University of Crete
Department of Computer Science

Neural Networks for the Quality and Intelligibility Enhancement of Speech

Ph.D. Thesis

Muhammed Shifas P. V.

Heraklion
July 2022

UNIVERSITY OF CRETE
SCHOOL OF SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

Neural Networks for the Quality and Intelligibility Enhancement of Speech

Submitted by

Muhammed Shifas P. V.

in partial fulfilment of the requirements for the
Doctor of Philosophy degree in Computer Science

Author:

Muhammed Shifas P. V.
Department of Computer Science

Examination Committee:

Supervisor

Yannis Stylianou, Professor, University of Crete, Greece

Member

Simon King, Professor, University of Edinburgh, United Kingdom

Member

Martin Cooke, Professor, University of the Basque Country, Spain

Member

Panagiotis Tsakalides, Professor, University of Crete, Greece

Member

Nikolaos Komontakis, Assistant Professor, University of Crete, Greece

Member

NKatsamanis

Athanasios Katsamanis, Principal Researcher, Athena Research Center, Greece

Member

Yannis Pantazis, Principal Researcher, FORTH, Greece

Departmental Approval:

Chairman
of the Department

Antonis Argyros, Professor, University of Crete, Greece

Heraklion, July 2022

Acknowledgements

During this fives years of Ph.D. research, I had the opportunity to work with exceptional people and scientists who helped me to evolve both in my research and personality. First of all, I would like to thank my supervisor, Professor Yannis Stylianou, for his continuous support and encouragement that have provided me with an insight to conduct research in a focussed and organised way without the loss of attention. More importantly, his training made me to view things from different angles when they were least expected.

At this point, I must also thank my secondary supervisors Martin Cooke, Professor, University of the Basque Country, Spain. and Simon King, Professor, University of Edinburgh, United Kingdom for their continuous support and timely monitoring of my work that helped to keep the research on track. Beyond that, for their valuables comments on shaping this thesis to the present state. I should not forget to thank Professor Panagiotis Tsakalides and Assist. Professor Nikolaos Komontakis form the University of Crete, Greece. Athanasios Katsamanis, Principal Researcher, Athena Research Center and Yannis Pantazis, Principal Researcher, Foundation for Research and Technology – Hellas (FORTH), Greece, for spending their valuable time to read and comment on the manuscript, beyond for accepting to be members of my thesis panel.

In the past four years of my Ph.D. journey, I shared my days with colleagues at the Speech Signal Processing Laboratory (SSPL) of the University of Crete, who made it a wonderful place to work that has helped me in so many ways to improve. I would extend my heartfelt thanks to Eirini Sisamaki, Dora Yakoumaki, Dr. Sunny Dayal, Dr. Nagraj Adiga, Dipjyoti Paul, Dr. Anna Sfakianaki for making the time together most rememberable. I extend a special thanks to Dr. George Kafentzis (Researcher, SSPL) for his continuoes support and collaborations in this duration.

While being a Ph.D. researcher at the University, I was also associated with the an European research network called ENRICH. At this point, I wish to thanks my friends in ENRICH circle who have really supported for the evolution of my research with timely exchange of research thoughts in various peer-to-peer meetings. I extend my thanks to Dr. Axel Winneke, Senior Scientist, Fraunhofer IDMT, Germany and Dr. Peter Derleth, Researcher, Sonova AG, Switzerland, for giving me the opportunity to visit and work in their team. Also thanks to Dr. Theano Chimona, ENT Department, Chania General Hospital, Greece and Dr. Catalin Zorila, Research Engineer, Toshiba Cambridge Research Laboratory, United Kingdom, for their close collaborations in different domain across this research period.

Last but not least, I extend the greatest "thank you" to my family: my parents, Aysha Valiyaparambh and Ismalutty Veettil, for their all prayers and support to reach this moment in life. My sisters Shadiya and Fathima Shirin for continuously seeking my physical and mental well being in this period. My elder brother Dr. Muhammed Shameem for sharing his knowledge and experience to shape this quest. Finally, heartfelt thanks to my wife Aquila Hamza who came to my life towards the end of the research period, but has given a great support and inspiration for the successful completion of this thesis.

Thank you all!

Ευχαριστίες

Κατά τη διάρκεια αυτής της πενταετούς διδακτορικής έρευνας, είχα την ευκαιρία να συνεργαστώ με εξαιρετικούς ανθρώπους και επιστήμονες που με βοήθησαν να εξελιχθώ τόσο στην έρευνα όσο και στην προσωπικότητά μου. Καταρχάς, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, καθηγητή Yannis Stylianou, για τη συνεχή υποστήριξή του και την ενθάρρυνση που μου παρείχαν μια διορατικότητα για να διεξάγω την έρευνα με εστιασμένο και οργανωμένο τρόπο χωρίς να χάνω την προσοχή μου. Το πιο σημαντικό, η εκπαίδευσή του με έκανε να βλέπω τα πράγματα από διαφορετικές οπτικές γωνίες εκεί που δεν περίμενα.

Σε αυτό το σημείο, πρέπει επίσης να ευχαριστήσω τους δευτερεύοντες επόπτες μου Martin Cooke, Καθηγητή, Πανεπιστήμιο της Χώρας των Βάσκων, Ισπανία. και τον Simon King, Καθηγητή, Πανεπιστήμιο του Εδιμβούργου, Ηνωμένο Βασίλειο για τη συνεχή υποστήριξή τους και την έγκαιρη παρακολούθηση της εργασίας μου που βοήθησε στη διατήρηση της έρευνας σε καλό δρόμο. Από εκεί και πέρα, για τα πολύτιμα τους σχόλια για τη διαμόρφωση αυτής της διατριβής στη σημερινή κατάσταση. Δεν πρέπει να ξεχάσω να ευχαριστήσω τον καθηγητή Panagiotis Tsakalides και τον Επίκ. Καθηγητής Nikolaos Komontakis από το Πανεπιστήμιο Κρήτης, Ελλάδα. Athanasios Katsamanis, Κύριος Ερευνητής, Ερευνητικό Κέντρο Αθηνά και Yannis Pantazis, Κύριος Ερευνητής, Ίδρυμα Έρευνας και Τεχνολογίας – Ελλάς (FORTH), Ελλάδα, που αφιέρωσαν τον πολύτιμο χρόνο τους για να διαβάσουν και να σχολιάσουν το χειρόγραφο, πέρα από την αποδοχή να γίνουν μέλη του πάνελ της πτυχιακής μου.

Τα τελευταία τέσσερα χρόνια του διδακτορικού μου. Το ταξίδι, μοιράστηκα τις μέρες μου με συναδέλφους στο Εργαστήριο Επεξεργασίας Σημάτων Λόγου (SSPL) του Πανεπιστημίου Κρήτης, οι οποίοι το έκαναν έναν υπέροχο χώρο εργασίας που με βοήθησε με πάρα πολλούς τρόπους να βελτιωθώ. Θα ήθελα να ευχαριστήσω από καρδιάς την Eirini Sisamaki, τη Dora Yakoumaki,, τον Dr. Sunny Dayal, τον Dr. Nagraj Adiga, τον Dipjyoti Paul, τη Dr. Anna Sfakianaki που έκαναν τον χρόνο μαζί πιο αξέχαστο. Εκφράζω ιδιαίτερες ευχαριστίες στον Dr. George Kafentzis (Ερευνητής, SSPL) για τη συνεχή υποστήριξη και τις συνεργασίες του σε αυτή τη διάρκεια.

Ενώ ήμουν Ph.D ερευνητής στο Πανεπιστήμιο, συνδέθηκα επίσης με το ευρωπαϊκό ερευνητικό δίκτυο που ονομάζεται ENRICH. Σε αυτό το σημείο, θα ήθελα να ευχαριστήσω τους φίλους μου στον κύκλο ENRICH που υποστήριξαν πραγματικά την εξέλιξη της έρευνάς μου με την έγκαιρη ανταλλαγή ερευνητικών σκέψεων σε διάφορες συναντήσεις πεερ-το-πεερ. Εκφράζω τις ευχαριστίες μου στον Dr. Axel Winneke, Ανώτερος Επιστήμονας, Fraunhofer IDMT, Γερμανία και τον Dr. Peter Derleth, ερευνητή, Sonova AG, Ελβετία, που μου έδωσαν την ευκαιρία να επισκεφθώ και να εργαστώ στην ομάδα τους. Ευχαριστούμε επίσης τη Dr. Theano Chimona, ENT Τμήμα, Γενικό Νοσοκομείο Χανίων, Ελλάδα και τον Dr. Catalin Zorila, Ερευνητικό Μηχανικό, Toshiba Cambridge Research Laboratory, Ηνωμένο Βασίλειο, για τις στενές συνεργασίες τους σε διαφορετικούς τομείς κατά τη διάρκεια αυτής της ερευνητικής περιόδου.

Τελευταίο αλλά εξίσου σημαντικό, απευθύνω το μεγαλύτερο ‘ευχαριστώ’ στην οικογένειά μου: τους γονείς μου, Aysha Valiyaparambh και Ismalutty Veettil, για όλες τις προσευχές και την υποστήριξή τους για να φτάσουν σε αυτή τη στιγμή της ζωής. Τις αδερφές μου Shadiya και Fathima Shirin που αναζητούσαν συνεχώς τη σωματική και ψυχική μου ευεξία αυτή την περίοδο. Ο μεγαλύτερος αδερφός μου Dr. Muhammed Shameem που μοιράστηκε τη γνώση και την εμπειρία του για να διαμορφώσει αυτήν την αναζήτηση. Τέλος, ευχαριστώ από καρδιάς τη σύζυγό μου Aquila Hamza που ήρθε στη ζωή μου προς το τέλος της ερευνητικής περιόδου, αλλά έδωσε μεγάλη υποστήριξη και έμπνευση για την επιτυχή ολοκλήρωση αυτής της διπλωματικής εργασίας.

Σας ευχαριστώ όλους!

Abstract

Speech is the most effective way to communicate ideas generated in human minds. However, spoken communication in real life is often affected by noise in the surroundings which can substantially reduce the intelligibility and perceived quality of the signal. Techniques to enhance the communication have been proposed in the past and successfully tested in modern engines like Amazon Alexa, allowing it to operate in adverse conditions. The ambient noise can disrupt both signal acquisition by a device as well as speech perception by the listener. Speech enhancement (SE) techniques are developed to restore speech from its disrupted observations, and listening enhancement (LE) techniques are designed to improve the perceived intelligibility by altering the speech before its presentation in noise as the naturally produced speech is not always very intelligible. Often SE and LE systems are operated as two independent modules in modern devices , which limit their performance. The effort in this thesis is to combine the SE and LE enhancement techniques to have an end-to-end system for communication applications. We approach the problem from neural networking perspective. As such, multiple novel architectures for SE and LE were invented, and the concepts from those models have been used to build the final end-to-end system.

Regarding speech enhancement (SE), three new architectures have been invented; two of which are in the feature domain and one in the waveform domain. The feature domain architectures formulate the enhancement task in the short-time Fourier transform (STFT) representation of speech, therefore, are parametrically less complex. Features from the two-dimensional (2D) representation of speech are extracted with the use of gruCNN neural cell, which is found effective in isolating noises with high variance. The gruCNN-SE model has outperformed state-of-the-art speech enhancement systems with standard convolution (CNN) and long short-term memory (LSTM) cells. Subsequently, a bidirectional extension of gruCNN module (BigruCNN) is proposed with the inclusion of backward dependencies among the 2D frames. Besides, a novel waveform domain network with a characteristic dilation pattern (SE-FFTNet) is presented. The SE-FFTNet is found efficient in learning the statistical dissimilarity of speech and noise in a noisy observation.

Regarding listening enhancement (LE), a novel WaveNet-like architecture to improve the listener's intelligibility in noise (wSSDRC) is proposed. The wSSDRC system performs both spectral shaping (SS) and dynamic range compression (DRC) of the input for intelligibility enhancement. The model is found to produce a median absolute intelligibility boost of 39% for normal hearing and 38% for hearing-impaired listeners in stationary noise over the unprocessed speech.

Subsequently, a novel end-to-end system which combines the objectives of SE and LE is proposed to enhance the intelligibility of noisy observations. The end-to-end system was found to increase the listeners' keyword correct rate in stationary noise from 2.5% to 60% at 0 dB input SNR, and from about 10% to 75% at 5 dB input SNR, compared with the unprocessed speech, while substantially outperforming the modular setup with SE followed by LE.

Περίληψη

Η ομιλία είναι ο πιο αποτελεσματικός τρόπος επικοινωνίας ιδεών που δημιουργούνται στο ανθρώπινο μυαλό. Ωστόσο, η προφορική επικοινωνία στην πραγματική ζωή συχνά επηρεάζεται από τον θόρυβο στο περιβάλλον, ο οποίος μπορεί να μειώσει σημαντικά την καταληπτότητα και την αντιληπτή ποιότητα του σήματος. Τεχνικές για τη βελτίωση της επικοινωνίας έχουν προταθεί στο παρελθόν και έχουν δοκιμαστεί με επιτυχία σε σύγχρονες συσκευές όπως το Αμαζον Αλεξα, επιτρέποντάς της να λειτουργεί σε αντίξοες συνθήκες. Ο θόρυβος περιβάλλοντος μπορεί να διαταράξει τόσο τη λήψη σήματος από μια συσκευή όσο και την αντίληψη της ομιλίας από τον ακροατή. Οι τεχνικές βελτίωσης ομιλίας (SE) αναπτύσσονται για την αποκατάσταση της ομιλίας από τις θορυβώδεις παρατηρήσεις της και οι τεχνικές βελτίωσης της ακρόασης (LE) έχουν σχεδιαστεί για να βελτιώνουν την καταληπτότητα αλλάζοντας την ομιλία πριν από την έκθεσή της σε θόρυβο, καθώς η φυσικά παραγόμενη ομιλία δεν είναι πάντα πολύ κατανοητή. Ως εκ τούτου, τόσο το SE όσο και το LE είναι απαραίτητα στις σύγχρονες συσκευές για να λειτουργήσουν σε διάφορες ακουστικές συνθήκες. Συχνά τα συστήματα SE και LE λειτουργούν ως δύο ανεξάρτητες μονάδες σε σύγχρονες συσκευές, οι οποίες περιορίζουν την απόδοσή τους. Η προσπάθεια σε αυτή τη διπλωματική εργασία είναι να συνδυαστούν οι τεχνικές βελτίωσης SE και LE ώστε να έχουμε ένα σύστημα από άκρη-σε-άκρη για εφαρμογές επικοινωνίας. Προσεγγίζουμε το πρόβλημα από τη σκοπιά των νευρωνικών δικτύων. Ως εκ τούτου, επινοήθηκαν πολλαπλές νέες αρχιτεκτονικές για SE και LE, και οι ιδέες από αυτά τα μοντέλα έχουν χρησιμοποιηθεί για την κατασκευή του τελικού συστήματος από άκρη-σε-άκρη. Τα παραδοσιακά συστήματα που βασίζονται σε στατιστικά είχαν περιορισμούς για την πλήρη μοντελοποίηση της δυναμικής της ομιλίας και του θορύβου. Τα νευρωνικά δίκτυα έχουν προκύψει ως εναλλακτική προσέγγιση για τη μοντελοποίηση δεδομένων. Ως εκ τούτου, αυτή η διατριβή επανεξετάζει τα προβλήματα SE και LE από την οπτική των νευρωνικών δικτύων.

Όσον αφορά τη βελτίωση ομιλίας (SE), έχουν εφευρεθεί τρεις νέες αρχιτεκτονικές, δύο από τις οποίες βρίσκονται στο χώρο των χαρακτηριστικών και ένα στο πεδίο της κυματομορφής. Οι αρχιτεκτονικές στο πεδίο των χαρακτηριστικών πραγματοποιούν την εργασία βελτίωσης της ομιλίας στην αναπαράσταση βραχυχρόνιου μετασχηματισμού Φουριερ (STFT), επομένως, είναι παραμετρικά λιγότερο περίπλοκες. Χαρακτηριστικά από τη δισδιάστατη (2D) αναπαράσταση της ομιλίας εξάγονται με τη χρήση νευρικού κυττάρου gruCNN, το οποίο βρέθηκε αποτελεσματικό στην απομόνωση θορύβων με υψηλή διακύμανση. Το μοντέλο gruCNN-SE έχει ξεπεράσει τα υπερσύγχρονα συστήματα βελτίωσης ομιλίας με τυπικά συνελικτικά νευρωνικά δίκτυα (CNN) και δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM). Στη συνέχεια, προτείνεται μια αμφίδρομη επέκταση της ενότητας gruCNN (BigruCNN) με τη συμπερίληψη εξαρτήσεων προς τα πίσω μεταξύ των 2D πλαισίων. Επιπλέον, παρουσιάζεται ένα νέο δίκτυο πεδίου κυματομορφής με χαρακτηριστικό μοτίβο διαστολής (SE-FFTNet). Το SE-FFTNet βρέθηκε αποτελεσματικό στην

εκμάθηση της στατιστικής ανομοιότητας της ομιλίας και του θορύβου σε μια θορυβώδη παρατήρηση.

Όσον αφορά τη βελτίωση της ακρόασης (LE), προτείνεται μια νέα αρχιτεκτονική παρόμοια με το WaveNet για τη βελτίωση της καταληπτότητας του ακροατή στο θόρυβο (wSSDRC). Το σύστημα wSSDRC εκτελεί τόσο φασματική διαμόρφωση (SS) όσο και συμπίεση δυναμικού εύρους (DRC) της εισόδου για βελτίωση της ευκρίνειας. Βρέθηκε ότι το μοντέλο έχει ως αποτέλεσμα μια μέση απόλυτη αύξηση καταληπτότητας 39% για κανονική ακοή και 38% για ακροατές με προβλήματα ακοής σε στάσιμο θόρυβο κατά τη διάρκεια της μη επεξεργασμένης ομιλίας.

Στη συνέχεια, προτείνεται ένα νέο σύστημα από άκρη-σε-άκρη το οποίο συνδυάζει τους στόχους του SE και του LE για να ενισχύσει την καταληπτότητα των θορυβωδών παρατηρήσεων. Το σύστημα από άκρη-σε-άκρη βρέθηκε να αυξάνει το ποσοστό σωστών λέξεων-κλειδιών των ακροατών σε στάσιμο θόρυβο από 2,5% σε 60% στην είσοδο SNR 0 δB και από περίπου 10% σε 75% σε SNR εισόδου 5 dB, σε σύγκριση με την μη επεξεργασμένη ομιλία, ενώ ξεπερνούσε σημαντικά το σύστημα με διαδοχική εφαρμογή της SE ακολουθούμενη από LE.

Abbreviations and symbols

STFT	Short-time Fourier transform
ISTFT	Inverse STFT
SNR	Signal to noise ratio
SE	Speech enhancement
LE	Listening enhancement
FFT	Fast Fourier transform
SS	Spectral shaping
DRC	Dynamic range compression
2D	Two-dimensional
Bi	Bi-directional
CNN	Convolutional neural networks
LSTM	Long short-term memory
GRU	Gated recurrent unit
TTS	Text-to-speech synthesis

Table 1 – Abbreviations.

Σ	Summation
Π	Multiplication
$p(\,.\,)$	Probability of
$f(\,.\,)$	Function of
$\lvert \,.\, \rvert$	Magnitude of
w	Frequency dimension
t	Time dimention

Table 2 – Symbols.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

Speech is the simplest medium to exchange thoughts between individuals. The production of speech involves both cognitive and articulatory processes. The cognitive process includes the selection of words and structuring of them to form meaningful sentences, while the articulatory functioning decides how individual segments of sound are produced with the controlled movements of speech organs. As such, the spoken ability evolves with changes in cognitive and articulative functionings. In addition to those, the acoustics of speech can vary based on factors such as the quality of speaker surroundings or the language proficiency of the speaker. For instance, non-native speakers are observed to produce speech at a slower rate compared to natives [BBM15]. Similarly, variations are observed when speaking in quiet and noisy ambiances [Jun96]. Such changes in speaking style can influence the perception of speech by listeners. In this modern industrialized world, speech signals are being produced and listened to in various sub-optimal acoustic conditions (e.g., mobile communication). Therefore, there is a high demand for noise-robust speech processing strategies, both for recovering speech from its noisy observations and delivering the message effectively to listeners in noise or at distance.

