SPEECH ENHANCEMENT USING DEEP NEURAL NETWORK BASED ON MASK ESTIMATION AND HARMONIC REGENERATION NOISE REDUCTION FOR SINGLE CHANNEL MICROPHONE

NOREZMI BINTI MD JAMAL

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy in Electrical Engineering

Faculty of Electrical and Electronic Engineering
Universiti Tun Hussein Onn Malaysia

JANUARY 2022

To my lovely daughter, beloved husband,

caring parents, and parents in law

who are strongly support me throughout PhD journey.

Thank you very much

ACKNOWLEDGEMENT

In the Name of Allah, the Most Gracious and the Most Merciful. All praises and many thanks to Allah, this thesis finally managed to be completed successfully with patience, hard work and moral supports from many people.

First and foremost, I would like to take this opportunity to highly express appreciation to my beloved supervisor, Dr. Norfaiza binti Fuad, for her enthusiasm and guidance throughout this research journey. I would also like to acknowledge Dr. Farhanahani binti Mahmud, Dr. Fadhilah binti Rosdi, Prof. Syed Abdul Rahman Al-Haddad and Associate Professor Dr. Nabilah binti Ibrahim in sharing their knowledge and ideas. Also, many thanks to Prof. Syed Rahman and Dr. Tien-Ping Tan that provided me Malay datasets to run the experiments. My special thanks also go to Dr. Shahnoor Shanta for her constructive advice especially in long-distance communication.

Not to forget to my colleagues in the multidiscipline research group (MDRG) for their kind help and support throughout the completion of this thesis. Moreover, acknowledgment also goes to the support staff at the Faculty of Electrical and Electronic Engineering for their support in providing me with such great guidance and facility. This thesis has been financially supported by Universiti Tun Hussein Malaysia under Geran Penyelidikan Pascasiswazah (GPPS) and Short-Term Grant (STG).

My deepest appreciation goes to my family, especially my lovely father, Md Jamal bin Md Yusof, my lovely mother, Leli binti Mohd Dewa, and my beloved parents-in-law, Sha'abani bin Sukepi and Kasirah binti Saijoh for their 'Dua', love, motivation, support, and encouragement. Finally, I would like to express my special thanks and appreciation to my lovely and handsome husband, Mohd Nurul Al-Hafiz for love, understanding, support, motivation and sacrifice. Not to forget to my lovely daughter, Nur Awatif, who is understanding me a lot. Thank you very much.

ABSTRACT

The development of speech-enabled mobile applications has greatly improved human-computer interaction in recent years. These applications are flexible and convenient for users. Since the speech signal is captured in mobile conditions, it may easily be contaminated by background noises, which may result in a complicated computation and require speech enhancement algorithm. Thus, the performance of speech applications can be degraded when signal-to-noise ratio (SNR) is low and non-stationary noise is present. Moreover, the task of removing noises without causing speech distortion is also challenging, in which the quality and intelligibility of speech are affected. In order to overcome these issues, a supervised Deep Neural Network (DNN) algorithm predicted constrained Wiener Filter (cWF) target mask algorithm based on extracted Gammatone filter bank power spectrum (GF-TF) features and trained model is developed. As a result, the trained model with GF-TF features and cross-speech dataset produced promising results, while the proposed target mask scored higher on the perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) tests. On top of that, a modified Harmonic Regeneration Noise Reduction (HRNR) algorithm is proposed as a post-filtering strategy to enhance speech signal due to residual noise being introduced after DNN prediction. Results from TIMIT dataset revealed that average STOI scores for the joint algorithm are higher than those of DNN, conventional HRNR and Log Minimum Mean Square Error (Log-MMSE) algorithms. With SNR of -5 dB, an improvement of 4% over DNN algorithm, 36% over conventional HRNR algorithm, and 12% over Log-MMSE algorithm are obtained. While the average PESQ score is less affected after post-filtering strategy. Thus, this work has contributed to improve speech intelligibility from noisy backgrounds at low SNR as it can be deployed in speech-enabled mobile applications.

ABSTRAK

Perkembangan aplikasi mudah alih yang diaktifkan oleh pertuturan telah meningkatkan interaksi manusia-komputer dalam beberapa tahun kebelakangan ini. Aplikasi ini fleksibel dan mudah untuk pengguna. Oleh kerana isyarat pertuturan dirakam dalam keadaan mudah alih, ia mungkin mudah dicemari oleh bunyi latar belakang, yang menyebabkan pengiraan yang rumit dan memerlukan algoritma peningkatan pertuturan. Oleh itu, prestasi aplikasi pertuturan boleh berkurang apabila nisbah isyarat-ke-bunyi (SNR) adalah rendah dan kehadiran hingar tidak pegun. Selain itu, proses membuang bunyi bising tanpa menyebabkan gangguan pertuturan juga mencabar, di mana kualiti dan kepintaran pertuturan boleh terjejas. Untuk mengatasi isu-isu ini, algoritma rangkaian saraf dalam (DNN) meramalkan penapis algoritma topeng sasaran (cWF) yang diekstrak berdasarkan ciri-ciri spektrum kuasa bank penapis Gammatone (GF-TF) dan model terlatih telah dibangunkan. Hasilnya, model yang dilatih dengan ciri GF-TF dan gabungan pertuturan menghasilkan keputusan yang memberangsangkan, manakala topeng sasaran yang dicadangkan mendapat skor yang lebih tinggi pada ujian persepsi penilaian kualiti pertuturan (PESQ) dan kefahaman objektif jangka pendek (STOI). Di samping itu, algoritma pengurangan hingar penjanaan semula harmonik (HRNR) yang diubah suai telah dicadangkan sebagai strategi pasca-penapisan untuk meningkatkan isyarat pertuturan yang mana baki hingar terhasil selepas ramalan DNN. Hasil daripada set data TIMIT mendedahkan bahawa purata skor STOI untuk algoritma bersama HRNR adalah lebih tinggi daripada algoritma DNN, konvensional HRNR dan ralat kuasa dua purata minimum log (Log-MMSE). Peningkatan 4% ke atas algoritma DNN, 36% ke atas algoritma konvensional HRNR, dan 12% ke atas algoritma Log-MMSE ditemui pada -5 dB SNR. Manakala skor purata PESQ yang diperolehi kurang terjejas. Oleh itu, hasil kerja ini dapat meningkatkan kepintaran pertuturan daripada latar belakang bising pada SNR rendah supaya ia boleh digunakan dalam aplikasi mudah alih yang diaktifkan oleh pertuturan.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AdaGrad	-	Adaptive stochastic gradient descent
AMS	-	Amplitude Modulation Spectrogram
ASR	-	Automatic Speech Recognition
CASA	-	Computational Auditory Scene Analysis
CNN	-	Convolutional Neural Network
CPU	-	Central Processing Unit
cWF	-	Constrained Wiener Filter
DBN	-	Deep Belief Network
DCT	-	Discrete Cosine Transform
DNN	-	Deep Neural Network
FFT	-	Fast Fourier Transform
GF	-	Gammatone Filter bank
GMM	-	Gaussian mixture models
GVE	-	Global Variance Equalization
HRNR	-	Harmonic Regeneration Noise Reduction
IBM	-	Ideal Binary Mask
IRM	-	Ideal Ratio Mask
LMS	-	Least Mean Square
Log-MMSE	-	Log Minimum Mean Square Error
LSTM	-	Long Short-Term Memory
MAP	-	Maximum A Posteriori Estimator
MFCC	-	Mel Frequency Cepstrum coefficient
MLP	-	Multi-layer Perceptron
MRCG	-	Multi-Resolution Cochleagram
MSE	-	Mean Square Error
MMSE	-	Minimum Mean Square Error

NH	-	Normal hearing
PESQ	-	Perceptual evaluation of speech quality
PNCC	-	Power Normalized Cepstral Coefficient
Rasta-PLP	-	Relative Spectral Transformed Perceptual Linear Prediction
ReLU	-	Rectified Linear Unit
RNN	-	Recurrent Neural Network
SGD	-	Stochastic Gradient Descent
SNR	-	Signal-noise-to ratio
STFT	-	Short Time Fourier Transform
STOI	-	Short-time objective intelligibility
SVM	-	Support Vector Machine
T-F	-	Time-Frequency
$y(t)$	-	Noisy speech signal in time domain
$x(t)$	-	Clean speech signal in time domain
$n(t)$	-	Ambient noise signal in time domain
ζ_k	-	Prior SNR
v_k	-	Posterior SNR
y[k]	-	Discrete noisy speech data
x[k]	-	Discrete clean speech signal
n[k]	-	Discrete noise signal
f_s	-	Sampling frequency
$X_{STFT}[f, k]$	-	Clean speech signal in time-frequency domain
$N_{STFT}[f, k]$	-	Noise signal in time-frequency domain
cWf $[f, k]$.	-	Constrained Wiener Filter in time-frequency domain
H	-	Learning rate
$w_{i,j}$	-	Weight between layers
a_j^n	-	Hidden nodes in every hidden layer

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

This chapter provides an overview of the research background while emphasizing current issues on speech enhancement algorithms in low signal-to-noise ratio (SNR) and non-stationary noise, described in Sections 1.1 and 1.3. Subsequently, the objectives along with the scopes of research are presented in Sections 1.4 and 1.5, respectively. Research contributions and thesis outline are then briefly explained in Sections 1.6 and 1.7, respectively.

1.1 Research background

The development of speech-enabled mobile application has significantly improved human beings' daily activities and offers more flexibility. For example, in a remote context or hands-free computing scenario, humans interact with the computer system via verbal communication for purposes of navigation and speech-to-speech translations for foreign language understanding. Practically, a single-channel microphone in mobile application system is used to capture human speech signals and a central processing unit (CPU) is used to process the acquired raw speech signals. Due to the nature of distant speech-enabled mobile applications, acquiring speech sources will be easily degraded by background noise or unwanted sound in various scenarios depending on the users' situation. To ensure that the acquired speech signal could be learned by the device, domain knowledge of digital signal processing

techniques, acoustic theories, and mathematical algorithms including speech enhancement algorithm are required.

Speech is a spoken word created by the phonetic combination of a limited set of vowel and consonant speech sound units, known as phonemes such as /ba/ or /sha/. In contrast to a computer system, a human perceived speech sound using both ears during a spoken conversation situation. Thus, the intelligibility of speech could not be guaranteed as the expected phonemes when perceiving and recognizing noisy speech signals and the quality of degraded speech might be low. A noise like non-stationary noise and higher intensity of noise, which are more dominant among spoken words remain an issue to be tackled in mobile speech applications, even though vast speech applications have been commercialized and made available in the marketplace. Moreover, the goal of reducing noise using speech enhancement algorithm without generating speech distortion, which affects speech quality and intelligibility, is extremely difficult.

