YAŞAR UNIVERSITY
GRADUATE SCHOOL

MASTER THESIS

NEW APPROACHES FOR SPEECH ENHANCEMENT WITH WAVELET TRANSFORM

ELİF ÖZEN

THESIS ADVISOR: ASSIST. PROF. DR. NALAN ÖZKURT

ELECTRICAL AND ELECTRONICS ENGINEERING

PRESENTATION DATE: 12.01.2022

BORNOVA / İZMİR
January 2022

We certify that, as the jury, we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Jury Members:	**Signature:**
Assist. Prof. (PhD) Nalan ÖZKURT Yaşar University	.....
Prof.(PhD) Cüneyt GÜZELİŞ Yaşar University	.....
Assoc. Prof.(PhD) Erkan Zeki ENGİN Ege University	.....

Prof. (PhD) Yücel Öztürkoğlu
Director of the Graduate School

ABSTRACT

NEW APPROACHES FOR SPEECH ENHANCEMENT
WITH WAVELET TRANSFORM

Özen, Elif

MSc, Electrical and Electronics Engineering

Advisor: Assist. Prof. Dr. Nalan ÖZKURT

January 2022

Today, in the light of technological developments, communication is gaining more and more importance. Although there are various communication methods, one of the most frequently used communication bases is speech. Today, communication takes place between humans and between humans and machines in many crucial applications. Therefore, speech signals must be clear and intelligible to ensure these communications are carried out smoothly. The speech enhancement application improves the quality and intelligibility of speech signals by removing the noise effect as much as possible. With the increase in speech-based applications, research in this field has gained momentum. Generally, speech enhancement methods are examined under two main classes: single-channel and multi-channel methods. In this study, In this study, we proposed a new approach for both types to increase the success of the method used up to now with the help of the wavelet transform.

The first proposed method is a wavelet transform domain adaptive filter system. Since speech signals and noise are non-stationary signals, adaptive filters are one of the most preferred methods to denoise them. However, the application of adaptive filter in the time domain has some deficiencies, such as lower convergence speed especially for large datasets. Therefore, Transform Domain Adaptive Filters (TDAF) have been used in some studies. With the proposed method, we aimed to eliminate deficiencies of existing TDAF in terms of convergence speed, denoising rate, and computational complexity with multiple sub-band adaptive filters fully applied in the wavelet transform domain. The performance of the proposed system was tested on speech signals under the effect of various noises such as white noise, pink noise, babble noise, engine idling noise, aircraft cockpit noise. The commonly used objective measures were used to evaluate results. However, as our primary focal point in the study is enhancing speech signals, our aim is not only decreasing noise on the signal but also increasing the quality and intelligibility of speech signals. Therefore, objective

measures such as Perceptual Evaluation of Speech Quality (PESQ) and the Short-Time Objective Intelligibility score (STOI) were used to evaluate processed speech signals. Finally, the results were compared with the studies in the literature.

The second method proposed in the thesis is a Convolutional Neural Network (CNN) combined with wavelet transform. This is a single-channel speech enhancement application, and the main challenge in this method is distinguishing speech signals from unknown noise. Many deep learning-based methods have been used to ensure this in recent years. CNN is one of the methods used for speech enhancement applications. Commonly, it is used for image processing in many applications. We trained CNN with scalograms obtained by the magnitude of Continuous Wavelet Transform (CWT) in this method. In this way, as scalograms are two-dimensional data like images, we aimed to utilize to best properties of CNNs. Also, wavelet transform is one of the best methods to observe signals in the time-frequency plane. By combining CNNs and wavelet transform, we investigated the contribution of wavelet transform in terms of increasing the success of the existing methods and decreasing computational complexity. Finally, we evaluated the results with standard speech evaluation criterias and presented them with comparisons.

Keywords: Speech enhancement, single-channel, double-channel, adaptive filters, transform domain adaptive filters (TDAF), discrete wavelet transform (DWT), continuous wavelet transform (CWT), scalograms, convolutional neural networks (CNN)

ÖZ

DALGACIK DÖNÜŞÜMÜ İLE KONUŞMA İYİLEŞTİRME İÇİN YENİ YAKLAŞIMLAR

Özen, Elif
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği
Danışman: Dr. Öğr. Üyesi Nalan Özkurt
Ocak 2022

Günümüzde teknolojik gelişmelerin ışığında iletişim giderek daha fazla önem kazanmaktadır. İletişim çeşitli yöntemlerle gerçekleşse de en sık kullanılan iletişim tabanlarından biri konuşmadır. Günümüzde iletişim sadece insanlar arasında değil, birçok önemli uygulamada insanlarla makineler arasında gerçekleşmektedir. Bu nedenlerden dolayı, iletişimin sorunsuz bir şekilde sağlanabilmesi için konuşma sinyalinin temiz ve anlaşılır olması gerekir. Konuşma iyileştirme uygulamaları, gürültü etkisini mümkün olduğunca ortadan kaldırarak konuşma sinyallerinin kalitesini ve anlaşılırlığını artırmak için kullanılır. Konuşma tabanlı uygulamaların artmasıyla bu alandaki araştırmalar da hız kazanmıştır. Bu amaçla kullanılan yöntemler, tek kanallı ve çok kanallı yöntemler olmak üzere iki ana sınıf altında incelenir. Bu çalışmada, dalgacık dönüşümü yardımıyla şimdiye kadar kullanılan yöntemin başarısını artırmak için her yöntem için yeni bir yaklaşım önerdik.

Önerilen ilk yöntem, bir dalgacık dönüşümü alan uyarlamalı filtre sistemidir. Konuşma sinyalleri ve gürültü, statik olarak durağan olmayan sinyaller olduğundan, uyarlanabilir filtreler, gürültüyü gidermek için en çok tercih edilen yöntemlerden biridir. Ancak, zaman alanında uyarlanabilir filtre uygulamasının, büyük veri kümeleri için daha düşük yakınsama hızı ve oranı gibi bazı eksiklikleri vardır. Bu nedenle bazı çalışmalarda Dönüşüm Alanında Uyarlanabilir Filtreler (DAUF) kullanılmıştır. Önerilen yöntemle, dalgacık dönüşümü alanında tam olarak uygulanan çoklu alt bant uyarlamalı filtreler ile mevcut DAUF'in yakınsama hızı, yakınsama oranı ve hesaplama karmaşıklığı açısından eksikliklerini gidermeyi amaçladık. Önerilen sistemin performansı, beyaz gürültü, pembe gürültü, gevezelik gürültüsü, motor rölanti gürültüsü, uçak kokpit gürültüsü gibi çeşitli gürültülerin etkisi altında konuşma sinyalleri üzerinde test edilmiştir. Sonuçları değerlendirmek için yaygın olarak

kullanılan objektif ölçümler kullanıldı. Ancak, çalışmadaki öncelikli odak noktamız konuşma sinyallerini iyileştirmek olduğundan, amacımız sadece sinyal üzerindeki gürültüyü azaltmak değil, aynı zamanda konuşma sinyallerinin kalitesini ve anlaşılırlığını artırmaktır. Bu nedenle, işlenmiş konuşma sinyallerini değerlendirmek için Konuşma Kalitesinin Algısal Değerlendirmesi (PESQ) ve Kısa Süreli Amaç Anlaşılabilirlik puanı (STOI) gibi nesnel ölçüler kullanıldı. Son olarak sonuçlar literatürdeki çalışmalarla karşılaştırıldı.

Tezde önerilen ikinci yöntem, dalgacık dönüşümü ile birleştirilmiş bir Evrişimsel Sinir Ağıdır (ESA). Bu yöntem, bir tek kanallı bir konuşma geliştirme uygulamasıdır ve bu yöntemdeki ana zorluk, konuşma sinyallerini bilinmeyen gürültüden ayırt etmektir. Bunu sağlamak için son yıllarda birçok derin öğrenme tabanlı yöntem kullanılmaktadır. ESA da son yıllarda konuşma iyileştirme için kullanılan yöntemlerden birisidir. ESA, normalde birçok uygulamada görüntü sinyallerini işlemek için kullanılır. Bu yöntemde, biz Sürekli Dalgacık Dönüşümünün (SDD) büyüklüğü ile elde edilen skalogramlarla ESA'yı eğittik. Bu şekilde, scalogramlar da görüntü gibi iki boyutlu veriler olduğu için ESA'nın en iyi özelliklerinden yararlanmayı amaçladık. Ayrıca dalgacık dönüşümü, sinyalleri zaman-frekans düzleminde gözlemlemek için en iyi yöntemlerden biridir. Çalışmanın bu bölünde, ESA'yı dalgacık dönüşümüyle birleştirerek, dalgacık dönüşümünün mevcut yöntemlerin başarısını artırma ve hesaplama karmaşıklığını azaltma açısından katkısını araştırdık. Son olarak, sonuçları standart konuşma değerlendirme ölçütleriyle değerlendirdik ve karşılaştırmalar ile sunduk.

Anahtar Kelimeler: Konuşma geliştirme, tek kanal, çift kanal, uyarlanabilir filtreler, dönüşüm alanı uyarlamalı filtreler (DAUF), ayrık dalgacık dönüşümü (ADD), sürekli dalgacık dönüşümü (SDD), skalogramlar, evrişimli sinir ağları (ESA)

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge various people who have been journeyed with me in recent years as I have worked on this thesis.

Firstly, I know that a master's education is a long and arduous journey. I feel fortunate about having an advisor who always supports me and allows me to consult and ask for help on every issue I was stuck with. I can't thank enough Asst. Prof. Dr. Nalan ÖZKURT, who always supported me endlessly in solving all the problems that I encountered during the progress of the thesis and was always by my side with her understanding, helpfulness, and sincerity. I would like to present my endless thanks for her unwavering support, patience, and guidance in making this work possible.

Many other people have contributed directly or indirectly to make this study possible. Secondly, I would like to thank my committee members Prof. Dr. Cüneyt GÜZELİŞ and Assoc. Prof. Dr. Erkan Zeki ENGİN for letting my defense be an enjoyable moment, and for their brilliant comments and suggestions. Also, I would like presents my special thanks to Prof. Dr. Cüneyt GÜZELİŞ for his great contribution to shaping thesis ideas by valuable information and suggestions. Then, I would like to express my gratitude to all my professors, whom I took courses from during my graduate education, for their contributions to my scientific development.

Also, I would like to thank my dear friends Yiğitcan ACARBAY and Hayriye DÖNMEZ, who were with me during my graduate education and gave me moral support whenever needed. This would have been a much more difficult feat without them. Thank them all for their unwavering support and help.

Finally, it is impossible to extend enough thanks to my family; my dear father Bektaş ÖZEN, my dear mother Fatma ÖZEN, and my dear brother Zafer ÖZEN. Thanks to them for being by my side in every moment of my life, supporting me, and giving me strength in every field.

Elif ÖZEN

İzmir, 2022

TEXT OF OATH

I declare and honestly confirm that my study, titled “NEW APPROACHES FOR SPEECH ENHANCEMENT WITH WAVELET TRANSFORM” and presented as a Master’s Thesis, has been written without applying to any assistance inconsistent with scientific ethics and traditions. I declare, to the best of my knowledge and belief, that all content and ideas drawn directly or indirectly from external sources are indicated in the text and listed in the list of references.

Elif Özen

12.01.2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1. Engineering Problem and Thesis Motivation

Speech enhancement can be defined as removing background noise from a speech by protecting the quality and intelligibility of speech signals. It is frequently used for voice communication applications. In the light of technological developments, communication is gaining more and more importance. Although the communication may be in text, audio, image or speech, speech signals are used more frequently in communication.

Speech signals are exposed to various noises during the recording and transmission stages. For instance, in cellular communication, especially in hand-free mode, the microphone which records the speech signal is located at a certain distance from the sound source. In this case, the speech signal recorded by the microphone also includes background noise, that is, ambient noise (Qin Linmei et al., 2001). When voice communication occurs in a high-noise environment, such as the aircraft cockpit, interior of the construction equipment, a crowded place, the speech signal is highly destroyed by this high noise effect, ambient noise. It is not easy to ensure successful voice communication because this signal will be distorted by many effects (such as channel noise) during communication. Considering the airplane cockpit example, under these conditions, the listener on the ground may misunderstand the information given by the pilot. Therefore, the value of smooth voice communication is emphasized more if it is thought about the importance of the information sent at this stage. Overall, background noise is one of the biggest obstacles to smooth voice communication. Hence, speech enhancement applications are needed to eliminate this obstacle and ensure smooth voice communication.

The field of application is not limited to communication only. Media/information sharing based on speech signals such as podcasts, audiobooks, and interviews is one

of the most exciting topics for social media creators today. However, a quiet environment without acoustic and ambiance noise is needed for the clearly understandable content offering. Thanks to successful speech enhancement applications, audiobooks, podcasts, and interview recordings can be recorded in any environment without special equipment (Xing Luo, O., 2019). It is also used for the smooth operation of robust speech recognition and voice command technologies. Furthermore, in the biomedical field, it is frequently preferred in hearing aid design.

As a result of this increasing demand for applications, as detailed above, the research about the speech enhancement application is motivated. Up to now, there have been several studies in this area. These will be detailed in the next section.

1.2. Literature Review

Today, many studies are carried out in the field of speech enhancement. Although the interest in each method used in these studies has changed over time, the approaches in speech enhancement are incredibly comprehensive. If speech enhancement is examined under the general heading of de-noising digital signals, the oldest source of these studies is conventional filters to remove noise. However, the traditional Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filter with constant filter coefficient will not be sufficient to de-noise speech signal because of speech and noise signals' statistically non-stationary properties. Therefore, Adaptive Filters were commonly preferred in former studies of speech enhancement application to remedy this deficiency of conventional filters. (Haykin, Adaptive filter theory 1996).

Adaptive filters can be defined as the filters that adjust filter coefficients according to input signals. Considering this feature, adaptive filters in filtering non-stationary signals (such as speech) give more efficient results. Various learning algorithms are used in the adaptive filters to perform adjustment of filter coefficients such as Least Mean Square (LMS), Normalized Least Mean Square (NLMS), and Recursive Least Squares (RLS) (Haykin, Adaptive filter theory 1996). The performance of different algorithms for speech enhancement applications was compared by several studies (Gupta et al., 2015 and Borisagar & Kulkarn, 2010). The results showed that LMS algorithms outperformed other learning algorithms for speech enhancement applications in terms of ease of application, computational complexity, and converge

speed. For this reason, we prefer this algorithm in the method proposed for the two-channel speech enhancement application in this study.

Applying adaptive filters for large data sets in the time domain increases computational complexity and decreases convergence speed. Because of this, many researchers have worked on using adaptive filters in a transform domain (Shams Esfand Abadi et al., 2017). The concept of applying the adaptive filters in the transform domain was introduced by Dentino in an article published in 1978 (Dentino et al., 1978). After this article, research on this topic has gained momentum (Donoho & Johnstone, 1994). In the former studies, some orthogonal transform methods such as Fourier Transform (FT), Discrete Cosine Transform (DCT), Walsh-Hadamard Transform (WHT) were used more frequently (Jenkins et al., 2009 and Huang, 1999). The results showed that applying adaptive filters in an orthogonal transform domain decreases computational complexity and increases the convergence speed of the filter. Especially for the LMS algorithm, the convergence speed of the filter is highly dependent on the eigenvalue spread of the autocorrelation matrix of the input signal. With the help of orthogonal transforms applied to the input signal, the eigenvalue spread is arranged thanks to the de-correlation of the input signal.

With the widespread use of wavelet transform applications, Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT) have become one of the main methods to enhance non-stationary signals such as speech. Furthermore, studies on the application of adaptive filters in the wavelet transform domain have gained momentum since the wavelet transform is orthogonal, helpful in observing the time-frequency resolution, and the processing complexity is less than other orthogonal transformations. For instance, the computational complexity is defined by "$\mathrm{Nlog}_2 N$ " for FFT, while for WT this computational complexity is equal to "N", where N is the length of the input signal. Thus, WT is much easier to apply in large data sets than FT (Burrus et al., 1998).

Some inspiring work on the application of adaptive filters in the WT domain can be listed as follows. In one of the most remarkable studies (Akhaee et al., 2005), a hybrid method to reduce the noise of the speech signal was proposed. In this method, the LMS algorithm is used for the signal's low-frequency components (approximation coefficient), while thresholding and Wiener filter are used for high-frequency components (Detail coefficients). In (Attallah, 2000 and Hosur & Tewfik, 1997), the

success of the WTD-LMS algorithm in removing noise from essential (sine, pulse, binary sequence, etc.) signals were tested. In these studies, error signal calculation is made in the time domain, and only one adaptive filter is applied. In this way, the inverse transformation has to be done at every iteration of the algorithm. This method increases computational complexity too much, especially for data with many samples. Furthermore, unfortunately, it will not be possible to completely filter the noise with a single adaptive filter since the same noise level does not affect all sub-bands of the signal in each noise type.

In the light of information given up to now, in the first part of our study, we worked on a two-channel speech enhancement method that uses Wavelet Transform Domain (WTD)-LMS/NLMS algorithm. For Two-channel speech enhancement applications, a speech recording system with two sensors or microphones is required. In other words, noisy speech signals should be recorded from two different sources. In this system, one microphone records the noisy speech signal, while the other is positioned closer to the noise source and records the noise signal. These systems can be used when the speaker and the noise source are almost stationary (such as aircraft cockpit, construction equipment interior, etc.) and require extra costs and equipment to record the noise (reference) signal. To avoid these additional requirements, researchers have studied single-channel speech enhancement applications. In these applications, there is no need for a second recording device that records the reference signal. However, although much work has been done in this area so far, due to some reasons that will be explained in the following sections, the success of two-channel speech enhancement applications, especially in terms of voice intelligibility, has not been achieved by single-channel systems.

DWT is one of the most preferred methods for single-channel speech enhancement applications. It is mainly used in speech enhancement applications with thresholding, spectral subtraction, Wiener filtering methods because it provides an excellent resolution to examine different frequency values of non-stationary speech signals. In former thresholding methods, threshold values are manually adjusted, which is hard to implement, especially for unknown noise. In the latter application, the noise estimated after estimating the noisy frames (Active Voice Detection (AVD)) in sub-bands of signal with various decision-making algorithms is used to perform spectral subtraction or thresholding (Özaydın & Alak, 2018 and Abd El-Fattah et al., 2013). Even though

this method is acceptably successful in increasing the Signal to Noise Ratio (SNR) value, they are mostly unsuccessful in enhancing speech intelligibility due to the loss of speeches' frequency component where the frequency value of speeches and noises overlapped.

In the last few years, due to the acceleration of artificial learning and deep learning applications, the use of *Deep Neural Networks* (DNN) in speech enhancement applications has become popular. Among the deep learning methods, the most commonly used speech enhancement methods are *Deep Auto Encoders* (DAE) (Feng et al., 2014), *Recurrent Neural Network* (RNN) (Maas et al., 2012), *Long Short-Term Memory* (LSTM) (Gao et al., 2018), *Speech Enhancement based on Generative Adversarial Network* (SEGAN) (Pascual et al., 2017), and *Convolutional Neural Network* (CNN) (Park & Lee, 2017 and Yuliani et al., 2021). In this study, we worked on a CNN-based network for a single-channel speech enhancement system.

CNN is a model inspired by the vision mechanism of animals and obtained by combining this mechanism with mathematical theory. It can extract the spatial relationship between image pixels with the help of sliding filters in each layer. In addition, it has been reported that it is more efficient than RNN-based speech enhancement applications and provides more successful results with fewer parameters (approximately ten times smaller network) than RNN (Park & Lee, 2017). Thanks to the CNN's ability to capture the pattern in two-dimensional data, it is predicted that the system will efficiently distinguish between speech and noise when the time-frequency distribution of speech signals is used as an input signal of CNNs. In light of this thought, many studies so far have used spectrograms of speech signals as the input signal of CNN (Shi et al., 2018). The spectrogram contains the time-frequency distribution information of the speech signal obtained by the Short Term Fourier Transform (STFT) of the speech signals. One of the most effective methods of observing the time-frequency distribution of signals is the scalograms obtained by the CWT of the signal. Thanks to the multi-resolution provided by WT, scalograms allow more efficient observation of all frequency components. Based on this information, a speech enhancement application with a CNN using scalograms of speech signals as input is proposed in this study. In the study, it is thought that increasing resolution in the time-frequency distribution, which is the input signal, will increase the network’s success in capturing the required pattern.

1.3. Aim of Study

There are three main purposes of presenting this thesis. The first and most important aim is to propose speech improvement methods with increased efficiency (in terms of error reduction, intelligibility increase, and quality improvement) by using the features offered by wavelet transform in signal analysis. The other two aims are to evaluate the success of the proposed methods with the evaluation criterias accepted for speech improvement applications and to present the results comprehensively and comparatively.

In the thesis, two different sound enhancement applications were proposed for two different recording systems. The study's contributions aimed to be achieved for each application are as follows.

- *Two-channel enhancement application: speech enhancement with WTD-LMS/NLMS algorithms*

The proposed method aims to increase the success of the applications done so far and eliminate the previously stated deficiencies of two-channel speech enhancement applications, especially for voice communication with hands-free mode. For this purpose, in the proposed method, after separating the signal into sub-bands with DWT, a separate adaptive filter is applied to each sub-band. This way aims to avoid speech's noise effect as much as possible, even for the noise with changing spectral properties. Also, in the proposed method, adaptive filtering is done entirely in transformation domain. Thus, avoiding inverse transformation at every step reduces the complexity of the process. The final aim of this application is to optimize the proposed method's parameters and obtain a closed-box two-channel speech enhancement system that provides de-noising of speech signals under variable noise effect.

- *Single-channel enhancement application: Speech Enhancement by CNN using Scalograms*

The key objective of this method is to investigate the success of the proposed CNN method in terms of speech enhancement ability. In the proposed method, scalograms were used as input of CNN to utilize the multi-resolution properties of CWT. To achieve the stated goal, obtaining scalograms of speech signals with optimal parameters, designing the CNN that will best process these features for speech

enhancement, comparing the results with previous studies, and measuring the system's success was carried out one by one.

1.4. Thesis Organization

The contents of the chapters of the thesis are as follows:

• Chapter 1 – Presents the motivation of the thesis and the engineering problem, provides brief information about the studies in the literature in the field of speech enhancement by discussing their advantages and disadvantages. Finally, it indicates the aim of the study for each application done in the study.

• Chapter 2 – Explains the theoretical background of the study briefly. Firstly, it gives short information about the scope of speech enhancement applications, nature of speech, commonly effective noises, and frequently used evaluation criterias. Then, it presents theoretical knowledge about the methods used in speech enhancement applications: Wavelet Transform, Adaptive filters, and Convolutional Neural Networks (CNN) in this study.

• Chapter 3 – Includes new approaches for speech enhancement application with obtained results after simulation of methods implemented on MATLAB. There were two new approaches to utilize the wavelet transformation's well performance in speech signal examination in the study. Firstly, it describes an adaptive filter system in the wavelet transform domain as a double-channel speech enhancement application. Then, it explains a single-channel speech enhancement application with CNN fed by wavelet scalograms. Moreover, it presents the data used in the implementation, methodology of the proposed methods or approaches, results, and discussions with comparisons for each new approach.

