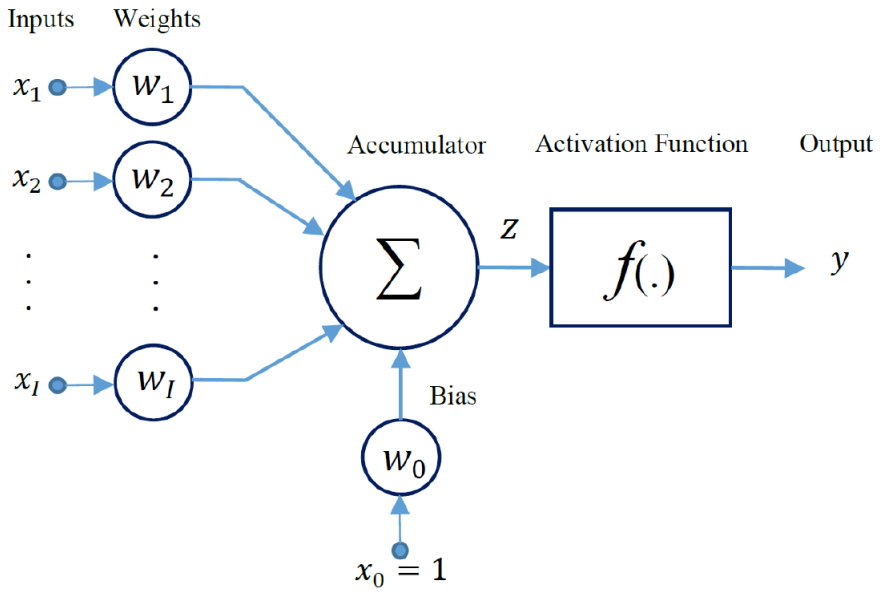

Fig. 2.2 Block diagram of an artificial neuron

Below, we provide a summary and mathematical description of the most used activation functions in the field of neural network, which are also shown plotted in Fig. 2.3 for illustration:

- The linear activation function is an identity function which is mostly used in linear problems or for regression purposes. The activation function is expressed as,

$$f_{linear}(x) = x. \tag{2.3}$$

- The Rectified Linear Unit (ReLU) only keeps the positive values of the input and outputs zero for any negative input. This function is commonly used because of its simplicity and effectiveness. In this thesis, we extensively use ReLU since our

inputs are non-negative quantities derived from power spectrum measurements and are greater than 0 in the input. This activation function is expressed as,

$$f_{ReLU}(x) = \max(0, x). \tag{2.4}$$

- The Sigmoid activation function is the most commonly used function in the literature because of its many desirable features, i.e.: continuous, non-linear, differentiable and outputs a value between 0 and 1. In our case, the sigmoid function is used in the last network layer to output a Wiener-type filtering value. It can be expressed as,

$$f_{sigmoid}(x) = \frac{1}{1 + e^{-x}}. \tag{2.5}$$

- The Hyperbolic Tangent (tanH) activation function is the ratio between the hyperbolic sine and cosine. Similar to the sigmoid function, it is continuous, non-linear and differentiable. Its outputs is always between -1 and 1 and is calculated as,

$$f_{tanh}(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{2.6}$$

2.2.2 Network Training

Connecting a group of artificial neurons in multiple layers gives us a deep neural network as depicted in Fig. 2.4. This goal of the network is to change its set of weight and bias parameters in a way to perform a certain mapping from inputs to the outputs. In the context of supervised learning, the corresponding inputs and outputs, represented by vectors $\boldsymbol{x} = [x_1, \ldots, x_I]$ and $\boldsymbol{y} = [y_1 \ldots, y_K]$, respectively, are given to the network. A cost

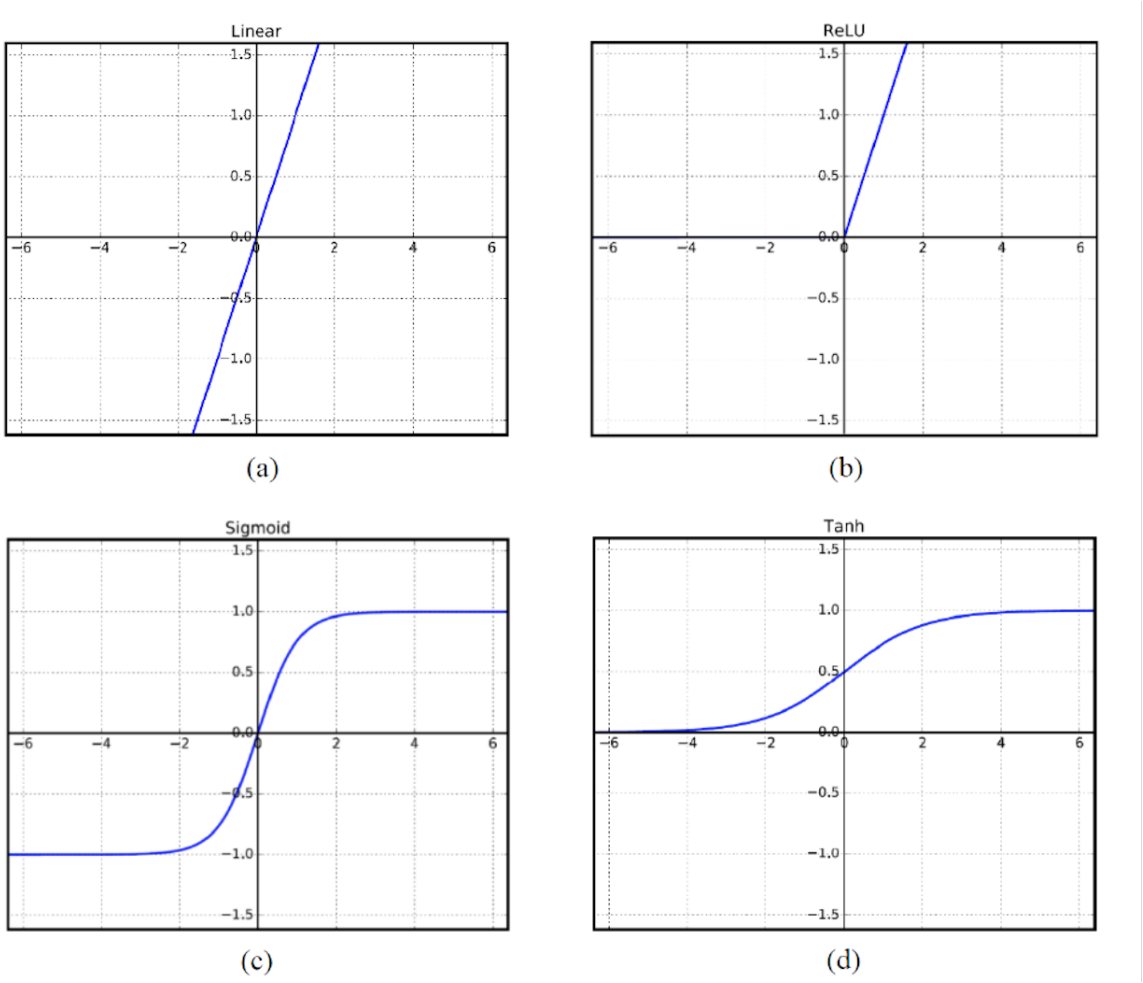

Fig. 2.3 Illustration of the most common activation functions: (a) linear, (b) rectified linear unit (ReLU), (c) sigmoid and (d) tanh

function is used to quantitatively measure the error (or loss) between the network's output vector $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_K]$ and the desired output. Two most common cost functions are the Mean Square Error (MSE) and the cross-entropy, which can be expressed as, respectively,

$$C_{\text{MSE}}(\boldsymbol{W}^{(l)}) = \sum_{i=1}^{K} |\hat{y}_i - y_i|^2, \tag{2.7}$$

$$C_{\mathrm{MSE}}(\boldsymbol{W}^{(l)}) = -\sum_{i=1}^{K} y_i \ln \hat{y}_i + (1 - y_i) \ln (1 - \hat{y}_i), \tag{2.8}$$

where y_i and $\hat{y}_i$ are the i-th target and estimated output, respectively. These quantities are functions of the network parameters, represented by $\boldsymbol{W} = [\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(l)}]$ where in turn, each matrix $\boldsymbol{W}^{(l)} = [w_{ij}^{(l)}]$, comprises the weights and bias parameters of the l-th layer, and L being the total number of layers. The network parameters are chosen in a way to minimize the cost function $C(\boldsymbol{W}^{(l)})$,

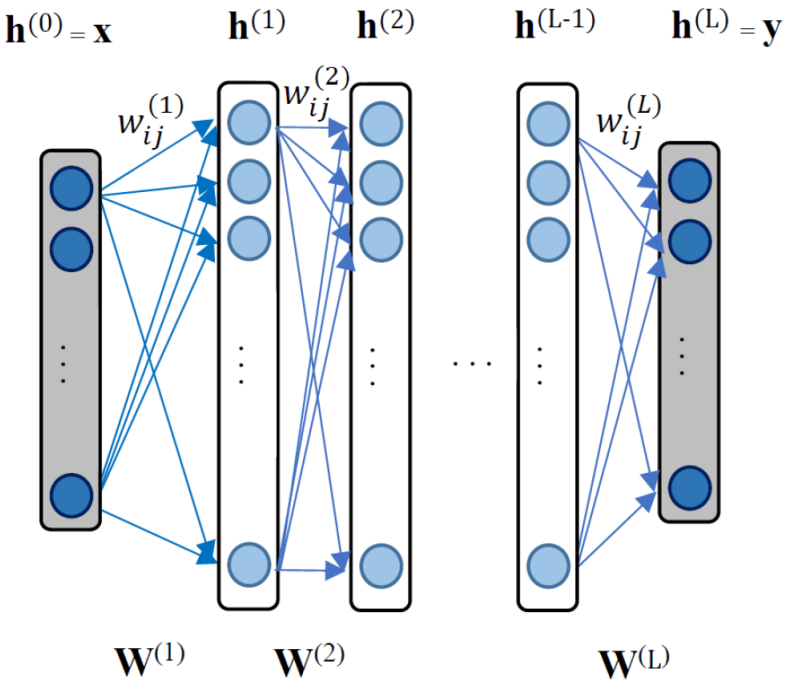

Fig. 2.4 Example of a feed-forward deep neural network

In neural networks, it is very important to find the optimal weights to achieve a better performance and lower the cost. Many algorithms exist for this purpose, which of which are iterative nature, i.e., wherein the weights are updated at each iteration with the goal of reducing the cost. Different algorithms use different weight update paradigms which for

the most part involve the weight gradient and possibly higher derivatives. In the literature, the gradient descent algorithm represents the general updating rule form as,

$$\boldsymbol{W}^{(l)}(\tau+1) = \boldsymbol{W}^{(l)}(\tau) - \mu\nabla\boldsymbol{W}^{(l)}(\tau), \tag{2.9}$$

where $\boldsymbol{W}^{(l)}(\tau)$ is the weight matrix of layer l at iteration τ, $\mu > 0$ is a step size controlling the learning rate, and $\nabla\boldsymbol{W}^{(l)}(\tau)$ is defined as the gradient of the cost function, as shown in,

$$\nabla\boldsymbol{W}^{(l)}(\tau) = \left.\frac{\partial C(\boldsymbol{W}^{(l)})}{\partial \boldsymbol{W}^{(\boldsymbol{l})}}\right|_{\boldsymbol{W}^{(l)}=\boldsymbol{W}^{(l)}(\tau)}. \tag{2.10}$$

The learning process consists of two stages, i.e.: feed-forward and feed-backward propagation. In the feed-forward stage, input data are supplied to the network and the output of each hidden layer is calculated until it reaches the last layer. In the last layer, the error is calculated and propagated back through the network in order to update the weights in a way to reduce the loss calculated from the cost function. Assuming the MSE cost function and gradient descent optimization, the rate of change of the error with respect to each weight in the hidden layer, $w_{ij}^{(l)}$, in the network is given by,

$$\delta_j^{(l)} = \frac{\partial C}{\partial z_j^{(l)}}, \tag{2.11}$$

$$\frac{\partial C}{\partial w_{ij}^{(l)}} = y_j^{(l-1)}\delta_j^{(l)} = y_j^{(l-1)} \sum_{k\in I_{l+1}} w_{kj}^{(l+1)} f'(z_j^{(l)})\delta_k^{(l+1)}, \tag{2.12}$$

where $\delta_j^{(l)}$ represents the error at each layer l and neuron j. Using this error change rate, we update each network weight using (2.9), to minimize the total network loss.