The restoration of speech from its noisy observations is widely known as *speech enhancement* (SE). SE helps to restore the quality and to some extend the intelligibility of the signal. Therefore, SE is seen as an essential front-end for any speech processing devices operating in practice. On the other hand, speech is perceived as non-linear bands of frequencies at the cochlear level [Pat95]. Thus, the perception can be damaged by the presence of background noise if the noise source frequency falls in the active speech band. The low sensitivity of cochlear filters to certain frequency components has been reported as the prime factor of sensorineural hearing impairment, which makes listening in noise more difficult for hearing-impaired population [GM86, Tyl86, GM88]. To improve the audibility in noise, modification of the spectro-temporal structure of speech has been proposed in the past and was found to enhance the listening experience in noise for both normal and hearing-impaired groups [BMG93, SMG90]. Since hearing-impaired listeners suffer from the low sensitivity to spectral contrast – the difference between peaks and valleys of the spectrum, the contrast enhancement with artificial modifications was observed to significantly improve the intelligibility in noise for hearing-impaired [SMG90, SM92]. Even normal-hearing population are being subjected to hearing difficulty (by noise or distance) in our day-to-day life . e.g., attending to public announcements at train stations or airports. Modification of naturally produced speech to boost its intelligibility in various sub-optimal listening conditions is called *listening enhancement* (LE).

Although many statistical approaches have been presented in the literature to address the noise influence on the production and perception of speech, most of them were based on the linear modeling of speech and stationarity of noise processes. However, speech as a complex dynamic process is produced and perceived with high-level nonlinear interactions, therefore, can not be approximated by first and second-order statistics. Besides, a great progress has been made in regards to the artificial neural networks (ANNs) domain in the recent years wherein better modeling of complex data structures is made possible Therefore, this thesis investigates the prospect of neural networking to model the nonlinear nature of speech and noise for novel *speech enhancement* and *listening enhancement* models. Subsequently, the effort has been made to combine the SE and LE systems as an end-to-end model, which can provide a generalised modelling of the real-world scenario as well as reduce the latency in the processing.

1.1 Nonlinear Speech Processing

Speech production in the design of early speech processing models was approximated with the source-filter model in which the output of sound sources is filtered by the resonant properties of the vocal tract [Fan81]. The most common source signal is the quasi-periodic vibration of vocal folds and the filter is characterized as a linear filter with time-variant properties. This approximates the physics of speech production with linear acoustics and one-dimensional propagation of sound waves in the vocal tract. Due to their simplicity, source-filter models have laid the foundation for many speech processing applications like speech coding, synthesis, and enhancement over the years.

However, this modeling of speech as the output of a linear system is only a first-order approximation of speech production dynamics, therefore neglects higher-order structures that are observed to be present in speech. For instance, in the modeling, the airflow through the vocal tract is assumed laminar with the excitation source and filter are operating independently. This negligence is manifested as distortions at the output of speech enhancement Kalman filters [PB87], or less natural speech synthesis in synthesizers [MTKI96]. Many studies in the past have reported the possible evidence for the presence of high order statistics in speech. In [TNH94], a comparison of linear and non-linear predictive models was conducted. A linear predictive coding analysis of order 10 was applied several times to the speech segment. It was observed that the prediction gain after the fifth iteration was 0 dB, indicating the remaining redundancies could not be linear. While subsequent application of the Volterra filter (a nonlinear filter) on the residue from the linear predictor was found to further reduce the prediction error, providing strong evidence of speech nonlinearity. The presence of higher-order statistics in the production dynamics was also observed on the bi-spectrum analysis of speech which measures variance beyond second-order [NM93]. Besides, speech is perceived as bands with a non-linear form at the human cochlea. Thus, nonlinear frameworks have indeed the potential of modeling or analysis of; 1.) nonlinearities in the speech generation process 2.) nonlinearities in the signal acquisition process 3.) nonlinearities in the transmission stage 4.) nonlinearities in the perception mechanism. Besides, some problems are difficult to solve with linear techniques and are more tractable with nonlinear ones [FZME^{+}02].

On the other hand, there are certain limitations when dealing with non-linear techniques. First, there is no unified theory among different non-linear methods. Second, they are computationally more complex and much more difficult to analyze with the traditional tools. Besides, in situations where the closed-form solution does not exist, iterative solution should be followed, therefore the local minimum problems exist.

The non-linear methods for speech processing is a rapid growing area of research mainly since the mid 1980s. A variety of techniques aimed at engineering applications have been proposed since then, a broad classification of which is possible as *parametric* – e.g., non-linear predictive vector quantization [WNF94] or codebook prediction [KG97], and *non-parametric* – e.g., quadratic filter [Sic92] or extended Kalman filter [Bir95]. Among the parametric models, artificial neural networks (ANNs) have gained popularity in recent years for their ability to better account for the high-level nonlinearities in the acoustic or electro-acoustic phases of speech processing. For instance, the integration of ANN modules to model speech and noise processes in the mixture in the Kalman filter-based speech enhancement was reported to reduce distortions on the enhanced signal in various adverse conditions [WNK^{+}99, RNP20]. Therefore, this thesis investigates the prospect of using ANNs for enhancing the perceived quality and intelligibility of speech recorded in various acoustic conditions. Since having basic knowledge about neural networks is required to follow the different neural architectures that are presented in the upcoming chapters of this thesis, a brief overview of artificial neural networks is presented next.

1.1.1 Artificial neural networks (ANN)

Humans with experience can visually recognize objects and perceptually segregate sound sources effectively in various real-life situations. The brain's ability to selectively switch attention from one to another makes communication possible in noisy environments, a phenomenon often referred to as cocktail party problem [HC05]. Indeed, this has to do with the evolution of the human brain over the centuries, but the experience individuals gain in the early stages of life is also a contributing factor. With the mass accumulation of human-centric data and rapid growth in computing powers in the last few decades, the machine can be directed to learn the experience factor from data to better predict events in the future – popularly known as artificial intelligence (AI) [RN02]. The AI has proven its success in a variety of applications affecting numerous aspects of human life; human-computer interaction [Xu19], biometric applications [SAK$^+$17], security and surveillance [GSD$^+$18], natural disasters monitoring [ORC18] and many more. A class of machine learning techniques called the artificial neural networks (ANNs) [Yeg09] plays the central role in modern AI realizations.

The design of ANNs was inspired by the hierarchical information processing in the primitive visual system of the human brain wherein the information is being processed at different levels of abstraction. ANN is built with hierarchically arranging layers of non-linear modules in a pre-defined order. The resulting deep hierarchical structure is called the deep neural networks (DNNs). DNNs have a dominant role in speech processing as, if trained effectively, the model can capture the nonlinearities involved in the production, acquisition, transmission, and perception of speech. With the availability of large acoustically and linguistically variant data sets for public use, we believe that a stable training of DNNs is possible. Although several variants of DNNs are present in the literature to date, an architectural classification of those is possible based on the basic blocks used to build the network, namely fully connected neural networks (FCNNs) [Mur91], convolutional neural networks (CNNs) [KSH17, SZ14] and recurrent neural networks (RNNs) [HS97, GSC99]. Each of which differs on the way information is being propagated from the lower level to high hierarchical levels, which is discussed in detail next.

The modeling of the biological neuron is fundamental to designing neural networks. An approximated mathematical representation of a biological neuron is depicted in Figure 1.1. The input vector $\boldsymbol{x} = [x_1, x_2, ..., x_4]$, that can be a series of speech samples, are being linearly weighted by the parameters $\boldsymbol{w} = [w_1, w_2, ..., w_4]$ and merged with the bias coefficient b. This linear representation is transformed by a non-linear transformation function ϕ to have the final non-linear output y. Mathematically, these transformations can be formulated as

$$y\left(\boldsymbol{x}\right) = \phi\left(z\right) = \phi\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right) = \phi\left(\sum_{j=1}^{n} w_j x_j + b\right) \tag{1.1}$$

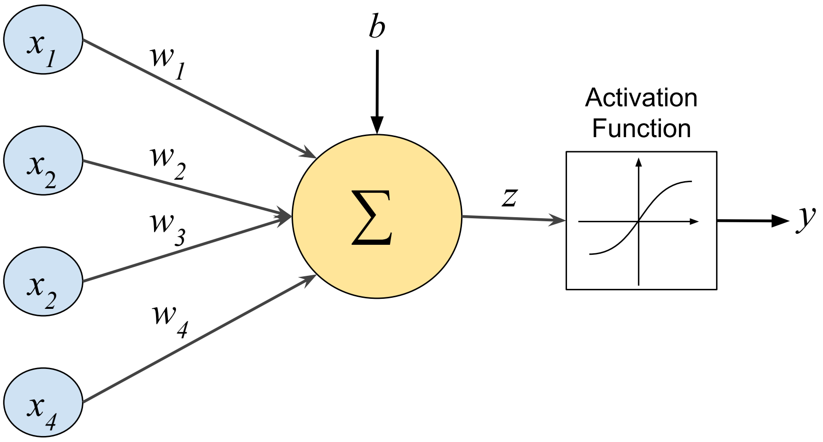


Figure 1.1 – Mathematical representation of biological neuron.

The non-linear function ϕ, often called activation function in the context of neural networks, can be any user-defined function in practice. One of the main objectives of using such an activation function is to limit the output overshooting which allows stable training of the model. However, few common choices have been observed to perform best in most applications, e.g. ReLU [Aga18] and Sigmoid activations [NIGM18], both of which compress the input dynamic range with different mapping functions. ReLU transforms its input z with the function $f(z) = max(0, z)$, while Sigmoid follows the transformation $f(z) = \frac{1}{1+e^{-z}}$. Since we don't use any parametric functions in the activation module, the final output y is purely a function of input random variable $\mathbf{x}$ and model parameter $\mathbf{w}$.

Once the mathematical modeling of the neuron is complete, building deep architectures is rather straightforward. By tiling multiple neurons in a 2D plane and connecting their receptive fields together, a deep network would have the structure in Figure 1.2. As a single neural cell is too basic to model complex non-linearities at the input, it is only through deep networks we can fully capture relevant patterns in the data. As the layers between input and output in Figure 1.2 are invisible to an observer outside, it is often called hidden layers of the network. The dimension of the hidden layer is determined by how many number of neural cells are used at each stage. Whereas, the input and output dimensions depend on the task for which the network is designed. For instance, in the case of automatic speaker verification network [PPG^{+}16], the input could be either continuous speech samples or hand-crafter features while the output will be a fixed vector with individual elements representing the probability of the input belonging to a certain speaker.

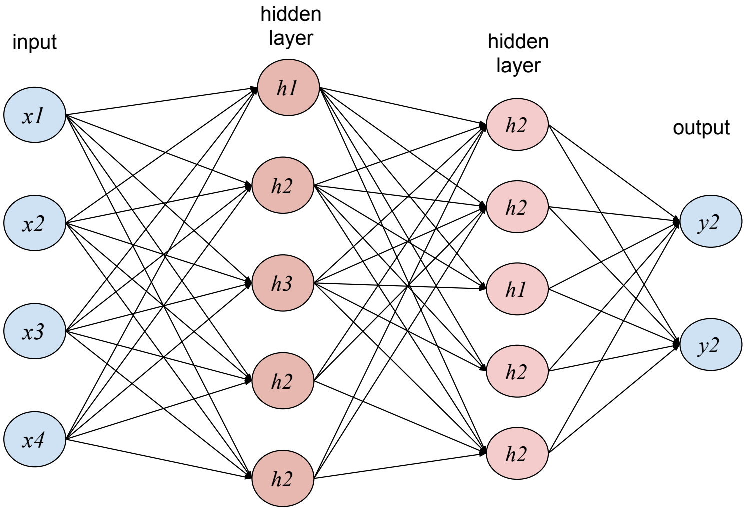


Figure 1.2 – Neural networking.

Since each node in the architecture is connected to every other node in the preceding and succeeding layer, this category of networks is called fully connected neural networks (FCNNs), or multilayer perceptrons (MLP). The individual layers are called fully connected layers. Although the choice of activation functions can be different at each layer and in each node, it has been a general practice to use the same activation across layers except in the final layer where the Softmax layer [GBC16] is often used to convert the hidden features to probabilistic distribution.

1.1.2 Training procedure

The most important non-structural aspect of a neural network is the optimization of its parameters for a specific task. This is often referred to as the training of the network. The training can be performed either in supervision (supervised learning) or non-supervision (unsupervised learning) [BMY19] fashion. Supervised learning requires the availability of labeled data wherein the input-output paired are marked for each training set sample. On the other hand, unsupervised learning does not require such prior labeling of data. A good example of unsupervised learning is the K-means clustering [VO^{+}13] where the data points are grouped such that each point belongs to the cluster with the nearest mean. Since supervised learning is more robust than unsupervised learning and due to the availability of large-scale labeled data, supervised training strategies are followed across this thesis. As such, the supervised optimization of network parameters is discussed next.

Imagine that we are given a set of input-output sample pairs

$$\left\{\left\{\boldsymbol{x}^{1}, \boldsymbol{q}^{1}\right\} \ldots, \left\{\boldsymbol{x}^{n}, \boldsymbol{q}^{n}\right\}\right\}, \tag{1.2}$$

which is a good representative of the true data distribution we are planning the model to be deployed on. For each input $\boldsymbol{x}^p$, we compare the prediction of the network $\boldsymbol{y}(\boldsymbol{x}^p)$ with the true label $\boldsymbol{q}^p$ to estimate the divergence of model from the true distribution. This deviation can be quantified in different ways on different metrics. One of the most common approaches, as well as the one that has been extensively used in our experiments, is the mean squared error (MSE) criterion. MSE is defined as

$$E = \frac{1}{2} \sum_{p=1}^{n} \sum_{i=1}^{K} \left(y_i\left(\boldsymbol{x}^p\right) - q_i^p\right)^2, \tag{1.3}$$

where K denotes the dimension of the output vector or simply the number of output neurons, and n is the size of the training set.

Once the objective function has been defined, the next step is to optimize the model parameters such that to minimize the loss function E, alternatively, to reduce the prediction divergence from the true distribution. Although it is not explicitly mentioned, note that $\boldsymbol{y}$ is a function of both the input data variable $\boldsymbol{x}$ and network parameters $\boldsymbol{w}$. To better understand the optimization process, let us consider the single neuron model presented in Figure 1.1 with sigmoid activation as a sample network. The inner summation in Eqn.1.3 which represents the sum of the squares of elements of the vector can be rewritten as vector product operation

$$E = \frac{1}{2} \sum_{p=1}^{n} \left(\boldsymbol{y}\left(\boldsymbol{x}^p\right) - \boldsymbol{q}^p\right)^{\top} \left(\boldsymbol{y}\left(\boldsymbol{x}^p\right) - \boldsymbol{q}^p\right) = \sum_{p=1}^{n} E^{(p)}. \tag{1.4}$$

Our objective is to minimize this multi-dimensional function in the parameter space $\boldsymbol{w}$, while keeping the data variable $\boldsymbol{x}$ untouched. For this, we first compute the derivative of the error function E with respect to each parameter w_i. With Figure 1.1 as the reference, this becomes

$$\frac{\partial E}{\partial w_i} = \left(\frac{\partial E}{\partial z}\right)\left(\frac{\partial z}{\partial w_i}\right). \tag{1.5}$$

The term $\frac{\partial E}{\partial z}$ can be further split into products of local derivatives as

$$\frac{\partial E}{\partial w_i} = \left(\frac{\partial E}{\partial y}\right)\left(\frac{\partial y}{\partial z}\right)\left(\frac{\partial z}{\partial w_i}\right). \tag{1.6}$$

Let us first compute the gradient for a single data point $(\boldsymbol{x}^p, \boldsymbol{q}^p)$, with the sigmoid activation function $y(\boldsymbol{z}) = \frac{1}{1+e^{-z}}$, the local derivatives are:

$$\frac{\partial E}{\partial y} = (y(\boldsymbol{x}^p) - \boldsymbol{q}^p), \tag{1.7}$$

$$\frac{\partial y}{\partial z} = y(\boldsymbol{x}^p)(1 - y(\boldsymbol{x}^p)), \tag{1.8}$$

$$\frac{\partial z}{\partial w_i} = x_i. \tag{1.9}$$

Therefore, the joint expression becomes

$$\frac{\partial E}{\partial w_i} = (y(\boldsymbol{x}^p) - \boldsymbol{q}^p) y(\boldsymbol{x}^p)(1 - y(\boldsymbol{x}^p)) x_i. \tag{1.10}$$

With the inclusion of the full training set samples in Eqn. 1.4, the gradient vector $\boldsymbol{\nabla} E$ for the network parameter w_i becomes

$$\boldsymbol{\nabla} E = \sum_{p=1}^{n} \left(y\left(\boldsymbol{x}^p\right) - \boldsymbol{q}^p\right) y\left(\boldsymbol{x}^p\right) \left(1 - y\left(\boldsymbol{x}^p\right)\right) x_i. \tag{1.11}$$

This computed gradient points the direction to which the error function E increases in the parametric space w_i. To reach at the minimum in this parametric space, one must walk on the direction to which the gradient descents. Therefore, the parameter w_i is being updated for the iteration t to $t+1$ as

$$w_i^{(t+1)} = w_i^{(t)} - \gamma \boldsymbol{\nabla} E, \tag{1.12}$$

where γ is a scalar constant that defines the size of updation. The same process is repeated for all i in the network. The entire process is known as the gradient descent optimization. Different extensions of the same algorithm have been proposed in the literature for stable optimisation of neural networks, interested readers are referred to [Rud16].

1.1.3 Convolutional neural networks (CNN)

The type of network that we saw in the above section is called the fully connected network (in Figure 1.2). In such networks, the activation of each node depends on the entire input signal. Therefore, the activation does not convey anything regarding the local statistics of the input data. For instance, speech signals only exhibit local stationarity and vary dynamically over time. Therefore, detecting local structures of the data is very important in the design of models for natural language processing. Convolutional neural networks (CNNs) have been demonstrated to be efficient in detection of spatial variations in data such as the recognition of differently posed faces in images [SZ14], [KSH17], [GWK+18]. This is because, in contrast to the matrix operations in fully connected networks which define the internal states, CNNs perform the local filtering of the input signal which helps isolate the local patterns such as the edge transitions in a 2D object. A brief overview of building convolutional networks for one-dimensional data such as speech is presented below.

Given an input speech segment $X = [x_1, x_2, ..., x_n]$ to a convolutional layer with kernel parameters $K = [k_1, k_2, k_3]$. The convolved output $Y = [y_1, y_2, ..., y_n]$ in mathematics is defined as:

$$y_i = \sum_{j=1}^{3} k_j x_{i+j} \text{ for } i = 0, 1, \ldots, n-3 \tag{1.13}$$

A visual representation of this filtering is depicted in Figure 1.3, where the kernel slides over the input to have a filtered representation of the original signal.

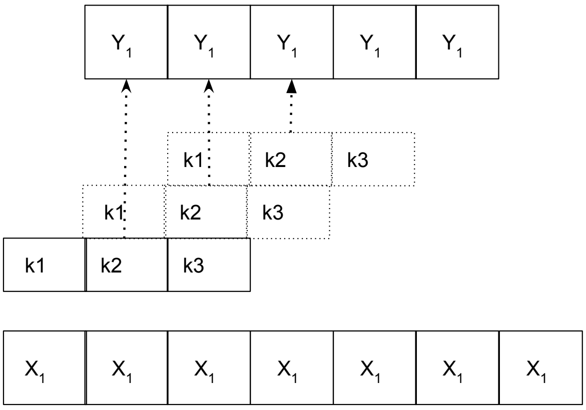


Figure 1.3 – A single layer convolution.