• Chapter 4 – Summarizes all results obtained in the thesis and discusses the study's contribution to state of art. Finally, it presents envisioned further studies and developments.

CHAPTER 2
THEORETICAL BACKGROUND

2.1. Speech Enhancement

Speech enhancement applications aim to improve both the quality and intelligibility of speech signals impaired by additive noise. From this point of view, it is possible to say that this field of study is a specialized sub-application of audio denoising applications and is sometimes known as the noise removal method (Loizou, 2017).

In many cases, speech enhancement is necessary, and it can be used as a pre or post-processor to ensure smooth application functionality. In most scenarios, the speech signal is corrupted by the noise from the ambiance, recording devices, or transmission channels. For example, in voice communication, ambient noise has a highly disruptive effect on the speech recorded in a boisterous environment at the transmitting end. Therefore, it is beneficial to apply speech enhancement at the receiver end to increase speech quality or before transmission to ensure smooth voice communication (Loizou, 2017). In a speech recognition or voice command system, with the help of speech enhancement applications, the recognition accuracy of the system can be increased. In hearing aid design, the background noise can be removed from speech before the amplification process by a speech enhancer to provide the best understanding of conversations to patients. These examples emphasize the importance of speech enhancement applications (Chaudhari & Dhonde, 2015).

Speech enhancement methods can be classified based on the number of recording sensors used in the system (Xu et al., 2015). There are single-channel or multi-channel recording methods. Two or more microphones are used to record speech signals in a multi-channel system. The most commonly used method among multi-channel recording is a two-channel system that includes separate microphones for reference noise and noisy speech signal recording. The microphone used to record the reference signal is closely located at the noise source. There is only one microphone in a single-channel system, and this records noisy speech signals. These systems' main challenge is distinguishing unknown noise from speech signals. In this case, characteristics of

the noise source, the relationship of noise with clean speech (interference, correlation) are gaining importance.

Furthermore, the number of sensors used in the system can affect the success of speech enhancement. In general, multi-channel systems with more sensor are more successful than single-channel systems. In other words, the increasing number of sensors in speech enhancement applications provides a rise in success (Zhang & Zhao, 2013 and Loizou, 2017). However, single-channel speech enhancement application is still one of the significant research areas because of ease of application, lower implementation cost, and convenience (Chaudhari & Dhonde, 2015).

This study proposes an advanced application of single-channel and double-channel speech enhancement with the help of wavelet transform. Thus, we aim to observe the contribution of wavelet transform for both methods.

2.1.1. Speech Signals and Noises

This section will present brief information about speech signals and noises primer subjects of speech enhancement applications.

Speech is a type of sound produced by humans. Sound waves are defined as waves transmitted by the compression and rarefaction of particles that cause pressure changes in the atmosphere. These waves that we cannot observe with the naked eye are likened to those that appear when a stone is thrown into a still pond. The primary source of sound waves is vibrations emanating from an entity. This entity can be an instrument string, the diaphragm of a speaker, or the vocal cords of a human being (Borisagar et al., 2019).

In general, speech is pressure waves created by reshaping the air from the human lungs by the vocal cords, mouth, tongue, teeth, and lips (Rabiner and Schafer, 2007). Speech signals, which form the basis of auditory communication, are an acoustic waveform of an analog message. The microphone converts this acoustic waveform into an electrical waveform for later analog or digital processing. However, the recorded signal by the microphone is still an analog or a continuous signal. Therefore, it is crucial to convert this signal to the digital form to store, transmit, or process in digital environments such as computers.

As an essential part of digital to analog converters, sampling rate affect the intelligibility, amount of information, and perceptual quality of speech. For digital speech processing, after several studies about the nature of speech signal, its production, and characteristics of speech(phonemes), it was observed that the optimum sampling frequency for speech signal is 44.1 kHz (CD quality) as the electromagnetic spectrum of the speech signal in between 20-20 kHz (Rabiner and Schafer, 2007). With this sampling rate, all speech features in terms of intelligibility are saved by the digital version, and the digital version of the speech is nearly the same as the original version. However, according to application demand, lower sampling frequencies can be selected in speech processing applications. For instance, for telephone communication, the sampling frequency of the speech signal is 8 kHz. Since, intelligibility reduces in lower rates, sampling frequencies less than 8 kHz is not appropriate for speech processing applications. In this study, we worked with 8 kHz as we focused on the case of mobile communication in a high noise environment with hands-free mode.

The time-amplitude representation of speech signals is called waveform representation. Because of the nature of speech, the speech signal's time and amplitudes are dynamic, continuously changing over time. Therefore, it is hard to distinguish most of the critical properties of speech only by observing the waveform of speech. In this case, the frequency domain examinations or frequency spectrums are a helpful tool. Although speech signals have changing frequency content over time, when the Fourier spectrum of the signal is examined, it is observed that 80% of the energy of speech signal lies below 1 kHz, and a negligible amount of energy exists above 8 kHz (Borisagar et al., 2019). Therefore, it can be said that a speech signal is a low-band signal with most of the energy is located in lower frequencies. However, in terms of intelligibility, all frequency components, including higher frequencies, have critical importance (Monson et al., 2014).

From the signal processing view, noise can be defined as unwanted additive signals that affect the desired signal and reduce its quality or processing capacity (Haykin, 1996). From the first moment that a speech signal comes out of the human mouth, it is exposed to various noises coming from the entire environment. For example, speech signals propagating from wireless medium come across various noises emitted or produced by different noise sources such as ambient acoustic noise, thermal noise of recording devices, channel noise, electromagnetic noise (Borisagar et al., 2019). The

source of noises can change, but they affect the quality and intelligibility of speech signals. In general, the success of speech processing in terms of noise reduction is highly dependent on knowledge about noise features (Vaseghi, 2008). Moreover, the spectral characteristics of the noises determine the noise reduction method to be used. For instance, while the noise reduction process with conventional filters gives successful results when the frequency components of the speech signal do not overlap with the noise, noise reduction methods such as adaptive filters, which require further investigation, are needed in case the frequency components overlap. Figure 1.1 shows the possible noise effect coming from different sources, affecting speech signals throughout any speech applications. According to Borisagar, the definitions of the noises emitted from different sources are as follows (Borisagar et al., 2019):

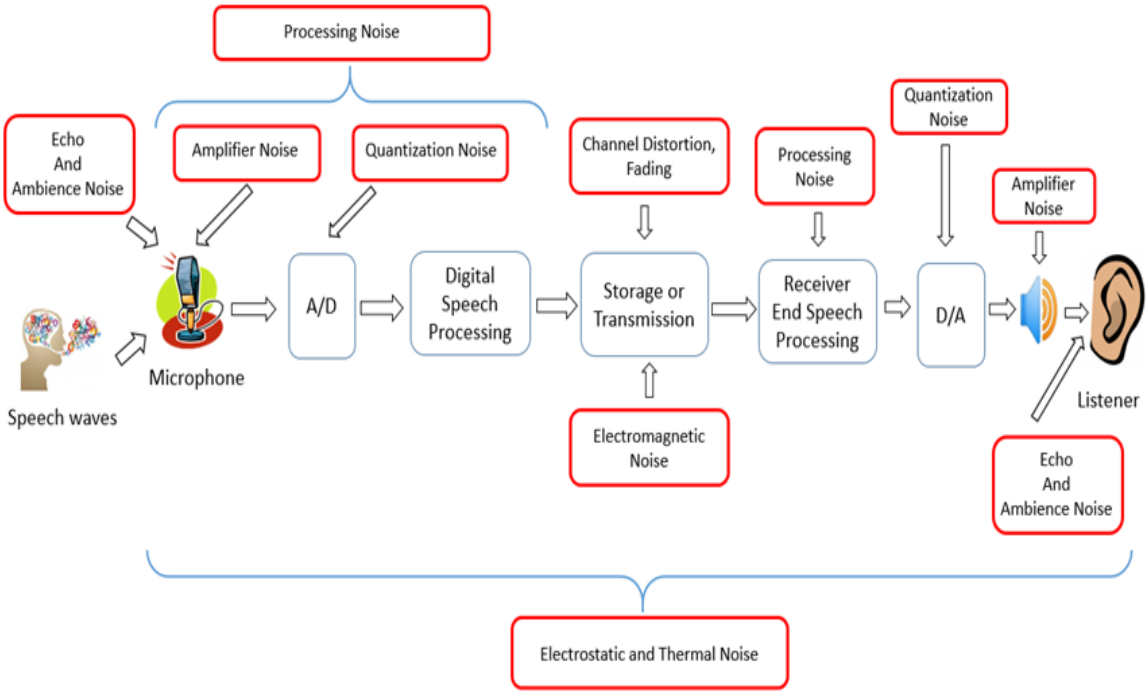


Figure 2.1. A schematic shows possible noises from different sources affecting the speech signals throughout any speech application.

- ***Ambient noise or Acoustic noise:*** Noises originating from the environment. These noises can be created by rotating, moving, or vibrating objects such as engines, cars, work machines, electrical devices such as air conditioners, fans, natural events such as wind, storms, or by people.

- ***Thermal noise:*** It is produced by electronic devices because of heat. The components of electronic devices move or vibrate because of heat caused by electric current and produce thermal noises. It is unavoidable for all electronic environments. Generally, additive white noise is used to simulate this noise in digital processing.
- ***Electrostatic noise:*** It is the noise caused by the presence of voltage or, in other words, caused by the flow of electric current. A well-known example is a noise produced by fluorescent lamps.
- ***Electromagnetic noise:*** The noise affects all frequency bands during the transmission or reception of speech or other data with radio frequencies. Therefore, all devices working with radio frequency are exposed to this noise.
- ***Channel Distortion, fading and echo:*** *Channel distortion* can be defined as the losses caused by the transmission medium while sending a signal from the transmitter to the receiver. *Fading* occurs when the receiver and transmitter are mobile during communication. The transmitted signal may be weakened or distorted due to fading. In addition, the reflection of the speech signals from the objects in the environment and returning to the recording device is defined as *echo* and has a disruptive effect on the intelligibility of the speech.
- ***Processing noise:*** It is the type of noise caused by the errors that occur during speech signal processing. It can be caused by quantization, especially in converting the analog speech signal to digital and converting it back to analog. One of the reasons for this error is data loss due to error-prone channels. It can also occur during encoding/compressing or decoding/decompressing stages for the same reason.

In addition, noises are examined according to their spectral properties under various classifications such as white noise, colored noise, narrow-band noise, band-limited noise (Borisagar et al., 2019). These classifications have been made based on the frequency band where the noise is effective.

It is critically important to have basic knowledge of speech signals and noise signals for successful speech enhancement application. In this study, we aimed to propose speech enhancement applications that can reduce all types of noise effects on speech

signals. For this purpose, we select a variety of noise that can simulate the majority of these noise effects on speech and tried to reduce this effect as much as possible.

2.1.2. Evaluation of Speech Enhancement Applications

Another critical issue for speech enhancement applications is the evaluation of speech quality or intelligibility, which is the most important criterion for measuring method success. In this section, we introduced some of the globally accepted evaluation metrics used in the study to measure and compare the study's success with similar studies in the literature

Speech evaluation studies are generally divided into two categories as subjective and objective assessment methods. In subjective methods, it is expected that a group of pre-trained listener rate the quality or intelligibility of speech signal under pre-determined limits after the real-listening process. These methods provide the most convenient, reliable, and robust assessment of speech quality and intelligibility. However, these methods are time and effort-consuming because of listeners' real-listening process and training (Yi Hu & Loizou, 2006). Therefore, objective methods are mostly preferred for the evaluation of speech processing applications.

In the objective methods, the quality or intelligibility of speech is measured by mathematical comparison of clean and processed speech signals. The main goal of this method is to evaluate speech quality by using the numerical distance between related signals. In this study, we used six different objective evaluation metrics to measure the success of the approaches offered.

The Mean Square Error (MSE) is the first method used in the study to measure the success of the speech enhancement process. It refers to the average energy of error on the speech signals. This error can be thought of as the amount of distortion on the signal. In this case, decreasing value of MSE refers to minimum distortion on speech signals. The formula used to calculate the MSE value is presented in (1) (Haykins, 1996) as;

$$\text{MSE}=\frac{1}{N}\sum_{n=0}^{N}\big(s(n)-y(n)\big)^2 \tag{1}$$

where N is the number of samples, s(n) is the n^{th} observation of the clean speech and y (n) is the de-noised signal. Since the magnitudes of signals are various in different

algorithms, the range of MSE will vary from one study to another. This metric is frequently used in learning algorithms to observe converges to optimum results.

Signal to Distortion Ratio (SDR) or *Signal to Noise Ratio (SNR*) is the other method used to measure the success of the speech enhancement method. It can be interpreted as the ratio of the energy of processed signal to the energy of distortion in decibel (dB). In this method, distortion or error signal is calculated by taking the difference between clean and processed signals. The equation used to calculate SDR is given in equation (2) (Park & Lee, 2017).

$$SDR = 10\log_{10}\frac{\sum_{n=0}^{N} s(n)^2}{\sum_{n=0}^{N}\left(s(n)-y(n)\right)^2} \quad (2)$$

Where y(n) is the clean signal, s(n) is the enhanced speech signal, N is the number of samples. A high SDR value indicates that we are getting closer to the value we desire, clean speech. This ratio is a measure that is frequently used as an evaluation criterion in noise reduction application for speech signals.

SNR is helpful to observe the ratio of the energy of error on the signal, but for further examination, for speech signal, a particular type of SNR is used called Segmental SNR (Seg-SNR). This metric can be calculated both in the time and frequency domain, but the calculation of Seg-SNR in the time domain is commonly preferred (Loizou, 2017, p.635-636). In this method, SNR values are calculated for short-time segments of speech, and by taking the average of these values Seg-SNR is obtained. The formula of Seg-SNR calculation is given in (3).

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1} log_{10}\frac{\sum_{n=Nm}^{Nm+N-1} s(n)^2}{\sum_{n=Nm}^{Nm+N-1}(s(n)-y(n))^2} \quad (3)$$

where s(n) is the original (clean) signal, y(n) is the enhanced signal, N is the frame length (32 milliseconds (ms) for this study), and M is the number of frames in the signal (Loizou, 2017, p.635).

Although MSE and SNR values are sufficient to measure the convergence of the processed speech signals to clean speeches by observing the energy of the error signal, they do not contain any information about the quality and intelligibility of the speech signal. For this purpose, the *Perceptual Evaluation of Speech Quality (PESQ)* and the *Short-Time Objective Intelligibility score (STOI)* measurement criteria were used in the study.

PESQ is a family of standards that includes a test method for automatic objective evaluation of speech quality (Al-Akhras et al., 2010). It is standardized as ITU-T Recommendation P.862 (02/01) (Hu & Loizou, 2008) and the code taken directly from the standards was used in this paper. In this standard, audio signals are scored between 0.5-4.5; 0.5 indicates that the sound quality is very poor, and 4.5 indicates that the sound quality is very high. STOI is a method used for the subjective intelligibility estimation of the audio signal. It is often used to prevent loss of time and workforce caused by real listening and evaluation practices. In this method, the intelligibility of audio signals was scored with correlation values ranging from 0-1. Therefore, it is possible to say that the intelligibility of the audio signal with a high STOI value is higher (Taal et al., 2011). The algorithm used at this thesis is taken from (Taal et al., 2011).

2.2. Wavelet Transform

Analyzing signals in the time domain does not always provide enough information for signal processing. Therefore, many transformation methods have been used to analyze and process signals, such as Fourier Transform (FT), Laplace Transform (LT), Fast Fourier Transform (FFT), Short Time Fourier Transform (STFT), and Wavelet Transform (WT).

Fourier transform is one of the former methods used to analyze signals. This method is used to transform time-domain signals into the frequency domain. So, we can observe the frequency content of a signal, but there is no information about the time that includes this frequency content (Huang, 1999). Therefore, FT is not appropriate for analyzing a non-stationary signal with changing properties both in the time and frequency domain, such as speech signals. Then, STFT, a type of Fourier transform calculated over the short signal time, was introduced for the frequency-time analysis of the signals. In this method, a windowing operation is applied to signal to divide signal into small segment then FT of this small segments calculated to ensure time and frequency information at the same time. The window size is vital for this method because it is directly related to the time-frequency resolution. Therefore, it should be nearly equal to stationary segments on the signal, and it is hard to determine for unknown non-stationary signals. Moreover, the size of the windows for STFT is the same for all frequency bands. Thus this method does not provide good resolution for

high frequencies of the signal. Therefore, WT with multi-resolution properties, detailed in the section, was proposed to overcome the drawbacks of other transform methods, especially for non-stationary signal processing.

Up to now, WT has been used for many signal processing applications such as heart monitoring, analyzing financial indices, video image compressing, denoising (Addison, 2002). Furthermore, because of its pretty good performance in signal analysis and feature extraction, WT contributes to the success of methods combined with it, such as various artificial learning applications and speech enhancement applications. In this section, we will briefly present wavelet transform methods used in this study.

2.2.1. Continuous Wavelet Transform

The wavelet transformation is an orthogonal time-frequency transformation and is generally used to separate the signal into high and low-frequency components. The WT represents the signal in terms of wavelets which are scaled and translated mother wavelets, like FT, which represents signal by superposition of sine and cosine. To calculate WT, we need a wavelet that is the function satisfying specific mathematical criterias. This wavelet (also called mother-wavelets) is used to localize the time and frequency properties of the signal by being manipulated through the process of translation (i.e., shifting over the time axis) and scaling (i.e., stretching or compressing the wavelet) (Addison, 2002). Therefore, selecting the mother-wavelets function has great importance on proper feature extraction by this method. There are several mother wavelet functions used to calculate WT and Figure 2.2 illustrates some of them.

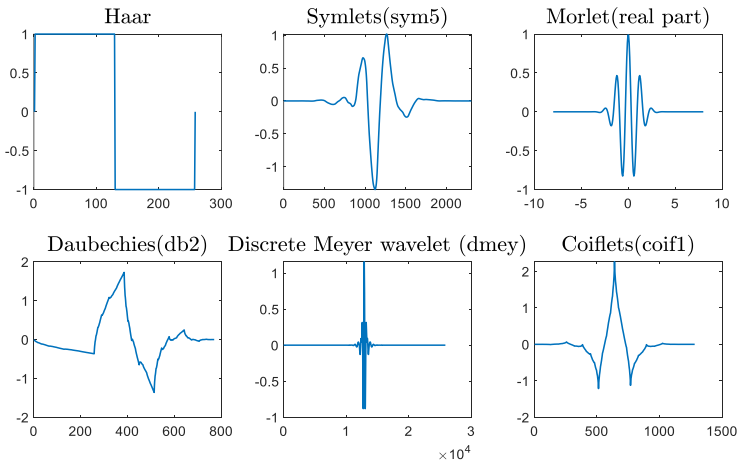


Figure 2.2. The graph of six different mother-wavelet functions used commonly.

The wavelet coefficient obtained after WT represents the measure of similarity or correlation in the time-frequency content between a signal and a selected mother-wavelet function and these coefficients are calculated by the convolution of the signal and the scaled mother-wavelet function (Ergen, 2012). The formula used to calculate continuous wavelet transform (CWT) is presented in (4) (Addison, 2002);

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \Psi^* \left(\frac{t-b}{a} \right) dt \tag{4}$$

Where $x(t)$ is the signal in time domain, $T(a,b)$ is the CWT of the signal, $\Psi\left(\frac{t-b}{a}\right)$ is translated and scaled mother-wavelet function, a is scale parameter, b is translation parameter and the asterisk indicated the complex conjugation operator.

In the study we use CWT to obtain scalograms of speech signal in feature extraction phase of speech enhancement application with CNN. Scalograms is a method used to observe time-frequency energy density of a signal. The formula used to calculate scalograms is given in (5) (Addison, 2002);

$$SC(a,b) = |T(a,b)|^2 \tag{5}$$

Where $SC(a,b)$ is known as two-dimensional energy density function of signal at a scale and b location. A plot of SC gives the scalogram. The resolution of the time-frequency distributions obtained with the scalogram and spectrogram were compared with the visuals shown in Figure 2.3.

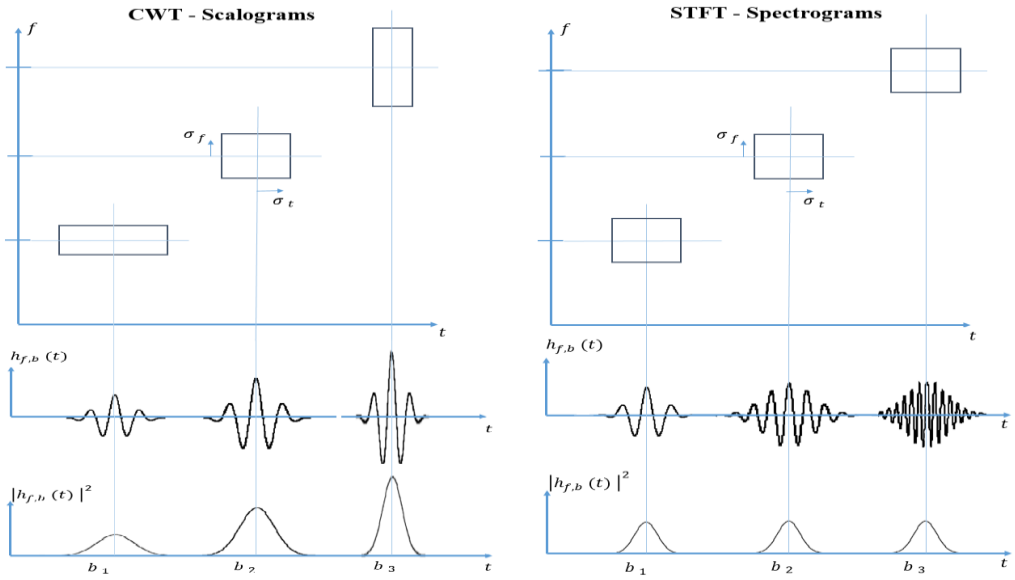


Figure 2.3. Time-frequency resolution comparison between spectrogram and scalograms (Addison, 2002).

As seen in the figure, the scalograms provide better time-frequency resolution for higher frequencies. However, with spectrograms, it is hard to observe rapid changes in higher frequency. Therefore, we can miss some speech features in the feature extraction phase, especially for over noisy speech, which might decrease the ability to learn in the neural network. Starting from this point, in our study, we expected that the learning of the neural network, that is, the success of speech enhancement, would increase as a result of using the scalogram instead of the spectrogram, which is frequently used in the literature during the feature extraction stage, and we developed our single-channel speech improvement approach accordingly.

Finally, the time-domain representation of the signal is obtained using Inverse CWT (ICWT). The formula used in the reconstruction phase is given in (6) and (7);

$$x(t) = \frac{1}{C_\Psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T(a,b)\, \Psi_{a,b}(t)\, \frac{dadb}{a^2} \tag{6}$$

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-b}{a}\right) \tag{7}$$

C_Ψ in (6) is called as admissibility constant. This formula allows to obtaining original signal by integrating over all scales and locations (Addison, 2002).

