UNIVERSITÀ
DI TRENTO

Neural Enhancement Strategies for Robust Speech Processing

By

MOHAMED NABIH ALI MOHAMED NAWAR

Supervisors

ALESSIO BRUTTI

DANIELE FALAVIGNA

DOCTORAL THESIS

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

in

Computer Science
Department of Information Engineering and Computer Science
IECS International Doctoral School
XXXV CYCLE

February 27, 2023

Approval Sheet

The thesis entitled *"Neural Enhancement Strategies for Robust Speech Processing"*, prepared by Mohamed Nabih Ali Mohamed Nawar in fulfillment of the requirements for the degree of Doctor of Philosophy is recommended for the final oral examination.

Dr. Alessio Brutti
Student Advisor

Examination Committee

Name	**Signature**	**Date**
Fabio Antonacci (External Member)		
Alberto Abad (External Member)		
Luca Turchet (Internal Member)		
Matteo Negri (Internal Member)		

Declaration of Authorship

I, Mohamed Nabih Ali Mohamed Nawar , declare that this thesis titled, "*Neural Enhancement Strategies for Robust Speech Processing*" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"We should be taught not to wait for inspiration to start a thing. Action always generates inspiration. Inspiration seldom generates action."

Frank Tibolt

Abstract

In real-world scenarios, speech signals are often contaminated with environmental noises, and reverberation, which degrades speech quality and intelligibility. Lately, the development of deep learning algorithms has marked milestones in speech-based research fields e.g. speech recognition, spoken language understanding, etc. As one of the crucial topics in the speech processing research area, speech enhancement aims to restore clean speech signals from noisy signals. In the last decades, many conventional speech enhancement statistical-based algorithms had been proposed. However, the performance of these approaches is limited in non-stationary noisy conditions. The raising of deep learning-based approaches for speech enhancement has led to revolutionary advances in their performance. In this context, speech enhancement is formulated as a supervised learning problem, which tackles the open challenges introduced by the speech enhancement conventional approaches. In general, deep learning speech enhancement approaches are categorized into frequency-domain and time-domain approaches. In particular, we experiment with the performance of the Wave-U-Net model, a solid and superior time-domain approach for speech enhancement.

First, we attempt to improve the performance of back-end speech-based classification tasks in noisy conditions. In detail, we propose a pipeline that integrates the Wave-U-Net (later this model is modified to the Dilated Encoder Wave-U-Net) as a pre-processing stage for noise elimination with a temporal convolution network (TCN) for the intent classification task. Both models are trained independently from each other. Reported experimental results showed that the modified Wave-U-Net model not only improves the speech quality and intelligibility measured in terms of PESQ, and STOI metrics, but also improves the back-end classification accuracy.

Later, it was observed that the dis-joint training approach often introduces signal distortion in the output of the speech enhancement module. Thus, it can deteriorate the back-end performance. Motivated by this, we introduce a set of fully time-domain joint training pipelines that combine the Wave-U-Net model with the TCN intent classifier. The difference between these architectures is the interconnections between the front-end and back-end. All architectures are trained with a loss function that combines the MSE loss as the front-end loss with the cross-entropy loss for the classification task. Based on our observations, we claim that the JT architecture with equally balancing both components' contributions yields better classification accuracy.

Lately, the release of large-scale pre-trained feature extraction models has considerably simplified the development of speech classification and recognition algorithms. However, environmental noise and reverberation still negatively affect performance, making robustness in noisy conditions mandatory in real-world applications. One way to mitigate the noise effect is to integrate a speech enhancement front-end that removes artifacts from the desired speech signals. Unlike the state-of-the-art enhancement approaches that operate either on speech spectrogram, or directly on time-domain signals, we study how enhancement can be applied directly on the speech embeddings, extracted using *Wav2Vec*, and *WavLM* models. We investigate a variety of training approaches, considering different flavors of joint and disjoint training of the speech enhancement front-end and of the classification/recognition back-end. We perform exhaustive experiments on the Fluent Speech Commands and Google Speech Commands datasets, contaminated with noises from the Microsoft Scalable Noisy Speech Dataset, as well as on LibriSpeech, contaminated with noises from the MUSAN dataset, considering intent classification, keyword spotting, and speech recognition tasks respectively. Results show that enhancing the speech embedding is a viable and computationally effective approach, and provide insights about the most promising training approaches.

Keywords: Deep Learning - Speech Enhancement - Speech Classification - Speech Embeddings

Acknowledgements

This achievement would not have been possible without the help and support of many people.

Firstly, I would like to express a special appreciation to my advisor Prof. Alessio Brutti, that has been an exceptional mentor to me. I would like to thank him for encouraging my research and for allowing me to grow as a research scientist and introducing me to the world of speech technology.

Special thanks to my co-supervisor Prof. Daniele Falavigna, for advising me during my Ph.D. thesis at the SpeechTek unit, Fondazione Bruno Kessler. I enjoyed and learned much from the too many interesting technical discussions that I had with him.

I wish to thank my colleagues at the SpeechTek unit, Fondazione Bruno Kessler especially Stefano Banno, and Umberto Capellazo for providing me with invaluable help and technical support since the beginning of the thesis. Their help was more than I could have asked for.

I would like to thank my family for their love and encouragement. Whatever I have achieved in my life, I owe it to my family. They are the pillars that I stand on today, always present even while on different continents. I am a better person because I have them in my life.

Lastly, My sincere to the soul of my father “Nabih Ali Mohamed” who I lost during my doctoral study and even the rest of my life. I believe he would be the proudest of me when I will be granted the Ph.D. degree.

List of Publications

- Ali, M.N., Brutti, A. and Falavigna, D., 2020, September. Speech enhancement using dilated wave-u-net: an experimental analysis. In 2020 27th Conference of Open Innovations Association (FRUCT) (pp. 3-9). IEEE.
- Ali, M.N., Schmalz, V.J., Brutti, A. and Falavigna, D., 2021, August. A Speech Enhancement Front-End for Intent Classification in Noisy Environments. European Signal Processing Conference (EUSIPCO) (pp. 471-475). IEEE.
- Ali, M.N., Falavigna, D. and Brutti, A., 2022. Time-Domain Joint Training Strategies of Speech Enhancement and Intent Classification Neural Models. Sensors, 22(1), p.374.
- Ali, M.N., Brutti, A. and Daniele, F., 2022. Enhancing Embeddings for Speech Classification in Noisy Conditions. Proc. Interspeech 2022, pp.2933-2937.
- Ali, M.N., Brutti, A. and Daniele, F., Direct Enhancement of Pre-trained Speech Embeddings for Speech Processing In Noisy Conditions, accepted for publication in the Computer Speech & Language journal, 2022.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
AMS	**A**mplitude **M**odulation **S**pectrogram
ASR	**A**utomatic **S**peech **R**ecognition
CER	**C**haracter **E**rror **Rate**
CNN	**C**onvolutional **N**eural **N**etwork
CRN	**C**onvolutional **R**ecurrent **N**etwork
DNN	**D**eep **N**eural **N**etwork
E2E	**E**nd **t**o **E**nd
FSC	**F**luent **S**peech **C**ommands
GAN	**G**enerative **A**dversarial **N**etwork
GMM	**G**aussian **M**ixture **M**odel
GSC	**G**oogle **S**peech **C**ommands
HMM	**H**idden **M**arkov **M**odel
IBM	**I**deal **B**inary **M**ask
IC	**I**ntent **C**lassification
IRM	**I**deal **R**atio **M**ask
ISTFT	**I**nverse **S**hort **T**ime **F**ourier **T**ransform
LSE	**L**east **S**quares **E**stimators
MCG	**M**ultiple **C**onfidence **G**ates
MFCC	**M**el **F**requency **C**epstral **C**oefficients
MS-SNSD	**M**icrosoft **S**calable **N**oisy **S**peech **D**ataset
MSE	**M**ean **S**quare **E**rror
PESQ	**P**erceptual **E**valuation **O**f **S**peech **Q**uality
RNN	**R**ecurrent **N**eural **N**etwork
S-SE	**S**upervised **S**peech **E**nhancement
SDC	**S**pectral **Constrained** **E**stimator
SE	**S**peech **E**nhancement
SNR	**S**ignal **T**o **N**oise **R**atio
SOTA	**State** **O**f **T**he **A**rt
SS	**S**pectral **S**ubtraction
SSL	**S**elf **S**upervised **L**earning
STFT	**S**hort **T**ime **F**ourier **T**ransform
STOI	**S**hort **T**ime **O**bjective **I**ntelligibility
TCN	**T**emporal **C**onvolutional **N**etwork
TDC	**T**ime **Constrained** **E**stimator
U-SE	**U**nsupervised **S**peech **E**nhancement
WER	**W**ord **E**rror **R**ate
WF	**W**iener **F**ilter

Dedicated to the soul of my father, no one will come like you, and no one else will take your place in my heart until the end of my life, may Allah have mercy on you.

Chapter 1

Introduction

This Chapter provides an overview of this thesis overview. The speech enhancement background is discussed in Section 1.1, highlighting the issues and trends of current speech enhancement approaches. The research problem is defined in Section 1.2. Our motivations are explained in Section 1.3. The thesis contributions are summarized in Section 1.4. Section 1.5 reports the datasets (i.e clean, and noise datasets) used in this thesis. Finally, the thesis organization is given in Section 1.6.

1.1 Overview of Speech Enhancement

Speech is the most common tool used for human communication [67]. Naturally, human speech conveys fundamental information e.g. context meaning, speaker information including speaker identity, emotion, gender, and age [96]. During the COVID-19 pandemic, most private and working commitments were done remotely depending on audio-visual platforms e.g. Zoom, Microsoft Teams, and Google Meet [153].

However, in real-world scenarios, speech signals are often contaminated by either stationary background noise mainly due to transmission equipment (electrical humming or blowing noises) or non-stationary environmental noise (public places, traffic

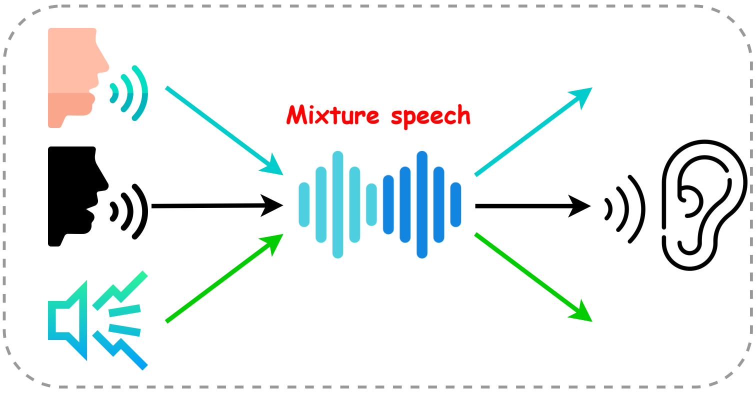


FIGURE 1.1: Cocktail party effect.

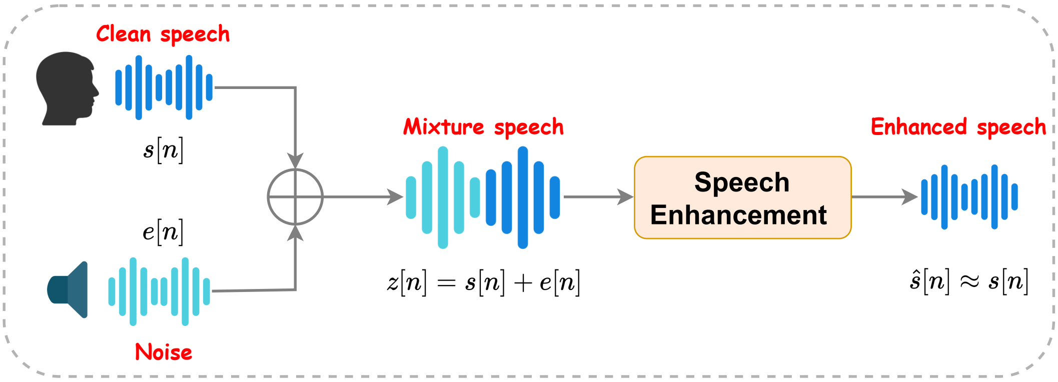


FIGURE 1.2: The basic diagram of speech enhancement system.

noise, background conversation) leads to a phenomenon known as the cocktail party effect [26, 42, 122] illustrated in Fig. 1.1.

Humans can extract the target speech signals among all interfering signals [200], although those who suffer from hearing impairment may have difficulty with speech quality and intelligibility under challenging noisy environments, especially when the signal-to-noise ratio (SNR) is less than or equal to $+10dB$ [75, 78].

In addition, the performance of speech-based applications, like automatic speech recognition (ASR), voice activity detection, and speaker recognition degrades in presence of these adverse noisy environments [1, 9, 125].

Hence, it is crucial to design computer algorithms to extract the target speech signal in the cocktail party scenario. In the last decades, countless kinds of research have been conducted to mitigate the noise effect and improve speech signal quality and intelligibility.

This is accomplished with speech enhancement, one of the most essential speech processing research areas [17], which aims to improve speech quality and intelligibility [115]. In practice, speech enhancement is widely integrated into many real-time applications [47] such as mobile communication [257], hearing aids [169], and speech recognition [167]. Fig. 1.2 shows the generic pipeline of the speech enhancement process.

Generally, speech enhancement approaches are categorized as unsupervised and supervised [189, 191]. Unsupervised approaches, considered as classical signal processing algorithms, (e.g. Wiener Filtering [54], spectral subtraction [162], etc.) usually depend on spectrogram transformation of speech signals and showed an acceptable performance in eliminating additive noise. However, they introduce distortion, especially in low SNR cases. In addition, their performance is insufficient in the case of non-stationary noisy environments [248].

Therefore, a series of supervised approaches (e.g. Gaussian mixture model [33], non-negative matrix factorization [93], etc.) were introduced in the past to mitigate speech distortion and residual noise issues. Despite the effective speech representation provided by these algorithms, their performance in presence of non-stationary noise is still a challenging task.

Recently, deep learning has shown outstanding performance in many research areas including speech enhancement [47]. Examples of deep learning architectures proposed for speech enhancement are: denoising autoencoder [121], convolutional neural networks [62, 63], recurrent neural networks [241, 262], and generative adversarial networks [129, 208]. All of them provide a better speech representation and enhancement performance. For these reasons, this thesis focuses on the application of deep learning-based approaches to speech enhancement. An extensive review of speech enhancement approaches is presented in detail in Chapter 2.

1.2 Problem Statement

A common dilemma with speech processing is speech perception. Common speech-based tasks are designed to work properly using clean speech. Regrettably, when these systems are exposed to noisy conditions, their performance considerably deteriorates. Considering the ASR scenario, an ASR system trained on isolated word clean speech signals achieves 100% accuracy, while in extremely noisy environments, the performance can be dropped by 30% [136]. This difference between the recognizer performance in clean and noisy environments causes a major obstacle in introducing ASR in real environments.

Speech enhancement approaches focus on retrieving clean speech signals from noisy ones either in the waveform or as hand-crafted features of clean speech embedded in noise. Literately, these techniques are not intended to improve the back-end performance. In particular, these techniques were originally aimed to improve speech quality. As mentioned above, this introduces signal distortion that is tolerable for humans but degrades the recognizers' performance.

1.3 Motivation

Deep learning-based approaches substantially alleviate the current problems introduced by classical approaches. However, there are still challenging tasks to be considered:

- **Effective speech representation:** Recently, large amounts of data are required to train deep learning-based models to achieve adequate performance. Thus, learning speech representation from the available data effectively gives rise to notable performance [39]. In specific, effective speech enhancement algorithms require learning how to represent the data as well as exploring the latent information [250].

- **Preservation of speech information:** Classical speech enhancement algorithms ignore some crucial speech information e.g. speech phonetic characteristics, and phase information which degrades the performance [43, 84]. Thus, it is important to utilize all of this information for further improvement.

- **Performance of speech back-end tasks:** In speech-based applications that involve noisy speech, speech enhancement is applied as a ”front-end” module followed by a ”back-end” one, which addresses the actual task, e.g. classifying speech. Often, the front-end and the back-end are trained dis-jointly. In this case, the front-end introduces distorted output-enhanced signals that deteriorate the back-end performance.

- **Computational complexity and resources:** Deploying deep learning models on edge devices e.g. mobile or embedded platforms is still a crucial need.

However, typical deep-learning algorithms exhaust these devices due to a large amount of multiply and accumulate (MAC) operations and memory access operations [34]. Consequently, matching the gap between deep learning requirements and low-resource devices is still a challenging task.

1.4 Contributions

To address the issues mentioned in Section 1.3 our novel contributions in this thesis are summarized below:

- For effective speech representation and utilizing speech information, we investigate the performance of a fully convolutional neural network called Wave-U-Net. This model was first proposed in [207] for audio source separation, later utilized for speech enhancement [146]. This approach technically sounds as it is a time-domain approach i.e. operates directly on the noisy raw waveform. Thus, no need for hand-crafted features, and it is effective in handling multi-noisy environments with affordable computational resources.
- To mitigate the front-end output distortion, that deteriorates the subsequent back-end speech classification performance. We introduce different joint training strategies in the time-domain and a novel method based on the domain of speech embedding. In this scenario, the whole training process is guided by the back-end model i.e. the front-end generates output signals desired for the back-end task. Hence, it improves not only the front-end performance but also the back-end.
- For further improvement in speech representation, we proposed using large-scale pre-trained speech models e.g. *Wav2Vec* [194], and *WavLM* [37]. In particular, we experiment with two enhancement strategies, the first is called **Wave-Enh** applies the time-domain enhancement at the beginning of the pipeline. Then speech embeddings are extracted from the enhanced signals and used to train the back-end. This solution makes use of the state-of-the-art model Wave-U-Net model. Conversely, the second strategy **Embeds-Enh**, applies the enhancement directly to speech embeddings and shows a positive impact both on the back-end performance and computational resources.

1.5 Thesis Datasets

1.5.1 Datasets for Speech Enhancement

Microsoft Scalable Noisy Speech Dataset (MS-SNSD)

The MS-SNSD dataset [180] [1] provides noise clips obtained from the DEMAND database [219] and Free Sound website [2]. The clips are carefully selected to ensure the quality of further noisy recordings. The chosen noise types are selected to be more relevant in realistic scenarios, but these types can always be scaled to accommodate new types.

[1] https://github.com/microsoft/MS-SNSD

[2] https://freesound.org/

Overall 14 different types of noise are available: air conditioner, announcements, appliances (washer/dryer), car noise, copy machine, door shutting, eating (munching), multi-talker babble, neighbor speaking, a squeaky chair, traffic, road, typing, and vacuum cleaner. It is worth mentioning that the dataset has a portion called test noise that conveys different noise recordings from the training noise although coming from similar categories. Hence, it is possible to investigate the robustness of their approach against unseen noisy conditions.

The dataset gives a wide range of options to generate noisy speech signals based on different selected options e.g. the number of speakers, noise types, and SNR desired levels.

MUSAN Dataset

The MUSAN corpus [201] includes approximately 109 hours of audio formatted as 16 kHz (.wav format) files. The dataset is in the US Public Domain or under a Creative Commons license and is publicly available at OpenSLR website [3]. The dataset is partitioned into speech, music, and noise (the category we used for our further experiments).

This noise corpus contains 929 different noise recordings, with approximately 6 hours of duration without including intelligible speech recordings. However, some recordings are crowd noises with indistinct voices. These range from technical noises e.g. dial tones, car idling, thunder, wind, footsteps, paper rustling, rain, animal noises, etc. The recordings were downloaded from Free Sound [2], and Sound Bible [4] websites.

1.5.2 Datasets for Speech Classification

Fluent Speech Commands Dataset

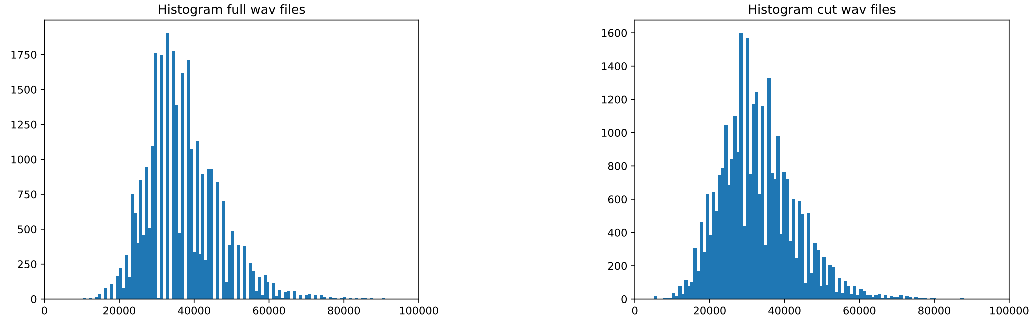
The Fluent Speech Commands (FSC) dataset proposed in [143] includes 30,043 English utterances obtained from 97 native and non-native speakers representing an interaction between smart-home devices or communicating with virtual assistants (e.g. “turn on the heat”, “switch on lights”, etc.). All signals are limited to a 4-sec duration and sampled at 16 kHz single-channel audio files.

Overall, the dataset provides 248 different utterances representing 31 different intents. On average, for each intent 8 different utterances are present. As mentioned above each intent comprise three slots: action, object, and location. For example, "Turn off the light" is labeled as `{action: "switch", object: "lights", location: "none"}`, and the combination of these three slots represents the utterance intent. Totally, the dataset includes 6 different actions, 14 objects, and 4 locations. The state-of-the-art reported on the clean testing portion for this dataset is around 99% intent classification accuracy [143, 175, 177, 220].

In order to avoid the presence of long silence in the original files, the ‘librosa.effects.trim()‘ module is employed to maintain signals duration to 4-sec long. Fig. 1.3(a) shows the histogram of the original length of the FSC dataset, while the histogram of the cut files is depicted in Fig. 1.3(b).

[3] `http://www.openslr.org/17/`

[4] `https://soundbible.com/`



(a)

(b)

FIGURE 1.3: Histogram representation (a) original FSC dataset. (b) cut FSC dataset.

Google Speech Commands Dataset v.1

The Google Speech Commands Dataset (GSC) has 65,000 recordings of 1-sec long utterances that provides 30 short words e.g. bed, three, digits from zero to nine, and robotic commands e.g. "Up, "Yes", "No", "Up", etc., contributed by members of the public through the AIY website. It's released under a Creative Commons BY 4.0 license. The recordings are organized into sub-folders according to the word they convey.

The dataset is used as a benchmark for training and evaluating keyword-spotting models. The goal is to detect a single spoken word in audio files from a set of different target words with as few false positives as possible from background noise or unrelated speech.

LibriSpeech-100 hours Dataset

LibriSpeech is a dataset specifically designed for ASR and commonly used in the related literature [163]. It features clean recordings of several different speakers reading segments of audiobooks that are a part of the LibriVox project. For training, we consider the "train-clean-100" set of the LibriSpeech corpus, containing 100 hours of clean speech signals uttered by 251 speakers and recorded at 16 kHz sampling frequency. For ASR validation and test we have used "dev-clean", and "test-clean" partitions, each including 40 speakers.

This corpus also provides an n-gram language model and the corresponding texts excerpted from the Project Gutenberg books, which contain 803M tokens and 977K unique words.

In all of our experiments, i.e. intent classification, keyword spotting, and speech recognition we consider the official split of the FSC, GSC, and LibriSpeech datasets described in Table 1.1.

1.6 Thesis Organization

In Chapter 2, we extensively survey different speech enhancement approaches. In particular, we overview both unsupervised and supervised algorithms highlighting

TABLE 1.1: Statistics (number of utterances, and duration in hours) of the FSC, GSC, and LibriSpeech datasets.

Data	FSC		GSC		LibriSpeech	
	Duration	# of utt.	Duration	# of utt.	Duration	# of utt.
Train set	14.7	23132	12.7	45931	100	28539
Validation set	1.9	3119	1.8	6799	5.4	2703
Test set	2.4	3793	1.9	6836	5.4	2620

many studies towards improving the performance of both categories. Finally, we discuss the main limitation of each algorithm.

In Chapter 3, we experiment with a pipeline that integrates the Wave-U-Net for speech enhancement with a back-end E2E intent classification model that operates on the 40-Mel filter-banks features of the enhanced signals.

Chapter 4, we proposed a fully time-domain joint training pipeline that integrates the Wave-U-Net model with the same intent classifier in this case our classifier is directly trained on the waveform of the enhanced signal.

Chapter 5, presents the proposed pipeline that integrated the large-scale pre-trained speech models e.g. *Wav2Vec* and *WavLM* for joint training speech enactment with different back-end speech classification tasks e.g. intent classification, keyword spotting, and speech recognition.

Finally, Chapter 6 concludes our work and discusses the possible future directions. Fig. 1.4 shows the graphical representation of the thesis organization.

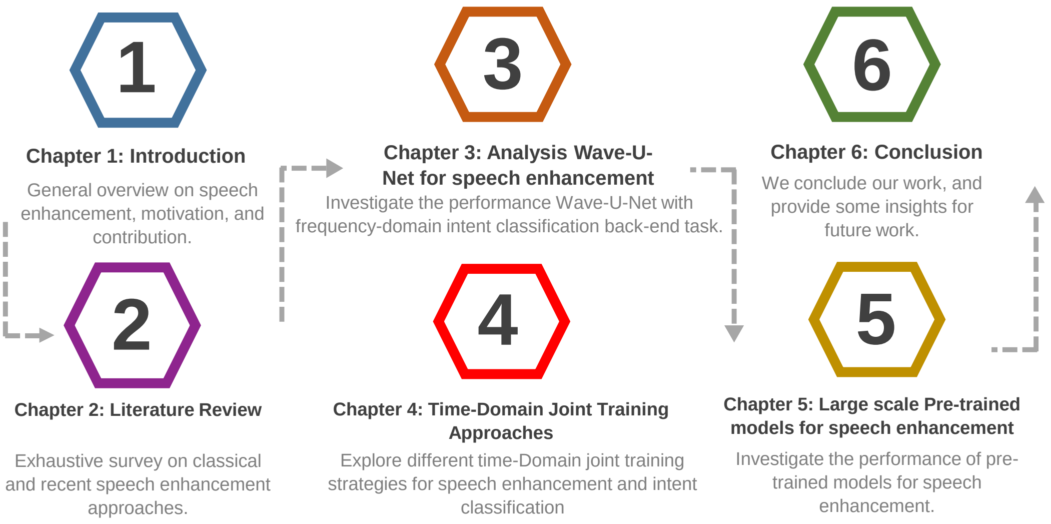

1
Chapter 1: Introduction
General overview on speech enhancement, motivation, and contribution.
3
Chapter 3: Analysis Wave-U-Net for speech enhancement
Investigate the performance Wave-U-Net with frequency-domain intent classification back-end task.
6
Chapter 6: Conclusion
We conclude our work, and provide some insights for future work.
2
Chapter 2: Literature Review
Exhaustive survey on classical and recent speech enhancement approaches.
4
Chapter 4: Time-Domain Joint Training Approaches
Explore different time-Domain joint training strategies for speech enhancement and intent classification
5
Chapter 5: Large scale Pre-trained models for speech enhancement
Investigate the performance of pre-trained models for speech enhancement.

FIGURE 1.4: Overview of this thesis.

Chapter 2

Literature Review

This chapter surveys different speech enhancement approaches. Section 2.1, gives an introduction to the main speech enhancement categories present in the literature. Section 2.2, and Section 2.3 the Unsupervised and Supervised speech enhancement approaches are reviewed respectively. At the end of each section,the main limitation of each technique are highlighted. In Section 2.4, the most common speech enhancement evaluation metrics are explained that will be used in later experiments. Finally, Section 2.5 surveys some recent research attempts to improve speech recognition performance using SE as a front-end.

2.1 Introduction

During the last decades, SE has received a lot of attention in the speech processing research area. The goal of this process is to improve both speech quality and intelligibility by mitigating the noise impact on the desired speech signal. In particular, the SE algorithms estimate the noise characteristics from the noisy signals and eliminate the undesired noise to provide clean speech signals.

Generally, single-channel SE algorithms are classified into two main categories: Unsupervised speech enhancement (U-SE), and Supervised speech enhancement (S-SE), as shown in Fig. 2.1. In this Chapter, we provide a detailed review of both categories highlighting the advantages and disadvantages of each technique.

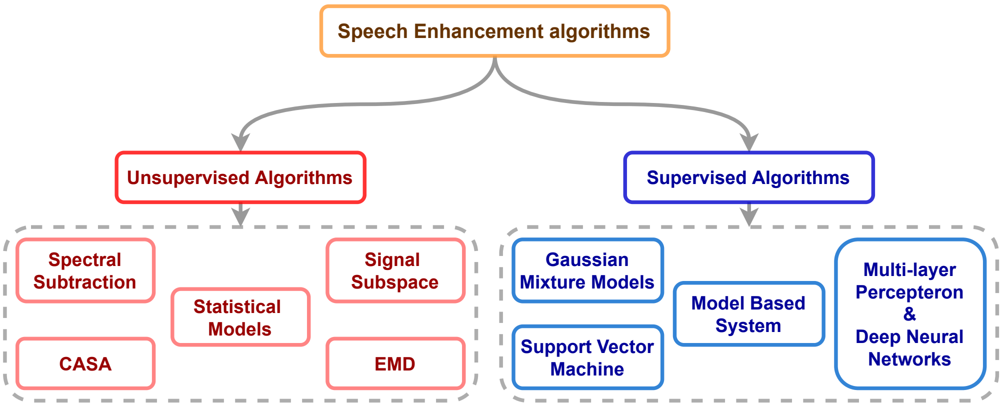


FIGURE 2.1: Classification of Speech Enhancement algorithms.