Compared to the input, the length of the output is shortened in the process. This can be overcome with the addition of extra zeroes at the ends of the input signal, which is popularly known as zero-padding in the neural network domain. Additionally, in the above example, the filter strides only a single sample at every operating instant, which helps keep redundant representations at the output. However, this can be increased to higher values to reduce the computational complexity at layers of the network without losing much of the representativeness. The stride of convolution (S) largely defines the dimension of the convolved output. Mathematically, the length of a convolved representation (O) for an input of length W filtered with convolution kernel of size K, zeros padding P and stride S is

$$O = \frac{W - K + P}{S} + 1. \tag{1.14}$$

Therefore the larger the stride, the smaller the output gets. In the above modeling, only a single convolution kernel was considered. Instead, there can be a number of kernels operating in parallel on the input returning individual feature streams as illustrated in Figure 1.4. The filtered representation of the input is called feature maps in neural network terminology. The number of feature maps is equivalent to the number of filters employed in that layer, also called the *channel dimension*.

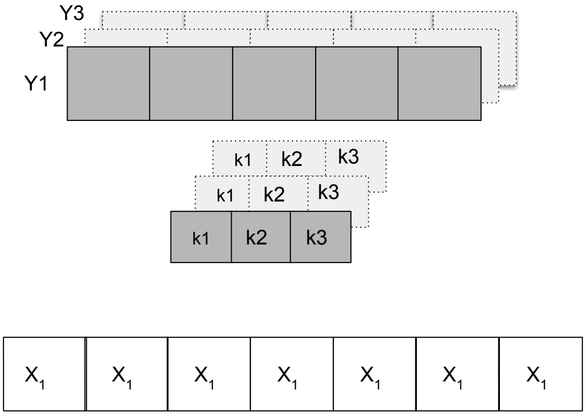


Figure 1.4 – Convolution feature maps.

Deep convolutional networks are built by stacking the convolutional blocks having the structure shown in Figure 1.4 one over the other. Since the kernels are of fixed size and are shared over the stretch of the input, the parameter complexity of the network is far lesser than that of the fully connected networks. Generally, in deep neural models, we usually observe a series of convolution layers at the beginning to extract features followed by a fully connected layer to merge the feature representations to a vector representation such as class scores in a classification task.

There is a main sub-class in the convolution category of networks, they are called dilated convolution networks. In contrast to the normal convolution, the kernels of dilated convolution layer have only a few active coefficients and the rest is set to zero. For instance, in Figure 1.5 we set the second coefficient of the filter to zero, whereby, the data points falling in the middle of the filter are not counted. However, this does not lead to the loss of such data points as they have been counted in the next instance of filtering. The main objective of using dilated convolution is to reduce the parameter complexity of models without affecting the data structure. As such, such dilated layers with varying dilation factors (the amount of dilation in the convolution operation) have been widely used for building very deep neural models [VOKK16, RPS18]. WaveNet, initially proposed by [ODZ+16], uses a dilated convolution structure with the dilation factor decreases by a factor of 2 over the layers starting from the beginning.

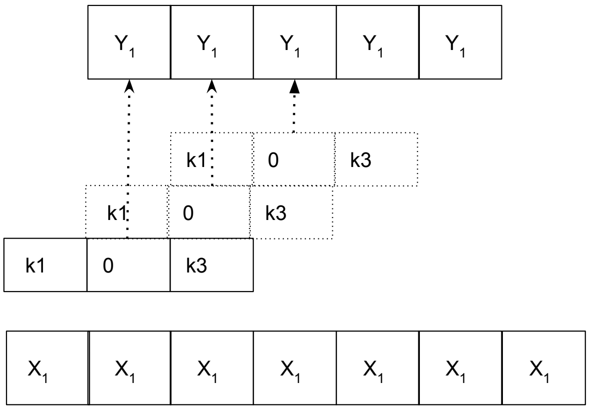


Figure 1.5 – Dilated convolution layer.

1.1.4 Recurrent neural networks (RNN)

In the above formulation of data structure, we ignored a very important aspect of speech or any time series for that matter, which is the temporal correlation of data. Speech is the pressure variations resulted of a continuous articulation, therefore, can be modeeled as an autoregressive (AR) process where the sample at each time instance is evolved out of the past samples. This is because speech is produced with the dynamic movement of vocal organs wherein the transition from one state to the next is highly correlated. Therefore, efficient modeling of the sequence of the samples is crucial when designing robust speech processing models. Many statistical models like linear predictive (LP) modeling have been tried in the past to model this correlated structure of speech and were successfully used in applications such as speech recognition, low bit rate coding.

In the context of neural networks, the temporal flow of time-series is modeled with the use of recurrent neural networks (RNNs) which can persist the long-term dependencies at the input. As an example, take a segment of speech samples $X = [x_1, x_2, ..., x_t]$. The joint distribution of the samples can be mathematically expressed as

$$p(X_k) = p([x_1, ..., x_t, ..., x_T]), \tag{1.15}$$

with the chain rule of probability, the joint distribution can be split into individual fractions as

$$p(X_k) = \prod_{t=1}^{T} p(x_t | x_{<t}). \tag{1.16}$$

Under the assumption that speech production process is a Markov process, the probability of next state can be fully defined with the knowledge of the current state. Therefore, with the replacement of $p(x_t|x_{<t}) = p(x_t|x_{t-1})$, the above formulation becomes

$$p(X_k) = \prod_{t=1}^{T} p(x_t | x_{t-1}). \tag{1.17}$$

This evolution of samples can be captured with the use of an RNN cell. At any given time instant t, RNN module takes inputs the current sample x_t and the previous state h_{t-1} that encodes the evolution of samples. Then, these are non-linearly combined to get the final prediction for the current time instance y_t. This is graphically illustrated in Figure 1.6.

Different types of RNN are possible based on how the input and past-state are mixed inside the module. Although

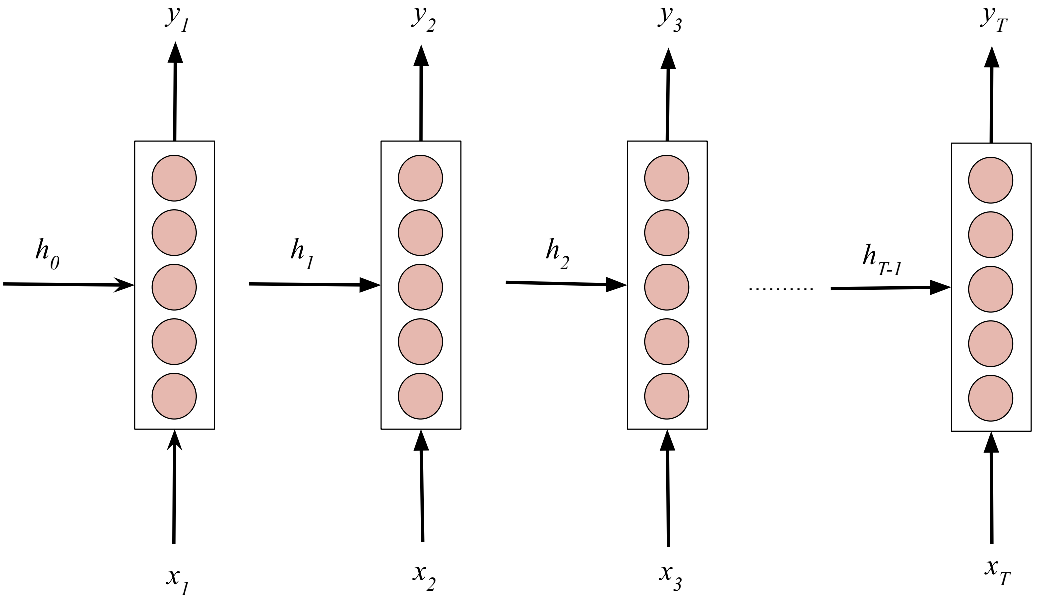


Figure 1.6 – Recurrent neural network (RNN) once unfolded over time.

RNNs can predict samples more accurately than CNNs or FCNNs, most of them are hard to train due to the vanishing of gradient – the gradient computed at the output node fails to propagate smoothly back through all instances of the network while training. This becomes a major problem when the network size increases, or as the input data size grows. Among different RNN architectures, Long Short-Term Memory (LSTM) [HS97] and Gated Recurrent Unit (GRU) [CGCB14] are popular for their ability to mitigate the vanishing gradient problem when training. As such, LSTM and GRU cells have been successfully used in many speech processing applications like automatic speech recognition and speech enhancement [WEW+15], [SSB14]. In the case of GRU, the input and state vectors are being merged by the non-linear transformations

$$z_t = \sigma(W_{zh}h_{t-1} + W_{zx}x_t + b_z) \tag{1.18}$$

$$r_t = \sigma(W_{rh}h_{t-1} + W_{rx}x_t + b_r) \tag{1.19}$$

$$\hat{r_t} = \tanh(W_{hh}(z_t \odot h_{t-1}) + W_{hx}x_t + b_h) \tag{1.20}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \hat{h_t}. \tag{1.21}$$

Whereas, the LSTM follows the formulation

$$i_t = \Phi(W_{xi}X_t + W_{hi}h_t + W_{ci} \odot c_{t-1} + b_i) \tag{1.22}$$

$$f_t = \Phi(W_{xf}X_t + W_{hf}h_t + W_{cf} \odot c_{t-1} + b_f) \tag{1.23}$$

$$o_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{xc}X_t + W_{hc}h_{t-1} + b_0) \tag{1.24}$$

$$y_t = \Phi(W_{xo}X_t + W_{ho}h_t + W_{co} \odot c_{t-1} + b_y). \tag{1.25}$$

1.2 General Objective

With such a considerable progress in machine learning research in the last decade, there is a ncessity to investigate the usefulness of neural models to facilitate speech communication in real-world. Noise in the communication background can deteriorate the quality and intelligibility of speech both for humans and machines. Speech intelligibility is broadly identified as the understandability of the underlying message in a speech sound, whereas the quality is more of a psychological factor that represents how natural the signal is. As such, preserving the quality and intelligibility of speech in communications is essential for novel human assistive devices in practice.

A typical speech communication scenario is depicted in Figure 1.7., where both the speaker (far-end) and listener (near-end) are being affected by noise. The design objective of speech processing models is to effectively deliver the message communicated by the speaker to the listener in the presence of the noises. Noise at the speaker side that affects the speech acquisition is called the *far-end noise* (far with respect to the listener). Similarly, noise at the listener's end which affects the speech perception of the listener is called *near-end noise*. Although distortions could occur in the processor module during speech transmission from one end to the other, only the impact of ambient noise on speech quality and intelligibility is considered in this thesis, assuming an ideal transmission channel. From a design perspective, the main difference between the two types of noise is that the impact of far-end noise on speech can be measured and manipulated at the processor module with the use of speech enhancement (SE) algorithms, whereas the impact of near-end noise is not quantifiable as it directly interacts with the speech at the listener's auditory periphery. Therefore, unless the listener wears a hearing assistive device that can cancel the background (which is not so common in practice), the intelligibility can be significantly reduced with the auditory masking from near-end noise.

However, to promote listeners' intelligibility, modification of speech spectral and/or temporal structure before its presentation to adverse listening conditions was proposed in the past. Such structural modifications of speech were also observed in humans when speaking in noisy ambiance, adjusting their articulations to alleviate the masking of background perceived through the auditory feedback, a phenomenon widely known as the Lombard reflex [Jun96]. The speech produced in noise (Lombard speech) has many characteristic features such as higher fundamental frequency, reduced speaking rate, and flatter spectral tilt compared to the plain speech produced in quiet speaking condition [CMV14, CMVB$^+$13c, LC09]. Lombard speech was found to be more intelligible than plain speech for listeners in noise even after removing the loudness variations [BC20, CL12]. As such, speech has to be adjusted based on the listening context to ensure its intelligibility for the interlocutor. Similarly, artificial modifications of speech are required in speech output devices to ensure their intelligibility in various operating conditions. This field of research is known as listening enhancement (LE).

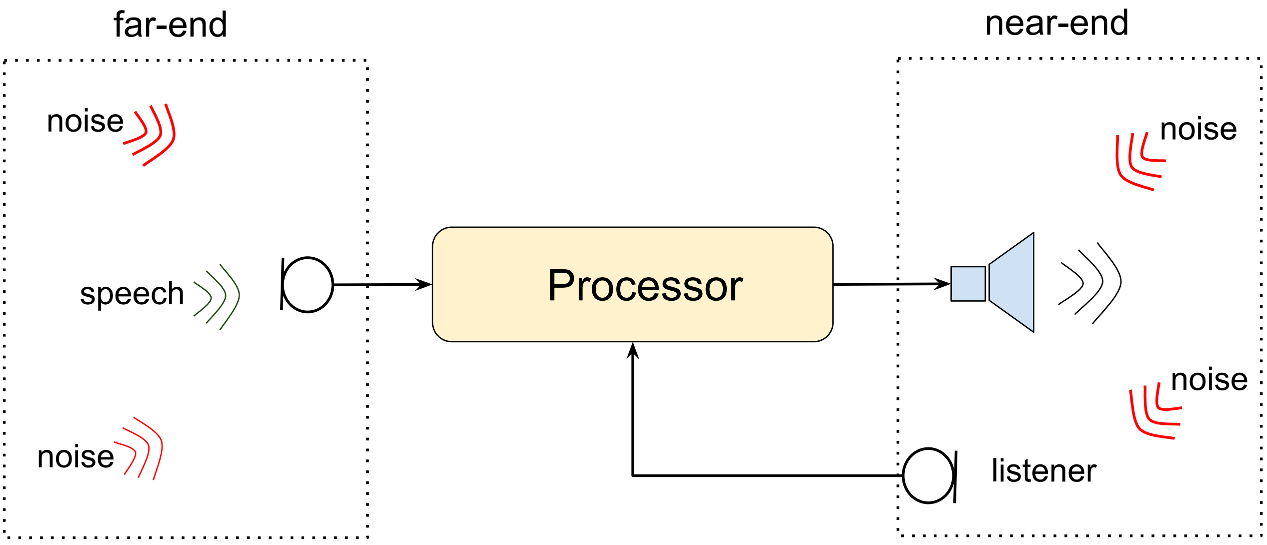


Figure 1.7 – A typical communication scenario.

Far-end	Near-end	Processor requirement
Quiet	Quiet	Idle
Noisy	Quiet	Speech enhancement (SE)
Quiet	Noisy	Listening enhancement (LE)
Noisy	Noisy	SE and LE

Table 1.1 – The possible combinations of noise in Figure 1.7.

In reality, the noise level at the far- and near- ends varies based on the operating environment. Therefore, SE and LE algorithms would have to be used in varying degrees. The four possible acoustic scenarios in reference to Figure 1.7 together with the optimal processing strategies are presented in Table 1.1. As such, the prime objective of this thesis is to build models to operate in each of those conditions in order to effectively deliver the message from the speaker to the listener with a minimal impact of noise on the listener.

Most of the past attempts considered speech enhancement (SE) and listening enhancement (LE) as two independent fields of research. However, in many applications, e.g. mobile communication in outdoor ambiance, both of them must be required to operate in parallel (the final condition in Table 1.1). Therefore, a great amount of effort has been made in this thesis to combine the SE and LE modelings as a single unified framework. Such an end-to-end system has few advantages: 1) it would avoid the propagation of error from the SE to LE module when operating in modular fashion, therefore would improve the performance, 2) the joint modeling would also reduce the computational complexity of the modular setup while providing a more generalized representation of the problem.

1.3 Motivation and Vision

Recent advancements in neural networks stitched together with the growth of computational power and the availability of linguistically diverse speech recordings provide us with the opportunity to develop a new class of speech processing models. With its inherent non-linear structure, neural networks can capture higher order dynamic variations of speech compared to traditional statistical models. This would particularly be useful to recover speech from its noisy observations where the statistical approaches often resorted to primitive assumptions on the noise process, e.g. Gaussian assumption in speech enhancement Kalman filter (KF) [PB87], which has resulted to poor estimation of the noise and distortions on the output speech. On the contrary, NNs provide the flexibility to learn complex noise patterns in supervised learning framework [SCS+20]. For instance, linear predictive modeling such as linear autoregressive (AR) models was endorsed in the past to better restore speech from noisy observations. Subsequent replacement of the AR module with a neural architecture was proven to further improve speech restoration of the enhancement system [WNK+99]. Besides, the introduction of residual networks [HZRS16] has made plausible the training of very deep networks, resolving the vanishing gradient problem. Since then, deep architectures that can operate on the waveform of speech, such as WaveNet [RPS18] and FFTNet [JFML18], have been proposed and have attracted considerable attention.

Most existing neural models introduced in the literature for speech enhancement ignore the phase information of speech while enhancing only the magnitude spectrum of the short-time Fourier transform (STFT) representation. This limits the quality of enhanced signal. Waveform models can overcome such limitations by cleaning both the magnitude and phase componenets jointly. Although models operating on the waveform domain of speech have been proposed in the past, the context covered with the receptive field of such networks was relatively short, like 5 – 10 ms, due to the liner increase of parameter complexity with fully connected layer that was popular back then. However, the introduction of dilated convolution layer gives us the flexibility to build deeper models owning to its weight-sharing property. Such models can capture longer dependencies among the waveform samples taking both the phase and amplitude of speech into consideration. In spite of that, modeling long-term temporal dependency with convolution layers requires very deep architecture, which limits their applicability in low-end devices like hearing aids. On the other hand, the long-term speech dependency can be more efficiently captured with recurrent cells [GSC99, HXY15]. However, such neural cells require a redesign to fully account for the statistical structure of speech and noise for the enhancement task. Therefore, the usability of recurrent cells for speech enhancement was investigated and structural changes have been suggested as part of this thesis.

Another factor of motivation originated from the pattern learning ability of neural networks. Neural networks are generally seen good at learning patterns from data. In the case of speech, some speaking styles are observed to be more intelligible than others in perception studies. With the availability of various style databases, neural models can be trained to produce customized styles based on the listener's (user) requirement. This is especially useful in noisy listening scenarios, where the intelligibility degradation of natural style is often observed. Artificial styles such as spectral shaping and dynamic range compression (SSDRC) [ZKS12] were found to produce the best intelligibility in noise for both normal and hearing-impaired listeners [ZSFM17]. Such systems requires clean speech at the input to synthesize matching high intelligibility voice. Besides, they are highly sensitive to the noise at input. This has limited their applicability in applications like telephony or hearing aids where the recorded speech is often noisy. We believe that noise robust listening enhancement systems can be built by fusing the noise robustness and pattern learning ability of neural networks. Therefore, we also investigate the prospect of training neural models to generate intelligible styles in sub-optimal acoustic settings.

1.4 Research Questions

Research findings reported in this thesis can be broadly classified as the answers to these three questions:

1. How to design neural architectures to learn better the statistical dissimilarity of speech and noise in a noisy mixture for novel noise removal ?
2. Can neural networks produce intelligible voices to enhance speech perception for listeners in adverse conditions ?
3. Can we develop an end-to-end neural system to jointly suppress the noise at input and boost the intelligibility at output ?

1.5 Contributions of this Dissertation

The key contributions of this thesis are outlined below:

1. Contributions to speech enhancement (SE) :
   - An efficient approach to model the temporal dependency among the input 2D spectrograms is proposed. The proposed method extracts features that are temporally relevant at each frame instance with the use of gruCNN recurrent cell. The gruCNN enhancement network (gruCNN-SE) learns the temporal flow of speech at multiple layers of abstraction starting from the inout, which helps extract robust noise discriminating features at the layers. gruCNN-SE is found to outperform substantially traditional enhancement networks with independent modeling of feature extraction and temporal dependency, as well as reduces the parameter complexity of the system considerably.
   - A bidirectional extension of gruCNN neural cell (BigruCNN) is proposed for modeling both the forward and backward dependencies of a time series. BigruCNN is tested in the context of speech enhancement for modeling the correlations among 2D spectrograms of input speech. The BigruCNN was observed to be highly efficient to detect and isolate locale events in the spectrum, therefore, has proven to generalize better to

non-stationary and unseen noise conditions. Meanwhile, it had a far lesser number of parameters compared to existing speech enhancement networks, making it a desirable architecture in practice.