2.2.2. Discrete Wavelet Transform

In CWT, the scale parameter a and translation parameter b have infinitely many values to represent the signal in the wavelet domain, and it is sometimes called a redundant transform. Discrete wavelet transform (DWT) is the discretized version of the CWT, and it is introduced to eliminate this redundancy and reduce computational complexity.

DWT is calculated as a result of discretizing scale a and translation b parameters in wavelet function. The equation representing the discrete version of the wavelet function is shown in (8).

$$\Psi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} \Psi\left(\frac{t - n b_0 a_0^m}{a_0^m}\right) \tag{8}$$

Where m and n are the integers that control scale and translation, respectively, a_0 is fixed scale step-size, b_0 is fixed translation parameter, and $\Psi_{m,n}(t)$ is the discretized version of the wavelet function.

By replacing discrete wavelet function wavelet transform formula given in (4) the equation of DWT can be obtained. The formula of DWT is;

$$T_{m,n} = \int_{-\infty}^{\infty} x(t)\Psi_{m,n}(t)\, dt \tag{9}$$

Where $T_{m,n}$ are discrete wavelet transform coefficients on the scale-location grid of index m, n (Addison, 2002).

The wavelet coefficient should satisfy the condition given in (10) to ensure the validity of inverse transform for DWT.

$$AE \leq \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \left|T_{m,n}\right|^2 \leq BE \tag{10}$$

Where A and B are upper and lower frame bounds and E is the energy of signal in time domain.

To obtain discrete wavelet function, one of the common choices for parameters a_0 and b_0 are 2 and 1, respectively. This scale-location frame is called a *dyadic grid*, and it is the simplest and most efficient way of discretization for many applications. Furthermore, the wavelet functions obtained as a result of this selection are orthonormal. Thanks to this property, after wavelet decomposition of the signal, we can observe and process subbands of the signal separately without any loss.

The basic idea of the DWT is to decompose the signal into sub-signals corresponding to different frequency band contents. In the decomposition step, a signal is expressed as a series of orthonormal wavelet functions that constitute a wavelet basis (Misiti, 2006). Starting from the formula given (9), it can be said that DWT is a filtering operation with a discrete wavelet function representing filters in varied scales (Huang, 1999). Therefore, DWT can be implemented using the filter bank to decompose the signal into different subbands. The decomposition of the signal into different subbands with different resolutions ensures multi-resolution ideas can be realized using successive low pass and high pass filtering. The schematic given in Figure 2.4 explains the two-level decomposition and reconstruction of a signal in DWT.

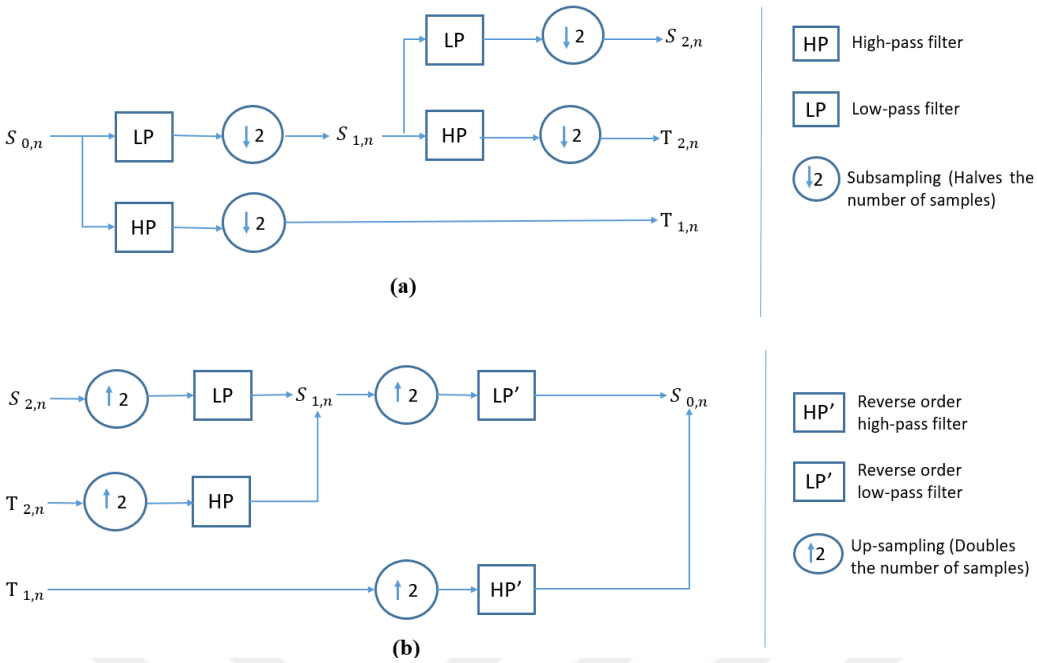


Figure 2.4. (a) 2-Level decomposition of the signal, (b) Reconstruction of the signal using detail and approximation coefficients.

In the decomposition phase given in Figure 2.4 (a) $S_{0,n}$ is called as 0^{th} level approximation coefficient and it is equal to x(t) which is the original signal, in general $S_{m,n}$ is the m^{th} level approximation coefficient for m=0,1,2.. and $T_{m,n}$ represents the m^{th} detail coeficents m=1,2… The signal can be represented using this approximation and detail coefficient obtained as a result of DWT. In general, approximation coefficients include information about the signal's lower frequency content, and the detail coefficients give information about the higher frequency content. The increasing number of decomposition levels allows observing higher frequencies of signal with increasing frequency resolution. However, after each decomposition level, the time resolution decreases because of the subsampling operation. Therefore, it is crucial to determine the correct decomposition levels to observe the signal's subbands with good resolution. The procedure shown in the figure can be repeated for further decomposition by adding successive low and high pass filters.

Figure 2.4 (b) describes the reconstruction of the original signal using this detail and approximation coefficients.

In general, the original signal in the time domain can be obtained by the formula given in (11) or (12) which is called Inverse DWT.

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{m,n} \Psi_{m,n}(t) \tag{11}$$

$$x(t) = \sum_{n=-\infty}^{\infty} S_{m',n} \emptyset_{m',n}(t) + \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{m,n} \Psi_{m,n}(t) \tag{12}$$

In (12) $\emptyset_{m',n}(t)$ is called as scaling function and ıt is represented as high-pass filter in figure 2.4. This formula summarizes the information that the original signal can be obtained by summing the approximation and detail coefficients at decomposition levels.

2.3. Adaptive Filters

Filters can be examined under two main headings: adaptive filters and non-adaptive filters (Gupta et al., 2015). Conventional filters, which are non-adaptive, are filters with constant filter coefficients. Because of this property, it is not possible to process statically non-stationary signals with these filters. Besides, to denoise signals with this type of filter, some characteristic information about noise signals such as the influential frequency band of noise should be known precisely. However, the signals used in real-life applications such as speech is generally non-stationary, and the characteristic of the noise signal that causes distortion may not be known in every case. Moreover, even the characteristic of noise is known, the frequency components of signal and noise can be overlapped. For example, the frequency content of a speech signal under the effect of low SNR broad-band noise mostly overlaps with the noise's frequency content. Therefore, if we try to denoise this speech signal with conventional filters, it is very probable to lose overlapping frequencies' that affect the speech's intelligibility. In such cases as in the example, adaptive filters are preferred. The diagram showing the overall functioning of the adaptive filters is shown in Figure 2.5.

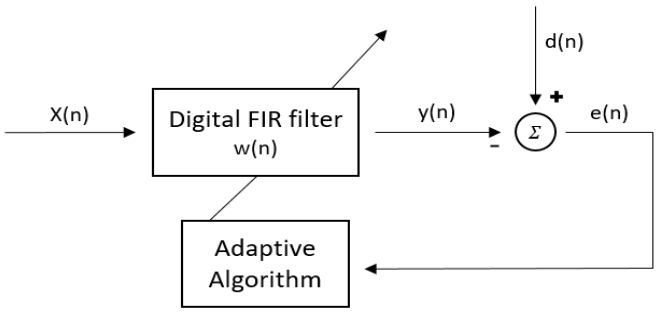


Figure 2.5. Block diagram of adaptive filtering

Adaptive filters can adjust the filter coefficients based on the current value of the input signal without having any prior knowledge of the characteristics of noise affecting the signal (Haykins, 1996). In these filters, the output signal y(n) is obtained by convolution of the input signal with x(n) and digital filter coefficients w(n). Then, the error signal is obtained by taking the difference of the desired signal d(n) and the output signal y(n). Finally, adaptive learning algorithms update the digital filter coefficients using the resulting error signal e(n) in every iteration of the filtering operation. This process continues until the desired performance criteria are met (Kumar & Rajan, 2012).

The main working principle of the adaptive filter is the minimizing squared error value. Wiener-Hopf equations are used to achieve optimum adaptive filter weight in general. These equations are accepted as the basis of adaptive filters and algorithms, and the representation of these equations in matrix format is as follows (Haykin, 1996):

$$\boldsymbol{R} \cdot \boldsymbol{w_o} = \boldsymbol{p} \tag{13}$$

$$\boldsymbol{w_o} = \boldsymbol{R^{-1}}.\boldsymbol{p} \tag{14}$$

The **R** symbol shown in the equations describes the auto-correlation matrix of the input sequence, the **p** symbol indicates the cross-correlation vector of the input signal and the desired signal. Finally, $\boldsymbol{w_o}$ include optimum filter coefficients. Thanks to these equations, we can achieve the adaptive filter's optimum filter coefficients (weights). However, it is not easy to achieve the optimum solutions analytically with these equations because of the computational complexity caused by statistical examinations and matrix inversion. Therefore, some adaptive learning algorithms that aim to achieve these optimum weights iteratively are preferred to eliminate these extra computational costs.

The Least Mean Squares (LMS) algorithm is an improved Steepest Descent *a*lgorithm, one of the most commonly preferred learning algorithms for adaptive filters. The main reason for choosing this algorithm for the proposed *adaptive speech enhancement approach* is that it provides ease of calculation, good converge speed, robust solutions in terms of stability. The equations used in the LMS algorithm can be listed as follows (Haykin, 1996);

$$y(n) = w(n)^H.x(n) \tag{15}$$

$$e(n) = d(n) - y(n) \tag{16}$$

$$w(n+1) = w(n) + \mu x(n) e^{*}(n) \tag{17}$$

Where y(n) is the output signal, w(n) is the initial value for filter coefficient, x(n) is the input signal, d(n) is the desired signal, μ is the step-size, w(n+1) represents updated filter coefficients, the superscript H denotes Hermitian transposition, and * denotes the complex conjugation. As frequently used for adaptive noise cancellation applications, a type of double-channel sound enhancement application, the noisy speech signal is used as the input signal x(n), and the noise signal is used as the reference or desired signal d(n). So, the error signal e(n) was estimated by running the adaptive algorithm giving out the noise-free speech signal.

The adjustment of step-size μ is of critical importance in stability of LMS algorithm. For the algorithm to function smoothly, the step-size value must satisfy the following condition (Haykin, 1996);

$$0 < \mu < {}^{1}/_{\lambda_{max}} \tag{18}$$

where λ_{max} is the maximum eigenvalue of the autocorrelation matrix of the input signal.

NLMS algorithm, another algorithm used in the study, is obtained by the normalization of the LMS algorithm. The set of equations used to implement the NLMS algorithm can be defined as follows (Haykin, 1996);

$$e(n) = d(n) - w(n)^{H}.x(n) \tag{19}$$

$$w(n+1) = w(n) + \frac{\tilde{\mu}}{\delta + \|x(n)\|^2} x(n) e^{*}(n) \tag{20}$$

where $e(n)$ is represented as the error signal, $d(n)$ is the desired signal, $w(n)$ is the initial value for filter coefficient, $x(n)$ is the input, $\tilde{\mu}$ is adaptation constant and $w(n+1)$ is represented as updated filter coefficient. The operation of the LMS and NLMS algorithms is very similar. The main difference between these two algorithms is that step-size is normalized with the energy of input signal.

LMS algorithms are mostly preferred adaptive learning algorithms, and these filters can be used for many applications, especially speech enhancement, thanks to their ease of application and robustness. However, these filters in the time domain have some critical drawbacks for processing large data sets or signals with many samples. For

example, the computational complexity and the converge time, the time required to meet desired performance criteria or filter weight, increase when a signal with an increasing sample number is processed with a time-domain adaptive filter. Therefore, the concept of *Transform Domain Adaptive Filter* (TDAF) was introduced to overcome these deficiencies.

The TDAF can be defined as a parallel application of an adaptive filter to the pre-processed input signal with an orthogonal transform and normalization (Beaufays, 1995). Many orthogonal transformation methods are used in TDAF, such as Fourier transform, Discrete Cosine Transform, Walsh-Hadamard transforms, and Wavelet Transform. Among these methods, WT steps forward because of less computational complexity and better time-frequency examination properties.

The general scheme of TDAF is illustrated in Figure 2.6.

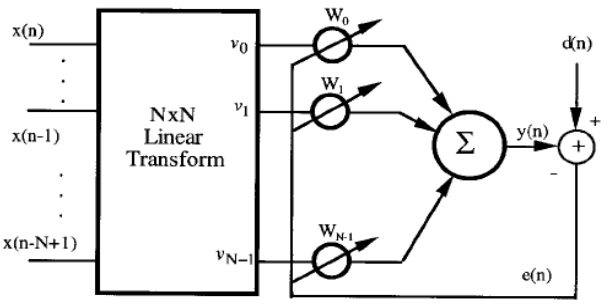


Figure 2.6. General diagram of adaptive filtering in transform domain (Jenkins & Marshall, 1999).

In transform domain adaptive filters, the input signal is first divided into parallel branches called sub-band signals using orthogonal transformations. Then, the application of adaptive algorithms to the obtained parallel branches is performed in the transformation domain. However, as seen from the figure, the error signal calculation is performed in the time domain in general. For this, firstly, the signal in the time domain is obtained using inverse transformation. After calculating the error signal in this domain, transform domain filter weights are updated using this error value.

The main disadvantage of the LMS adaptive filter in the time domain is the dependency of converge speed of the adaptive filter on the eigenvalue spread of the autocorrelation matrix of the input signal. The eigenvalue spread can be defined as the ratio of maximum eigenvalue to minimum eigenvalue (Haykin, 1996), and the

optimum converge speed for this algorithm can be achieved when the eigenvalue spread of the autocorrelation matrix is equal to one (Beaufays, 1995). Thanks to the orthogonal transformation applied to the input signal, the signal is de-correlated as much as possible, i.e., the eigenvalue distribution of the autocorrelation matrix of sub-signals approaches unity (Jenkins & Marshall, 1999 and Akhaee et al., 2005). Thus, the maximum convergence speed of the algorithm can be achieved. Furthermore, in this phase, power normalization contributes to obtaining unity eigenvalue spread and regulating error surface that increase convergence speed and stability of the algorithm. In our study, we aim to achieve power normalization only using the NLMS algorithm without extra normalization, and we observed an increase in success with the help of normalization integrated in the algorithm.

Moreover, decomposition of the signal into subband signals with orthogonal transform provides the opportunity to process subband signals separately because orthogonal transformations minimize cross-correlation of subband signals. Thereby, with the parallel application of the adaptive filter, the adaptive filter length and the time required for convergence can be reduced because of fewer samples included by the subband signal.

In this study, we proposed a double-channel speech enhancement method using wavelet transform domain adaptive filters, a type of TDAF. To create this method, we utilize the background information presented up to now. The detail about the proposed method and results will be given in the next chapter.

2.4. Speech Enhancement with CNN

Convolutional Neural Networks (CCN or Conv-Net) is a type of deep learning network frequently used in visual estimation (Park & Lee, 2017). Due to the success of the method in image processing, it has been used in recent years to improve speech signals. In speech enhancement methods with CNN, firstly, one-dimensional speech signals are pre-proceed with time-frequency transformation, called feature exaction phase, to convert it into two-dimensional signals. Then, the data obtained after pre-process is used in CNN as an input to utilize the pretty good performance of CNN's in two-dimensional data (signal) processing. CNN is a model inspired by the vision mechanism of animals and obtained by combining this mechanism with mathematical theory (Tüfekçi & Karpat, 2019). Generally, it aims to use the spatial relationship

between image pixels. It is based on the discrete convolution of the image pixels with the filter sliding over the image to detect relationships among the pixels. The discrete convolution process in CNN is frequently used to determine the features of the image and classify the images according to these features (Shahriyar et al., 2019).

2.4.1. Learning Methods used in CNN

CNN has a multi-layered architecture with an input, an output layer, and hidden layers. There are generally three types of learning models classified as supervised, unsupervised, and semi-supervised learning (Koushik, 2016). Supervised learning can also be called mapping in general (Koushik, 2016). In this type of training, inputs and desired outputs are given to the system during the training phase. The system is expected to create a function explaining these examples' relationships. In short, it maps inputs to output.

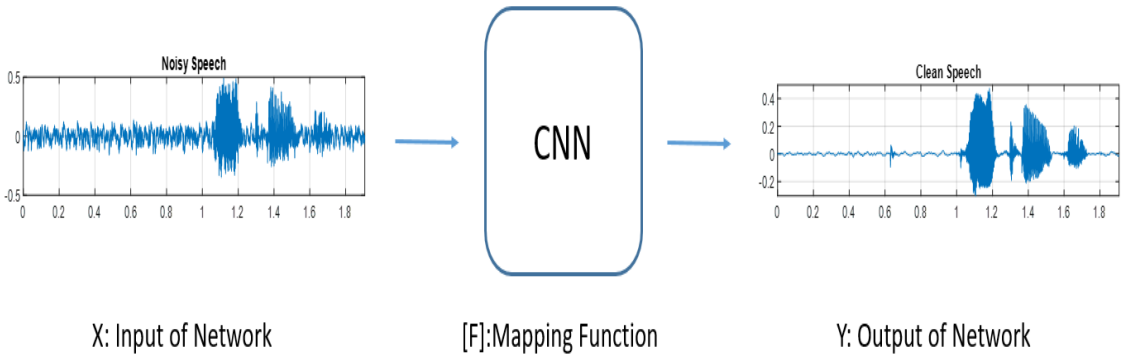


Figure 2.7. Diagram of supervised-learning based speech de-nosing with CNN.

In Figure 2.7. the X and Y are the input and desired output sample pairs to be used to train the network as shown in equations (21) and (22) in general (Shahriyar et al., 2019). In this type of learning, neural networks create a mapping function that sets the relation between input and desired output values.

$$T_s := \{ (x_s, y_s) \quad 1 \leq s \leq N \} \tag{21}$$

$$\hat{y}_s = F(x_s) \tag{22}$$

For Equation (1), T_s shows the training data set in x_s and y_s are the input and desired output sample pairs in this data set, and N is the number of samples. For Equation (22), $\hat{y}_s$ is the actual output value calculated by the system with the learned parameters. The

F function can be called a mapping function depending on system parameters.

Training CNN aims to minimize the difference between actual output values and desired output values. For this purpose, the mean square error (MSE) as a loss function is calculated in each training iteration. Then, some optimization algorithms are used to minimize this loss function. The loss function MSE equation is the same as the Equation presented in (1). Moreover, several functions calculate loss during training according to the application to be used. The RMSE function calculated by taking the square root of the MSE is one of the frequently used functions for speech enhancement applications. There are different training models or protocols for calculating the error or loss during training. These methods can be listed as follows (Stutz, 2014);

- **Stochastic training;** in this model, a random input is selected from the input set and the network parameters are updated using the error or loss function of this input.
- **Batch training;** In this model, the system parameters are updated by using the error function or loss function obtained as a result of processing the entire input set.
- **Mini-batch training;** In this training model, the error value obtained from processing the sub-input set containing a certain number of input values selected within the input set is used to update the system parameters.

There are many optimization algorithms used in minimizing the error function. The most preferred optimization algorithm is the Gradient-Descent algorithm. The algorithm generally allows updating the system parameters by using the gradient function of the loss function depending on the system parameters. This method is called the first-order optimization method. Because while obtaining the Equation used to update the system parameters, the first derivative of the error signal depending on the system parameters is calculated. The Equation of the Gradient-Descent algorithm is given in Equation (4) (Stutz, 2014);

$$\Delta MSEw = -\gamma \frac{\partial MSE_n}{\partial W} \tag{23}$$

Where γ is the learning rate constant in [0, 1] interval, w is the connection weights or general system parameters. As seen in the Equation (23), in the Gradient-Descent algorithm, the same learning rate constant is used to update all system connection weights or system parameters.

Today, an advanced Gradient-Descent algorithm, Adaptive Moment Estimator (ADAM) optimization algorithm, is used to optimize many deep learning processes (Kingma & Ba, 2015). This algorithm was used for optimization within the scope of the study. In ADAM optimization, unlike the Gradient-descent algorithm, a different learning rate obtained by using the first and second-order moments of the gradient is used to update each system parameter (Kingma & Ba, 2015). In other words, it is based on the logic that the learning rate per parameter is regulated and this learning rate is used to update system parameters as connection weights.

2.4.2. Network Architecture and Layers of CNN

As mentioned earlier, CNN has a multi-layer architecture. It also consists of several layers with different functions and contributions for the CNN architecture. Various CNN network architectures can be obtained with combinations of these layers. Some known and frequently used CNN architectures can be listed as LeNet, AlexNet, VGG Net, GoogLeNet, ResNet from simple to complex. As the complexity of the architecture increases, the number of parameters to be learned will increase, so the size of the data set and the number of learning steps (epochs) to be used for training the system should be increased. In this study, we create our network architecture, a type of CNN with the skipped connection for speech enhancement applications. The details about the architecture of the proposed network will be given in the next section. In the continuation of this section, brief information about some layers and the operations performed by layers in traditional CNN architecture and related hyper-parameters will be given.

- **Convolutional Layer:** This layer is the essential layer for CNN. The extraction of the visual features is provided by the operations performed on this layer. Generally, 3D filters are used in this layer with a size of N x M x K. Here N symbolizes the height of the filter matrix, M represents the width of the matrix, and K refers to the depth of the matrix. The output image is obtained by sliding these filters starting from the top corner of the image and summing the product of the overlapping pixels (O'Shea & Nash, 2015). This shift, multiplication, and addition process are mathematically defined as the two-dimensional discrete convolution operation, and the name of this layer comes from precisely here. A diagram showing how to obtain the output, in other words, the feature map, using the filter and the input image, is given in Figure 2.8. As shown in

figure, the matrix obtained by convolving the input matrix or image and filter or kernel is called the feature map. At the points where the filter and the image are similar, the feature map gets higher values, so it is detected where the feature represented by the filter is in the image. By sliding the filter over the image, desired features are described by the filter.