Generally, the perceived speech signal, $x(t)$ is mixed with ambient noises, $n(t)$ or acoustic environment, which is known as a noisy speech signal, $y(t)$. Noisy speech signals can be modelled by additive noise and mixed noise as shown in Equation 1.1 and Equation 1.2 (Zhao, 2000; Zhang *et al.*, 2018b). These noises may cause a mismatch between trained speech models and spectral features of test speech, and therefore often lead to severe degradation of recognition accuracy in speech-enabled mobile applications (Zhao, 2000).

$$y(t) = x(t) + n(t) \tag{1.1}$$

$$y(t) = x(t) * h(t) + n(t) \tag{1.2}$$

For information, the convolutive noise only exists when the distortions introduced by echo or reverberation are correlated with desired speech signal by the impulse response of the surrounding, $h(t)$ (Wölfel & McDonough, 2009). Since this thesis focuses on single-channel microphones during far-field speech sound acquisition, hence, the noisy speech signal with additive noise or background noise is badly affected as illustrated in Figure 1.1 (Wölfel & McDonough, 2009). Speech sound event occurred between 0 s and 0.5 s, which changes over time. While background noise is always there over time, it is possible that an overlapping occurrence happened.

This noise is also known as ambient noise, which is unwanted sound judged to be unpleasant, loud, or disruptive to hearing as well as causing difficulties for the computer to recognize and perceive the spoken words when noise signal is more dominant than speech signal. From a physics perspective, both noise and speech sound are vibrations through a medium, such as air or water. Hence, removing background noise is a challenging task due to different environmental characteristics, especially the non-stationary noise dominating speech signals (Vincent et al., 2017; Yuan, 2020), which is constantly changing compared to stationary noise (Parchami et al., 2016).

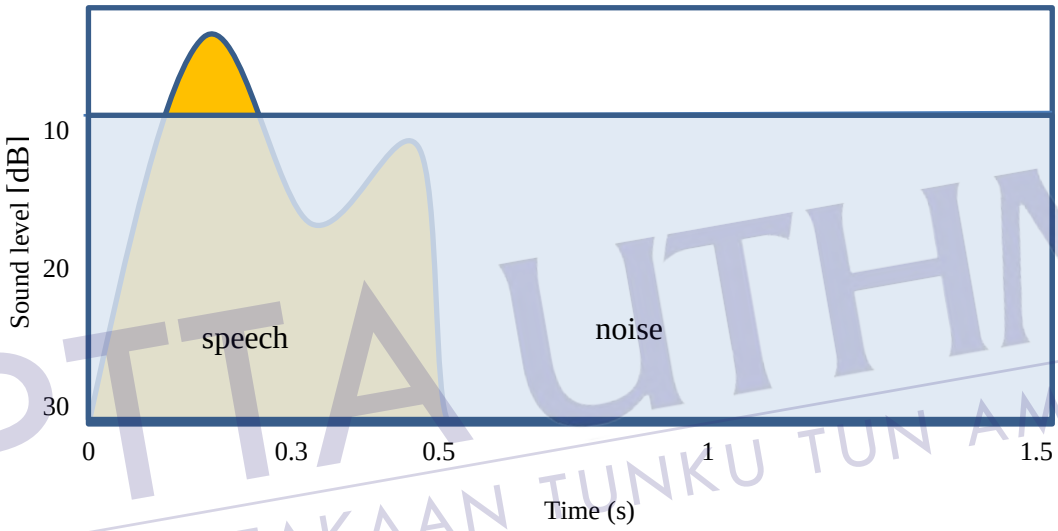


Figure 1.1: Noisy speech signal with background noise (Wölfel & McDonough, 2009)

Due to the aforementioned problems, this research focus on the improvement of a speech enhancement algorithm for a single-channel microphone. Hence, a supervised Deep Neural Network (DNN)-based mask estimation approach is utilized to predict a new target output mask from noisy speech signals. The DNN-based approach was chosen due to its capability to automatically learn the complex relationship between noisy and target mask speech signals. It is required to select the most significant features that are suitable and more discriminant to be used as the DNN algorithm input for speech enhancement processes. Besides, the DNN algorithm is used to generalize different datasets and also the different speakers at higher duration length. A post-filtering strategy is also a possible solution to rectify residual noise after speech reconstruction. This is to ensure that high-quality and more intelligible

enhanced speech signals can be perceived. Further elaboration can be found in the next session to better understand the research.

1.2 Research motivation

With the advancement of digital technology in the present era, most speech applications such as speech recognition and hearing aids are in high demand among consumers due to the significant impact of their usage. Mobility and portability will lead to the distortion of speech by background noise during the acquisition of speech signals. In the past several decades, several mathematical algorithms regarding speech enhancement have been proposed to deal with the issues. Basically, the speech enhancement techniques include two modes: (1) single channel processing; and (2) microphone-array processing. The single-channel microphone is more ideal compared to the microphone array in terms of deployment in mobile speech applications. However, the single-channel speech enhancement remains a challenging task when having higher noises than speech sources due to no reference channel to get noise information. Therefore, Deep Neural Network (DNN)-based mask estimation is preferred to alleviate noisy background due to the ability to construct the complex models for nonlinear processing (Saleem & Khattak, 2019). Moreover, it only depends on minimum mean squared error (MSE) between target mask reference and estimated target mask without consideration about the distribution of clean speech and noise signal.

1.3 Problem statement

The performance of speech applications such as speech recognition systems and a hearing aid in achieving high recognition accuracy in a clean environment is somewhat better than in a noisy environment. Recent studies show that there is room for improvement for speech application during noisy conditions (Wang, 2015; Chen, 2017; Odelowo, 2018). This is because, unlike the speech data of clean speech signals,

the contaminated speech signals have distinct thresholds in adverse environments and the speech signals vary in different environments with unexpected acoustic conditions (Zhang *et al.*, 2018b). The contaminated speech signals consist of the utterance of words or phrases that may be included by different types of background noises and signal-to-noise-ratio (SNR) level values, which may lead to the noisy acoustic effect. The intelligibility of consonants sounds may be easily affected due to its lowest magnitude of intensity compared to that of the vowel sounds. Hence, the greatest challenge in speech application systems is mainly during distant speech perception or recognition, whenever the speaker far from the microphone and the acquisition of speech is degraded by non-stationary noise and low SNR value.

Several researchers have proposed speech enhancement algorithms to tackle the issue of the noisy speech signal with background noise. However, most of them were capable of overcoming speech quality issues during high SNR and stationary noise rather than speech intelligibility issues in non-stationary noise and low SNR cases (Tseng, 2015; Chen, 2017). Then, supervised Deep Neural Network (DNN)-based mask estimation has been proposed to tackle the speech intelligibility issue to replace the traditional approach. But, a majority of past related works focused on their targeted population (Tseng, 2015; Wang, 2015; Chen, 2017) and excluded the cross-dataset model contains different language utterances due to the limited time of sound recording and the high cost. Thus, it is a good idea to generalize some conditions that have not been explored yet and observed during a training session using the DNN-based mask approach. Otherwise, residual noise is also introduced after speech reconstruction that may cause some speech harmonic losses and hinder oral speech from the point of relevance.

1.4 Research objectives

This research aims to improve the enhancement of speech signal from background noise for single-channel microphone without distorting speech intelligibility using a supervised machine learning algorithm. The specific objectives of this research are as follows:

1. To design an algorithm for enhancement of speech using supervised Deep Neural Network (DNN)-based mask estimation with a new target mask.
2. To propose a post-filtering algorithm that optimizes the speech estimation and reconstruction of residual noise.
3. To evaluate the performance of the proposed algorithms with that of other speech enhancement algorithms.

1.5 Research scopes

The scopes of this research are set out as follows:

1. The single channel-based speech enhancement is the focus of this research, which uses only a single microphone in the experiment. Basically, it is less affected by the room reverberation and spatial sources (Saleem *et al.*, 2019a).
2. Ten types of noise background are considered, namely subway noise, factory noise, train noise, car noise, station bus noise, street noise, restaurant noise, exhibition noise, babble noise, and airport noise to analyze the speech enhancement.
3. Five audio datasets that are used in this research: (1) self-recorded audio from Universiti Tun Hussein Onn Malaysia (UTHM), (2) recorded audio from Universiti Putra Malaysia (UPM), (3) MASS dataset from Universiti Sains Malaysia (USM), (4) TIMIT dataset and (5) IEEE dataset. Noted that IEEE and TIMIT datasets are commonly used among researchers to validate speech enhancement algorithm.
4. A supervised Deep Neural Network (DNN) is used to learn the proposed target mask algorithm, especially for the cross-dataset model and multi-speaker model.
5. MATLAB 2020a software with Intel (R) Core (TM) i5-8250U CPU @ 1.6) GH, 8 Gigabyte RAM is used in this research.

1.6 Research contributions

This research involves designing an improved algorithm in a supervised speech enhancement framework to improve speech quality and speech intelligibility during noise background using a hybrid approach between supervised Deep Neural Network (DNN)-based mask estimation and Harmonic Regeneration Noise Reduction (HRNR). Specifically, the contributions of this research are as follows:

1. An optimize DNN-cWf Masked Estimation is proposed to improve speech intelligibility and quality from noisy backgrounds.
2. An effective post-filtering strategy based on the modified HRNR algorithm was proposed to rectify residual noise.
3. Produce a MALISH model by training the DNN algorithm with cross-speech datasets, which consist of both Malay and English language utterances and different types of single channel microphones.
4. Produce multi-speaker model by training the DNN algorithm with various speakers

1.7 Thesis organization

Chapter 1 describes the background and motivation of the research. The direction of the research such as research problems, objectives, significance, contributions, and scopes are also presented.

Chapter 2 explains the theoretical study on a physiological signal of speech, acoustic environment, and past related works on speech enhancement algorithms. The speech signal characteristics and acoustic environment are briefly explained. Next, state-of-art speech enhancement algorithms and particularly, deep neural network (DNN)-based mask estimation for speech enhancement is presented to find room for improvement for speech intelligibility and quality during noisy conditions.

Chapter 3 elaborates the system overview of the research study. A new framework of Deep Neural Network (DNN)-based mask estimation is also introduced.

A description of theoretical and research methodology is also explained briefly. Several experiments for model validation are also presented.

Chapter 4 presents the results of the experiments presented in Chapter 3. The performance results of the proposed method based on the DNN-based mask estimation and post-filtering process are presented and discussed. The performance evaluation of enhanced speech using a performance metric based on two objective measures is also presented in this chapter.

Chapter 5 concludes this study. Several future works related to this research are also discussed in this chapter. This study concludes with a summary of the main concepts related to this research.

CHAPTER 2

LITERATURE REVIEW

This chapter provides an overview of past related works as background knowledge for this research. A brief discussion on the acoustic theories and speech enhancement that are relevant to this research is presented in Section 2.1 and Section 2.2, respectively. These include the speech physiology and speech enhancement algorithms that were used by researchers to handle non-stationary noise issues and intelligibility issues, particularly in the single-channel microphone. Subsequently, Section 2.3 explains an overview of supervised deep neural network-based mask estimation, in terms of their architectures, challenges, and limitations.