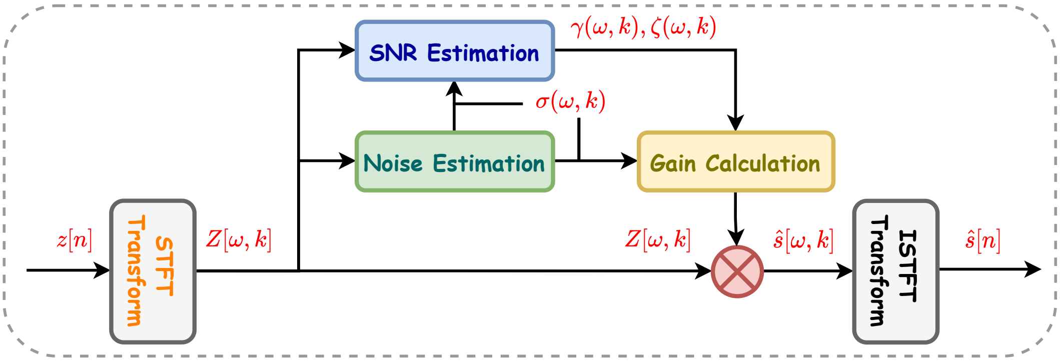


FIGURE 2.2: Block diagram of unsupervised systems [191].

2.2 Unsupervised Speech Enhancement Algorithms

In U-SE algorithms, a statistical-based model is employed to estimate the target speech signals from the noisy signal, while ignoring other information e.g. noise type, and speaker identity. In the following sub-sections, we provide a review of U-SE algorithms. Fig. 2.2 shows the block diagram of a generic unsupervised approach for single-channel speech enhancement.

2.2.1 Spectral Subtraction

The spectral subtraction (SS) algorithm, proposed in [20], is one of the earliest and most effective solutions to mitigate the noise effect. In this algorithm, the noise is assumed to be additive, and the enhanced speech spectrum is obtained by subtracting the noise spectrum from the mixture, as depicted in Fig. 2.3. This algorithm is designed based on the hypothesis that the noise spectra are stationary [224]. The enhanced speech signals are then reconstructed by computing the inverse discrete Fourier transform of the enhanced spectrum, using however the phase of the noisy signal. Mathematically, denoting as $s[n]$ the clean speech signal, and as $e[n]$ the additive noise at time index n, the noisy speech signal $z[n]$ can be formulated as:

$$z[n] = s[n] + e[n] \tag{2.1}$$

Applying the short-time Fourier transform (STFT) to Eq. 2.1, we obtain the spectral

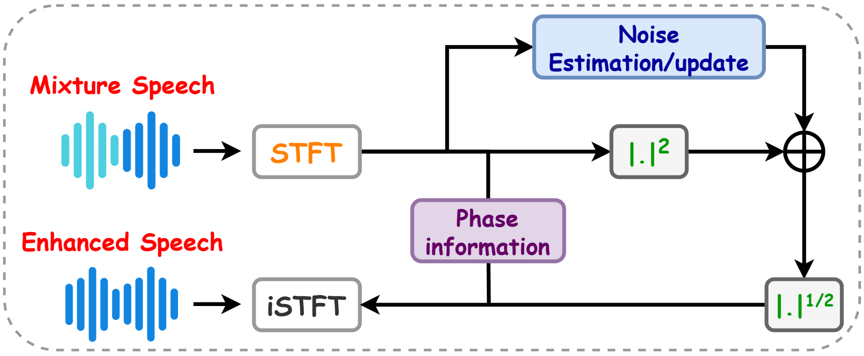


FIGURE 2.3: Block diagram of SS approach.

formula expressed as:

$$Z(\omega,k) = S(\omega,k) + E(\omega,k) \tag{2.2}$$

By subtracting the noise magnitude spectrum $|E(\omega,k)|$ from the noisy magnitude spectrum $|Z(\omega,k)|$, we can obtain an estimate of the clean signal spectrum.

$$\hat{S}(\omega,k) = [Z(\omega,k) - E(\omega,k)]e^{j\phi_z(\omega,k)} \tag{2.3}$$

Finally, the inverse Fourier transform is applied to retrieve the time-domain speech signals $\hat{s}[n]$ such that:

$$s[n] \approx \hat{s}[n] \tag{2.4}$$

Since the SS algorithm assumes the noise signal as stationary or slowly time-variant, it tends to introduce negative values in the enhanced magnitude spectrum, which result in musical noise artifacts [58]. To mitigate the effect of musical noise, the authors in [141, 157] proposed an improved SS algorithm using the geometric approach for SE. In this approach, the cross-terms are estimated involving the phase differences between noisy, clean speech signals and noise. An experimental analysis of the proposed algorithm shows that it outperforms the conventional SS approach.

Unlike the conventional approach that performs the subtraction on the magnitude spectrum in the frequency-domain, the authors in [261] proposed to perform the subtraction separately on the real and imaginary spectra. Exhaustive analysis showed that fewer musical noise artifacts were observed, which improves speech quality and intelligibility.

Finally, [12] proposed a single-channel blind dereverberation algorithm based on the SS approach for remote-talking speech recognition applications. Subsequently, the Viterbi-decoding method was employed on the output of the reverberation model to find out the most likely word sequence. In [112], the authors showed that the SS algorithm outperformed the ideal reverberant masking approach in terms of late reflections suppression.

2.2.2 Statistical Model-based Algorithms

Similar to SS algorithms, statistical model-based algorithms assume that speech and noise are stationary signals, hence their statistics remain constant and can be easily estimated. The noise signals are eliminated by utilizing either Finite Impulse response (FIR) or Infinite Impulse Response (IIR) filters [22]. Typically, the filter weighting gains are computed using the short-time power spectral density (PSD) of the noisy mixture $Z(\omega,k)$ and an SNR estimate in the frequency-domain. As shown in Eq. 2.5, the clean spectrum is estimated by multiplying the noisy spectrum with the weight gain $G(\omega,k)$:

$$\hat{S}(\omega,k) = G(\omega,k)Z(\omega,k) \tag{2.5}$$

The filter $G(\omega,k)$ is computed using particular SE algorithms as a function of short-time noise PSD estimate $P_D^2(\omega,k)$ and an estimate of the SNR. Assuming that the PSDs of the clean speech and of the noise $P_S^2(\omega,k)$ and $P_E^2(\omega,k)$ are available, the SNR can be computed as in Eq. 2.6:

$$\xi(\omega,k) = \frac{P_S^2(\omega,k)}{P_E^2(\omega,k)} \tag{2.6}$$

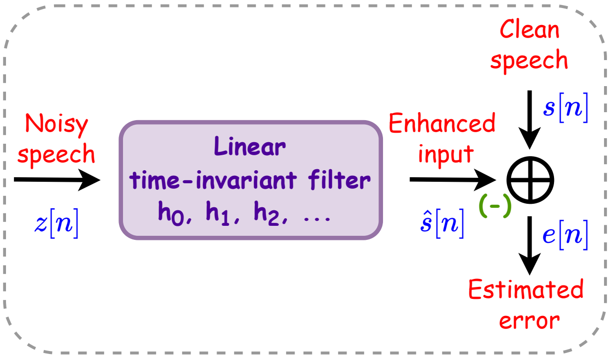


FIGURE 2.4: Block diagram of the statistical filtering problem.

However, since the two quantities are not available, $P_E^2(\omega, k)$ is calculated during the non-speech and silence periods using the following recursive equation:

$$\hat{P}_E^2(\omega, k) = \beta \hat{P}_E^2(\omega, k-1) + (1-\beta)\hat{P}_z^2(\omega, k-1) \tag{2.7}$$

Where β is a smoothing factor, and $\hat{P}_z^2(\omega, k-1)$ is the estimated noise in the previous frame. Then, the prior SNR can be estimated using the Decision Direct approach [56], and illustrated in [139], which linearly combines the prior and post SNRs as follows:

$$\xi(\omega, k) = \alpha \xi(\omega, k-1) + (1-\alpha) \max \left[\frac{{P_z}^2(\omega, k)}{\hat{P}_E^2(\omega, k)} - 1, 0 \right] \tag{2.8}$$

Where α is a weighting coefficient and $\xi(\omega, k-1)$ is the prior SNR at the previous iterations. Two main statistical-based model approaches are widely used namely the Wiener filter (WF) and minimum mean square error (MMSE).

- Wiener Filter

Analogous to the conventional filtering approaches, Wiener filter [35], depicted in Fig. 2.4, applies a linear and time-invariant system on the input noisy signals $z[n]$ to estimate the enhanced signals $\hat{s}[n]$. This can be done by minimizing the estimation error between clean signals, and enhanced ones. The optimal Wiener filter gain is formulated as follows [3]:

$$G(\omega, k) = \frac{\xi(\omega, k)}{\xi(\omega, k) + 1} \tag{2.9}$$

In the last decades, several contributions had been conducted to improve the Wiener filtering performance. The approach proposed in [50], utilized a hybrid 1D and 2D Wiener filter [205] to eliminate the noise in the speech spectrogram. Then, a post-processor is applied to the noisy regions to remove the residual noise components. Reported experiments showed that the hybrid filter approach is more effective than the conventional SS, and Wiener filter approaches in terms of speech quality.

Unlike the conventional frequency-domain Wiener filter approach, the authors in [54] proposed an adapted time-domain Wiener filtering approach. This method considers the local statistics of the speech signal. The proposed approach results showed performance superiority with respect to other approaches e.g. spectral subtraction and wavelet denoising in case of in the case of Additive White Gaussian Noise (AWGN), and colored noise.

In [11] a speech-distortion weighted inter-frame Wiener filter (SDW-IFWF) is proposed, for single-channel noise reduction based on filter-banks features. The filter employed a parameter μ that controls the trade-off between noise reduction, and speech distortion. This strategy is widely used in multi-channel applications under the term multi-channel speech-distortion weighted Wiener filter. Reported experiments show, that larger values of μ provide better enhancement performance in terms of segmental SNR metric, and it is computationally effective compared with the conventional approaches.

Recently, the authors in [182] proposed a Wiener filter estimation based on deep learning. In particular, the optimal parameter of the Wiener filter (i.e. SNR estimation, and gain function) are estimated by a deep neural network to improve the Wiener filter performance. Reported experiments show that incorporating data-driven approaches (i.e. deep learning approaches) for estimating the filter parameters outperforms the statistical-based speech estimator algorithm.

- MMSE Estimator

As discussed in the previous part, the Wiener filter estimates the enhanced speech signals by minimizing the error between the clean spectrum and the enhanced spectrum. Unfortunately, the Wiener filter is considered to be optimal for complex spectral estimators but is not optimal for spectral magnitude estimators, which degrades the SE performance [227].

Thus, the MMSE estimator exploits the performance of the short-time spectral amplitude (STSA) on speech quality and intelligibility. In literature, optimal MMSE estimators proposed to minimize the MSE between the enhanced and clean magnitudes

$$E\{(S(\omega,k)-\hat{S}(\omega,k))^2\} \rightarrow Min \tag{2.10}$$

The authors in [100] proposed an algorithm for joint MMSE estimation of speech coefficients using phase uncertainty to estimate the signal amplitude. Furthermore, new phase-blind estimators are developed based on the Nagakami power spectral density function and the generalized Gamma function for speech and noise priors.

A different approach for the MMSE estimator is presented in [65]. In contrast to the other estimators, the MMSE approach is used to estimate the clean phase from the noisy one. In this way, the estimated clean phase can provide additional information that can be exploited to improve the resulting speech quality.

To improve the performance of the MMSE algorithm, the authors in [4] employed β-order MMSE STSA. The motivation is to exploit the advantages of both Laplacian speech modeling and β-order cost function in MMSE estimation of clean speech. In particular, the proposed solution for $\beta-$order MMSE-STSA taking into account Laplacian priors for clean speech DFT coefficients leads to better adaptation for the estimators.

2.2.3 Signal Subspace-based Algorithms

Signal subspace-based algorithm proposed in [57, 80] utilizes Eigen Value Decomposition (EVD) and Karhunen-Loeve transform (KLT) to decompose the noisy signal into two subspaces, for clean and noise signals [137] as depicted in Fig. 2.5. Hence, the clean signal could be estimated by removing the noise subspace.

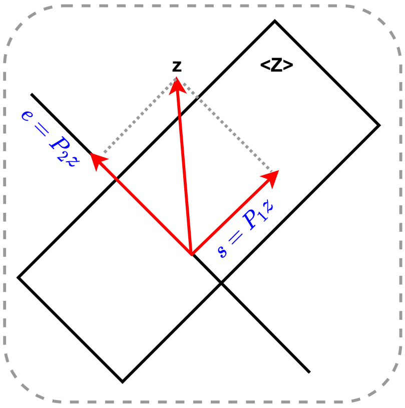


FIGURE 2.5: Decomposition of the noisy vector z into its orthogonal components s, and e represents clean, and noise respectively.

Using Eq. 2.1, the noisy covariance matrix (R_z) is represented as the sum of clean, and noise covariance matrices R_s, and R_e respectively, where R_z have a higher rank than R_s:

$$R_z = R_s + R_e \tag{2.11}$$

The EVD of R_z and R_s is given by:

$$R_z = U\Lambda_z U^T \qquad R_s = U_p\Lambda_s U_P^T \tag{2.12}$$

where Λ_z and Λ_s represent the eigen values diagonal matrices of R_z and R_s respectively:

$$\Lambda_z = diag(\lambda_1, \lambda_2, ..., \lambda_Q) \qquad \Lambda_s = diag(\lambda_{s,1}, \lambda_{s,2}, ..., \lambda_{s,P}) \tag{2.13}$$

Q and P are the dimensions of U and U_P respectively, such that $Q > P$. The noise covariance matrix is defined as

$$R_e = \sigma^2 I \tag{2.14}$$

where σ^2 is the noise variance and I is the identity matrix. Using Eq. 2.12 and Eq. 2.14, R_z can be computed as

$$R_z = U(\Lambda_s + \sigma^2 I)U^P \tag{2.15}$$

where $U = [U_P U_{Q-P}]$, $U_P = [u_1 u_2, ..., u_P]$ represents the signal subspace, $U_{Q-P} = [u_{P+1}, u_{P+2}, ..., u_Q]$ represents the noise subspace, u_i denotes the eigen vector corresponding to the eigen value λ_i. Finally, a linear filter ψ is designed using different estimators (i.e LSE, Liner-MMSE, TDC, SCD, etc.) to estimate the clean subspace from the noisy one.

$$\hat{s} = \psi z \tag{2.16}$$

The matrix ψ is defined as:

$$\psi = U_p G U_p^T \tag{2.17}$$

where G is a gain matrix. The residual error (r) is defined as

$$r = \hat{s} - s = \psi z - Is \tag{2.18}$$

$$r = \psi s - \psi e - Is = (\psi - I)s + \psi e = r_s + r_e \tag{2.19}$$

Where r_s and r_e represent the signal distortion and residual noise respectively.

$$r_s = (\psi - I)s \qquad r_e = \psi e \tag{2.20}$$

One of the most important aspects in the subspace algorithms is dimensionality reduction, achieved by reducing the noise matrix rank. Furthermore, the appropriate choice of some parameters (e.g. window size, matrix rank. etc.) shows competitive performance with respect to classical SE algorithms [212].

In literature, obtaining an optimal estimator to retrieve clean speech signals, gained a lot of attention in the case of colored noise, as it is a challenging task. A pioneer work that attempted to solve the colored noise issue is proposed in [57]. The authors suggested whitening the noisy speech. However, in this case, the performance of the estimators significantly deteriorates. The reason is that the estimators focus on minimizing the whitened speech distortions rather than the clean speech distortions. Other methods reported in [150, 181] mitigate the colored noise effect, by proposing approximations of the noise covariance matrix. The approach proposed in [85] employs a joint diagonalization of noise and speech covariance matrices that show promising performance in colored noise conditions. However, these approaches are highly-dependent on Lagrange multipliers that need to be carefully set to obtain desired filter performance. One possible solution is to set the Lagrange multipliers to a fixed value as proposed in [85]. Alternatively in [23], the residual power noise spectrum is used to estimate the Lagrange multipliers accurately. Then, the estimated Lagrange multipliers are utilized to modify the spectral-domain-estimator. This approach yielded high noise reduction and improved speech quality, at the cost, however, of increased computational complexity. More recent approaches in [94] and [212] utilized the Rayleigh quotient method. In particular, this method replaces the noise variance with the Rayleigh quotient. This approach better shapes the noise matrix with respect to the conventional approaches and decreases the computational complexity [212].

2.2.4 Computational Auditory Scene Analysis (CASA) Algorithms

The Computational Auditory Scene Analysis (CASA) approaches [184, 230] have been widely used in the SE task. These approaches employ the auditory perception mechanism without prior information about the noise. The CASA models are trained to estimate binary or ratio masks in the time-frequency domain [206]. These masks are used to remove the noise components from the noisy mixture.

The authors in [154] proposed an approach using the ideal binary mask (IBM) for speech separation in the time-frequency domain. In particular, an SNR threshold on the energy in speech and noise regions is used to define the binary masks. Finally, an SNR transform is introduced to estimate the true broadband SNR of the noisy signal.

The approach in [148] estimates the IBM using the amplitude modulation spectrogram (AMS) features and modulation filter-banks features. A spectro-temporal integration stage was employed to obtain speech activity information in neighboring time-frequency units.

In [97], a novel feature enhancement approach based on CASA was proposed. Unlike the other approaches, that focus on eliminating the noise from the noisy speech, the definition of IBM includes aspects related to speech recognition performance.

Exhaustive experiments showed an improvement in robust speech recognition, with a definite improvement from 20% to 40% at 5 dB SNR.

The approaches reported above are based on estimating the IBM, which is computed through thresholding with a local SNR criterion. In particular, each T-F unit is labeled as a target speech if the signal power is greater than the SNR threshold value. Thus, IBM is based on a hard decision approach, labeling the T-F units as 0 or as 1. Consequently, often the IBM approach removes the background noise in the weak speech T-F units, with negative effects on the speech quality. An alternative solution is to use the ideal ratio mask (IRM): a soft decision mask whose values smoothly vary between 0, and 1. For example, the authors in [13, 14] proposed a novel IRM in the Gammatone domain. The proposed approach is more effective in eliminating noise while preserving the speech components using the inter-channel correlation (ICC) between the noisy speech, clean speech, and noise power spectra. The ICC is assigned a larger value in case of a strong correlation between noisy speech, and noise. This means that the noise components are predominant in the noisy speech signal with respect to the clean ones.

The authors in [119] applied the shape analysis techniques originally introduced for image processing to CASA-based approaches. This approach extracts the desired speech signals from the noisy signals, while the missing speech signals are complemented using shape analysis techniques. This approach improves the final performance by 22% for speech recognition contaminated with stationary noise.

2.2.5 Empirical Mode Decomposition Algorithms

Empirical Mode Decomposition (EMD) is an approach that is designed for multiscale decomposition and signal analysis in the time-frequency domain. [202]. In particular, EMD employs the shifting process that decomposes an input signal into a finite set of oscillating components called Intrinsic Mode Functions (IMFs). Differently, from the conventional decomposition approaches (i.e. Fourier or Wavelet Transform), the IMFs are not set analytically but are obtained using only the analyzed sequences. The estimated IMFs from the EMD have to justify two criteria [90]:

- the number of zero-crossing and IMF extrema must either be equal or differ by one in the whole dataset.
- at any point, the envelopes mean value defined by the local maxima and local minima is zero at any point of an IMF.

The shifting process mentioned above repeatedly subtracts the input signal from its local mean until a zero mean is obtained. Typically a stopping criterion is applied to stop the shifting process. This is based on the relative variation between two consecutive shiftings and a threshold. The following steps summarize the EMD algorithm and are depicted in Fig. 2.6.

(i) Estimate the local maxima and minima of the input signal $y[n]$.

(ii) Employ an interpolation method to generate the upper and lower signal envelope by connecting the local maxima and minima as depicted in Fig. 2.7.

(iii) Averaging the upper and lower envelopes to determine the local mean $\mu[n]$.

(iv) Subtract the local mean from the input signal $h[n] = z[n] - \mu[n]$.

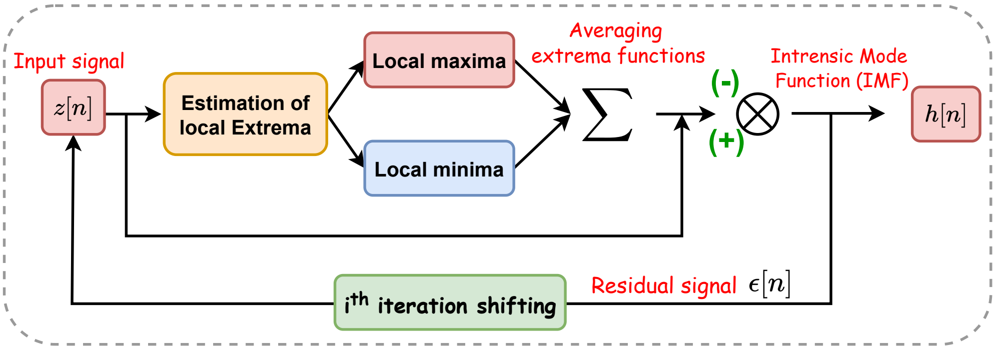

Input signal
$z[n]$
Estimation of local Extrema
Local maxima
Local minima
Averaging extrema functions
$\sum$
(-)
(+)
Intrensic Mode Function (IMF)
$h[n]$
Residual signal $\epsilon[n]$
i^{th} iteration shifting

FIGURE 2.6: Block diagram of EMD approach.

(v) If $h[n]$ complies with the stopping criterion, the IMF is defined as ($d[n] = h[n]$), otherwise, $y[n] = h[n]$ and repeat the process.

Finally, the EMD of the signal decomposition $z[n]$ can be formulated as:

$$z[n] = \sum_{t=1}^{m} IMF_t[n] + \epsilon_m[n] \tag{2.21}$$

Where m and $\epsilon_m[n]$ are the extracted IMFs, and the residual signal after decomposition respectively.

Several research were reported in the literature that employs the EMD approach for SE task. The approach proposed in [102] investigates the performance of the EMD algorithm combined with the Teager-Kaiser energy operator, which uses an adaptive threshold method. However, this approach is designed to perform in white noisy conditions. Conversely, the approach proposed in [255] combines the EMD approach with the Hurst exponent. This approach shows a substantial improvement in the case of highly non-stationary noise. However, it does not bring significant improvements in white noise conditions. Towards solving this issue, the approach presented in [32], namely the EMD-based filtering approach (EMDH) employed to eliminate the low-frequency noise components. Despite the promising results, its performance drastically deteriorates in presence of babble noise. Alternatively, EMD

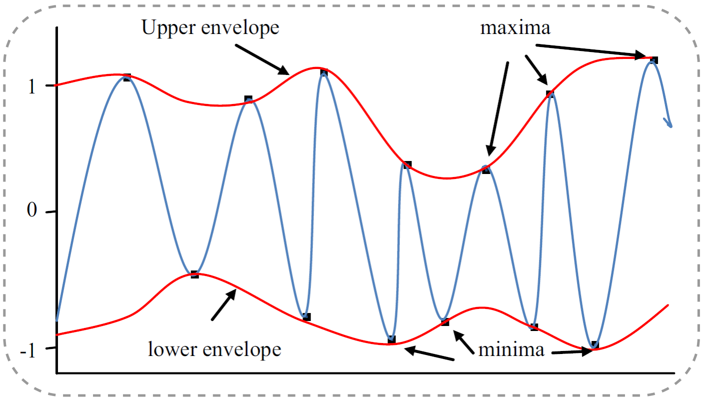


FIGURE 2.7: Extrema, upper, and lower envelope for time-domain signal [55].

was combined with other filtering approaches, such as MMSE [103] or spectral subtraction [55]. The key idea is that the filter is used to denoise each IMF separately, and later the enhanced IMFs are used to reconstruct the enhanced signals. These approaches show promising performance when employed in white Gaussian noise environments. Recently, the combination of EMD with variation mode decomposition (VMD), originally introduced in [51], was proposed in [135] for fiber optic gyroscope signals denoising and later employed in SE area [223], proving effective in reducing both high and low-frequency noise.

2.2.6 Limitation of Unsupervised Speech Enhancement Algorithms

The U-SE algorithms bring notable improvement in terms of speech quality and noise reduction in real-world noise sources, these algorithms have some limitations summarized as follows:

- **Performance with non-stationary noise**: Despite the promising performance of U-SE algorithms in terms of speech quality. These algorithms assume that the noise is stationary. Thus, its performance is negatively affected in the case of non-stationary noise. Hence, it still needs effective noise estimation for further performance improvement.
- **Speech intelligibility, and distortion**: As mentioned in the previous point the U-SE algorithms obtained high-quality enhanced speech signals. However, these algorithms introduce signal distortions at their output, leading to low speech intelligibility. Thus, it is still challenging to provide more effective approaches to remove the residual noise artifacts.

2.3 Supervised Speech Enhancement Algorithms

S-SE algorithms are trained using a labeled dataset (i.e pair of clean and noisy speech samples). The goal of these algorithms is to learn the relationship between the clean and noisy versions of the speech signals and use this knowledge to enhance the quality of noisy speech signals. The training process can be done on appropriate speech signals transformation (in the case of frequency-domain S-SE algorithms), or directly on the raw waveform (in the case of time-domain S-SE algorithms).

2.3.1 Gaussian Mixture Model for Speech Enhancement

The Gaussian Mixture Model (GMM) introduced in [242] uses a probabilistic model based on the assumption that the data points are generated from a finite number of Gaussian distributions. The GMM for a process (N) is defined as:

$$f(N) = \sum_{m=1}^{M} p_m G_m \left(N; \lambda_m, \sum_{m} \right) \tag{2.22}$$

Where f represents the probability density function (PDF), G_m is the Gaussian PDF of the m_{th} mixture component, λ_m, $\sum_m$, and p_m are the mean vector, covariance matrix, and prior probability respectively. We highlight some research papers that study GMM for speech enhancement.

In [107], the GMM model is trained to predict the IBM mask in multi-noisy conditions utilizing the AMS with its delta feature augmentation. The resulting feature vector is defined as:

$$A(\tau, k) = [a(\tau, k), \Delta a_T(\tau, k), \Delta a_K(\tau, k)] \tag{2.23}$$

Where $\Delta a_T(\tau, k)$, and $\Delta a_K(\tau, k)$ denote the delta feature vectors computed across time (τ) and frequency (k), respectively. The obtained results show a notable performance in terms of speech intelligibility when the model is trained and tested on matched noisy conditions. However, this approach lacks generalization i.e. the performance deteriorates in case of unseen noisy (mismatched) conditions.

This problem has been addressed in [149] where the authors examine the sensitivity of their speech segregation model to different noise parameters (i.e noise variations, duration) during the training and testing phases. In addition, they consider the complex interaction between noise variations, Gaussian component numbers, and feature space dimensionality. Exhaustive experiments show clear robustness against unseen noisy conditions during the testing phase.

The authors in [106], proposed an alternative algorithm to estimate the IBM tailoring speech intelligibility. In detail, the noisy speech is decomposed into time-frequency units and the GMM is used to take binary decisions about whether each unit belongs to the target speech signal or noise. Based on these decisions, the target speech units are retained, while other units are discarded.

Another contribution was done in [108] to improve GMMs performance against noise variations. In detail, frequency-dependent masking classifiers are developed to estimate the missing features. Finally, an adaptive approach estimates the prior values of the mask classifiers to decide whether the T-F segment is enhanced or not. This method showed a promising improvement in terms of WER.

Recent research was done in [147], where the noise is modeled using the GMM approach with a multi-stage process incorporated with a parametric Wiener filter. In this way, the model estimates the noise power spectral density accurately for better generalization, hence better speech quality and intelligibility in terms of PESQ and STOI metrics.

2.3.2 Support Vector Machine for Speech Enhancement

Support vector machine (SVM), as depicted in Fig. 2.8, is a discriminative classifier that uses a hyper-plane to differentiate among all classes [29, 168]. In the SE task, this hyper-plane separates the noisy training data into two parts belonging to the target speech and noise.

The authors in [72], proposed a classification approach to estimate IBM. The SVM classifier is trained on a combination of AMS, and pitch-based features to classify the time-frequency units either target speech or noise. For further improvement, a re-thresholding method is integrated to robust classification accuracy and maximize hit minus false alarm rates.

Furthermore, the authors in [73], addressed the problem of generalization to unseen noise conditions using a small training corpus for voice activity detection task. The system employs SVM for the classification task followed by a thresholding technique to estimate the IBM. Systematic evaluation shows that the proposed approach estimates high-quality IBM under unseen conditions.