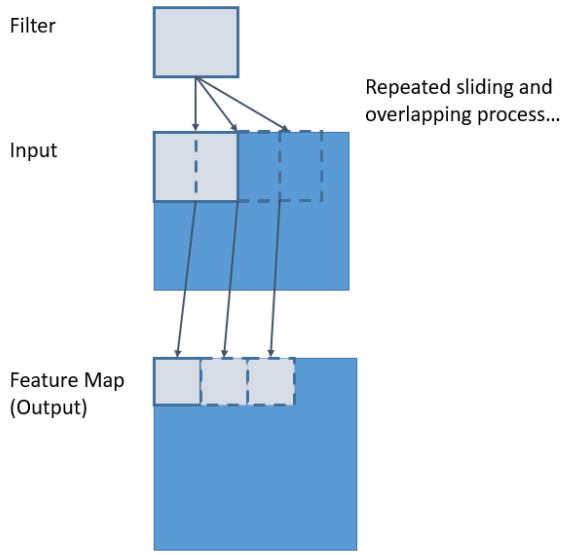


Figure 2.8. Diagram of filter applied to a two-dimensional input to create output in convolutional layer (Brownlee, 2020).

As shown in the figure, the size of the feature map may not be same as the input. The dimensions of the output matrix are calculated depending on the size of the input and filter matrices. For example, if the size of the input matrix for each layer is m x n and the kernel (filter) size is k x l, the size of the output matrix is determined as (m-k + 1) x (n-1 + 1). If the system has M layer, this process is repeated M times. Depending on the properties of the application, the dimensions of the output can be kept the same or reduced. In this case, two hyper-parameters are effectively used in this layer to regulate the dimensions of the output matrix (O'Shea & Nash, 2015).

- ***Padding:*** The output of a 5-layer convolutional network with an input matrix of 250x250 and a filter matrix of 10x10 is found as 205x205. Considering that the system will have more layers, a large part of the input matrix is slid off due to these operations. One of the procedures to get rid of this situation is padding. In the padding method, extra pixels with a value of 0 are added around the input matrix, as shown in Figure 2.9. The size of the output matrix to be obtained after padding

with size s x t becomes (m-k-s + 1) x (n-l-t + 1) (Wang et al., 2020). In this process, when s = k-1 (k: filter height), t = l-1 (l: filter width) is selected, input and output sizes are equal to each other.

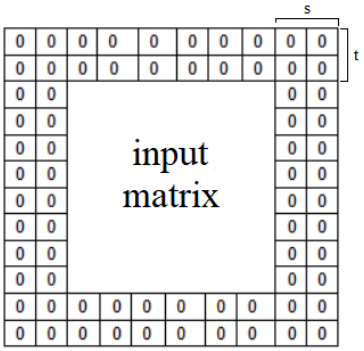


Figure 2.9. Padding example with s x t padding size

- ***Stride:*** We know that the convolutional layer's output matrix is obtained from the convolution process, which is calculated by shifting the filter over the image at each step starting from the upper left corner of the image. In the section so far, this scrolling action has been considered 1 pixel per step. The method used to determine how many pixels this filter will shift down or sideways in each step is called stride. As the filter will scan the image with certain pixel ranges due to the stride process, the dimensions and properties of the output matrix also change. If the stride size is determined as a x b, the size of the output matrix to be obtained using the m x n input matrix and k x l filter will be [(m-k + a) / a] x [(n-l + b) / b] (Wang et al., 2020).

- **Non-Linearity Layer:** In CNN architecture, a nonlinear layer is generally used after all layers (Ergin, 2020). If the nonlinear layer is not used in multi-layer CNN, the output values of the neural network cannot go beyond being a linear combination of input values. This layer is essential because not all learning processes performed with neural networks are linear. This layer is also called the activation layer because it is the layer where nonlinear activation functions are applied (O'Shea & Nash, 2015). Some nonlinear activation functions commonly used in neural networks are sigmoid, tanh, and rectifier. In the CNN field, the Rectifier (ReLu) function is generally preferred because it gives the best results in terms of training speed (Wang et al., 2020). The Equation of the ReLu function is as shown in (24).

$$f(x) = \max(0, x), \quad x \geq 0 \tag{24}$$

This function equals 0 for all input values less than zero.

- **Pooling Layer:** In CNN, this layer is generally used after the activation layer (Tüfekçi & Karpat, 2019). The primary purpose of using the layer is to reduce the number of samples in the output matrix while keeping the image features detected by the filter in the output matrix (O'Shea & Nash, 2015). The process performed on this layer is a nonlinear sample reduction process. There is no learned parameter in this layer. It is a layer that is often used to reduce computational complexity. However, it does not give successful results, especially in applications where the feature desired to be detected on the visual is essential. Therefore, it is not preferred to be used in speech enhancement applications. Various methods can be used in this pooling process. The most common of these are max-pooling and average-pooling.
- **Flattening Layer:** This layer is generally used to transform the matrix-shaped input into a one-dimensional array. It is the layer in which the connection between fully-connected layers to convolutional layers in applications such as image recognition and image captioning is made (Wang et al., 2020). Since there is no image recognition or captioning process in the study, this layer was not used.
- **Fully-Connected Layer:** This layer is generally used to transform the matrix-shaped input into a one-dimensional array. It is the layer in which the connection between fully-connected layers to convolutional layers in applications such as image recognition and image captioning is made (Wang et al., 2020). Since there is no image recognition or captioning process in the study, this layer was not used.

CHAPTER 3
EXPERIMENTAL STUDIES AND RESULTS

As explained in the first two chapters of the thesis, two new approaches for speech enhancement applications were presented in this study. One of them is a double-channel speech enhancement application named wavelet transform domain adaptive filters. The other is a single-channel speech enhancement application that combines CNN and wavelet transform. The main aim of both studies was to benefit from the wavelet transform's outperforming features in terms of signal examination for speech enhancement. This chapter will present the methods offered for achieving the study's main goal with illustrations, obtained results, tables, and comparisons.

3.1. A Two-Channel Speech Enhancement Application: Speech enhancement with Wavelet Domain LMS-NLMS algorithms

This study used the WTD-LMS algorithms to improve the speech signals with the proposed adaptive noise canceling method. The proposed method aims to increase the success of the applications done so far and eliminate the previously stated deficiencies. For this purpose, in the proposed method, after separating the signal into sub-bands with DWT, a separate adaptive filter is applied to each sub-band. This method was inspired by one of the architectures based on the different use of the WTD-LMS algorithm presented in a review study (Huang, 1999). It is aimed to avoid the noise effect on speech as much as possible by using multiple sub-band adaptive filters in parallel. Also, in the proposed method, adaptive filtering is done entirely in transformation domain. Thus, avoiding inverse transformation at every step reduces the complexity of the process. Finally, decomposing speech signal into de-correlated sub-band offers the opportunity to process fewer samples in parallel filters. So, processing time and filter order can be reduced. As a result, it is aimed to increase the convergence rate and success of the adaptive algorithm with the proposed method. Although experiments and tests were only applied in speech enhancement in this study, it is predicted that the obtained filter will give successful results in all systems where

two-channel or sensor recording are available thanks to its high convergence speed and low computational complexity.

Two experiments were made in this study to investigate the success of the proposed method. In the first experiment, the speech signal recorded in a high noise environment was improved using WTD-NLMS and WTD-LMS algorithms. As given in Section 2.3, normalization is crucial for TDAF to arrange error surface and increase converge speed. Therefore, we foresaw that normalization in the NLMS algorithm would ensure this effect without extra computational cost. To test the contribution of normalization on convergence speed of the proposed method, highly disruptive aircraft engine noise with different SNR values was added to the speech signals. Thus, we aimed to simulate a scenario of a speech taking place in an aircraft cockpit. Besides, the speech with a short duration was selected at this stage to create a challenging condition for the adaptive filter's convergence speed. Overall, the success of the proposed method has been observed in challenging conditions for adaptive noise canceling applications, and the contribution of normalization has been proven.

In the second experiment, the proposed method's success in improving speech signals under the effect of different noise signals was investigated. For this purpose, distorted speech signals were obtained by adding noise signals with different characteristics such as white noise, pink noise, engine idling sound, siren sound, cafe ambiance noise to have a low SNR value (high noise level). The proposed WTD-NLMS filter system with optimized parameters has improved these noisy speech signals. At this stage, the selected speech signal's duration is longer, and the SNR value of the loud speech was arranged to be 0 dB. This SNR value is one of the most challenging conditions for sound enhancement or noise-canceling applications. Finally, the success of the fixed system was measured only by changing the input signals, and results were compared with the studies in the literature. All applications in the study were carried out using the MATLAB program.

3.1.1. Information About the Data

Two different audio signals were used in this study to visualize the results. These audio signals are speech signals of different lengths recorded in a quiet environment. Also, noise signals recorded in the natural environment distort these speech signals for different scenarios. The noise signals used are aircraft engine noise, white noise, pink

noise, siren noise, cafe ambiance noise, and engine idling noise. These noises are preferred because; white noise represents the thermal noise (electro-magnetic noise) that recorders have; Pink noise is often preferred for testing audio applications (processing noise); Siren, cafe ambiance and engine idle noise are background noise that can often interfere with the speech signal in voice communication in hands-free mode. All audio signals used in the study were obtained through "www.freesound.com." This site offers audio signals recorded in natural environments, especially for application development and scientific research, without copyright issues (Kumar & Rajan, 2012).

In the first experiment, a scenario of improving the speech signal recorded in the aircraft cockpit was tried to be realized. The aircraft cockpit is an environment with high levels of aircraft engine noise. For voice communication to occur smoothly in this environment, the sound signal recorded must be enhanced before communication. Since speech signals are non-stationary signals and the aircraft engine noise has a spectral characteristic that covers the entire frequency band in which the human voice is present, conventional filtering is not expected to succeed in this area. The magnitude spectrum of the speech signal and aircraft engine noise used in this part of the study is as shown in Figure 3.1.

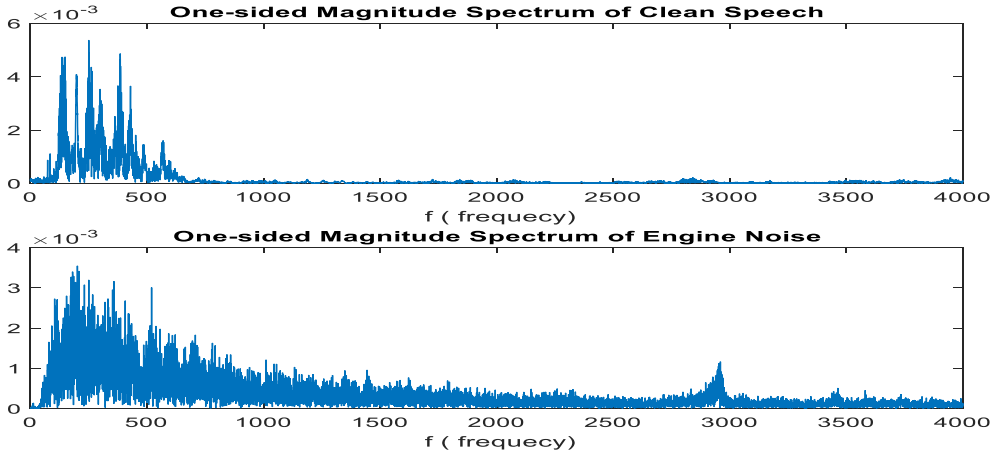


Figure 3.1. One-sided magnitude spectrum of noiseless speech and the aircraft engine noise.

As can be seen from the graph, a human voice is generally in the 0-4 kHz frequency band, while aircraft engine noise has a characteristic that completely covers this frequency band. Therefore, adaptive filtering to filter such noise from the speech signal gives more successful results. A short speech signal and aircraft engine noise were

used in the first experiment. The SNR value of the noisy speech signals was adjusted as 0, 5, 15, 30 dB using random noise segments taken from the noise signal. These refer to very high, high, medium, and low noise levels for speech signals, respectively. The time-amplitude graphics of the noiseless speech signal and the noisy speech signals obtained after the arrangements are shown in Figure 3.2.

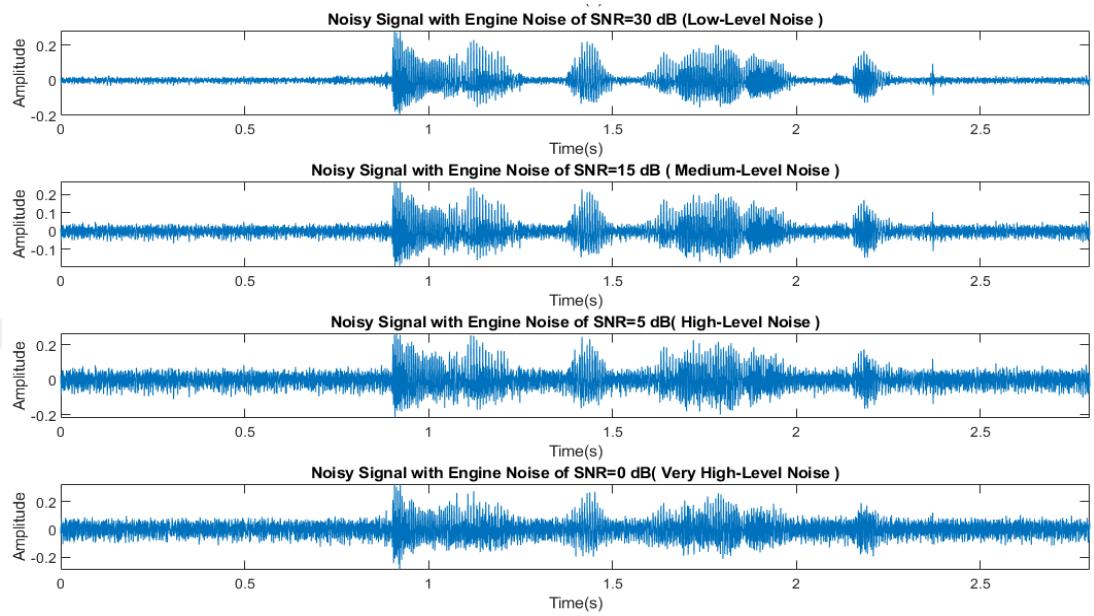


Figure 3.2. Time-amplitude graph of the clear speech signal and low, medium, high and very high noise versions of this signal, respectively.

The noisy speech signals are used as input signals for the proposed adaptive filter. As can be seen from the graph, the input signals obtained are approximately 2.8 seconds long. The signal's sampling frequency is 8 kHz, and the signal contains 22376 samples in total. The reference noise signal used in the filter is a delayed version from the randomly selected noise segment to realize acoustic delay in the virtual environment since all applications are carried out on the MATLAB program.

In the second experiment, the success of the WTD-NLMS algorithm in filtering noise signals with a very high level of noise and different characteristics was examined. For this purpose, white noise, pink noise, siren noise, engine idle noise, cafe ambiance noise was added to the speech signal with an SNR value of 0 dB. The time-amplitude graphics of the obtained noisy audio signals are presented in Figure 3.3.

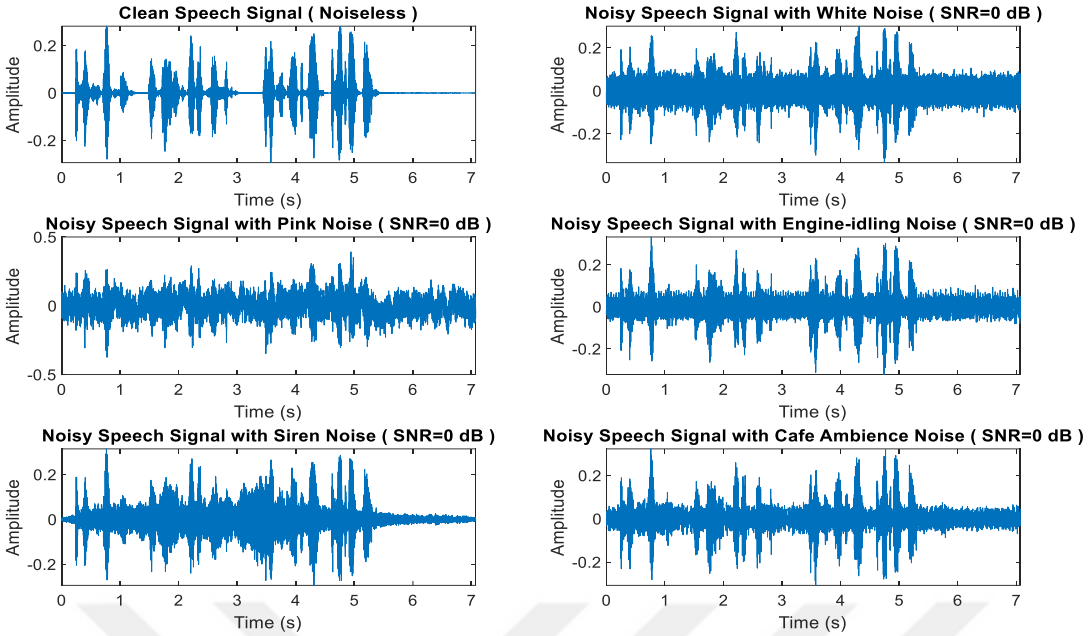
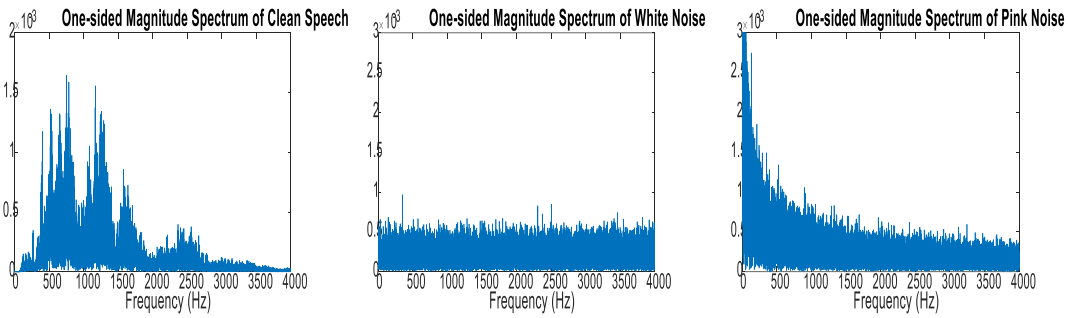


Figure 3.3. The time-amplitude graph of clean and noisy speech signals used in the second phase of the study.

Figure 3.3 shows how different noises affect the speech signal. When the graphs were examined, it was clearly observed that the sound was exposed to different distortions at different time intervals depending on the type of noise. However, it is difficult to clean such a rapidly changing noise signal with the adaptive filter applied in the time domain due to the problems arising from the convergence time of the adaptive filter. Thus, we aimed to eliminate this type of noise with the help of WT's sub-band decomposition properties. The input signals used at this stage are about 7 seconds long. The sampling frequency is arranged as 8 kHz and contains 56563 samples in total. The frequency characteristics of these noise signals are shown in Figure 3.4.



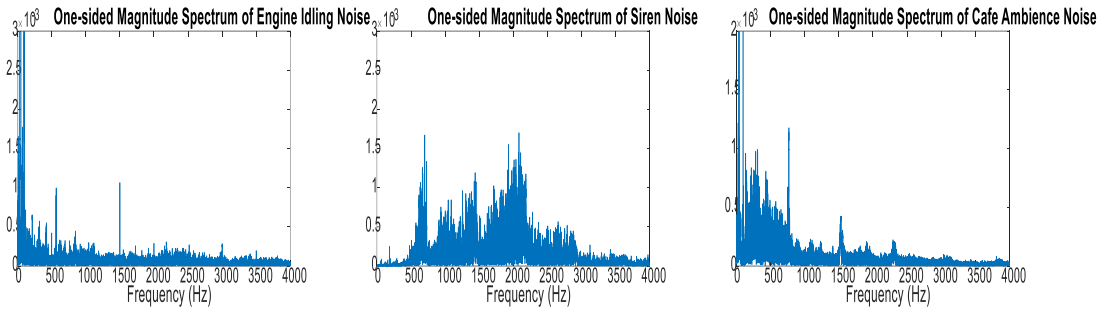


Figure 3.4. One-sided magnitude spectrum of noiseless speech and the noise signal which are white noise, pink noise, engine idling noise, siren noise, and café ambience noise, respectively.

As can be seen from the spectrums, all the noise signals used at this stage completely cover the frequency band in which the human voice is present and have a high distortion effect on the human voice. After filtering the noisy input signals with the WTD-NLMS algorithm, it is aimed to obtain a signal as close as possible to the speech signal is shown as clear speech.

3.1.2. Proposed Method and Implementation

In this part of the study, adaptive filtering in the transformation domain has been worked. As mentioned in Section 2.3, the general name of this type of application is TDAF, and these are commonly used as double-channel speech enhancement applications. After several studies and research on TDAF, the results obtained showed that the application of adaptive filters in the transform domain decreases the process complexity and increases the convergence speed of the filters. The main reason for this is that the applied orthogonal transform increases the decorrelation of the input signal, thus increasing the adaptive algorithm's convergence speed and rate.

In literature there are several orthogonal transform method has been used for TDAF. One of them is WT which is our focus point. WT has been preferred because of its good time-frequency resolution and less computational complexity. However, after examining the method used up to now, we realized that there are some deficiencies of the method because of methodologies used to utilize WT. These drawbacks are detailed in Section 1.2, with examples of studies in the literature. The block diagram of the method proposed WT-LMS algorithm in this paper is shown in Figure 3.5.

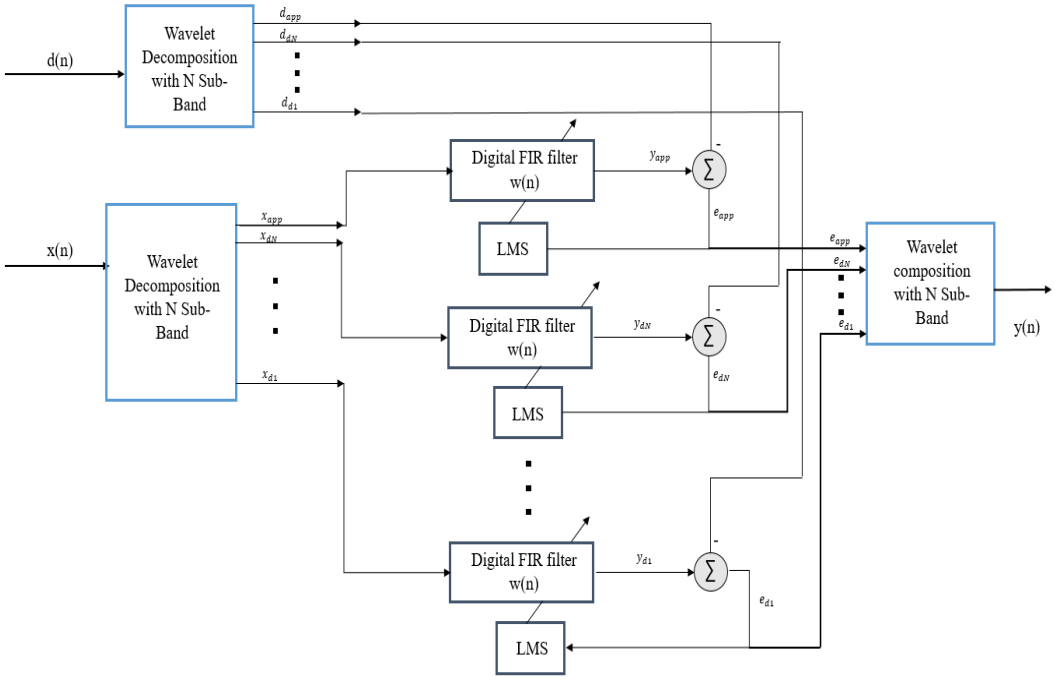


Figure 3.5. Blok diagram of proposed WTD-LMS algorithm.