2.3 Deep Learning and Speech Enhancement

In this section, we first present an overview of how deep learning and neural networks can be applied to speech enhancement problems in general. Next, we summarize the basic concepts and operation of GAN and explain how it can be adapted for use in the speech enhancement application.

2.3.1 Overview

The majority of single-channel speech enhancement methods use the Analysis-Modification-Synthesis (AMS) as the underlying processing framework. In this approach, the noisy speech, which results from an additive combination of acoustic noise to the desired speech, is first analyzed into fundamental components (e.g. via STFT analysis), the components are then processed to remove noise artifacts, and finally, the processed components are used to synthesize the enhanced speech at the system's output. This approach is typically employed for both classical (i.e., statistical) and learning-based methods. In the latter case, the modification stage amounts to extracting a set of desired features and processing them with a trained neural network. Thus, unlike more conventional deep learning-based classification methods which predict a label or category at their output, deep learning in speech enhancement applications aims to predict a clean speech signal from a sequence of input feature vectors by enhancing with the network a sequence of input feature extracted from the noisy speech signal.

As previously outlined in Chapter 1, the extracted features (used as input to the network) could be in time-domain, as in e.g. SEGAN [43] or in frequency-domain [46], which is a special case of AMS. The only advantage of time-domain features over the AMS framework is the lack feature-extraction step, which saves the resources allocated to the signal analysis

and synthesis (i.e., STFT and inverse STFT). However, the frequency-domain analysis of audio signals naturally mimics the processing taking place in the human auditory system and lends itself to several lower-dimensional representations such as the MFCC. In turn, these representations make it possible to significantly reduce the processing and training complexity of the network without critically affecting performance. Such representations, which are furthered discussed in Chapter 3, play a key role in this thesis.

In the modification stage, depending on the considered method, the deep learning speech enhancement model is used to output either the magnitude spectrum of the enhanced speech or a Wiener-type filter to be used for enhancing the noisy speech as in the statistical-based methods. In neural network-based models, this procedure is performed as discussed in Section 2.2 by applying the non-linear mapping from the trained network to the to a temporal sequence of input feature vectors extracted from the noisy speech. Finally, in the synthesis stage, the modified and enhanced speech representation in the frequency domain is employed to reconstruct the speech signal. This stage involves the application of the inverse STFT to the frequency domain data along with the overlap-add method in the time-domain, since the data have initially been processed in separated window frames.

In this thesis, we study a learning-based speech enhancement model developed based on the GAN framework and evaluate it with different feature sets derived from frequency-domain, i.e., STFT analysis of the noisy input speech, with emphasis on so-called audio fingerprinting features. In particular, the enhanced features predicted by the trained neural network model will be used as Wiener filter, which in turn will be applied to the STFT magnitudes of the noisy speech in order to remove unwanted acoustic noise.

2.3.2 Generative Adversarial Network

The Generative Adversarial Network (GAN) is a class of machine learning methods introduced by Goodfellow *et al.* [42] in 2014. The GAN aims to solve certain problems which arise due to the difficulty of approximating intractable probabilistic computations in deep *generative* models, by proposing an adversarial setting. GANs have proven to be useful for supervised [67] and reinforcement learning [66] in applications such as image enhancement and synthesis [68].

Specifically, GANs are generative models designed to map noisy sample vectors, say $\boldsymbol{z}$, from a prior distribution into outputs that resemble those generated from the real (i.e., actual) data distribution. To achieve this, a generator (G) learns to effectively imitate the real data distribution under adversarial conditions. The adversary in this case is the discriminator (D) which is a binary classifier whose inputs are either samples from the real distribution, or *fake* samples made up by G. The training process is a game between G and D: G is trying to fool D to accept its outputs as *real*, and D gets better in detecting fake inputs from G and distinguishing them from real data. As a result, G adjusts its parameters to move towards the real data manifold described by the training data [42].

The adversarial training described above can be formulated as the following minmax problem,

$$\min_G \max_D V(D,G) = \mathbb{E}[\log D(\boldsymbol{x})] + \mathbb{E}[\log(1 - D(G(\boldsymbol{z})))], \tag{2.13}$$

where $V(D,G)$ is the value function of the system, referred to as sigmoid cross entropy loss function, $\boldsymbol{x}$ is the feature vector from the real data distribution, $\boldsymbol{z}$ is the latent vector generated from a noisy distribution, $D(\boldsymbol{x})$ and $G(\boldsymbol{x})$ are the outputs of D and G, and $\mathbb{E}$ denotes expected value.

In speech enhancement applications, it has been observed that Conditional GAN

(CGAN) [67, 69] results in better performance than conventional GAN [43, 46, 47]. CGAN uses an additional data vector $\boldsymbol{x}_c$ in both G and D for regression purposes, while the value function from (2.13) is changed to,

$$\min_G \max_D V_{\mathrm{C}}(D, G) = \mathbb{E}[\log D(\boldsymbol{x}, \boldsymbol{x}_c)] + \mathbb{E}[\log(1 - D(G(\boldsymbol{z}, \boldsymbol{x}_c), \boldsymbol{x}_c))]. \tag{2.14}$$

The training steps of CGAN are depicted in Fig. 2.5. The training consists of three consecutive steps: First, D is trained with a concatenation of the vector $\boldsymbol{x}$ and the additional conditional feature vector $\boldsymbol{x}_c$, in such a way that it recognizes $\boldsymbol{x}$ as real (or output 1). Next, D learns to categorize the concatenation of $\hat{\boldsymbol{x}} = G(\boldsymbol{z}, \boldsymbol{x})$ and $\boldsymbol{x}_c$ as fake data distribution (or output 0). Finally, D's variables are frozen and G is trained with the $\boldsymbol{x}_c$ features to fool the D.

The GAN methods based on (2.13) and (2.14) use the sigmoid cross entropy loss function which causes vanishing gradients problem for some fake samples far from the real data, which in turn leads to saturation of the loss function. Alternatively, CGAN can be combined with the Least-Squares GAN (LSGAN) [70], which solves this problem by stabilizing GAN training and increasing G's output quality. This is achieved by substituting the cross-entropy loss with a binary-coded least-squares function, and training G and D individually. The objective function of the resulting modified GAN expressed by,

$$\min_D V(D) = \mathbb{E}[(D(\boldsymbol{x}, \boldsymbol{x}_c) - 1)^2] + \mathbb{E}[(D(G(\boldsymbol{z}, \boldsymbol{x}_c), \boldsymbol{x}_c))^2], \tag{2.15}$$

$$\min_G V(G) = \mathbb{E}[(D(G(\boldsymbol{z}, \boldsymbol{x}_c), \boldsymbol{x}_c) - 1)^2]. \tag{2.16}$$

The use of the objective function in (2.15) and (2.16) which is a combination of CGAN and LSGAN, alleviates the saturation and convergence problems occurring in the conven-

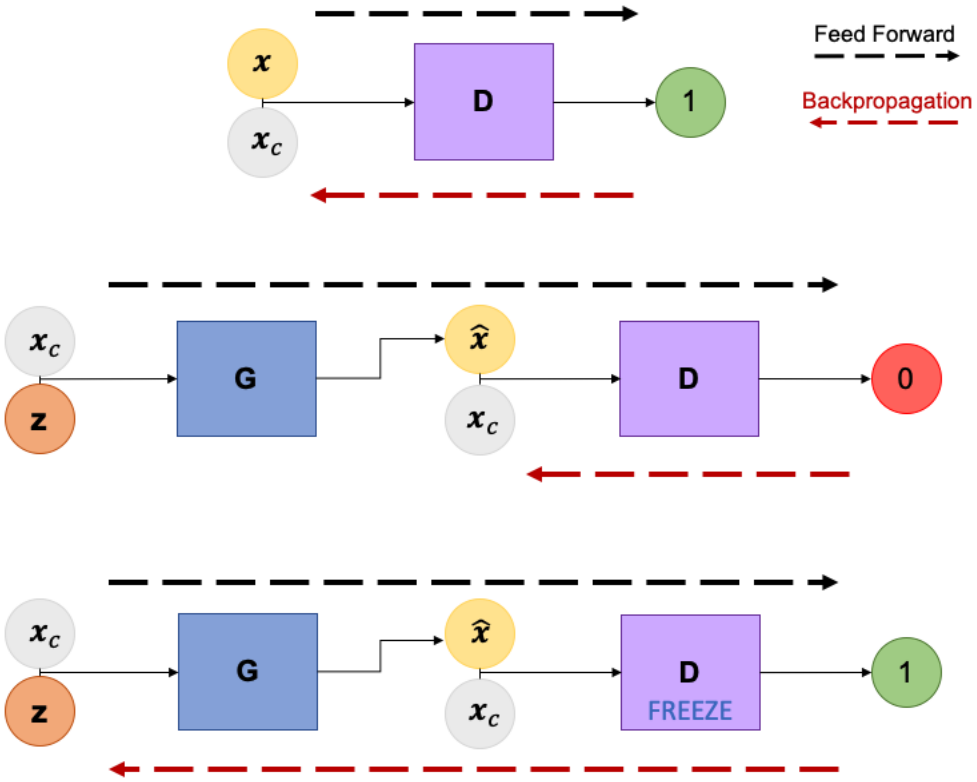

Fig. 2.5 Training steps of Conditional GAN.

tional GAN model. This final architecture is implemented in this thesis to study and compare the performance of different feature sets, including the proposed combination of audio fingerprinting features, in learning-based speech enhancement.

Chapter 3

Speech Features

In this chapter, we first present the discrete-time speech signal model and its Short-Time Fourier Transform (STFT) representation. Subsequently, different feature extraction methods based on the STFT are presented and briefly explained. This includes the Mel-Frequency Cepstral Coefficients (MFCC), Spectral Subband Centroids (SSC), Spectral Energy Peaks (SEP), Spectral Band Energies (SBE), and Spectral Flatness Measures (SFM).

3.1 Speech Model

Let $y[m]$ denote the observed noisy speech signal, where $m \in \mathbb{Z}$ is the discrete-time index. The noisy speech results from the contamination of a desired, clean speech signal $s[m]$ with an additive noise signal $n[m]$, i.e.,

$$y[m] = s[m] + n[m], \quad m \in \mathbb{Z}, \tag{3.1}$$

where no particular assumptions are made on the noise type. We represent the signals of interest in the time-frequency domain, as obtained from application of the STFT to (3.1).

Specifically, the STFT coefficients of the noisy speech signal $y[m]$ are defined as,

$$\text{STFT}\{y[m]\} \equiv Y(k,f) = \sum_{m=0}^{M-1} y[m+kL]h[m]e^{-j2\pi fm/M}, \tag{3.2}$$

where $k \in \mathbb{Z}$ is the frame index, L is the frame advance, $f \in \{0,1,2,...,M/2\}$ is the frequency bin index, M is the frame size and $h[m]$ is a non-negative window function. In practice, the calculation in (3.2) is implemented by means of an M-point Fast Fourier Transform (FFT) algorithm. Applying the STFT formula from (3.2) on the time-domain model (3.1) yields the time-frequency model representation as,

$$Y(k,f) = S(k,f) + N(k,f), \tag{3.3}$$

where $S(k,f)$ and $N(k,f)$ are the STFT of the clean speech and noise signals, respectively.