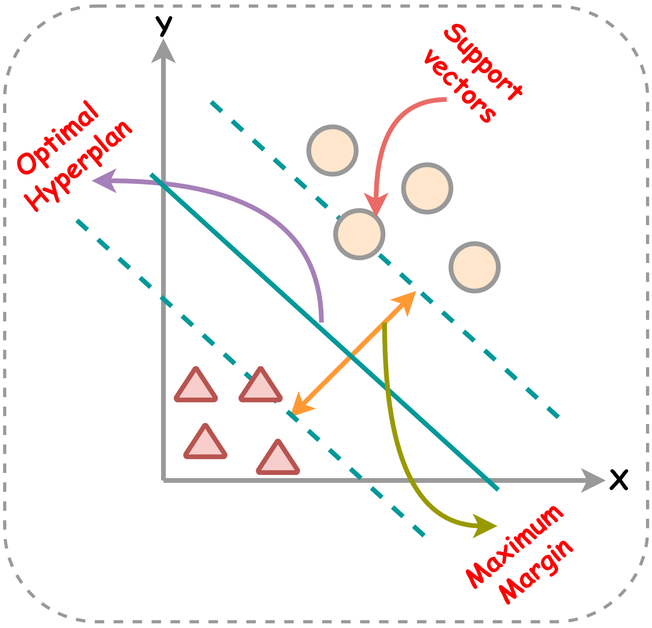


FIGURE 2.8: Support Vector Machine

In [36], the authors introduced new features named multi-resolution cochleagram features (MRCC) that outperformed the complementary features proposed in [234]. The SVM classifier is utilized to classify clean and noise T-F units. Experimental results show better speech intelligibility in the case of extreme non-stationary noisy conditions.

2.3.3 Non-Negative Matrix Factorization

Model-based algorithms characterize the SE problem as a supervised learning task by constructing models that estimate speech and noise characteristics. The core idea of these algorithms is to detect the appropriate time-frequency area for signal reconstruction, thus a few time-frequency areas must be contaminated with high SNR [113].

Non-negative matrix factorization (NMF) is a well-known model-based speech enhancement approach [5, 152]. In this algorithm, the input speech signals are decomposed into activation and basis matrices under the assumption that both matrices and the signals are positive. Consider $H \in \mathbb{R}_{+}^{I \times T}$ represents the data low dimension non-negative representation i.e. the activation matrix and W is the basis matrix. The NMF is defined as the product of these two non-negative matrices that gives an accurate estimate of the signal [151] as follows:

$$Y = HW \tag{2.24}$$

Several studies investigated the NMF performance for speech enhancement and separation [120, 142, 240], we report the recent work for NMF-based SE.

The authors in [244], proposed a novel NMF-HMM algorithm based on the Kullback-Leibler (KL) divergence. Compared to the conventional NMF approach, the proposed approach exploit the speech signal temporal dynamics to perform the enhancement task. In this way, the time information is considered during the enhancement process. Moreover, they employ the sum of Poisson distribution as the state-conditioned likelihood for the HMM rather than the general GMM. They motivated this, as the sum of Poisson distribution leads to the KL divergence measure used for NMF measurement.

The author in [88] proposed an improved semi-supervised NMF algorithm based on the frame level. In particular, they estimated the bases coefficient matrices of speech and noise computed using pre-trained speech, and noise bases to avoid noisy speech variability over time. When a new noisy frame is processed the proposed NMF approach is used to train the noise bases, hence it can maintain the dimension reduction, and the computational complexity. Thus, the proposed algorithms can be implemented for real-time speech-processing tasks.

In this research [44], the authors proposed an improved NMF algorithm based on basis compensation. In the enhancement phase, extra basis vectors for clean and noise signals are used in order to capture the features that are missed in the training phase. Especially, the free basis vectors of the clean speech are estimated by utilizing a priori knowledge based on Gamma distribution, Conversely, the free basis vector of the noise relies on prior knowledge of a regularization approach respectively. In this way, it forces the noise-free vector basis to be orthogonal to the clean speech and noise basis vectors estimated during the training stage.

The authors in [45] proposed an algorithm for regularizing the NMF approach. In particular, the speech and noise magnitude spectrum likelihood functions are used as regularization parameters in the NMF cost function. Finally, to improve the speech quality they integrate a masking model based on the human auditory mechanism. The final results showed a clear improvement in terms of speech quality and noise suppression.

In this research [245], they introduced an approach for enhancing speech signals spectrogram using NMF and sparse NMF algorithm. Unlike the traditional spectrograms estimated using STFT, which has a frequency resolution lacking at low frequency. Constant Q-transform is used the provide high resolution at low frequencies, while the back-end remains the sparse NMF. The proposed method outperformed the conventional STFT approaches at low SNR values.

2.3.4 Multi-layer Perceptron Algorithms

A Multi-layer Perceptron (MLP) is a machine learning approach based on feed-forward artificial neural network (ANN) [161, 217]. Generally, the MLP network comprises at least three layers namely input, hidden, and output layers. Each layer contains neurons followed by a non-linear activation function. During the training, the MLP network uses the back-propagation algorithm as a supervised training approach. Thus, the MLP networks can distinguish non-linear data due to the multiple layers with non-linear activation. In the following sections, we review studies that investigate MLP applied to speech enhancement.

In [218], the AMS features are used as inputs for the MLP network to estimate the local SNR, based on it the noise is canceled. In particular, mixture speech signals are represented by spectro-temporal patterns e.g. AMS features, since it carries the information of center and modulation frequencies for the analysis frames.

The authors in [98] introduced novel features based on the pitch for MLP training to estimate IBM. The network is trained using 128-channel Gammatone filterbank features to separate noise signals from the voice segments with input SNR $= 0$, and reverberation time extending from 0.1 sec to 0.6 sec. for each mixture. An objective cost function is used based on maximizing the input SNR during the training showed a notable improvement in the classification process.

Another algorithm is proposed in [83] using CASA as MLP for robust pitch-tracking and segregation. In particular, the tandem approach estimates the target speech pitch and eliminates other segments. The calculated target speech pitch is used to separate the target speech from the noise using temporal and harmonic continuity. A systematic evaluation revealed that the target speech pitch is accurately extracted without noise artifacts.

2.3.5 Deep Neural Networks Algorithms

Deep neural networks (DNN) is a kind of ANN with more hidden layers between input and output layers [116, 190, 249]. Recently, DNN-based algorithms for different research areas gain a lot of attention due to the remarkable improvement in parallel computing resources as well as software and hardware. Regarding the software, the powerful computing platforms introduced by NVIDIA e.g. compute unified device architecture (CUDA) [60] and current deep learning frameworks e.g. Tensorflow [2], and PyTorch [173]. Concerning the hardware, graphics processing units (GPUs) and tensor processing units (TPUs) substantially improve DNN performance. DNN algorithms emerged in many research areas e.g. computer vision [229], machine translation [134], and automatic speech recognition [204].

In the last decades, DNN-based approaches brought a notable improvement in the speech processing area, including speech enhancement, as it doesn't take into account any prior assumptions for speech and noise. Additionally, it outperforms the traditional algorithms discussed in section 2.2. However, DNN-based algorithms show low latency processing which is critical for real-time applications [76].

Generally, DNN-based speech enhancement algorithms are categorized into two main categories namely Frequency-domain and Time-domain according to the network input. In the following, we overview the SOTA of both categories.

Frequency-Domain Approaches

Different DNN architectures including fully connected networks (FC), convolutional neural networks, and recurrent neural networks (RNN) were investigated to analyze 2D or 3D images in the computer vision field. Contrary, speech signals are time-domain 1D signals with correlations between successive samples. Hence, STFT is calculated for 1D speech signals to obtain a 2D representation e.g. spectrogram in the time-frequency (TF) representation. Thus, DNN-based approaches adopted for computer vision can be applied without extra adjustment [160].

Typically DNNs are able to learn the complex relationship between the signal input features, either in time or frequency domains, and the desired training target e.g. speech spectra or spectral masks. Speech enhancement-based STFT representation

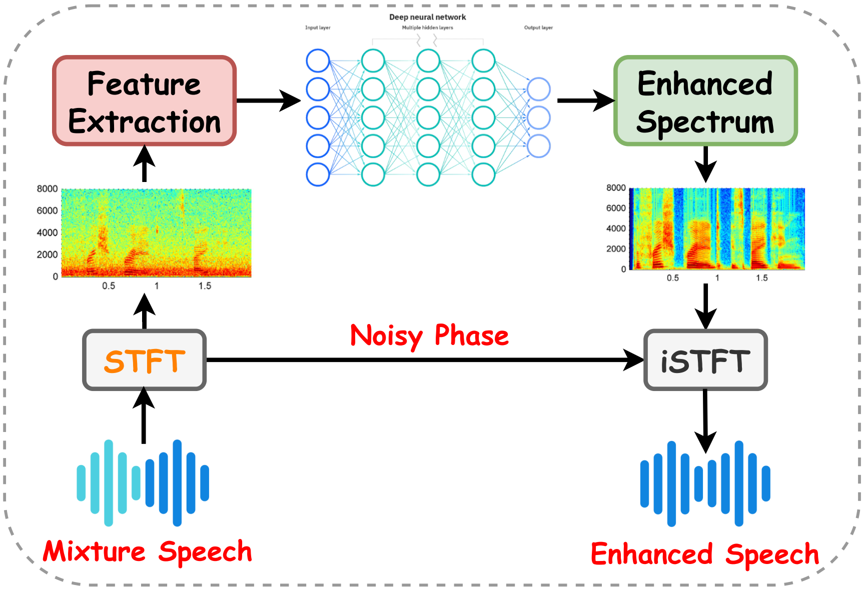


FIGURE 2.9: Schematic diagram of frequency-domain speech enhancement approach.

is concerning either magnitude or phase enhancement as depicted in Fig 2.9. We overview the SOTA in magnitude and phase enhancement.

- **Magnitude Enhancement**
  The authors in [238], which is an extension of [236, 237], proposed DNN-based SE using a masking-based framework. A fully-connected network is utilized for sub-band classification for IBM estimation. The network parameters were initiated using a restricted Boltzmann machine (RBM) which is a stochastic generative neural network. The motivation behind that, during training FC network starts with random parameters, which slows reaching the local minima especially when the model consists of a large number of hidden layers [82]. The network is trained using 64-channel Gammatone filterbank features extracted from each TF unit to learn discriminative features. Finally, SVM is trained using the earned features concatenated with the input to estimate the sub-band IBM. The obtained enhancement results showed substantial improvement with respect to the SOTA approaches.

  The mapping-based approach for speech enhancement was first introduced in [140], where a deep auto-encoder model is used to map the noisy power spectrum to the clean ones. The authors reported that increasing the model complexity leads to better enhancement results. Comparing the proposed model with the MMSE approach, the proposed one shows superior performance. The same approach is employed in [249], where an FC network is used to map the noisy power spectrum to its clean version with RBM used for the same reason explained above. During training, the dropout technique is used to avoid over-fitting. Finally, the authors utilized the acoustic context information i.e. full-frequency band and context frame expansion to reduce discontinuity and achieve better speech quality. The proposed approach not only outperforms the traditional MMSE approach and is able to eliminate non-stationary noise but also generalizes to unseen noise.

Other studies tried to investigate other different training targets in their DNN-based approaches. The authors in [256] attempted to take advantage of the masking-based approach and the mapping-based approach. In particular, their FC network jointly estimated IBM, IRM, and the spectrogram of the clean speech. Obtained results demonstrated that joint mask-spectrum estimation leads to better enhancement performance. The same approach was introduced in [249] adding a multi-input framework. This framework is able to integrate the learned acoustic features e.g. MFCCs, mask representation, and jointly optimize all the parameters. An analogous approach is proposed in [155, 235], the trained a DNN to estimate the ratio mask in the Gammatone domain. The estimated mask is then used to obtain the enhanced signal in the time-domain. A similar contribution is also reported in [243], while the difference is that the estimated mask is based is the Discrete Fourier Transform (DFT) domain.

Different DNN-based based on recurrent neural networks e.g. RNN and LSTM had been investigated in the speech enhancement area. The authors in [241] investigated the LSTM network as a front-end SE for robust ASR in noisy conditions. Precisely, employing the LSTM network as a front-end to provide enhanced speech signals leads to a 13.76% improvement in terms of WER. The same approach was also applied in [133]. In [192], an RNN is trained to learn spectral masking from the magnitude spectrograms of the noisy speech signals integrated with an intelligibility improvement filter used to improve speech intelligibility. Reported experimental results show a notable improvement in speech quality and intelligibility with 17.6%, 5.22, and 19% for STOI, SDR, and PESQ metrics over noisy scores.

The authors in [233] attempted to avoid problems of gradient disappearance and gradient explosion using a gated recurrent network (GRU). Firstly, a DNN network with three hidden layers is used to learn the mapping function between the logarithmic power spectrum (LPS) features of noisy and clean speech signals. Then a GRU network is trained with a feature fusion between the LPS features and noisy speech signals to learn the mapping relationship between LPS features and log power spectrum features of the clean speech spectrum. Obtained experimental results showed that the PESQ, SSNR, and STOI are improved by 30.72%, 39.84%, and 5.53%, respectively, with respect to the noisy metrics.

CNN-based architectures are widely investigated in speech enhancement areas because the weight-sharing property leads to fewer parameters with respect to FC and RNN architectures. The authors in [170] employed a fully CNN network as the mapping-based framework. Each layer consists of a convolutional layer followed by a batch normalization operation and ReLU activation function. The network is trained based on the noisy spectrogram, while the clean speech spectrogram is used as the training target. Obtained results demonstrated promising enhancement results. In [214], the authors combined the convolutional and recurrent layers to form a novel architecture called convolutional recurrent network (CRN). The motivation is to introduce a causal system for real-time applications that is noise and speaker-independent. Experiments showed better quality and intelligibility with fewer trainable parameters.

- **Phase Enhancement**
  Most of the discussed approaches are concerned with enhancing the speech

magnitude, while the phase remains noisy and deployed during signals reconstruction [6], which deteriorates the final performance. Later, it proved the effectiveness of investigating the phase information for SE [211], where the authors proposed that enhancing the phase spectrum along with magnitude improves speech quality and intelligibility in terms of objective and subjective measures.

An MMSE phase estimator approach was proposed in [187] to estimate the phase information for enhanced speech signals reconstruction with prior knowledge of signal spectrum amplitude. In [243] the authors jointly trained a DNN to estimate the real and imaginary part of a complex ideal ratio mask (cIRM), which can be considered a multi-target approach.

The authors in [188], propose a SE-DNN approach combined with a phase estimator to improve speech quality and intelligibility. During the training stage, the DNN learns a mapping function from the noisy speech to estimate the IRM for the spectral magnitude. Then, the temporal smoothing unwrapped spectral phase estimation is employed and transformed into a structured spectral phase during signal reconstruction. In the enhancement stage, the enhanced speech magnitude is reconstructed with the estimated structured spectral phase.

Recently the authors in [77], reports major limitation of SE DNN-based approaches: (a) Most of these approaches discard the phase spectrum information. (b) More computational resources and memory requirements are required to train these models. Thus they have limited usage in real-time applications. Towards solving these issues, they proposed a phase-aware composite deep neural network (PACDNN) to simultaneously estimate the magnitude processing with a spectral mask and phase reconstruction using the phase derivative approach. Exhaustive experiments yielded better enhancement performance with respect to SOTA approaches with lower computational complexity and memory consumption. A similar approach is proposed in [174], the authors conducted a systematical study on the contribution of phase and magnitude in modern SE DNN-based approach at different frame lengths. Systematical analysis showed that adequate choice of frame length is a critical parameter of designing SE STFT-based systems as it controls the system latency needed for real-time applications. In particular, short frame length deteriorates the algorithm performance, while a large number of frames significantly increases the computational complexity.

Time-Domain Approaches

Time-domain approaches are alternative algorithms that operate directly on the raw waveform [8]. Recently, approaches operating in the time domain have emerged to mitigate the phase estimation problem, introduced in frequency domain approaches, which improve speech quality and intelligibility [231]. In [164], the authors proposed an approach based on a fully-CNN architecture. The network input is the noisy speech signals and the output is the corresponding enhanced ones. During the training phase, the MSE-based frequency domain loss function was used to train the time domain framework. In this way, they avoid the STFT phase estimation problems leading to better enhancement performance. The same approach was utilized in [166], improving the generalization as the model is trained in a speaker- and noise-independent.

A Dense-CNN model is proposed in [165] with a self-attention mechanism. The model architecture is an encoder-decoder with skip connections, each layer in both the encoder and the decoder consists of a dense block followed by an attention block that helps in features extraction. Finally, the model is trained based on a novel loss function based on the magnitudes of the enhanced speech and the predicted noise.

As the Convolution-augmented transformer (conformer) showed a substantial improvement in speech-domain applications, such as automatic speech recognition, and speech separation. The authors in [105] exploit the Conformer performance in the SE task, as it is able to capture both the short and long-term temporal sequence information by attending to the whole sequence at once with multi-head self-attention and convolutional neural network. The experimental results showed the proposed model outperforms other baselines (i.e. HiFi-GAN [209], DeepMMSE [258], and DEMUCS [49]) in terms of standard SE evaluation metrics.

SEGAN proposed in [171, 172], was the first attempt that investigates the use of generative adversarial networks (GAN) for speech enhancement. Practically, GAN architecture has two sub-networks namely Generator (G), and Discriminator (D). The G component is trained for mapping tasks, while the D component, which is a binary classifier decided that the inputs are either real samples or synthetic ones. The same approach was later developed towards UNet-GAN architecture [74], where the G component is replaced by the U-Net model, an encoder-decoder model that employs dilated convolution in the bottleneck of it.

2.3.6 Limitation of Supervised Speech Enhancement Algorithms

Supervised speech enhancement algorithms outperform Unsupervised algorithms in terms of speech quality, intelligibility and generalization capability for unseen noisy conditions especially in case of DNN-based approaches. However, they have serious limitations discussed below:

- **GMM algorithm:** The model is statistically ineffective in the case of modeling the data that is located on or located near a nonlinear manifold in the data space. Thus, it fails in modeling the acoustics of speech signals.

- **SVM algorithm:** SVM can not handle large data sets. Moreover, its performance deteriorates in case of high overlapping between speech and noise classes. Finally, the SVM adjusts the data points, above and below the defined hyperplane, there is no probabilistic clarification for the classification.

- **Non-Negative Matrix Factorization:** These algorithms have a lack of generalization capabilities, especially in presence of multi-noisy conditions. Moreover, it needs powerful computational resources.

- **DNN-based algorithms:** Generally, they require large datasets to obtain good performance. In the case of speech processing applications e.g. speech recognition and speech enhancement, large datasets are available online with very few stereo data available for speech enhancement. Therefore DNNs are implausible to perform better than other competing methods. Moreover, these algorithms are extremely computationally complex, as it needs long time periods to train the model using powerful GPUs.

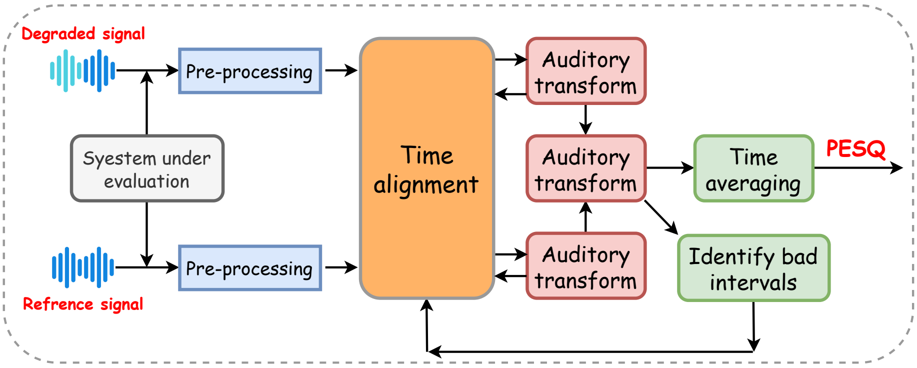


FIGURE 2.10: Block diagram of PESQ measure computation [137].

2.4 Speech Enhancement Evaluation Metrics

Quality and intelligibility are two attributes used to evaluate the processed speech signals [137, 138]. Quality is a subjective metric used to measure how utterances are produced and involve some attributes e.g. natural, raspy, hoarse, and scratchy [95]. Unlike quality, intelligibility can be measured by giving processed speech signals (sentences, words, etc.) to a group of listeners and asking them to identify these spoken words. Then intelligibility is computed using the correct identified words or phonemes [137].

Many research studies had been conducted to develop objective evaluation measurements to estimate speech quality and intelligibility with high correlation. These measurements are based on mathematical representations between clean and enhanced speech signals [137]. In general, most of the objective metrics are based on the time or frequency domain features that are extracted from the clean and noisy signals for similarity index measurement.

Frequency-weighted segmental SNR (fwsegSNR) relies on calculating the geometric mean of the SNR for all speech frames [137]. Other types of objective metrics depend on the speech features e.g. spectral distance of Linear prediction coefficients (LPC). For example, the log-likelihood ratio (LLR) metric, is used for similarity prediction between clean and processed speech signals [176] based on all-poles models estimated from LPC coefficients. The weighted spectral slope (WSS) is used to determine the difference either of formants or spectral peaks location, by searching for the spectral slope for all bands of speech frequencies [110].

Perceptual evaluation of speech quality (PESQ), as depicted in Fig. 2.10, is a common objective metric that is used to evaluate speech distortion, packet loss, codec distortion, and speech quality [137]. This metric is recommended by the international telecommunication union (ITU) to be used as an objective metric for speech quality (P.862 standard), and its range lies between (-0.5 to 4.5) i.e. higher score means signals quality. This metric is based on the time alignment approach to compensate for the delay between the clean and noisy signals. Hence, it applies a transformation to equalize the linear filtering and gain variation to achieve the loudness spectra [179]. Despite the effectiveness of the PESQ metric, it has some limitations including listening levels, loudness loss, effects of delay in conversational tests, talker echo and side tones [81].

In [86], the authors exploit the performance of PESQ, SNR, LLR, and WSS metrics to measure speech signal quality obtained using spectral subtraction, subspace, and Wiener filter algorithms. This research also proposed other objective metrics called composite metrics (Csig, Cbak, and Covl), which is a linear combination of the above-mentioned objective metrics to predict speech quality [87].

The short-time objective intelligibility measure (STOI) proposed in [213], uses 384 ms long blocks containing excitation spectra of the clean and processed signals. This metric computes the average of the correlations across all 1/3-octave bands and 384-ms blocks and uses it to predict speech intelligibility assigning a score $\in (0, 1)$ i.e. higher score means signals intelligibility. Prior to the correlation computation, the processed envelope was normalized and clipped as follows:

$$\bar{y}_{(j,m)} = min\left(\frac{||x||_2}{||\hat{x}||_2}\hat{x}, (1 - 10^{-\beta/20}x\right) \tag{2.25}$$

Where $\beta = -15dB$.
x, and $\hat{x}$ denote the clean and enhanced envelope vectors respectively.
$||.||_2$ represents the vector 2-norm.

The β parameter that controls the clipping operation is effective primarily in noise-only regions. Thus, it is employed to mitigate the impact of those regions on speech intelligibility.

Table 2.1 summarizes the most common speech enhancement metrics, their mathematical representation, and the purpose of using them.

TABLE 2.1: Common speech enhancement evaluation metrics

Metric	Equation	Measure
PESQ	$\alpha_0 - \alpha_1.A_{ins} - \alpha_2 B_{ins}$	Speech quality
LLR	$\log \frac{\overrightarrow{b_x} R_x \overrightarrow{b_x^T}}{\overrightarrow{b_{\hat{x}}} R_x \overrightarrow{b_{\hat{x}}^T}}$	Speech quality & Spectral distance
segSNR	$\frac{10}{M}\sum_{m=0}^{M-1} \log_{10}\left(\frac{\lvert S(m,\omega_m\rvert^2}{\lvert S(m,\omega_m\rvert - \lvert\hat{S}(m,\omega_m\rvert\rvert^2}\right)$	Speech quality & noise suppression
fwsegSNR	$\frac{10}{M}\sum_{m=0}^{M-1}\left(\frac{\sum_{j=1}^{k} B_j \log_{10}[\frac{F^2(m,j)}{F(m,j)-\hat{F}(m,j)}]}{\sum_{j=1}^{k} B_j}\right)$	Speech intelligibility & Speech quality
Csig	$3.093 - 1.029$ LLR $+0.603$ PESQ -0.009 WSS	Speech Distortion & Residual noise
Cbak	$1.634 + 0.478$ PESQ -0.007 WSS $+0.093$ segSNR	
Covl	$1.549 + 0.805$ PESQ -0.512 LLR -0.007 WSS	
STOI	$\bar{y}_{(j,m)} = min\left(\frac{\lvert\lvert x\rvert\rvert_2}{\lvert\lvert\hat{x}\rvert\rvert_2}\hat{x}, (1 - 10^{-\beta/20}x\right)$	Speech intelligibility
SNR	$10\log_{10}\frac{X(k,m)^2}{\hat{X}(k,m)^2}$	Speech intelligibility

2.5 Speech Enhancement for Robust Speech Classification: (ASR Case Study)

Most of the SOTA SE approaches are designed to improve the perceptual quality measured using SE metrics. Nevertheless, due to the inconsistency between the training objectives of the SE and ASR modules, these improvements do not always have a positive impact on the ASR performance, especially in case of multi-noise conditions. Recently, different research attempts have been made to optimize the SE module to maximize the performance of the subsequent downstream ASR task. Table 2.2 summarizes part of these research paper.

TABLE 2.2: Summary of research papers investigate SE for ASR in noisy conditions

SE architecture	Dataset	Domain			Average (WER) (CER)%	
		Frequency	Time	SSL Feat.	Noisy	Enhanced
CycleGAN [123]	CHiME-3	✓			61.46	52.80
TENET [31]	DEMAND QUT-NOISE	✓			23.76 82.32	6.76 26.50
6-layers DNN [198]	MATBAN	✓			68.77	57.47
CRN (MCG) [232]	Noisy AISHELL1	✓			20.984	16.882
GAN [132]	Noisy AISHELL1	✓			51.5	49.1
Dense CRN [167]	Social media English video	✓			17.4	11.2
DCCRN [118]	Noisy Librispeech	✓			16.43 9.26	15.54 9.07
Dense CNN [101]	Noisy Librispeech	✓			34.04	15.46
Conformer [111]	Noisy Librispeech	✓			10.5	9.2
BiLSTM [193]	Noisy Libri-light	✓			11.0	10.0
TASNet [109]	CHiME-4 Aurora-4		✓		12.23 8.5	8.19 6.3
MC Conv-Tas Net [259]	CHiME-4		✓		19.5	10.7
Residual-CNN [158]	CHiME-4		✓		13.44	8.56
Conv-TAS Net [30]	CHiME-4		✓		6.36	4.93
Attention Wave-U-Net [66]	VCTK Simulated Data		✓		11.55 34.42	10.69 26.33
CRN [128]	In-house corpus		✓		7.43	6.19
DEMUCS [53]	DNS		✓		7.5	5.01
CNN+LSTM [203]	DNS-3	✓		✓	24.72	15.982

2.5.1 Discussion

With the advent of deep learning, research on noise-robust ASR has increased significantly. However, improving the ASR performance in noisy conditions is still a challenging task. Multiple frequency-domain models are employed as a pre-processing SE stage; subsequently the ASR is trained on the enhanced features. The drawbacks of these approaches, as mentioned previously, is that they employ the noisy phase while reconstructing the enhanced signals. Moreover, most of these SE front-end modules are trained separately from the ASR module. For this reasons, they often introduce speech distortions that degrade the ASR performance. A clear evidence of this is that the enhanced WER metric is relatively high, often approximately similar to the noisy WER as in [111, 118, 193] .

Motivated by the unprecedented breakthroughs of time-domain SE deep learning approaches, that showed outstanding enhancement performance, recent SE models are trained directly to map the noisy speech into its clean counterpart. The key strength of these approaches is that they get rid of the hand-crafted features e.g. spectrograms, F-banks features, commonly used in the frequency-domain approaches. These approaches improves the quality metrics as well as the WER compared with noisy ones.

Latterly, features augmentations bring a remarkable improvement in speech processing research as in [203], where a combination between the noisy speech embeddings, obtained from a large scale pre-trained *WavLM* model [37], with the STFT features are fed the SE module to estimate the enhanced speech. This approach shows promising performance even with a limited amount of training data.

Chapter 3

Robust Intent Classification in Noisy environments

This Chapter gives some insights into the robustness of back-end speech classification tasks, especially intent classification in noisy environments. Section 3.1 provides a general overview of the intent classification task. Section 3.2 describes each module in the proposed pipeline that integrates the speech enhancement with the intent classifier. Experimental results are presented and analyzed in Section 3.3, followed by the conclusion in Section 3.4. The results discussed in this chapter were reviewed and published in the Proceedings of European Signal Processing Conference (EUSIPCO), 2021 [9] [1].