- To avoid the phase distortions associated with conventional feature domain enhancement models, a new waveform domain enhancement network called SE-FFTNet is proposed. SE-FFTNet processes noisy speech samples in time domain and returns the predictions of underline clean speech at the same resolution. Compared to other waveform domain enhancement models, SE-FFTNet produces better quality enhancement due to its unique dilation pattern in the convolutions. Besides, the SE-FFTNet has fewer parameters than other waveform domain SE architectures.

2. Contributions to listening enhancement (LE) :

   - A WaveNet-like architecture to enhance speech intelligibility for listeners affected by adverse conditions is proposed. The system called wSSDRC is trained with a new training paradigm, where a signal processing technique called spectral shaping and dynamic range compression (SSDRC) is being used as a teacher module to educate the wSSDRC student network about speech intelligibility features. Once trained adequately, the wSSDRC is observed to produce intelligibility boosts comparable to the teacher SSDRC for a range of listening groups – native, non-native, normal, and hearing-impaired listeners.
   - The concept from wSSDRC training framework inspired us to build intelligible text-to-speech (TTS) synthesizers that can generate various intelligibility styles from text in an end-to-end manner. When evaluated in stationary noise, the best TTS system is found to produce relative intelligibility improvements of up to 455% compared to the traditional TTS model.

3. Contributions to joint SE and LE :

   - An end-to-end intelligibility enhancement system aimed to improve the intelligibility of noise-corrupted recordings is proposed. The system jointly models the noise suppression (of the input) and intelligibility modification (at output) as a single unified learning task. The system has a similar convolutional pattern of SE-FFTNet in the waveform domain. Both causal and non-causal extensions of the system were tested. The training framework of wSSDRC was partly adapted to optimize the model parameters. Subjective and objective evaluations were conducted in a range of noise conditions. The results indicate that the neural model can produce up to 140% average intelligibility boost for listeners in noise compared to the state-of-the-art statistical approaches.

1.6 Outline of Dissertation

The rest of this dissertation is organized as follows. The content is split into three main parts. In the first part, we talk about SE techniques to restore speech from its noisy observations. A review of the state-of-the-art methods is conducted and new architectures both in feature and waveform domains are presented. In the second part, we discuss the existing LE techniques and propose neural network models to generate intelligible speech to improve the listening experience in noise. In the final part, we combine the knowledge from the previous parts to design an end-to-end approach for joint speech and and listening enhancement, unveiling the noise-robust intelligibility modification system.

Part I

Speech Enhancement

Chapter 2

Introduction

Speech acquisition in the real world is often affected by noise in the background, which can degrade both the perceived quality and intelligibility of the signal. Speech enhancement (SE) is concerned with improving the quality and/or intelligibility of speech by suppressing noise components in a noisy observation. Therefore, SE is required in varying degree in speech processing systems such as automatic speech recognition (ASR) which comes under constant noise attacks in real-world operations. As such, there is a great demand for robust enhancement systems to model various challenging acoustic conditions a device would have to operate. This chapter is devoted to the discussion of speech enhancement with the approaches that have been presented in literature and the new alternatives suggested in this thesis.

First, let us define the problem more mathematically based on which the solutions can be presented. The degradation of speech $s[t]$ articulated by a speaker can be broadly formulated in time domain as

$$y[t] = h[t] \star s[t] + n[t] = s^h[t] + n[t], \tag{2.1}$$

where $y[t]$ is the observation of the signal $s[t]$, and $n[t]$ is the additive noise in the background which is assumed uncorrelated to speech. The $h[t]$ denotes the coupling between the speaker and sensing device. The symbol $*$ represents the convolution operation which transforms the true speech to its representation $s^h[t]$ at the input of the sensor. There are special occasions in practice where the noise would be correlated to speech, e.g. reverberation (reflection of the speech from the surroundings), which demands a different formulation of the restoration task. However, our objective in this thesis is constrained at suppressing the uncorrelated background $n[t]$ for it is the most common in practice. Besides, the methods covered in this chapter do not compensate for channel distortion but are designed purely for removing the background noise $n[t]$. This is under the assumption that the channel was designed properly to ensure the ideal coupling between the speaker and sensor, i.e., $h[t]$ = δ[t]. In practice, multiple sensors can be used to collect speech at different spatial locations and later be combined them using *beam forming* [Com92] and *noise cancellation* [Ste85] techniques. However, we limit ourselves to single sensor observation.

Having a good knowledge about the statistical characteristics of speech and noise would help develop robust noise reduction strategies. A sample speech signal observed at 16 kHz sampling rate is displayed in Figure 2.1(a) together with its frequency content over time as a spectrum. It is observable that some regions are more oscillatory in nature with information mostly at low frequencies. They are called voiced phones – examples are sounds like */a/, /i/* produced with vibrating vocal folds. The frequency of oscillation for the vocal folds is called the fundamental frequency or pitch ($F0$) of speech. Female voices are characterized with a higher pitch of about 165 – 225 Hz compared to the male which typically ranges from 85 – 155 Hz. However, the rise in $F0$ does not directly translate to the high intelligibility of female voices. On the other hand, there are articulations that are more random in nature with nearly uniform spectral information in 0 – 8 kHz, these are called unvoiced phones (sounds like */s/, /p/*).

Another important thing that is noticeable from Figure 2.1(a) is the presence of silence between phones, e.g., at around 0.75 seconds there is no speech content. These silent regions are generally referred to as the voice offset regions

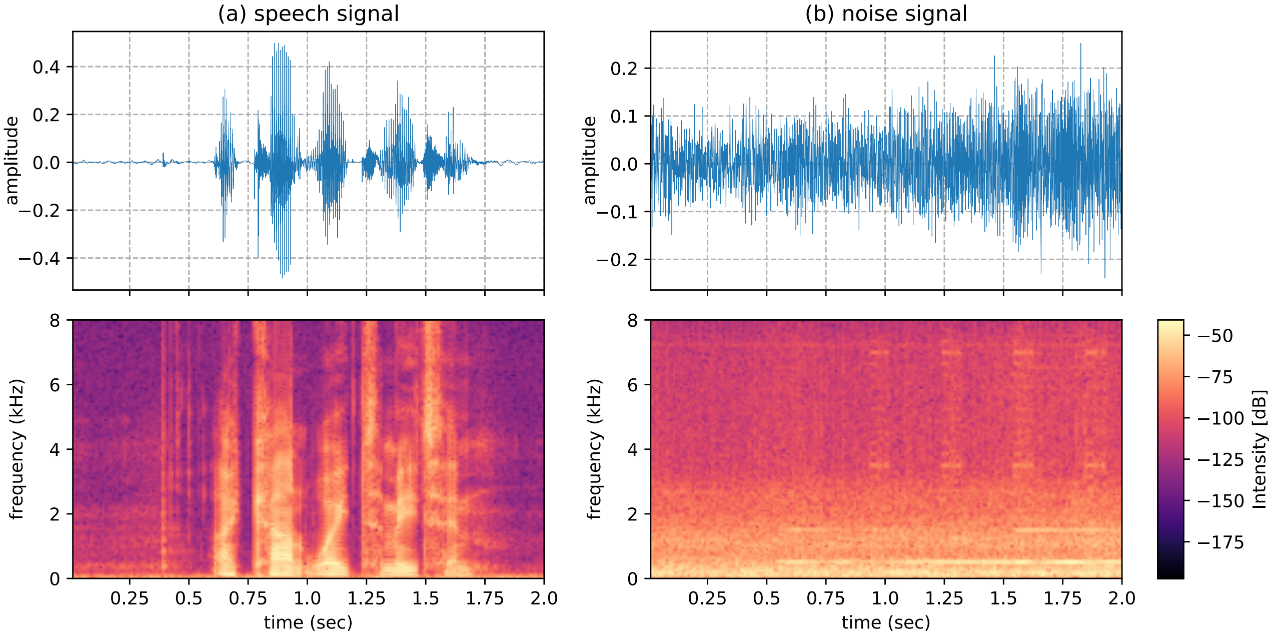


Figure 2.1 – A 16 kHz sampled speech and noise signals. The bottom panels shows the frequency content over time.

and they play an important role in analyzing the phonetic structure of speech. In the context of speech enhancement, those silent regions provide a glimpse of background events, therefore, can be marked to estimate the non-speech events in the recording. The resonance property of the vocal tract appears as the peaks of the spectrum at each time frame, generally known as the formant frequency in linguistic research.

Unlike speech, a generalization of background noise characteristics is not an easy task as it is produced by various sources with different acoustic properties. Noise sources can be broadly classified into two categories based on the nature of their frequency behavior over time; stationary and non-stationary noise. Stationary noise would have a fixed spectral content over time, e.g, wind blow, while the non-stationary noise is having a time-varying frequency characteristic, eg., the noise produced in a Cafeteria. The temporal variation can also happen in the intensity of noise while having a fixed spectral content, e.g, a vehicle running away from the observer. As an example, the noise produced by a running car recorded at 16 kHz sampling frequency is plotted in Figure 2.1(b). It is visible that the spectral and temporal characteristics largely differ from speech. Although the intensity of the noise increases over time, the spectral pattern remains largely stationary. This stationary nature of noise sources provides the opportunity to use the estimate of the background in speech offset regions as a good replacement for noise in the speech onset regions for speech enhancement models.

Much progress has been made in the speech enhancement domain over the past years. The initial approaches were based on signal estimation theories where the speech and/or noise processes were modeled with parametric models representing the individual process. However, in recent years, the use of neural networks has become prevalent for speech enhancement because of its ability to model complex noise processes faced in practice. An overview of progress in regards to speech enhancement is provided in the following sections.

2.1 Mathematical formulation of the problem

The objective of speech enhancement models is to recover the speech from its noisy observation. Although speech degradations can happen through non-linear ways as with the reverberation of voice in the speaking surroundings, the speech enhancement models target at removing the linear distractions on the speech. Given the noisy observations $y[t]$, the objective is to restore the clean components $s[t]$ which has been linearly degraded with noise $n[t]$ as

$$y[t] = s[t] + n[t]. \tag{2.2}$$

Robust modeling of speech and noise distributions is required to fully restore the speech from the mixture. Although multiple realizations of $y[t]$ can be collected with the use of a sensor array, discussions here are restricted to single sensor observation. Therefore, all the discussion following is on the single-channel speech enhancement.

2.2 Statistical approaches

Approaches based on the estimation framework have been the dominant solution to restore speech from the noisy mixture, which explores the statistical dissimilarity of speech and noise on the basis of mathematical estimation theorems. Although there is a wide category of approaches in the literature, the most popular of which are discussed in this chapter.

2.2.1 Spectral subtraction technique

The spectral subtractive algorithms, initially proposed by Boll[Bol79], is the historical and by far the simplest noise reduction model to implement. As the name indicates, it removes the background noise activities on a speech by subtracting an estimate of the noise from the noisy speech in a transformed domain such as in short-time Fourier transform (STFT)[Bol79] or Modulation domain [PWS10]. The noise spectrum estimation is done in the periods where the speech is absent. Such algorithms operate on the hypothesis that the noise statistics do not change drastically over time so that the noise estimated in speech absent frames can be a fair replacement for noise in speech active frames. Therefore, the subtraction should be done in a controlled manner to avoid speech distortions in the enhanced signal. If too much is subtracted, then some speech information will be lost, whereas if too little is subtracted, then much of the interfering noise remains.

We may consider the Fourier domain spectral subtraction method as an examples, in which the signal in Equation 2.2 is transformed with the short-term Fourier transform (STFT) algorithm with an analysis windows of size 20 – 30 ms as

$$Y(w,t) = X(w,t) + N(w,t) \tag{2.3}$$

The observed signal $Y(w,t)$ can be expressed in polar form as a combination of the magnitude and phase as $Y(w,t) = |Y(w,t)|e^{j\phi_y(w,t)}$, where $|Y(w,t)|$, $\phi_y(w,t)$ are the magnitude and phase spectrum of the noisy signal, respectively. Similarly the noise signal spectrum can be expressed as $N(w,t) = |N(w,t)|e^{j\phi_n(w,t)}$. Thus, there are two parameters to be estimated to model the noise process; the magnitude $|N(w)|$ and the phase $\phi_n(w,t)$. However, on the belief that the phase does not contribute much to the perception of speech as human ears are less sensitive to the phase distortions [OL81, OLKP79], the noise phase is substituted with the phase of noisy signal $\phi_n(w,t) = \phi_y(w,t)$ in spectral subtraction algorithms. Therefore, under the additive degradation of speech the clean spectrum can be restored by the subtraction of the noise spectrum from the noisy spectrum as

$$X(w,t) = \Big\{ |Y(w,t)| - |\hat{N}(w,t)| \Big\} e^{j\phi_y(w,t)} \tag{2.4}$$

Now we are left with only one unknown quantity to be estimated, the noise magnitude $|N(w,t)|$. Although many advanced approaches have been suggested recently to better estimate the noise magnitude [CCHM93, BK03, MM80], the basic concept is to utilize speech absent regions where the background activities can be cited. This requires effective detection of speech activity. We can observe in Figure 2.2 that there are regions in a noisy recording where speech is absent to estimate background events if identified properly. Since it is not practical to manually annotate the speech present and absent regions for the input in an enhancement system, automated detection of such regions are done with

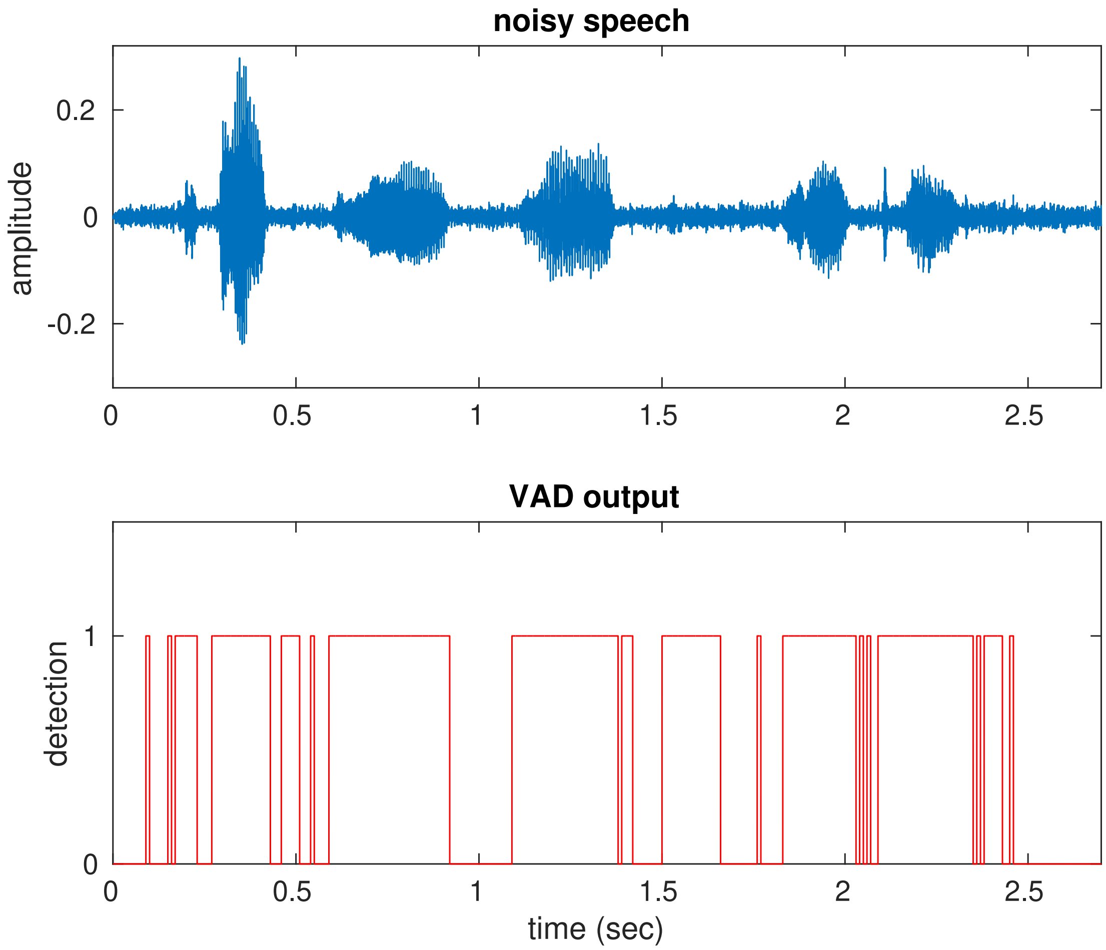


Figure 2.2 – Voice activities in a noisy recording is detected by a voice activity detector (VAD).

the use of voice activity detectors (VAD) in spectral subtraction algorithms which can predict the presence (VAD = 1) and absence (VAD = 0) of speech locally over time. VAD algorithms typically rely on features like short-term energy or zero-crossings of the signal, which in turn are compared against a threshold to make the decision at each time frame. The decision is updated in frames of 20 – 40 ms duration. An example of the predictions by an automatic VAD algorithm is plotted in Figure 2.2. The updating of noise estimate for frame indexed t ($|\hat{N}(w,t)|$) is done based on the VAD state as

$$|\hat{N}(w,t)| = \begin{cases} |Y(w,t)| & \text{if} \quad VAD(t) = 0 \\ |\hat{N}(w,t-1)| & \text{if} \quad VAD(t) = 1 \end{cases} \tag{2.5}$$

where the estimate from the previous frame is used in speech presence regions. This is a simple and direct estimation of noise magnitude spectrum, however, smoothing of the estimates in multiple past frames is used in modern spectral substation models. An important boundary condition to be met is that the magnitude of estimated spectral coefficients should be positive. Proper consideration must be given to prevent such cases in the subtraction in Equation 2.4 which may result from overestimation of the noise spectrum $|\hat{N}(w,t)|$. One simple and most common approach is to use a rectifier that clips any non-positive values in the estimated magnitudes to zero with

$$|X(w,t)| = \begin{cases} |Y(w,t)| - |\hat{N}(w,t)| & \text{if} \quad |Y(w,t)| > |\hat{N}(w,t)| \\ 0 & \text{else} \end{cases} \tag{2.6}$$

Spectral subtraction (SS) is a linear approach that works very well at high signal-to-noise ratio (SNR) conditions. However, as the input SNR reduces (or the noise level increases) most of the VAD algorithms faces problems in detecting the speech active events due to intense background noise on speech, therefore their performance diminishes. Besides, the estimation of noise in SS relied on the stationarity nature of the noise process. The replacement of noise in the current frame with the past frame may not be optimal in non-stationary noise conditions. Therefore, the performance of SS algorithms is reduced in non-stationary noises. To solve this issue, extended spectral subtraction (ESS) algorithms were presented in [Sov95, SPK96]. The key feature of ESS algorithms is that they can continuously update the background noise spectrum estimate, even during speech active instances. Among this category of algorithms, Wiener filtering is a popular one in which the Wiener filter coefficients are derived based on the noise power spectrum estimate [LO79, AEFDA+14]. For more variations of spectral subtractive algorithm, interested readers are referred to [Loi13].

2.2.2 Maximum likelihood estimator

Spectral subtraction was just an intuitive approach to the problem of speech enhancement that does not fully utilize the statistical properties of the involved signals. Whereas modeling of the distribution of speech and noise must be imperative for robust enhancement strategies. As such, the statistical model-based approaches for enhancement were developed as a more generalized mathematical formulation of the problem in the estimation framework. In such frameworks, given a set of noisy observations that are dependent on an unknown parameter (clean samples in SE), we wish to find a non-linear estimator of the parameter. There are different non-linear estimators proposed in the literature under the estimation theory framework such as *maximum likelihood* (ML) estimator, Bayesian *minimum mean square error* (MMSE) estimator and *maximum a posteriori* estimator [Kay93]. They mainly differ on the assumptions made on the unknown parameter to be estimated; whether it is a deterministic unknown or random variable.