As shown in the diagram, the input signal x(n) and the desired or reference signal d(n) are first divided into N sub-bands using DWT filter-bank where x(n) is noisy speech signal, d(n) is the delayed noise segment (reference noise signal). An output signal is obtained for each sub-band by applying the formulas given in (8) to (20). Then, error signals (e(n)) provided the enhanced sub-band signal. As a result of iterations in the learning algorithm, each sub-band is denoised individually. Finally, a noise-free signal is reconstructed from the filtered sub-band signals by IDWT. This application is a multi-sub-band application of adaptive filter in the wavelet transform domain.

So far, many studies have been done on the application of adaptive filters in the transformation domain. In general, error signal calculation for TDAF is made in the time domain to eliminate the transformation of the reference signal, as shown in figure 2.6, especially in filters used for noise removal or reduction operations. In this case, an inverse transform and transform of the input signal must be calculated for each iteration of the adaptive learning algorithm. This approach increases the computational complexity of the method for cases where digital signals with high sample numbers are processed. Unlike the architecture of the filters using the WTD-LMS algorithms previously used, the proposed method offers to calculate the error signal in the WT domain for each sub-band signal. In other words, the multi-subband adaptive filters

are applied entirely in the transformation domain. Therefore, it is anticipated that the number of operations and computational complexity will be reduced since inverse transformation operations will be applied only once. In addition, the idea of simultaneous application of separate adaptive filters to each sub-signal was advocated in the proposed method to remove the noise on the relevant sub-signal as much as possible. Although many noise signals cover a wide frequency band, they also have a complex frequency-time distribution. Thus, the noise affecting different sub-bands of noisy signals does not have the same disruptive effect. With the help of DWT, we desired to observe different subbands of signal separately and reduce these changing effects as much as possible with the multi-subband adaptive filters. In summary, an adaptive filter design with high convergence speed and success has been obtained by maximizing the use of the high frequency-time resolution that WT will provide.

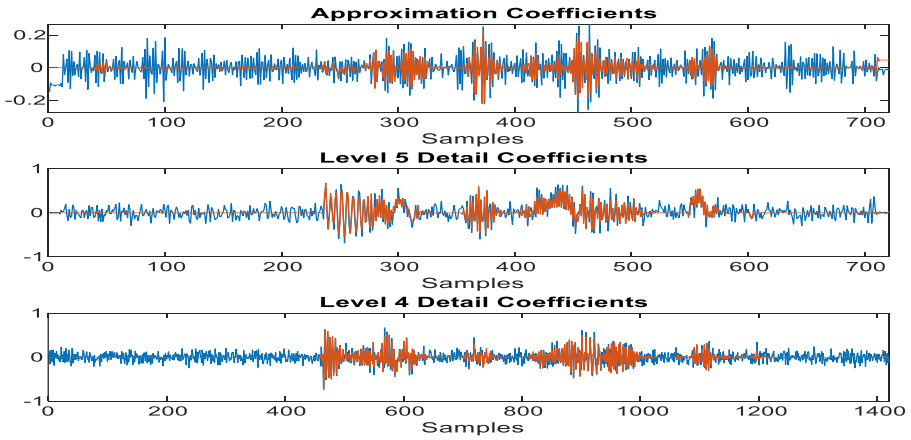
Thanks to its good converge properties, the proposed method can be used for all noise reduction applications if a two-channel recording system is available. However, the proposed method is optimized and specialized for speech enhancement application in this study. First, the Symlets and Meyer mother wavelet functions are selected for DWT application to obtain the best system. As it is known, the mother wavelet function selection is essential to extract correct features. Previous studies and our trials showed that a type of Symlet and Meyer mother wavelet function gives the best output for this application, sym5, and dmey (Özaydın & Alak, 2018 and Yan Long et al., 2004). Another critical factor in the DWT phase is deciding the decomposition level of the signal. The increasing number of levels will provide better observation for higher frequencies, but it will cause an increase in the computational complexity of the system as the number of adaptive filters is increased. After observations were done for various decomposition levels, it was decided that the best decomposition level is 5 (N=5). However, the decomposition above this level did not sufficiently contribute to the system's success rate. Then, step size and order of subband filters were selected to optimize the system after several observations and trials. After that, the system parameter was fixed, and black box filter systems using WTD-LMS/NLMS were obtained. The experiments used these systems to measure the method's success in reducing various noise effects.

3.1.3. Results and Discussions About the Experiments

3.1.3.1. Experiment 1: Cockpit Noise Removal with WTD-LMS/NLMS

The proposed WTD-Adaptive filtering algorithm's success in clearing aircraft engine noise at different noise levels from the speech signal was tested in the first experiment of the implementation. For this purpose, noisy speech signals with different noise levels were used, but the visualized results were presented for the 0 dB SNR value, which can be considered a very high noise level. Furthermore, this experiment enhanced the signals using the WTD-LMS and WTD-NLMS algorithms to observe the normalization process's contribution to the transform domain adaptive algorithm's convergence speed. Finally, results were obtained for each noise level and evaluated with previously explained criterias.

The Symlets (sym5) were used as the mother-wavelet function since it is usually preferred for speech enhancement applications and outperformed many other wavelets in this application. Therefore, all visual results were achieved by using sym5. However, measures were obtained using both dmey and sym5 since dmey is offered as the best mother-wavelet signal for speech signals in English in the study (Yan Long et al., 2004). We also heuristically observed that 5-level DWT was sufficient for decomposition signals in this study. Then, six sub-signals were obtained, including one approximate and five detail coefficients for each signal. At this stage, depending on the frequency-time distribution of the noise used, each sub-band is exposed to noise at different distortion rates. The graphics of the sub-signals (subband signals) obtained as a result of the decomposition of the input signal with 0 dB SNR value are presented in Figure 3.6.



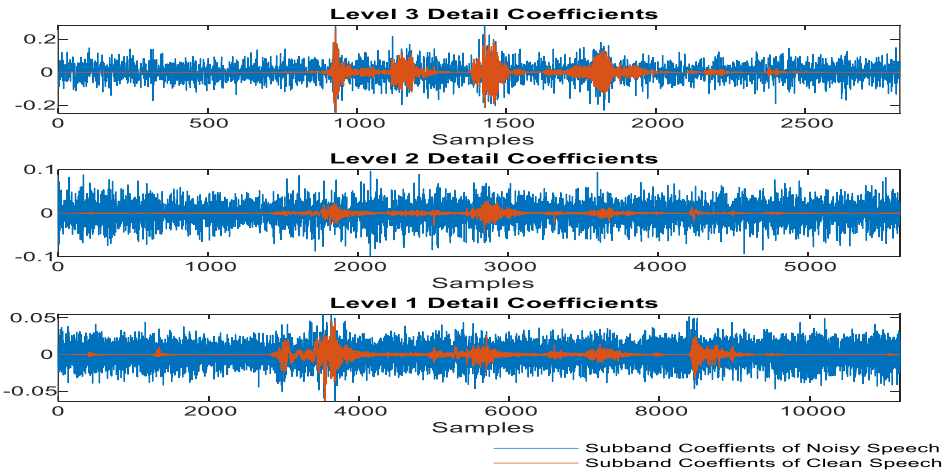


Figure 3.6. The graph of sub-signals for the noisy signal with 0 dB of SNR value

The horizontal axis of the graphs shows the number of samples contained in each sub-signal, and the number of samples is halved for each increasing number of decomposition levels according to the previous level. Therefore, the length of adaptive filters can be reduced for each sub-band application. Also, the number of transactions made during filtering is reduced. For this application, the sub-signals most affected by noise are detail-1 and detail-2 sub-signals which refer to lower speech frequencies. After separating the signals into sub-signals, adaptive filters using given algorithms are applied in parallel branches for all noisy sub-signals, and it is aimed to obtain the output signal as close as possible to the clear speech signal shown in red on the graph. The LMS algorithm's convergence speed is highly dependent on the eigenvalue distribution of the input signal, so it is also envisioned to increase the convergence speed and rate of the adaptive filter by enabling the sub-signals to be de-correlated in this way thanks to the orthogonality of the WT. Furthermore, this decorrelation provides the opportunity of processing each sub-band of signal separately.

The graphs of the sub-band signals obtained after applying adaptive filters using LMS and NLMS algorithms in the DWT domain were shown in Figures 3.7 (a) and (b), respectively.

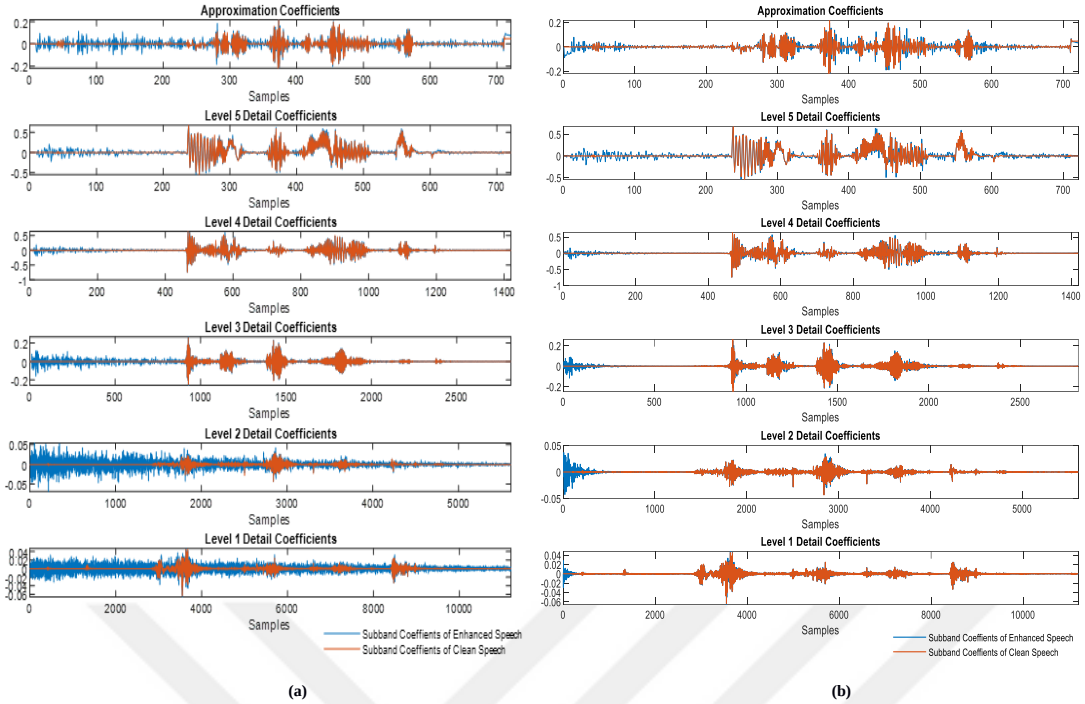


Figure 3.7. The graph of output subband signals ($e(n)_N$) obtained using (a) WTD-LMS adaptive filter (b) WTD-NLMS adaptive filter

When the obtained results are examined, it is seen that the noise on sub-signals is significantly reduced. However, it is impossible to say that the input signal is completely noise-free. As the number of iterations of the algorithm increases, the adaptive filter coefficients converge to the optimum filter coefficients. Therefore, the enhanced sub-band signals converge to clean speech. In this stage, the convergence speed of filters is critical because there is a need for time to adapt filter coefficients to changes in the input signals. If a short speech is enhanced, the convergence speed of the filter must be maximized to reduce the noise on the speech in this limited time. When the results obtained with both algorithms are examined, it can be easily seen that the convergence speed of the NLMS algorithm is much higher than the LMS algorithm. Naturally, the NLMS algorithm is much more successful. As explained before, the main reason for this situation is the energy normalization used in the NLMS algorithm. As a result of this normalization, the algorithm's convergence speed is increased by arranging the eigenvalue distribution and error surface. This case makes a significant difference in improving short-duration speech signals exposed to high noise, such as this example. In addition, this high convergence rate/speed adaptive filter will also provide successful results for filtering all signals using a real-time two-channel recording system. Overall, these results prove that normalization integrated into

NLMS algorithms successfully increases convergence speed and rate without extra transactions. The graphics of the output signals obtained by reconstruction of subband signals shown in Figure 3.7 with IDWT were given in Figure 3.8.

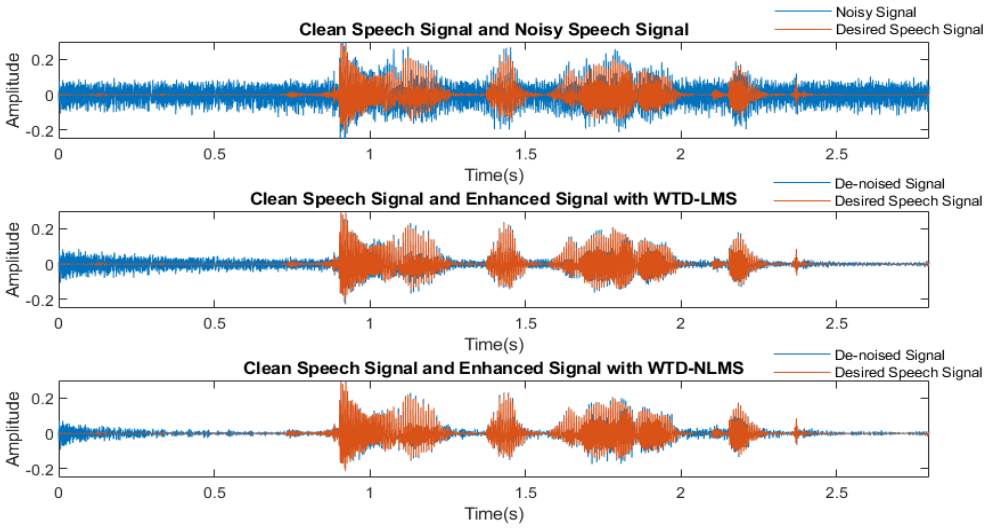


Figure 3.8. The graphs of output signals (y(n)) obtained using WTD-LMS and WTD-NLMS in time domain

When the results presented in the graph are examined, as expected, the speech signals obtained in the time domain are primarily free from noise signals. Therefore, improved speech signals are presented in the figures compared to the noiseless speech signal. Also, visually, it was seen that the results obtained with the NLMS algorithm were more successful than those obtained with the LMS algorithm. These results were obtained by selecting the step size of the filters and the filter order in both methods to obtain optimum results. The filter order selected for the NLMS algorithm varies between 4 and 5, and the filter order selected for the LMS algorithm varies between 10 and 15. Therefore, the better results of the NLMS algorithm with the smaller filter order are another proof that the NLMS algorithm will be more successful in terms of application.

The results presented so far have been obtained to improve the speech signal with a very high noise effect (SNR = 0dB) with the proposed method. Then, the same processes are applied to speech signals with SNR values of 5 dB, 15 dB, and 30 dB, respectively. Finally, the MSE, SDR, PESQ, and STOI values calculated due to the improvement of these audio signals are presented in Table 3.1.

Table 3.1. Evaluation Results of the Process Applied to Noisy Speech with 0dB, 5dB, 15dB, 30dB Aircraft Engine Noise

	Final Values with WTD-NLMS		**Final Values with WTD-LMS**		
Mother-wavelet type:	"smy5"	"dmey"	"smy5"	"dmey"	
MSE	**$9.871\ x10^{-5}$**	$1.069x10^{-4}$	$2.101\ x10^{-4}$	**$1.8268\ x10^{-4}$**	**0 dB**
PESQ	3.086	**3.193**	2.290	**2.351**	
STOI	0.961	**0.964**	0.816	**0.8362**	
SDR	**27.370 dB**	26.788 dB	19.932 dB	**21.335 dB**	
MSE	$7.596\ x10^{-5}$	**$6.445\ x10^{-5}$**	$1.820\ x10^{-4}$	**$1.517\ x10^{-4}$**	**5 dB**
PESQ	3.115	**3.208**	2.368	**2.388**	
STOI	0.9615	**0.9745**	0.8119	**0.8267**	
SDR	30.109 dB	**31.7538 dB**	21.369 dB	**25.896 dB**	
MSE	$4.994x10^{-5}$	**$4.495x10^{-5}$**	$1.029x10^{-4}$	**$1.022\ x10^{-4}$**	**15 dB**
PESQ	3.173	**3.306**	2.390	**2.440**	
STOI	0.9617	**0.9749**	0.8135	**0.8215**	
SDR	34.304 dB	**35.135 dB**	27.073 dB	**27.080dB**	
MSE	$3.860x10^{-5}$	**$3.741x10^{-5}$**	**$5.132\ x10^{-5}$**	$5.187\ x10^{-5}$	**30 dB**
PESQ	3.251	**3.398**	**2.588**	2.486	
STOI	0.9619	**0.9751**	**0.8760**	0.8608	
SDR	36.879 dB	**37.192 dB**	**34.0302 dB**	33.5451 dB	

The objective measures indicated how our proposed method achieved our aims. For speech enhancement, smaller values of MSE, increasing values of SDR, closer values of PESQ to 4.5, and STOI values getting closer to 1 indicate that the application is successful in terms of noise reduction, and enhancing speech quality and intelligibility. When the evaluation criterias presented in the table are examined, it is seen that the success of the NLMS algorithm is better than the LMS algorithm for all noise levels. However, in terms of speech enhancement in both methods, it offers acceptable, successful results in difficult conditions by selecting a short-term speech signal. With the help of selecting different mother-wavelet functions in the DWT stage, it is observed that the method's success can be increased for different language applications by determining the best mother-wavelet function for a specified language. As can be observed in the Table 3.1 both mother-wavelet functions offered satisfactory results in all measures and the best results were obtained with dmey in some measures and for sym5 for others. When looking at the PESQ and STOI values of the improved speech signal obtained by the NLMS algorithm, it is possible to say that the obtained speech

signal is very successful in terms of intelligibility and quality. Another point that draws attention to the methods applied is that the application's success in improving the speech signal decreases as the noise level affecting the speech signal decreases. This situation shows that the method has an improvement limit. The primary and the most important reason for this limit is the short duration of the audio signal used. Filtering does not provide successful results until the filter coefficients obtained during the algorithm's operation converge to the optimum filter coefficients. Thus, the noise level in the first seconds of the audio signal is higher in both methods, limiting the method's success. The results presented in the table are obtained using these algorithms at optimum convergence speed. Therefore, it seems that the maximum success limit of the LMS algorithm is lower than the NLMS algorithm. So, further examination of the method's success would be continued on the NLMS algorithm and dmey mother-wavelet function.

3.1.3.2. Experiment 2: Speech Denoising with WTD-NLMS for Various Noises

In the second experiment, the speech enhancement with noises that may frequently be exposed, such as white noise, pink noise, engine idling noise, siren noise, cafe ambiance noise, was performed. The main reason for choosing the NLMS algorithm is the proven success of the algorithm with results obtained in the first experiment. A longer speech signal than the first one was used at this stage. It is thought that the success limit of the algorithm will increase due to the longer speech sound used. In this application, all noise signals used as input signals have an SNR value of 0 dB, which is one of the most challenging cases for speech enhancement applications. The reference noise signal used in the application is a slightly delayed version of the selected segment from the noise signal. The time-amplitude and magnitude spectrum graphs of the input signals are presented in Figures 3.3 and 3.4.

Noisy speech signals were improved by using the WTD-NLMS algorithm used in the first part of the study. System parameters such as decomposition level, step-size, and filter-order are kept fixed in the test process. Thus, a black box filter system with the proposed method was obtained. The output signal was obtained as a system response to changing input and reference signals. In this way, the proposed filter's success in improving speech signals affected by various noise signals has been observed in a virtual environment. The noise signals used at this stage have different time-frequency

characteristics. However, considering the SNR values, it is possible to say that the distortion created by all noise signals on speech is high. Spectrograms of noisy signals used as filter inputs and spectrograms of filtered speech are presented in Figure 3.9.

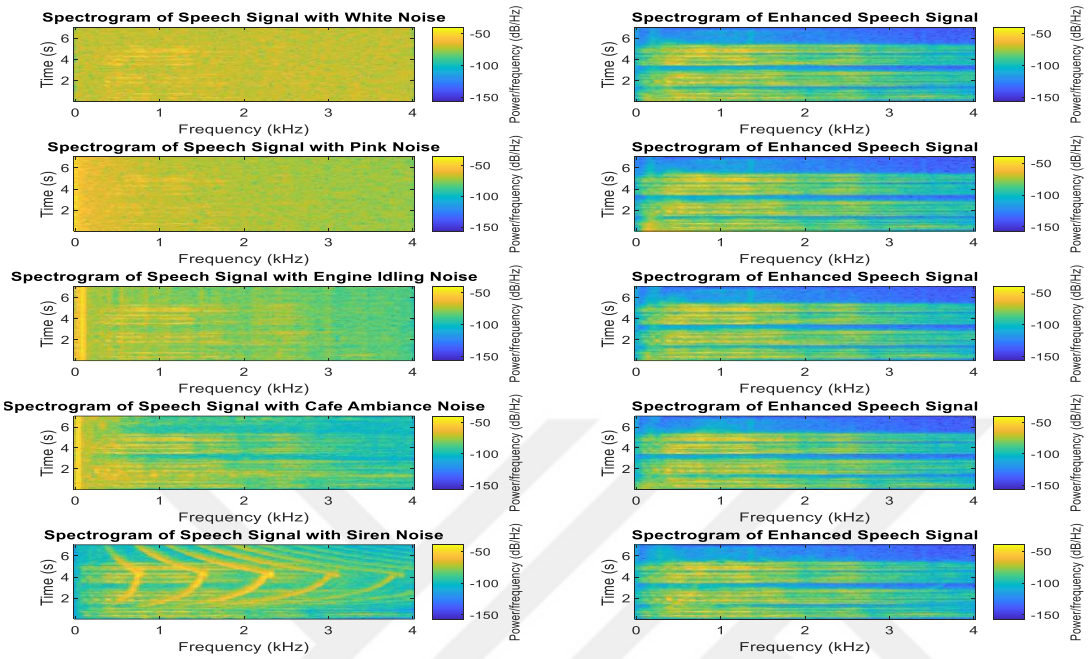
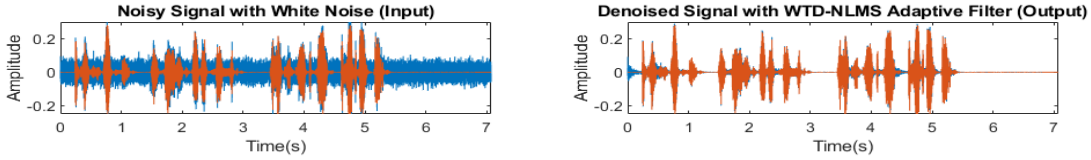


Figure 3.9. The spectrograms of input signals with various noise effect and the spectrogram of output signal obtain with proposed filter system

The spectra of the proposed WTD-LMS input and output signals are shown in Figure 3.9. When spectrograms of noisy speech signals, i.e., input signals, were examined, it was observed that the frequency-time distribution of each noise signal is different from each other, but all noise signals have a high distortion effect on speech. Furthermore, the noise effect on speech signal is not steady for each sub-band of the speech signal. Therefore, it can be said that the speech signal is entirely distorted, especially with the effect of white and pink noise. However, the output signals obtained were almost completely recovered from this disturbance. It is proof that the filter used adaptively provides successful results in all noise types. Then, obtaining the amplitude-time graphs before and after filtering was presented in Figure 3.10. The figure shows the graphs of the noisy audio signals in the first column and the enhanced versions of the signal in each row in the second column.