3.1 Overview on Intent Classification Task

Spoken Language Understanding (SLU) is a research field that has inspired the interest of scientific communities referring to the natural language processing (NLP) area for many years. Nowadays, spoken dialogue interaction, in a natural way, is possible with several commercial products, such as the most known personal assistants (Google Home, Amazon Alexa, Siri, Microsoft Cortana, etc), and can be implemented with a set of toolkits, both commercial (e.g. dialog flow [186]) and open source (e.g. Rasa, Opendial, [19, 131]).

The fundamental function of SLU systems is to understand the intents of the users, which causes the execution of "actions" aimed to fulfill their requests. For example, in smart home applications an utterance like "increase the sound" might correspond to an intent represented with the following filled slots: action: "increase", type: "sound", count: "None", place: "None". A survey reporting fundamentals of SLU technology can be found in [64, 222].

The Intent Classification (IC) task is usually accomplished by applying natural language understanding (NLU) techniques to the output of an ASR system, to produce a semantic interpretation of the input speech as described in Fig. 3.1(a). Recently, approaches that perform this task in an end-to-end (E2E) fashion, shown in Fig. 3.1(b), have started to be investigated and produced an excellent performance on several datasets. The E2E paradigm uses a single neural model to map a spoken

[1] https://ieeexplore.ieee.org/abstract/document/9616322

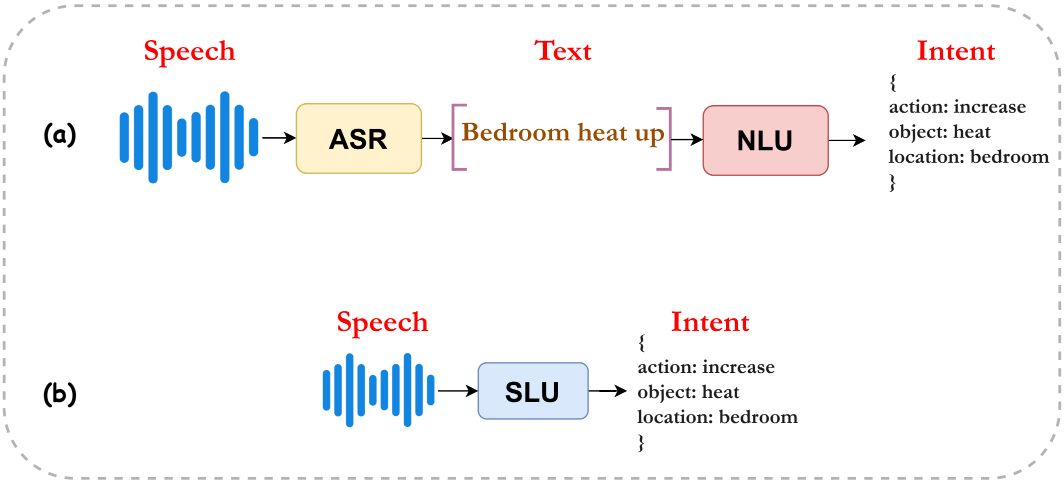


FIGURE 3.1: Conventional SLU pipeline versus E2E2 SLU pipeline.

input into the corresponding intents, thus optimizing directly the classification metrics and avoiding error propagation caused by ASR errors. Some interesting models and related results in this direction can be found in the works reported in [71, 143, 175, 197].

Unfortunately, as in ASR systems, environmental noise deteriorates the quality and intelligibility of speech signals, resulting in low intent classification accuracy [225]. To mitigate the impact of noise, a possible approach consists in training, or adapting, the classification model on the noisy data [253]. This can be done either by collecting application-specific data or through the usage of data augmentation strategies [25]. However, acquiring large sets of noisy data is costly and time-consuming while, in general, all possible noisy conditions cannot be known a priori making unfeasible the data augmentation-based approach. Therefore, an alternative method is to use a speech enhancement front-end to improve classification accuracy.

To tackle the problem of IC in noisy environments, we employ a speech enhancement front-end to mitigate the noise impact on the speech signals before processing them with the IC back-end. Fig. 3.2 shows the complete pipeline of the proposed approach.

More in detail, we use an improved version of the Wave-U-Net: a deep learning speech enhancement front-end [146] which is an extension of the model introduced for audio source separation in [207]. Regarding the IC task, we exploit here in after a convolutional deep neural network with residual layers named temporal convolutional network (TCN), which allows for achieving state-of-the-art performance. Finally, to make the IC task robust against out-of-vocabulary sentences, we introduce a multi-task learning framework by predicting each intent element disjointly.

3.2 System Description

The following subsections describe each component of the pipeline depicted in Fig. 3.2.

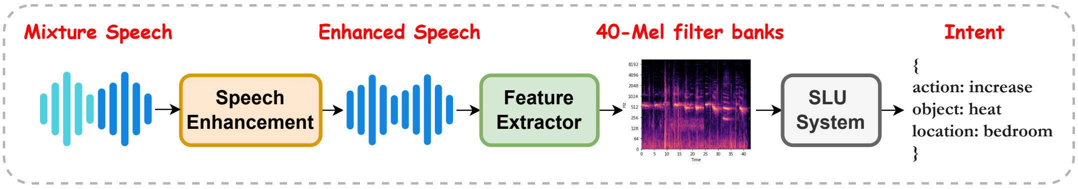


FIGURE 3.2: The full pipeline of our intent classification scheme, including speech enhancement and intent classifier.

3.2.1 Wave-U-Net for Speech Enhancement

As discussed in Chapter 2, time-domain speech enhancement approaches allow the achievement of promising results in comparison with other techniques. Among them, the U-Net architecture proposed in [183], later successively improved towards Wave-U-Net [146, 164] has obtained encouraging results.

Wave-U-Net comprises 3 components [6] as depicted in Fig. 3.3: (a) an encoder network consisting of multiple 1-D fully convolutional down-sampling blocks; (b) a bottleneck 1-D convolutional layer; and (c) a decoder network made by a stack of 1-D fully convolutional up-sampling blocks. Note that skip connections are used between each down-sampling block and its corresponding up-sampling counterpart.

In detail, the network input is a vector of noisy speech signals $z[n] \in [-1,1]^{L\times C}$, $n = 0, ..., L-1$, where L represents the number of samples and C is the number of input channels. During training, low-dimensional high-level features are computed at different time scales through a series of down-sampling blocks. These features are then concatenated with their corresponding local, and high-resolution features extracted through the up-sampling blocks. In the case of monaural speech enhancement, the network is trained to map noisy signals $z[n]$ to its enhanced counterpart

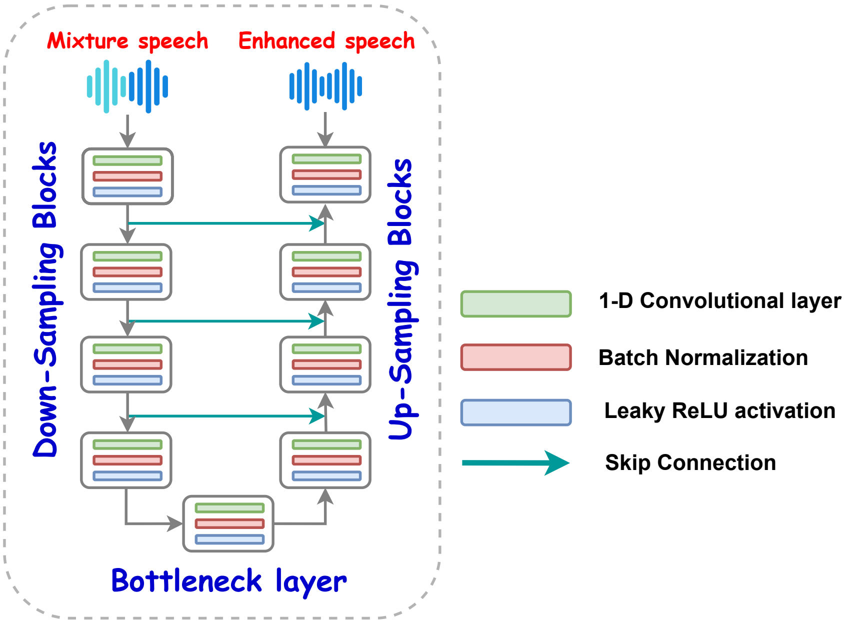


FIGURE 3.3: Schematic diagram of the Wave-U-Net model for speech enhancement.

$\hat{s}[n]$ using clean signals $s[n]$ as the training target. The model attempts to minimize the MSE loss between $\hat{s}[n]$ and $s[n]$ i.e.:

$$\mathcal{L}_{SE} = \sum_{n} \|s[n] - \hat{s}[n]\|^2 \tag{3.1}$$

3.2.2 Dilated Encoder Wave-U-Net

The proposed modified model follows the basic Wave-U-Net architecture. Recently, using dilated convolutional layer achieves promising performance on time-series data as it captures long-term information without increasing the computational complexity [24, 251].

In particular, the model comprises four downsampling and four upsampling blocks. Each downsampling block consists of three 1D convolutional layers with "kernel size = 15", and "stride = 1". While in the original Wave-U-Net architecture the padding value is fixed and equal to 7, in the modified model the padding value is doubled successively i.e " padding = 7, 14, 28".

The key difference between the original and the adopted architecture is that the original Wave-U-Net architecture uses a constant dilation factor i.e. "dilation = 1", while in the modified architecture the dilation factor is increased exponentially from layer to layer i.e. "dilation = 1, 2, 4" respectively.

In both architectures, each convolutional layer is followed by a 1D-Batch normalization layer and the Leaky ReLU activation function with a negative slope "α = 0.1". The bottleneck layer is a 1D convolutional layer with "kernel size = 15", "stride = 1", and "padding = 7". The network's right side consists of the same number of blocks i.e up-sampling blocks with the same number of non-dilated convolutional layers. Finally, a 1-D convolutional layer with "kernel size = 1", and "stride = 1" is set on top of the model followed by the Tanh activation function to produce the enhanced speech signals.

3.2.3 Temporal Convolutional Network for Intent Classification

For intent classification, we propose to use a multi-class architecture to directly map the enhanced input features into the corresponding intents. Our proposed model is based on the separation part of Conv-TAS-Net, originally introduced for the speech separation task [145].

The model depicted in Fig. 3.4, processes 40-Mel filter banks computed on a 20 ms window size, with a 10 ms step. It applied a global layer normalization (gln)(see Eq. 3.2) and a 1-D convolutional layer (Conv 1×1 as depicted in Fig. 3.4(a)) that maps the input features into 64 bottleneck channels.

$$gln(F) = \frac{F - E[F]}{\sqrt{Var[F] + \epsilon}} \odot \quad \gamma + \beta \tag{3.2}$$

Where $F \in \mathbb{R}^{N \times T}$ is the tensor of features, $\beta, \gamma \in \mathbb{R}^{N \times 1}$ are trainable parameters, $E[F]$ represents the mean feature vector, $Var[F]$ is the related variance, and ϵ is a small constant added for numerical stability.

This layer is followed by two repetitions (R = 2) of five consecutive 1-D dilated convolutional residual blocks (B = 5) with skip connections. Each residual block as

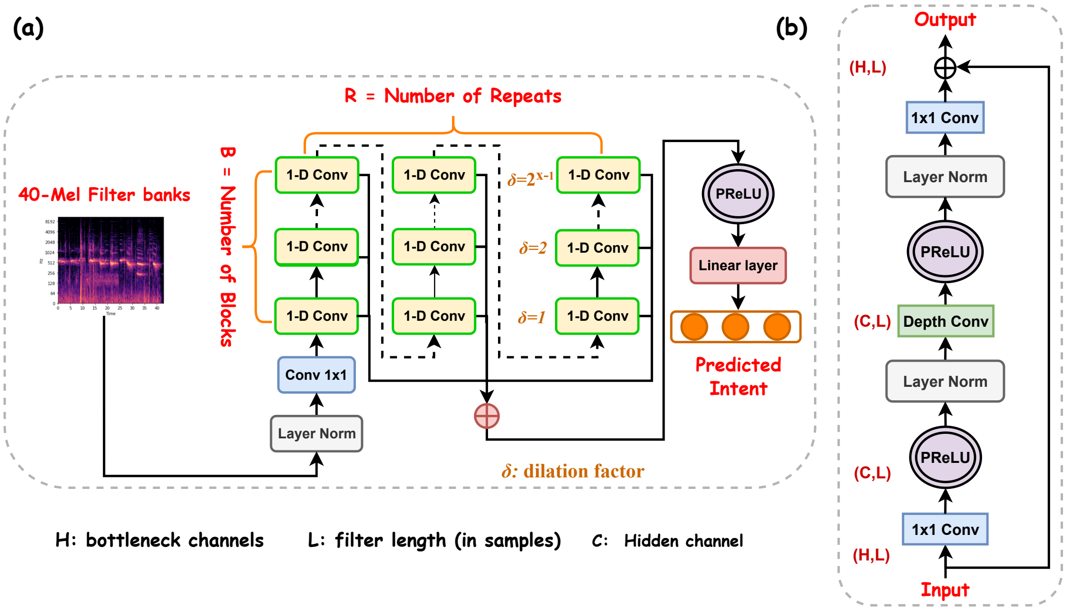


FIGURE 3.4: (a) Block diagram of the TCN classifier. (b) of a single 1-D dilated convolutional block.

shown in Fig. 3.4(b) consists of two symmetrical pipelines with a depth-wise separable convolutional layer that maps the 64 bottleneck channels into 128 channels. Each pipeline has a pointwise convolution (1x1 Conv block) followed by a gLN with Parametric Rectified Linear Unit (PReLU) activation function. A pointwise convolution is applied at the input and as a final operation. A residual branch connects the original input to the output. Mean pooling is applied to the output of the last block, followed by gLN and a linear layer. It is worth mentioning that, the dilation factor is increased exponentially in every successive residual block.

The IC classifier is trained to estimate the target intent by minimizing the cross entropy loss between the predicted and actual labels as illustrated in Eq. 3.3.

$$\mathcal{L}_{IC} = -\frac{1}{T}\sum_{t}\log(p_t) \tag{3.3}$$

where T is the number of training samples and p_t is the probability of the t^{th} target sample.

We consider two different training strategies: joint and dis-joint classification. In the joint classification task, shown in Fig. 3.5(a), the model estimates the whole components associated with the desired intent simultaneously. In the case of a disjoint classification strategy, the three components of each intent (i.e. action, object, location) are classified independently [61]. The latter strategy can be considered as a multi-task learning approach since the final classification layer is split into three distinct tasks as depicted in Fig. 3.5(b). During inference, the three predicted parts are combined to form the predicted intent. On one hand, this is a more difficult task as it allows the prediction of non-existing intents when joining the three parts. On the other hand, it is supposed to be more robust in case of out-of-vocabulary utterances (e.g. ways to express intents that are not available in the training dataset) or in case of unseen intents are present in the test dataset.

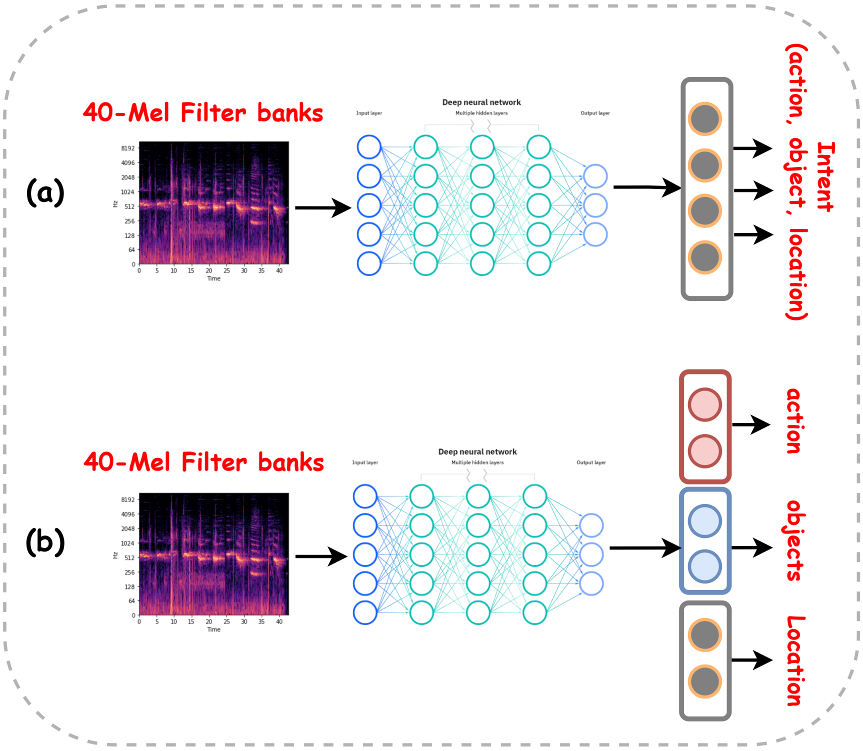


FIGURE 3.5: The intent classification strategies: (a) Dis-joint strategy. (b) Joint strategy.

3.3 Experimental Results

The speech enhancement front-end is trained using a noisy version of the Librispeech-100 dataset described in Section 1.5. Randomly, 10 hours of clean speech are selected, and contaminated by adding noise from the MS-SNSD dataset described in Section 1.5.

The noisy Librispeech is generated by randomly selecting a random noise file available in the MS-SNSD dataset and is added to the clean signal with one out of five SNRs: 5 dB, 7.5 dB, 10 dB, 12.5 dB, and 15 dB. The dataset was split into three portions: 6 hours, 2 hours, and 2 hours for training, validation, and testing respectively. The FSC dataset, described in Section 1.5, is also contaminated with similar procedures using the MS-SNSD library with the same SNR ranges plus two more levels -5 dB, and 0 dB.

Both the original Wave-U-Net [2] and the modified model are trained using the noisy Librispeech dataset. To handle signal length variation, the network is designed to process fixed-length input signals taking 16384 continuous samples randomly selected from the noisy and clean speech signals. Both models are trained using Adam optimizer with learning rate =10^{-4}, decay rates $\beta 1$ = 0.9, and $\beta 2$ = 0.999, "batch size = 10", and as previously mentioned the MSE is used as loss function. Finally, to investigate the generalization capability of the trained models to out-of-domain noisy data, we utilize the Librispeech-trained Wave-U-Net models to denoise the noisy FSC dataset.

[2] https://github.com/haoxiangsnr/Wave-U-Net-for-Speech-Enhancement

Algorithm 1 Pseudo-code for training models

```
Require: Wave-U-Net initialization.
Require: Number of Epochs = 500.
  for $i \in$ Number of Epochs do
      Forward Pass:
      Starting from the input layer do a forward pass.
      (with batch normalization) through DNNs.
      Compute the speech enhancement loss function $\mathcal{L}_{SE}$.
      $\mathcal{L}_{SE} = \sum_n \|x[n] - \hat{x}[n]\|^2$          ▷ MSE loss based on the waveform
      Backward Pass:
      Compute the gradient of the enhancement loss $\nabla\mathcal{L}_{SE}$ and backpropagate it.
      Parameter Update:
      $\Theta_{SE} \leftarrow \Theta_{SE} - \lambda_1[\alpha\nabla\mathcal{L}_{SE}]$
  end for
Require: Evaluate the front-end with Enhancement metrics
Require: Generate the Enhanced signals.
Require: Intent classifier initialization.
Require: Number of Epochs = 100.
  for $i \in$ Number of Epochs do
      Forward Pass:
      Extract 40-Mel Filter banks features.
      Compute the intent classifier loss function $\mathcal{L}_{cl}$.
      $\mathcal{L}_{IC} = -\frac{1}{T}\sum_t \log(p_t)$          ▷ Cross-entropy loss
      Backward Pass:
      Compute the gradient of the classifier loss $\nabla\mathcal{L}_{IC}$ and backpropagate it.
      Parameter Update:
      $\Theta_{IC} \leftarrow \Theta_{IC} - [(1-\alpha)\lambda_2\nabla\mathcal{L}_{IC}]$
  end for
  Compute the accuracy on the validation dataset.
  if $\{acc_{dev}^{i+1}\} > \{acc_{dev}^{i}\}$ then
      Save the back-end model.
  end if
  $i+=1$
```

3.3.1 Speech Enhancement Results

The performance of the enhancement process is evaluated using a set of quality and intelligibility metrics: PESQ, STOI, and SNR discussed in Section 2.4.

First, we experiment with the performance of both Wave-U-Net models (Basic and Dilated), results are reported in Table 3.1. We trained both models with MSE loss, L1 loss, and a combination of them using $\alpha = 0.2$ and, 0.8 to control the weight of each loss as follows:

$$\mathcal{L}_{comb} = \alpha\mathcal{L}_{MSE} + (1-\alpha)\mathcal{L}_{L1} \tag{3.4}$$

As shown in Table 3.1, for both models, the MSE loss function outperforms the L1 norm loss, especially in terms of the PESQ metric. Despite the noticeable improvement in the PESQ score, both STOI and SNR metrics don't exhibit large differences between the two losses.

The use of the combined loss function does not improve with respect to the MSE loss, especially in the PESQ and SNR metrics. We can conclude that the highest scores are

TABLE 3.1: PESQ, STOI, and SNR for both Wave-U-Net models using MSE, L1 and combined Loss Functions

Model	loss	PESQ	STOI	SNR
	Unproc.	1.30	0.70	11.52
Basic Wave-U-Net	L1	1.93	0.74	12.89
	MSE	2.13	0.76	13.42
	$\alpha = 0.8$	1.95	0.76	12.95
	$\alpha = 0.2$	1.94	0.74	12.91
Dilated Wave-U-Net	L1	2.48	0.78	13.86
	MSE	2.37	0.78	13.95
	$\alpha = 0.8$	2.38	0.78	13.72
	$\alpha = 0.2$	2.24	0.78	12.93

TABLE 3.2: PESQ, STOI, and SNR for the Wave-U-Net models using Librispeech and FSC datasets based on MSE loss

Data sets		**PESQ**	**STOI**	**SNR**
Librispeech	Unproc.	1.30	0.70	11.52
	SE-GAN [171]	1.85	0.72	11.71
	Wave-U-Net	2.13	0.76	13.42
	Dilated Wave-U-Net	2.37	0.78	13.95
FSC	Unproc.	1.79	0.62	8.68
	SE-GAN [171]	2.15	0.64	11.35
	Wave-U-Net	2.68	0.67	11.06
	Dilated Wave-U-Net	3.09	0.73	11.30

obtained using the MSE loss function.

Table 3.2 reports the enhancement metrics on the contaminated Librispeech-100 and FSC datasets using both the original Wave-U-Net model and our proposed dilated encoder Wave-U-Net, considering SEGAN approach [171] as the baseline. For Librispeech, evaluation metrics are computed on the 2-hours official testing partition. Conversely, for the FSC dataset, the metrics are reported considering the whole dataset (as the models are trained on the noisy Librispeech-100 training set).

The dilated encoder Wave-U-Net clearly outperforms the conventional Wave-U-Net model as well as the SEGAN model in all three metrics. Despite the clear improvement achieved in terms of PESQ, and STOI metrics especially in the case of the FSC dataset. This improvement points out that the dilated encoder Wave-U-Net model not only removes noise but also preserves the spectro-temporal properties of the signals.

3.3.2 Intent Classification Results

We evaluate the impact of speech enhancement on the intent classifier performance using IC accuracy, which measures the actual match between the predicted and the ground-truth intent slots.

TABLE 3.3: Intent classification accuracy on FSC, using clean, noisy, and enhanced signals. Models are trained on clean data

	Full Data		**50% out of voc.**	
Evaluation Data	Disjoint	Joint	Disjoint	Joint
Clean	98.3%	98.8%	88.1%	84.8%
Noisy	63.2%	61.1%	42.3%	41.6%
Wave-U-Net	61.6%	64.2%	50.4%	47.7%
Dilated Wave-U-Net	75.3%	77.7%	65.1%	62.5%

Table 3.3 reports the classification accuracy when applying the model trained on the clean FSC dataset and evaluated on clean, noisy, and enhanced signals in the FSC test dataset.

First of all, we highlight the solidity of our back-end model as the performance on the FSC clean dataset is 98.8%, which is in line with the state-of-the-art. Although the conventional Wave-U-Net model brings significant improvement in terms of the signal quality metrics as reported in Table 3.2, the same trend is not observed in the intent classification in the case of noisy data.

Contrary, the proposed dilated Wave-U-Net model shows a substantial improvement, lifting the classification accuracy from 61.1% to 77.7%. Considering the two training strategies, the "joint" one is in general better, as expected, but the gap with the "disjoint" approach is not so wide.

To evaluate the generalization capabilities of the proposed classifier, we consider an experimental training setup where 50% of the utterances for each intent are removed from the training set. Therefore, for each intent, an average of 4 utterances out of 8 in the test set haven't been seen in training. This 50% is randomly selected and results are averaged on the two halves. Results are reported in the right part of Table 3.3.

As expected in this case, we observe a performance deterioration with respect to using the full dataset. Speech enhancement provides similar improvements to the full data case. Note that the disjoint classification strategy provides a small but consistent improvement with respect to the joint approach. This supports our hypothesis based on the fact that predicting the intent components disjointly helps in the case of unseen utterances.

3.4 Concluding Remarks

In this chapter, we propose a pipeline that integrates a speech enhancement front-end based on a modified version of Wave-U-net called dilated encoder Wave-U-Net. Both front-end models i.e. the conventional Wave-U-Net and the modified architecture are used as a pre-processing stage to robust the intent classification task in noisy environments.

Exhaustive experiments reported that our proposed speech enhancement not only improves the speech quality and intelligibility metrics but also it improves the final intent classification accuracy calculated based on a noisy version of the FSC dataset.

A natural extension of the proposed approach is to investigate the joint training strategy of speech enhancement and intent classification models. The key idea is to concatenate both modules and jointly optimize their parameters. In this way, the intent classification model can guide the enhancement front-end to provide more suitable and more discriminative enhanced signals.

Chapter 4

Time-Domain Joint Training Approaches

This Chapter gives some insights into how can joint training approach mitigate the conventional dis-joint training approach as illustrated in Section 4.1. Section 4.2 overviews the recent contributions on joint training SE with different speech-based tasks. The proposed joint training approach is explained in detail in Section 4.3. finally, the experimental results are presented, discussed, and concluded in Section 4.4 and Section 4.5, respectively. The results discussed in this chapter were reviewed and published in MDPI, Sensors Journal [8] [1].

4.1 Introduction

Building upon our work discussed in Chapter 3, we continue to address the IC task in noisy environments. In detail, we propose a pipeline that integrates both time-domain approaches: Wave-U-Net, for SE, and the TCN for IC as depicted in Fig. 4.1. In detail, we investigate different configurations to jointly optimize end-to-end neural models for both SE and IC in the time domain.

This Chapter extends our previous published research in [9] and discussed in Chapter 3, where we investigated the impact of employing pre-trained SE models on the intent classifier performance in noisy conditions. The key difference is the experiments presented in Chapter 3 did not consider joint training of the two models. Moreover, the back-end is trained based on the 40-Mel filter banks features.

4.2 Overview on Jointly Training SE with Speech Tasks

In the literature, three approaches have been considered to jointly optimize a SE front-end with different speech-based applications. The first approach requires to train the back-end component (i.e. IC task in the case of this work) on clean speech signals, while at the inference phase a SE front-end is employed to mitigate the noise effect [156]. The main limitation of this approach is that the SE introduces signal distortion that is unseen in the training phase. However, this approach is still effective in robust speech-based applications in noisy environments.

[1] https://www.mdpi.com/1424-8220/22/1/374

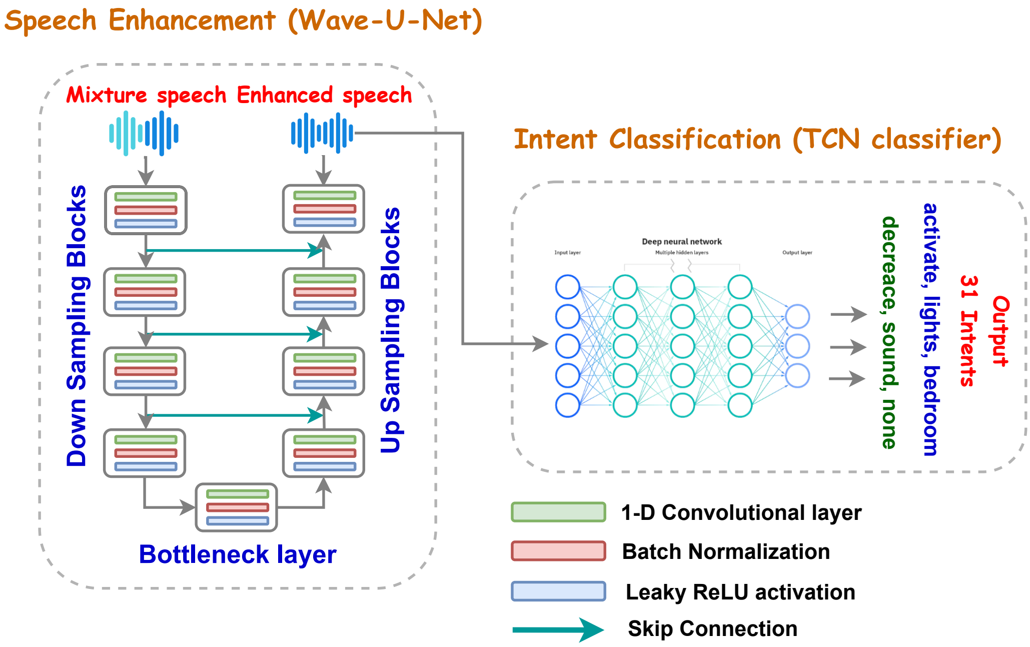


FIGURE 4.1: The full pipeline of our IC scheme, including SE and intent classifier.