3.2 Audio Fingerprinting Features

To train the GAN architecture, we will propose and study in the next chapter, a new feature set obtained by combination of MFCC and NSSC. In this part, we explain the calculation and combination of these and other AFP features.

3.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC are widely used in speech recognition and enhancement due to their remarkable capabilities to compress speech while preserving its essential information [53, 54, 71]. As a first step in the calculation of the MFCC features, the time-domain signal $y[m]$ is passed through a first-order FIR filter to boost the highband formants in a so-called pre-emphasis

stage, as given by,

$$y'[m] = y[m] - \alpha y[m-1], \tag{3.4}$$

where α is the pre-emphasis coefficient, typically in the range $0.95 \leq \alpha \leq 1$.

Next, the STFT of the filtered signal $y'[m]$ is calculated as in (3.2), yielding the STFT coefficients $Y'(k, f)$. For each data frame, these STFT coefficients are used to calculate a set of Spectral Subband Energies (SSE) defined in terms of a bank of overlapping narrow-band filters. Specifically, the SSE of the k-th frame are calculated as,

$$\mathrm{SSE}_y(k, b) = \sum_{f=l_b}^{h_b} w_b(f)|Y'(k, f)|^2, \tag{3.5}$$

where $b \in \{0, 1, \ldots, B-1\}$ is the subband index, B is the number of subbands in the filterbank, and $w_b(f) \geq 0$ is the spectral shaping filter of the b-th subband, with l_b and h_b denoting the lower and upper frequency limits of $w_b(f)$. More specifically, the filters $w_b(f)$ together form a mel-spaced filterbank, i.e., they are characterized by triangular shapes with peak frequencies distributed according to the mel-scale of frequency [72].

Finally, the Discrete Cosine Transform (DCT) - Type III [73] is applied to the logarithm of the SSE to obtain the desired MFCC features, which can be expressed as,

$$\mathrm{MFCC}_y(k, p) = \sqrt{\frac{2}{B}} \sum_{b=0}^{B-1} \log_{10}(\mathrm{SSE}_y(k, b)) \cos\left(\frac{p\pi}{B}(b - 0.5)\right), \tag{3.6}$$

where $p \in \{0, 1, \ldots, P-1\}$ is the DCT index, and P is the number of coefficients. We define the MFCC feature vector of the current data frame as,

$$\mathbf{MFCC}_y = [\mathrm{MFCC}_y(k, 0), \ldots, \mathrm{MFCC}_y(k, P-1)]. \tag{3.7}$$

3.2.2 Spectral Subband Centroids (SSC)

The SSC were introduced in [57] to measure the center of mass of a subband spectrum in terms of frequency, using a weighted average technique. These features exhibit robustness against the equalization, data compression and additive noise which do not significantly alter the peak frequencies at moderate to high Signal-to-Noise Ratio (SNR) [56]. In [74], the SSC outperform MFCC when used as inputs in a audio recognition task based on dictionary matching. To generate SSC values, the noisy speech signal $y[m]$ is pre-emphasized as in (3.4) and the corresponding STFT coefficients $Y'(k, f)$ are computed. For each frame, a set of SSC is obtained by calculating the centroid frequencies of a bank of narrowband filters as in the MFCC. Specifically, the SSC of the k-th frame are calculated as,

$$\mathrm{SSC}_y(k, b) = \frac{\sum_{f=l_b}^{h_b} f\, w'_b(f)|Y'(k, f)|^2}{\sum_{f=l_b}^{h_b} w'_b(f)|Y'(k, f)|^2}, \tag{3.8}$$

where $b \in \{0, 1, \ldots, B-1\}$ and $w'_b(f)$ is the corresponding subband filter. In this work, to simplify implementation, we use the same bank of triangular mel-scale filters for both MFCC and SSC calculations, i.e. $w'_b(f) = w_b(f)$, but this constraint could be relaxed.

Finally, following [74], the SSC values are normalized within the range $[-1, 1]$, which is more convenient for use in neural network layers and activation functions. The normalized SSC (NSSC) features are obtained as,

$$\mathrm{NSSC}_y(k, b) = \frac{2\,\mathrm{SSC}_y(k, b) - (h_b + l_b)}{h_b - l_b}. \tag{3.9}$$

For later reference, we define the NSSC feature vector of signal $y[m]$ at the current frame k as,

$$\mathbf{NSSC}_y = [\mathrm{NSSC}_y(k,0), \ldots, \mathrm{NSSC}_y(k,B-1)]. \tag{3.10}$$

3.2.3 Spectral Energy Peaks (SEP)

SEP have been used for music identification systems in [75] where a time-frequency point is considered as a peak if it has higher amplitude than its neighboring points. SEP is argued to be intrinsically robust to even high-level background noise and can provide discrimination in sound mixtures [76]. Shazam's system [75] is a very good real-world application of this type of audio features where time-frequency coordinates of the energy peaks are described as sparse landmark points. Furthermore, by using pairs of landmark points rather than single points, SEP can be exploited to characterize the spectral structure of sound sources. In [77], start times of the SEP, referred to as onsets, are used for the automatic alignment of audio occurrences in an audio fingerprinting system.

3.2.4 Spectral Band Energies (SBE)

In addition to SEP, the SBE have been widely used in fingerprinting algorithms [78]. Let us denote $Y(k,f)$ as the STFT coefficients of an audio signal at time frame index k and frequency bin index f, $0 \leq f \leq M/2$. Let us also consider an auditory-motivated filterbank denoted with $w_b(f)$, e.g., in either Mel, Bark, Log, or Cent scale, with l_b and h_b as the lower and upper frequencies and $b \in \{0, 1., ..., B-1\}$ as the subband index subband. The SBE are then computed as,

$$\mathrm{SBE}_y(k,b) = \frac{\sum_{f=l_b}^{h_b} w_b(f)|Y(k,f)|^2}{\sum_{f=0}^{M/2} w_b(f)|Y(k,f)|^2}. \tag{3.11}$$

3.2.5 Spectral Flatness Measures (SFM)

SFM, also known as Wiener entropies, characterize the tonality aspect of an audio signals within different subbands and are therefore often used as an audio matching feature to distinguish different recordings [79]. The SFM for each time-frequency subband point (k, b) is computed as,

$$\mathrm{SFM}_y(k,b) = \frac{(\prod_{f=l_b}^{h_b} |Y(k,f)|^2)^{\frac{1}{h_b - l_b + 1}}}{\frac{1}{h_b - l_b + 1} \sum_{f=l_b}^{h_b} |Y(k,f)|^2}. \tag{3.12}$$

A high SFM in a given subband indicates the similarity of signal power over all frequencies within that subband, while a low SFM means that signal power is concentrated in a relatively small number of frequency bins over the full subband. This feature is shown to be a measure of multiplicative noise in [80].

Chapter 4

Proposed Method

In this chapter, the proposed Audio FingerPrinting (AFP) feature combination made out of MFCC and NSSC is presented, and its incorporation into GAN is explained. Finally, the procedure for synthesizing the final enhanced speech using these features within GAN is explained.

4.1 Proposed Feature Combination

In this thesis, we propose to use the concatenation of MFCC and NSSC vectors, along with some of their first and second differences (i.e., delta and double-delta) for training the GAN architecture. In the sequel, we refer to this extended feature set as AFP Combination (AFPC). Other AFP features, such as SEP are not included in our combination, since their information content is redundant when combined with the SSC. The MFCC and their deltas have long been used as an efficient alternative to the STFT, as they contain crucial information about the spectral subband energies and their temporal evolution [81].

Nevertheless, due to the smoothing nature of (3.5), the MFCC ignore the dynamics of

the formant present in each subband. In contrast, the NSSC and their deltas can provide critical information about the formant locations and their temporal variations. At the same time, the NSSC tend to be more noise-robust, compared to the MFCC, since the formant locations are not significantly disturbed by the additive noise distortion [57]. Thence, the proposed AFPC features have the ability to capture information about the distribution of energy, both across and inside spectral subbands. The NSSC and MFCC use the same STFT spectrum and mel-filterbank and they share the same processing unit for computing the STFT coefficients and the SSE. Thus, compared to the MFCC, the AFPC maintains the computational efficiency, even though it increases the memory requirements.

To obtain the AFPC, the MFCC and NSSC are both extracted from the STFT of the noisy signal, $Y(k, f)$ as described in Chapter 3. The proposed AFPC feature vector at the k-th time frame for signal $y[m]$ is then defined as,

$$\mathbf{AFPC}_y = [\mathbf{MFCC}_y, \Delta\mathbf{MFCC}_y, \Delta^2\mathbf{MFCC}_y, \mathbf{NSSC}_y, \Delta\mathbf{NSSC}_y, \Delta^2\mathbf{NSSC}_y], \quad (4.1)$$

where $\Delta\mathbf{MFCC}_y$ and $\Delta^2\mathbf{MFCC}_y$ are the deltas and double-deltas of the MFCC. Similarly, $\Delta\mathbf{NSSC}_y$ and $\Delta^2\mathbf{NSSC}_y$ are the deltas and double deltas of the NSSC.

4.2 Incorporation of AFPC within GAN

We assume that the magnitude spectrum of the noisy speech can be approximated by the sum of the clean speech and noise magnitude spectra, i.e, $|Y(k, f)| \approx |S(k, f)| + |N(k, f)|$. The generator in the adversarial setting is trained to predict a *real* output, which is taken as the Ideal Ratio Mask (IRM) generated from the known clean speech and noise signals [32],

i.e.,

$$\mathrm{IRM}(k,f) = \sqrt{\frac{|S(k,f)|^2}{|S(k,f)|^2 + |N(k,f)|^2}}, \tag{4.2}$$

where $\mathrm{IRM}(k,f)$ is the IRM value at the k-th frame and frequency bin f. We define the IRM vector at the current frame k as $\mathbf{IRM} = [\mathrm{IRM}(k,0), ..., \mathrm{IRM}(k,M/2)]$. Then, the generator produces the estimated IRM whose patterns and distribution should be close to the real IRM, as expressed by,

$$\widehat{\mathbf{IRM}} = G(\boldsymbol{z}, \mathbf{AFPC}_y^j), \tag{4.3}$$

where $\mathbf{AFPC}_y^j$ represents the *extended* AFPC feature vector at the current frame, obtained by concatenating the AFPC feature vectors from a subset of $2j+1$ consecutive context frames centered at the current one (i.e., by including the j adjacent frames to its left and right). The estimated output $\widehat{\mathbf{IRM}}$ in (4.3) is only calculated for the current frame. By examining $\widehat{\mathbf{IRM}}$ and the $\mathbf{AFPC}_y$ of the current frame, D decides whether its input is the real IRM from (4.2), or the *fake* output $\widehat{\mathbf{IRM}}$, this decision is reflected in,

$$D(\widehat{\mathbf{IRM}}, \mathbf{AFPC}_y) \in \{fake \equiv 0, real \equiv 1\}. \tag{4.4}$$

In [43], it is reported that having an extra term in training the generator using CGAN is very useful. Pandey *et al.* [47] show that using a penalty term based on the ℓ_1-norm gives a better performance compared to the ℓ_2-norm in speech enhancement applications. This approach allows adversarial component to produce more refined and realistic results. The weight of the ℓ_1-norm component in the objective function is controlled by a parameter

$\lambda > 0$. Therefore, the objective functions from (2.16) are modified as,

$$\min_{D} V(D) = \mathbb{E}[(D(\mathbf{IRM}, \mathbf{AFPC}_y) - 1)^2] + \mathbb{E}[(D(G(\boldsymbol{z}, \mathbf{AFPC}_y^j), \mathbf{AFPC}_y))^2], \quad (4.5)$$

$$\min_{G} V(G) = \mathbb{E}[(D(G(\boldsymbol{z}, \mathbf{AFPC}_y^j), \mathbf{AFPC}_y) - 1)^2] + \lambda \|G(\boldsymbol{z}, \mathbf{AFPC}_y^j) - \mathbf{IRM}\|_1, \quad (4.6)$$

where $\|.\|_1$ denote the ℓ_1 norm of its vector argument.