The maximum likelihood (ML) approach [Whi82] is conceivably the most popular signal estimation approach in practice, and has been widely used even in most complex estimation problems [Myu03]. It was first used for speech enhancement by McAulay and Malpass [MM80]. Imagine that we observed N-points of noisy speech $y = \{y(0), y(1), ..., y(N-1)\}$ which correspond with an unknown parameter θ (the hidden clean speech). Also assume that we know the probability distribution function (pdf) of y that is conditioned on θ, denoted as $p(y;\theta)$. Since the pdf curve is parameterized by the unknown quantity θ, one must search for " *the θ that most likely produces the observed y*". Mathematically, this is posted as an optimization problem:

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(\mathbf{y};\theta) \tag{2.7}$$

The estimate $\hat{\theta}_{ML}$ is called the ML estimate of θ, and the function $p(y;\theta)$ is called *the maximum likelihood* function. The optimization is performed by differentiating the function with respect to θ and then finding the maxima. Often, it is a common practice to convert the likelihood function into logarithmic form for optimization convenience, and since the log transformation is a monotonically increasing function, it does not alter the point of maxima in the parameter space.

McAulay and Malpass [MM80] formulated speech enhancement problem in the spectral domain. Thus, the noisy mixture is represented in polar form as:

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + N_k e^{j\theta_n(k)} \tag{2.8}$$

where $\{Y_k, X_k, N_k\}$ are the the magnitudes and $\{\theta_y(k), \theta_x(k), \theta_n(k)\}$ are the phases of noisy, clean and noise segments, respectively. Since the signals are processed in digital domain, the frequency axis is discretised such that $w_k = \frac{2\pi k}{K}$, for $k = 1, 2, 3, ..., K$.

In the ML estimation framework, both the magnitude spectrum X_k and phase spectrum $\phi_x(k)$ are considered to be deterministic unknowns. Additionally, under the assumption that the noise is a Gaussian process with zero mean and $\lambda_n(k)/2$ variance, the likelihood function for observed signal $Y(w_k)$ can be derived as

$$p\left(Y\left(\omega_k\right);X_k,\theta_x(k)\right)=\frac{1}{\pi\lambda_n(k)}\exp\left[-\frac{\left|Y\left(\omega_k\right)-X_ke^{j\theta_x(k)}\right|^2}{\lambda_n(k)}\right]. \tag{2.9}$$

To get the ML estimator of X_k and $\theta_x(k)$, one must compute the maxima of the distribution $p\left(Y\left(\omega_k\right);X_k,\theta_x(k)\right)$. It is not a direct task as the function is with two independent variables. Under the general assumption that the phase will not contribute much into the overall quality of enhanced samples, the phase parameter ϕ is truncated by integrating the above function with respect to the phase axis with a uniform distribution $p\left(\theta_x(k)\right)=\frac{1}{2\pi}$ for $\theta_x(k)\in[0,2\pi]$

$$p_L\left(Y\left(\omega_k\right);X_k\right)=\int_0^{2\pi}p\left(Y\left(\omega_k\right);X_k,\theta_x(k)\right)p\left(\theta_x(k)\right)d\theta_x(k). \tag{2.10}$$

This integration simplifies the likelihood distribution to a conditional function conditioned only on the magnitude of clean spectra X_k as

$$p_L\left(Y\left(\omega_k\right);X_k\right)=\frac{1}{\pi\lambda_n(k)}\frac{1}{\sqrt{2\pi\frac{2X_kY_k}{\lambda_n(k)}}}\exp\left[-\frac{Y_k^2+X_k^2-2Y_kX_k}{\lambda_n(k)}\right]. \tag{2.11}$$

This is transformed to Logarithmic domain as the log-likelihood function ($\log p_L\left(Y\left(\omega_k\right);X_k\right)$). Then, differentiating it with respect to $\hat{X}_k$ and setting it to zero, we get the ML estimate of the clean magnitude spectrum

$$\hat{X}_k=\frac{1}{2}\left[Y_k+\sqrt{Y_k^2-\lambda_n(k)}\right]. \tag{2.12}$$

With the replacement of clean phase content with noisy phase as was in the case of spectral subtraction, the clean spectral coefficients are estimated as

$$\begin{aligned}\hat{X}\left(\omega_k\right)&=\hat{X}_ke^{j\theta_y(k)}=\hat{X}_k\frac{Y\left(\omega_k\right)}{Y_k}\\&=\left[\frac{1}{2}+\frac{1}{2}\sqrt{\frac{Y_k^2-\lambda_n(k)}{Y_k^2}}\right]Y\left(\omega_k\right).\end{aligned} \tag{2.13}$$

In the above formulation, the unknown parameter (X_k) to be estimated is assumed to be a deterministic unknown. However, speech production can not be considered as a deterministic process as the articulation of syllables varies based on the context of their appearance, besides, there will be speaker variability in the articulation of speech. Therefore, such variabilities must be accounted for in enhancement models. The Bayesian approach in statistics is known for the estimation of unknown quantities that are more random in nature.

2.2.3 Bayesian Estimator

In the case of Bayesian estimation [EW95], the unknown variable to be estimated is assumed as a random variable with a certain distribution. As such, the objective is to estimate the realization of that random variable. The approach is called Bayesian because its derivation is based on Bayes' theorem of probability [Joy03]. The incorporation of *prior* knowledge about the estimating variable θ into the estimation framework is the key difference from the ML approach. In applications such as speech enhancement, one must have prior knowledge about the range and distribution pattern to which the parameter θ belongs. With the inclusion of such knowledge, Bayesian estimators refine the estimation to be more voracious than the ML method.

Among different variations of Bayesian estimators [Mar02, LV05] the minimum mean square error (MMSE) estimator is the most commonly used in applications [EM85]. In the MMSE approach for speech enhancement, the general objective is to retrieve a clean speech spectrum from the noisy observations in the minimum mean square sense by minimizing the error function

$$e = E\left\{\left(\hat{X}_k - X_k\right)^2\right\}, \tag{2.14}$$

where $X_k, \hat{X}_k$ are the true and estimated spectral magnitudes, respectively. The symbol E denotes the expectation operation in probability theory. In Bayesian sense, this expectation is computed with reference to the joint *pdf* of the observed noisy process $\mathbf{Y}$ and the unknown parameter X_k represented as $p(\mathbf{Y}, X_k)$. Therefore, the Bayesian MMSE estimator is

$$\mathrm{BMSE}\left(\hat{X}_k\right) = \iint \left(X_k - \hat{X}_k\right)^2 p\left(\mathbf{Y}, X_k\right) d\mathbf{Y} dX_k. \tag{2.15}$$

Minimisation of the Bayesian MSE function $\mathrm{BMSE}(\hat{X}_k)$ with respect to X_k leads to the optimal MMSE estimator of the variable as given by [Kay93]

$$\begin{aligned}\hat{X}_k &= \int X_k p\left(X_k \mid \mathbf{Y}\right) dX_k \\ &= E\left[X_k \mid \mathbf{Y}\right] \\ &= E\left[X_k \mid Y\left(\omega_0\right) Y\left(\omega_1\right) \ldots Y\left(\omega_{N-1}\right)\right].\end{aligned} \tag{2.16}$$

Hence, the optimal estimator $\hat{X}_k$ is the mean of *a posteriori* pdf of X_k computed once the full spectrum of Y has been observed. It differs from the prior pdf $p(X_k)$ which was the distribution before observing the signal Y.

If one has the prior distributions of speech and noise spectra, the *a posteriori* distribution can be easily computed with the use of Bayes' theorem [Joy03]. However, the accurate computation of the distributions of Fourier spectra for speech and noise is a challenging task due to the non-stationarity of the signals. To resolve this issue, Ephraim and Malah [EM85] suggested a statistical model utilizing the asymptotic properties of the Fourier transform coefficients. The modeling was based on two statistical assumptions: 1) the Fourier transform coefficients have a Gaussian distribution with zero mean and time varying variances owing to the nonstationarity of speech, 2) the Fourier transform coefficients are uncorrelated, i.e., each coefficient is statistically independent from its neighbours.

As a result of which, the Bayesian MMSE estimator in Equation. 2.16 becomes

$$\begin{aligned}\hat{X}_k &= E\left[X_k \mid Y\left(\omega_k\right)\right] \\ &= \int_0^\infty x_k p\left(x_k \mid Y\left(\omega_k\right)\right) dx_k\end{aligned} \tag{2.17}$$

Under the assumption that the speech and noise are Gaussian processes with zero mean and variances λ_x and λ_n, respectively, in the spectral domain, the above integration will be reduced to

$$\hat{X}_k = \frac{\sqrt{\pi}}{2} \frac{\sqrt{v_k}}{\gamma_k} \exp\left(-\frac{v_k}{2}\right)\left[\left(1+v_k\right) I_o\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] Y_k, \tag{2.18}$$

where the variable v_k is defined by

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \tag{2.19}$$

with

$$\gamma_k = \frac{Y_k^2}{\lambda_n(k)}, \quad \xi_k = \frac{\lambda_x(k)}{\lambda_n(k)}. \tag{2.20}$$

The functions $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zeros and first order, respectively. The involved variables, ξ_k is called *a priori* SNR as it is the SNR estimate prior to the observation, and γ_k is called *a posteriori* SNR as it is the observed SNR.

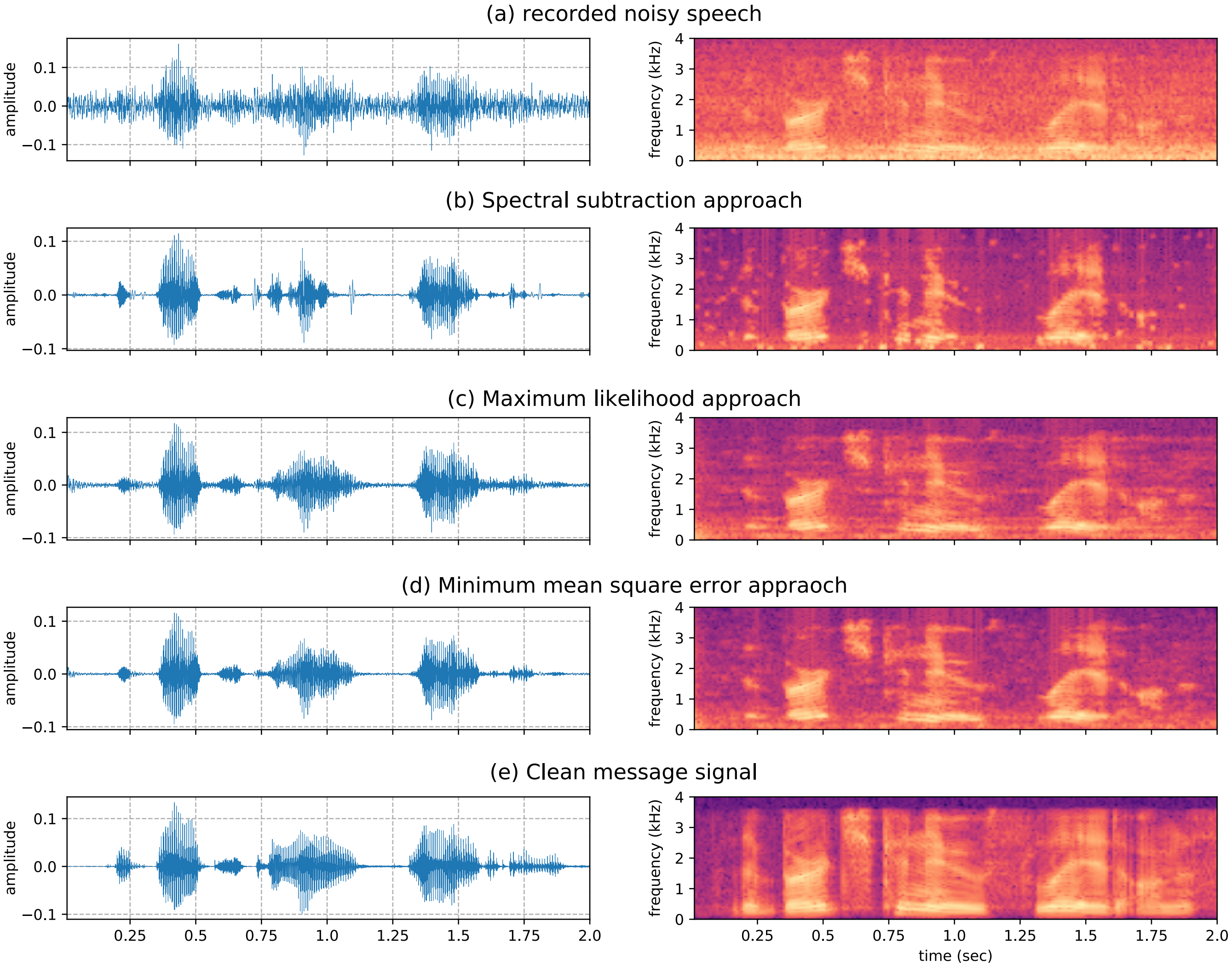


Figure 2.3 – Speech enhancement by various statistical approaches. Right panels display the spectral content over time computed by the short-time Fourier transform (STFT).

The enhancement performances of different algorithms discussed are illustrated in Figure 2.3 with a sample speech recording at 2.5 dB SNR. Few things are evident when we compare the enhanced signals to the clean speech. First, the spectral subtraction introduces large-scale distortions on the speech signal compared to the ML and MMSE approaches. Whereas, the maximum likelihood (ML) estimator reduced the distortions, while, the minimum mean square error (MMSE) approach yielded the closest to the clean enhancement.

Beyond these methods, there are more advanced statistical model-based algorithms for speech enhancement in the literature such as Kalman filters (KF) [PB87] which can better model the non-stationary nature of the involved signals with a dynamic state model. Another category of enhancement algorithms is the subspace-based enhancement [EVT95] that are derived by theorems of linear algebra. In the subspace approach, the noisy signal is considered to reside in a wider Euclidean space, while the clean speech is spanning only a subspace of the bigger space. Subsequently, given a method for decomposing the vector space of noisy observations as clean and noise subspaces, we could easily compute the clean speech by nulling the components of noise attributes. The decomposition of noisy space can be done with the concepts from linear algebra, like the singular value decomposition (SVD) [WRR03], or eigenvalue decomposition [PS03].

The aforementioned algorithms have demonstrated good at improving speech quality and were found to increase listeners' preference, but failed in improving speech intelligibility. However, a new class of algorithms motivated by the study on auditory scene analysis (ASA) [BC94] was presented for noise reduction, which has been demonstrated successful at improving speech intelligibility [KLHL09]. These algorithms estimate a binary mask in the time-frequency domain deciding on which spectral bins to keep and which is to remove for the enhancement, therefore, called binary mask algorithms. The estimation of the mask requires prior knowledge about the involved speech and noise signals, which makes them difficult to be used in practice. For more on signal processing for speech enhancement, interesting readers are referred to [Loi13].

2.3 Conclusions

In this chapter, an introduction to speech enhancement objectives and a few most popular statistical methods for enhancement were discussed; spectral subtraction (SS), Maximum likelihood (ML) estimator, Bayesian minimum mean square (BMMSE) estimator. SE is the simplest approach where we estimate the noise spectrum and subtract it from the noisy spectrum to restore the clean speech. Although the SS method provides an improvement in terms of noise attenuation, it often produces a new randomly fluctuating noise, referred to as musical noise due to its tone-like characteristics in the spectrum as shown in Figure 2.3(b). This is originated from high variance in the noise estimation, which becomes more prominent in non-stationary noise conditions. Whereas, ML and BMMSE approaches produce more stable enhancement of speech due to their more statistical oriented formulation of the problem. Although both of them make the same assumption on the noise process – Gaussian process with zero mean and fixed variance, the BMMSE treats the speech as a random process with a Gaussian distribution while the ML presume it as a deterministic unknown. This has led to a better restoration of speech by the BMMSE method. However, such assumptions about speech and noise would not be a good representation of many real-world scenarios, therefore, would result in a degraded performance in deployment. In the next chapter, we will discuss the advantage of neural networks in this regard to learn the distribution of speech and noise from data.

Chapter 3

Neural Networks for Speech Enhancement in Feature Domain

The statistical approaches for speech enhancement involved estimation of critical quantities such as the noise spectrum in spectral subtraction, or the *a priori* SNR in the MMSE method. Besides, these models are derived under various assumptions on the production of speech and noise, like the zero-mean Gaussian assumption in the Bayesian estimator. However, these assumptions can not be fulfilled in practice because of the diversity of noise conditions encountered and the nonlinearities in speech production. Therefore, models that can detect high-level statistical patterns of the signals are required to generalize well to practical scenarios. Neural networks have been demonstrated efficient at detection and classification of objects in images or video sequences, achieving close to human performance [SZ14, KSH17, GWK+18]. Thanks to its parametric non-linear modules, which made it possible to isolate complex patterns in the data.

With the advent of deep learning, signs of progress have been made in the speech enhancement domain. Different varieties of architectures are presented in the literature with the objective of finding the underlying patterns connecting noisy speech to the clean. Two broad classifications of such methods are possible; one based on mask estimation and the second based on mapping estimation. The mask estimation approaches estimate the binary mask in the time-frequency domain of speech – it can be seen as a filter, then be applied onto the noisy signal to retrieve the speech [WNW14]. Whereas, the mapping estimation approaches directly estimate the mapping function representing noisy observations to the clean speech in either time-frequency (TF) domain or in the time domain without any intermediate stage. This thesis focuses on mapping-based speech approaches. As such, we first discuss the TF domain enhancement models available in the literature as well as improvements suggested as part of this thesis, while the time domain models are presented in the next chapter.

3.1 Enhancement framework

Speech is redundant with information in its raw form. As such, neural processing of the raw samples without any transformations to lower dimension would require extra computational as well as time (latency) cost, therefore, may not be desirable for many real-world enhancement applications like hearing aids. Therefore, it has been a common practice to manually transform the signal to a lower two-dimensional (2D) time-frequency representation with the use of transformations such as the short-time Fourier transform (STFT) where the enhancement task is then formulated. For a speech signal $x[n]$, the STFT is computed as

$$\mathrm{STFT}\{x[n]\} \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{3.1}$$

where $w[n]$ is a window that defines the time resolution of the transformation. As m varies from the beginning to end the window selects different regions of the signal. The variable ω denotes the continuous frequency axis. However, in the context of digital signal processing, the frequency axis is sampled at selected points that are equally separated. Therefore, the 2D representation would be a compressed representation of the raw waveform. Subsequently, this 2D representation is presented to the network to learn the noise reduction in the STFT domain. However, since the phase does not contribute much to the perceptual quality of speech, it has become a common practice to ignore the phase information while processing only the magnitude spectrum, as illustrated in Figure 3.1.

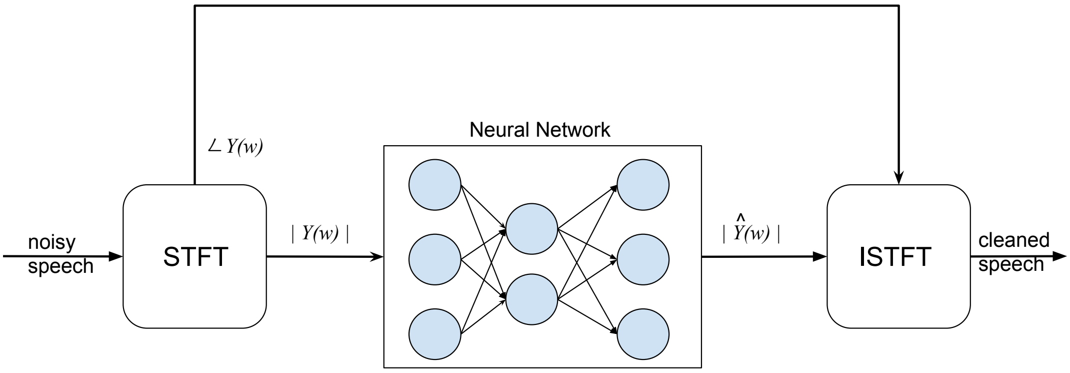


Figure 3.1 – Schematic of speech enhancement in feature domain.

The input to the neural network will be noisy magnitude spectrum, while the target is the corresponding clean magnitude spectrum. The noisy phase is used for reconstruction of the clean samples once the enhanced magnitude spectrum has been predicted. The reconstruction of speech waveform from the combined STFT coefficients is done with the use of Overlap and Add (OLA) method [RG75].