To tackle this issue, the second approach firstly enhances the noisy features, which are then used to train the back-end component. Although it was demonstrated somehow effective [195], it was found that it is better to train the back-end on noisy datasets if they contain enough samples of the noise present in the operating field conditions.

In the third approach, the back-end component is trained on the noisy speech features, while at the inference the noisy features are either enhanced first using the enhancement module or fed directly to the back-end. Despite some promising performance achieved with this approach [228], it exhibits poor performance in unmatched conditions [124]. In summary, each approach has its own strengths and weaknesses, depending on the desired application domain.

The joint training approach proposed in this thesis attempts to jointly optimize the parameters of the neural (front-end) of a SE task and of the neural (back-end) model designed for a speech classification task (e.g., ASR, keyword spotting, or IC). In this way, the back-end model guides the whole process and forces the SE front-end to provide a more discriminative "enhanced speech" desired by the back-end.

Fig. 4.2 shows the conventional joint training schematic diagram. The two losses introduced for the front-end and the back-end will be combined in a total loss, as explained later in Section 4.3. To the best of our knowledge, most joint training approaches use ASR, voice activity detection, or keyword spotting as back-ends. In the following subsections, we overview the recent research that addresses joint training SE with these back-ends.

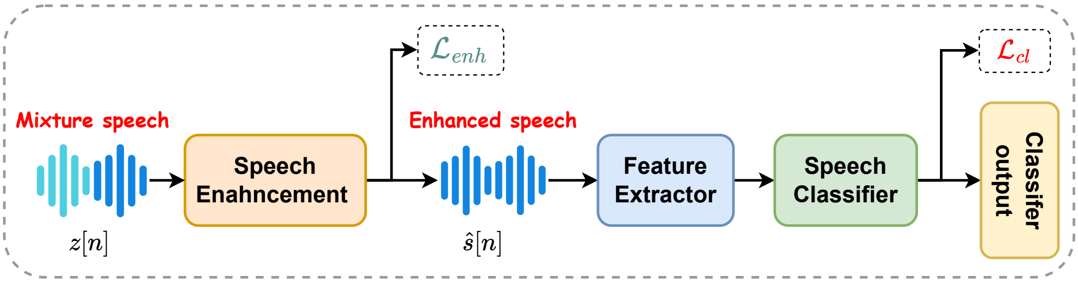


FIGURE 4.2: Block diagram of the conventional joint training approach including SE with generic back-end speech-based task.

4.2.1 Jointly training SE for Voice Activity Detection

Recently, improving the performance of voice activity detection (VAD) systems in noisy conditions gain a lot of attention, typically by employing a SE front-end as a pre-processing stage to eliminate the noise [260].

In [130, 231] the enhanced speech signals obtained from the front-end SE are used to train the VAD network in which both components are jointly optimized and fine-tuned. Further analysis shows that the poor performance of the enhancement module deteriorates the VAD performance [246]. Later the authors in [216] employ an advanced SE module to provide more enhanced features for VAD training.

Motivated by the performance of the U-Net model, the authors in [117] employed a frequency-domain SE based on U-Net to estimate both enhanced and noise spectra simultaneously, while the VAD is trained directly on the enhanced spectrum.

In [99], the authors exploit a variational auto-encoder (VAE) architecture for SE, while the VAD model is trained on VAE latent representations. Conversely, the authors in [246] train the VAD models on noisy acoustic features concatenated with the enhanced features estimated from a convolutional recurrent neural network.

A multi-objective approach is proposed in [215, 263] to jointly train SE and VAD modules to boost their performance. In particular, the same network is shared for both tasks with different loss functions. Unfortunately, this approach weakens the performance of the VAD model.

4.2.2 Jointly training SE for Keyword Spotting

Like other speech-based applications, keyword spotting performance can be negatively affected in the presence of noise. In [91], the authors addressed the problem of noise reduction for KWS by employing a microphone-array SE front-end working in the frequency-domain. The front-end is trained to optimize a KWS loss, leading to an approximately 32% improvement over their baseline. A similar contribution was conducted in [70], where the authors propose to jointly train a pre-trained SE model with a CNN-based KWS classifier.

In [21, 27], the authors addressed the "wake-up" word detection task in noisy conditions by using a linear combination of a reconstruction-based loss computed either on the log-mel filter-banks or directly on the raw waveform. Exhaustive experiments with different classifiers show that joint training improves overall performance.

The contribution in [254] addresses the KWS task in multi-speaker environments. In particular, the authors propose a joint training approach incorporating a multi-look enhancement front-end that combines spectral, inter-channel phase difference, and directions associated features with a back-end KWS.

It is worth noting that the direct enhancement speech embedding representations is still an open issue, will be investigated in detail in Chapter 5, barely investigated in literature so far. To our knowledge, the only work addressing SE with direct usage of speech embeddings is the one reported in [221]. In this work, the loss function, i.e. MSE loss is computed between the enhanced and clean speech embeddings. Recently, in [92] speech embeddings were utilized to predict the T-F masks to be applied to the noisy spectrogram. However, they are not employed in successive speech recognition or classification tasks.

4.2.3 Jointly training SE for ASR

An early contribution for jointly training SE with ASR was proposed in [52], where a feature extraction front-end module was jointly trained with an ASR based on Hidden Markov Model. Both modules were optimized with the maximum mutual information criterion.

To overcome the distortion issue introduced in conventional approaches, the authors in [239] proposed a novel joint training approach that concatenates a speech separation DNN, a filterbank feature extractor followed by an acoustic model. To strengthen their approach, linguistic information was also used. In addition, multiple features are used (e.g. log Mel-spectrogram, multi-resolution cochleagram (MRCG), etc.) to increase the acoustic model performance.

In [178], the authors observed that the front-end output distribution changes dramatically during joint training optimization, causing a negative effect on the ASR performance. Thus, they proposed a joint-training approach based on a fully batch-normalized architecture.

Motivated by the adversarial training technique, the authors in [132] proposed a joint training scheme including a mask-based enhancement module, an ASR-based encoder-decoder architecture employing the attention mechanism, and a discriminating network. The motivation behind the usage of the discriminator module is to produce features that allow better distinction between the clean and the enhanced features. A similar approach was proposed in [126] by replacing the front-end with a self-attention GAN network.

More recent research was done in [132], where a pre-trained SE module is jointly trained with a self-supervised ASR back-end. In the pre-training stage, the output waveform of the SE module is used to train a self-supervised model to learn the contextual representation using clean speech signals as the training target. Then the enhanced and noisy features are fused using a dual-attention fusion mechanism to balance the information loss.

Recently, research works in [118, 199] address a more complicated task i.e speech separation, and how can jointly train speech separation models to robust ASR systems in noisy environments using similar surveyed approaches.

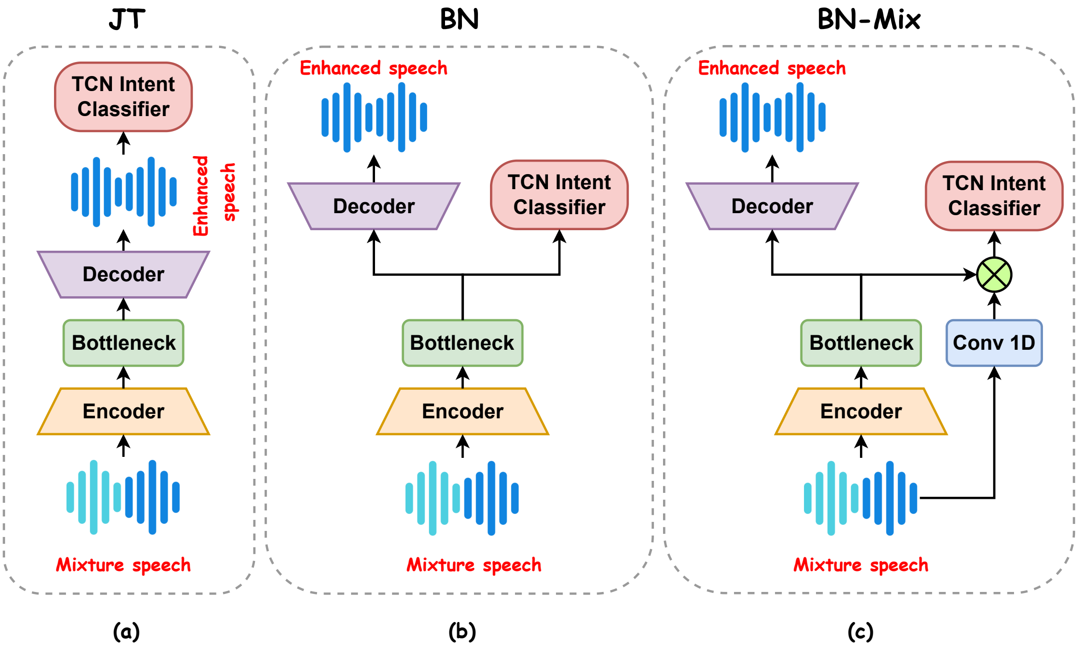


FIGURE 4.3: The proposed three joint training approaches: (**a**) based on the mixture signals (JT). (**b**) based on bottleneck representation (BN). (**c**) based on the concatenation between mixture signals and bottleneck representation (BN-Mix).

4.3 Time-Domain Joint Training Architectures

As mentioned above training a front-end SE module independently from the back-end task (IC as in our case) often introduces signal distortion that deteriorates the final performance. Hence, jointly training both two components has the capability to alleviate this issue. [99, 215].

Therefore in this Chapter, we have investigated different joint training architectures based on the Wave-U-Net for SE and the TCN for IC. The key difference between these architectures is varying the interconnection between both the SE and the intent classifier modules as shown in Fig. 4.3.

The **Joint Training** (JT) approach, shown in Fig. 4.3(a), is the most straightforward combination strategy where the intent classifier is trained on the enhanced speech signals. An alternative connection is depicted in Fig. 4.3(b) called the **Bottleneck** approach (BN), where the back-end intent classifier is trained on the SE bottleneck features. Finally, a more articulated combination called **Bottleneck-Mix** (BN-Mix), as depicted in Fig. 4.3(c), concatenates the mixture waveforms with the bottleneck representations.

All these three end-to-end joint training approaches were trained using the following total loss ($\mathcal{L}_{TOT}$):

$$\mathcal{L}_{TOT} = \alpha\mathcal{L}_{SE} + (1-\alpha)\mathcal{L}_{IC} \tag{4.1}$$

Where $\mathcal{L}_{SE}$ and $\mathcal{L}_{IC}$ are the MSE loss for the SE and the cross-entropy loss for the IC defined in Eq. 3.1 and Eq. 3.3, respectively. The weight coefficient $\alpha \in (0,1)$ is a hyper-parameter that determines the weight of each loss. In all of our experiments,

Algorithm 2 Pseudo-code for joint training

```
Require: DNNs initialization.
Require: Number of Epochs = 100.
  for $i \in$ Number of Epochs do
      Forward Pass:
      Starting from the input layer do a forward pass.
      (with batch normalization) through DNNs.
      Compute the SE loss function $\mathcal{L}_{SE}$.
      $\mathcal{L}_{SE} = \sum_n \|x[n] - \hat{x}[n]\|^2$        ▷ MSE loss based on the waveform
      Compute the intent classifier loss function $\mathcal{L}_{cl}$.
      $\mathcal{L}_{IC} = -\frac{1}{T}\sum_t \log(p_t)$        ▷ Cross-entropy loss
      Compute total loss $\mathcal{L}_{TOT}$
      $\mathcal{L}_{TOT} = \alpha\mathcal{L}_{SE} + (1-\alpha)\mathcal{L}_{IC}$        ▷ $\alpha \in (0,1)$
      Backward Pass:
      Compute the gradient of the enhancement loss $\nabla\mathcal{L}_{SE}$ and backpropagate it.
      Compute the gradient of the classifier loss $\nabla\mathcal{L}_{IC}$ and backpropagate it.
      Parameter Update:
      $\Theta_{SE} \leftarrow \Theta_{SE} - \lambda_1[\alpha\nabla\mathcal{L}_{SE} + (1-\alpha)\nabla\mathcal{L}_{IC}]$
      $\Theta_{IC} \leftarrow \Theta_{IC} - [(1-\alpha)\lambda_2\nabla\mathcal{L}_{IC}]$
  end for
  Compute the accuracy on the validation dataset.
  if $\{acc_{dev}^{i+1}\} > \{acc_{dev}^{i}\}$ then
      Save the front-end and Back-end models.
  end if
  $i+ = 1$
```

we investigate the impact of this coefficient (α) using a grid of values $\alpha \in$ (0, 0.1, 0.5, 0.9).

Although these architectures are trained using the same loss functions, their components (i.e., $\mathcal{L}_{SE}$ and $\mathcal{L}_{IC}$) affect differently the model parameters depending on the architecture interconnections and giving different performance trends for both *SE*, and *IC* tasks, as will be discussed later in Section 4.4.

The SE model parameters Θ_{SE} are updated as follows:

$$\Theta_{SE} \leftarrow \Theta_{SE} - \lambda_1[\alpha\nabla\mathcal{L}_{SE} + (1-\alpha)\nabla\mathcal{L}_{IC}] \tag{4.2}$$

Where $\nabla\mathcal{L}_{SE}$ and $\nabla\mathcal{L}_{IC}$ represent the gradients of *SE* and *IC* respectively, and λ_1 is the learning rate for the SE model.

Hence, the SE module is supposed to provide enhanced signals that match the target clean signals and maximize the intent classifier performance. Conversely, in BN and BN-Mix architectures, the $\mathcal{L}_{IC}$ does not affect the SE decoder part. Unlike the front-end, the IC model is optimized using its own loss function and its parameters are updated as:

$$\Theta_{IC} \leftarrow \Theta_{IC} - [(1-\alpha)\lambda_2\nabla\mathcal{L}_{IC}] \tag{4.3}$$

where Θ_{IC} denotes the *IC* parameters, and λ_2 is the *IC* learning rate.

4.4 Experimental Analysis

4.4.1 Dataset

For our experimental analysis, we have used the FSC dataset, described in Section 1.5. To emulate the noisy conditions in realistic scenarios where the presence of environmental noise deteriorates the classification performance, the FSC dataset is contaminated by 6 different types of noise (air conditioner, airport announcement, traffic, neighbor speaking, shutting a door, and restaurant) obtained from the MS-SNSD dataset (see Section 1.5 for more details). The clean FSC dataset is contaminated using the "maracas" library [2] by superimposing each clean signal with a noise signal using a random SNR value selected from 3 possible values: -5 dB, 0 dB, and 5 dB. Thus, the resulting noisy FSC dataset includes a uniformly distributed variety of conditions in terms of noise types and values.

4.4.2 Model Hyper-parameters

Following the same architecture of the Wave-U-Net explained in Section 3.2.1, the model is designed to process a fixed length chunk with length 16384, concatenated once enhanced. The encoder part uses 12 1-D convolutional layers with $kernelsize = 15$, $stride = 1$, and $padding = 7$, while the decoder has the same number of layers with $kernelsize = 5$, $stride = 1$, and $padding = 2$. The model is trained to minimize MSE loss ($\mathcal{L}_{SE}$) with learning rate $\lambda_1 = 10^{-4}$.

For the intent classifier, the TCN model explained in Section 3.2.3 is trained using the cross-entropy loss $\mathcal{L}_{IC}$ with learning rate $\lambda_2 = 10^{-3}$. For the BN-MIX architecture, the first layer is a 1-D convolutional layer with $kernelsize = 1$. Both models are trained with ADAM optimizer with decay rates are $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and batch size 2.

4.4.3 Experimental Results

Although our final goal is to improve the classification accuracy, we also evaluate the performance of the enhancement component in terms of PESQ, STOI, and MSE metrics illustrated in Section 2.4.

Table 4.1 reports the classification accuracy obtained from the different joint-training strategies described in Fig. 4.3, considering different values of α in Eq. 4.1. The table also reports the JT-Clean approach applied to the clean FSC dataset considering it as the upper bound, while the lower bound accuracy is reported in the "noisy" column obtained when the back-end is trained on the noisy dataset.

The JT architecture evidently brings a notable improvement in the back-end performance. The best accuracy is obtained with $\alpha = 0.5$, indicating that $\mathcal{L}_{SE}$, and $\mathcal{L}_{IC}$ are equally contribute.

Note that a value of $\alpha = 0$ corresponds to updating the SE model parameters considering only the classification loss $\mathcal{L}_{IC}$, i.e. both models are considered as a larger classifier model. In this case, we achieve relative improvement in the classifier performance with respect to the noisy case as the classifier is actually deeper.

[2] https://github.com/jfsantos/maracas

TABLE 4.1: Classification acc. for different architectures with different α.

Noisy		Jt-Clean	Jt	BN	BN-Mix
53.2%	$\alpha = 0$	73.37%	72.80%	72.39%	-
	$\alpha = 0.1$	91.53%	80.50%	77.80%	58.02%
	$\alpha = 0.5$	92.77%	86.02%	77.53%	54.99%
	$\alpha = 0.9$	-	82.52%	77.90%	66.67%

Giving more importance to the front-end module, i.e. $\alpha = 0.9$, tends to improve the output signal quality and intelligibility rather than making it suitable for the classifier module. Thus, as we expected we note in this case a performance drop.

A similar trend is also observed in the Jt-Clean approach, injecting the intermediate loss i.e. $\mathcal{L}_{SE}$ helps to improve the classifier performance. In detail, when $\alpha = 0$ we notice a small improvement with respect to the noisy case, while larger values of α improve the performance lifting the accuracy from 73.37% to 92.77%.

Conversely, the other two architectures BN, and BN-Mix show different behavior. Regarding the BN architecture, the performance is quite similar to the JT approach with $\alpha = 0$. The reasons behind that are: (a) the Wave-U-Net decoder does not interact with the classifier which leads to a negligible effect of $\mathcal{L}_{SE}$, (b) the bottleneck is a signal compact representation that does not convey enough information for the classifier. For the BN-Mix architecture, the performance limitation is back to the dimensionality gap between the combined features i.e. the bottleneck representation and the output of the 1-D convolutional layer. This leads to a limited contribution of the bottleneck representation. We also observe that in this interconnection giving more weight to the enhancement module, i.e. $\alpha = 0.9$, has a positive influence on the performance.

Tables 4.2, 4.3, and 4.4 show the enhancement performance achieved with the proposed three architectures. As discussed above in both architectures BN and BN-Mix the Wave-U-Net decoder is optimized independently based on the $\mathcal{L}_{SE}$ loss. Thus, it is not strange if we notice a substantial improvement in the intelligibility metrics with respect to the Jt architecture. However, the enhancement performance does not dependent on the value of α except $\alpha = 0$. Basically, the decoder is capable of reconstructing the signal counterbalancing the impact of the classifier on the encoder. Concerning the JT architecture, we observe a direct relation between α and intelligibility metrics. Finally, it is remarkable to observe that reconstruction quality and classification accuracy are in contrast with each other and it is not possible to effectively optimize both. Finally, for better interpretation, we report all the evaluation metrics in graphical representation as shown in Fig. 4.4.

TABLE 4.2: PESQ metric for different architectures with different α.

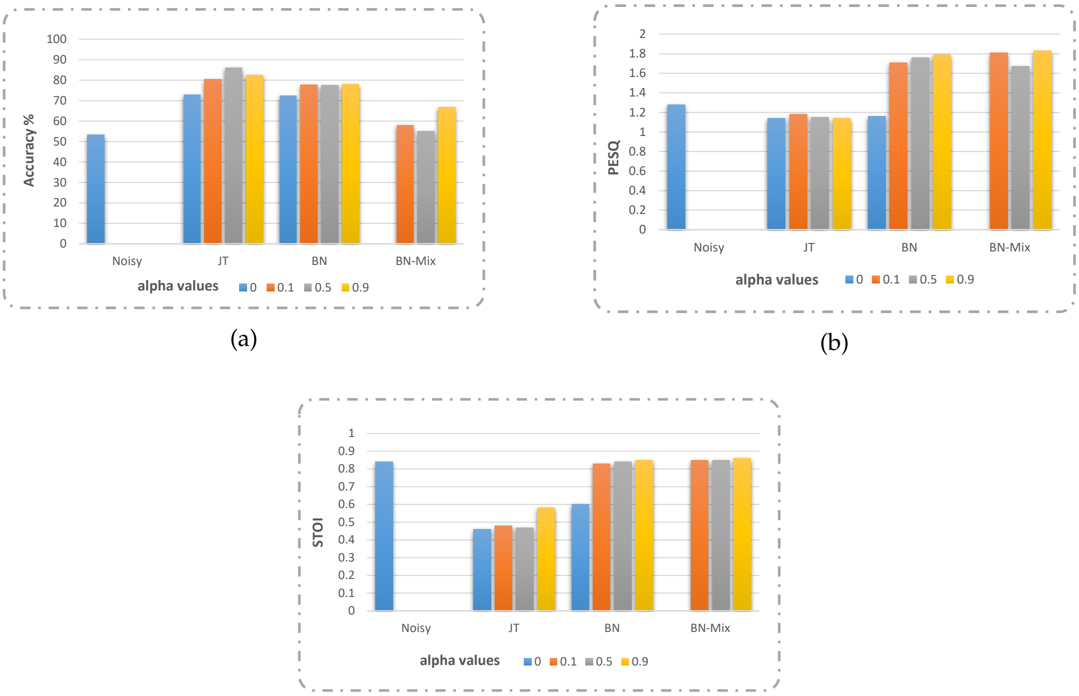
Noisy		JT	BN	BN-Mix
1.28	$\alpha = 0$	1.14	1.16	-
	$\alpha = 0.1$	1.18	1.71	1.81
	$\alpha = 0.5$	1.15	1.76	1.67
	$\alpha = 0.9$	1.14	1.79	1.83

TABLE 4.3: STOI metric for different architectures with different α.

Noisy		JT	BN	BN-Mix
0.84	$\alpha = 0$	0.46	0.60	-
	$\alpha = 0.1$	0.48	0.83	0.85
	$\alpha = 0.5$	0.47	0.84	0.85
	$\alpha = 0.9$	0.58	0.85	0.86

TABLE 4.4: MSE metric for different architectures with different α.

Noisy		JT	BN	BN-Mix
3.5×10^{-3}	$\alpha = 0$	7.6×10^{-1}	1.4×10^{-1}	-
	$\alpha = 0.1$	2.7×10^{-2}	1.7×10^{-3}	1.8×10^{-3}
	$\alpha = 0.5$	9×10^{-3}	1.7×10^{-3}	1.8×10^{-3}
	$\alpha = 0.9$	6×10^{-3}	1.7×10^{-3}	1.8×10^{-3}



(a)

(b)

(c)

FIGURE 4.4: Graphical representation for (a) the classification accuracy. (b) PESQ metric. (c) STOI metric. against α values for all experiments.

4.5 Concluding Remarks

In this chapter, we proposed three end-to-end time-domain joint training approaches namely JT, BN, and BN-Mix to robust the IC task in multi-noisy conditions. The joint training scheme integrates a neural-based SE front-end (i.e. Wave-U -Net) with a back-end intent classifier (i.e. TCN-based model). The key difference between these proposed architectures is the interconnections that combine the two components.

All experiments were conducted on a noisy version of the FSC dataset contaminated with different a set of noises obtained from the MS-SNSD dataset. Contrary to what was observed in the other speech-based classification tasks, exhaustive experiments showed the efficacy of the proposed joint training approach, in which the pre-processing enhancement stage has a positive influence on the classifier performance, especially in the case of matched noisy conditions.

Moreover, we observe that at $\alpha = 0.5$ (i.e. equally balancing both components' loss contribution) gives the best classification accuracy. In addition, we can claim that injecting an intermediate loss is always beneficial, as observed in the case of JT-clean experiments. The motivation could be the deeper model with a relatively small amount of training material the intermediate loss guides the network toward its optimal configuration. Finally, we also observed that the sequential nature of JT is better than the multi-task structure used in BN and BN-mix.

Chapter 5

Pre-trained Models for Speech Enhancement & Classification

This Chapter sheds light on directly enhancing and classification speech embeddings using large-scale pre-trained models. Section 5.1 highlights our contribution to this task. Section 5.2 discusses the utilized pre-trained models *Wav2Vec* and *WavLm*. In Section 5.3 and Section 5.4, we explain in detail our joint training approaches, giving more details on our system. In Section 5.5 and Section 5.6 our experimental results are presented and concluded, respectively. The first part of the results discussed in this Chapter was reviewed and published in the Proceedings of INTERSPEECH conference, 2022 [7] [1]. While the remaining part is accepted for publication in the Computer Speech & Language journal [2].

5.1 Introduction

Recently, the use of large-scale pre-trained models that embed speech information has become extremely popular, due to: (a) easiness and effectiveness of fine-tuning towards a specific task [41, 104, 114, 196] (b) possibility to use them as a feature extractor for successive processing [196].

In the first part of this Chapter, we extend the work in Chapter 4, by employing pre-trained speech embeddings (as *Wav2Vec* [194]) for different speech classification tasks in noisy environments, as depicted in Fig. 5.1.

Moreover, unlike the SOTA approaches, that use either frequency-domain or time-domain speech processing, we propose an approach that directly enhances speech embeddings. To do this and similarly to what shown in in Chapter 4, we employ different CNN architectures inside a joint training framework. More in detail, we compare two different joint training strategies summarized below and discussed in Section 5.3:

1. Wave-Enh: speech embeddings are extracted using *Wav2Vec* (see Section 5.2.1 for more details) on enhanced waveforms on top of the enhancement network (i.e. Wave-U-Net).

[1] https://www.isca-speech.org/archive/pdfs/interspeech_2022/ali22_interspeech.pdf

[2] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4222034

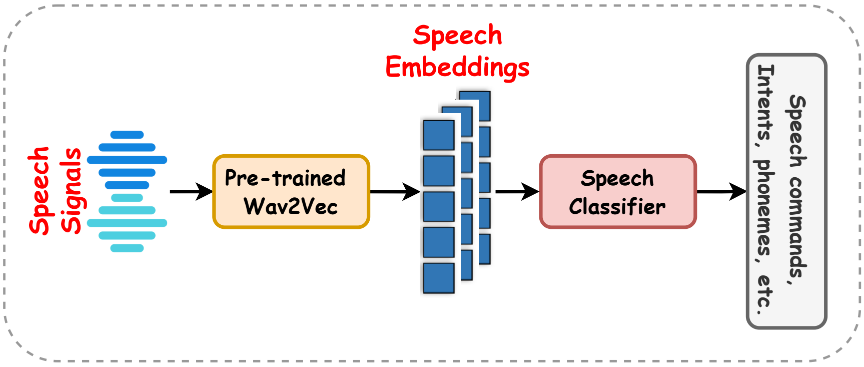


FIGURE 5.1: Speech classification based on pre-trained speech embeddings.

2. Embeds-Enh: speech embeddings are extracted from noisy waveforms and we directly enhance these noisy embeddings.

In the second part of this Chapter, we further investigate and consolidate our approach aimed to directly enhance speech embeddings, particularly:

1. We extend experiments, by employing more recent pre-trained embeddings, i.e. *WavLM* [37] (see Section 5.2.2 for more details), which are more robust to noise.
2. Besides the keyword spotting and intent classification tasks reported in the first part, we apply the proposed approach to an ASR task employing recurrent models and a **connectionist temporal classification (CTC) loss** [69].
3. We provide a more comprehensive analysis, considering different training strategies and different architectures of the embedding enhancement network.

5.2 Pre-trained Speech Models

5.2.1 Wav2Vec: Unsupervised Pre-trained Model for Speech Recognition

The *Wav2Vec* model proposed in [194] is a pre-trained unsupervised fashion to learn the speech representation from the waveform speech signals. The model, depicted in Fig. 5.2(a), consists of two networks: (a) An encoder that embeds the raw waveform speech signals ($\mathcal{X}$) to a latent space ($\mathcal{Z}$), (b) A context network that combines encoder output at multiple time-steps to estimate the contextualized speech representations ($\mathcal{C}$).

The model is trained to differentiate between k-steps future elements (z_k) and other elements ($\hat{z}$) belonging to some distribution (p_n). This can be achieved by minimizing the contrastive loss function defined as:

$$\mathcal{L}_k = -\sum_{i=1}^{T-k}(\log\sigma(z_{i+k}^T h_k(c_i)) + \lambda E_{\hat{z}\sim p_n}[\log\sigma(-\hat{z}^T h_k(c_i))]) \tag{5.1}$$

Where σ represents the sigmoid function, $\sigma(z_{i+k}^T h_k(c_i))$ is the probability that z_{i+k} is a true sample and h_k is the affine function used at step k ($h_k(c_i) = W_k c_i + b_k$).