A schematic diagram of this adversarial training procedure is illustrated in Fig. 4.1. The training consists of three consecutive steps: First, D is trained with a concatenation of the **IRM** vector and the $\mathbf{AFPC}_y$ feature vector, in such a way that it recognizes the **IRM** as real (or output 1). Next, D learns to categorize the concatenation of the $\widehat{\mathbf{IRM}}$ and $\mathbf{AFPC}_y$ feature vector as fake data distribution (or output 0). Finally, the D variables are frozen and the G is trained with the $\mathbf{AFPC}_y^j$ features to fool the D.

4.3 Overall System

A block diagram of the system architecture is depicted in Fig. 4.2. The operation consists of two stages: training and enhancement, where the corresponding processing paths are shown by continuous and dashed blue lines, respectively. During the training stage, the system uses the AFPC feature set to train the D and G as shown in Fig. 4.1 in an adversarial setting. Using G, the model learns estimates the output IRM.

In the enhancement stage, the estimated IRM by G for every frame and frequency index is used as a Wiener type of filter on the STFT magnitude of the noisy speech. This method only enhances the amplitude of the signal and uses the phase from the noisy speech to reconstruct the time-domain enhanced signal using the overlap-add and Inverse STFT

(ISTFT) as shown in,

$$|\hat{S}(k,f)| = \widehat{\text{IRM}}(k,f)|Y(k,f)|, \tag{4.7}$$

$$\hat{s}[m] = \text{ISTFT}\{|\hat{S}(k,f)|e^{jk\angle Y(k,f)}\}. \tag{4.8}$$

This method could be further improved to enhance the speech phase as well, but in this work we only limit our model to enhancing the speech magnitude. In [4], the authors study the importance of phase in speech enhancement. A possible solution to exploit the speech phase in the present ANN context is by using Complex IRM (cIRM) proposed by [82] or phase-sensitive solutions such as [37, 83].

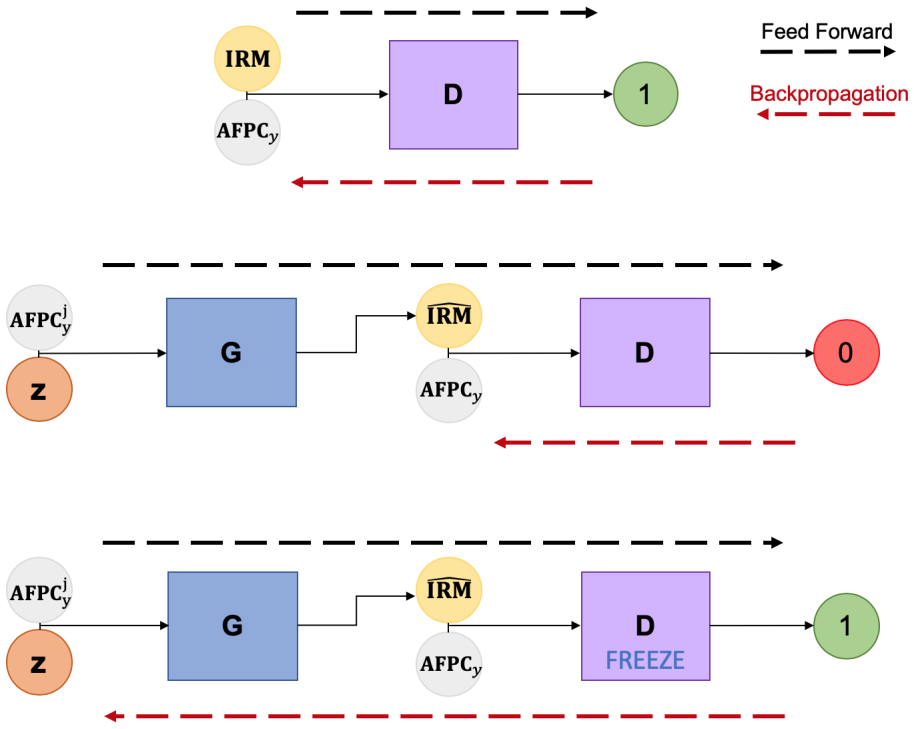

Fig. 4.1 The Proposed GAN training procedure used with the AFPC. First, the discriminator is trained with a concatenation of real IRM and AFPC features of the noisy signal. Next, the discriminator is trained with the estimated IRM and the noisy AFP features. Finally, the discriminator is frozen and the generator is trained with AFPC features so that it fools the discriminator.

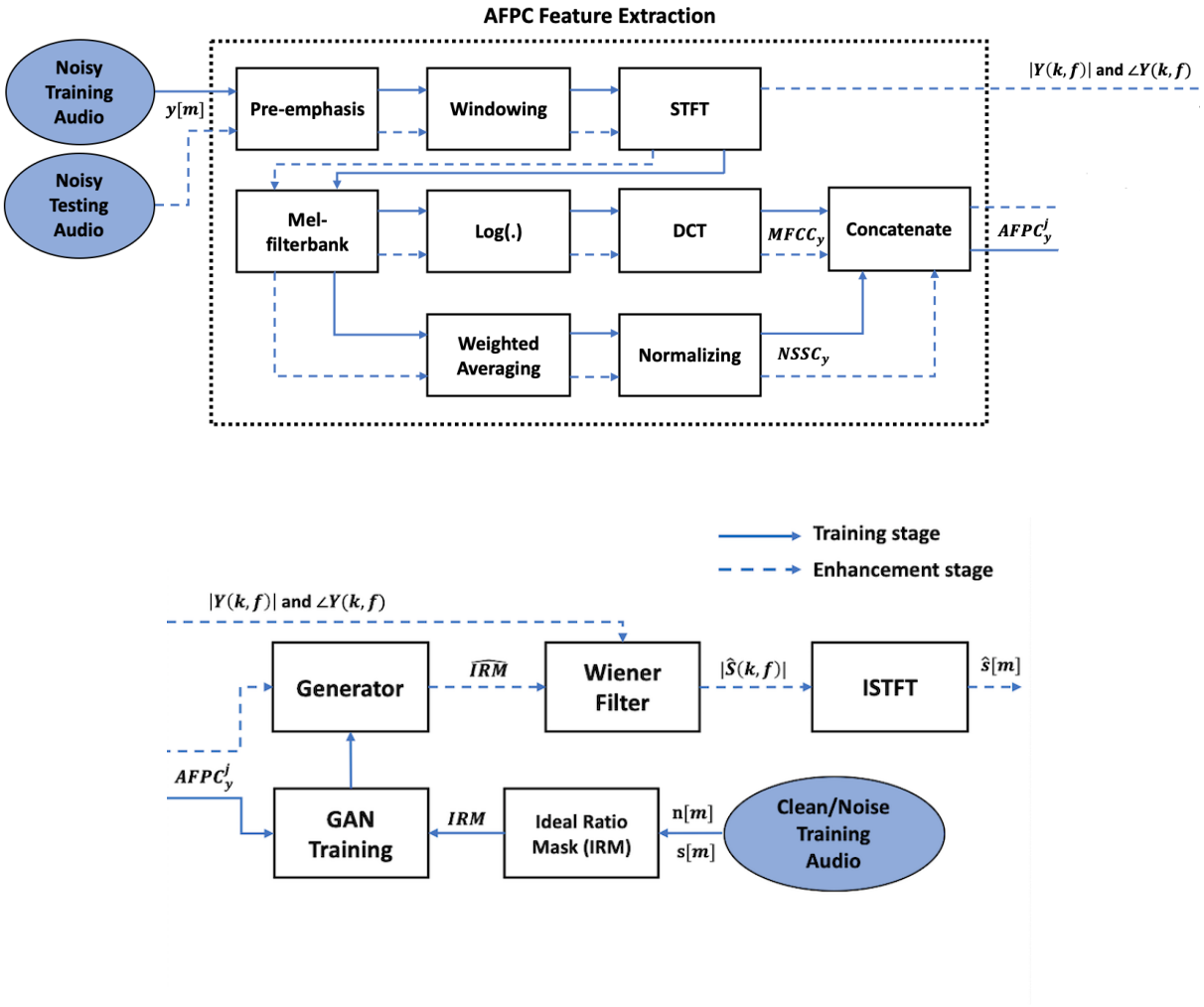

Fig. 4.2 Block diagram of the proposed AFPC feature set extraction (top) and its incorporation into GAN (bottom).

Chapter 5

Experiments and Results

In this chapter, we first describe the experimental setup and methodology used to train and evaluate the performance of the proposed GAN system for speech enhancement. We then present and discuss experimental results where the performance of the system is compared for different combination of STFT-based feature sets, including the proposed Audio Finger-Printing (AFP) combination, consisting of MFCC and NSSC. The performance is evaluated in terms of objective measures, i.e., PESQ, SDR and STOI, as well as model complexity and training time.

5.1 Experimental Setup

5.1.1 Dataset

We use the LibriSpeech [84] dataset which is an open corpus based on audio books and containing 1000 hours of relatively noise-free speech in English. The corpus provides us with the speaker and word diversity required for a speaker-general speech enhancement system. For training, 1755 utterances are randomly selected from 250 speakers (half male,

half female) for a total of 6 hours of speech. For testing, 255 different utterances are selected from 40 speakers (half male, half female), for a total of 30 minutes of speech. The clean files are contaminated with additive noise at -5dB, 0dB and 5dB SNRs for both training and testing sets, while two extra SNRs of 10dB and 15dB are added for testing under unmatched SNR conditions. Five different noise types from NOISEX-92 [85] are used for both training and testing: babble, pink, buccaneer2, factory1 and hfchannel.

All the audio files are sampled at 16 KHz. The STFT coefficients are extracted with an $M = 512$ STFT, using a 32ms Hanning window, overlap of 50% ($L = 256$) and three context frames (i.e. $j = 1$). The MFCC and NSSC are computed from the STFT parameters using $B = 64$ subbands with mel-frequency triangular filters $w_b(f)$ distributed between 0Hz and 8KHz. The number of MFCC is set to $P = 22$ while for NSSC, only the first 22 coefficients are kept in the feature vector. The pre-emphasis factor $\alpha = 0.97$ is used in (3.4). The delta and double-delta variations are included in the feature sets for each context frame [57]. The estimated IRM (4.3) is calculated only for the middle STFT frame. For each feature set, one model is trained for all noise types, SNRs and speakers.

5.1.2 Training

The parameters used to build each system is explained here. Most of the parameters in this section are empirically found or dynamically tuned for the specific conditions to avoid under/over-fitting. The generator's architecture has three hidden layers, each including 512 nodes. The ReLU activation function is used after each hidden layer with a dropout rate of 0.2. The discriminator has the same structure as the generator but uses instead the leaky ReLU activation function. Both employ the sigmoid activation at the output layer because they predict the IRM. A batch normalization layer with momentum 0.8 is used after each dense layer [86]. The latent vector $\boldsymbol{z}$ has 15 elements generated randomly from a normal

Gaussian distribution. The GAN architecture is trained in 50 epochs with a learning rate of 10^{-4} for the first half and 10^{-5} for the second half of the epochs. The batch size is set to 128 and ADAM optimizer [87] is used for training. We set $\lambda = 100$ in (4.6), which provides good convergence.

5.1.3 Evaluation

In our evaluation, we compare the effect of different combinations of feature sets on the overall performance of the GAN-based speech enhancement system. Specifically, we consider the previously discussed STFT coefficients, MFCC and NSSC as basic feature sets, along with various combinations thereof. The MFCC and NSSC features always include the delta and double delta coefficients. The various combinations are designated with "+", which means concatenation of the indicated feature vectors. Out of the seven distinct possible combinations, MFCC+NSSC corresponds to the proposed AFPC feature vector in (4.1). For each comparative experiment, the same GAN architecture is trained independently for each combination of features using all SNRs and noise types, audio training, and hyper-parameters.