2.1 Introduction

Over the past few decades, speech-enabled mobile application had gained a surge of interest among researchers in the speech processing area. This is due to its wide use in many speech applications. For example, Automatic Speech Recognition (ASR) system has been used in telephony (Garberg & Yudkowsky, 1998), military (Beek *et al.*, 1977), and customer service (Gusler *et al.*, 2005). The ASR system-based approaches are also increasingly demanded recently in rehabilitation helping people suffering from communication disorder such as aphasic patients to do speech therapy and cognitive exercise (Abad *et al.*, 2013; Le *et al.*, 2016; Lee *et al.*, 2016). The hearing aid also is an example of a speech application that was invented to enable inaudible sound for the hearing impaired to be amplified and perceived by ear (Nossier *et al.*,

2019). Although both applications were already used in several commercial applications, there are still challenges to be tackled inaccurate remote context or distant speech recognition (Wölfel & McDonough, 2009) since the ASR system is sensitive to the acoustical environments (Rabiner & Schafer, 2010; Vincent *et al.*, 2017), similar to hearing aid application in clearly perceiving speech sound without any distortion of speech signal caused by background noise.

Basically, to capture human speech signal, a single microphone is practically enough to be used as an input device to acquire audio signal while the computer will be the system to process the captured signal to realize human-computer interaction (HCI) in the ASR system and also in hearing aid applications which involve real-time situation. Even though most researchers proposed several mathematical models and digital signal processing techniques to overcome the issue of noise in these applications, there is still room for improvement to enhance speech signal from noisy background, especially involving non-stationary noise instead of stationary noise and to tackle the issue related to low Signal-to-noise ratio (SNR) (Loizou, 2013; Hurmalainen *et al.*, 2013). For that reason, the following sections will provide a review on speech physiology and noisy speech signal. A review of previous related works for speech enhancement in a noisy environment is discussed in this chapter to discover the research gap. The detailed descriptions of the elements in the Deep Neural Network (DNN)-based mask estimation are also discussed.

2.1.1 Physiology of speech and noisy speech signal

Learning about the physiology and characteristics of speech, as well as the noisy speech signal is required in speech processing. Naturally, a human is more effective than machines at recognizing speech during aural conversations. While the machine-recognized spoken language is transpired by human speech communication between two people as shown in Figure 2.1 (Tunali & Dogruel, 2005). The communication begins when a speaker delivers his speech by formulating the message and then, the listener will try to listen and understand the conversation (Lawrence, 2008). The machine can be very complicated and sensitive to the acoustical environment and

speech variants. To design a much more robust speech-enabled mobile applications, the fundamental knowledge of speech and noise are required.

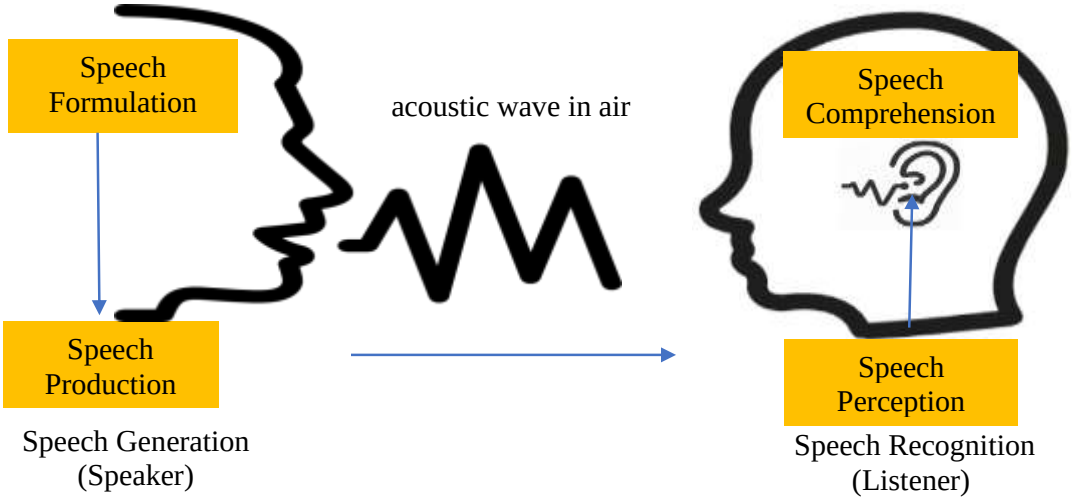
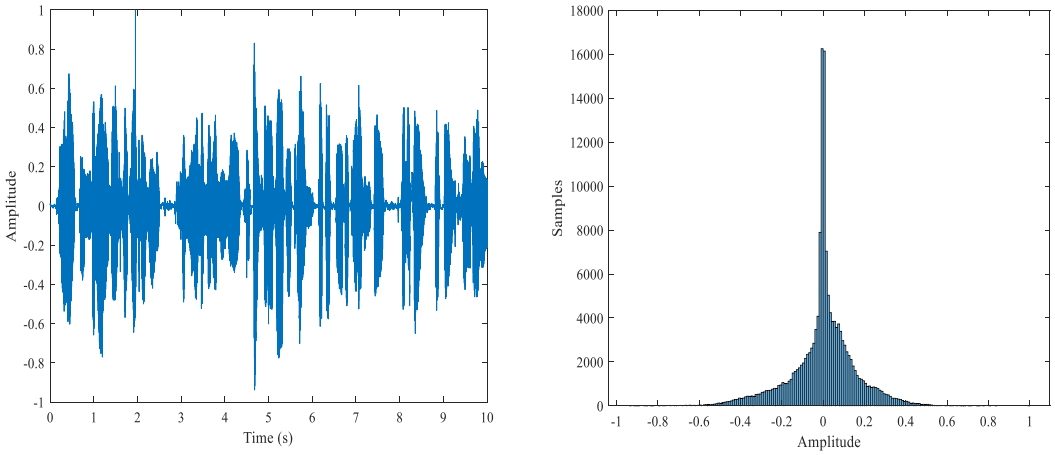
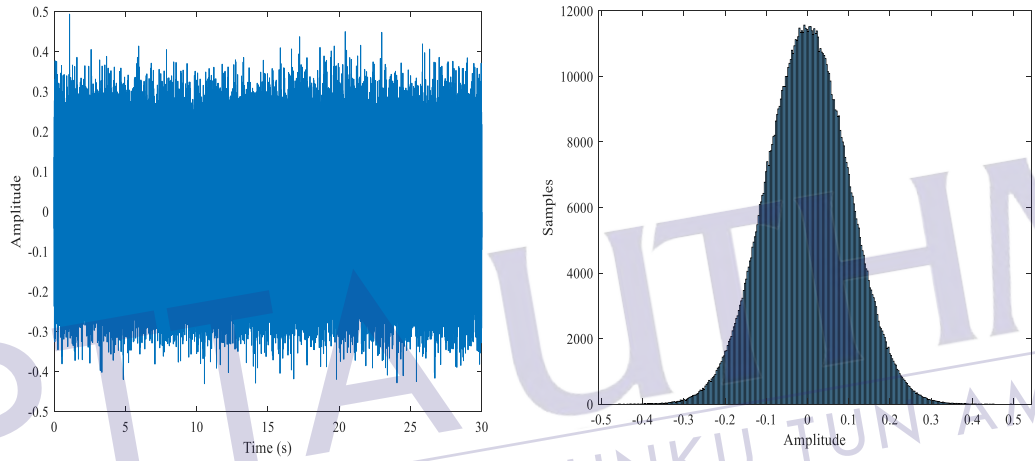


Figure 2.1: Schematic diagram of human speech communication (Tunali & Dogruel, 2005)

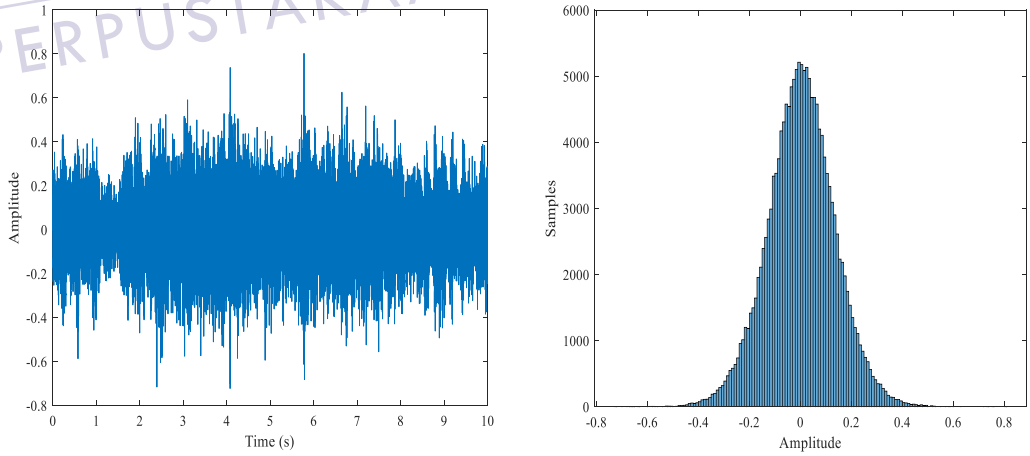
Machine speech recognition has a higher error rate compared to humans (Huang *et al.*, 2001). It showed that the machine is not robust enough to recognize speech tasks such as connected digits, alphabet letters, and spontaneous speech especially during distant speech recognition or remote condition using single microphone. The acquired speech signal by machine could be easily mixed with noise during that condition, which is known as noisy speech signal. Thresholds of noise may vary with frequency and from the environment-to-environment condition. Noise often impacts 3000, 4000, or 6000 Hz thresholds during speech in noisy environments, but not those at 500, 1000, or 2000 Hz (Le Prell & Clavier, 2017). Moreover, noise can distort the spectrum's shape, slope, and spectral dynamic range. However, the frequency positions of the lower formant peaks can be preserved to some extent. Statistically, speech signal or noise signal could be individually visualized based on its distribution in time domain and histogram chart as illustrated in Figure 2.2. In signal processing analysis, the speech signal is highly non-stationary compared to noise signal. Background noises or ambient noises do exist in our daily life in different forms. It was varied to be steady-state noise, noise that modulated with the amplitude envelope of speech, or a competing single talker (Brown & Bacon, 2010).



(a) Highly non-stationary speech



(b) Stationary noise (white noise)



(c) Non-stationary noise (babble noise)

Figure 2.2: Characteristics of speech and noise with their distribution waveform (Toroghi, 2016)

Noises can be categorized into stationary noise and non-stationary noise forms (Wölfel & McDonough, 2009) as illustrated in Figure 2.2 according to their distribution pattern using histogram representation in time series. Stationary noise is represented as a highly normal distribution compared to non-stationary noise. Unlike stationary noise signals, the highly non-stationary speech signal is Laplacian distribution. So, noises from computer fans and white noise are examples of stationary noises, in which the event does not change over time. On the other hand, noises from train and babble sound are examples of non-stationary noise, in which event is keenly changing in time. Thus, noise is difficult to be removed from noisy speech signal and a challenging task due to different environmental characteristics, especially during higher intensity of non-stationary noise (Vincent *et al.*, 2017; Yuan, 2020), which is constantly changing compared to stationary noise (Parchami *et al.*, 2016). So, it can be concluded that speech signals could easily be distorted by noise during distant speech recognition or other speech applications and leads to a challenging speech recognition process faced by machines. Hence, the required speech enhancement is discussed in Section 2.2 in studying speech enhancement algorithms proposed by several researchers to overcome the degraded speech signal issue.