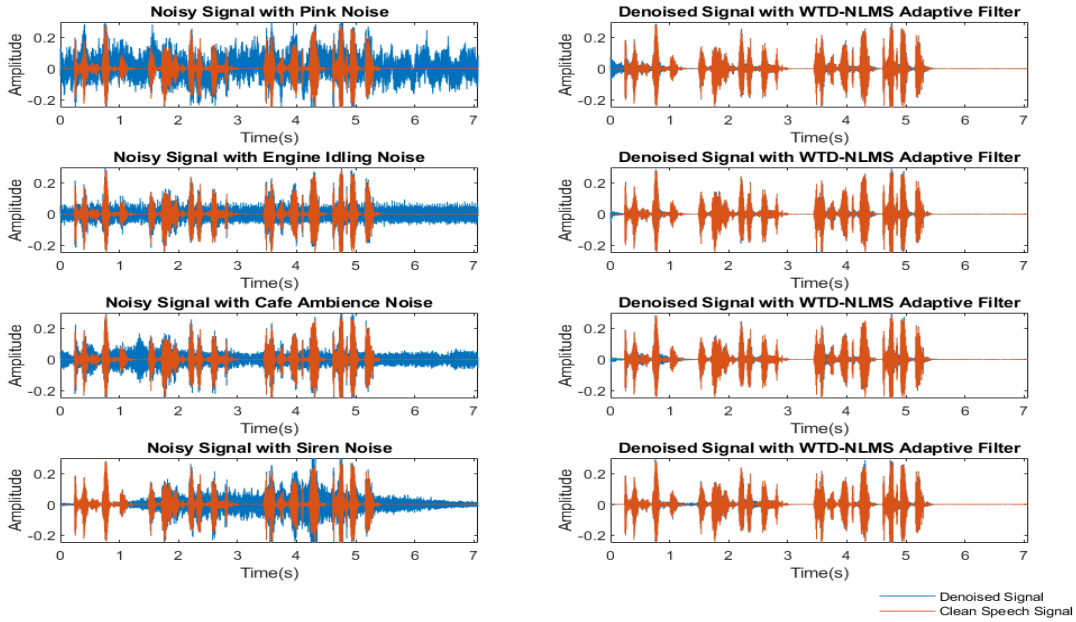


Figure 3.10. The amplitude-time graph of noisy and de-noised signal as a result of adaptive filtering

The visuals show that the filtering process is thriving despite the varying noise effect over time. The main reason for this situation is the positive contribution of filtering applied to each sub-band signal. If the filtering had been done in the time domain, the sudden changes of the noise signal over time would be affected by the adaptation of the filter coefficients, and therefore the filtering would have been less successful. Finally, the tests are repeated for 100 noisy speech signals disturbed with different noise segments selected randomly for each noise type. Then MSE, SDR, PESQ, and STOI values of audio signals were calculated. Given results are average values of the repeated test. The variance for the results presented in the table is not given because the values are too small to affect only the thousands or ten-thousands digits. The success of the method and its contribution to improving speech were observed by evaluating the pre-filtering (initial) measurement values and post-filtering (final) values of the noisy audio signal. The results obtained are as represented in Table 3.2.

Table 3.2. Evaluation Results of the Process Applied to Noisy Speech Signal with Different Noises

		MSE	**PESQ**	**STOI**	**SDR**
White Noise	Initial Value	$8.26x10^{-4}$	1.90	0.64	$5.55x10^{-15}$~ 0dB
	Final Value	$3.28x10^{-5}$	3.33	0.96	32.65 dB

Pink Noise	Initial Value	$8.26x10^{-4}$	2.16	0.68	$1.18\ x10^{-15}$~ 0dB
	Final Value	$3.87x10^{-5}$	3.25	0.95	31.46 dB
Engine Idling Noise	Initial Value	$8.26x10^{-4}$	2.48	0.79	$3.10\ x10^{-15}$~ 0dB
	Final Value	$3.22x10^{-5}$	3.35	0.96	32.50 dB
Café Ambience Noise	Initial Value	$8.26x10^{-4}$	2.58	0.76	$-9.88\ x10^{-15}$~ 0dB
	Final Value	$4.29x10^{-5}$	3.23	0.96	30.42 dB
Siren Noise	Initial Value	$8.26x10^{-4}$	2.26	0.86	$2.22\ x10^{-15}$~ 0dB
	Final Value	$2.74x10^{-5}$	3.35	0.97	33.03 dB

When the data presented in the table are examined, the most damaged speech signals regarding the intelligibility and quality of the speech signal (STOI and PESQ values, respectively) are the speech signals under the influence of white noise and pink noise. The audio signal's PESQ value under the influence of white noises increased from 1.90 to 3.33, while the STOI value was improved from 0.64 to 0.96. In general, the SDR value was improved by more than 30 dB after the improvement processes. This improvement in SDR value means that the audio signal has recovered from the high noise effect. It is observed that some noise due to the convergence delay of the filter still affects the audio signal. Also, the MSE value decreased approximately 25 times compared to the initial value. This is another proof that the signal is highly convergent to the desired signal. Consequently, the proposed adaptive filter system's success in clearing various noise signals with a high interference effect from the speech signal is admirably good.

The results obtained with the proposed method have been compared with the recent speech denoising method. In (Chiluveru & Tripathy, 2020), the application of clearing the speech signal from babble noise and factory noise was made by the WTD-Thresholding method. The results obtained in this application are presented with both PESQ and STOI criterias. The PESQ values obtained from improving the speech signal disturbed by factory noise with 0 and 5 SNR values are presented as 1.3839 and 1.8481, and the STOI values for babble noise are presented as 0.41 and 0.8. In our experiment, these results were compared with the results obtained with cafe-ambiance noise and white noise, and in this application, PESQ values obtained with speech

signals with the same SNR values were presented as 3.20 and 3.3 STOI values as 0.96. When the values are given are compared, it is possible to say that the method proposed in the article based on PESQ and STOI criterias has a superior success. The main reason for this is that the thresholding process loses the overlapping frequency component of the speech signal. In this case, the lost frequency content may affect the intelligibility of the speech. For this reason, even if the error value is reduced with the wavelet thresholding method, the intelligibility of the speech is not improved according to the given evaluation criterias.

Table 3.3 compares our best results with previous adaptive two-channel speech enhancement applications except (Özaydın & Alak, 2018) which is a type of thresholding application so it is a single-channel model.

Table 3.3. Comparison of Methods Used in Literature with Proposed Method

Method	Noise Type	Data	Outputs	
In wavelet domain LMS for approximation coefficient, Thresholding for detail coefficients in (Akhaee et al., 2005).	Noisex-92 database	Speech signal (fs=16kHz)	Initial SNRs (dB)	Final SNRs (dB)
			-5	6.04
			0	7.34
			5	7.86
In wavelet domain LMS for approximation coefficient, Wiener filter for detail coefficients in (Akhaee et al., 2005).	Noisex-92 database	Speech signal (fs=16kHz)	Initial SNRs (dB)	Final SNRs (dB)
			-5	8.48
			0	9.22
			5	9.91
Maximal Overlap Discrete Wavelet Transform (with types of thresholdings) in (Özaydın & Alak, 2018).	AWGN Restaurant Noise Car Noise	Speech signal (fs=8kHz)	Initial SNRs (dB)	Final SNRs (dB)
			AWGN→5	10.19
			Restaurant→5	7.51
			Car →5	8.44
Proposed Method (WTD-LMS)	Aircraft Engine Noise in Cockpit	Speech signal (fs=8kHz) 3 second long	Initial SNRs (dB)	Final SNRs (dB)
			0	21.33
			5	25.89
			15	27.08
			30	34.03

			Initial SNRs (dB)	Final SNRs (dB)
Proposed Method (WTD-NLMS)	AWGN Pink Noise Engine Idling Café Ambience Siren	Speech signal (fs=8kHz) 7 second long	AWGN→0 Pink Noise →0 Engine →0 Ambience→0 Siren→0	32.65 32.46 32 32.42 33.03

In Table 3.4, the contributions of deep learning-based methods, which are currently gaining momentum on speech improvement, are compared with the proposed method by considering PESQ and STOI metrics. When the data presented in the table is examined, it is observed that the proposed method offers much more successful results than deep learning-based methods, especially in improving speech intelligibility and sound quality. However, the proposed method has some disadvantages as it requires a two-channel audio recording system, and as it has known, an increasing number of recording sensors increases the success of the method for speech enhancement application. Still, it also has much less processing complexity and higher convergence speed and significantly increases the speech's intelligibility and quality than deep learning methods. To obtain a fair comparison these results are also compared with the results obtained using CNN which will be presented in the next section.

Table 3.4. Performances of Proposed Method Against State-Of-Art Based on Deep Learning Methods

Method	**STOI**	**PESQ**
DNN (Xu et al., 2015)	0.8120	2.450
TSN (Kim & Hahn, 2019)	0.8745	**2.939**
SEGAN (Pascual et al., 2017)	0.9300	2.160
DSEGAN (Phan et al., 2020)	**0.9358**	2.420
ISEGAN (Phan et al., 2020)	0.9348	2.270
MMSE-GAN (Soni et al., 2018)	0.9300	2.530
CNN-GAN (Shah et al., 2018)	0.9300	2.340
PROPOSED METHOD WTD-NLMS	**0.9615**	**3.308**

3.2. A Single-Channel Speech Enhancement Application with DNN: Speech Enhancement by CNN Using Scalograms

In the previous section, we proposed a double-channel speech enhancement application that outperforms the success of the speech enhancement applications used up to now, thanks to the contributions of WT. However, the biggest drawback of this method is that it requires two-channel recording. As explained before, this causes an extra cost and narrows down application areas. Therefore, we tried to utilize WT's impressing properties of signal examination to obtain a successful one-channel speech enhancement application.

A CNN-based speech improvement application was presented thanks to CNN's artificial learning and WT's contributions within this study's scope. This method provides a versatile and cost-effective solution to the problems arising from single-channel recording in high noise environments. We designed a CNN obtained with the skipped layers using supervised learning in the study. For this purpose, a data set containing scalograms of noisy and noiseless speech signals were obtained and used to train the neural network. In other words, one-dimensional speech data have been transformed into two-dimensional images. Thus, the success of CNN in image processing has been utilized. Similar methods using spectrogram to train CNN have been proposed in previous studies. As shown in the comparison given in Figure 2.3, scalograms obtained with CWT provide better observation for higher frequencies of the signals than spectrograms. From our point of view, using scalograms instead of spectrograms to train CNN will increase the learning ability of the system as better feature extraction is applied.

In this part of the study, the data set was first rearranged to obtain clean and noisy scalograms pairs to train the proposed network. The noise signal used to contaminate clean speech signals was the same as the noises used in the previous section. The SNR values of noisy speech signals were 0 dB to create a demanding condition in the training phase. Then hyper-parameter optimization of system parameters was accomplished. Finally, the trained network's success was tested using unseen noisy speech signals with different noise effects. The results were evaluated using the given measures to observe performance of proposed method and compared with the results obtained in previous studies from the literature.

3.2.1. Information about the Data Set

In this part of the study, firstly, the clean speech signals were polluted using white noise, pink noise, cockpit noise, engine idling noise, siren noise, café ambiance noise (containing babble noise), and all noise signals used in this study are long signals more than 3 minutes long. The spectral properties of the noise signals were presented in previous section. Noisy speech signals were obtained by adding randomly selected noise segments with the same length as the clean speech from the noise signals to the clean speech signals.

The clean speech signals needed for the test and the training phase were obtained from The Device-Recorded Voice Bank Corpus (DR-VCTK) (Sarfjoo & Yamagishi, 2018). This data set is a small sub-set of Voice Bank Corpus that includes high-quality speech signals are recorded in the quiet environment offered for particular speech processing applications and published by the University of Edinburgh School of Informatics' The Centre for Speech Technology Research (CSTR) (Sarfjoo & Yamagishi, 2018). The reason for selecting this data is that it offers speech signals with high quality, completely free from the noises caused by recording devices. The training set contains 400 different sentences from published scripts voiced by 28 different speakers with English accents. For this set, the ratio of men and women speakers is the same, and a total of 11200 speech signals with 16 kHz sampling frequency exists in the training set. Also, the test set contains 824 clean speech signals with the same sampling frequency. There is no intersection between training and test sets in terms of speakers and sentences to perform the test process fairly.

We used these noise-free speech signals to obtain noisy and clean scalograms pairs to train the proposed network. Also, the network test proceeded through the speech signal taken from the test set.

3.2.2. Pre-Process Applied to the Dataset

In order to obtain the signals to be used for training and testing the proposed network in the project, the following processes were applied respectively;

i. The sampling frequency of the speech signals obtained from the DR-VCTK data set is 16 kHz. In order to reduce the computational complexity in the learning process, the sampling frequency of the speech signals has been reduced to 8 kHz.

ii. A different segment of the noise signal with the speech signal's length was taken for each noise signal. This segment of noise was chosen at random. Then the noise signal was arranged to be SNR value of 0dB. Then, noisy signals were obtained by summing up the edited noise signal with the down-sampled speech signal. The graphs of the noisy and clean sample pair obtained for a single speech signal after the first two processes are shown in Figure 3.11.

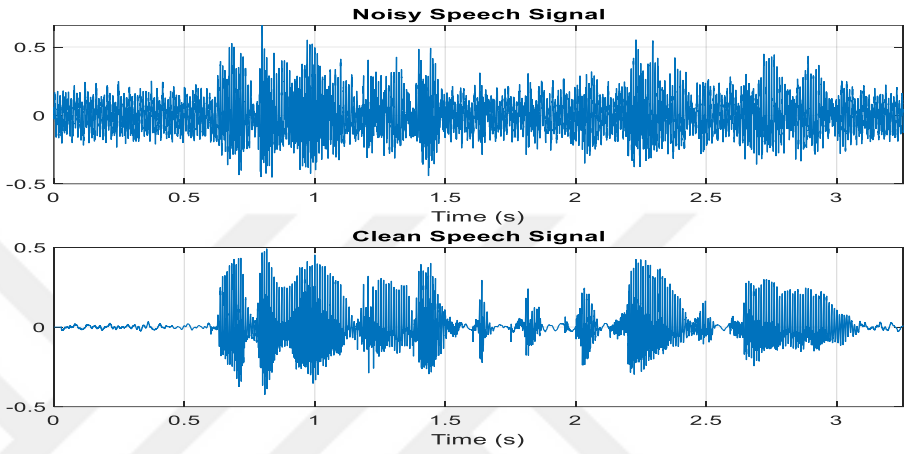
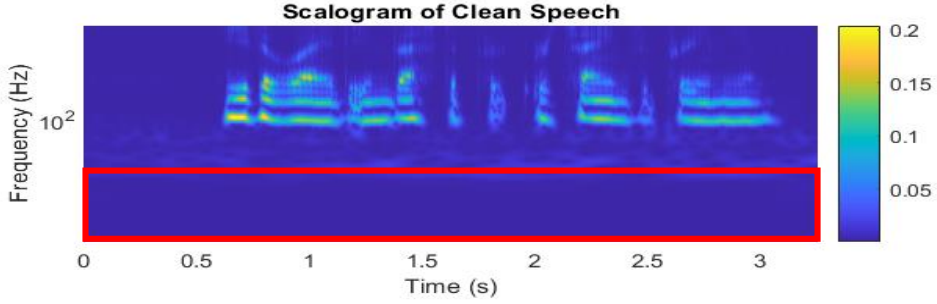


Figure 3.11. Amplitude time graph of a noisy (input), a noiseless (desired output) speech signals

iii. After this process, scalograms containing the time-frequency distribution of both noisy and noiseless audio signals were obtained by using CWT. First, the complex Morlet wavelets, a type of complex-valued wavelet helpful to observe signals with time-varying amplitude and frequencies, were preferred to calculate wavelet coefficients. Then, scalograms were obtained by taking the absolute value of the wavelet coefficient. In this way, one-dimensional audio signals have been transformed into 2-dimensional time-frequency visuals containing the essential feature of the signals. The scalograms pairs obtained for given sample pair are shown in Figure 3.12.



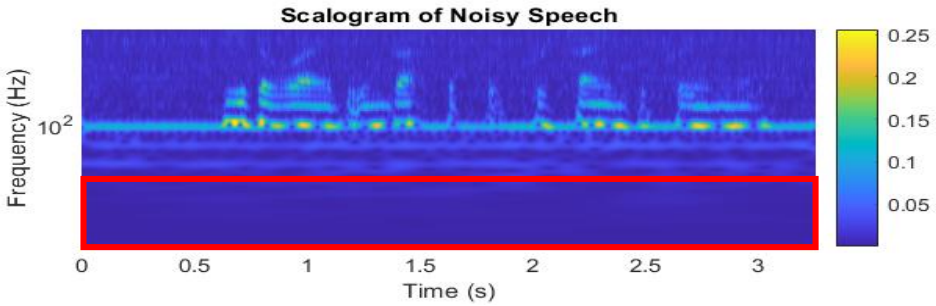


Figure 3.12. Scalograms of a noisy (input), a noiseless (desired output) speech signals obtained with CWT

As seen from the visual scalograms offers good resolution for speech signal. However, the CWTs infinite scaling and shifting process causes redundant information, especially for lower frequency values marked with red rectangulars. Therefore, scalograms are windowed by clipping frequency values lower than 80 Hz to eliminate this redundant information. This process does not remarkably affect speech signals' intelligibility and quality, and it helps reduce noise effect, especially for some low band noises. Furthermore, the size of data nearly halved by this windowing operation reduces the computational complexity of the system. The windowed scalograms are presented in figure 3.13.

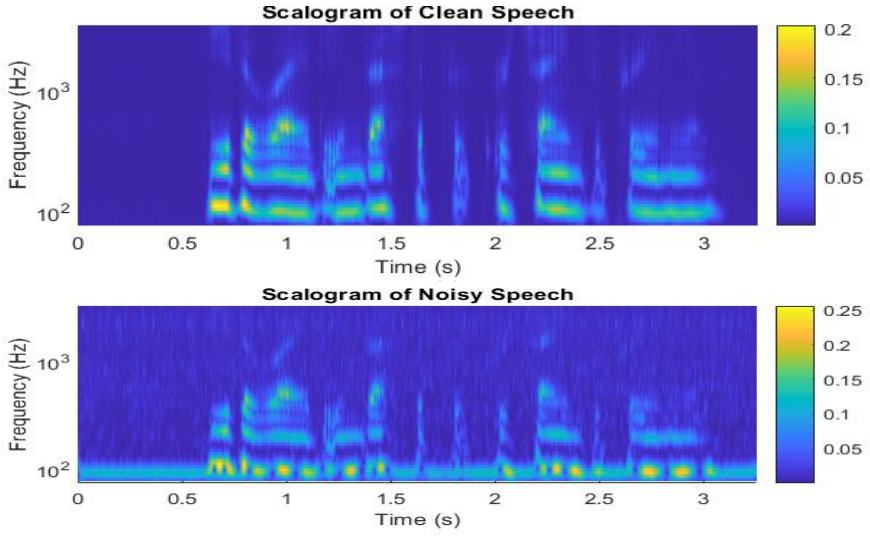


Figure 3.13. Windowed Scalograms of a noisy (input), a noiseless (desired output) speech signals obtained with clipping frequencies below 80 Hz.

The scalograms of speech signals have the size of 55 x N for each speech signal, where N is equal to the sample number of the speech signals. After

segmentation was applied to scalograms, input and desired output pairs were obtained for the training process. Desired output samples were equal to each time segment of clean speech signal's scalograms with the size of 55 x 1. Input samples were obtained taking 16 consecutive time delays of each segment, so the input size is 55x16. A sample of scalograms segments used to train network as input and desired output pairs are shown in Figure 3.14. These small images contain features of the speech signals. For every 16 consecutive segments of the noisy speech signal, one segment of clean speech is given to the system. With the help of CNN, it is tried to map these 16 noisy segments into one clean segment. Thus, by combining cleaned segments, noise-free speech scalograms could be obtained.

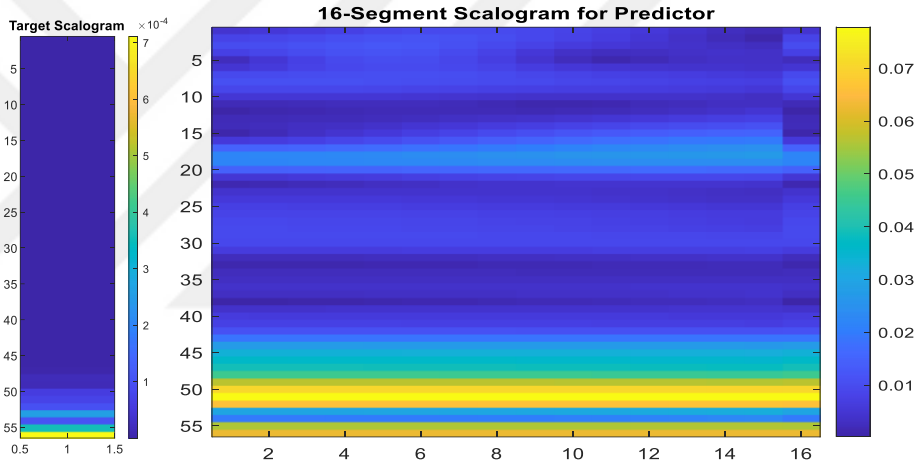


Figure 3.14. The segment taken from noisy and clean spectrum to be used as target (desired output) and predictors (input) in training data set.

iv. Finally, all the scalograms values arranged as targets and predictors are standardized with a mean of zero variance of 1 which called as normal distribution scaling. This is a type of data scaling process, thanks to this process, the parameters of the network to be trained will take more standard values, which will increase the convergence speed of the system and ensure that the system remains more stable during the learning phase (Shi et al., 2018).