The feature sets are compared objectively in terms of PESQ, which provides a measure of signal quality between -0.5 and 4.5, Signal-to-Distortion Ratio (SDR) which measures the speech quality in dB based on the introduced speech distortion, and Short-Time Objective Intelligibility (STOI), which provides a measure of intelligibility between 0 and 1. Here, we use a version of PESQ called ITU-T P.862.2 which is the second version of the original PESQ [88]. The comparative performance results demonstrate the effectiveness of each combination as well as the amount of information present in the concatenated feature vectors. Besides these performance measures, we also compare the different feature combinations in terms of system efficiency, i.e. feature vector size, training time per epoch, and

number of network parameters.

5.2 Results and Discussion

In this section, we study how STFT, MFCC and NSSC perform when used individually or when combined into an extended feature vector. In particular, the proposed system with audio fingerprinting features, MFCC+NSSC, is objectively compared to other alternatives and their combinations in terms of PESQ, SDR and STOI measures. The baseline system is STFT-based GAN which is widely used in the recent literature. The comparative results demonstrate to some extent the information overlap between different feature sets and their combinations. Finally, the different combinations of features are compared in terms of the processing complexity and training time for the underlying GAN system.

5.2.1 Number of Context Frames

The number of context frames directly affects the overall system latency and complexity by increasing the number of inputs and the processing time, for both training and enhancement stages. Our goal here is to choose the least number of context frames while maintaining the performance. Ideally, increasing the number of context frames improves the enhancement performance as more information in input to the system. However, as we increase the number of context frames, the level of correlation (or mutual information) between the middle frame and the more distant ones decreases, so that the potential benefit of additional frames is diminished. Thus, the performance improvement resulting from increasing the number of context frames saturates at a certain level while the complexity continues to grow [89].

To select the optimal number of context frames (i.e., $2j + 1$), the PESQ, SDR and

STOI performances of three selected feature sets are studied and representative results are presented in Fig. 5.1, 5.2 and 5.3. According to the results in these figures, when the number of context frames increases from 1 to 9 (i.e., $j \in \{0, 1, 2, 3, 4\}$), the performance tends to improve for each feature set. However, since most of the gains for MFCC+NSSC and STFT+MFCC are obtained with 3 context frames, we use the value of $j = 1$ for all subsequent experiments.

Another interesting observation from these figures is that the performance increases sharply from 1 to 3 context frames for MFCC+STFT and STFT+MFCC and nearly saturates after 3 context frames. However, in the case of the STFT, the performance tends to increase more "linearly" (or less sharply) as j increases.

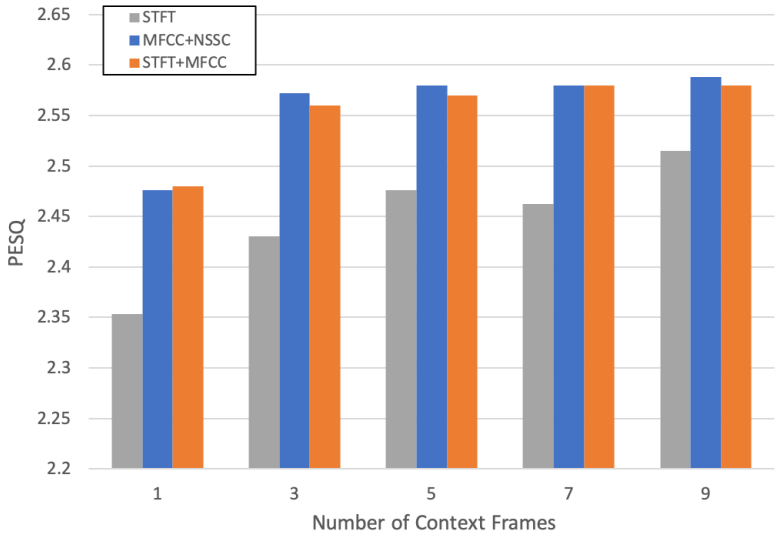

Fig. 5.1 Average PESQ performance for three feature sets: STFT (baseline), MFCC+NSSC and STFT+MFCC versus number of context frames $2j + 1$.

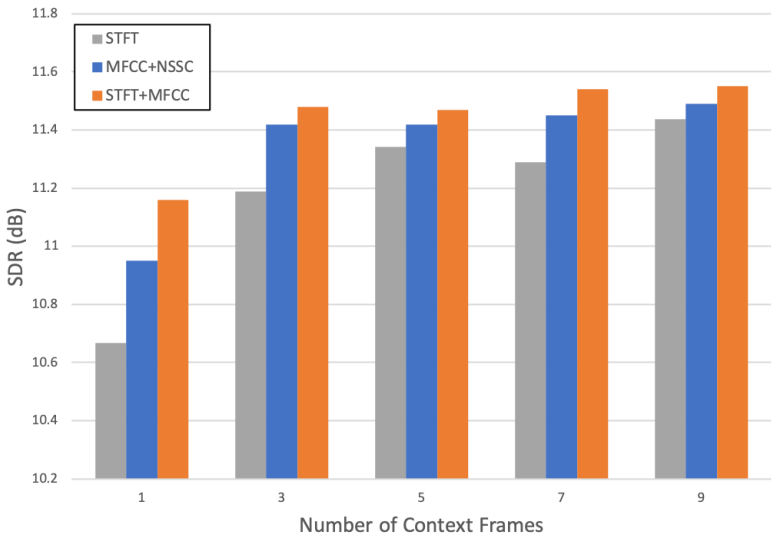

Fig. 5.2 Average SDR performance for three feature sets: STFT (baseline), MFCC+NSSC and STFT+MFCC versus number of context frames $2j + 1$.

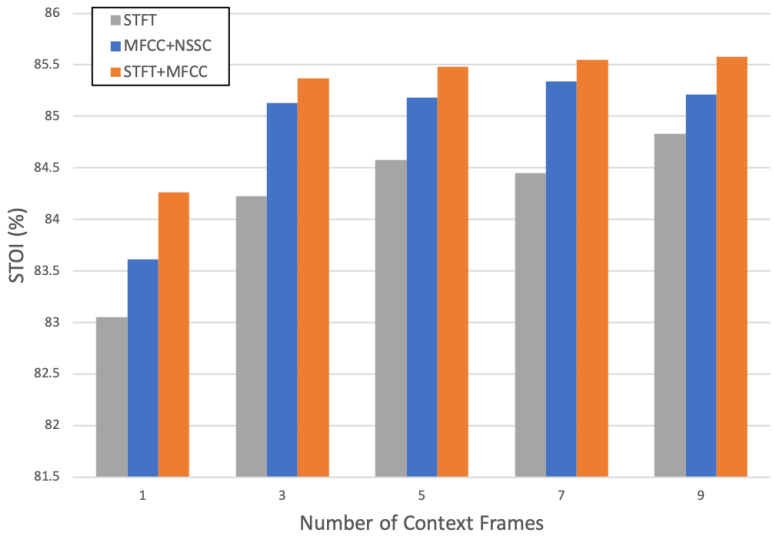

Fig. 5.3 Average STOI performance for three feature sets: STFT (baseline), MFCC+NSSC and STFT+MFCC versus number of context frames $2j + 1$.

5.2.2 Enhancement Performance

As explained in section 5.1, the performance results in terms of speech enhancement are reported for every combination of feature sets. Various noise types, i.e., babble, pink, buccaneer2, factory1 and hfchannel at various SNR levels, i.e., -5dB, 0dB, 5dB, 10dB and 15dB SNRs are used to conduct the experiments. For the purpose of applications, it is important to characterize the performance of the difference combinations of features for different noise types and SNR levels. Indeed, different applications such as aviation, hearing aids, etc. may involve quite different and specific noise environments and it is therefore of great interest to see how the trained GAN system for speech enhancement performance on different noise types. In this study, we train our model using all the noise types and SNR levels in a general format. However, studying noise types individually gives us a better idea of what to expect from our model under different conditions. The different noise types considered herein differ in terms of their fundamental attribute, i.e.: stationary versus non-stationary in the time domain and white versus colored in the frequency domain.

Babble noise: This non-stationary non-white noise is the most challenging noise type in speech enhancement scenarios since its temporal and spectral characteristics are similar to the target speech. This noise type is encountered in many different places such as restaurants, streets, etc. and it is therefore very useful to have a high performance for this noise type, specially for hearing aids and automatic speech recognition applications. Tables 5.1, 5.2 and 5.3 present the PESQ, SDR and STOI results obtained with the different feature combinations for the enhancement results of this noise type. In terms of PESQ, the more computationally demanding STFT+MFCC+NSSC combination achieves the best overall performance. The proposed AFPC=MFCC+NSSC outperforms the latter at low SNR, i.e. -5dB, while achieving near best performance at other SNRs. In terms of SDR,

there is a close competition between the last 4 combinations of feature sets, while a similar observation can be made for the STOI metric.

Table 5.1 Average PESQ results at various SNRs - babble Noise

Feature Set	**PESQ**				
	-5dB	0dB	5dB	10dB	15dB
Noisy	1.33	1.62	1.98	2.33	2.67
STFT	1.51	1.97	2.45	2.87	3.20
NSSC	1.40	1.90	2.36	2.73	3.04
MFCC	1.60	2.01	2.43	2.82	3.18
STFT+NSSC	1.56	2.04	2.52	2.95	3.26
STFT+MFCC	1.64	2.12	2.56	2.94	3.26
MFCC+NSSC	**1.66**	2.11	2.53	2.92	3.26
STFT+MFCC+NSSC	1.64	**2.13**	**2.57**	**2.95**	**3.27**

Table 5.2 Average SDR results at various SNRs - babble noise

Feature Set	SDR(dB)				
	-5dB	0dB	5dB	10dB	15dB
Noisy	-5.26	-0.40	4.55	9.54	14.53
STFT	0.56	5.44	9.83	13.94	17.40
NSSC	-0.25	4.58	8.75	12.33	15.14
MFCC	1.02	4.90	9.01	12.91	16.16
STFT+NSSC	0.95	5.68	9.94	**14.05**	17.56
STFT+MFCC	1.16	5.72	9.94	**14.05**	17.64
MFCC+NSSC	**1.47**	5.62	9.77	13.91	**17.70**
STFT+MFCC+NSSC	1.11	**5.76**	**9.95**	13.93	17.32

Table 5.3 Average STOI results at various SNRs - babble noise

Feature Set	STOI				
	-5dB	0dB	5dB	10dB	15dB
Noisy	0.56	0.67	0.78	0.87	0.93
STFT	0.63	0.75	0.85	0.91	0.94
NSSC	0.61	0.72	0.81	0.88	0.92
MFCC	0.66	0.76	0.84	0.90	0.94
STFT+NSSC	0.64	0.76	**0.86**	**0.92**	**0.95**
STFT+MFCC	0.66	**0.77**	**0.86**	**0.92**	**0.95**
MFCC+NSSC	**0.67**	**0.77**	**0.86**	**0.92**	**0.95**
STFT+MFCC+NSSC	0.66	**0.77**	**0.86**	**0.92**	**0.95**

Pink noise: Pink noise is a stationary colored noise which is often used as a reference signal in audio engineering applications. The power spectral density of pink noise decreases linearly on a logarithmic scale so that every octave contains the same amount of energy. It is important to study the effects of speech noise in the context of speech enhancement, since the latter is perceived similarly in different frequency bands. Tables 5.4, 5.5 and 5.6 illustrate the GAN-based enhancement results for this noise type when using different combinations of features. In this scenario, STFT+MFCC, MFCC+NSSC and STFT+MFCC+NSSC achieve the best and nearly similar performance for each one of the three measures, i.e. PESQ, SDR and STOI, while the proposed combination MFCC+NSSC slightly outperforms the other two at 15dB SNR. It is noteworthy that the STFT can achieve a decent performance with this noise type at lower SNR levels.