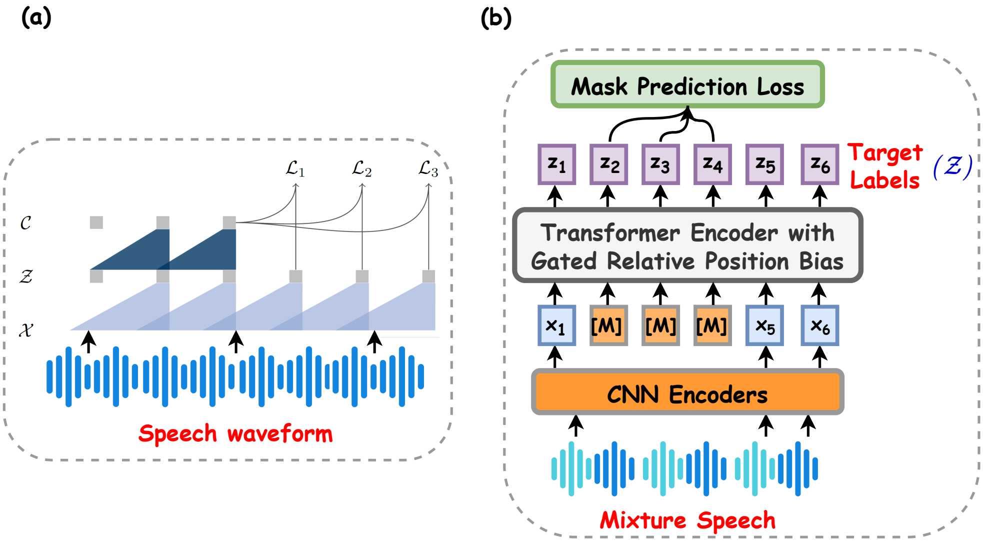


FIGURE 5.2: Block diagram of: (a) *Wav2Vec*. (b) *WavLM* models. [37, 194].

5.2.2 WavLM: Self-Supervised Pre-Trained Model for Speech Processing

The *WavLM* is a large-scale model pre-trained with a self-supervised approach [37], and similarly to *Wav2vec* it has been experimented on different speech processing tasks (e.g. speech recognition, speech enhancement, and separation) [38, 40, 203]. The model architecture, depicted in Fig 5.2(b), comprises two main networks a CNN encoder and a Transformer [226] with L blocks. During training, some output frames of the CNN encoder (x) are masked (M) and used as Transformer input. The Transformer is trained to estimate the target discrete sequence (z), where $z \in \mathcal{Z}$. The classes distribution is calculated as follows:

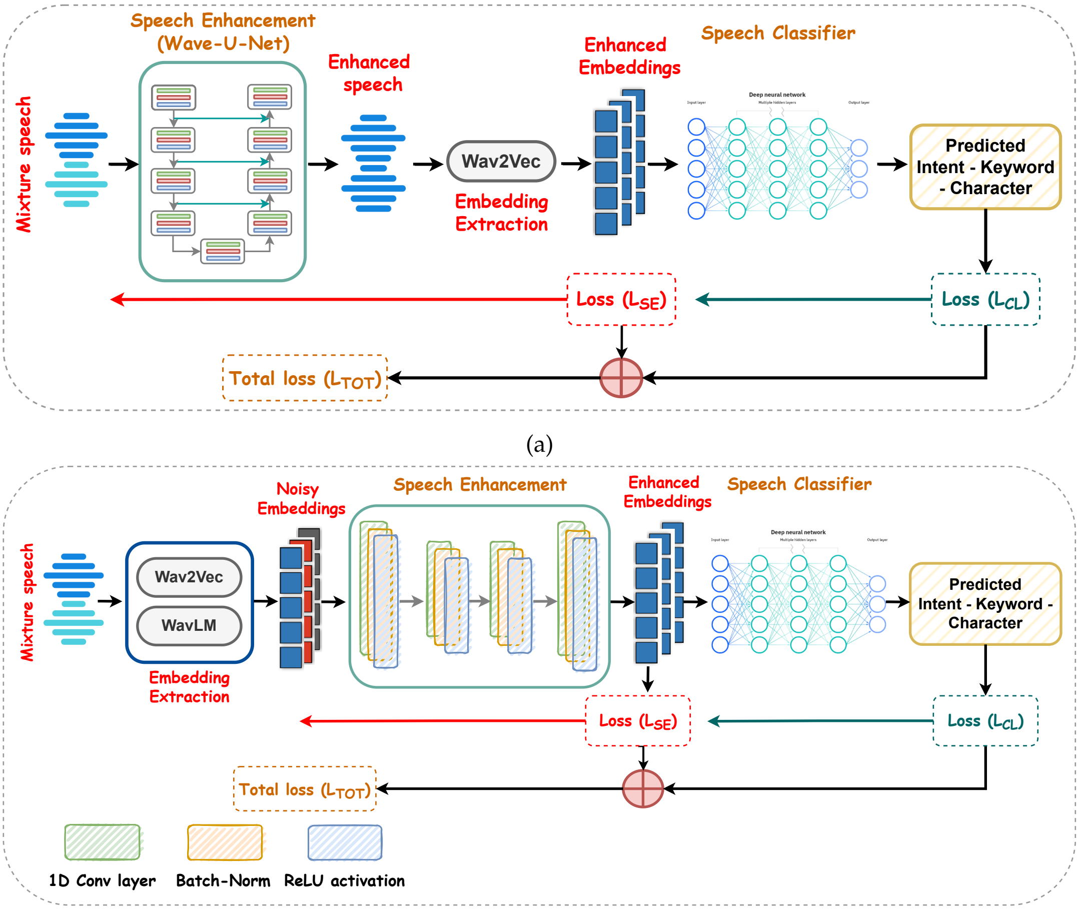
$$p(z \mid h_t) = \frac{exp(sim(W^P h_t^L, e_z/\tau)}{\sum_{z=1}^{Z} exp(sim(W^P h_t^L, e_z/\tau)} \tag{5.2}$$

Where W^P represents a projection matrix, h_t^L is the hidden state output at step t, e_z is the estimated embedding for class z, *sim* denotes the cosine similarity between two vectors, and $\tau = 0.1$ is a parameter the scales the logit. Finally, the mask prediction loss is applied on the masked frames that allows the model to learn a combination between acoustic and language models over continuous output [203].

5.3 Joint Training Schemes

As mentioned in Section 5.1, we investigate two different neural architectures, shown in Fig. 5.3, for speech classification in noisy environments. Both architectures include a stack of front-end and back-end modules for SE, and speech classification (i.e. intent classification, keywords spotting, and ASR) respectively.

The difference between the two architectures is where speech embeddings are computed. As depicted in Fig. 5.3(a) the *Wav2Vec* module is applied on top of the enhancement network, we refer to this approach as **Wave-Enh**. In the other approach,



(a)

(b)

FIGURE 5.3: The two proposed enhancement strategies: (**a**) Wave-Enh, speech waveforms are enhanced prior to embedding extraction; (**b**) Embeds-Enh, embedding are enhanced instead.

shown in Fig. 5.3(b), the *Wav2Vec* model is applied before (i.e. on bottom of) the SE module, we denote this solution as **Embeds-Enh**. Note that in this case, the enhancement stage operates directly on the embedding representation. Both architectures are trained by optimizing a joint loss function that combines the SE loss ($\mathcal{L}_{SE}$) with the classification loss in ($\mathcal{L}_{cl}$) defined in Eq. 3.1, and Eq. 3.3 respectively. Moreover, when extending our experiments to the ASR task in the **Embeds-Enh** pipeline, we adopt the CTC loss ($\mathcal{L}_{CTC}$) [68].

Similarly to what was discussed in Chapter 4, we adopt a coefficient $\alpha \in (0, 1)$ as a hyper-parameter that adjusts the weight of each component in the joint loss. In this research, we experiment with a grid of values for α, i.e. (0.1, 0.5, 0.9).

5.4 System Description

The computation of the speech embeddings as shown in Fig. 5.1 is carried out with the pre-trained *Wav2Vec* [194] [3] and *WavLM* [37] [4] models discussed in Section 5.2.1,

[3] `https://github.com/pytorch/fairseq/tree/main/examples/wav2vec`

[4] `https://github.com/microsoft/unilm/tree/master/wavlm`

Algorithm 3 Pseudo-code for joint training (Wave-Enh)

```
Require: DNNs initialization.
Require: Number of Epochs = 100.
  for $i \in$ Number of Epochs do
      Forward Pass:
      Starting from the input layer do a forward pass.
      (with batch normalization) through DNNs.
      Compute the speech enhancement loss function $\mathcal{L}_{SE}$.
      $\mathcal{L}_{SE} = \sum_n \|x[n] - \hat{x}[n]\|^2$          ▷ MSE loss based on the waveform
      Extract embeddings from enhanced signals using *Wav2Vec*.
      Compute the intent /keyword classifier loss function $\mathcal{L}_{CL}$.
      $\mathcal{L}_{CL} = \mathcal{L}_{IC} = -\frac{1}{T}\sum_t \log(p_t)$          ▷ Cross-entropy loss
      Compute total loss $\mathcal{L}_{TOT}$.
      $\mathcal{L}_{TOT} = \alpha\mathcal{L}_{SE} + (1-\alpha)\mathcal{L}_{CL}$          ▷ $\alpha \in (0,1)$
      Backward Pass:
      Compute the gradient of the enhancement loss $\nabla\mathcal{L}_{SE}$ and backpropagate it.
      Compute the gradient of the classifier loss $\nabla\mathcal{L}_{CL}$ and backpropagate it.
      Parameter Update:
      $\Theta_{SE} \leftarrow \Theta_{SE} - \lambda_1[\alpha\nabla\mathcal{L}_{SE} + (1-\alpha)\nabla\mathcal{L}_{CL}]$
      $\Theta_{CL} \leftarrow \Theta_{CL} - [(1-\alpha)\lambda_2\nabla\mathcal{L}_{CL}]$
  end for
  Compute the accuracy on the validation dataset.
  if $\{acc_{dev}^{i+1}\} > \{acc_{dev}^{i}\}$ then
      Save the front-end and Back-end models.
  end if
  $i+=1$
```

and Section 5.2.2. Recently, these models have been demonstrated effective for tackling ASR tasks, even when few supervised data are available [5] for fine-tuning.

5.4.1 Datasets

We have evaluated our proposed enhancing strategies on: *a)* an intent classification task using the FSC dataset, *b)* a keyword spotting task, using the GSC dataset v.1, and *c)* a speech recognition task using the LibriSpeech corpus (mainly investigated with the **Embeds-Enh** approach).

To emulate a realistic scenario, equivalent noisy versions of the three datasets are generated. Both FSC, and GSC datasets are contaminated with 6 types of noise (i.e. "air conditioner", "airport announcements", "traffic", "neighbor speaking", "shutting doors", and "restaurant") obtained from the MS-SNSD [180] dataset, while in the case of LibriSpeech we use 3 types of noise (i.e. noise-free-sound-0836.wav, noise-free-sound-0304.wav, and noise-free-sound-0131.wav) obtained from the MUSAN [201] dataset. More in detail, each clean signal in train, validation, and test datasets is selected and contaminated by a noise signal with an SNR randomly selected from 3 possible values: -5dB, 0dB, and 5dB using the "maracas" [6] library. For more details related to these datasets see Section 1.5.

[5] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

[6] https://github.com/jfsantos/maracas

Algorithm 4 Pseudo-code for joint training (Embeds-Enh)

```
Require: Extract embeddings from noisy, and clean waveform signals using
  Wave2Vec or WavLM.
Require: DNNs initialization.
Require: Number of Epochs = 100.
  for $i \in$ Number of Epochs do
    Forward Pass:
    Starting from the input layer do a forward pass.
    (with batch normalization) through DNNs.
    Compute the speech enhancement loss function $\mathcal{L}_{SE}$.
    $\mathcal{L}_{SE} = \sum_n \|\chi[n] - \hat{\chi}[n]\|^2$        ▷ MSE loss based on embeddings
    Compute the classifier loss function $\mathcal{L}_{CL}$.
    $\mathcal{L}_{CL} = \mathcal{L}_{IC} = -\frac{1}{T}\sum_t \log(p_t)$        ▷ Cross-entropy loss
    $\mathcal{L}_{CL} = \mathcal{L}_{CTC} = -\sum_{\hat{\mathbf{e}}_x \in \mathcal{T}} \log(p[\mathbf{s}_x \mid \hat{\mathbf{e}}_x])$        ▷ CTC loss for ASR task
    Compute total loss $\mathcal{L}_{TOT}$.
    $\mathcal{L}_{TOT} = \alpha\mathcal{L}_{SE} + (1-\alpha)\mathcal{L}_{IC}$        ▷ $\alpha \in (0,1)$
    Backward Pass:
    Compute the gradient of the enhancement loss $\nabla\mathcal{L}_{SE}$ and backpropagate it.
    Compute the gradient of the classifier loss $\nabla\mathcal{L}_{IC}$ and backpropagate it.
    Parameter Update:
    $\Theta_{SE} \leftarrow \Theta_{SE} - \lambda_1[\alpha\nabla\mathcal{L}_{SE} + (1-\alpha)\nabla\mathcal{L}_{CL}]$
    $\Theta_{CL} \leftarrow \Theta_{CL} - [(1-\alpha)\lambda_2\nabla\mathcal{L}_{CL}]$
  end for
  Compute the accuracy on the validation dataset.
  if $\{acc_{dev}^{i+1}\} > \{acc_{dev}^{i}\}$ then
    Save the front-end and Back-end models.
  end if
  $i+=1$
```

5.4.2 Enhancement

Similar to experiments presented in Chapter 3, and Chapter 4, we employ the Wave-U-Net model as the SE front-end module in the **Wave-Enh** strategy.

While mentioned in Chapter 2, many approaches had been proposed that can achieve competitive performance either in the frequency-domain or the time-domain, there are no established approaches for directly enhancing speech embeddings. Therefore for the **Embeds-Enh** strategy, we have investigated different architectures for enhancing the speech embedding in the pipeline of Fig. 5.3(b).

The **U-Net** architecture follows the model proposed in [183]. Similar to the Wave-U-Net model, U-Net is a fully CNN network with three main parts: encoding network (contracting path), bottleneck layer, and decoding network (expansive path). The encoder consists of a stack of 1D-convolutional blocks followed by ReLU activation functions and a max-pooling operation. In our implementation, we use the same configuration proposed in [183] with four downsampling blocks in the encoder network followed by the same number of blocks in the decoder network, with one bottleneck block between the encoder and the decoder networks.

Building upon the **U-Net** model, the **U-Net-2** model follows the same architecture, but without skip connections. In addition, the encoder has 4 1D-convolutional layers where the output feature channels are successively doubled from layer to layer. The

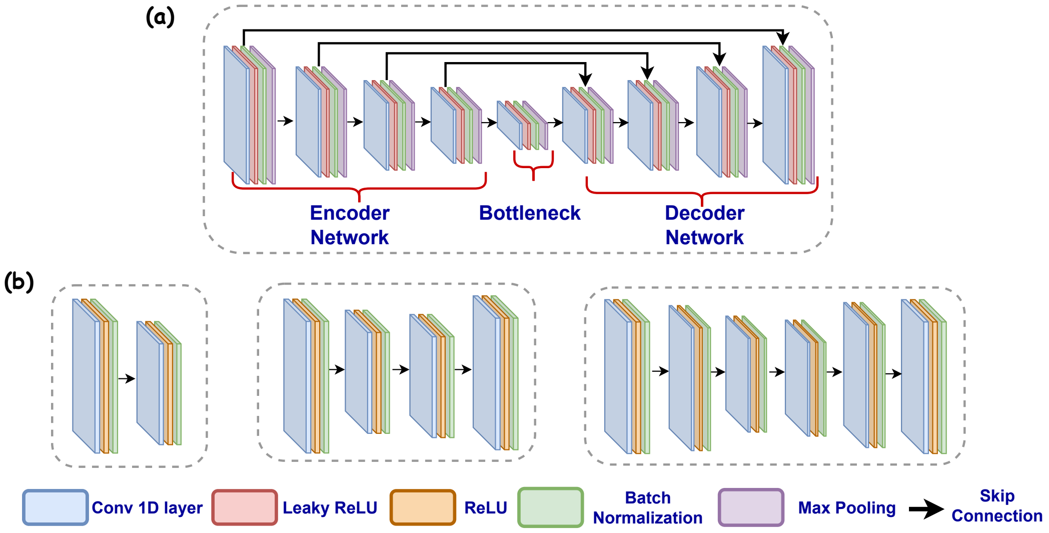


FIGURE 5.4: Graphical representation of architectures used for embedding enhancement (*Embeds-Enh*): (a) U-Net architecture. (b) CNN-K architecture, $K = 2, 4, 6$.

decoder has the same architecture as the encoder, but the output feature channels are halved sequentially.

The U-Net architecture was specifically designed to handle high dimensional and correlated feature vectors, as the samples of speech waveforms. However, in our case, speech embeddings are already low-dimensional and compact representations of the speech signal. Thus, both U-Net-based architectures may not be suitable for embedding enhancement.

Therefore, we have investigated simpler architectures for mapping embeddings. **CNN-2** consists of a stack of 2 1D-convolutional layers: the first 1D-convolutional layer has an input size $\kappa = 512$, or $\kappa = 1024$ for *Wav2Vec* or *WavLM*, respectively, while the output size is $\kappa/2$. The subsequent layer is a transposed 1D-convolutional layer with input size $\kappa/2$, and output size κ. Both layers use "kernel size" = 3, with "stride" = 1, and "padding" = 1.

In order to extend **CNN-2**, we define **CNN-4**, a model with 4 1D-convolutional layers, instead of 2. Finally, given the higher dimension of embeddings extracted with *WavLM*, we also consider **CNN-6**, which feature two further 1D-convolutional layers, at both the beginning and end of the network with input size $\kappa = 1024$ and output size $\kappa/2$.

Fig. 5.4 shows the generic architectures of the used SE front-end. All the models employ batch normalization and either leaky ReLU or ReLU activation functions after each convolutional layer. They are trained with the ADAM optimizer, with learning rate $\lambda_1 = 10^{-4}$.

5.4.3 Speech Classifier

As mentioned above, we investigate three speech classification tasks: intent classification, keyword spotting, and ASR (only the **Embeds-Enh** strategy is applied to the latter).

For both intent classification and keyword spotting, we use the TCN classifier illustrated in Section 3.2.3 to map each utterance to its corresponding 31 possible intents, or 30 keywords, respectively. The classifier processes the embeddings of size 512, or 1024, depending on whether *Wav2Vec* or *WavLM* models are employed respectively.

For the ASR scenario, we consider a character classification task. We employ a model based on Deep Speech2 [10]. It has two main parts: a series of stacked residual CNN networks (ResCNN) with GELU activation function, and layer normalization operation, followed by a set of bidirectional recurrent neural networks (BiRNN) to leverage the learned output features from the ResCNN module. Finally, a fully connected layer is inserted on top of the model having a number of units equal to the number of characters to recognize. The model is trained in order to minimize the CTC loss function [69]. Given the sequence of enhanced embeddings $\hat{\mathbf{e}}_x = \{\hat{e}_x[0], \ldots, \hat{e}_x[T-1]\}$ of a training utterance and the corresponding sequence of target characters $\mathbf{s}_x$, the CTC loss is defined over the whole training set $\mathcal{T}$ as:

$$\mathcal{L}_{CL} = \mathcal{L}_{CTC} = -\log \sum_{\hat{\mathbf{e}}_x \in \mathcal{T}} P_\pi(\mathbf{s}_x \mid \hat{\mathbf{e}}_x) \tag{5.3}$$

$$P_\pi(\mathbf{s}_x \mid \hat{\mathbf{e}}_x) \approx \Pi_{0 \leq t \leq \mathcal{T}} P_t(\pi[t] \mid \hat{\mathbf{e}}_x) \tag{5.4}$$

Where the posterior probability $P(\mathbf{s}_x \mid \hat{\mathbf{e}}_x)$ is the sum of posterior probabilities of all possible paths π that align $\mathbf{s}_x$ with $\hat{\mathbf{e}}_x$, and can be computed with the forward-backward algorithm.

The learning rate adopted for intent classification, and keyword spotting tasks is $\lambda_2 = 10^{-3}$, while for the speech recognition task it is set to $\lambda_2 = 5e^{-4}$. The optimization is carried out with the ADAM optimizer with decay rates: $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

5.5 Experimental Analysis

We evaluate our proposed enhancement approaches considering two training strategies, i.e. dis-joint and joint training. In the dis-joint training, the speech enhancement module is trained individually and then the classifier is trained on the enhanced embeddings. For joint training, both modules are trained simultaneously with three possible values of parameter α: 0.1, 0.5, and 0.9. All experiments use 100 training epochs. For the **Wave-Enh** strategy the batch size is set to 4, and 10 for FSC and GSC, respectively. While for **Embeds-Enh** the batch size is set to 10, and 20 for FSC and GSC, respectively.

TABLE 5.1: Classification accuracy using dis-joint training on FSC for different embedding enhancing models (no. of model parameters are also reported). The performance using Wave-Enh is reported as a reference.

	Wave-Enh	**Embeds-Enh**			
	Wave-U-Net	U-Net	U-Net-2	CNN-2	CNN-4
Acc	93.40%	79.88%	81.07%	92.51%	92.96%
#. of Para.	10 M	1 M	38 M	788 K	986 K

5.5.1 Results Part 1 (Wave-Enh vs. Embeds-Enh)

First of all, since an established solution for directly enhancing the speech embeddings is not yet available in the literature, we compare in Table 5.1 the performance achieved with the different architectures described in section 5.4.2 on the FSC dataset, applying dis-joint training. The best performance is obtained with the CNN-4 topology, therefore it will be used in the next experiments.

Table 5.2 gives the classification accuracies on the two classification tasks for both **Wave-Enh** and **Embeds-Enh** approaches. First, the rows "clean", and "noisy" report the upper, and lower bounds performance on clean and noisy data respectively. We point out that, our classifier is in-line with the SOTA. The other rows of the table show the performance on the noisy data considering both dis-joint and joint training.

It is worth observing the substantial improvement obtained by employing speech enhancement, especially with the joint training strategy. Interestingly, the **Embeds-Enh** approach shows a competitive performance with respect to the **Wave-Enh** counterpart not only in terms of the classification accuracy but also considering the computational complexity. In particular, the **Embeds-Enh** strategy remarkably reduces the computational complexity, being the total number of parameters of the CNN-4 model much lower ($\approx \frac{1}{10}$) than that of Wave-U-Net. The motivation behind that is the smaller dimensionality of the speech embeddings with respect to the audio waveforms.

The **Wave-Enh** approach obtains the highest classification accuracy when $\alpha = 0.5$ i.e both loss components $\mathcal{L}_{SE}$, and $\mathcal{L}_{cl}$ equally contribute to the joint loss. A similar trend is also observed in the **Embeds-Enh** strategy only on the GSC dataset. Conversely, for the FSC dataset, the optimal value is $\alpha = 0.9$ i.e. the $\mathcal{L}_{SE}$ is the predominant component in the joint loss optimization. Nevertheless, the difference between the corresponding accuracies is very small.

TABLE 5.2: Accuracy for the two speech classification tasks using different enhancement strategies. The enhancement based on embeddings uses the CNN-4 model.

Data		**Acc - FSC**	**Acc - GSC**
Clean		98.94%	96.22%
Noisy		89.63%	88.00%
Wave-Enh	Dis-joint training	93.40%	89.42%
	Joint training		
	$\alpha = 0.1$	90.95%	89.78%
	$\alpha = 0.5$	**94.88%**	**90.13%**
	$\alpha = 0.9$	93.30%	89.52%
Embeds-Enh	Dis-joint training	92.96%	89.20%
	Joint training		
	$\alpha = 0.1$	92.32%	89.08%
	$\alpha = 0.5$	93.25%	**90.28%**
	$\alpha = 0.9$	**93.30%**	89.81%

TABLE 5.3: Speech enhancement evaluation metrics on FSC dataset using Wave-Enh strategy.

		PESQ	**STOI**
Noisy		1.28	0.840
Dis-Joint training		2.26	0.890
Joint training	$\alpha = 0.1$	3.50	0.95
	$\alpha = 0.5$	2.34	0.894
	$\alpha = 0.9$	3.94	0.970

Finally, in Table 5.3, we report the enhancement performance obtained with the **Wave-Enh** pipeline, in terms of PESQ and STOI metrics defined in Section 2.4, on the FSC dataset. Noting that, the joint approach, compared with the dis-joint one not only improves the final classification accuracy but it also, improves as a by-product the SE metrics. Also in this case the optimal value is $\alpha = 0.9$, as in this case the $\mathcal{L}_{SE}$ component is the predominant loss.

5.5.2 Results Part 2 (Dive into Embeds-Enh Strategy)

Motivated by the performance of the **Embeds-Enh** strategy in this section, we give an in-depth experimental analysis considering a more recent pre-trained model i.e. *WavLM* for embedding extraction, and apply it to an ASR task. We use the character error rate as the performance metric.

We evaluate the **Embeds-Enh** strategy using the previously mentioned approaches i.e.: dis-joint training, as shown in Fig. 5.5(a), joint training, depicted in Fig 5.5(b), and warm-up denoted as the green-dash line in Fig. 5.5. In the warm-up training approach, firstly, the front-end module is trained individually and then fine-tuned jointly with the back-end. Note that this approach is different from that reported in [70] because we use the same training set for the front-end pre-training.

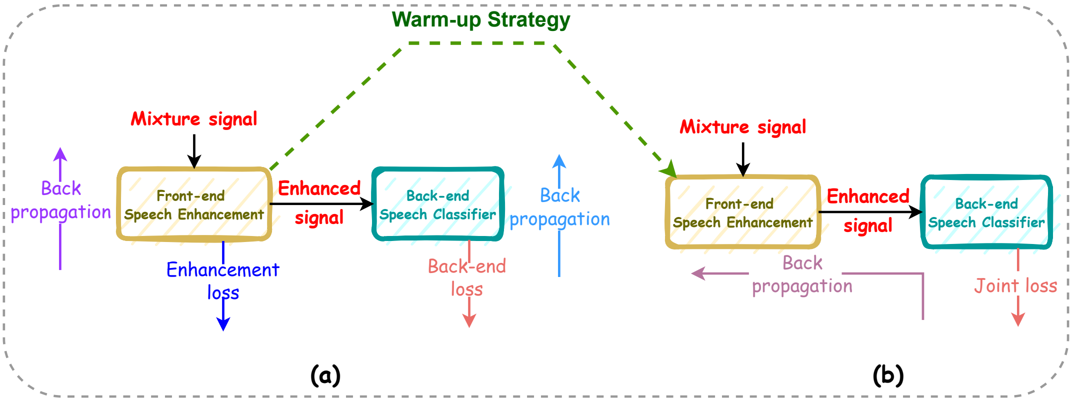


FIGURE 5.5: Graphical representation of the training strategies adopted in our experimental analysis. (a) Dis-joint training; (b) E2E joint training. The green dash line indicates that we used a front-end warmed-up with a dis-joint approach in the Joint-training approach (warm-up strategy).

TABLE 5.4: Performance achieved with different sizes of *WavLM* model on clean, noisy, enhanced FSC dataset and using CNN-4 model.

Model	**WavLM Base**	**WavLM Base+**	**WavLM Large**
# of Params.	94.70 M	94.70 M	316.62 M
# of Features	768	768	1024
Clean	99.13%	98.99%	99.44%
Noisy	58.15%	65.70%	94.77%
Dis-Joint training	74.55%	74.87%	95.70%
Joint training ($\alpha = 0.9$)	58.76%	68.78%	95.67%
Warm-up Training	68.25%	76.64%	96.22%

We introduced the warm-up approach because we observed that the front-end training was heavily penalized in the case of large back-end models. All experiments are done with a batch size equal to 10 and early stopping on the accuracy of the validation set. The maximum number of epochs is 100.

To understand how the proposed training schemes behave on different speech embedding models, besides comparing *Wav2Vec* with *WavLM*, Table 5.4 also reports the performance achieved with different *WavLM* models sizes [37] [7] on the FSC dataset. Despite using the same feature dimension as *WavLM Base*, *WavLM Base+*, shows better performance, especially in the noisy case. This is due to the fact that *WavLM Base+* is trained on a larger dataset (i.e. the same used to train *WavLM Large*). Note that, despite the different behavior on noisy data and different performance in absolute terms, we can observe the same trends in how our proposed scheme improves the performance. Note also the large improvements on noisy data achieved with *WavLM Large*, at the cost of tripling the number of parameters of the model.) Since the goal of our work is not to minimize the model size but instead to maximize the performance, we consider *WavLM Large* for our successive experiments.

Table 5.5 reports the classification accuracy considering dis-joint, joint, and warm-up training strategies on FSC. Analogously to what we reported in our previous experiments, we show the upper and lower bound accuracy in the rows "clean", and noisy (note that the performance on clean is in line with the current state-of-the-art: 99.70% in [18] and 99.30% in [28]).

First of all, we observe a substantial improvement when the enhancement front-end is applied prior to the back-end classifier on the *noisy* model in both dis-joint and joint training approaches. This gain is clearly evident when *WavLM* is employed instead of *Wav2Vec*, as the first trained using data covered by environmental noise. Nevertheless, enhancing the *WavLM* embeddings still provides a significative performance improvement, from 94.77% accuracy to $\approx 96\%$.