All these steps explained were repeated for all speech signals in the data set. As a result, a sample set consisting of input- desired output pairs with the size of 55x1 desired

output and 55x16 the size of the input to be used for training the network. The number of training pairs obtained from each speech signal is equal to the number of samples in the speech. Training of the proposed network model was performed using this set. Nearly 250 different speech signals were selected randomly to obtain the training set due to the computer's capabilities where the learning was performed. 1,867,558 training pairs were obtained and used using these speech signals. 5% of the input and desired output pairs obtained from these speech signals are reserved for validation to calculate the error during training and avoid overfitting (56,027 sample pairs). In the test phase, speech signals unseen and untrained with the network selected from the test set are used. The test process was repeated with 500 noisy speech signals, and the results were evaluated with selected measures.

3.2.3. Proposed Network and Implementation

Within the scope of this study, it is aimed to remove noise from speech signals. For this purpose, we tried to extract features of a clean speech signal from the noisy speech's scalogram by the CNN. To accomplish it, a simplified CNN network model obtained by skipping some layers of CNN was used, a network that has been tested and accepted with success with previous studies such as (Park & Lee, 2017 and Shi et al., 2018). The general diagram of this type of CNN network is visualized in Figure 3.15.

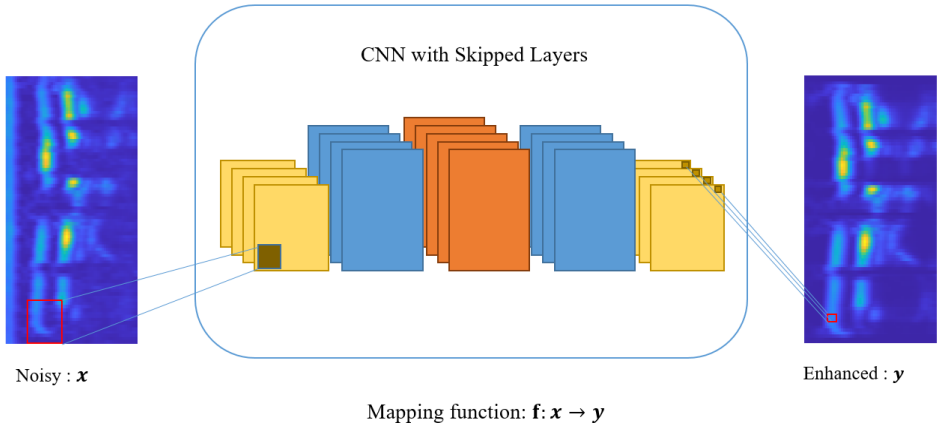


Figure 3.15. General schemes of speech enhancement application with skipped layers CNNs (Park & Lee, 2017).

As seen in the figure, in this CNN architecture, the pooling, fully-connected, and flattening layers, which are the classical CNN layers described in the previous section,

are not used. The pooling layer is not used because the data as input and output in the project is a visual that includes speech frequency-time contents. Furthermore, during the pooling phase, any possible frequency-time component loss or positional information deterioration while reducing the sample will cause deterioration in speech signals. It has also been shown in previous studies. Also, the system's success in noise-cleaning does not change due to removing these layers. Besides, since the system parameters are reduced, the system's convergence speed increases.

The operations carried out during this study can be discussed under three main headings. These topics can be listed as preparing and splitting the data set with pre-processing, creating and training the neural network model, testing the trained neural network and interpreting its performance. These stages of the study are presented in figure 3.16.

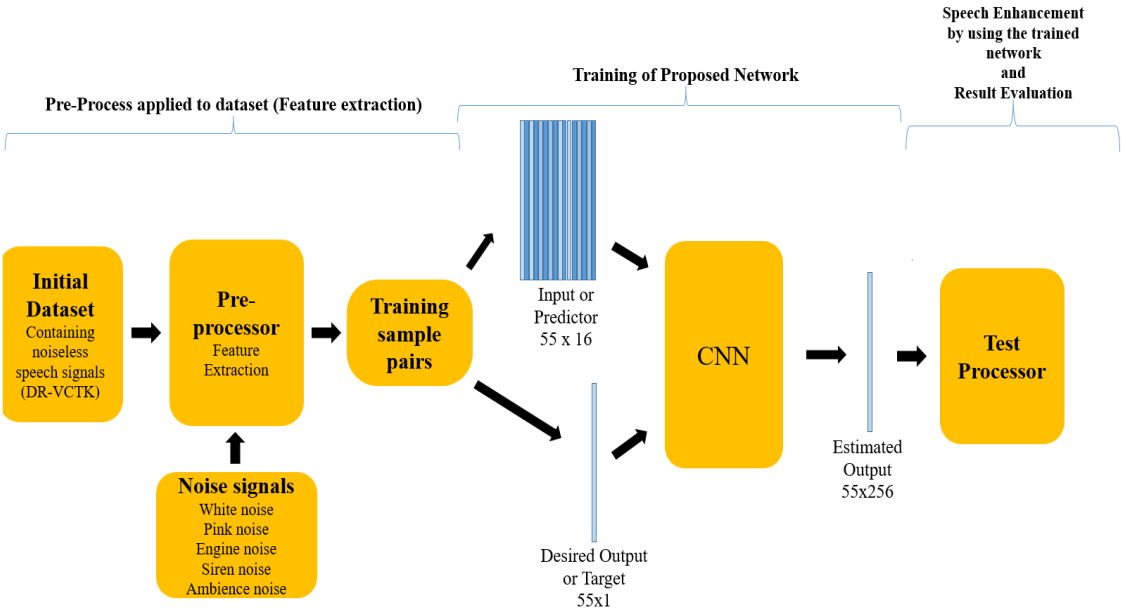


Figure 3.16. The main stages of the study of speech enhancement with CNN

All operations carried out in pre-process applied to dataset were presented in the previous section with examples and visuals. The processes applied to test the network's success were summarized by the "Test Processor" in the diagram. The detail about the inside of this processor will be given in Figure 3.20.

Obtained training sample pairs after pre-process stage that contain the predictor and the target scalograms were used to train the proposed CNN network. In the proposed CNN model the network architecture consists entirely of combinations of

convolutional and activation layers. The layers of the network architecture used are as shown in Figure 3.17.

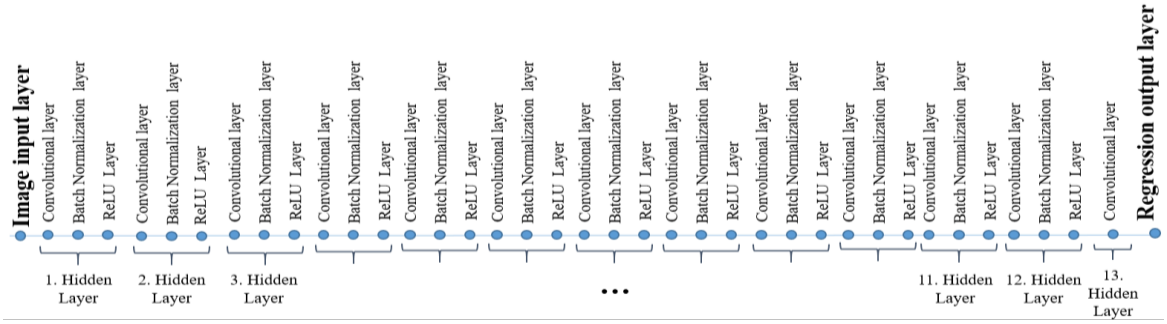


Figure 3.17. Architecture of proposed CNN model

As seen in figure 3.17, an input, an output layer, and 13 hidden convolutional layers are used in the selected model. Each convolutional layer, except the last convolutional layer, was combined with the batch normalization layer that performs normalization for each mini-batch set. In this layer, the input values for each mini-batch set are normalized to a mean of 0 and a variance of 1. Thus, it aims to reduce the network's sensitivity to the initial values and keep the system stable during the training. Besides, a non-linear activation process was carried out in the ReLU layers. The filters used in these layers and the number of filters in each layer are shown in Figure 3.18.

		Layers	Information about the layers
	1	Image Input	55×16×1 images with 'zerocenter' normalization
1. Hidden L.	2	Convolution	108 5×8 convolutions with stride [1 1] and padding 'same'
	3	Batch Normalization	Batch normalization
	4	ReLU	ReLU
2. Hidden L.	5	Convolution	96 3×1 convolutions with stride [1 1] and padding 'same'
	6	Batch Normalization	Batch normalization
	7	ReLU	ReLU
3. Hidden L.	8	Convolution	72 3×1 convolutions with stride [1 1] and padding 'same'
	9	Batch Normalization	Batch normalization
	10	ReLU	ReLU
4. Hidden L.	11	Convolution	72 5×1 convolutions with stride [1 1] and padding 'same'
	12	Batch Normalization	Batch normalization
	13	ReLU	ReLU
5. Hidden L.	14	Convolution	96 3×1 convolutions with stride [1 1] and padding 'same'
	15	Batch Normalization	Batch normalization
	16	ReLU	ReLU
6. Hidden L.	17	Convolution	72 3×1 convolutions with stride [1 1] and padding 'same'
	18	Batch Normalization	Batch normalization
	19	ReLU	ReLU
7. Hidden L.	20	Convolution	72 5×1 convolutions with stride [1 1] and padding 'same'
	21	Batch Normalization	Batch normalization
	22	ReLU	ReLU
8. Hidden L.	23	Convolution	96 3×1 convolutions with stride [1 1] and padding 'same'
	24	Batch Normalization	Batch normalization
	25	ReLU	ReLU
9. Hidden L.	26	Convolution	72 3×1 convolutions with stride [1 1] and padding 'same'
	27	Batch Normalization	Batch normalization
	28	ReLU	ReLU
10. Hidden L.	29	Convolution	72 5×1 convolutions with stride [1 1] and padding 'same'
	30	Batch Normalization	Batch normalization
	31	ReLU	ReLU
11. Hidden L.	32	Convolution	72 3×1 convolutions with stride [1 1] and padding 'same'
	33	Batch Normalization	Batch normalization
	34	ReLU	ReLU
12. Hidden L.	35	Convolution	56 5×1 convolutions with stride [1 1] and padding 'same'
	36	Batch Normalization	Batch normalization
	37	ReLU	ReLU
13. Hidden L.	38	Convolution	1 55×1 convolutions with stride [1 1] and padding 'same'
	39	Regression Output	mean-squared-error

Figure 3.18. Outline of the CNN architecture

In figure 3.18, the dimensions of the filters used in each convolutional layer and the number of filters are given. For instance, 108 filters with the size of 5x8 were used in the first hidden layer. The figure contains the related information about all hidden layers, as explained in the example. Furthermore, information about padding and stride factors for each layer is also provided in the figure. Padding has been applied to keep the input and output sizes the same in the conventional layer, and the stride factor is determined as 1. For this network, the total number of weights to be learned after training is 255,656 This network structure was developed heuristically to obtain best performance for the specific problem.

After this process, the initial values were determined for the start of the training. The mini-batch training method was chosen for training, and the mini-batch size was chosen as 64. As a result of the experiments made with different mini-batch sizes, this size was chosen because the most successful results were obtained with this mini-batch size. During the training, ADAM optimization was preferred as the optimization algorithm. The initial learning rate was chosen as 0.003. Besides, it was planned to reduce the learning rate by 0.6 after each mini-batch. Thus, the system was aimed to remain stable. The automatic-early stopping was not used in the system. Instead, a validation error was calculated for every 3500 iterations. While the system was being trained, the validation error was monitored, and if it increased, it was planned to stop the learning to prevent overfitting. However, due to the device's technical inadequacies in which it was applied, the learning process was determined to make a maximum of 16 epochs. It was not observed that the system was overfitting within these 16 epochs. The graph showing the change of the error function according to the number of iterations obtained during the training of the network is shown in figure 3.19.

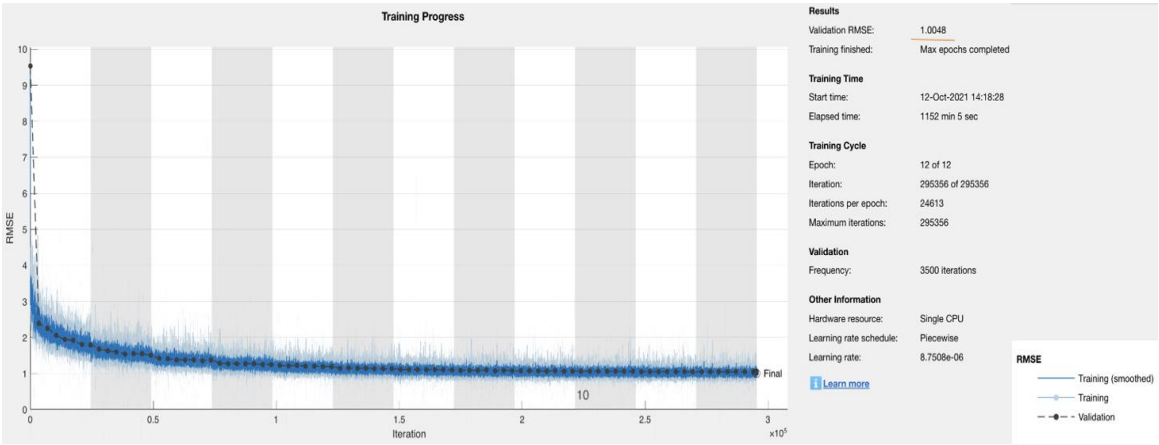


Figure 3.19. The graph of RMSE change during training progress

The first two stages of study, which are pre-process applied to the data set and training of the proposed network, were explained in detail. After the training process, the parameters of the network were estimated. This parameter determines the mapping function, which maps noisy scalograms segments into clean scalograms segments. Then, this network was used to remove noise from speech signals without knowing any information about noise data. The diagram showing how the trained network was used to enhance speech signal and tested is illustrated in figure 3.20.

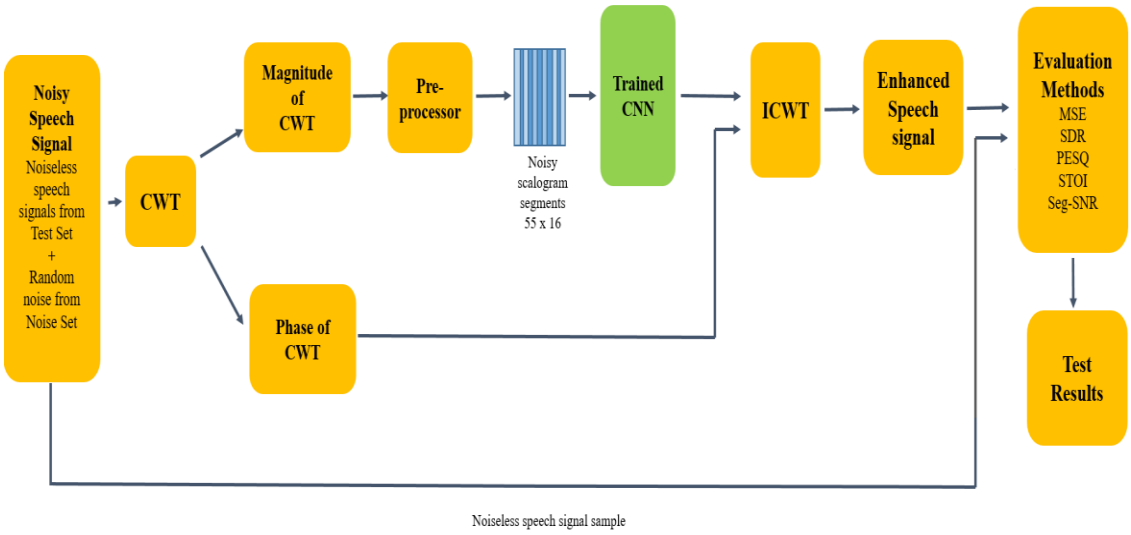


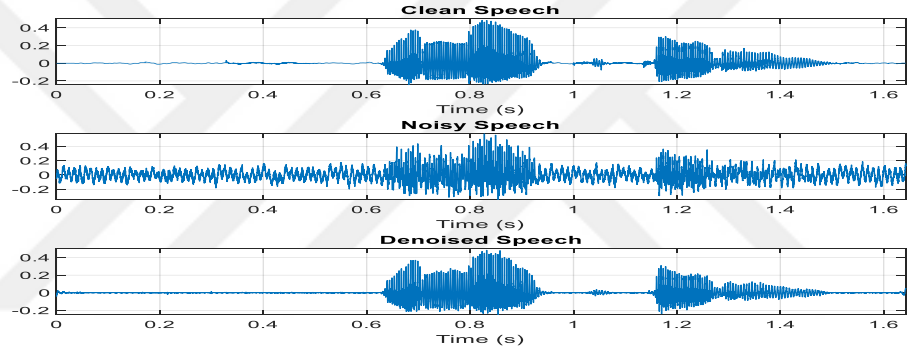
Figure 3.20. The diagram illustrates processes applied in the Test Processor, which enhances noisy speech signals by proposed CNN and obtaining test results.

As is known, only the magnitude spectrogram of noise and noiseless signals was used in training phase. The main reason for this is that the human ear is insensitive to phase changes smaller than 45 degrees (Park & Lee, 2017), and the distortion in phase mainly contains information about the speakers' position. In this study, our focus point is commonly enhancing speech signals quality and intelligibility of speech signal. Based on the notion that the phase scalograms has no discernible effect on speech intelligibility, no process has been applied for phase spectrogram in order to simplify the system. The first operation in testing the network is the noisy signal's pre-processing. These operations are the same as in section 3.2.2. Here, the differences are that the phase scalogram of the noisy signal was obtained after CWT and only predictors were calculated. Then, this phase scalogram were used to calculate the ICWT. The noisy magnitude scalogram was processed by the network, and finally, a noise-free signal was obtained as a result of the ICWT calculation.

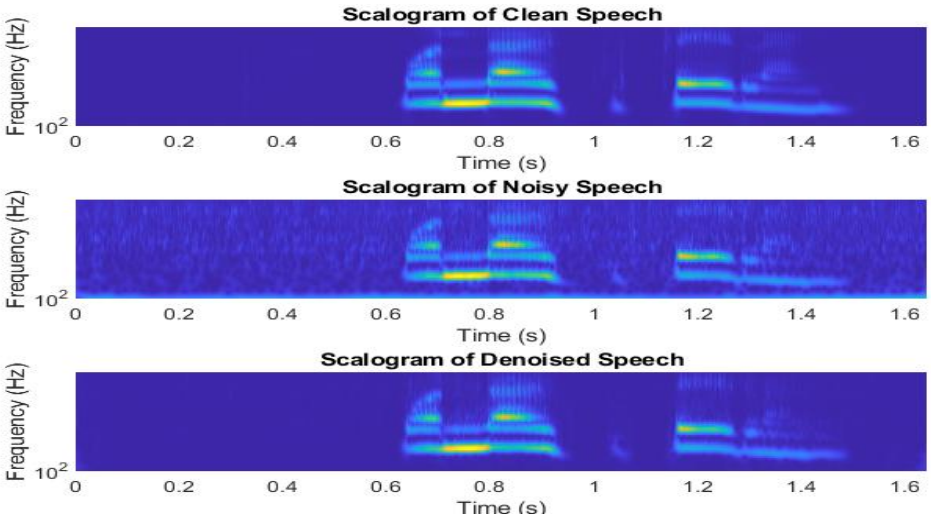
3.2.4. Results and Discussion of the Study

In the test phase, the success of the proposed CNN model on speech enhancement was investigated. For this purpose, speech signals were randomly selected from the test data set. Then, these speech signals were corrupted by the random noise selected from the noise data set, and the initial SNR value of noisy speech signals was arranged as 0 dB. After this stage, noisy speech signals under the unknown noise effect were obtained. Finally, the processes shown in figure 3.20 were applied to noisy speech signals, and enhanced speech signals were obtained. This section will present the results obtained in the test phase.

Figure 3.21 shows the graphical results obtained for a sample noisy speech signal using the trained network.



(a)



(b)

Figure 3.21. Test results for a random speech chosen from the test data set (a) the time-amplitude graph, (b) The scalograms

When the sample results in the graphs are examined, it is easily seen that the network gives acceptably successful results in term of speech enhancement. The noisy speech signal under effect of engine idling noise was highly distorted by noise. However, with the help of proposed network, it was mostly saved from the noise. This fact can be easily observed on both spectrogram and amplitude-time graphs.

The visual results are not enough to prove the success of the method. Therefore, the results evaluated using objective speech enhancement measures are given in Table 3.5. These results were obtained to observe the network's success in enhancing noisy speech signals under the effect of different noises. The enhancement process was repeated 100 times to obtain the presented average results. For each case, speech signals were dirted by different segments of selected noise. So, this result refers to the average success of the network for each noise type.

Table 3.5. The Evaluation of the Trained CNN in Enhancing the Noisy Speech Signals with 0 dB SNR Under the Effect of Different Noises

Noise Types		MSE	SDR	STOI	PESQ
Siren Noise	Initial	$5.2x10^{-3}$	~0 dB	0.71	1.75
	Final	$4.5x10^{-4}$	24.519 dB	0.83	2.36
Engine Idling Noise	Initial	$5.2x10^{-3}$	~0 dB	0.75	1.95
	Final	$3.8x10^{-4}$	27.392 dB	0.87	2.56
Café Ambience Noise	Initial	$4.5x10^{-3}$	~0 dB	0.80	2.30
	Final	$2.7x10^{-4}$	29.326 dB	0.88	2.73
White Noise	Initial	$4.2x10^{-3}$	~0 dB	0.66	1.45
	Final	$5.4x10^{-4}$	23.690 dB	0.77	2.11
Pink Noise	Initial	$4.6x10^{-3}$	~0 dB	0.7	1.54
	Final	$5.9x10^{-4}$	24.151 dB	0.81	2.31

When the result given in Table 3.5 are examined, it can be said that acceptable improvement was achieved for each noise type using the proposed network. For example, the final MSE values are ten times lower than the initial values, and the SDR values are improved by more than 23 dB. Furthermore, a good improvement in STOI and PESQ was also provided. In terms of these measures, the best results were

achieved for café ambiance noise, and the worst was obtained using white noise. Since corruption given by white noise to speech signal was the highest, the ensured improvement for both noise types is nearly the same.

Then, the general success of the network in improving noisy speech signals with selected noises was measured using 500 randomly selected speech signals from the test set. The number of noisy speech signals distorted with each noise type was equal in this stage. The average results are presented in Table 3.6. The improvement achieved in each evaluation measure is also given in the table. The variance of the measurements was in the order of 10^{-3} and omitted in the table.

Table 3.6. Results Obtained by Testing the Network with 500 Noisy Speeches (SNR= 0db)

MSE		**SDR(dB)**		**STOI**		**PESQ**		**Seg-SNR(dB)**		**LSD**	
Initial	Final	Initial	Final	Initial	Final	Initial	Final	Initial	Final	Initial	Final
5.1 $\boldsymbol{x10^{-3}}$	4.3 $x10^{-4}$	~0	26.09	0.71	0.84	1.61	2.45	-4.59	3.54	2.63	1.47
Improvement		Improvement		Improvement		Improvement		Improvement		Improvement	
reduction more than **10 times**		**+26 dB**		**+0.13**		**+0.84**		**+8.13 dB**		**+1.16**	

*LSD: Log-squared Distance

Finally, results are compared with results of some novel studies published recent years. The comparative results are presented in Table 3.7.