Table 5.4 Average PESQ results at various SNRs - pink noise

Feature Set	**PESQ**				
	-5dB	0dB	5dB	10dB	15dB
Noisy	1.09	1.39	1.74	2.12	2.50
STFT	1.83	2.25	2.65	2.87	3.01
NSSC	1.63	2.20	2.61	2.93	3.19
MFCC	1.76	2.24	2.64	2.96	3.18
STFT+NSSC	1.88	2.33	2.73	2.96	3.10
STFT+MFCC	1.93	**2.40**	**2.77**	3.04	3.18
MFCC+NSSC	1.90	2.37	2.75	**3.06**	**3.28**
STFT+MFCC+NSSC	**1.94**	**2.40**	**2.77**	3.02	3.17

Table 5.5 Average SDR results at various SNRs - pink noise

Feature Set	SDR(dB)				
	-5dB	0dB	5dB	10dB	15dB
Noisy	-5.10	-0.20	4.77	9.76	14.76
STFT	4.47	8.28	12.03	15.38	17.98
NSSC	3.49	7.55	11.19	14.23	16.51
MFCC	3.58	7.43	11.23	14.52	17.12
STFT+NSSC	4.81	8.49	12.22	15.55	18.16
STFT+MFCC	4.81	**8.49**	**12.19**	**15.65**	18.45
MFCC+NSSC	4.58	8.28	12.00	15.58	**18.68**
STFT+MFCC+NSSC	**4.83**	**8.50**	**12.20**	15.61	18.35

Table 5.6 Average STOI results at various SNRs - pink noise

Feature Set	STOI				
	-5dB	0dB	5dB	10dB	15dB
Noisy	0.55	0.68	0.80	0.89	0.94
STFT	0.71	0.81	**0.89**	**0.93**	0.95
NSSC	0.64	0.78	0.86	0.91	0.94
MFCC	0.69	0.80	0.88	0.92	0.95
STFT+NSSC	**0.72**	0.82	**0.89**	**0.93**	0.95
STFT+MFCC	**0.72**	**0.83**	**0.89**	0.93	**0.96**
MFCC+NSSC	0.71	0.82	**0.89**	**0.93**	**0.96**
STFT+MFCC+NSSC	**0.72**	**0.83**	**0.89**	**0.93**	0.95

Buccaneer2 noise: This noise type, a travelling jet noise recorded in a cockpit, is useful to study aviation applications, especially noise reduction for crew members operating in flights. This noise is colored and non-stationary due to changes in noise frequency emitted from the engine. Tables 5.7, 5.8 and 5.9 illustrate the results for this noise type. Similar results as in the case of pink noise can be observed for this noise type: better performance of AFPC at high SNRs and close performance for APFC, STFT+MFCC and STFT+MFCC+NSSC at lower SNR values. Again, the STFT performance is relatively good at low SNR, although not as good as the above three feature sets.

Table 5.7 Average PESQ results at various SNRs - buccaneer2 noise

Feature Set	**PESQ**				
	-5dB	0dB	5dB	10dB	15dB
Noisy	1.03	1.32	1.68	2.06	2.43
STFT	1.81	2.21	2.60	2.85	2.88
NSSC	1.59	2.11	2.54	2.85	3.09
MFCC	1.74	2.18	2.57	2.91	3.13
STFT+NSSC	1.87	2.29	2.68	2.91	2.91
STFT+MFCC	**1.91**	**2.36**	**2.72**	2.99	3.11
MFCC+NSSC	1.88	2.32	2.71	**3.01**	**3.21**
STFT+MFCC+NSSC	**1.91**	2.35	**2.72**	2.98	3.12

Table 5.8 Average SDR results at various SNRs - buccaneer2 noise

Feature Set	SDR(dB)				
	-5dB	0dB	5dB	10dB	15dB
Noisy	-4.99	-0.10	4.86	9.85	14.85
STFT	4.43	7.92	11.60	15.02	17.55
NSSC	3.48	7.25	10.90	14.22	16.84
MFCC	3.58	7.14	10.82	14.32	16.93
STFT+NSSC	**4.78**	8.12	**11.73**	15.14	17.57
STFT+MFCC	4.75	**8.17**	**11.72**	**15.22**	18.08
MFCC+NSSC	4.60	8.00	11.57	15.20	**18.37**
STFT+MFCC+NSSC	**4.78**	8.16	**11.73**	15.19	18.01

Table 5.9 Average STOI results at various SNRs - buccaneer2 noise

Feature Set	STOI				
	-5dB	0dB	5dB	10dB	15dB
Noisy	0.54	0.65	0.77	0.86	0.92
STFT	0.70	0.80	0.87	0.91	0.94
NSSC	0.62	0.75	0.84	0.90	0.93
MFCC	0.66	0.78	0.86	0.91	0.94
STFT+NSSC	0.71	**0.81**	**0.88**	0.91	0.94
STFT+MFCC	**0.72**	**0.81**	**0.88**	**0.92**	0.94
MFCC+NSSC	0.69	0.80	**0.88**	**0.92**	**0.95**
STFT+MFCC+NSSC	0.71	**0.81**	**0.88**	**0.92**	0.94

Factory1 noise: This noise was recorded in a factory near plate-cutting and electrical welding equipment. These machines produce a non-stationary colored noise with a number of spectral peaks over a wide frequency band. It is useful in noise reduction applications intended for factory and construction workers. Tables 5.10, 5.11 and 5.12 illustrate the results for this noise type. In terms of PESQ, AFPC equals or outperforms all the other features sets over the complete SNR range. This is an important result considering the importance of PESQ as a speech quality metric and the challenges posed by this particular type of noise. In terms of SDR and STOI, the results are similar to the pink and buccaneer2 results, where the proposed AFPC only outperform the other feature sets at higher SNRs.

Table 5.10 Average PESQ results at various SNRs - factory1 noise

Feature Set	**PESQ**				
	-5dB	0dB	5dB	10dB	15dB
Noisy	1.02	1.34	1.68	2.04	2.41
STFT	1.48	1.94	2.38	2.73	2.97
NSSC	1.40	1.93	2.37	2.72	3.02
MFCC	1.55	1.99	2.40	2.78	3.08
STFT+NSSC	1.54	2.02	2.46	2.79	3.04
STFT+MFCC	1.61	**2.09**	**2.50**	2.83	3.08
MFCC+NSSC	**1.63**	**2.09**	**2.50**	**2.86**	**3.15**
STFT+MFCC+NSSC	1.60	**2.09**	**2.50**	2.83	3.09

Table 5.11 Average SDR results at various SNRs - factory1 noise

Feature Set	SDR(dB)				
	-5dB	0dB	5dB	10dB	15dB
Noisy	-5.60	-0.75	4.21	9.19	14.18
STFT	2.69	6.78	10.59	14.25	17.19
NSSC	2.35	6.30	10.00	13.31	15.91
MFCC	1.94	5.97	9.92	13.61	16.81
STFT+NSSC	3.08	6.99	10.78	14.45	17.45
STFT+MFCC	3.04	6.99	10.77	**14.50**	17.67
MFCC+NSSC	2.84	6.74	10.60	14.44	**17.96**
STFT+MFCC+NSSC	**3.09**	**7.02**	**10.79**	14.48	17.65

Table 5.12 Average STOI results at various SNRs - factory1 noise

Feature Set	STOI				
	-5dB	0dB	5dB	10dB	15dB
Noisy	0.55	0.67	0.79	0.87	0.93
STFT	0.67	0.78	0.86	0.91	0.94
NSSC	0.63	0.76	0.85	0.90	0.93
MFCC	0.67	0.77	0.85	0.91	0.94
STFT+NSSC	0.68	0.79	**0.87**	0.91	0.94
STFT+MFCC	**0.69**	**0.80**	**0.87**	**0.92**	**0.95**
MFCC+NSSC	0.68	0.79	**0.87**	**0.92**	**0.95**
STFT+MFCC+NSSC	**0.69**	**0.80**	**0.87**	**0.92**	**0.95**

Hfchannel noise: This is a classic noise type which has been since the beginning of telecommunications. The hfchannel noise is acquired from an High Frequency (HF) radio channel after demodulation at the receiver. It is useful in the study of noise reduction for traditional radio applications, since it embodies the main features of HF noise characteristics. This noise is stationary but colored: its power spectrum is nearly flat below (similar to white noise), exhibits a peak around 2kHz, and decreases thereafter. Tables 5.13, 5.14 and 5.15 illustrate the results for this noise type. In this case, it is interesting to note that AFPC=MFCC+NSSC, STFT+MFCC and STFT+MFCC+NSSC achieve the best performance, while the proposed AFPC slightly outperforms the other two combinations at 15dB SNR.

Table 5.13 Average PESQ results at various SNRs - hfchannel noise

Feature Set	**PESQ**				
	-5dB	0dB	5dB	10dB	15dB
Noisy	1.18	1.32	1.54	1.81	2.13
STFT	1.93	2.21	2.50	2.78	2.88
NSSC	1.76	2.19	2.51	2.79	3.01
MFCC	1.92	2.25	2.56	2.86	3.11
STFT+NSSC	2.00	2.31	2.61	2.87	2.88
STFT+MFCC	**2.04**	**2.36**	**2.64**	2.91	3.06
MFCC+NSSC	**2.04**	2.34	2.63	**2.93**	**3.16**
STFT+MFCC+NSSC	**2.04**	2.35	**2.64**	2.89	3.06

Table 5.14 Average SDR results at various SNRs - hfchannel noise

Feature Set	SDR(dB)				
	-5dB	0dB	5dB	10dB	15dB
Noisy	-5.13	-0.25	4.71	9.70	14.70
STFT	6.84	10.13	13.50	16.83	19.07
NSSC	6.19	9.80	13.15	16.01	17.91
MFCC	6.32	9.73	13.11	16.37	18.92
STFT+NSSC	7.17	10.46	13.78	16.92	18.87
STFT+MFCC	7.16	**10.45**	**13.76**	**17.08**	19.62
MFCC+NSSC	7.05	10.34	13.64	16.96	**19.71**
STFT+MFCC+NSSC	**7.19**	**10.45**	**13.77**	17.05	19.65

Table 5.15 Average STOI results at various SNRs - hfchannel noise

Feature Set	STOI				
	-5dB	0dB	5dB	10dB	15dB
Noisy	0.59	0.69	0.77	0.85	0.91
STFT	0.74	0.82	0.88	**0.92**	0.94
NSSC	0.70	0.80	0.86	0.90	0.93
MFCC	0.74	0.82	0.88	**0.92**	0.94
STFT+NSSC	**0.76**	0.83	0.88	**0.92**	0.93
STFT+MFCC	**0.76**	**0.84**	**0.89**	**0.92**	0.94
MFCC+NSSC	**0.76**	0.83	0.88	**0.92**	**0.95**
STFT+MFCC+NSSC	**0.76**	**0.84**	**0.89**	**0.92**	0.94

Average Performance: For each feature set, results are obtained for five different noise types at five SNR levels from -5dB to 15dB. Average PESQ, SDR and STOI measures over all noise types are reported in Tables 5.16, 5.17 and 5.18, where the best results (within 2% of the observed maximum) are highlighted for each SNR. When used separately, MFCC and NSSC improve the overall speech quality compared to the noisy speech but do not generally outperform STFT. The NSSC are weaker than STFT and MFCC in all three measures, although they lead to improvement over the noisy speech. This is due to the fact that NSSC is only a normalized weighted average and does not represent the energy contained in each subband. Comparing STFT with STFT+NSSC and STFT+MFCC indicates that both AFP features add important information to the STFT features, from the perspective of noise reduction. In particular, STFT+MFCC outperforms STFT+NSSC in terms of both PESQ and STOI, while achieving a similar SDR performance. Interestingly, the combination of the three feature sets STFT+MFCC+NSSC does only matches the performance of the previous pairs of features, but does outperform them in any significant way, suggesting that it only brings redundant information to the training of the GAN system.