Exhaustive experiments in Table 5.5 show that applying SE allows significantly improving the accuracy with respect to noisy embeddings (89.63% and 94.77%) in dis-joint training, except in cases where U-Net and CNN-6 models are used. The reason could be due to the fact that these models have much more training parameters with respect to other models. Therefore, they tend to "overfit" the enhancement task, introducing critical artifacts that cannot be learned by the subsequent classifier model.

[7] https://github.com/microsoft/unilm/tree/master/wavlm

TABLE 5.5: Classification accuracy on the FSC dataset using Wav2Vec and WavLM embeddings and the enhancement networks depicted in Figure 5.4.

		Wav2Vec			**WavLM Large**		
Clean		98.94%			99.44%		
Noisy		89.63%			94.77%		
		U-Net	**CNN-2**	**CNN-4**	**CNN-2**	**CNN-4**	**CNN-6**
Dis-joint training		79.88%	92.51%	92.96%	**96.50%**	95.70%	93.48%
Joint training	$\alpha = 0.1$	92.64%	92.24%	92.32%	95.01%	95.83%	95.04%
	$\alpha = 0.5$	93.03%	93.09%	93.25%	95.30%	95.33%	95.14%
	$\alpha = 0.9$	93.00%	93.11%	93.30%	95.93%	95.67%	95.25%
Warm-up training		-	93.25%	**93.43%**	96.07%	**96.22%**	90.20%

A clear performance gap between dis-joint and joint training approaches is not observable, although joint training always allows increasing accuracy compared with noisy embeddings. Going more in detail, joint training provides a notable improvement in terms of classification accuracy when $\alpha = 0.9$, either using *Wav2Vec* or *WavLM*, meaning that the enhancement loss largely predominates over the classification loss.

We have observed that, in general, joint training first brings the front-end to convergence, and then it adapts the back-end. Using larger values of α tends to force this behavior. Conversely, smaller values of α let the training switch earlier, without bringing the front-end to full convergence.

Based on these observations, we experiment with the warm-up training approach where the front-end is firstly trained independently, and then, differently from the dis-joint approach, the model is jointly trained with the back-end. Results reported in the bottom row of Table 5.5, demonstrate that warm-up improves the final performance with respect to the conventional joint-training approaches. Finally, it is important to observe that enhancing embeddings is effective when the largest pre-trained model, i.e. *WavLM*, is employed. The fact behind this is that *WavLM* was trained over large sets of noisy data. The best accuracy is above 96.0% which is remarkable given the extremely low SNR conditions considered here.

The experimental results for the keyword spotting task, are reported in Table 5.6. First of all, let us point out that our back-end model achieves the SOTA performance on clean data (95.5 %[48] and 97.2% [185]), therefore it can be considered a solid baseline. As in the previous case, we report results using both clean, and noisy embeddings followed by **Embeds-Enh** results. We experiment only with a subset of the enhancement architectures depicted in Fig. 5.4, considering only the most performing ones, given the results obtained on FSC.

We observe a similar performance trend to the intent classification task. The *WavLM* pre-trained model improves the classification accuracy, in particular in noisy conditions, with respect to *Wav2Vec*: 92.90% versus 88.00%. For **Embeds-Enh** strategy, conclusions similar to those of the FSC task can be drawn. Also in this case, an optimal value of $\alpha = 0.9$ can be noticed. Again, the warm-up approach always improves

TABLE 5.6: Classification accuracy on GSC dataset applying different enhancement strategies to embeddings obtained using Wav2Vec and WavLM.

		Wav2Vec	WavLM Large	
Clean		96.22%	96.85%	
Noisy		88.00%	92.90%	
Dis-Joint training		**CNN-4**	**CNN-2**	**CNN-6**
		89.20%	93.75%	93.40%
Joint training	$\alpha = 0.1$	89.08%	93.37%	93.01%
	$\alpha = 0.5$	90.28%	93.12%	93.03%
	$\alpha = 0.9$	89.81%	93.77%	93.66%
Warm-up training		**90.32%**	**94.16%**	94.13%

the joint-training results, providing the best accuracy of 94.16%, which is equivalent for both architectures.

To further investigate our proposed approach and to better consolidate our conclusions, we also consider an ASR task. With respect to the tasks investigated so far, ASR involves a more complex back-end model based on a sequence-to-sequence architecture. To do this, we have used the DeepSpeech2 model [10], trained to optimize the CTC loss (see section 5.4.3), and a beam search decoder [68].

Table 5.7 reports the performance in terms of CER, obtained on the noisy version of the "test-clean" set of LibriSpeech corpus. As a matter of fact, the higher complexity of the model and the larger amount of training data make this experimental analysis much more time and computation-demanding than the previous ones. As a consequence, we limited our experiments to the sole use of the CNN-6 architecture.

Results reported in the row "clean" show that the achieved performance is in line with the SOTA, confirming also in this case the solidity of the employed back-end model. Concerning enhancement, we observe trends similar to those of Table 5.5 and Table 5.6.

However, could be, joint training fails to converge in this case. The reason as mentioned above, is the high complexity of the model and the employment of the CTC loss, which takes into account many alignment paths with the input sequence, probably moving too early the focus of the training from the front-end to the back-end.

TABLE 5.7: CER on the "test-clean" set using different embeddings and enhancement strategies based on CNN-6.

	Wav2Vec	WavLM Large
Clean	6.8%	2.5%
Noisy	19.0%	7.5%
Dis-Joint training	12.5%	4.3%
Joint training $\alpha = 0.9$	14.0%	14.8%
Warm-Up training	13.0%	**3.4%**

TABLE 5.8: CER under unseen noisy data using "test-other" of LibriSpeech using *WavLM Large*.

Model	Clean	Noisy	Dis-Joint	Warm-up
WavLM	5.1	8.4	4.8	**3.8**

Note however how the warm-up strategy, which trains the front-end alone at the beginning, substantially improves the performance of *WavLM*, reaching a very low 3.4% CER on test-clean.

Finally, to further affirm the ASR performance, we report the recognition performance on "test-other" in Table 5.8 using the pre-trained models from the previous experiments. This test set is characterized by higher noise and lower recognition performance. Note, in fact, that the CER using the clean model increases from 2.5% to 5.1%. In this way, we can explore the behavior of the proposed approach in presence of mild, unseen noise. To do this, we have used the warm-up approach. Although the model is trained on different and stronger noises, only a very minor deterioration is observed (i.e. from 3.4% to 3.8%), demonstrating that the proposed **Embeds-Enh** approach exhibits good generalization capabilities.

5.6 Concluding Remarks

In this Chapter, we proposed two joint training approaches namely **Wave-Enh**, and **Embeds-Enh** to robust intent, keywords classification, and speech recognition based on character level in noisy conditions. The difference between the two approaches is where the speech embeddings are computed. The jointly compositional scheme consists of a neural speech enhancement front-end based on the Wave-U-Net model, and a CNN model for **Wave-Enh**, and **Embeds-Enh** respectively combined with intent, keywords, and speech classifier. Note that the speech recognition task is addressed only in the **Embeds-Enh** strategy, as it is a time-consuming experiment.

All the experiments are conducted on noisy versions of FSC, GSC datasets (contaminated with noises from MS-SNSD), and the LibriSpeech dataset (contaminated with noises from MUSAN) for intent/keywords classification, and speech recognition tasks respectively. Exhaustive experiments prove the efficacy of embedding enhancement. In particular, the embedding enhancement approach shows a competitive performance not only in terms of classification accuracy or character error rate metrics but also in terms of computational complexity. The proposed CNN enhancement reduces the computational resources being the total number of parameters are reduced by ($\approx \frac{1}{10}$) with respect to the state-of-the-art models (e.g. Wave-U-Net used in **Wave-Enh** strategy. Finally, we observe that giving more weight to the front-end loss i.e. $\alpha = 0.9$ tends to improve the back-end performance as in the case of intent/keyword classification tasks.

In addition, we observe that the warm-up strategy brings notable improvements in terms of the back-end performance as it can mitigate the effects of distortions, often introduce by dis-joint training strategies, or of incomplete convergence of the front-end, that can occur in joint training as observed in our speech recognition task.

Chapter 6

Conclusion & Future Work

In this Chapter, our work is concluded in Section 6.1, followed by future research direction in Section 6.2.

6.1 Conclusion

Speech enhancement is an essential pre-processing stage for different speech-based applications such as intent classification, keyword spotting, and ASR (the tasks we addressed in this dissertation) to robust their performance in noisy conditions. In this dissertation, we have investigated different joint training strategies that combine a neural-based speech enhancement front-end utilizing the Wave-U-Net model or CNN-based architectures with a neural-based back-end speech classifier either using a Temporal convolutional network as in the case of intent classification, and keyword spotting tasks or a convolutional recurrent architecture for ASR task.

In Chapter 2, we gave an extensive overview of different speech enhancement algorithms i.e unsupervised and supervised approaches highlighting the advantages and disadvantages of both categories. Among them, time-domain deep learning approaches showed a substantial improvement concerning other algorithms. Finally, we gave more details on the commonly used speech enhancement evaluation metrics that could be used to evaluate speech quality and intelligibility.

In Chapter 3, we employed a well-known time-domain approach called Wave-U-Net and an improved version based on this model namely Dilated Encoder Wave-U-Net. The motivation behind this modified model is the dilation factor is increased exponentially successfully from layer to layer. Thus, it allows for increasing the receptive field for exploiting the contextual signal representation effectively. We evaluate the performance of both Wave-U-Net models on a back-end intent classifier based on TCN architecture. In, particular the back-end classifier is trained on the 40-Mel filter banks features extracted from the enhanced speech signals. Exhaustive experiments shed light that integrating a speech enhancement module has a positive effect on the back-end performance.

In Chapter 4, as the dis-joint training approach often introduces signal distortion at the output of the speech enhancement module, which deteriorates the back-end performance. Thus, we proposed different fully time-domain joint training strategies namely JT, BN, and BN-Mix that combine a front-end speech enhancement (Wave-U-Net) with a back-end intent classifier (TCN classifier). The key difference between these strategies is the interconnections between both models. Both models are trained to optimize a loss function that combines both front-end and back-end loss functions with a hyper-parameter α that controls the weight of each loss. Exhaustive experiments based on a noisy version of the FSC dataset showed that the speech enhancement robust the classifier performance in multi-noisy conditions, especially in the case of matched noisy conditions.

In Chapter 5, we investigated the performance of large-scale pre-trained models in the robustness of several speech classification tasks. In detail, we proposed two joint training strategies namely *Wave-Enh*, and *Embeds-Enh*. The difference between the two strategies is where speech embeddings are computed. In the *Wave-Enh* strategy, the pre-trained speech model is applied on top of the enhancement module, while in *Embeds-Enh*, the pre-trained model is applied the Wav2Vec model is on the bottom of the speech enhancement module in this case speech embeddings are directly enhanced. In the first part of our experimental analysis, we investigate the performance of *Wav2vec* model applied with both strategies. Experimental results showed that pre-trained models bring substantial improvement not only in terms of the classification accuracy addressed based on intent classification and keyword spotting tasks but also the computational complexity, especially the *Embeds-Enh* strategy. In the second part, we investigated a recent pre-trained speech representation model *WavLM* employing the model in the *Embeds-Enh* strategy. Moreover, besides the intent/keyword classification tasks, we addressed a more complicated task e.g. ASR-based on character level. Experimental results supported our claim that directly enhancing speech embeddings improve the back-end speech tasks. Additionally, the utilized model for embedding enhancement has fewer trainable parameters concerning conventional time-domain approaches (Wave-U-Net as in our case).

6.2 Future Works

In this section, we propose some future work for speech enhancement toward improving downstream tasks performance in noisy conditions.

- **Enlarge experiments scale:** We will further conduct experiments with sufficient speech data including more realistic scenarios to evaluate the effectiveness and robustness of our methods. The training data should consider as many scenarios as possible to reflect the realistic environments and improve the adaptability of the speech enhancement model. Moreover, we aim to evaluate the proposed approach on different more complex datasets for intent classification and slot-filling tasks e.g. the ATIS corpus [79], the Almawave-SLU corpus [16], the SLURP corpus [15].
- **Features fusion for speech representation learning:** Multiple feature fusion approaches can provide multiple hierarchies of data representation for model training and mapping learning. Recently, feature fusion methods are used to achieve a more robust and effective model [89, 233, 247, 252]. Thus, further exploration of multiple-feature fusion in speech enhancement will be one of our future projects.
- **Powerful neural networks for speech enhancement:** The models we investigated for embedding enhancement are stacked with 1D convolutional layers,

so the future direction is to utilize more articulated models to further improve the performance. In particular, several novel architectures were proposed and made a breakthrough in many research areas such as attention-based transformer architecture [59, 127, 159]. Those models show revolutionary performance by eliminating recurrent or convolutional portions to improve information learning.

- **Multi-channel Speech Enhancement:** In this dissertation, we mainly focus on single-channel speech enhancement algorithms. However, microphone arrays are widely used in many modern speech-processing systems, including smartphones, personal assistants, and other smart devices. With multiple microphones, spatial information can be exploited to complement spectral information for better de-noising and dereverberation. Thus, how incorporating this information into the proposed systems could be an interesting research direction.

- **Real-time speech enhancement:** The proposed speech enhancement technologies perform offline, which does not consider causal setting and latency processing. However, real-world applications require processing in real time. For example, a delay of 3 ms is noticeable in real-time applications, and delays the time of over 10 ms are unacceptable. To meet real-time applications, we need to adapt the proposed systems to causal systems and reduce the processing latency for inference, while keeping the performance at a high level. This could be a useful direction to explore.

- **Language models to robust ASR performance:** Our *Embeds-Enh* experiments for speech recognition task didn't consider a language model while training the ASR, which may degrade the performance in terms of WER. Thus, a possible direction to robust its performance is to include a language model. In detail, the classifier's output is fed into the decoder integrated with the language model. Thus, it helps to generate top words, which are then passed to language models to predict the correct sentence.

- **Directly separate speech embeddings:** Finally, we plan to extend our approach for embedding enhancement to embedding separation. In detail, we intend to adopt SOTA models for speech separation i.e. Conv-TAS-Net model [145], Dual-path-RNN [144] to directly separate speech embeddings. For this task, several benchmark datasets can be utilized i.e. LibriMix [46], WSJ2Mix investigated in this research [210].

Bibliography

[1] Khamis A Al-Karawi et al. "Automatic speaker recognition system in adverse conditions—implication of noise and reverberation on system performance". In: *International Journal of Information and Electronics Engineering* 5.6 (2015), pp. 423–427.

[2] Martín Abadi et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv:1603.04467* (2016).

[3] M Abd El-Fattah et al. "Speech enhancement using an adaptive wiener filtering approach". In: *Progress In Electromagnetics Research M* 4 (2008), pp. 167–184.

[4] Hamid Reza Abutalebi and Mehdi Rashidinejad. "Speech enhancement based on β-order mmse estimation of short time spectral amplitude and laplacian speech modeling". In: *Speech communication* 67 (2015), pp. 92–101.

[5] KA Akarsh and R Senthamizh Selvi. "Speech enhancement using non negative matrix factorization and enhanced NMF". In: *International Conference on Circuits, Power and Computing Technologies*. IEEE. 2015, pp. 1–7.

[6] Mohamed Nabih Ali, Alessio Brutti, and Daniele Falavigna. "Speech enhancement using dilated wave-u-net: an experimental analysis". In: *Conference of Open Innovations Association*. IEEE. 2020, pp. 3–9.

[7] Mohamed Nabih Ali, Daniele Falavigna, and Alessio Brutti. "Enhancing Embeddings for Speech Classification in Noisy Conditions". In: *Proc. of Interspeech* (2022), pp. 2933–2937.

[8] Mohamed Nabih Ali, Daniele Falavigna, and Alessio Brutti. "Time-Domain Joint Training Strategies of Speech Enhancement and Intent Classification Neural Models". In: *Sensors* 22.1 (2022), p. 374.

[9] Mohamed Nabih Ali et al. "A Speech Enhancement Front-End for Intent Classification in Noisy Environments". In: *European Signal Processing Conference*. IEEE. 2021, pp. 471–475.

[10] Dario Amodei et al. "Deep speech 2: End-to-end speech recognition in english and mandarin". In: *International conference on machine learning*. PMLR. 2016, pp. 173–182.

[11] Kristian Timm Andersen and Marc Moonen. "Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.1 (2017), pp. 97–107.

[12] Sung Min Ban and Hyung Soon Kim. "Weight-space viterbi decoding based spectral subtraction for reverberant speech recognition". In: *IEEE Signal Processing Letters* 22.9 (2015), pp. 1424–1428.

[13] Feng Bao and Waleed H Abdulla. "A new ratio mask representation for CASA-based speech enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.1 (2018), pp. 7–19.

[14] Feng Bao and Waleed H Abdulla. "A Novel Training Target of DNN Used for Casa-Based Speech Enhancement". In: *International Workshop on Acoustic Signal Enhancement*. IEEE. 2018, pp. 346–350.

[15] Emanuele Bastianelli et al. "SLURP: A Spoken Language Understanding Resource Package". In: *Proc. of the Conference on Empirical Methods in Natural Language Processing)*. Nov. 2020, pp. 7252–7262.

[16] Valentina Bellomaria et al. "Almawave-SLU: a new dataset for SLU in Italian". In: *arXiv:1907.07526* (2019).

[17] Jacob Benesty et al. *Speech enhancement: A signal subspace perspective*. Elsevier, 2014.

[18] Daniel Bermuth, Alexander Poeppel, and Wolfgang Reif. "Finstreder: Simple and fast Spoken Language Understanding with Finite State Transducers using modern Speech-to-Text models". In: *arXiv:2206.14589, 2022* (2022).

[19] Tom Bocklisch et al. "Rasa: Open Source Language Understanding and Dialogue Management". In: *ArXiv* abs/1712.05181 (2017).

[20] Steven Boll. "Suppression of acoustic noise in speech using spectral subtraction". In: *IEEE Transactions on acoustics, speech, and signal processing* 27.2 (1979), pp. 113–120.

[21] David Bonet et al. "Speech Enhancement for Wake-Up-Word detection in Voice Assistants". In: *Proc. IberSPEECH 2021*. 2021, pp. 41–45.

[22] Bengt J Borgström, Michael S Brandstein, and Robert B Dunn. "Improving statistical model-based speech enhancement with deep neural networks". In: *International Workshop on Acoustic Signal Enhancement*. IEEE. 2018, pp. 471–475.

[23] Adam Borowicz and Alexandr Petrovsky. "Signal subspace approach for psychoacoustically motivated speech enhancement". In: *Speech communication* 53.2 (2011), pp. 210–219.

[24] Amélie Bosca et al. "Dilated U-net based approach for multichannel speech enhancement from First-Order Ambisonics recordings". In: *European Signal Processing Conference*. IEEE. 2021, pp. 216–220.

[25] Sebastian Braun and Ivan Tashev. "Data augmentation and loss normalization for deep noise suppression". In: *International Conference on Speech and Computer*. Springer. 2020, pp. 79–86.

[26] Adelbert W Bronkhorst. "The cocktail-party problem revisited: early processing and selection of multi-talker speech". In: *Attention, Perception, & Psychophysics* 77.5 (2015), pp. 1465–1487.

[27] Guillermo Cámbara et al. "TASE: Task-Aware Speech Enhancement for Wake-Up Word Detection in Voice Assistants". In: *Applied Sciences* 12.4 (2022), p. 1974.

[28] Yiran Cao, Nihal Potdar, and Anderson R. Avila. "Sequential End-to-End Intent and Slot Label Classification and Localization". In: *Proc. Interspeech 2021*. 2021, pp. 1229–1233.

[29] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: a library for support vector machines". In: *ACM transactions on intelligent systems and technology* 2.3 (2011), pp. 1–27.

[30] Xuankai Chang et al. "End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation". In: *Proc. of Interspeech*. 2022, pp. 3819–3823.

[31] Fu-An Chao et al. "TENET: A time-reversal enhancement network for noise-robust ASR". In: *IEEE Automatic Speech Recognition and Understanding Workshop*. IEEE. 2021, pp. 55–61.

[32] Navin Chatlani and John J Soraghan. "EMD-based filtering (EMDF) of low-frequency noise for speech enhancement". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.4 (2011), pp. 1158–1166.

[33] Sarang Chehrehsa and Tom James Moir. "Speech enhancement using maximum a-posteriori and gaussian mixture models for speech and noise periodogram estimation". In: *Computer Speech & Language* 36 (2016), pp. 58–71.

[34] Chunlei Chen et al. "Deep learning on computational-resource-limited platforms: a survey". In: *Mobile Information Systems* (2020).

[35] Jingdong Chen et al. "New insights into the noise reduction Wiener filter". In: *IEEE Transactions on audio, speech, and language processing* 14.4 (2006), pp. 1218–1234.

[36] Jitong Chen, Yuxuan Wang, and DeLiang Wang. "A feature study for classification-based speech separation at low signal-to-noise ratios". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.12 (2014), pp. 1993–2002.

[37] Sanyuan Chen et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing". In: *IEEE Journal of Selected Topics in Signal Processing* (2022).

[38] Sanyuan Chen et al. "Why does Self-Supervised Learning for Speech Recognition Benefit Speaker Recognition?" In: *Proc. of Interspeech* (2022).

[39] Xi Chen et al. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets". In: *Advances in neural information processing systems* 29 (2016).

[40] Zhengyang Chen et al. "Large-scale self-supervised speech representation learning for automatic speaker verification". In: *Proc. of ICASSP*. IEEE. 2022, pp. 6147–6151.

[41] Zhuo Chen et al. "Speech separation with large-scale self-supervised learning". In: *arXiv:2211.05172* (2022).

[42] E Colin Cherry. "Some experiments on the recognition of speech, with one and with two ears". In: *The Journal of the acoustical society of America* 25.5 (1953), pp. 975–979.

[43] Hyeong-Seok Choi et al. "Phase-aware speech enhancement with deep complex u-net". In: *International Conference on Learning Representations*. 2018.

[44] Hanwook Chung, Eric Plourde, and Benoit Champagne. "Basis compensation in non-negative matrix factorization model for speech enhancement". In: *Proc. of ICASSP*. IEEE. 2016, pp. 2249–2253.

[45] Hanwook Chung, Eric Plourde, and Benoit Champagne. "Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement". In: *Speech Communication* 87 (2017), pp. 18–30.

[46] Joris Cosentino et al. "Librimix: An open-source dataset for generalizable speech separation". In: *arXiv:2005.11262* (2020).

[47] Nabanita Das et al. "Fundamentals, present and future perspectives of speech enhancement". In: *International Journal of Speech Technology* 24.4 (2021), pp. 883–901.

[48] Douglas Coimbra De Andrade et al. "A neural attention model for speech command recognition". In: *arXiv:1808.08929, 2018* (2018).

[49] Alexandre Défossez. "Hybrid spectrogram and waveform source separation". In: *arXiv:2111.03600* (2021).

[50] Huijun Ding et al. "A spectral filtering method based on hybrid wiener filters for speech enhancement". In: *Speech Communication* 51.3 (2009), pp. 259–267.

[51] Konstantin Dragomiretskiy and Dominique Zosso. "Variational mode decomposition". In: *IEEE transactions on signal processing* 62.3 (2013), pp. 531–544.

[52] Jasha Droppo and Alex Acero. “Joint discriminative front end and back end training for improved speech recognition accuracy”. In: *Proc. of ICASSP*. Vol. 1. IEEE. 2006, pp. I–I.

[53] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. “Real Time Speech Enhancement in the Waveform Domain”. In: *Proc. Interspeech*. 2020, pp. 3291–3295.

[54] Abd El-Fattah et al. “Speech enhancement with an adaptive Wiener filter”. In: *International Journal of Speech Technology* 17.1 (2014), pp. 53–64.

[55] Abd El-Moneim et al. “Hybrid speech enhancement with empirical mode decomposition and spectral subtraction for efficient speaker identification”. In: *International journal of speech technology* 18.4 (2015), pp. 555–564.

[56] Yariv Ephraim and David Malah. “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator”. In: *IEEE Transactions on acoustics, speech, and signal processing* 32.6 (1984), pp. 1109–1121.

[57] Yariv Ephraim and Harry L Van Trees. “A signal subspace approach for speech enhancement”. In: *IEEE Transactions on speech and audio processing* 3.4 (1995), pp. 251–266.

[58] Thomas Esch and Peter Vary. “Efficient musical noise suppression for speech enhancement system”. In: *Proc. of ICASSP*. IEEE. 2009, pp. 4409–4412.

[59] Junyi Fan et al. “Real-time single-channel speech enhancement based on causal attention mechanism”. In: *Applied Acoustics* 201 (2022), p. 109084.

[60] Massimiliano Fatica. “CUDA toolkit and libraries”. In: *IEEE hot chips 20 symposium*. IEEE. 2008, pp. 1–22.

[61] Mauajama Firdaus et al. “A deep multi-task model for dialogue act classification, intent detection and slot filling”. In: *Cognitive Computation* 13.3 (2021), pp. 626–645.

[62] Szu-Wei Fu, Yu Tsao, and Xugang Lu. “SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement.” In: *Proc. of Interspeech*. 2016, pp. 3768–3772.

[63] Szu-Wei Fu et al. “Raw waveform-based speech enhancement by fully convolutional networks”. In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE. 2017, pp. 006–012.

[64] Jianfeng Gao, Michel Galley, and Lihong Li. *Neural Approaches to Conversational AI: Question Answering, Task-oriented Dialogues and Social Chatbots*. Now Foundations and Trends, 2019.

[65] Timo Gerkmann and Martin Krawczyk. “MMSE-optimal spectral amplitude estimation given the STFT-phase”. In: *IEEE Signal Processing Letters* 20.2 (2012), pp. 129–132.

[66] Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy. “Attention wave-u-net for speech enhancement”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE. 2019, pp. 249–253.

[67] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[68] Alex Graves. “Connectionist temporal classification”. In: *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 61–93.

[69] Alex Graves et al. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. In: *Proc. of international conference on Machine learning*. 2006, pp. 369–376.

[70] Yue Gu et al. “A Monaural Speech Enhancement Method for Robust Small-Footprint Keyword Spotting”. In: *arXiv:1906.08415* (2019).

[71] Parisa Haghani et al. "From audio to semantics: Approaches to end-to-end spoken language understanding". In: *IEEE Spoken Language Technology Workshop*. 2018, pp. 720–726.

[72] Kun Han and DeLiang Wang. "A classification based approach to speech segregation". In: *The Journal of the Acoustical Society of America* 132.5 (2012), pp. 3475–3483.

[73] Kun Han and DeLiang Wang. "Towards generalizing classification based speech separation". In: *IEEE transactions on audio, speech, and language processing* 21.1 (2012), pp. 168–177.

[74] Xiang Hao et al. "UNetGAN: A Robust Speech Enhancement Approach in Time Domain for Extremely Low Signal-to-Noise Ratio Condition". In: *Proc. Interspeech*. 2019, pp. 1786–1790.

[75] Richard W Harris and David W Swenson. "Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment". In: *Audiology* 29.6 (1990), pp. 314–321.

[76] Mojtaba Hasannezhad. "Speech Enhancement with Improved Deep Learning Methods". PhD thesis. Concordia University, 2021.

[77] Mojtaba Hasannezhad et al. "PACDNN: A phase-aware composite deep neural network for speech enhancement". In: *Speech Communication* 136 (2022), pp. 1–13.

[78] Karen S Helfer and Laura A Wilber. "Hearing loss, aging, and speech perception in reverberation and noise". In: *Journal of Speech, Language, and Hearing Research* 33.1 (1990), pp. 149–155.

[79] Charles T. Hemphill and others. "The ATIS Spoken Language Systems Pilot Corpus". In: *Workshop on Speech and Natural Language*. 1990.

[80] Kris Hermus, Patrick Wambacq, and Hugo Van Hamme. "A review of signal subspace speech enhancement and its application to noise robust speech recognition". In: *EURASIP journal on advances in signal processing* (2006), pp. 1–15.

[81] Andrew Hines et al. "ViSQOL: an objective speech quality model". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1 (2015), pp. 1–18.

[82] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *science* 313.5786 (2006), pp. 504–507.

[83] Guoning Hu and DeLiang Wang. "A tandem algorithm for pitch estimation and voiced speech segregation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.8 (2010), pp. 2067–2079.

[84] Yanxin Hu et al. "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement". In: *Proc. Interspeech*. 2020, pp. 2472–2476.

[85] Yi Hu and Philipos C Loizou. "A generalized subspace approach for enhancing speech corrupted by colored noise". In: *IEEE transactions on speech and audio processing* 11.4 (2003), pp. 334–341.

[86] Yi Hu and Philipos C Loizou. "Evaluation of objective quality measures for speech enhancement". In: *IEEE Transactions on audio, speech, and language processing* 16.1 (2007), pp. 229–238.

[87] Yi Hu and Philipos C Loizou. "Subjective comparison of speech enhancement algorithms". In: *Proc. of ICASSP*. Vol. 1. IEEE. 2006, pp. I–I.