2.1.2 Speech production process

Speech is a mechanism to communicate and express thoughts and feelings in spoken words via articulated sounds, either in the form of an isolated word, continuous sentence, and spontaneous spoken word. It consists of combined lexical and names drawn from very large vocabularies. Each spoken word is created from the phonetic combination of a limited set of vowels and consonant speech sound units known as phonemes (Moore, 2000; Doire, 2016). Mouth is a main organ to generate informative speech signals. The human speech production process initially took place inside the vocal tract extending from the epiglottis to the lips as shown in Figure 2.3.

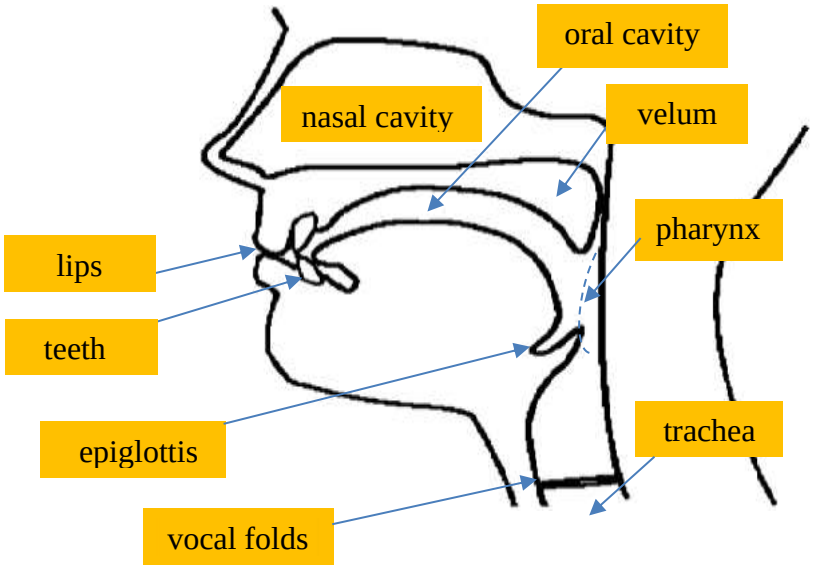


Figure 2.3: Vocal tract anatomical structure (Huang *et al.*, 2001)

Speech is produced by exhaled air from the lungs. The vocal tract is a chamber of an extremely complicated geometrical shape whose dimensions and configuration may continuously vary in time and whose walls consist of tissues with various properties. Thus, the speech sound begins at the epiglottis and ends at the lips. In short, speech wave production can be divided into three stages: sound source generation, articulation by vocal tract, and radiation from the lips or nostrils to produce different letter sounds and phonemes (Lawrence, 2008). Speech sound is determined by the position of articulators such as tongue, teeth, and lips that changes over time. It can be fricatives due to turbulent airflow and plosives due to constriction, then released in the vocal tract. So, characteristic of speech signals is non-stationary signals (Lawrence, 2008).

As illustrated in Figure 2.4, the acoustic waveform for segmented speech signal can be represented as silent, voiced, and unvoiced speech. A silent speech signal represents no speech activity or speech sound. Next, voiced speech normally occurs when vocal folds vibrate at a fundamental frequency or known as pitch, and air freely passes through articulators (Loizou, 2013). This voiced speech waveform is a quasi-periodic sound. For example, vowels are the most prominent instances of voiced speech due to their periodicity and denote high energy when the vocal tract remains relatively open. The vowel sounds are normally dominant at low frequency.

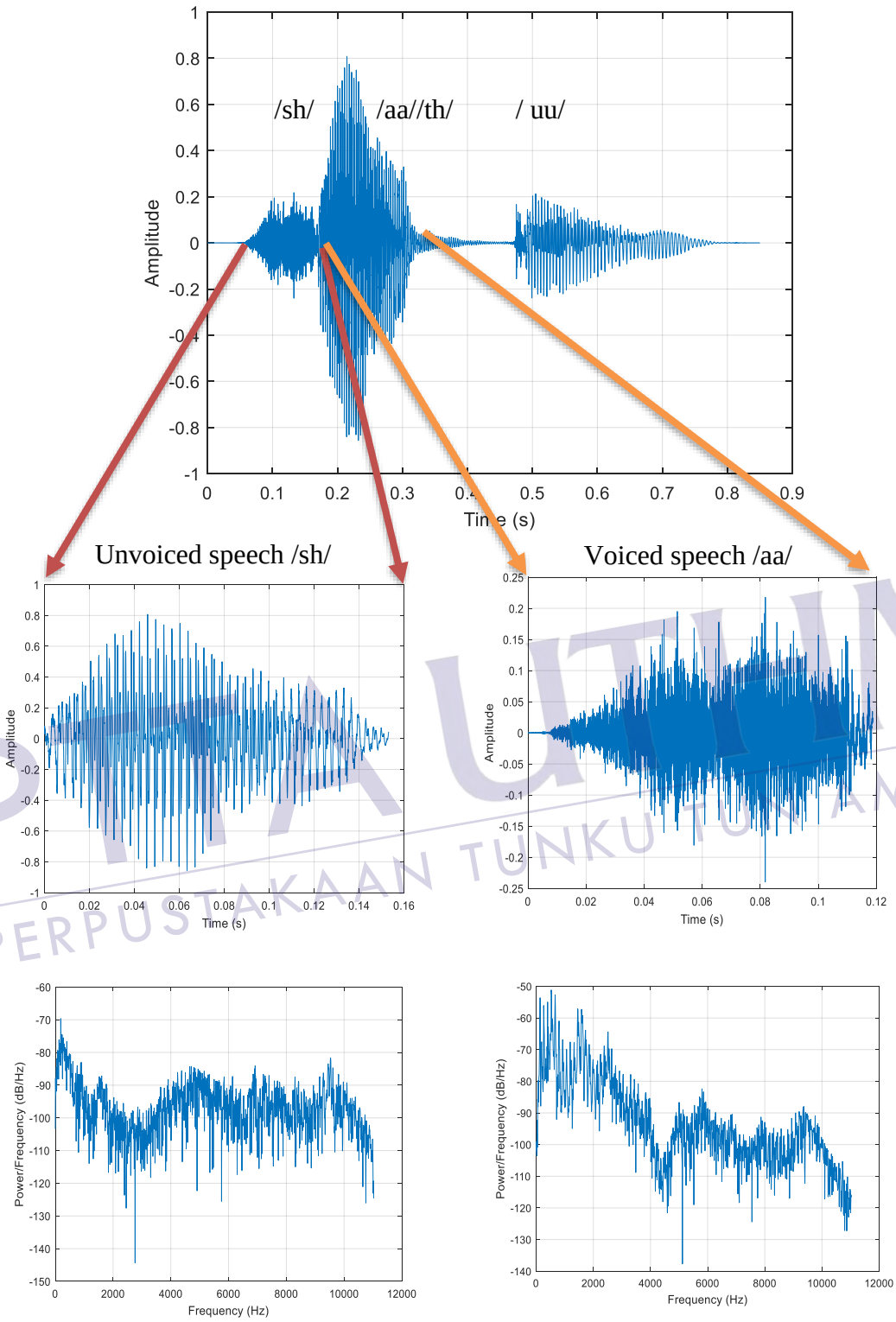


Figure 2.4: Voice and unvoiced speech waveform in time domain and frequency domain for “satu” spoken word

Different vowels are characterised mainly by tongue position on the axes front-back and open-closed, and lip rounding (Juvela, 2015). While several consonants like ‘s’ and ‘t’ sound are examples of unvoiced speech due to aperiodic sound, where air partially passes or is obstructed by one or more places as it passes through the articulators (Loizou, 2013). Fricatives voiceless and voiced, nasal-voiced, stop voiceless and voiced, glide voiced, and liquid voiced are the manner of articulation for consonant sounds (Lawrence, 2008).

2.1.3 Speech perception process

During the speech perception process, ear is the main organ involved in perceiving speech signals during speech conversation. The ear consists of an outer part, middle part, and inner part as depicted in Figure 2.5. The outer ear directs and transmits speech sound waves into the middle ear, which acts as a mechanical transducer. The inner then ear transduces the vibrations transmitted from the middle ear into neural firings or known as cochlea (Irwin, 2006). Noted that the function of the outer ear is to direct and amplify speech sound coming from the environment. While the middle ear is an air-filled cavity connected to the outer ear at the eardrum and to the inner ear at the oval window. In terms of hearing, the main functional part of the middle ear is the transduction system comprised of the ossicular bones. In the inner ear, the functional part relevant for hearing is the cochlea, an organ of 32–35 mm length resembling a snail, coiled in approximately 2.5 turns (Juvela, 2015).

A linearised schematic of the cochlea where the coil has been unwound is depicted in Figure 2.6. The mechanical vibrations enter the cochlea via the oval window, which is connected to stapes in the middle ear. The motion of the oval window creates a traveling wave that reaches its maximum amplitude at a position depending on the frequency so that the maxima near the base of the cochlea correspond to high frequencies and the maxima near the apex to low frequencies. The majority of lower-frequency sounds between 250 and 500 Hz correspond to the first formant of vowel sounds, while most higher frequencies between 2000 and 4000 Hz correspond to the consonant sounds (Lazim *et al.*, 2020). However, some studies showed that the second and third formant of vowel sounds frequency range could reach up to 3000 Hz

(Lawrence, 2008; Monahan & Idsardi, 2010; Viegas et al., 2019). Frequency response for a specific position resembles a bandpass filter response centered at the characteristic frequency (Juvela, 2015).