Table 3.7 Performance Comparison of the Proposed CNN Model with The Previously Presented Methods

Method	**Noise Type**	**Data**	**Outputs**					
DFT + DNN (Xu et al., 2015)	15 noise from NOISEX-92 (white, pink, car, siren, engine, restaurant …)	TIMIT	PESQ		STOI		Seg-SNR(dB)	
			Initial	Final	Initial	Final	Initial	Final
			1.91	2.74	0.7	0.82	-4.59	-1.5
			+0.83		**+0.12**		**+3.09 dB**	
FFT + TSN (Kim & Hahn, 2019)	15 noise from NOISEX-92 (white, pink, car, siren, engine, restaurant …)	TIMIT	PESQ		STOI		Seg-SNR(dB)	
			Initial	Final	Initial	Final	Initial	Final
			1.92	2.63	0.7	0.81	-5.63	1.037
			+0.71		+0.11		+6.67 dB	

Raw speech + SEGAN (Pascual et al., 2017)	10 Noise (2 artificial and 8 from the Demand database	Voice Bank corpus (VCTK)	PESQ		STOI		Seg-SNR(dB)	
			Initial	Final	Initial	Final	Initial	Final
			1.97	2.16	-	-	1.68	7.73
			+0.19		-		6.05	
Gammatone spectrum + GAN (Soni et al., 2018)	10 noise (2 artificial and 8 from Demand database)	Voice Bank corpus (VCTK)	PESQ		STOI		Seg-SNR(dB)	
			Initial	Final	Initial	Final	Initial	Final
			1.91	2.53	0.91	0.93	-	-
			+0.56		+0.02		-	
T-F Mask(Gammatone spectrum) + CNN + GAN(Shah et al., 2018)	10 noise (2 artificial and 8 from Demand database	Voice Bank corpus (VCTK)	PESQ		STOI		Seg-SNR(dB)	
			Initial	Final	Initial	Final	Initial	Final
			1.97	2.34	0.91	0.93	-	-
			+0.37		+0.02		-	
Wavelet Scalogram + CNN (Proposed Method)	5 noise from www.freesound.com (white, pink, siren, babble + restaurant (café), engine idling,)	Voice Bank corpus (DR-VCTK)	PESQ		STOI		Seg-SNR(dB)	
			Initial	Final	Initial	Final	Initial	Final
			1.61	2.45	0.71	0.84	-4.59	3.54
			+0.84		**+0.13**		**+8.13 dB**	

As can be seen from the table, the best results in given measures received by (Xu et al., 2015) study and our results are slightly better than this study. However, since we used a limited number of speech signals in the study because of technical deficiencies, the results indicate that the proposed model is acceptably successful in speech enhancement. It is anticipated that the success of the given method can be increased with further studies, such as the increasing number of samples and epochs, enhancement applied to the phase of scalograms.

CHAPTER 4
CONCLUSIONS AND FUTURE RESEARCH

Speech enhancement applications are commonly used pre or post-process in many areas where speech signals are used. The primary purpose of these applications is to reduce noise on speech signals to increase the quality and intelligibility of the speech. In this thesis, we tried to increase the success of speech enhancement applications done so far by using the excellent performance of WT in terms of signal analysis. For this purpose, we offered two new approaches for single and double-channel speech enhancement.

The double channel speech enhancement application proposed in the thesis was an application of adaptive filtering in the wavelet transform domain. Adaptive filters are preferred for statistically changing signals and environments, in which the filter coefficients are determined according to the statistical properties of the input signal. The most preferred adaptive filters use the LMS algorithm. The main reason for this is that the LMS algorithm is easy to apply and has good convergence features. However, there are still difficulties in applying adaptive filters in the time domain for large data sets. When applying adaptive filters for large data sets, computational complexity increases and convergence speed decreases. Using the transfer domain increases the convergence speed of the adaptive filter and reduce the processing complexity.

In our transform domain adaptive filter, the DWT is first applied to the input signal. In this way, the signal is divided into orthogonal sub-signals, thus increasing the de-correlation of the input signal. In other words, the eigenvalue distribution of the auto-correlation matrix of the input signal is approximated by 1. This is the case where the LMS algorithm has a maximum rate of convergence. Then, adaptive filters are applied to all sub-signals in parallel branches. The output of the adaptive filter is obtained by passing the obtained output signals to the time domain as a result of inverse transformation. The main superiority of WT over other transforms its lower processing complexity, easier applicability and offering a better time-frequency resolution than FT.

Two basic implementations were made in this project. In the first experiment, the purpose of the application was to compare the success of NLMS and LMS algorithms in the proposed method to show the contribution of normalization integrated into the NLMS algorithm to the method's success. For this purpose, it was tried to recover the speech signal from the ambient noise effect to ensure that voice communication can be performed smoothly in the hands-free mode. The communication area is preferred as the aircraft cockpit, a unique area of this application. Therefore, the audio signal containing a high aircraft engine noise as the ambient noise recorded in the aircraft cockpit was accepted as the input signal of the adaptive filter. The value of SNR is arranged as 0 dB. At this stage, it is aimed to create a challenging condition for the convergence speed of the filter using a short speech signal. Adaptive filters using WTD-LMS and WTDN-NLMS algorithms improved this noisy signal. The variables of this system, such as filter order, decomposition level in DWT, step size of the adaptive algorithm, have been selected to give optimum results due to various investigations and applications. Also, different mother wavelet functions are used to detect the best mother wavelet function. It is observed that the NLMS algorithm is more successful than the results of the filters applied for all sub-signals obtained with DWT. Likewise, the evaluation criterias (MSE, SDR, PESQ, STOI) calculated with the output signals also showed that the convergence speed and ratio of the NLMS algorithm were better. So the NLMS algorithm is significantly more successful than the LMS algorithm. The main reason for this situation is that the energy of each sub-signal is not the same. The normalization process helps increase the filter's convergence speed as it helps regulate the eigenvalue distribution of the autocorrelation matrix of the input signal. Additionally, the speech signal used in the study is in the English language, so dmey (discrete approximation of Meyer function) offered the best results.

In the second experiment, the adaptive filter's success using the WTD-NLMS algorithm and dmey mother wavelet function, which gave successful results in the first experiment, in improving noisy speech signals with noises that have different frequency-time characteristics were tested. For this purpose, longer speech signals have been contaminated using noise signals such as white noise, pink noise, engine relay noise, cafe ambiance noise, siren noise, which can be frequently affected by speech signals in various applications. The filter parameters were kept the same as in

the previous stage. When the results obtained are examined, it is seen that the convergence speed and ratio of the proposed filter system are very high in filtering the noises with different statistical characteristics. Much more successful results were obtained at this stage than applying the adaptive algorithm in the time domain. It is challenging to clean suddenly changing noise signals with adaptive filters applied in the time domain, such as siren sound noise. The complete change of the noise signal in the time required to adapt the filter coefficients reduces the convergence rate of the filter. However, in this method, applying the filter separately to sub-band signals contributes to reducing this sudden variability feature of the noise in the time domain, thus increasing the convergence rate of the adaptive algorithm.

The performance limits of the adaptive filter proposed in the application were tested through the speech improvement application. In addition to the enhancement achieved for speech signals, the convergence speed and success of the adaptive filters increased. Also, reducing the computational complexity has made it easier to apply the filter to large data sets or signals with many samples. Especially in the second part of the study, the proposed method provided successful results thanks to its rapid adaptation ability. Based on these results, it is predicted that the proposed method will be successful in digital signal processing and filtering applications where two-channel recording systems are used, such as detecting the fetal ECG from ECG contaminated by the mother's ECG.

Based on the results obtained in this study, it is possible to say that the WTD-NLMS algorithm is a successful method in signal improvement. In addition, the most significant advantages of the method are that the convergence rate is higher, and the computational complexity is less than the applications in the time domain. However, the method still has its shortcomings. Ambient noise as a reference signal must be known or estimated precisely to improve the signal, and this is not possible in all environments. The most significant disadvantage of double-channel speech enhancement is that it requires a dual-channel system which may add an extra cost to the system, and the application area is narrow. Many studies have been conducted on single-channel speech enhancement applications to eliminate these deficiencies.

The single-channel speech enhancement method proposed in the thesis is a fully convolutional neural network that uses scalograms as input. The CNNs are commonly preferred in image processing applications because it is beneficial to detect a feature

on the image with the 2-D convolution process. This study aimed to take advantage of CNN's outperforming properties in image processing by converting speech signals (1-D signal) into scalograms (2-D signal) with the CWT. Many time-frequency(T-F) transformation methods have been combined with CNN in the literature. However, we prefer WT because of its better T-F resolution with the multi-resolution property. From our point of view, this feature, which provides better monitoring of the signal, will increase the learning capacity of the deep learning network used, as it will be more successful in extracting the signal features. This will result in a more successful speech improvement application. Also, the computational complexity of WT is less than other transformation methods.

This study aims to purify the speech signal, which is exposed to various noises (white, pink, chatter, restaurant, engine idle) with the help of CNN's learning feature, from the related noise without noise information. In this method, noisy and noiseless speech sound pairs must be used in the training phase. The noiseless speech sounds used in the method were taken from the Voice Bank corpus (DR-VCKT) data set, which is frequently preferred in speech improvement applications. The applications in this study are carried out under three main parts, pre-processing applied to the data set, training of the proposed network, and test of the network in speech enhancement application.

In the pre-processing phase of the study, firstly, clean speech signals from the dataset were corrupted by different noise signals. The SNR value of noisy speech signals was 0 dB, one of the most challenging conditions for speech enhancement application. Then, the CWT of the speech signals was calculated. Scalograms can be defined as magnitude information of CWT. After windowing and splitting obtained scalograms, noisy and clean scalogram segments pairs referred to the target and predictor in the training phase were obtained. In the second part, the proposed CNN network was created and trained with obtained target and predictor pairs. The CNN network has 13 hidden layers with an input and an output layer. All hidden layers of the network were convolutional layers combined with activation and normalization. This network was trained with 1,867,558 training sample pairs, and 256,656 parameters tried to be estimated. An increasing number of training sample pairs will increase the learning or estimation ability in this stage. This sample pairs number is the highest number that can be reached by the computer where the application was performed.

Finally, the trained network was tested with the unseen noisy speech signals. As the proposed network was trained using only magnitude information of the CWTs, the magnitude of CWT is enhanced with CNN. The phase information of the noisy speech signal is used to reconstruct the speech signal in the time domain. In this phase, the noisy phase information does not affect the success of the enhancement process too much. However, we know that it will limit success at a point. In the test phase, firstly, the performance of the proposed method to reduce the effect of each noise type was measured with evaluation criterias. It was observed that the maximum value achieved for each noise type was not the same. Since the harmful effects (initial values) were not the same for each noise type, it can be said that the improvement achieved was equal. So, we can say that the proposed network has a stable improvement ability for each noise type. Also, the network's general performance was measured and compared with the studies based on the deep learning method. It was seen that the performance of the proposed methods was better. So, it can be said that a successful single-channel speech enhancement method was obtained with help of wavelet transform.

When single-channel application was compared with the double-channel, however, it is observed that better improvements were obtained at the expense of increasing the cost and reducing the convergence speed. In further studies, the success of the single-channel system might be tried to improve by increasing pair samples, number epoch, and enhancing phase information with additional learning methods.

REFERENCES

Abd El-Fattah, M. A., Dessouky, M. I., Abbas, A. M., Diab, S. M., El-Rabaie, E.-S. M., Al-Nuaimy, W., Alshebeili, S. A., & Abd El-samie, F. E. (2013). Speech enhancement with an adaptive Wiener filter. International Journal of Speech Technology, 17(1), 53–64. https://doi.org/10.1007/s10772-013-9205-5

Addison, P. S. (2002). The Illustrated Wavelet Transform Handbook, Introductory Theory and Applications in Science, Engineering, Medicine and Finance. Institute of Physics.

Akhaee, M. A., Ameri, A., & Marvasti, F. A. (n.d.). Speech enhancement by adaptive noise cancellation in the wavelet domain. *2005 5th International Conference on Information Communications & Signal Processing*. https://doi.org/10.1109/icics.2005.1689142

Al-Akhras, M., Daqrouq, K., & Al-Qawasmi, A. R. (2010). Perceptual Evaluation of Speech enhancement. 2010 7th International Multi- Conference on Systems, Signals and Devices. https://doi.org/10.1109/ssd.2010.5585514

Attallah, S. (2000). The wavelet transform-domain LMS algorithm: A more practical approach. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *47*(3), 209–213. https://doi.org/10.1109/82.826747

Beaufays, F. (1995). Transform-domain Adaptive Filters: An analytical approach. *IEEE Transactions on Signal Processing*, *43*(2), 422–431. https://doi.org/10.1109/78.348125

Borisagar, K. R., & Kulkarn, G. R. (2010). Simulation and Comparative Analysis of LMS and RLS Algorithms Using Real Time Speech Input Signal. *Global Journal of Researches in Engineering*, *10*(5), (pp. 44-47).

Borisagar, K. R., Thanki, R. M., & Sedani, B. S. (2019). *Speech enhancement techniques for digital hearing aids*. Springer International Publishing.

Brownlee, J. (2020, April 16). *How do convolutional layers work in Deep Learning Neural Networks?* Machine Learning Mastery. Retrieved December 1, 2021, from https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/.

Burrus, C. S., Gopinath, R. A., & Guo, H. (1998). *Introduction to wavelets and wavelet transforms: A Primer*. Prentice Hall.

Chaudhari, A., &; Dhonde, S. B. (2015). A review on Speech Enhancement Techniques. 2015 International Conference on Pervasive Computing (ICPC). https://doi.org/10.1109/pervasive.2015.7087096

Chiluveru, S. R., & Tripathy, M. (2020). Speech enhancement using a variable level decomposition DWT. *National Academy Science Letters*, *44*(3), 239–242. https://doi.org/10.1007/s40009-020-00983-3

Davis, G. M. (2002). *Noise reduction in speech applications*. CRC Press.

Dentino, M., McCool, J., & Widrow, B. (1978). Adaptive filtering in the frequency domain. *Proceedings of the IEEE*, *66*(12), 1658–1659. https://doi.org/10.1109/proc.1978.11177

Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation by Wavelet Shrinkage. *Biometrika*, *81*(3), 425–455. https://doi.org/10.1093/biomet/81.3.425

Ergen, B. (2012). Signal and image denoising using wavelet transform. *Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology*. https://doi.org/10.5772/36434

Ergin, T. (2020, February 22). *Convolutional Neural Network (convnet yada CNN) Nedir, Nasıl çalışır?* Medium. Retrieved December 1, 2021, from https://medium.com/@tuncerergin/convolutional-neural-network-convnet-yada-cnn-nedir-nasil-calisir-97a0f5d34cad.

Feng, X., Zhang, Y., & Glass, J. (2014). Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2014.6853900

Gao, T., Du, J., Dai, L.-R., & Lee, C.-H. (2018). Densely Connected Progressive Learning for LSTM-based speech enhancement. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2018.8461861

Gupta, P., Patidar, M., & Nema, P. (2015, September). Performance analysis of speech enhancement using LMS, NLMS and UNANR algorithms. In *2015 International Conference on Computer, Communication and Control (IC4)* (pp. 1-5). IEEE.

Haykin, S. (1996). Adaptive filter theory 3rd edition Prentice-Hall.

Hosur, S., & Tewfik, A. H. (1997). Wavelet transform domain adaptive fir filtering. *IEEE Transactions on Signal Processing*, *45*(3), 617–630. https://doi.org/10.1109/78.558477

Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 16(1), 229–238. https://doi.org/10.1109/tasl.2007.911054

Huang, W. (1999). *Wavelet transform adaptive signal detection* (thesis). North Carolina State University, Raleigh, NC.

Jenkins, W. K., & Marshall, D. F. (1999). Transform Domain Adaptive Filtering. In *Digital Signal Processing Handbook.* essay, CRC Press LLC.

Jenkins, W., Radhakrishnan, C., & Marshall, D. (2009). Transform domain adaptive filtering. *Digital Signal Processing Fundamentals*, 1–22. https://doi.org/10.1201/9781420046076-c22

Kim, J., & Hahn, M. (2019). Speech enhancement using a two-stage network for an efficient boosting strategy. *IEEE Signal Processing Letters*, *26*(5), 770–774. https://doi.org/10.1109/lsp.2019.2905660

Kingma, D., & Ba, J. (2015, April 23). Adam: A method for stochastic optimization. arXiv.org. Retrieved December 1, 2021, from https://arxiv.org/abs/1412.6980v5.

Kızrak, A. (2020, January 7). *Derine Daha DERİNE: Evrişimli Sinir Ağları*. Medium. Retrieved December 1, 2021, from https://ayyucekizrak.medium.com/deri%CC%87ne-daha-deri%CC%87ne-evri%C5%9Fimli-sinir-a%C4%9Flar%C4%B1-2813a2c8b2a9.

Koushik, J. (2016, May 30). Understanding convolutional neural networks. arXiv.org. Retrieved December 1, 2021, from https://arxiv.org/abs/1605.09081.

Kumar, T. L., & Rajan, K. S. (2012). Noise Suppression in speech signals using Adaptive algorithms. *International Journal of Engineering Research and Applications (IJERA)* , *2*(1), 718–721.

Linmei, Q., Guangrui, H., & Chongni, L. (2001, May). A new speech enhancement method. In *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)* (pp. 92-94). IEEE.

Loizou, P. C. (2017). *Speech enhancement: Theory and practice* (2nd ed.). CRC Press.

Maas, A. L., Le, Q. V., O'Neil, T. M., Vinyals, O., Nguyen, P., & Ng, A. Y. (2012). Recurrent neural networks for noise reduction in robust ASR. *Interspeech 2012*. https://doi.org/10.21437/interspeech.2012-6

Misiti, M. (2006). *Wavelet toolbox: For use with Matlab®: User's guide*. MathWorks, Inc.

Monson, B. B., Hunter, E. J., Lotto, A. J., & Story, B. H. (2014). The perceptual significance of high-frequency energy in the human voice. Frontiers in Psychology, 5. https://doi.org/10.3389/fpsyg.2014.00587

O'Shea, K., & Nash, R. (2015, December 2). *An introduction to Convolutional Neural Networks*. arXiv.org. Retrieved December 1, 2021, from https://arxiv.org/abs/1511.08458.

Özaydın, S. & Alak, İ. K. (2018). Speech Enhancement using Maximal Overlap Discrete Wavelet Transform. Gazi University Journal of Science Part A: Engineering and Innovation, 5(4), 159-171. Retrieved from https://dergipark.org.tr/en/pub/gujsa/issue/41914/451683

Park, S. R., & Lee, J. W. (2017). A fully convolutional neural network for speech enhancement. *Interspeech 2017*. https://doi.org/10.21437/interspeech.2017-1465

Pascual, S., Bonafonte, A., & Serrà, J. (2017). Segan: Speech enhancement generative adversarial network. Interspeech 2017. https://doi.org/10.21437/interspeech.2017-1428

Phan, H., McLoughlin, I. V., Pham, L., Chen, O. Y., Koch, P., De Vos, M., & Mertins, A. (2020). Improving gans for speech enhancement. IEEE Signal Processing Letters, 27, 1700–1704. https://doi.org/10.1109/lsp.2020.3025020

Rabiner, L. R., & Schafer, R. W. (2007). Introduction to digital speech processing. *Foundations and Trends in Signal Processing*, *1*, 1–194. https://doi.org/10.1561/9781601980717

Sarfjoo, Seyyed Saeed; Yamagishi, Junichi. (2018). Device Recorded VCTK (Small subset version), [sound]. University of Edinburgh. School of Informatics. The Centre for Speech Technology Research (CSTR). https://doi.org/10.7488/ds/2316.

Shah, N., Patil, H. A., & Soni, M. H. (2018). Time-frequency mask-based speech enhancement using convolutional generative Adversarial Network. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). https://doi.org/10.23919/apsipa.2018.8659692

Shahriyar, S. A., Akhand, M. A., Siddique, N., & Shimamura, T. (2019). Speech enhancement using convolutional denoising autoencoder. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). https://doi.org/10.1109/ecace.2019.8679106

Shams Esfand Abadi, M., Mesgarani, H., & Khademiyan, S. M. (2017). The wavelet transform-domain LMS adaptive filter employing dynamic selection of subband-coefficients. *Digital Signal Processing*, *69*, 94–105. https://doi.org/10.1016/j.dsp.2017.05.012

Shi, Y., Rong, W., & Zheng, N. (2018). Speech enhancement using convolutional neural network with Skip Connections. *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. https://doi.org/10.1109/iscslp.2018.8706591

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). https://doi.org/10.1109/icassp.2018.8462068

Stutz, D. (2014, August 30). *Understanding convolutional neural networks*. Wordpress - Seminar Report. Retrieved December 1, 2021, from https://www.davidstutz.de/wordpress/wp-content/uploads/2014/07/seminar.pdf.

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(7), 2125–2136. https://doi.org/10.1109/tasl.2011.2114881

Tüfekçi, M., & Karpat, F. (2019). International Conference on Human-Computer Interaction, Optimization and Robotic Applications (HORA). In *SETSCI Conference Processing* (5th ed., Vol. 4, pp. 28–31). Nevşehir, Turkey.

Upadhyay, N., & Karmakar, A. (2013). An improved multi-band spectral subtraction algorithm for enhancing speech in various Noise Environments. *Procedia Engineering*, *64*, 312–321. https://doi.org/10.1016/j.proeng.2013.09.103

Vaseghi, S. V. (2008). *Advanced Digital Signal Processing and noise reduction*. John Wiley & Sons.

Wang, J., Chen, Y., Chakraborty, R., & Yu, S. X. (2020). Orthogonal Convolutional Neural Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11505–11515. https://doi.org/10.1109/cvpr42600.2020.01152

Xing Luo, O. (2019). Deep Learning for Speech Enhancement: A Study on WaveNet, GANs and General CNN-RNN Architectures.

Xu, Y., Du, J., Dai, L.-R., & Lee, C.-H. (2015). A regression approach to speech enhancement based on Deep Neural Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(1), 7–19. https://doi.org/10.1109/taslp.2014.2364452

Yan Long, Lin Gang, & Guo An. (2004). Selection of the best wavelet base for speech signal. Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004., 218–221. https://doi.org/10.1109/isimp.2004

Yi Hu, & Loizou, P. C. (2006). Subjective comparison of speech enhancement algorithms. 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings. https://doi.org/10.1109/icassp.2006.1659980

Yuliani, A. R., Amri, M. F., Suryawati, E., Ramdan, A., & Pardede, H. F. (2021). Speech enhancement using Deep Learning Methods: A Review. Jurnal Elektronika Dan Telekomunikasi, 21(1), 19. https://doi.org/10.14203/jet.v21.19-26

Zhang, Y., & Zhao, Y. (2013). Real and imaginary modulation spectral subtraction for speech enhancement. *Speech Communication*, *55*(4), 509–522. https://doi.org/10.1016/j.specom.2012.09.005