Different neural architectures have been presented in the literature to date in the above framework. The initial neural network architectures considered were fully connected networks (FNNs) [XDDL14, LTMH13]. FNNs predict an output frame from the corresponding input frame or from a small window of frames around it. In tasks that require long receptive fields, such as separating a target speaker from babble noise, their performance drops [TW18]. Alternative architectures that are capable to model time dependencies efficiently include convolutional neural network (CNN) layers and/or recurrent neural network (RNN) layers. A CNN layer captures the local dependencies and a network of CNN layers can capture longer dependencies. As a result, CNNs perform better than FNNs [PL16]. Also, CNNs are more memory efficient than FNNs due to their weight-sharing property. When applied on a 2D spectrum, convolutional layers with a two-dimensional kernel can detect local patterns in the input. The 2D transformation produced by a convolutional layer of CNN speech enhancement (CNN-SE) network is depicted in Figure 3.2. As visible in the figure, the kernel strides over the input detecting local activities in the spectrum. Deep convolutional enhancement networks are built by stacking multiple such 2D convolution layers.

Different variations of CNN-SE are seen in the literature. One popular approach is the convolutional encoder-decoder (CED) architecture [VLL$^+$10]. The CED has an encoder module comprised of CNN layers which compresses the input 2D spectrum into a lower dimension representation from which the original dimension is reconstructed by the decoder module, which usually involves upsampling layers. The encoder in the CED architecture can be seen as a dimensionality reduction network, like the principal component analysis (PCA) in statistical approaches [WRR03], where the representation of speech and noise would be more separable. Besides, expansion of encoder feature representations with the use of more number of kernels towards the depth of encoder architecture was also found to contribute for improved enhancement [PL16]. Such a large number of filters would ensure redundant representation of the input spectrum as multitudes of resolution in the depth of the network.

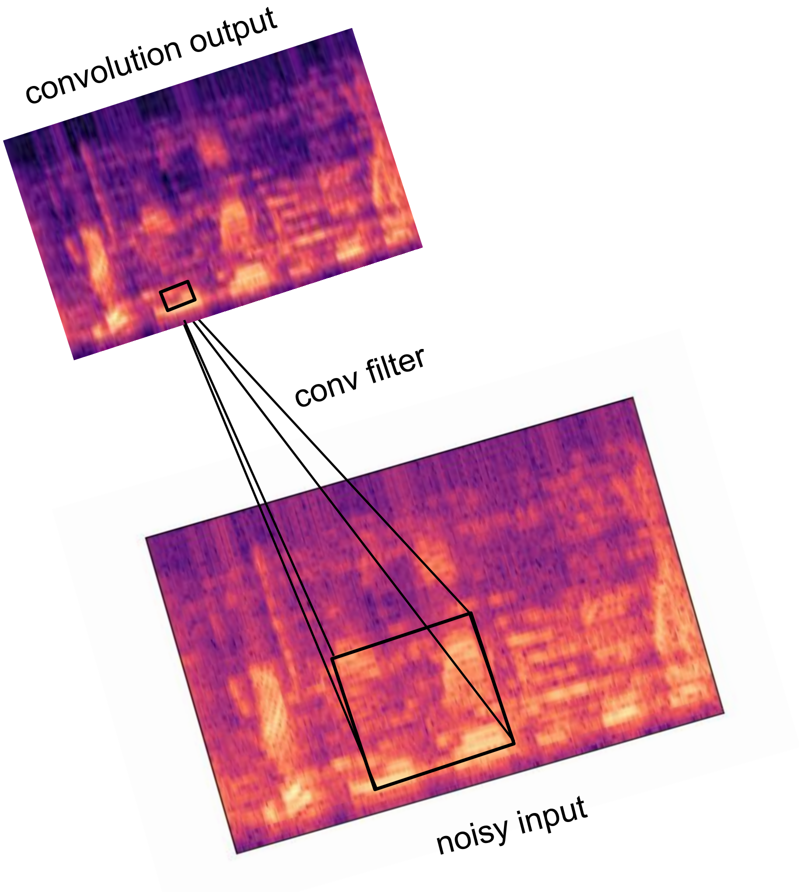


Figure 3.2 – Transformation in a 2D convolution layer of CNN-SE.

However, as the depth of CNNs grows, the number of their parameters also grows and this limits their applicability in low-end and embedded devices. On the other hand, RNNs are capable to model long dependencies with only one or a few stacking layers. Weninger et al. [WEW+15] used LSTMs recurrent neural networks to model the long-term dependencies in the frames. Networks that combine CNN and RNN layers (CNN_RNN-SE) were also considered recently [ZZTL18], [NBP+17]. In these networks, the CNN layers specialize in feature extraction and the RNN layers in modeling the longer dependencies among frames. The CNN_RNN-SE architectures have been shown to exhibit better generalization of speakers and noise in unseen test conditions [TW18] because of the fact that the RNN module can constructively integrate information among the frames, like formant transitions carrying the speaker information. A typical CNN_RNN-SE architecture is shown in Figure 3.3 with LSTM as the RNN module. The LSTM layer placed at the top of the architecture models the temporal dependency among the frames on the feature maps produced by the final convolution layer. However, due to the repeated applications of convolution kernels on the input, the time-frequency resolution of the input spectrum would have been damaged. Therefore, the LSTM layer placed at the top could only see an abstract representation of the input noisy spectrum. Hence, it is not an optimal modeling of the temporal dependency among successive frames. To this end, we presented in [SCS+20] a new approach that combined the feature extraction and dependency modeling into a single unified framework, in which features are being extracted recurrently over time. This modeling is discussed next.

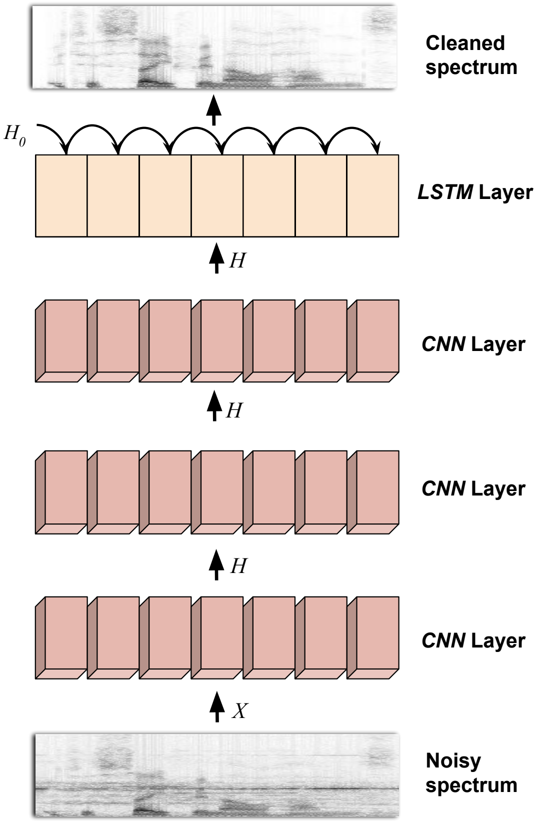


Figure 3.3 – Convolutional recurrent speech enhancement network (CNN_RNN-SE) with LSTM as the RNN module [ZZTL18], [NBP+17].

3.2 A recurrent feature extraction for speech enhancement

The temporal dependency modeling with LSTM or GRU layers have been used in the past enhancement models. However, such models with CNN layers as the front-end smooth the input spectral information being presented to the recurrent layer towards the end of the model architecture. Therefore, limits their visibility to model finer – frequency localized– temporal dependencies in the spectrum. Convolutional LSTM [SCW+15] or GRU [BYPC15] have been proposed to better preserve the spatial resolution of a 2D representation while modelling the temporal dependency. Similarly, TS Hartmann, in [Har18], added recurrent connections into the CNN layers, whereby has improved the accuracy of image classification in noise. The new feature extraction module was then called gruCNN.

Investigating the prospect of gruCNN neural module to model speech, we introduce a new feature extraction framework for speech enhancement – where the features are being extracted recurrently over time with the reinforcement of context-awareness. With the inclusion of the recurrency factor into the feature extracting layers, the proposed gruCNN enhancement model (gruCNN_FC–SE) learns to extract features that are maximally relevant in each time instance. In contrast to the CNN models with an architecture presented in Figure 3.3, gruCNN_FC–SE model is robust of having refined features in the layers of the network, while at the same time reducing the parameter complexity considerably. This is because of the fact that the two-stage modeling with initial convolution layers followed by a fully connected layer does not give attention to the local recurrency patterns in the input spectrum, leading to the lack of qualitative features at the front-end – as elaborated in detail below.

gruCNN_FC–SE defines the speech enhancement on the STFT domain of the signal. Let X_k be the slice of k^{th} frequency bin values over time, from the noisy input spectrum X, such that $X_k = [x_1, ..., x_{T-1}, x_T]$; where T is the total number of frames considered. Then, the probability of X_k to happen can be expressed as

$$p(X_k) = p(x_1, ..., x_{T-1}, x_T) \tag{3.2}$$

with the product rule of probability, the joint distribution can be redefined as the product of individual probabilities:

$$p(X_k) = \prod_{t=1}^{T} p(x_t / x_{t-1}, ..., x_{t-T}) \tag{3.3}$$

Preserving this statistical structure is essential when designing speech enhancement models to ensure the auto-regressive nature of predictions. Moreover, the quality of enhancement will be determined by how accurately this dependency is being modelled. Though there may have been some inter-bin dependencies between the spectral bins within a frame, as k varies from 1 to K (the final bin), present modelling has not considered that for it may be trivial compared to the temporal dependency. With this decomposition, only the past dependencies are considered for the model to be causal.

In conventional speech enhancement neural models [ZZTL18][NBP+17] with the architecture displayed in Figure 3.3, the temporal recurrency of speech was modelled by fully connected recurrent neural network (FC-RNN) layers such as LSTM [HS97] or GRU [CGCB14] cells, employed towards the end of model architecture. Therefore, the front-end feature extraction with CNN layers and the back-end recurrency modelling with FC-RNN operate independently. Such a modeling, without counting recurrency factor at the feature extraction level, leads to the lack of qualitative features at front-end. Further, due to the inherent fully connected nature of FC-RNN, the bin-wise recurrency factor described in (3.3) has been ignored.

In contrast, the gruCNN layers are designed to model the local recursion over time with the use of convolution kernels of fixed dimension to trace the local statistics of previous frame to be integrated into the current feature estimation. At a given frame index t, the new feature extraction layer (gruCNN) has inputs the previous layer output $\boldsymbol{x}_t$ – which is the noisy speech spectrum at the beginning layer – and the feature status of the previous frame ($\boldsymbol{h}_{t-1}$). This is being processed through the nonlinear transformations in (3.4) – (3.7) to get the feature representation of the current frame ($\boldsymbol{h}_t$). Whereby, the feature map $\boldsymbol{h}_t$ encodes information from the current frame together with the past context.

$$\boldsymbol{z}_t = \sigma(W_{zh} * \boldsymbol{h}_{t-1} + W_{zx} * \boldsymbol{x}_t) \tag{3.4}$$

$$\boldsymbol{r}_t = \sigma(W_{rh} * \boldsymbol{h}_{t-1} + W_{rx} * \boldsymbol{x}_t) \tag{3.5}$$

$$\hat{\boldsymbol{h}}_t = \tanh(W_{hh} * (\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + W_{hx} * \boldsymbol{x}_t) \tag{3.6}$$

$$\boldsymbol{h}_t = \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{z}_t) \odot \hat{\boldsymbol{h}}_t \tag{3.7}$$

where the operations $*$ and $\odot$ indicate convolution and element-wise matrix multiplication, respectively. While training in this setting, the network will learn the optimal kernels (W_{zh}, W_{zx}, W_{rh}, W_{rx}, W_{hh} and W_{hx}) that maximize the local bins recurrency, whereby ensure the best features at the layers. It is worth to note that unlike fully connected RNN cells [CVMG+14], [HS97] that use matrix operation to model the long-term context, the gruCNN has kernel coefficients that are shared, which in turn reduces the parameter complexity.

By layering a set of gruCNN modules one after another, the gruCNN_FC–SE network has the final structure shown in Figure 3.4. At the end of model architecture, it is a time distributed fully connected layer which regresses the recurrently extracted features into the enhanced spectral bins. These predictions are combined with the noisy phase information to reconstruct back the enhanced speech samples.

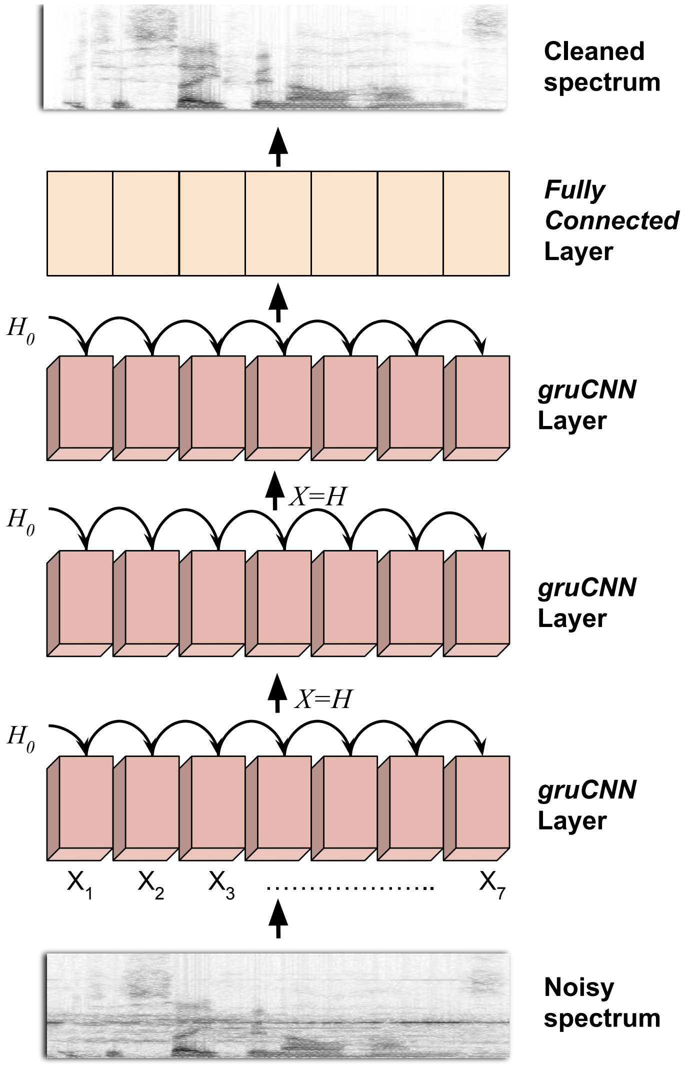


Figure 3.4 – gruCNN_SE: Speech enhancement with gruCNN as layers.

3.2.1 Model configuration and database selection

As the primary focus is on evaluating the efficacy of suggested recurrent feature extraction strategy over the conventional CNN architecture, the comparing models should have had the same structural setting. To this purpose, a model without any recurrent connections in the feature extracting CNN layers is considered (CNN_FC–SE). Since it does not incorporate any form of temporal recurrency at all in its modeling, the architecture is similar to Figure 3.3, but with the replacement of the LSTM with a fully connected (FC) layer. Secondly, to quantify the benefits of recurrency modelled precisely at the feature extraction stage, a model rather having the front-end CNN layers followed by the standard fully connected LSTM cell [GSC99] (CNN_LSTM–SE), having the exact structure shown in Figure 3.3, is implemented. The LSTM cell was selected instead of GRU for they have shown better enhancement, as have been reported in the past studies [ZZTL18][NBP$^+$17].

All the models considered have six convolutional layers (recurrent/casual) followed by the final fully connected (recurrent/casual) layer. The convolutional kernels of each layer are set to be of [3 × 3] size. The filter size was selected to be of basic for swiftly isolate the performance gain by different models. Each layer of the models has had channel depth of 256 with Parametric ReLU (PReLU) activation. Further details about the individual layers are highlighted in Table 3.1, for an input tensor of shape $[1, 161, 128, 1]$.

Table 3.1 – Layer-wise description of different models

Layer	CNN_FC-SE	CNN_LSTM-SE	gruCNN_FC-SE	Output shape
1	[3 × 3] CNN	[3 × 3] CNN	[3 × 3] gruCNN	[1, 161, 128, 256]
2	[3 × 3] CNN	[3 × 3] CNN	[3 × 3] gruCNN	[1, 161, 128, 256]
3	[2 × 1] Maxpool	[2 × 1] Maxpool	[2 × 1] Maxpool	[1, 81, 128, 256]
4	[3 × 3] CNN	[3 × 3] CNN	[3 × 3] gruCNN	[1, 81, 128, 256]
5	[3 × 3] CNN	[3 × 3] CNN	[3 × 3] gruCNN	[1, 81, 128, 256]
6	[2 × 1] Maxpool	[2 × 1] Maxpool	[2 × 1] Maxpool	[1, 41, 128, 256]
7	[3 × 3] CNN	[3 × 3] CNN	[3 × 3] gruCNN	[1, 41, 128, 256]
8	[3 × 3] CNN	[3 × 3] CNN	[3 × 3] gruCNN	[1, 41, 128, 256]
9	FC	LSTM + FC	FC	[1, 161, 128, 1]

Data Set (Training and Testing): The speech set is a selection of ten British English speakers – both male and female – from the Voice Bank speech corpus [VYK13], each of which has around 400 clean utterances. Eight speaker's data were used for training, and the remaining two (one male and one female) were reserved for performance testing. The noisy mixtures were created manually. The noises are from [Loi13], which contains 20 different types of common environmental noises. Fourteen of which were used for the training, and the remaining six were used as the unseen noises, under which the models are tested. For training set mixtures, each speech sample was masked by a random training set noise at a random SNR point from [0, 5, 10, 15, 20] dB. A similar process has been followed for the test set, but with the unseen noises at unseen SNR points of [2.5, 12.5, 22.5] dB.

Although the original speech was sampled at 48 kHz, it was down-sampled to 16 kHz for our experiment as in [MSATS19] [RPS18]. The 16 kHz sampled signals were framed into 20 ms frames with 10 ms overlap. The frames were Fourier transformed into 320 points. The log-power spectra feature is the domain on which the enhancement task is modeled [Por80]. Therefore, the frequency dimension of input spectrum is halved to 161 points, due to the spectral symmetry.

Model Training: All the comparing models are trained in an end-to-end mode, where the losses are computed directly between the magnitudes of predicted ($\hat{Y}(k,t)$) and target ($Y(k,t)$) STFT components. For each noisy-clean training set pair (X,Y), the model parameters are optimized by minimizing the mean square error (MSE) objective function

$$L_{X,Y} = \frac{1}{T \times K} \sum_{t=1,f=1}^{t=T,f=K} (|Y(k,t)| - |\hat{Y}(k,t)|)^2, \tag{3.8}$$

where K denotes the dimension of frequency axis that is 161, and the variable T is the number of time frames recurrently generated in the training process; which has been set to T = 128. The T value for testing varies based on the input signal duration for the recurrency is being modeled over the temporal axis. The loss was minimized by the Adam optimizer [KB14] with an exponentially decaying learning rate method with learning rate = 0.001, decay steps = 20,000 and decay rate = 0.99.

Procedure: Typically, speech processing models have to be evaluated both in terms of quality and intelligibility of enhanced samples, as both the factors play its' role in understanding speech. Particularly, in the case of speech enhancement, the masking of noise could seriously inflict the intelligibility and perceived quality of the signals. As such, we had to have multiple objective metrics to objectively measure quality and intelligibility.

For the quality assessment, the perceptual evaluation of speech quality (PESQ) [HL07] metric based on the ITU P.862.2 standard in a wide-band setting is used. The PESQ ranges between 1.02 and 4.56, as higher the value better the sample quality. Secondly, a set of composite quality metrics that have shown high correlations with mean opinion score (MOS) are chosen [Loi13]; 1) composite measure for signal distortion (CSIG) which predicts the MOS by speech distortion independently from background noise. 2) composite measure for background noise (CBAK), which predicts the MOS by background noise intrusiveness, solitarily. 3) composite measure for overall sample quality (COVL) which predicts the joint MOS by counting speech restoration and noise intrusiveness. The MOS predictions vary between 1 to 5, with higher the number better the samples upon each category.