[88] Yonggang Hu et al. "Improved semi-supervised NMF based real-time capable speech enhancement". In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 99.1 (2016), pp. 402–406.

[89] Yuchen Hu et al. "Interactive feature fusion for end-to-end noise-robust speech recognition". In: *Proc. of ICASSP*. IEEE. 2022, pp. 6292–6296.

[90] Norden E Huang et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis". In: *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences* 454.1971 (1998), pp. 903–995.

[91] Yiteng Huang et al. "Supervised noise reduction for multichannel keyword spotting". In: *Proc. of ICASSP*. IEEE. 2018, pp. 5474–5478.

[92] Zili Huang et al. "Investigating self-supervised learning for speech enhancement and separation". In: *Proc. of ICASSP*. IEEE. 2022, pp. 6837–6841.

[93] Md Shohidul Islam et al. "Supervised single channel dual domains speech enhancement using sparse non-negative matrix factorization". In: *Digital Signal Processing* 100 (2020), p. 102697.

[94] Firas Jabloun and Benoît Champagne. "Incorporating the human hearing properties in the signal subspace approach for speech enhancement". In: *IEEE Transactions on Speech and Audio Processing* 11.6 (2003), pp. 700–708.

[95] Wissam A Jassim and Muhammad SA Zilany. "Speech quality assessment using 2D neurogram orthogonal moments". In: *Speech Communication* 80 (2016), pp. 34–48.

[96] AR Jayan. *Speech and Audio Signal Processing*. PHI Learning Pvt. Ltd., 2017.

[97] Yi Jiang et al. "Feature Enhancement Based on CASA for Robust Speech Recognition". In: *Proc. of Second International Conference on Electric Information and Control Engineering-Volume 01*. 2012, pp. 712–715.

[98] Zhaozhang Jin and DeLiang Wang. "A supervised learning approach to monaural segregation of reverberant speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.4 (2009), pp. 625–638.

[99] Youngmoon Jung et al. "Joint Learning Using Denoising Variational Autoencoders for Voice Activity Detection." In: *Proc. of Interspeech*. 2018, pp. 1210–1214.

[100] Ravi Kumar Kandagatla and PV Subbaiah. "Speech enhancement using MMSE estimation of amplitude and complex speech spectral coefficients under phase-uncertainty". In: *Speech Communication* 96 (2018), pp. 10–27.

[101] Gil Keren, Jing Han, and Björn Schuller. "Scaling speech enhancement in unseen environments with noise embeddings". In: *Proc. International Workshop on Speech Processing in Everyday Environments (CHiME 2018)*. 2018, pp. 25–29.

[102] Kais Khaldi, Abdel-Ouahab Boudraa, and Ali Komaty. "Speech enhancement using empirical mode decomposition and the Teager–Kaiser energy operator". In: *The Journal of the Acoustical Society of America* 135.1 (2014), pp. 451–459.

[103] Kais Khaldi et al. "Speech enhancement via EMD". In: *EURASIP Journal on Advances in Signal Processing* 2008 (2008), pp. 1–8.

[104] Waad Ben Kheder et al. "A unified joint model to deal with nuisance variabilities in the i-vector space". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.3 (2018), pp. 633–645.

[105] Eesung Kim and Hyeji Seo. "SE-Conformer: Time-Domain Speech Enhancement Using Conformer." In: *Proc. of Interspeech*. 2021, pp. 2736–2740.

[106] Gibak Kim and Philipos C Loizou. "Improving speech intelligibility in noise using environment-optimized algorithms". In: *IEEE transactions on audio, speech, and language processing* 18.8 (2010), pp. 2080–2090.

[107] Gibak Kim et al. "An algorithm that improves speech intelligibility in noise for normal-hearing listeners". In: *The Journal of the Acoustical Society of America* 126.3 (2009), pp. 1486–1494.

[108] Wooil Kim and Richard M Stern. "Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise". In: *Speech Communication* 53.1 (2011), pp. 1–11.

[109] Keisuke Kinoshita et al. "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network". In: *Proc. of ICASSP*. IEEE. 2020, pp. 7009–7013.

[110] Dennis Klatt. "Prediction of perceived phonetic distance from critical-band spectra: A first step". In: *Proc. of ICASSP*. Vol. 7. IEEE. 1982, pp. 1278–1281.

[111] Yuma Koizumi et al. "SNRi Target Training for Joint Speech Enhancement and Recognition". In: *Proc. Interspeech 2022*. 2022, pp. 1173–1177.

[112] Kostas Kokkinakis et al. "Evaluation of a spectral subtraction strategy to suppress reverberant energy in cochlear implant devices". In: *The Journal of the Acoustical Society of America* 138.1 (2015), pp. 115–124.

[113] Mohamed Krini and Gerhard Schmidt. "Model-based speech enhancement". In: *Speech and Audio Processing in Adverse Environments*. Springer, 2008, pp. 89–134.

[114] Yotaro Kubo, Shigeki Karita, and Michiel Bacchiani. "Knowledge Transfer from Large-scale Pretrained Language Models to End-to-end Speech Recognizers". In: *Proc. of ICASSP*. IEEE. 2022, pp. 8512–8516.

[115] Ying-Hui Lai and Wei-Zhong Zheng. "Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users". In: *Biomedical Signal Processing and Control* 48 (2019), pp. 35–45.

[116] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[117] Geon Woo Lee and Hong Kook Kim. "Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection". In: *Applied Sciences* 10.9 (2020), p. 3230.

[118] Geon Woo Lee and Hong Kook Kim. "Two-Step Joint Optimization with Auxiliary Loss Function for Noise-Robust Speech Recognition". In: *Sensors* 22.14 (2022), p. 5381.

[119] Yun-Kyung Lee and Oh-Wook Kwon. "Application of shape analysis techniques for improved CASA-based speech separation". In: *IEEE Transactions on Consumer Electronics* 55.1 (2009), pp. 146–149.

[120] Simon Leglaive, Laurent Girin, and Radu Horaud. "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization". In: *Proc. of ICASSP*. IEEE. 2019, pp. 101–105.

[121] Simon Leglaive et al. "A recurrent variational autoencoder for speech enhancement". In: *Proc. of ICASSP*. IEEE. 2020, pp. 371–375.

[122] Adrien Leman, Julien Faure, and Etienne Parizet. "Influence of informational content of background noise on speech quality evaluation for VoIP application". In: *Journal of the Acoustical Society of America* 123.5 (2008), p. 3066.

[123] Chia-Yu Li and Ngoc Thang Vu. "Improving Speech Recognition on Noisy Speech via Speech Enhancement with Multi-Discriminators CycleGAN". In: *Automatic Speech Recognition and Understanding Workshop*. IEEE. 2021, pp. 830–836.

[124] Feipeng Li et al. “A long, deep and wide artificial neural net for robust speech recognition in unknown noise”. In: *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.

[125] Jinyu Li et al. “An overview of noise-robust automatic speech recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (2014), pp. 745–777.

[126] Lujun Li et al. “Adversarial joint training with self-attention mechanism for robust end-to-end speech recognition”. In: *EURASIP Journal on Audio, Speech, and Music Processing* (2021), pp. 1–16.

[127] Yi Li, Yang Sun, and Syed Mohsen Naqvi. “U-shaped transformer with frequency-band aware attention for speech enhancement”. In: *arXiv:2112.06052* (2021).

[128] Ju Lin et al. “A time-domain convolutional recurrent network for packet loss concealment”. In: *Proc. of ICASSP*. IEEE. 2021, pp. 7148–7152.

[129] Ju Lin et al. “Improved speech enhancement using a time-domain GAN with mask learning”. In: *Proc. of Interspeech* (2020).

[130] Ruixi Lin et al. “Optimizing Voice Activity Detection for Noisy Conditions.” In: *Proc. of Interspeech*. 2019, pp. 2030–2034.

[131] Pierre Lison and Casey Kennington. “OpenDial: A Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules”. In: *Proc. of ACL System Demonstrations*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 67–72. DOI: 10.18653/v1/P16-4012. URL: https://www.aclweb.org/anthology/P16-4012.

[132] Bin Liu et al. “Jointly Adversarial Enhancement Training for Robust End-to-End Speech Recognition.” In: *Proc. of Interspeech*. 2019, pp. 491–495.

[133] Ming Liu et al. “Speech enhancement method based on LSTM neural network for speech recognition”. In: *IEEE International Conference on Signal Processing*. IEEE. 2018, pp. 245–249.

[134] Yang Liu and Jiajun Zhang. “Deep learning in machine translation”. In: *Deep Learning in Natural Language Processing*. Springer, 2018, pp. 147–183.

[135] Yuanyuan Liu et al. “Variational mode decomposition denoising combined the detrended fluctuation analysis”. In: *Signal Processing* 125 (2016), pp. 349–364.

[136] Philip Lockwood and Jérôme Boudy. “Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars”. In: *Speech communication* 11.2-3 (1992), pp. 215–228.

[137] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.

[138] Philipos C Loizou. “Speech quality assessment”. In: *Multimedia analysis, processing and communications*. Springer, 2011, pp. 623–654.

[139] Ching-Ta Lu. “Enhancement of single channel speech using perceptual-decision-directed approach”. In: *Speech communication* 53.4 (2011), pp. 495–507.

[140] Xugang Lu et al. “Speech enhancement based on deep denoising autoencoder.” In: *Proc. of Interspeech*. Vol. 2013. 2013, pp. 436–440.

[141] Yang Lu and Philipos C Loizou. “A geometric approach to spectral subtraction”. In: *Speech communication* 50.6 (2008), pp. 453–466.

[142] Jimmy Ludeña-Choez and Ascensión Gallardo-Antolín. “Speech denoising using non-negative matrix factorization with kullback-leibler divergence and sparseness constraints”. In: *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2012, pp. 207–216.

[143] Loren Lugosch et al. “Speech model pre-training for end-to-end spoken language understanding”. In: *Proc. of Interspeech*. 2019, pp. 814–818.

[144] Yi Luo, Zhuo Chen, and Takuya Yoshioka. "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation". In: *Proc. of ICASSP*. IEEE. 2020, pp. 46–50.

[145] Yi Luo and Nima Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8 (2019), pp. 1256–1266.

[146] Craig Macartney and Tillman Weyde. "Improved speech enhancement with the wave-u-net". In: *arXiv:1811.11307* (2018).

[147] Wageesha Manamperi et al. "GMM based multi-stage Wiener filtering for low SNR speech enhancement". In: *International Workshop on Acoustic Signal Enhancement*. IEEE. 2022, pp. 1–5.

[148] Tobias May and Torsten Dau. "Computational speech segregation based on an auditory-inspired modulation analysis". In: *The Journal of the Acoustical Society of America* 136.6 (2014), pp. 3350–3359.

[149] Tobias May and Torsten Dau. "Requirements for the evaluation of computational speech segregation systems". In: *The Journal of the Acoustical Society of America* 136.6 (2014), EL398–EL404.

[150] Udar Mittal and Nam Phamdo. "Signal/noise KLT based approach for enhancing speech degraded by colored noise". In: *IEEE Transactions on speech and audio processing* 8.2 (2000), pp. 159–167.

[151] Nasser Mohammadiha. "Speech Enhancement Using Nonnegative Matrix-Factorization and Hidden Markov Models". PhD thesis. KTH Royal Institute of Technology, 2013.

[152] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon. "Supervised and unsupervised speech enhancement using nonnegative matrix factorization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (2013), pp. 2140–2151.

[153] Giovanni Morrone. "Deep Learning Methods for Audio-Visual Speech Processing in Noisy Environments". PhD thesis. University of Modena and Reggio Emilia Modena, Italy, 2021.

[154] Arun Narayanan and DeLiang Wang. "A CASA-based system for long-term SNR estimation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.9 (2012), pp. 2518–2527.

[155] Arun Narayanan and DeLiang Wang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition". In: *Proc. of ICASSP*. IEEE. 2013, pp. 7092–7096.

[156] Arun Narayanan and DeLiang Wang. "Investigation of speech separation as a front-end for noise robust speech recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.4 (2014), pp. 826–835.

[157] S Nasir et al. "Speech enhancement with geometric advent of spectral subtraction using connected time-frequency regions noise estimation". In: *Research Journal of Applied Sciences, Engineering and Technology* 6.6 (2013), pp. 1081–1087.

[158] Zhaoxu Nian et al. "A Time Domain Progressive Learning Approach with SNR Constriction for Single-Channel Speech Enhancement and Recognition". In: *Proc. of ICASSP*. IEEE. 2022, pp. 6277–6281.

[159] Koen Oostermeijer, Qing Wang, and Jun Du. "Lightweight Causal Transformer with Local Self-Attention for Real-Time Speech Enhancement." In: *Proc. of Interspeech*. 2021, pp. 2831–2835.

[160] Zhiheng Ouyang. "Single-Channel Speech Enhancement Based on Deep Neural Networks". PhD thesis. Concordia University, 2019.

[161] Sankar K Pal and Sushmita Mitra. "Multilayer perceptron, fuzzy sets, classifiaction". In: *IEEE Transactions on Neural Networks* (1992), pp. 68–697.

[162] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain". In: *Speech communication* 52.5 (2010), pp. 450–475.

[163] Vassil Panayotov et al. "Librispeech: an asr corpus based on public domain audio books". In: *Proc. of ICASSP*. IEEE. 2015, pp. 5206–5210.

[164] Ashutosh Pandey and DeLiang Wang. "A new framework for CNN-based speech enhancement in the time domain". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.7 (2019), pp. 1179–1188.

[165] Ashutosh Pandey and DeLiang Wang. "Dense CNN with self-attention for time-domain speech enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1270–1279.

[166] Ashutosh Pandey and DeLiang Wang. "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain". In: *Proc. of ICASSP*. IEEE. 2019, pp. 6875–6879.

[167] Ashutosh Pandey et al. "Dual application of speech enhancement for automatic speech recognition". In: *IEEE Spoken Language Technology Workshop*. IEEE. 2021, pp. 223–228.

[168] Wanida Panup, Wachirapong Ratipapongton, and Rabian Wangkeeree. "A novel twin support vector machine with generalized pinball loss function for pattern classification". In: *Symmetry* 14.2 (2022), p. 289.

[169] Gyuseok Park et al. "Speech enhancement for hearing aids with deep learning on environmental noises". In: *Applied Sciences* 10.17 (2020), p. 6077.

[170] Se Rim Park and Jin Won Lee. "A Fully Convolutional Neural Network for Speech Enhancement". In: *Proc. Interspeech*. 2017, pp. 1993–1997.

[171] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. "SEGAN: Speech Enhancement Generative Adversarial Network". In: *Proc. Interspeech*. 2017, pp. 3642–3646.

[172] Santiago Pascual, Joan Serra, and Antonio Bonafonte. "Time-domain speech enhancement using generative adversarial networks". In: *Speech Communication* 114 (2019), pp. 10–21.

[173] Adam Paszke et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32 (2019).

[174] Tal Peer and Timo Gerkmann. "Phase-aware deep speech enhancement: It's all about the frame length". In: *JASA Express Letters* 2.10 (2022), p. 104802.

[175] Yao Qian et al. "Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system". In: *IEEE Automatic Speech Recognition and Understanding Workshop*. 2017, pp. 569–576.

[176] Schuyler R Quackenbush, Thomas P Barnwell, and Mark A Clements. *Objective measures of speech quality*. Prentice-Hall, 1988.

[177] Martin Radfar, Athanasios Mouchtaris, and Siegfried Kunzmann. "End-to-End Neural Transformer Based Spoken Language Understanding". In: *Proc. of Interspeech*. 2020, pp. 866–870.

[178] Mirco Ravanelli et al. "Batch-normalized joint training for DNN-based distant speech recognition". In: *IEEE Spoken Language Technology Workshop*. IEEE. 2016, pp. 28–34.

[179] ITU-T Recommendation. "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs". In: *Rec. ITU-T P. 862* (2001).

[180] Chandan KA Reddy et al. "A Scalable Noisy Speech Dataset and Online Subjective Test Framework". In: *Proc. of Interspeech* (2019), pp. 1816–1820.

[181] Afshin Rezayee and Saeed Gazor. "An adaptive KLT approach for speech enhancement". In: *IEEE Transactions on Speech and Audio Processing* 9.2 (2001), pp. 87–95.

[182] Dayana Ribas et al. "Wiener Filter and Deep Neural Networks: A Well-Balanced Pair for Speech Enhancement". In: *Applied Sciences* 12.18 (2022), p. 9000.

[183] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[184] Tomasz Rutkowski, Andrzej Cichocki, and Allan Kardec Barros. "Speech enhancement from interfering sounds using CASA techniques and blind source separation". In: *ICA'01* (2001), pp. 728–733.

[185] Oleg Rybakov et al. "Streaming Keyword Spotting on Mobile Devices". In: *Proc. Interspeech*. 2020, pp. 2277–2281.

[186] Navin Sabharwal and Amit Agrawal. "Introduction to Google dialogflow". In: *Cognitive virtual assistants using Google Dialogflow*. Springer, 2020, pp. 13–54.

[187] R Saeidi, P Mowlaee, and R Martin. "Phase estimation for signal reconstruction in single-channel source separation". In: *Proc. of Interspeech* (2012).

[188] N Saleem, Muhammad Irfan Khattak, and EV Perez. "Spectral phase estimation based on deep neural networks for single channel speech enhancement". In: *Journal of Communications Technology and Electronics* 64.12 (2019), pp. 1372–1382.

[189] Nasir Saleem and Muhammad Irfan Khattak. "A review of supervised learning algorithms for single channel speech enhancement". In: *International Journal of Speech Technology* 22.4 (2019), pp. 1051–1075.

[190] Nasir Saleem and Muhammad Irfan Khattak. "Deep neural networks for speech enhancement in complex-noisy environments". In: *International journal of interactive multimedia and artificial intelligence* (2020).

[191] Nasir Saleem, Muhammad Irfan Khattak, and Elena Verdú. "On improvement of speech intelligibility and quality: A survey of unsupervised single channel speech enhancement algorithms". In: *International Journal of Interactive Multimedia and Artificial Intelligence* (2020).

[192] Nasir Saleem et al. "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks". In: *IEEE Access* 8 (2020), pp. 160581–160595.

[193] Ryosuke Sawata, Yosuke Kashiwagi, and Shusuke Takahashi. "Improving Character Error Rate is Not Equal to Having Clean Speech: Speech Enhancement for ASR Systems with Black-Box Acoustic Models". In: *Proc. of ICASSP*. IEEE. 2022, pp. 991–995.

[194] Steffen Schneider et al. "wav2vec: Unsupervised Pre-Training for Speech Recognition". In: *Proc. Interspeech*. 2019, pp. 3465–3469.

[195] Michael L Seltzer et al. "An investigation of deep neural networks for noise robust speech recognition". In: *Proc. of ICASSP*. IEEE. 2013, pp. 7398–7402.

[196] Deokjin Seo, Heung-Seon Oh, and Yuchul Jung. "Wav2kws: Transfer learning from speech representations for keyword spotting". In: *IEEE Access* 9 (2021), pp. 80682–80691.

[197] Dmitriy Serdyuk et al. "Towards end-to-end spoken language understanding". In: *Proc. of ICASSP*. 2018, pp. 5754–5758.

[198] Yih-Liang Shen et al. "Reinforcement learning based speech enhancement for robust speech recognition". In: *Proc. of ICASSP*. IEEE. 2019, pp. 6750–6754.

[199] Jing Shi et al. "Train from scratch: Single-stage joint training of speech separation and recognition". In: *Computer Speech & Language* 76 (2022), p. 101387.

[200] Barbara G Shinn-Cunningham and Virginia Best. "Selective attention in normal and impaired hearing". In: *Trends in amplification* 12.4 (2008), pp. 283–299.

[201] David Snyder, Guoguo Chen, and Daniel Povey. "Musan: A music, speech, and noise corpus". In: *arXiv:1510.08484* (2015).

[202] Jordi Solé-Casals et al. "Speech Enhancement: a multivariate empirical mode decomposition approach". In: *International Conference on Nonlinear Speech Processing*. Springer. 2013, pp. 192–199.

[203] Hyungchan Song et al. "Exploring WavLM on Speech Enhancement". In: *IEEE Spoken Language Technology Workshop*. 2023, pp. 451–457.

[204] William Song and Jim Cai. "End-to-end deep neural network for automatic speech recognition". In: *Standford CS224D Reports* (2015).

[205] Yann Soon and Soo Ngee Koh. "Speech enhancement using 2-D Fourier transform". In: *IEEE Transactions on speech and audio processing* 11.6 (2003), pp. 717–724.

[206] Soundararajan Srinivasan, Nicoleta Roman, and DeLiang Wang. "Binary and ratio time-frequency masks for robust speech recognition". In: *Speech Communication* 48.11 (2006), pp. 1486–1501.

[207] Daniel Stoller, Sebastian Ewert, and Simon Dixon. "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation". In: *International Society for Music Information Retrieval Conference*. 2018.

[208] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features". In: *Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE. 2021, pp. 166–170.

[209] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks". In: *Proc. Interspeech*. 2020, pp. 4506–4510.

[210] Cem Subakan et al. "Attention is all you need in speech separation". In: *Proc. of ICASSP*. IEEE. 2021, pp. 21–25.

[211] Akihiko Sugiyama and Ryoji Miyahara. "Phase randomization-a new paradigm for single-channel signal enhancement". In: *Proc. of ICASSP*. IEEE. 2013, pp. 7487–7491.

[212] Sudeep Surendran and T Kishore Kumar. "Perceptual subspace speech enhancement with variance normalization". In: *Procedia Computer Science* 54 (2015), pp. 818–828.

[213] Cees H Taal et al. "A short-time objective intelligibility measure for time-frequency weighted noisy speech". In: *Proc. of ICASSP*. IEEE. 2010, pp. 4214–4217.

[214] Ke Tan and DeLiang Wang. "A convolutional recurrent neural network for real-time speech enhancement." In: *Proc. of Interspeech*. Vol. 2018. 2018, pp. 3229–3233.

[215] Xu Tan and Xiao-Lei Zhang. "Speech enhancement aided end-to-end multi-task learning for voice activity detection". In: *Proc. of ICASSP*. IEEE. 2021, pp. 6823–6827.

[216] Zheng-Hua Tan et al. "rVAD: An unsupervised segment-based robust voice activity detection method". In: *Computer speech & language* 59 (2020), pp. 1–21.

[217] Jiexiong Tang, Chenwei Deng, and Guang-Bin Huang. "Extreme learning machine for multilayer perceptron". In: *IEEE transactions on neural networks and learning systems* 27.4 (2015), pp. 809–821.

[218] Jürgen Tchorz and Birger Kollmeier. "SNR estimation based on amplitude modulation analysis with applications to noise suppression". In: *IEEE Transactions on Speech and Audio Processing* 11.3 (2003), pp. 184–192.

[219] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments". In: *Proc. Meetings Acoust.* 2013, pp. 1–6.

[220] Yusheng Tian and Philip John Gorinski. "Improving End-to-End Speech-to-Intent Classification with Reptile". In: *Proc. of Interspeech*. 2020, pp. 891–895.

[221] Viet Anh Trinh and Sebastian Braun. "Unsupervised Speech Enhancement with Speech Recognition Embedding and Disentanglement Losses". In: *Proc. of ICASSP*. 2022, pp. 391–395.

[222] Gokhan Tur and Renato De Mori. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons, 2011.

[223] Abhay Upadhyay and RB Pachori. "Speech enhancement based on mEMD-VMD method". In: *Electronics Letters* 53.7 (2017), pp. 502–504.

[224] Navneet Upadhyay and Abhijit Karmakar. "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study". In: *Procedia Computer Science* 54 (2015), pp. 574–584.

[225] Peter Vary and Rainer Martin. *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

[226] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[227] Siddala Vihari et al. "Comparison of speech enhancement algorithms". In: *Procedia computer science* 89 (2016), pp. 666–676.

[228] Emmanuel Vincent et al. "The second 'CHiME'speech separation and recognition challenge: An overview of challenge systems and outcomes". In: *Workshop on Automatic Speech Recognition and Understanding*. IEEE. 2013, pp. 162–167.

[229] Athanasios Voulodimos et al. "Deep learning for computer vision: A brief review". In: *Computational intelligence and neuroscience* 2018 (2018).

[230] DeLiang Wang and Guy J Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[231] Qing Wang et al. "A universal VAD based on jointly trained deep neural networks". In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[232] Tianrui Wang et al. "Multiple Confidence Gates For Joint Training Of SE And ASR". In: *arXiv:2204.00226* (2022).

[233] Youming Wang et al. "Speech enhancement from fused features based on deep neural network and gated recurrent unit network". In: *EURASIP Journal on Advances in Signal Processing* 2021.1 (2021), pp. 1–19.

[234] Yuxuan Wang, Kun Han, and DeLiang Wang. "Exploring monaural features for classification-based speech segregation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.2 (2012), pp. 270–279.

[235] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. "On training targets for supervised speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 22.12 (2014), pp. 1849–1858.

[236] Yuxuan Wang and DeLiang Wang. "Boosting classification based speech separation using temporal dynamics". In: *Annual Conference of the International Speech Communication Association*. 2012.

[237] Yuxuan Wang and DeLiang Wang. "Cocktail party processing via structured prediction". In: *Advances in Neural Information Processing Systems* 25 (2012).

[238] Yuxuan Wang and DeLiang Wang. "Towards scaling up classification-based speech separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.7 (2013), pp. 1381–1390.

[239] Zhong-Qiu Wang and DeLiang Wang. "Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition". In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[240] Felix Weninger et al. "Discriminative NMF and its application to single-channel source separation." In: *Proc. of Interspeech*. 2014, pp. 865–869.

[241] Felix Weninger et al. "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR". In: *International conference on latent variable analysis and signal separation*. Springer. 2015, pp. 91–99.

[242] Jürgen Wiest et al. "Probabilistic trajectory prediction with Gaussian mixture models". In: *IEEE Intelligent Vehicles Symposium*. IEEE. 2012, pp. 141–146.

[243] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. "Complex ratio masking for monaural speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 24.3 (2015), pp. 483–492.

[244] Yang Xiang et al. "An nmf-hmm speech enhancement method based on kullback-leibler divergence". In: *Proc. of Interspeech*. 2020, pp. 2667–2671.

[245] Longting Xu et al. "Speech enhancement based on nonnegative matrix factorization in constant-Q frequency domain". In: *Applied Acoustics* 174 (2021), p. 107732.

[246] Tianjiao Xu et al. "Joint training ResCNN-based voice activity detection with speech enhancement". In: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE. 2019, pp. 1157–1162.

[247] Xinmeng Xu and Jianjun Hao. "Multi-layer Feature Fusion Convolution Network for Audio-visual Speech Enhancement". In: *arXiv:2101.05975* (2021).

[248] Yong Xu et al. "A regression approach to speech enhancement based on deep neural networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.1 (2014), pp. 7–19.

[249] Yong Xu et al. "An experimental study on speech enhancement based on deep neural networks". In: *IEEE Signal processing letters* 21.1 (2013), pp. 65–68.

[250] Fan Yang et al. "Improving generative adversarial networks for speech enhancement through regularization of latent representations". In: *Speech Communication* 118 (2020), pp. 1–9.

[251] Omolbanin Yazdanbakhsh and Scott Dick. "Multivariate time series classification using dilated convolutional neural network". In: *arXiv:1905.01697* (2019).

[252] Moujia Ye and Hongjie Wan. "Improved Transformer-Based Dual-Path Network with Amplitude and Complex Domain Feature Fusion for Speech Enhancement". In: *Entropy* 25.2 (2023), p. 228.

[253] Shi Yin et al. "Noisy training for deep neural networks in speech recognition". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1 (2015), pp. 1–14.

[254] Meng Yu et al. "End-to-End Multi-Look Keyword Spotting". In: *Proc. Interspeech*. 2020, pp. 66–70.

[255] Leonardo Zão, Rosângela Coelho, and Patrick Flandrin. "Speech enhancement with EMD and hurst-based mode selection". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.5 (2014), pp. 899–911.

[256] Hui Zhang, Xueliang Zhang, and Guanglai Gao. "Multi-Target Ensemble Learning for Monaural Speech Separation." In: *Proc. of Interspeech*. 2017, pp. 1958–1962.

[257] Qian Zhang et al. "Sensing to hear: Speech enhancement for mobile devices using acoustic signals". In: *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.3 (2021), pp. 1–30.

[258] Qiquan Zhang et al. "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 1404–1415.

[259] Wangyou Zhang et al. "Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE. 2021, pp. 146–150.

[260] Xiao-Lei Zhang and Ji Wu. "Denoising deep neural networks based voice activity detection". In: *Proc. of ICASSP*. IEEE. 2013, pp. 853–857.

[261] Yi Zhang and Yunxin Zhao. "Real and imaginary modulation spectral subtraction for speech enhancement". In: *Speech Communication* 55.4 (2013), pp. 509–522.

[262] Naijun Zheng and Xiao-Lei Zhang. "Phase-aware speech enhancement based on deep neural networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.1 (2018), pp. 63–76.

[263] Yimeng Zhuang et al. "Multi-task joint-learning for robust voice activity detection". In: *International Symposium on Chinese Spoken Language Processing*. IEEE. 2016, pp. 1–5.