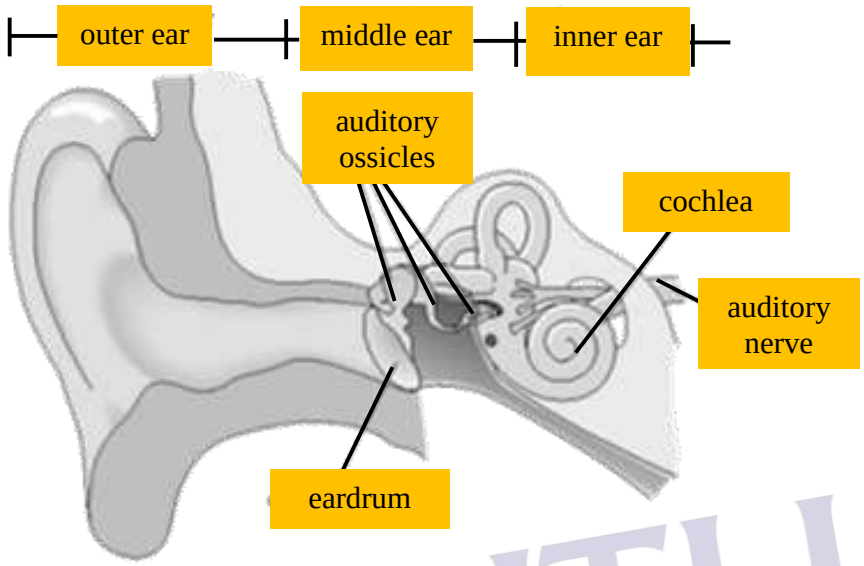


Figure 2.5: Anatomy of the human ear (Irwin, 2006)

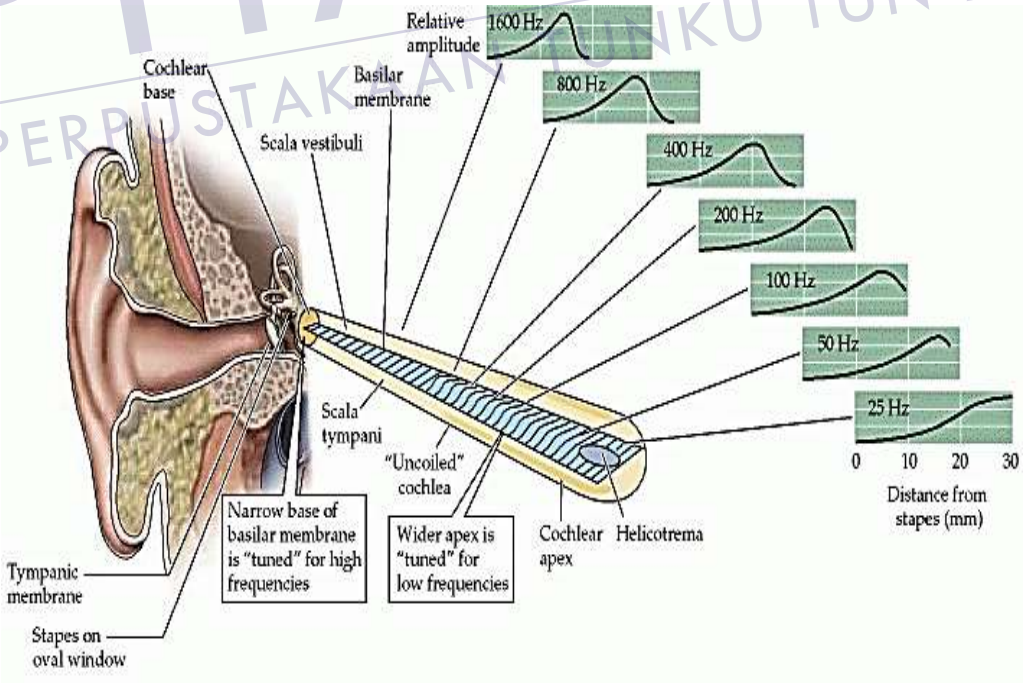


Figure 2.6: Regions of basilar membrane respond to the different frequency (Purves D, 2001)

Thus, studies have shown that human perception of the frequency content of sounds, either for pure tones or speech signals, does not follow a linear scale (Lawrence, 2008). The majority of the speech and speaker recognition systems have used the vector features derived from a filter bank that has been designed based on a non-linear scale according to the auditory system's model (Rabiner & Schafer, 2010). As explained by (Wölfel & McDonough, 2009), human perception can be represented by physical representation for relevant information hence becomes measurable. For example, a pitch can be measured by fundamental frequency, loudness can be measured by sound pressure level or sound intensity in decibel (dB) unit, location can be represented by phase difference and timbre is indicated by spectral shape (Wölfel & McDonough, 2009).

2.2 State-of-art speech enhancement in noisy environments

Figure 2.7 shows an illustration of the speech research area that will be discussed in this section. Hearing aid, speaker recognition, and speech recognition are examples of speech applications. As technology advances, these applications are widely used by humans either as an assistive gadget or communication device to communicate with others. Specifically, speaker and speech recognition are basically used in human-computer interaction while hearing aid is used to assist deaf people. Unfortunately, before these applications can be adopted and robustly utilized with being less sensitive to noise, there are still several unresolved design issues pertaining to this technology that needs to be addressed and rectified. Thus, it can be highlighted that there is emerging research in the field of speech separation to tackle the noise issue, especially related to supervised speech enhancement since speech in noise is still a remaining issue in speech research area. Noted that noise cannot be easily eliminated but it can be rectified and suppressed in many ways. So, speech enhancement algorithms are essential to reduce or suppress the background noise to some degree and at the same time speech signal must be intelligible to ensure speech application achieves higher accuracy in distant speech acquisition.

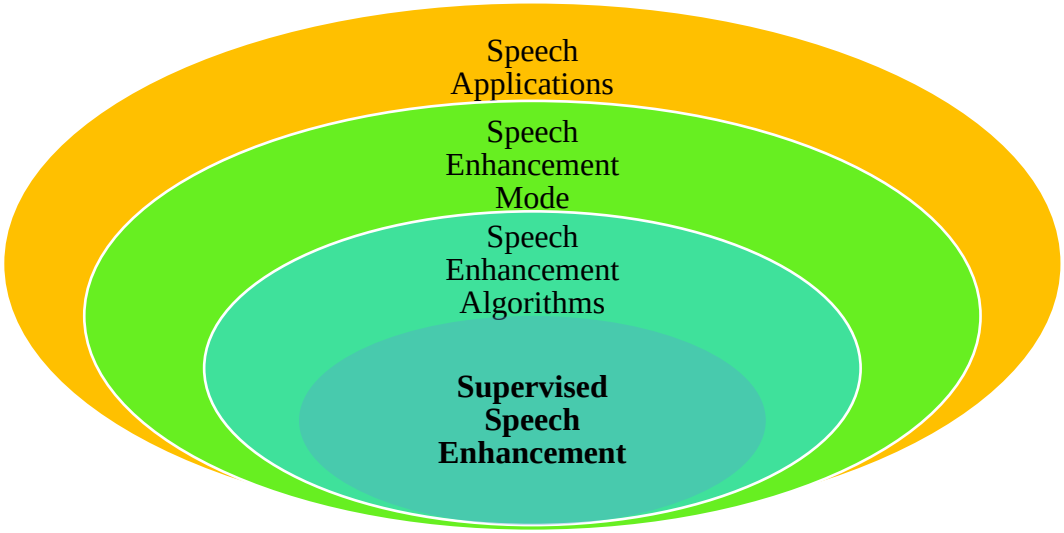


Figure 2.7: Speech research area

Numerous studies have attempted to enhance speech signal from background noise based on either single-channel (monaural) (Paliwal *et al.*, 2010; So & Paliwal, 2011; Mohammadiha *et al.*, 2011; Mohammadiha *et al.*, 2012) or multi-channel (array-based) microphone (Florêncio & Malvar, 2001; Acero *et al.*, 2009; Valin *et al.*, 2004; Kawase *et al.*, 2016; Tesch *et al.*, 2019; Flores *et al.*, 2018). Multiple channel microphones applied a beamforming approach to determine active speakers (Matheja *et al.*, 2013). State-of-the-art speech enhancement mode is discussed in Section 2.2.1 and Section 2.2.2 elaborates several speech enhancement algorithms proposed by a few researchers either to enhance speech signal or to remove noise signal. Finally, past related works on supervised speech enhancement which is superior to other speech enhancement algorithms are also discussed in this section to further understand how it works in increasing the speech quality and intelligibility as well in separating speech signal from background noise.

2.2.1 Speech enhancement modes

Speech enhancement or speech separation is one of the extensive researches in the audio processing research field. Wang & Chen (2018) reported speech separation as a source separation from the noisy background or mixture condition that is essential in speech applications. It can be categorized into two modes according to the number of

microphones used such as a single microphone (monaural) and an array-based microphone (Wang & Chen, 2018; Parchami *et al.*, 2016) as illustrated in Figure 2.8. The number of microphones used basically depends on the development and function of applications. For example, the array-based microphone was applied in humanoid robots for surface enabled estimation of the robot motion (Tourbabin & Rafaely, 2015) and also used in real-time meeting recording for detection and separation of speech events (Asano *et al.*, 2007).

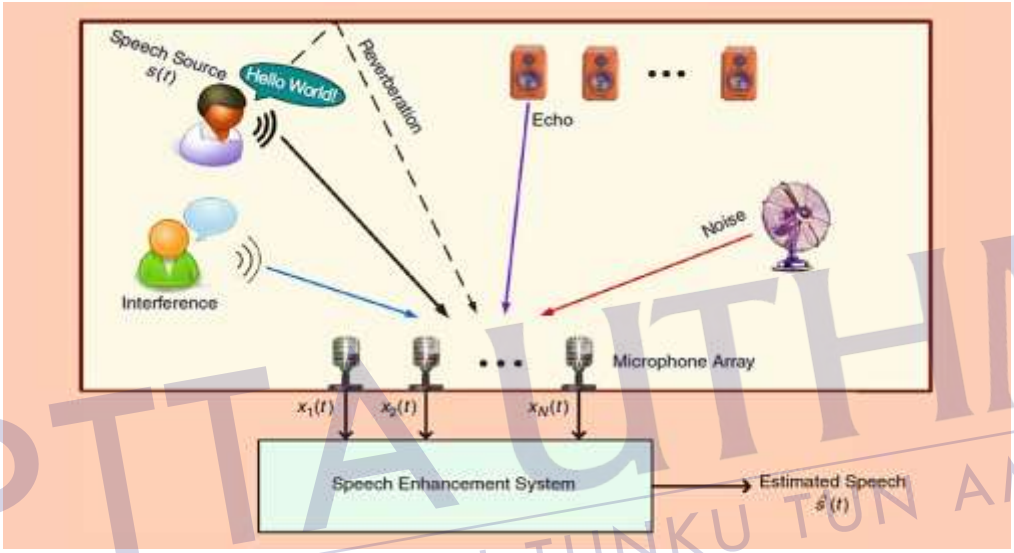


Figure 2.8: Microphone array with speech enhancement system for speech application (Parchami *et al.*, 2016)

While the single-channel microphone is basically applied in mobile phones, hearing aid, automatic speech recognition, and speaker identification for its portability and flexibility (Parchami *et al.*, 2016). A higher number of microphone arrays will lead to the higher complexity of speech enhancement (Nahma, 2018) and cost of the microphone as well, which would not be practical (Ashwini & Kumaraswamy, 2013). Since a single-channel microphone is more ideal to be applied in a lot of speech applications, the improvement of single-channel speech enhancement captured researchers' attention in achieving a reliable and robust speech processing system for adverse conditions (Mowlaee *et al.*, 2012). Moreover, it is still a challenging task when the location of the microphone is far away from the sound source, which is known as a far-field sound acquisition or distant speech recognition (Ashwini & Kumaraswamy,

2013; Bentsen, 2018), but it is less affected by the room reverberation and spatial sources (Saleem *et al.*, 2019a; Bao & Abdulla, 2018b).

As a result, the speech signal is easily degraded by background noises that will lessen the speech quality and intelligibility as well. Thus, high degradation will affect the performance of speech intelligibility for hearing-impaired listeners (Alvarez, 2013) and recognition accuracy in automatic speech recognition systems (Wu & Liu, 2012). Speech quality is more related to the intensity of sound level while speech intelligibility is the comprehensiveness of speech signal or understanding the utterance (Gonzalez, 2013; Kandagatla & Potluri, 2020). Specifically, several researchers tried to simultaneously enhance speech signals in terms of their quality and intelligibility as well as various improvement in speech enhancement algorithms (Kandagatla & Potluri, 2020).