For the intelligibility assessment of enhanced samples, the short-time objective intelligibility (STOI) [THHJ11] metric is employed, as it has proved to be highly correlated to the subjective intelligibility scores. The STOI ranges in 0 to 1, with higher values representing more intelligible speech. Ultimately, the most direct way to measure the intelligibility gain is with Signal to Noise Ratio (SNR), which directly computes the ration of the speech power to the noise power. The SNR computed on the segmental level – Segmental SNR (SSNR) metric – [Loi13] is used as the second intelligibility metric. SSNR is measured in the logarithmic domain, where the range is of any real value.

After all the model has to be validated with the listeners as they are the final beneficiary. A wide-scale subjective evaluation of the quality of enhanced samples has been conducted. The mean opinion score (MOS) evaluation was with normal hearing native English listeners. Where, the listeners were asked to rate the quality of the samples based on the noise artifacts present, on a scale of 1-5 (0 – very annoying artifacts, 5 – no artifacts at all). In total, 20 participants had participated in the MOS evaluation study.

3.2.2 Observations and Discussion

The mean objective scores on 220 test samples at each noise condition are displayed in Table 3.2. Along with the processing types, the scores of unprocessed noisy speech have also been included to better understand the relative gain. Compared to the CNN_FC–SE architecture, which does not incorporate any form of recurrency described in Equations (3.2) to (3.3), the suggested gruCNN_FC–SE model with recurrency modelled in the feature extraction layers, has distinctly outperformed in all the metrics. This gain is almost consistent across the noise conditions. With the inclusion of global recurrency, the performance of CNN_LSTM–SE has improved over CNN_FC–SE. This broadly conveys the benefits that can be achieved through temporal inclusive modeling of speech.

When comparing the two recurrent models, the proposed gruCNN_FC–SE, that is concerned of the bin-wise recurreny factor, has shown better enhancement over CNN_LSTM–SE. Even at the higher SNR point of 22.5dB, where the noise attributes are expected to be mild, gruCNN_FC–SE model elicited noticeable enhancement, showing an SSNR intelligibility gain of up to 1.5 dB over the other methods. This gain must be attributed to the qualitative restoration of speech components with the suggested feature extraction strategy.

Regarding the consistency of model predictions in different noise types, the model enhancements under the two unseen noise conditions are plotted in Figure 3.5. The upper panel shows construction noise (type–1) while the lower panel refers to street noise (type–2). Since type–1 noise is quite stationary and has the spectral energy that is distributed uniformly in a very wide frequency band (0 - 3 kHz), it is straightforward for a network to get a frequency smoothed estimate of the noise statistics. While type–2 noise (street) are highly localized at the lower band (0 - 0.5 kHz) of the spectrum (marked by a straight line in Figure 3.5). Unless the model looks into the local statistics of the spectrum, these noise activities could easily be miss-classified as speech events. We suggest that this explains the performance of CNN_FC–SE and CNN_LSTM–SE, whereas gruCNN_FC–SE seems to be successful in disentangling out the noise activities by exploiting the local patterns.

The subjective scores of different models are displayed in Table 3.3. In line with the objective measures, the suggested gruCNN_FC–SE model is ranked closer to the clean speech with a score of 3.16 on the five point scale, while there was not any statistically observable difference between the scores of the other two methods.

Pragmatically, the performance gain of a neural model could be argued by the additional parameters that are floated into the modeling. To this end, the parameter counts of different models are shown in Table 3.4. CNN_FC–SE is the least complex among the models and indeed its performance has been much lower than the other two models. On the other

Table 3.2 – Objective measures enumerating the performance

Noise level	Metric	Noisy	CNN_FC-SE	CNN_LSTM-SE	gruCNN_FC-SE
2.5 dB	PESQ	1.20	1.41	1.51	**1.57**
	STOI	0.68	0.71	0.72	**0.74**
	COVL	1.58	1.96	2.15	**2.22**
	SSNR	- 3.63	2.39	3.20	**3.94**
12.5 dB	PESQ	1.49	1.87	2.01	**2.08**
	STOI	0.77	0.78	0.79	**0.80**
	COVL	2.11	2.59	2.74	**2.83**
	SSNR	3.24	7.61	7.85	**8.96**
22.5 dB	PESQ	2.27	2.47	2.58	**2.66**
	STOI	0.85	0.83	0.84	**0.85**
	COVL	3.05	3.20	3.30	**3.41**
	SSNR	12.26	11.21	11.14	**12.83**

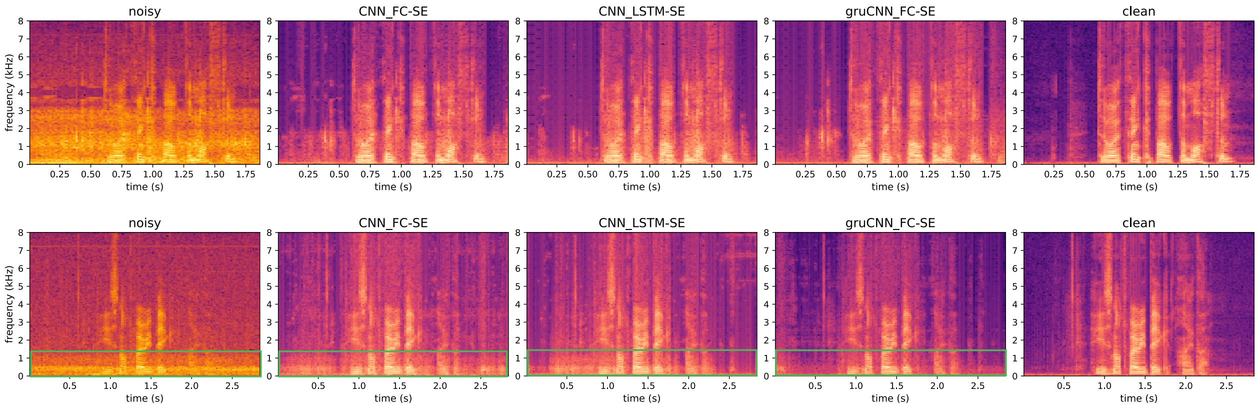


Figure 3.5 – Model enhancement under construction (upper panel) and street (lower panel) noise

Table 3.3 – Mean opinion score (MOS) with standard error

Metric	**Noisy**	**CNN_FC-SE**	**CNN_LSTM-SE**	**gruCNN_FC-SE**	**Clean**
MOS	2.01±0.97	2.75±0.92	2.77±0.89	**3.16±0.92**	4.86±0.42

Table 3.4 – Model parameters count in Million (M)

Metric	**CNN_FC-SE**	**CNN_LSTM-SE**	**gruCNN_FC-SE**
Parameters	11.13M	36.10M	27.22M

hand, the suggested gruCNN_FC–SE produces far better enhancement with only 75% parameters of the CNN_LSTM–SE. This reduction in parameter complexity is from the replacement of the fully-connected LSTM cell with the fixed kernels of gruCNN to model the temporal flow. All of which indicates the potentiality to have it implemented on computationally constrained applications, like hearing aids. A Tensorflow implementation and enhanced samples from the model are provided at [1] [2].

3.3 A bidirectional recurrent feature extraction technique

The usefulness of the gruCNN layer was explored for speech enhancement in the above section for modeling the forward dependencies among 2D spectral frames. However, in applications where the processing can be done offline using the whole sequence of spectrograms, it would be beneficial to include the future context into the modeling to further reinforce the context-awareness of the enhancement model. Motivated from this fact, we introduce the bidirectional extension of gruCNN, called BigruCNN cell, that can model dependencies both in the forward and backward directions of time. To test the efficacy of the suggested BigruCNN layer, a deep enhancement network with BigruCNN layers (BigruCNN-SE) was implemented. BigruCNN-SE avoids the initial CNN layers and relies only on BigruCNN layers for the feature extraction. Besides, as in the gruCNN-SE model, it is expected that this architecture performs well in noisy speech enhancement over the convolutional recurrent network, because it integrates the relevant features over time before the signal is being too abstract. Apart from its ability to extract information from noisy data, the BigruCNN-SE has a relatively small number of parameters due to shared weights in the GRU kernels as in the case of gruCNN, which is highly desirable for practical applications.

Extensive evaluations of the BigruCNN-SE have been conducted both in seen and unseen noise conditions on a multi-speaker data set. In objective evaluation, the suggested BigruCNN-SE model shows better generalization to unseen conditions, with the extended short-time objective intelligibility (ESTOI)[JT16] scores of 2.5 to 5.0% points higher than the traditional CNN based model. Compared to data-independent statistical enhancement models, the performances of both CNN and BigruCNN models slack in unseen noise cases relative to the seen condition. However, the relative performance gain of BigruCNN-SE over the CNN network was retained even at the lowest *SNR*. The subjective evaluation of the perceptual quality yielded a relative Mean Opinion Score (MOS) improvements of 0.2 and 0.4 at 0 dB and 5 dB *SNR*s, respectively. Equally important is the reduction in the number of parameters, where the BigruCNN-SE model has had only 19% parameters than that of the CNN architecture. The mathematical modeling that leads to BigruCNN cell and the BigruCNN-SE architecture is discussed next.

[1] https://www.csd.uoc.gr/~shifaspv/IEEE_Letter-demo

[2] https://github.com/shifaspv/gruCNN-speech-enhancement-tensorflow

3.3.1 Bidirectional gruCNN (BigruCNN) module

Our analysis is based on speech data, but it can be applied to other one-dimensional time series. Let $X = |STFT(s)|^2 \in \mathbb{R}^{N \times T}$ be the squared magnitude of the short-time Fourier transform (STFT) of the signal $s(t)$, where N is the number of frequency bins and T is the total number of frames. Also, let $X_k = X[k, :]$ be the slice of k^{th} frequency bin. To simplify the notation the index k will be omitted from the frames and we write vector X_k as $X_k = [x_1, ..., x_t, ..., x_T]$.

As such, the joint probability of the vector X_k can be expressed as

$$p(X_k) = p([x_1, ..., x_t, ..., x_T]). \tag{3.9}$$

Using the chain rule of probability, $p(X_k)$, can be decomposed as

$$p(X_k) = \prod_{i=1}^{T} P(x_i | x_{<I}), \tag{3.10}$$

where $x_{<i} = [x_1, \ldots, x_{(i-1)}]$. This is the forward decomposition.

Alternatively, $p(X_k)$, can be expressed as

$$p(X_k) = \prod_{i=1}^{T} p(x_i | x_{>I}), \tag{3.11}$$

where $x_{>i} = [x_{(i+1)}, \ldots, x_T]$. This is the backward decomposition.

The probability $p(x_i | x_{<i})$ is conditioned on the past samples and can be modelled with a forward RNN, while the probability $p(x_i | x_{>i})$ is conditioned on the future samples and can be modelled with a backward RNN.

In many speech processing applications, including speech enhancement, future samples are generally available. Therefore, in order to make more informed predictions the joint conditional probability $p(x_i | x_{<i}, x_{>i})$ is modelled with a bidirectional recurrent network. The above formulation is a strong model of speech auto-regression, for it has taken into account the frequency modulation over time.

Although different recurrent cells like bidirectional LSTM [HXY15] and GRU [CGCB14] are commonly used to model the recursion patterns in a 2D spectrum, these RNN cells ignore the spatial resolution over the frequency axis due to the inherent fully connected nature of the cells. To this end, we now introduce the proposed Bidirectional gruCNN (BigruCNN) neural module which can trace the local spectral statistics.

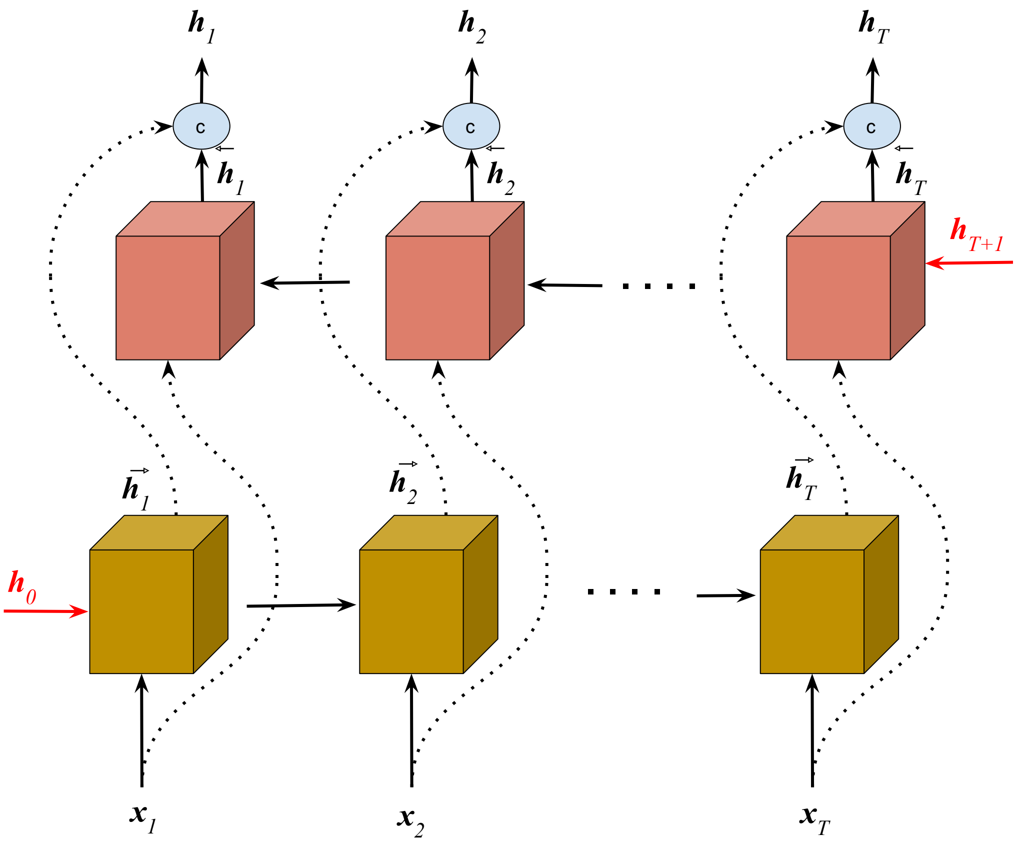
BigruCNN module has internally a set of kernels with fixed dimension that model the forward, $p(x_t | x_{<t})$, and the backward, $p(x_t | x_{>t})$, local dependencies. The kernels search for local features in the past and future frames to be integrated into the current feature estimation.

The forward and backward information streams are modelled separately with two sets of kernels, as depicted in Figure 3.6. Information propagates forward in the bottom sub-module, while the direction is reversed in the top sub-module. The input series $\{\boldsymbol{x}_t\}_{t=1}^{T}$ is a common input to both sub-modules, while the second input is sub-module specific. This is the past, $\boldsymbol{h}_{t-1}$, and future, $\boldsymbol{h}_{t+1}$, feature states for the forward and backward modules, respectively. Subsequently, the two inputs at the forward block are being subjected to the non-linear transformations

$$\overrightarrow{\boldsymbol{z}_t} = \sigma(W_{zh}^{f} * \overrightarrow{\boldsymbol{h}}_{t-1} + W_{zx}^{f} * \boldsymbol{x}_t) \tag{3.12}$$

$$\overrightarrow{\boldsymbol{r}_t} = \sigma(W_{rh}^{f} * \overrightarrow{\boldsymbol{h}}_{t-1} + W_{rx}^{f} * \boldsymbol{x}_t) \tag{3.13}$$

$$\overrightarrow{\hat{\boldsymbol{h}}_t} = \tanh(W_{hh}^{f} * (\boldsymbol{r}_t \odot \overrightarrow{\boldsymbol{h}}_{t-1}) + W_{hx}^{f} * \boldsymbol{x}_t) \tag{3.14}$$



$\boldsymbol{h}_1$
$\boldsymbol{h}_2$
$\boldsymbol{h}_T$
c
c
c
$\overleftarrow{\boldsymbol{h}}_1$
$\overleftarrow{\boldsymbol{h}}_2$
$\overleftarrow{\boldsymbol{h}}_T$
$\boldsymbol{h}_{T+1}$
$\overrightarrow{\boldsymbol{h}}_1$
$\overrightarrow{\boldsymbol{h}}_2$
$\overrightarrow{\boldsymbol{h}}_T$
$\boldsymbol{h}_0$
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_T$

Figure 3.6 – Block diagram of the BigruCNN cell. Red arrows indicate the point from which the processing start at each sub-module with $\boldsymbol{h}_0, \boldsymbol{h}_{T+1} = 0$.

$$\overrightarrow{\boldsymbol{h}}_t = \boldsymbol{z}_t \odot \overrightarrow{\boldsymbol{h}}_{t-1} + (1 - \boldsymbol{z}_t) \odot \hat{\overrightarrow{\boldsymbol{h}}}_t, \tag{3.15}$$

to get the forward feature representation, $\overrightarrow{\boldsymbol{h}}_t$, for the current frame. Similarly, the inputs at the backward sub-module are being transformed with

$$\overleftarrow{\boldsymbol{z}}_t = \sigma(W^b_{zh} * \overleftarrow{\boldsymbol{h}}_{t+1} + W^b_{zx} * \boldsymbol{x}_t) \tag{3.16}$$

$$\overleftarrow{\boldsymbol{r}}_t = \sigma(W^b_{rh} * \overleftarrow{\boldsymbol{h}}_{t+1} + W^b_{rx} * \boldsymbol{x}_t) \tag{3.17}$$

$$\hat{\overleftarrow{\boldsymbol{h}}}_t = \tanh(W^b_{hh} * (\boldsymbol{r}_t \odot \overleftarrow{\boldsymbol{h}}_{t+1}) + W^b_{hx} * \boldsymbol{x}_t) \tag{3.18}$$

$$\overleftarrow{\boldsymbol{h}}_t = \boldsymbol{z}_t \odot \overleftarrow{\boldsymbol{h}}_{t+1} + (1 - \boldsymbol{z}_t) \odot \hat{\overleftarrow{\boldsymbol{h}}}_t, \tag{3.19}$$

to generate the backward feature representation, $\overleftarrow{\boldsymbol{h}}_t$. The operations $*$ and $\odot$ indicate convolution and element-wise matrix multiplication, respectively.

The kernel set $\{W^f_{zh}, W^f_{rh}, W^f_{hh}\}$ operates on the forward feature map tracing any spatial patterns in the forward direction. Similarly, the other set $\{W^b_{zh}, W^b_{rh}, W^b_{hh}\}$ operates on the backward feature map searching for the backward spatial patterns. On the other hand, the input-to-hidden kernels $\{W^f_{zx}, W^f_{rx}, W^f_{hx}, W^b_{zx}, W^b_{rx}, W^b_{hx}\}$ search for local input patterns. The outputs of the forward and backward feature representations are concatenated to form the final features

$$\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_t], \tag{3.20}$$

where the feature map, $\boldsymbol{h}_t$, encodes the statistics of current frame in the context of the past and future frames. During training, the kernel coefficients are optimized so that to maximize the localized dependencies in the frequency axis, and propagate the best features along the frames. Unlike the fully connected RNNs that use matrix multiplications in the input-to-state and state-to-state transitions, the BigruCNN has fixed size kernels which considerably reduces the model parameters.