According to Tables 5.16-5.18, the proposed AFPC, i.e., MFCC+NSSC, substantially increases the performance of the GAN-based speech enhancement system in all three measures compared to MFCC or STFT. Furthermore, MFCC+NSSC achieves the best PESQ performance (within the error margin) and demonstrates a performance close to STFT+MFCC in terms of SDR and STOI. In particular, MFCC+NSSC outperforms the other feature sets in all three measures at high unmatched SNR of 15dB. This is due to the fact that at such high SNR, the additive noise does not significantly corrupt the extraction of formant frequencies with NSSC.

Table 5.16 Average PESQ Results for all noise types at various SNRs

Feature Set	PESQ				
	-5dB	0dB	5dB	10dB	15dB
Noisy	1.13	1.40	1.72	2.07	2.43
STFT	1.71	2.12	2.52	2.82	2.99
NSSC	1.56	2.07	2.48	2.80	3.07
MFCC	1.69	2.11	2.50	2.84	3.12
STFT+NSSC	1.77	2.20	2.60	2.90	3.04
STFT+MFCC	**1.83**	**2.27**	**2.64**	**2.94**	3.14
MFCC+NSSC	**1.82**	**2.25**	**2.63**	**2.96**	**3.21**
STFT+MFCC+NSSC	**1.83**	**2.26**	**2.64**	**2.93**	3.14

Table 5.17 Average SDR Results for all noise types at various SNRs

Feature Set	SDR(dB)				
	-5dB	0dB	5dB	10dB	15dB
Noisy	-5.21	-0.34	4.62	9.61	14.6
STFT	3.80	7.71	11.5	15.1	17.8
NSSC	3.05	7.10	10.8	14.0	16.5
MFCC	3.17	6.96	10.7	14.3	17.2
STFT+NSSC	**4.16**	**7.95**	**11.7**	**15.2**	17.9
STFT+MFCC	**4.18**	**7.96**	**11.7**	**15.3**	18.3
MFCC+NSSC	**4.11**	**7.80**	**11.6**	**15.2**	**18.5**
STFT+MFCC+NSSC	**4.20**	**7.98**	**11.7**	**15.2**	18.2

Table 5.18 Average STOI Results for all noise types at various SNRs

Feature Set	STOI				
	-5dB	0dB	5dB	10dB	15dB
Noisy	0.56	0.67	0.78	0.87	0.93
STFT	0.69	0.79	**0.87**	**0.92**	**0.94**
NSSC	0.64	0.76	0.85	0.90	0.93
MFCC	0.68	0.79	0.86	0.91	**0.94**
STFT+NSSC	**0.70**	**0.80**	**0.88**	**0.92**	**0.94**
STFT+MFCC	**0.71**	**0.81**	**0.88**	**0.92**	**0.95**
MFCC+NSSC	**0.70**	**0.80**	**0.88**	**0.92**	**0.95**
STFT+MFCC+NSSC	**0.71**	**0.81**	**0.88**	**0.92**	**0.95**

Fig. 5.4 shows the spectrograms of: (a) clean speech; (b) noisy speech after contamination with babble noise at 0dB SNR; (c) enhanced speech using GAN with STFT; (d) enhanced speech using MFCC features; (e) enhanced speech using the combination of STFT and MFCC, and; (f) enhanced speech using the proposed AFPC. It can be seen that the proposed AFPC features preserve the speech formants while removing more noise during non-speech segments. Visually the difference between AFPC and STFT+MFCC is not detectable, except for some small parts where the AFPC removes more isolated speech formants in the spectrogram. Apart from that, the visualization of the signals with the spectrograms is consistent with the results in Table 5.16, 5.17 and 5.18.

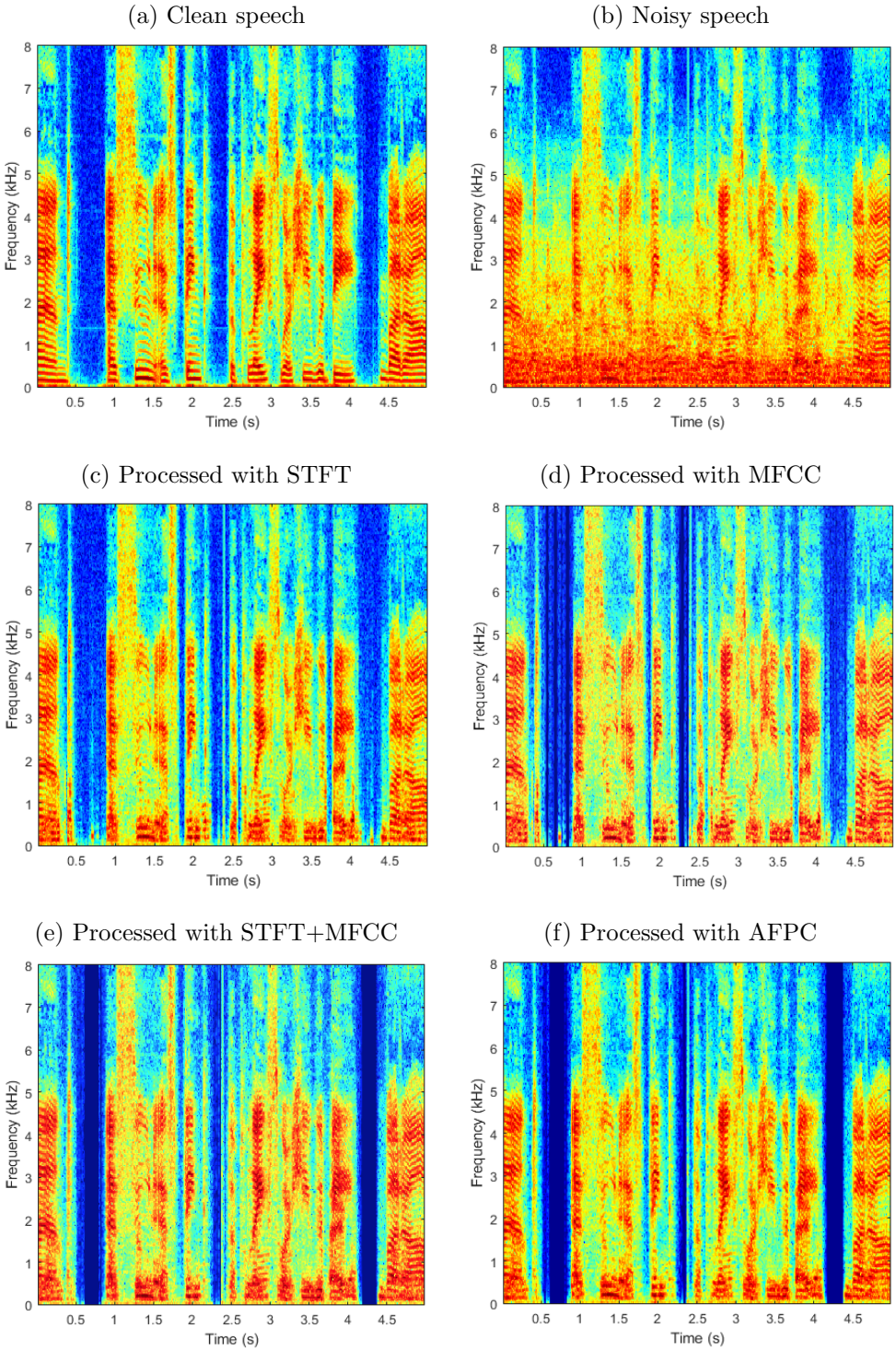

Fig. 5.4 Spectrograms of (a) Clean speech (b) Noisy speech (0dB babble noise) (c) Processed speech using STFT features (d) Processed speech using MFCC features (e) Processed speech using STFT+MFCC (f) Processed speech using the AFPC features.

5.2.3 Complexity Analysis

While the bottom 3 feature sets in Tables 5.16, 5.17 and 5.18, i.e., STFT+MFCC, MFCC+NSSC and STFT+MFCC+NSSC, achieve the best performance in terms of average PESQ, STOI and SDR, the cost of this improvement for a GAN-based system using STFT in combination with other features is much more than for the proposed AFPC=MFCC+NSSC. As shown in Table 5.19, the latter significantly outperforms the STFT-based combinations in terms of feature size, training time and number of network parameters. Specifically, compared to STFT+MFCC, the AFPC leads to reductions of 59.1% in memory storage for the training data, 43.3% in training time for the GAN system, and 25.0% in the number of network parameters. Compared to the STFT baseline, MFCC+NSCC requires 49.6% less memory storage for features and 30.1% less training time, while achieving significant performance improvements. The savings in training time and network size with the proposed AFPC become larger when we add more context frames (i.e., $j > 1$). The testing time is not reported in Table IV since it is almost the same for all systems. In testing, most of the processing time is allocated to the STFT computation which is needed for all feature combinations.

Table 5.19 Size of Feature Vector, Training Time per Epoch and number of Network Parameters for Different Combinations of Features.

Feature Set	Average PESQ	Feature Size	Training Time per epoch	Network Param.
STFT	2.43	257	17.6 mins	1.06M
NSSC	2.39	**66**	**10.5 mins**	**770K**
MFCC	2.47	**66**	**10.5 mins**	**770K**
STFT+NSSC	2.50	323	21.7 mins	1.16M
STFT+MFCC	**2.56**	323	21.7 mins	1.16M
MFCC+NSSC	**2.57**	**132**	**12.3 mins**	**870K**
STFT+MFCC+NSSC	**2.56**	389	24.9 mins	1.26M

Chapter 6

Conclusion and Future Work

This chapter provides some concluding remarks about the research presented in this thesis. Specifically, Section 6.1 presents a brief summary of the thesis work and contributions, while Section 6.2 provides suggestions for possible future work in this active area.

6.1 Conclusion

In this work, we proposed using a compact set of features obtained from the combination of two AFP techniques, i.e., MFCC and NSSC, to implement a speech enhancement system based on GAN and trained to predict the IRM of the noisy speech. The NSSC capture the speech formants and the distribution of energy in each subband, and therefore complement the MFCC in a crucial way.

Below, we present a chapter-wise sequential overview of the main topics discussed in this work.

- In Chapter 1, the speech enhancement problem was exposed. This was followed by a literature survey on the conventional (i.e., statistical) and machine learning methods.

Finally, an overview of the audio features used for the latter type of methods was presented.

- In Chapter 2, a brief introduction to deep learning was first presented. Then, motivated by biological networks, the principles of neural networks were reviewed. This was followed by the presentation of Generative Adversarial Network concepts.

- In Chapter 3, a detailed description of the underlying speech model and the basic STFT audio feature extraction was given. This was followed by a comprehensive explanation of two important AFP features, namely: MFCC and NSSC. Other AFP features of interest were also briefly introduced.

- In Chapter 4, the proposed AFPC feature set, which consists of the combination of the MFCC, NSSC, their deltas and double deltas, was presented. This was followed by detailed explanations regarding the incorporation and use of this feature set within the GAN framework, with particular attention on the enhancement and reconstruction procedures.

- In Chapter 5, experimental results for several different audio feature combinations were presented and discussed in terms of three objective measures, i.e, PESQ, SDR and STOI. The results showed that the AFPC feature sets significantly outperform the two conventional STFT and MFCC features and perform as well as the more complex combinations of STFT+MFCC and STFT+MFCC+NSSC.