2.2.2 Conventional single-channel speech enhancement

The goal of the speech enhancement algorithm is to remove noise and recover the original signal with lesser distortion and residual noise (Nahma, 2018) as well as to improve speech quality and intelligibility (Alvarez, 2013). Previously, conventional speech enhancement algorithms such as spectral subtraction, Wiener filtering, statistical model-based approach, and subspace algorithm have shown improvement in speech quality but their capability in increasing the intelligibility of speech in the noisy background has remained a challenging task at low SNR (Loizou & Kim, 2010; Alvarez, 2013; Tseng, 2015; Chen, 2017; Kolbæk, 2018). One of the reasons that contributed to the lack of intelligibility improvement in single microphone speech enhancement is statistically assuming the noise and speech signal are normal distribution in noisy speech signal that may lead to lack of speech or noise estimation (Kim & Loizou, 2011; Loizou & Kim, 2010). Then, speech distortions introduced by the frequency-specific gain functions could at times be more damaging than the background noise itself, and removal of low-intensity speech sounds such as unvoiced consonants may hamper the oral speech (Kim & Loizou, 2011). Moreover, the main drawback of the conventional approach is the introduction of spectral artifacts (Shoba & Rajavel, 2020). Thus, the conventional speech enhancement algorithms did not

benefit from non-stationary noise and low signal-to-noise ratios (SNRs) conditions (Tseng, 2015; Chen, 2017).

The spectral subtraction algorithm is the earliest speech enhancement algorithm (Pardede *et al.*, 2019), widely used in numerous speech applications due to its simplicity. The general idea of spectral subtraction is simply by subtracting the noise power spectrum, $\left|\widehat{N}(\omega)\right|^2$from the mixture power spectrum, $|Y(\omega)|^2$ to get the estimated speech signal, $\left|\hat{X}(\omega)\right|^2$ when there is no correlation between speech and noise is assumed since the noise is additive and stationary as shown in Equation 2.1 (Upadhyay & Karmakar, 2015). The spectral subtractive method is often formulated in the power of the Short-Time Fourier transform (STFT) domain rather than in the amplitude domain (Parchami *et al.*, 2016). However, it can easily fail when the noise is highly non-stationary (Srinivasan *et al.*, 2006). Another problem with spectral subtraction is that the resulting speech spectrum after subtraction could be negative. In short, the spectral subtraction approach suffered some high residual noise when underestimating noise power (Ingale & Nalbalwar, 2019; Yao *et al.*, 2016) and loss of useful information when overestimating noise power. Due to these constraints, extended spectral subtraction was proposed (Borský, 2016).

$$\left|\hat{X}(\omega)\right|^2 = |Y(\omega)|^2 - \left|\widehat{N}(\omega)\right|^2 \tag{2.1}$$

Next, Wiener filtering is the optimal complex filter in the STFT domain since the spectral subtraction algorithm is difficult to claim its optimality (Parchami *et al.*, 2016). Several alternative methods such as square-root and parametric Wiener filtering have been proposed are shown in Equation 2.2 until Equation 2.3 for calculating the Wiener filter gain function, $\widehat{W}(\omega)$ from the noisy speech signal. The $P_x(\omega)$ denotes as power spectrum of noise free signal while $P_n(\omega)$ denotes as power spectrum of noise signal.

Simplified square root:

$$\widehat{W}(\omega) = P_x(\omega) \tag{2.2}$$

Parametric:

$$\widehat{W}(\omega) = \left(\frac{P_x(\omega)}{P_x(\omega) + \alpha P_n(\omega)}\right)^{\beta} \tag{2.3}$$

The concept of ratio masking in supervised speech separation is very similar to parametric Wiener filter when β and α parameters are set to 0.5 and 1, respectively (Wang, 2015). The use of parameter β allows the concession between noise reduction and speech distortion. Although the residual noise might considerably be reduced by increasing β, it is likely resulted in speech distortion too (Loizou, 2013). Another popular speech enhancement algorithm is a statistical model-based approach (Loizou, 2013), which implements statistical distributions of speech and noise based on maximum likelihood estimator (Kuklasiński *et al.*, 2016), minimum mean square error (MMSE) (Momeni *et al.*, 2016) and maximum a posteriori estimator (MAP) (Su *et al.*, 2013). Generally, MAP and MMSE are also known as Bayesian estimators. The disadvantage of this algorithm is related to computational complexity in estimate noise distributions (Kawamura *et al.*, 2012).

Basically, statistical speech enhancement performs estimation based on the speech distribution conditioned on noise observation (Loizou, 2013). Like Wiener filtering, statistical speech enhancement typically relies on the precise estimation of speech or noise variances, which is a challenging task for non-stationary noises. Otherwise, a statistical approach is often difficult to deal with nonstationary noise in an unknown real environment (Liang *et al.*, 2020). Hence, these approaches usually improve quality but not intelligibility (Li, 2016). While subspace algorithm applied linear Algebra theory which is best performed in most noise condition but more complicated to perform noise suppression. Particularly, single value decomposition is applied in subspace algorithm which is widely used in speech recognition application and image processing (Loizou, 2013).

In short, spectral subtractive algorithms and statistical model-based algorithms are suitable for improving speech quality but not for intelligibility (Loizou, 2013). On the other hand, feature enhancement is applied for robust speech recognition applications. The feature-based approach includes enhancement of extracted features such as using enhanced Mel Frequency Cepstrum Coefficient (MFCC) (Ittichaichareon *et al.*, 2012), Least Mean Square (LMS) filter and cochleagram (Dou *et al.*, 2019) and Power Normalized Cepstral Coefficient (PNCC) (Chang, 2016; Kim & Stern, 2016) are widely proposed by researchers. Most researchers modified the auditory features to suppress noises (Tamazin *et al.*, 2019). However, several researchers argued on the feature-based approach due to the selection of which type of auditory information is important for robust speech recognition (Li *et al.*, 2014).

Otherwise, its performance is not always producing good results on speech recognition accuracy, especially at low SNR (Narayanan & Wang, 2014).

2.2.3 Machine learning-based speech enhancement

To alleviate the limitations in conventional speech enhancement algorithms, supervised speech enhancement has been recently proposed by several researchers due to numerous applications applying machine learning algorithms to operate them as an automated system (Wang & Chen, 2018) which is known as a data-driven approach. This supervised speech enhancement was inspired by time-frequency (T-F) masking in Computational Auditory Scene Analysis (CASA) (Wang & Chen, 2018), which determines either speech dominant or noise dominant in each T-F masking frame (Wang & Chen, 2018) without statistical assumptions (Kolbæk *et al.*, 2018). This recent supervised-based mask estimation caught the attention of several researchers after vast development in the computing system since conventional supervised speech enhancement suffered limited memory and time training. The conventional supervised speech enhancement used fewer hidden nodes in Multilayer Perceptron (MLP) algorithm to predict a short window (Tamura, 1989) and log power spectra (Xie & Van Compernolle, 1994) of the clean speech signal from mixture signals. Otherwise, the performance of speech enhancement using more hidden layers and nodes without the mask is still not promising (Fu *et al.*, 2017).

The main advantage of CASA is that this method incorporates the auditory perception mechanism without assuming any properties or models of the noise (Lang & Yang, 2020a). Due to the advantages of CASA in numerous speech applications (Wang & Brown, 2006), several machine learning algorithms had been proposed to be operated with CASA. For example, Kim *et al.* (2009) proposed a Bayesian classifier based on Gaussian mixture models (GMM) in a speaker and masker-dependent way with amplitude modulation spectrum (AMS) features, and the performance of speech intelligibility is evaluated with normal-hearing (NH) listener. Next, (Chang *et al.*, 2008; Han & Wang, 2011) used the Support Vector Machine (SVM) classifier in the speech enhancement. Han & Wang (2011) classified the T-F units of the noise-masked signal into two classes: target-dominated and masker-dominated. The individual T-F

REFERENCES

Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., Farrajota, L. & Martins, I. P. (2013). Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech & Language,* 27**,** 1235-1248.

Abdullah, S., Zamani, M. & Demosthenous, A. (2021). Towards more efficient DNN-based speech enhancement using quantized correlation mask. *IEEE Access,* 9**,** 24350-24362.

Acero, A., Tashev, I. J. & Seltzer, M. L. (2009). *Spatial Noise Suppression for a Microphone Array*. U.S. Patent. 7,565,288.

Aggarwal, R. K. & Dave, M. (2011). Acoustic modeling problem for automatic speech recognition system: conventional methods (Part I). *International Journal of Speech Technology,* 14**,** 297.

Al-Haddad, S., Samad, S., Hussain, A. & Ishak, K. (2008). Isolated Malay digit recognition using pattern recognition fusion of dynamic time warping and hidden Markov models. *American Journal of Applied Sciences,* 5**,** 714-720.

Alvarez, D. A. (2013). *Speech Enhancement Algorithms for Audiological Applications.* University of Alcalá Ph.D. Thesis.

Anusuya, M. & Katti, S. (2011). Front end analysis of speech recognition: a review. *International Journal of Speech Technology,* 14**,** 99-145.

Asano, F., Yamamoto, K., Ogata, J., Yamada, M. & Nakamura, M. (2007). Detection and separation of speech events in meeting recordings using a microphone

array. *EURASIP Journal on Audio, Speech, and Music Processing,* 2007**,** 027616.

Ashwini, J. K. & Kumaraswamy, R. (2013). Single-Channel Speech Enhancement Techniques for Distant Speech Recognition. *Journal of Intelligent Systems,* 22**,** 81-93.

Baby, D. & Verhulst, S. (2019). Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 106-110.

Bao, F. & Abdulla, W. H. (2018a). A new ratio mask representation for CASA-based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 27**,** 7-19.

Bao, F. & Abdulla, W. H. (2018b). Noise masking method based on an effective ratio mask estimation in Gammatone channels. *APSIPA Transactions on Signal and Information Processing,* 7.

Beek, B., Neuberg, E. & Hodge, D. (1977). An assessment of the technology of automatic speech recognition for military applications. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 25**,** 310-322.

Bentsen, T. (2018). *Computational Speech Segregation Inspired by Principles of Auditory Processing.* Technical University of Denmark, Ph.D Thesis.

Boldt, J. B. & Ellis, D. P. (2009). A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. *2009 17th European Signal Processing Conference*. 1849-1853.

Borský, M. (2016). *Robust Recognition of Strongly Distorted Speech.* Czech Technical University Ph.D. Thesis.

Brown, C. A. & Bacon, S. P. (2010). Fundamental frequency and speech intelligibility in background noise. *Hearing research,* 266**,** 52-59.

Brownlee, J. (2016). Machine learning mastery with python. *Machine Learning Mastery Pty Ltd,* 527**,** 100-120.

Chang, J.-H., Jo, Q.-H., Kim, D. K. & Kim, N. S. (2008). Global soft decision employing support vector machine for speech enhancement. *IEEE Signal Processing Letters,* 16**,** 57-60.

Chang, S.-Y. (2016). *Feature Design for Robust Speech Recognition: Nurture and Nature.* University of California, Berkeley, Ph.D. Thesis.

Chari, N., Herman, G. & Danhauer, J. L. (1977). Perception of one-third octave-band filtered speech. *The Journal of the Acoustical Society of America,* 61**,** 576-580.

Chen, J. (2017). *On Generalization of Supervised Speech Separation.* The Ohio State University, Ph.D. Thesis.

Chen, J. & Wang, D.(2018). Dnn based mask estimation for supervised speech separation. *Audio source separation.* Springer.

Chen, J., Wang, Y. & Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 22**,** 1993-2002.

Delikaris-Manias, S. & Pulkki, V. (2013). Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing,* 21**,** 2356-2367.

Dendani, B., Bahi, H. & Sari, T. (2020). Speech Enhancement Based on Deep AutoEncoder for Remote Arabic Speech Recognition. *International Conference on Image and Signal Processing*. 221-229.

Doire, C. (2016). *Single-channel Enhancement of Speech Corrupted by Reverberation and Noise.* Imperial College London, Ph.D. Thesis.

Dou, W., Wang, H. & Yang, R. (2019). Cochleagram-based identification of electronic disguised voice with pitch scaling in the noisy environment. *Proc. of the ACM Turing Celebration Conference*, China. 1-8.

Erdogan, H., Hershey, J. R., Watanabe, S. & Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 708-712.

Fadhilah, R. (2016). *Fuzzy Petri Nets as a Classification Method for Automatic Speech intelligibility Detection of Children with Speech Impairments.* University of Malaya, Ph.D. Thesis.

Florêncio, D. A. & Malvar, H. S. (2001). Multichannel filtering for optimum noise reduction in microphone arrays. *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. 197-200.

Flores, C. G., Tryfou, G. & Omologo, M. (2018). Cepstral distance based channel selection for distant speech recognition. *Computer Speech & Language,* 47**,** 314-332.

Fu, S.-W., Tsao, Y., Lu, X. & Kawai, H. (2017). Raw waveform-based speech enhancement by fully convolutional networks. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 006-012.

Gao, B., Woo, W. L. & Khor, L. (2014). Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic

sparsity adaptation. *The Journal of the Acoustical Society of America,* 135**,** 1171-1185.

Garberg, R. B. & Yudkowsky, M. (1998). *Method for automatic speech recognition in telephony*. U.S. Patent. 5,822,727.

Géron, A.(2019). Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow:. *Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media.

Giannakopoulos, T. (2009). A method for silence removal and segmentation of speech signals, implemented in Matlab. *University of Athens, Athens,* 2.

Gonzalez, S. (2013). *Analysis of Very Low Quality Speech for Mask-based Enhancement.* Imperial College London Ph.D. Thesis.

Grezes, F., Ni, Z., Trinh, V. A. & Mandel, M. (2020). Enhancement of Spatial Clustering-Based Time-Frequency Masks using LSTM Neural Networks. *arXiv preprint arXiv:2012.01576.*

Gusler, C. P., Hamilton, I. R. A. & Waters, T. M. (2005). *Employing speech recognition and capturing customer speech to improve customer service*. US Patent. 6,915,246.

Han, K. & Wang, D. (2011). An SVM based classification approach to speech separation. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4632-4635.

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE transactions on speech and audio processing,* 2**,** 578-589.

Hu, Y. (2007). Subjective evaluation and comparison of speech enhancement algorithms. *Speech Communication,* 49**,** 588-601.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M. & Smaragdis, P. (2014a). Deep learning for monaural speech separation. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1562-1566.

Huang, X., Acero, A., Hon, H.-W. & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice hall PTR.

Huang, X., Baker, J. & Reddy, R. (2014b). A historical perspective of speech recognition. *Communications of the ACM*, 57**,** 94-103.

Huimin, Z., Xupeng, J. & Dongmei, L. (2019). An Iterative Post-processing Approach for Speech Enhancement. *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*. 130-134.

Hurmalainen, A., Gemmeke, J. F. & Virtanen, T. (2013). Modelling non-stationary noise with spectral factorisation in automatic speech recognition. *Computer Speech & Language*, 27**,** 763-779.

Ingale, P. P. & Nalbalwar, S. L. (2019). Deep neural network based speech enhancement using mono channel mask. *International Journal of Speech Technology*, 22**,** 841-850.

Irwin, J. (2006). Basic anatomy and physiology of the ear. *Infection and hearing impairment***,** 8-13.

Ittichaichareon, C., Suksri, S. & Yingthawornsuk, T. (2012). Speech recognition using MFCC. *International Conference on Computer Graphics, Simulation and Modeling*. 135-138.

Juvela, L. (2015). *Perceptual spectral matching utilizing mel-scale filterbanks for statistical parametric speech synthesis with glottal excitation vocoder*. Aalto University, Master's Thesis.

Kandagatla, R. K. & Potluri, V. S. (2020). Performance analysis of neural network, NMF and statistical approaches for speech enhancement. *International Journal of Speech Technology***,** 1-21.

Kawamura, A., Thanhikam, W. & Iiguni, Y. (2012). Single channel speech enhancement techniques in spectral domain. *ISRN Mechanical Engineering,* 2012.

Kawase, T., Niwa, K., Fujimoto, M., Kamado, N., Kobayashi, K., Araki, S. & Nakatani, T. (2016). Real-time integration of statistical model-based speech enhancement with unsupervised noise PSD estimation using microphone array. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 604-608.

Keshavarzi, M., Goehring, T., Turner, R. E. & Moore, B. C. (2019). Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction. *The Journal of the Acoustical Society of America,* 145**,** 1493-1503.

Kim, C. & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing,* 24**,** 1315-1329.

Kim, G. & Loizou, P. C. (2010). Improving speech intelligibility in noise using environment-optimized algorithms. *IEEE transactions on audio, speech, and language processing,* 18**,** 2080-2090.

Kim, G. & Loizou, P. C. (2011). Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms. *The Journal of the Acoustical Society of America,* 130**,** 1581-1596.

Kim, G., Lu, Y., Hu, Y. & Loizou, P. C. (2009). An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America,* 126**,** 1486-1494.

Kolbæk, M. (2018). *Single-Microphone Speech Enhancement and Separation Using Deep Learning.* Aalborg University, Denmark, Ph.D. Thesis.

Kolbæk, M., Tan, Z.-H. & Jensen, J. (2016). Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 25**,** 153-167.

Kolbæk, M., Tan, Z.-H. & Jensen, J. (2018). Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5059-5063.

Kuklasiński, A., Doclo, S., Jensen, S. H. & Jensen, J. (2016). Maximum likelihood PSD estimation for speech enhancement in reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 24**,** 1599-1612.

Lang, H. & Yang, J. (2020a). Learning Ratio Mask with Cascaded Deep Neural Networks for Echo Cancellation in Laser Monitoring Signals. *Electronics,* 9**,** 856.

Lang, H. & Yang, J. (2020b). Speech Enhancement Based on Fusion of Both Magnitude/Phase-Aware Features and Targets. *Electronics,* 9**,** 1125.

Lawrence, R. (2008). *Fundamentals of speech recognition*, Pearson Education India.

Lazim, R. Y., Yun, Z. & Wu, X. (2020). Improving Speech Quality for Hearing Aid Applications Based on Wiener Filter and Composite of Deep Denoising Autoencoders. *Signals,* 1**,** 138-156.

Le, D., Licata, K., Persad, C. & Provost, E. M. (2016). Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE/ACM transactions on audio, speech, and language processing,* 24**,** 2187-2199.

Le Prell, C. G. & Clavier, O. H. (2017). Effects of noise on speech recognition: Challenges for communication by service members. *Hearing research,* 349**,** 76-89.

Lee, T., Liu, Y., Huang, P.-W., Chien, J.-T., Lam, W. K., Yeung, Y. T., Law, T. K., Lee, K. Y., Kong, A. P.-H. & Law, S.-P. (2016). Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6475-6479.

Li, A., Peng, R., Zheng, C. & Li, X. (2020). A Supervised Speech Enhancement Approach with Residual Noise Control for Voice Communication. *Applied Sciences,* 10**,** 2894.

Li, D. (2016). *Deep Neural Network Approach for Single Channel Speech Enhancement Processing.* University of Ottawa, Master's Thesis.

Li, J., Deng, L., Gong, Y. & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 22**,** 745-777.

Li, R., Sun, X., Liu, Y., Yang, D. & Dong, L. (2019). Multi-resolution auditory cepstral coefficient and adaptive mask for speech enhancement with deep neural network. *EURASIP Journal on Advances in Signal Processing,* 2019**,** 22.

Li, X. & Horaud, R. (2020). Online monaural speech enhancement using delayed subband lstm. *arXiv preprint arXiv:2005.05037*.

Li, X., Li, J. & Yan, Y. (2017). Ideal Ratio Mask Estimation Using Deep Neural Networks for Monaural Speech Segregation in Noisy Reverberant Conditions. *Interspeech*. 1203-1207.

Li, Z. & Gao, Y. (2016). Acoustic feature extraction method for robust speaker identification. *Multimedia Tools and Applications,* 75**,** 7391-7406.

Liang, R., Kong, F., Xie, Y., Tang, G. & Cheng, J. (2020). Real-Time Speech Enhancement Algorithm Based on Attention LSTM. *IEEE Access,* 8**,** 48464-48476.

Loizou, P. C. (2013). *Speech enhancement: theory and practice*, CRC press.

Loizou, P. C. & Kim, G. (2010). Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE transactions on audio, speech, and language processing,* 19**,** 47-56.

Matheja, T., Buck, M. & Fingscheidt, T. (2013). A dynamic multi-channel speech enhancement system for distributed microphones in a car environment. *EURASIP Journal on Advances in Signal Processing,* 2013**,** 191.

Melve, O. K. (2016). *Speech Enhancement with Deep Neural Networks.* Norwegian University of Science and Technology, Master's Thesis.

Michelsanti, D., Tan, Z.-H., Zhang, S.-X., Xu, Y., Yu, M., Yu, D. & Jensen, J. (2020). An overview of deep-learning-based audio-visual speech enhancement and separation. *arXiv preprint arXiv:2008.09586*.

Miller, C. W., Bentler, R. A., Wu, Y.-H., Lewis, J. & Tremblay, K. (2017). Output signal-to-noise ratio and speech perception in noise: effects of algorithm. *International journal of audiology,* 56**,** 568-579.

Mitra, S. K. & Kuo, Y. (2006). *Digital signal processing: a computer-based approach*, McGraw-Hill New York.

Mohammadiha, N., Gerkmann, T. & Leijon, A. (2011). A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization. *2011 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*. 45-48.

Mohammadiha, N., Taghia, J. & Leijon, A. (2012). Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4561-4564.

Momeni, H., Abutalebi, H. R. & Habets, E. A. (2016). Conditional MMSE-based single-channel speech enhancement using inter-frame and inter-band correlations. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5215-5219.

Moore, D. R. (2000). Auditory neuroscience: Is speech special? *Current Biology*, 10**,** R362-R364.