In experiments with diverse speakers and noise types, GAN-based speech enhancement with the proposed AFPC (MFCC+NSCC) achieved the best average performance in terms of PESQ, STOI and SDR objective measures. Furthermore, compared to the STFT+MFCC combination with nearly similar performance, AFPC led to reductions of about 60% in

memory storage, 45% in training time, and 25% in network size. Hence, the proposed AFPC set is a promising feature-extraction method in learning-based speech enhancement systems.

6.2 Future Works

In this section, we point out some possible directions for future research work. In this work we proposed using a combination of two AFP features within GAN, allowing us to achieve near best performance while reducing the number of parameters and training time in our system. Possible avenues for research include the following:

- It is possible to explore the use of AFP features with other neural network models and architectures such as CNN and RNN. It is difficult to claim that the same improvement is achievable with other deep learning architectures, so it would be interesting to further explore how different architectures perform with the proposed AFPC features.

- It is important to know the effect of unseen noise types when the system is tested on other noise types that are not seen during the training phase. Also, it is suggested to use more realistic and recent noise recordings compared to the ones used in this study.

- Statistical significance testing is usually overlooked in speech enhancement studies and it is useful to perform such a test in future speech enhancement model studies including future studies involving the AFP features.

- Besides the MFCC and NSSC, there are other AFP features which were shortly introduced in this thesis, but studying and analyzing them fell beyond the scope

of this research. It would be of interest to study how these other AFP features and their combinations can affect the performance of GAN-based and other learning-based speech enhancement systems.

- Finally, our study of APFC has shown that it is possible to achieve good performance in speech enhancement while using a reduced number of input features. A challenging area for possible study would be to explore and create a new feature set with even smaller dimensionality than the AFPC that can yet achieve a similar performance in the speech enhancement task. For applications with reduced vocabulary where processing complexity is a key factor, e.g, wake-up word detection on smart home devices, this type of investigation is indeed of great interest.

References

[1] J. Benesty, S. Makino and J. D. Chen, *Speech Enhancement*, Berlin, Germany, Springer, 2005.

[2] R. Hendriks, T. Gerkmann and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement*, Morgan & Claypool, 2013.

[3] P. C. Loizou, *Speech Enhancement - Theory and Practice*, Taylor and Francis, 2007

[4] K. Paliwal, K. Wójcicki and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465-494, 2011.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113-120, 1979

[6] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137, March 1999

[7] H. Hermansky, E. A. Wan and C. Avendano, "Speech enhancement based on temporal processing," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 405-408, May, 1995.

[8] T. H. Falk, S. Stadler, W. B. Kleijn and W. Y. Chan, "Noise suppression based on extending a speech-dominated modulation band," in *Proc. INTERSPEECH*, Antwerp, Belgium, pp. 2-5, August, 2007.

[9] K. Paliwal, K. Wójcicki and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain" *Speech communication*, vol. 52, no. 5, pp. 450-475.

[10] N. Dionelis and M. Brookes "Phase-aware single-channel speech enhancement with modulation-domain Kalman filtering," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 937-950, 2018.

[11] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 28, pp. 127-145, 1980.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[13] E. Plourde and B. Champagne, "Auditory-Based Spectral Amplitude Estimators for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1614-1623, 2008.

[14] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, USA, pp. 4266–4269, Mar. 2010.

[15] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, USA, pp. 177-180, April 1987.

[16] R. Ishaq, B. G. Zapirain, M. Shahid, and B. Lovstrom, "Subband Modulator Kalman filtering for Single Channel Speech Enhancement," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, pp. 7442-7446, May 2013.

[17] S. K. Roy, W. P. Zhu and B. Champagne, "Single channel speech enhancement using subband iterative Kalman filter," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Montreal, Canada, pp. 762-765, 2016.

[18] M. Dendrinos, S. Bakamides and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.*, vol. 10, pp. 45-57, 1991.

[19] Y. Ephraim and H. L Van Trees, "A signal subspace approach for speech enhancement," in *Proc. of IEEE Int. Conf. on Acoustic, Speech Signal Processing*, vol. 3, no. 4, Minneapolis, USA, pp. 251-266, July 1995.

[20] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 700-708, 2003.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, Salt Lake City, USA, vol.2, pp. 749-752, May 2001.

[23] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I - temporal alignment," *Journal of the Audio Engineering Society* vol. 61 no. 6, pp. 366-384, May 2013.

[24] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[25] Y. H. Tu, I. Tashev, S. Zarar and C. Lee, "A Hybrid Approach to Combining Conventional and Deep Learning Techniques for Single-Channel Speech Enhancement and Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 2531-2535, April 2018.

[26] S. Nie et al., "Exploiting spectro-temporal structures using NMF for DNN-based supervised speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 469-473, Mar. 2016.

[27] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[28] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, USA, pp. 4562–4565, Mar. 2010.

[29] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Vancouver, Canada, pp. 556–562, Sep. 2001.

[30] H. Chung, E. Plourde and B. Champagne, "Discriminative training of NMF model based on class probabilities for speech enhancement," in *IEEE Signal Process Lett*, vol. 3 no. 4, pp. 502-506, 2016.

[31] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii and T. Kawahara, "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 716-720, April 2018.

[32] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, Vancouver, Canada, pp. 7092–7096, May 2013.

[33] Y. Xu, J. Du, L. R. Dai and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Proc.*, vol. 23, no. 1, pp. 7-19, Jan. 2015.

[34] Y. Zhao, B. Xu, R. Giri and T. Zhang, "Perceptually Guided Speech Enhancement Using Deep Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 5074-5078, April 2018.

[35] K. Zhen, A. Sivaraman, J. Sung and M. Kim, "On Psychoacoustically Weighted Cost Functions Towards Resource-Efficient Deep Neural Networks for Speech Denoising," arXiv e-prints, 2018.

[36] F. Weninger, F. Eyben and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 3709-3713, May 2014.

[37] H. Erdogan, J. R. Hershey, S. Watanabe and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, pp. 708-712, April 2015.

[38] P. S. Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136-2147, Dec. 2015.

[39] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, Dec. 2014.

[40] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," arXiv:1703.08019, 2017.

[41] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in arXiv preprint arXiv:1609.07132, 2016.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde- Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets, in advances in neural information processing systems," in *Proc. NIPS*, Montreal, Canada, pp. 2672-2680, Dec. 2014.

[43] S. Pascual, A. Bonafonte, and J. Serr, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech*, Stockholm, Sweden, pp. 36423646, Aug. 2017.

[44] S. Pascual, M. Park, J. Serra, A. Bonafonte and K. Ahn, "Language and Noise Transfer in Speech Enhancement Generative Adversarial Network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 5019-5023, April 2018.

[45] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," arXiv:1711.05747, 2017.

[46] M. H. Soni, N. Shah and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP*, Calgary, Canada, pp. 5039-5043, April 2018.

[47] A. Pandey and D. Wang, "On Adversarial Training and Loss Functions for Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 5414-5418, April 2018.

[48] J. Abdulbaqi, Y. Gu, I. Marsic, "RHR-Net: A residual hourglass recurrent neural network for speech enhancement," arXiv:1904.07294, 2019.

[49] S. Mallat, *A wavelet tour of signal processing*, Academic Press, Second Edition, 1998.

[50] M. Gokhale et al., "Time domain signal analysis using wavelet packet decomposition approach," *Int. Journal of Commun. Netw. Syst. Sci.*, vol. 3, no. 3, p. 321, 2010.

[51] A. Bouzid, M. Ben Messaoud and N. Ellouze, "Speech enhancement based on wavelet packet of an improved principal component analysis," *Computer Speech Language*, vol. 35, pp. 58–72, 2016.

[52] D. Ribas, J. Llombart, A. Miguel, and L. Vicente, "Deep speech enhancement for reverberated and noisy signals using wide residual networks," arXiv:1901.00660, 2019.

[53] R. Razani, H. Chung, Y. Attabi, B. Champagne, "A reduced complexity MFCC-based DNN approach for speech enhancement," in *Proc. IEEE Symp. on Signal Process. and Information Tech.*, pp. six, Dec. 2017.

[54] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[55] Y. Wang, K. Han and D. Wang, "Exploring Monaural Features for Classification-Based Speech Segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 270-279, Feb. 2013.

[56] N. Q. Duong and H. T. Duong, "A review of audio features and statistical models exploited for voice pattern design," arXiv preprint arXiv:1502.06811. Feb. 2015.

[57] K. K. Paliwal, "Spectral subband centroids features for speech recognition," in *Proc. ICASSP*, vol. 2, Seattle, U.S, pp. 617–620, May 1998.

[58] N. Poh, C. Sanderson, and S. Bengio, "An investigation of spectral subband centroids for speaker authentication," in *Proc. Int. Conf. Biometric Authent. (ICBA)*, Hong Kong, pp. 631–639, 2004.

[59] A. Nicolson, J. Hanson, J. Lyons, and K. Paliwal, "Spectral subband centroids for robust speaker identification using marginalization-based missing feature theory," *Int. Journal of Signal Processing Systems*, vol. 6, no. 1, pp. 12–16, 2018.

[60] T. Kinnunen, B. Zhang, J. Zhu, and Y. Wang, "Speaker verification with adaptive spectral subband centroids," *Advanced Biometrics*, pp. 58–66, 2007.

[61] N. Thian, C. Sanderson, and S. Bengio, "Spectral subband centroids as complementary features for speaker authentication," in *Biometric Authenticat.*, pp. 1-38, 2004.

[62] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature* 521, pp. 436–444, 2015.

[63] P. Suganthan, "Pattern classification using multiple hierarchical overlapped self-organising maps," *Pattern Recognition*, vol. 34, no. 11, pp. 2173-2179, 2001.

[64] L. Brito da Silva, I. Elnabarawy and D. Wunsch, "A survey of adaptive resonance theory neural network models for engineering applications," *Neural Networks*, vol. 120, pp. 167-203, 2019.

[65] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[66] J. Ho, S. Ermon, "Generative adversarial imitation learning," *Advances in Neural Information Processing Systems*, pp. 4565–4573, 2016.

[67] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," arXiv: 1611.07004, 2016.

[68] G. Antipov, M. Baccouche and J. Dugelay, "Face aging with conditional generative adversarial networks," in Proc. IEEE Int. Conference on Image Processing (ICIP), Beijing, China, pp. 2089-2093, 2017.

[69] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784, 2014.

[70] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, and Z. Wang, "Least squares generative adversarial networks," arXiv: 1611.04076, 2016.

[71] M. Kolbæk, Z. Tan and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153-167, 2017.

[72] D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley, 1987.

[73] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Englewood Cliffs, Prentice-Hall, 1989.

[74] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, C. D. Yoo, "Audio fingerprinting based on normalized spectral subband centroids," in *Proc. ICASSP*, Philadelphia, USA, vol. 3, pp. 213-216, Mar. 2005.

[75] A. L.-C. Wang, "An industrial-strength audio search algorithm," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, pp. 1-4, Oct. 2003.

[76] J. Ogle and D. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 233–236, May 2011.

[77] C. V. Cotton and D. P. W. Ellis, "Audio fingerprinting to identify multiple videos of an event," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2386–2389, Mar. 2010.

[78] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, Oct. 2002.

[79] J. Herre, E. Allamanche, and O. Hellmuth, "Robust matching of audio signals using spectral atness features," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 127-130, 2001.

[80] T. H. Falk, and W. Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14 no. 6, pp. 1935-1947.

[81] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[82] D. S. Williamson, Y. Wang and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483-492, March 2016.

[83] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 8, pp. 1283–1294, Aug. 2015.

[84] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Brisbane, Australia, pp. 5206-5210, April 2015.

[85] A. Varga, H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol.12, no.3, pp. 247-252, 1993.

[86] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv:1502.03167, 2015.

[87] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014.

[88] ITU-T, "P.862: Revised Annex A - Reference implementations and conformance testing for ITU-T Recs P.862, P.862.1 and P.862.2", 2005, available at: https://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en

[89] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26 no. 7, pp. 1236-1246, 2018.