Mowlaee, P., Saeidi, R., Christensen, M. G., Tan, Z.-H., Kinnunen, T., Franti, P. & Jensen, S. H. (2012). A joint approach for single-channel speaker identification and speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20**,** 2586-2601.

Nahma, L. (2018). *A Study into Speech Enhancement Techniques in Adverse Environment.* Curtin University, Ph.D. Thesis.

Narayanan, A. & Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7092-7096.

Narayanan, A. & Wang, D. (2014). Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 22**,** 826-835.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M. & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE Access,* 7**,** 19143-19165.

Neumeyer, L. & Weintraub, M. (1994). Microphone-independent robust signal processing using probabilistic optimum filtering. *Human Language Technology: Proceedings of a Workshop*, Plainsboro, New Jersey.

Nossier, S. A., Rizk, M., Moussa, N. D. & El Shehaby, S. (2019). Enhanced smart hearing aid using deep neural networks. *Alexandria Engineering Journal,* 58**,** 539-550.

O'shaughnessy, D. (2013). Acoustic analysis for automatic speech recognition. *Proceedings of the IEEE,* 101**,** 1038-1053.

Odelowo, B. (2018). *Development of a Neural Network-based Speech Enhancement System.* Georgia Institute of Technology, Ph.D Thesis.

Odelowo, B. O. & Anderson, D. V. (2017). A Mask-Based Post Processing Approach for Improving the Quality and Intelligibility of Deep Neural Network Enhanced Speech. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 1134-1138.

Paliwal, K., Wójcicki, K. & Schwerin, B. (2010). Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech communication,* 52**,** 450-475.

Pandey, A. & Wang, D. (2019). A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 27**,** 1179-1188.

Pandey, A. & Wang, D. (2020). On cross-corpus generalization of deep learning based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 28**,** 2489-2499.

Parchami, M., Zhu, W.-P., Champagne, B. & Plourde, E. (2016). Recent developments in speech enhancement in the short-time Fourier transform domain. *IEEE Circuits and Systems Magazine,* 16**,** 45-77.

Pardede, H., Ramli, K., Suryanto, Y., Hayati, N. & Presekal, A. (2019). Speech Enhancement for Secure Communication Using Coupled Spectral Subtraction and Wiener Filter. *Electronics*, 8**,** 897.

Plapous, C., Marro, C. & Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing,* 14**,** 2098-2108.

Purves D, A. G., Fitzpatrick D,. (2001). Neuroscience. 2nd edition. The Inner Ear. National Center for Biotechnology Information (NCBI).

Rabiner, L. & Schafer, R. (2010). *Theory and applications of digital speech processing*, Prentice Hall Press.

Rao, K. S. & Vuppala, A. K. (2014). *Speech processing in mobile environments*, Springer.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., De Berker, A. & Ganguli, S. (2019). A deep learning framework for neuroscience. *Nature neuroscience,* 22**,** 1761-1770.

Routray, S. & Mao, Q. (2021). Phase sensitive masking-based single channel speech enhancement using conditional generative adversarial network. *Computer Speech & Language***,** 101270.

Russo, M., Stella, M., Sikora, M. & Pekić, V. (2019). Robust cochlear-model-based speech recognition. *Computers,* 8**,** 5.

Saleem, N., Irfan Khattak, M., Ali, M. Y. & Shafi, M. (2019a). Deep neural network for supervised single-channel speech enhancement. *Archives of Acoustics,* 44.

Saleem, N. & Khattak, M. I. (2019). A review of supervised learning algorithms for single channel speech enhancement. *International Journal of Speech Technology,* 22**,** 1051-1075.

Saleem, N., Khattak, M. I. & Perez, E. (2019b). Spectral Phase Estimation Based on Deep Neural Networks for Single Channel Speech Enhancement. *Journal of Communications Technology and Electronics,* 64**,** 1372-1382.

Samui, S., Chakrabarti, I. & Ghosh, S. K. (2019). Time–frequency masking based supervised speech enhancement framework using fuzzy deep belief network. *Applied Soft Computing,* 74**,** 583-602.

Seltzer, M. L., Yu, D. & Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. *2013 IEEE international conference on acoustics, speech and signal processing*. 7398-7402.

Sheela, K. G. & Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in neural networks. *Mathematical Problems in Engineering*, 2013.

Shen, T. W. & Lun, D. P. (2017). A speech enhancement method based on sparse reconstruction on log-spectra. *HKIE Transactions,* 24**,** 24-34.

Shoba, S. & Rajavel, R. (2020). A new Genetic Algorithm based fusion scheme in monaural CASA system to improve the performance of the speech. *Journal of Ambient Intelligence and Humanized Computing,* 11**,** 433-446.

So, S. & Paliwal, K. K. (2011). Modulation-domain Kalman filtering for single-channel speech enhancement. *Speech Communication,* 53**,** 818-829.

Soni, M. H., Shah, N. & Patil, H. A. (2018). Time-frequency masking-based speech enhancement using generative adversarial network. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5039-5043.

Srinivasan, S., Roman, N. & Wang, D. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication,* 48**,** 1486-1501.

Strake, M., Defraene, B., Fluyt, K., Tirry, W. & Fingscheidt, T. (2019). Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages. *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 239-243.

Su, Y.-C., Tsao, Y., Wu, J.-E. & Jean, F.-R. (2013). Speech enhancement using generalized maximum a posteriori spectral amplitude estimator. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7467-7471.

Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing,* 19**,** 2125-2136.

Tamazin, M., Gouda, A. & Khedr, M. (2019). Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients. *Applied Sciences,* 9**,** 2166.

Tamura, S. (1989). An analysis of a noise reduction neural network. *International Conference on Acoustics, Speech, and Signal Processing*. 2001-2004.

Tan, K. & Wang, D. (2018). A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. *Interspeech*. 3229-3233.

Tan, T.-P., Xiao, X., Tang, E. K., Chng, E. S. & Li, H. (2009). MASS: A Malay language LVCSR corpus resource. *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*. 25-30.

Tchorz, J. & Kollmeier, B. (2003). SNR estimation based on amplitude modulation analysis with applications to noise suppression. *IEEE Transactions on Speech and Audio Processing,* 11**,** 184-192.

Tesch, K., Rehr, R. & Gerkmann, T. (2019). On Nonlinear Spatial Filtering in Multichannel Speech Enhancement. *INTERSPEECH*. 91-95.

Toroghi, R. M. (2016). *Blind Speech Separation in Distant Speech Recognition Front-end Processing.* Saarland University, Saarbrücken, Germany, Doctor of Engineering.

Tourbabin, V. & Rafaely, B. (2015). Direction of arrival estimation using microphone array processing for moving humanoid robots. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 23**,** 2046-2058.

Tseng, H.-W. (2015). *A Combined Statistical and Machine Learning Approach for Single Channel Speech Enhancement.* University of Minnesota Ph.D. Thesis.

Tseng, H.-W., Hong, M. & Luo, Z.-Q. (2015). Combining sparse NMF with deep neural network: A new classification-based approach for speech enhancement. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2145-2149.

Tunali, V. & Dogruel, M. (2005). A speaker dependent, large vocabulary, isolated word speech recognition system for turkish. *Marmara Üniversity, Thesis for Degree of Master of Science*.

Une, M. & Miyazaki, R. (2020). Musical-noise-free noise reduction by using biased harmonic regeneration and considering relationship between a priori SNR and sound quality. *Applied Acoustics,* 168**,** 107410.

Upadhyay, N. & Karmakar, A. (2015). Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study. *Procedia Computer Science,* 54**,** 574-584.

Valin, J.-M., Rouat, J. & Michaud, F. (2004). Enhanced robot audition based on microphone array source separation with post-filter. *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*. 2123-2128.

Vestur, C.(2017). Building a performing Machine Learning model from A to Z.

Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J. & Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language,* 46**,** 535-557.

Wang, D. & Brown, G. J. (2006). *Fundamentals of computational auditory scene analysis*, Wiley-IEEE Press.

Wang, D. & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 26**,** 1702-1726.

Wang, L., Hon, T.-K., Reiss, J. D. & Cavallaro, A. (2016). An iterative approach to source counting and localization using two distant microphones. *IEEE/ACM transactions on audio, speech, and language processing,* 24**,** 1079-1093.

Wang, P. & Tan, K. (2019). Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 28**,** 39-48.

Wang, Y. (2015). *Supervised Speech Separation using Deep Neural Networks.* The Ohio State University Ph.D. Thesis.

Wang, Y., Han, K. & Wang, D. (2012). Exploring monaural features for classification-based speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing,* 21**,** 270-279.

Wang, Y., Narayanan, A. & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing,* 22**,** 1849-1858.

Wang, Y. & Wang, D. (2013). Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing,* 21**,** 1381-1390.

Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R. & Schuller, B. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. *International Conference on Latent Variable Analysis and Signal Separation*. 91-99.

Weninger, F., Hershey, J. R., Le Roux, J. & Schuller, B. (2014). Discriminatively trained recurrent neural networks for single-channel speech separation. *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 577-581.

Williamson, D. S., Wang, Y. & Wang, D. (2015). Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing,* 24**,** 483-492.

Wölfel, M. & Mcdonough, J. W. (2009). *Distant speech recognition*, Wiley Online Library.

Wu, C.-H. & Liu, C.-H. (2012). Robust Speech Recognition for Adverse Environments. *Modern Speech Recognition: Approaches with Case Studies***,** 1.

Xie, F. & Van Compernolle, D. (1994). A family of MLP based nonlinear spectral estimators for noise reduction. *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*. II/53-II/56 vol. 2.

Xu, Y., Du, J., Dai, L.-R. & Lee, C.-H. (2013). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters,* 21**,** 65-68.

Yao, R., Zeng, Z. & Zhu, P. (2016). A priori SNR estimation and noise estimation for speech enhancement. *EURASIP journal on advances in signal processing,* 2016**,** 101.

Yu, D. & Deng, L. (2016). *Automatic Speech Recognition: A Deep Learning Approach*, Springer.

Yu, D., Kolbæk, M., Tan, Z.-H. & Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 241-245.

Yu, H., Zhu, W.-P., Ouyang, Z. & Champagne, B. (2020). A hybrid speech enhancement system with DNN based speech reconstruction and Kalman filtering. *Multimedia Tools and Applications***,** 1-21.

Yuan, W. (2020). A time-frequency smoothing neural network for speech enhancement. *Speech Communication*, 124**,** pp.75-84.

Zhang, H., Zhang, X. & Gao, G. (2018a). Training supervised speech separation system to improve STOI and PESQ directly. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5374-5378.

Zhang, Z., Deng, C., Shen, Y., Williamson, D. S., Sha, Y., Zhang, Y., Song, H. & Li, X. (2020). On loss functions and recurrency training for GAN-based speech enhancement systems. *arXiv preprint arXiv:2007.14974*.

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A. E.-D., Jin, W. & Schuller, B. (2018b). Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9**,** 1-28.

Zhao, Y. (2000). Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises. *IEEE Transactions on Speech and Audio Processing*, 8**,** 255-266.

Zheng, N. & Zhang, X.-L. (2018). Phase-aware speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27**,** 63-76.