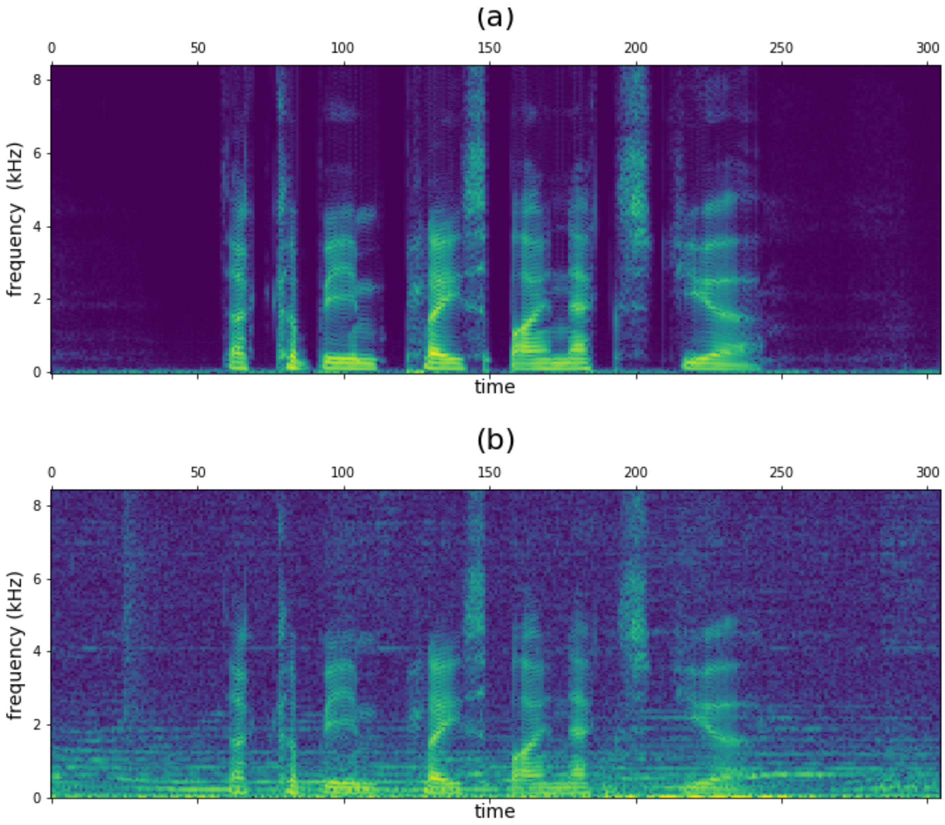


Figure 3.7 – a) clean spectrum. b) noise masked spectrum.

3.4 BigruCNN speech enhancement network (BigruCNN-SE)

A deep network with BigruCNN layers is constructed for speech enhancement in the spectral domain of speech. Only the magnitude of short-time Fourier transforms (STFT)[Por80] is considered, leaving the phase untouched. As such, the network objective is to extract relevant features from noisy magnitude spectrum at the input to restore the clean spectral magnitudes.

An example of clean and noise masked spectra is plotted in Figure 3.7, with spectral intensity as brightness. It is observable from Fig. 3.7(a) that the speech activities are localized in time. Besides, the transitions of spectral activities over time, e.g., transition of harmonics or formants, which contain essential information about the speaker, are frequency localized. While considering the masking of noise, Fig. 3.7(b), spectral regions are masked randomly without any clear pattern. Therefore, exploring the time-frequency localized characteristics in the spectrum in its full resolution is crucial for effective restoration of the noise-masked speech components.

In conventional speech enhancement models in time-frequency domain [TW18, ZZTL18, NBP$^+$17, HOZC20], [XSWN20] (refer to Figure 3.3), the temporal dependency among frames is modeled with the use of fully connected recurrent neural network (FC-RNN) layers, like LSTM [HXY15] or GRU [CGCB14], after a series of CNN front-end feature extraction layers. This filtering with CNN layers compresses the resolution of input spectrum to the FC-RNN layer, which hence could only see an abstract representation of the noisy input. Moreover, the FC-RNN can not preserve the spatial resolution over the frequency axis due to its inherent fully connected operations in the input-to-state and state-to-state transitions. To this end, we now present the use of BigruCNN module that are designed to preserve the spatial resolution in both axis when modelling the temporal dependency.

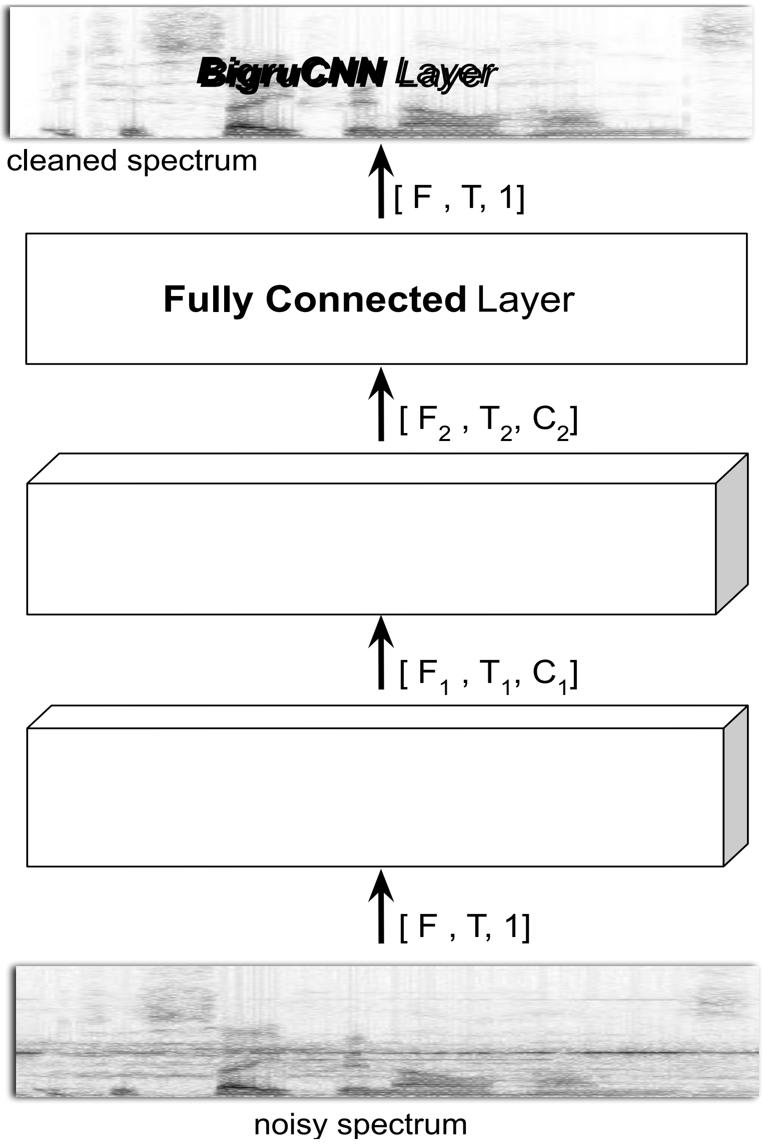


Figure 3.8 – BigruCNN-SE: Speech enhancement network with the BigruCNN as layers.

The BigruCNN cells are placed as the fundamental feature extraction layers of the proposed enhancement model (Figure 3.8). Since the temporal dependency is modelled at each layer starting from the input, transitions in the spectrum are captured as different levels of abstraction. Every layer has a set of 2D convolution kernels with different dimensions operating on the input and hidden states, the details of which are provided in the experimental section. The output dimensions at each layer, for an input spectrum with frequency resolution F and time resolution T, are also marked. The frequency dimension of the i^{th} layer output F_i is determined by the BigruCNN kernel parameters as

$$F_i = [(F_{i-1} - H_i + 2P_i^f)/S_i^f] + 1, \tag{3.21}$$

where H_i is the height of the 2D convolution kernels, $\{W_{zx}^f, W_{rx}^f, W_{hx}^f, W_{zx}^b, W_{rx}^b, W_{hx}^b\}$ in Equations 3.12 – 3.19, in the layer i. P_i^f is the size of zero-padding at the input along the frequency dimension, that was set to be $[H_i - 1]/2$. Hence, Equation 3.21 reduces to

$$F_i = [(F_{i-1} - 1)/S_i^f] + 1. \tag{3.22}$$

Similarly, the time axis is transformed with the kernels width W_i as

$$T_i = [(T_{i-1} - W_i + 2P_i^t)/S_i^t] + 1, \tag{3.23}$$

with setting zero-padding along time $P_i^t = [W_i - 1]/2$, the above expression reduces to

$$T_i = [(T_{i-1} - 1)/S_i^t] + 1. \tag{3.24}$$

where S_i^t defines the stride along the time axis.

In summary, the size of frequency and time axis at the output of a BigruCNN module is determined purely by the stride factors S_i^f and S_i^t, respectively. Since the compression of time resolution would lead to the loss of time localized events in the spectrum, we preserved the full resolution of the time axis across the layers. This is done by setting $S_i^t = 1$, in Equation 3.24, for all i. The channel dimensions C_1 and C_2 are layer hyper-parameter.

Also note that the output of BigruCNN module will have the channel dimension double that of the forward and backward sub-modules due to the reason that the concatenation in Equation 3.20 was performed over the channel axis. By looking at the already extracted features in the neighboring frames, the model will distill the feature representation for current frame to be the best fit in the context. At the top of the architecture, there is a fully connected layer that regresses the recurrently extracted features into the enhanced spectral bins. These predictions are combined with the noisy phase information passed from the input to reconstruct the enhanced waveform.

3.4.1 Model configuration and database selection

As the primary focus is on evaluating the efficacy of suggested recurrent feature extraction module against the conventional CNN module, a convolution-recurrent network (CNN_BiLSTM-SE), with CNN layers as front-end to extract features followed by a Bidirectional LSTM (BiLSTM) [GSC99] back-end to model the dependency over frames, was constructed. The LSTM was selected instead of GRU for recurrency modelling as they were found performing better in our initial experiments. This corresponds closely with the past speech enhancement studies where LSTM have been reported for modelling the temporal dependency in speech spectrum [TW18], [ZZTL18]. Besides, in order to see the performance gain of data-driven models over signal processing approach, we considered the Wiener filtering for speech enhancement suggested in [PMS06] as the Wiener filtering has been widely used in many real world applications.

Both the neural models considered have five layers in total; initially a series of casual or recurrent convolution followed by the fully connected regression. The frequency-time resolutions of convolution filter at each layer are provided in Table 3.5. The strides of convolution for CNN_BiLSTM-SE layers are set to [2,1], [2,1], [1,1], starting from the beginning. The layers have channel depth of 128 kernels. Parametric ReLU (PReLU) activation was used across the layers, except in the final layer where a normal ReLU activation is used to limit the predictions to positive value.

Table 3.5 – Layer-wise description of the two neural models

Layer	**CNN_BiLSTM-SE**	$Bigru$**CNN-SE**
1	[5 × 5] CNN	[5 × 5] $Bigru$CNN
2	[5 × 5] CNN	[5 × 5] $Bigru$CNN
3	[3 × 3] CNN	[3 × 3] $Bigru$CNN
4	BiLSTM (1024)	[3 × 3] $Bigru$CNN
5	FC	FC

Since each BigruCNN layer outputs concatenated tensor of the forward and backward feature maps, the channel dimension at the layer output is doubled, as depicted in Figure 3.9. Which will be the input to the next layer. Hence, the kernels of BigruCNN operating in the input space have double channel dimensions than that of the kernels operating on the feature maps; except at the beginning layer where the input spectrum is of unit depth. To be explicit, the exact dimensions of the kernels for the BigruCNN-SE layers are presented in Table 3.6. The naming is followed from Equations (3.12) – (3.19), where the superscripts have been omitted for notational simplicity as the dimensions of kernels are similar in the forward and backward sub-modules. By setting stride = 2 along the frequency dimension, based on Equation 3.22, the frequency resolution is nearly halved in the initial two layers. While the kernels operating on feature domain W_{zh}, W_{rh}, W_{hh} have a unit stride across the layers. Figure 3.9 provides a detailed view of how a two dimensional input spectrum is transformed and combined in BigruCNN module for the speech enhancement network.

Table 3.6 – Dimensions of the kernels at each layer of the BigruCNN-SE model. The kernels alignment is [Frequency × Time × Input channels × Output channels]

Layer	Specification	W_{zx}, W_{rx}, W_{hx}	W_{zh}, W_{rh}, W_{hh}
1	filter size	$[5 \times 5 \times 1 \times 128]$	$[3 \times 5 \times 128 \times 128]$
	filter stride	$[2 \times 1]$	$[1 \times 1]$
2	filter size	$[5 \times 5 \times 256 \times 128]$	$[3 \times 5 \times 128 \times 128]$
	filter stride	$[2 \times 1]$	$[1 \times 1]$
3	filter size	$[3 \times 3 \times 256 \times 128]$	$[1 \times 3 \times 128 \times 128]$
	filter stride	$[1 \times 1]$	$[1 \times 1]$
4	filter size	$[3 \times 3 \times 256 \times 128]$	$[1 \times 3 \times 128 \times 128]$
	filter stride	$[1 \times 1]$	$[1 \times 1]$

Data Set (Training and Testing) : The speech data-set is a selection of twelve British English speakers – both male and female – from the Voice Bank speech corpus [VYK13], each of which has around 400 clean utterances. Ten speaker's data were used for training, and the remaining two (one male and one female) were reserved for testing. Although the original speech were sampled at 48kHz, it was down-sampled to 16 kHz for our experiment as in [MSATS19][RPS18]. The noisy mixtures were simulated manually. The noises are from the DEMAND data set [TIV13], which contains 6 main categories of real world noise – Domestic, Nature, Office, Public, Street, Transportation – recorded at 16 kHz. In each of these categories there were recordings at different acoustics, adding up to 15 noise recordings in total. Ten of which were used for the training and the remaining five, selectively picked from each category, were used as the unseen noises to test the model. Although the actual DEMAND set had multiple channels, we extracted the first microphone recordings for the experiments. Before adding the noise, all clean speech files were root mean square (RMS) normalized to -26 dB level to avoid any loudness issues while training. To create the noisy training mixtures, each clean training sample was added with a random noise sample from the training set at a random *SNR* point from [0, 5, 10, 15] dB. Testing of the trained models has been done under both seen and unseen noise conditions; the seen are the noises from the training noise set and the unseen are the left-out noises from training. Subsequently, speech samples from the clean test set are mixed with random seen and unseen noises at [0, 5, 10, 15] dB to create the corresponding seen and unseen noisy mixtures.

The 16 kHz sampled signals were short-time Fourier transformed (STFT) with frame size 20 ms, hope size 10 ms, and 320 FFT points. Only the magnitude spectrum was enhanced while keeping the phase unchanged for the reconstruction. Because of the spectral symmetry, only the first half (161 points) of the STFT magnitude spectra was cleaned by the networks.

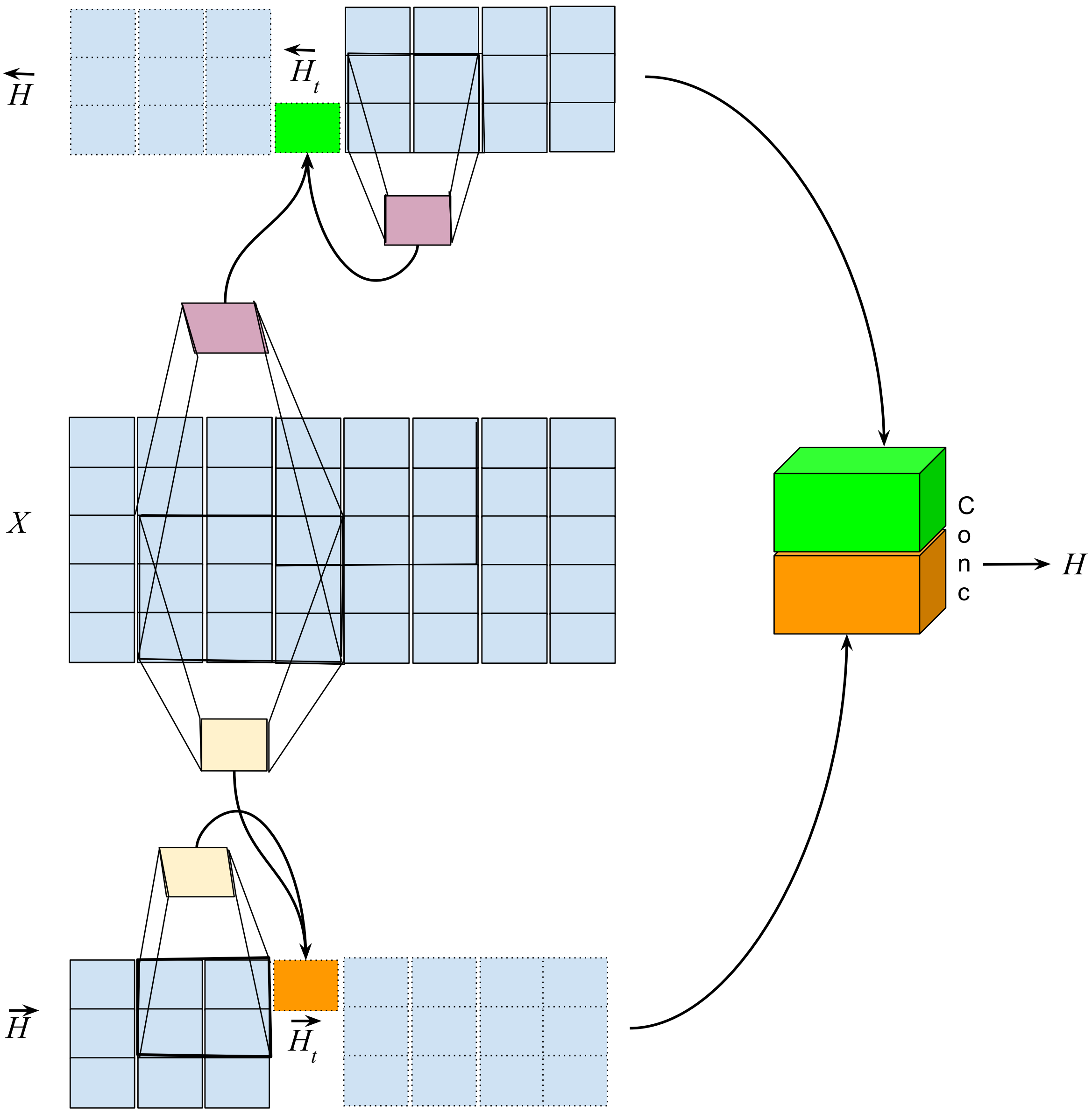


Figure 3.9 – Feature extraction in the BigruCNNs of the enhancement network; where X is the layer input, and $\overrightarrow{H}$ and $\overleftarrow{H}$ are the forward and backward feature streams, respectively, while H is the layer output.

Model Training: Both neural models have been trained in an end-to-end mode, where the losses have been computed directly between the magnitudes of predicted ($\hat{Y}(k,t)$) and target ($Y(k,t)$) STFT components. For each noisy-clean training set pair (X, Y), the model parameters are optimized by minimizing the mean square error (MSE) objective function

$$L_{X,Y} = \frac{1}{T \times K} \sum_{t=1,f=1}^{t=T,f=K} (|Y(k,t)| - |\hat{Y}(k,t)|)^2, \tag{3.25}$$

where K denotes the dimension of frequency axis that is 161, and the variable T is the number of time frames recurrently generated in the training process; which has been set to T = 128. The T value for testing varies based on the input signal duration. Note that the recurrency is being modeled over the temporal axis. The loss was minimized by the Adam optimizer [KB14] with an exponentially decaying learning rate method with learning rate = 0.001, decay steps = 20,000 and decay rate = 0.99. The models were continuously trained for 80 epochs and the model returned at the final epoch is considered as the final model.

Procedure: For objective evaluation, the same set of metrics used to evaluate the gruCNN-SE model are used for the performance assessment of the enhancement network with only a replacement – the STOI is replaced by the extended STOI (ESTOI)[JT16] for intelligibility measurement which has been proven effective in capturing the temporal modulation of speech in the measurement, therefore, shows stronger correlation with subjective scores.

The subjective evaluation on the quality of enhanced samples has been conducted among a group of 15 native English speakers with normal hearing. Individual tests were conducted at input SNRs of 0 dB, 5 dB and 15 dB in unseen noise condition. The subjects were asked to rate the quality of samples on a scale of 1 to 5, based on the presence of noise artifacts (0 – very annoying artifacts, 5 – no artifacts at all). Together with the two neural models and Wiener filter, unprocessed noisy and clean reference files were also included in the evaluation. The responses are averaged and considered as the mean opinion score (MOS) in each category.

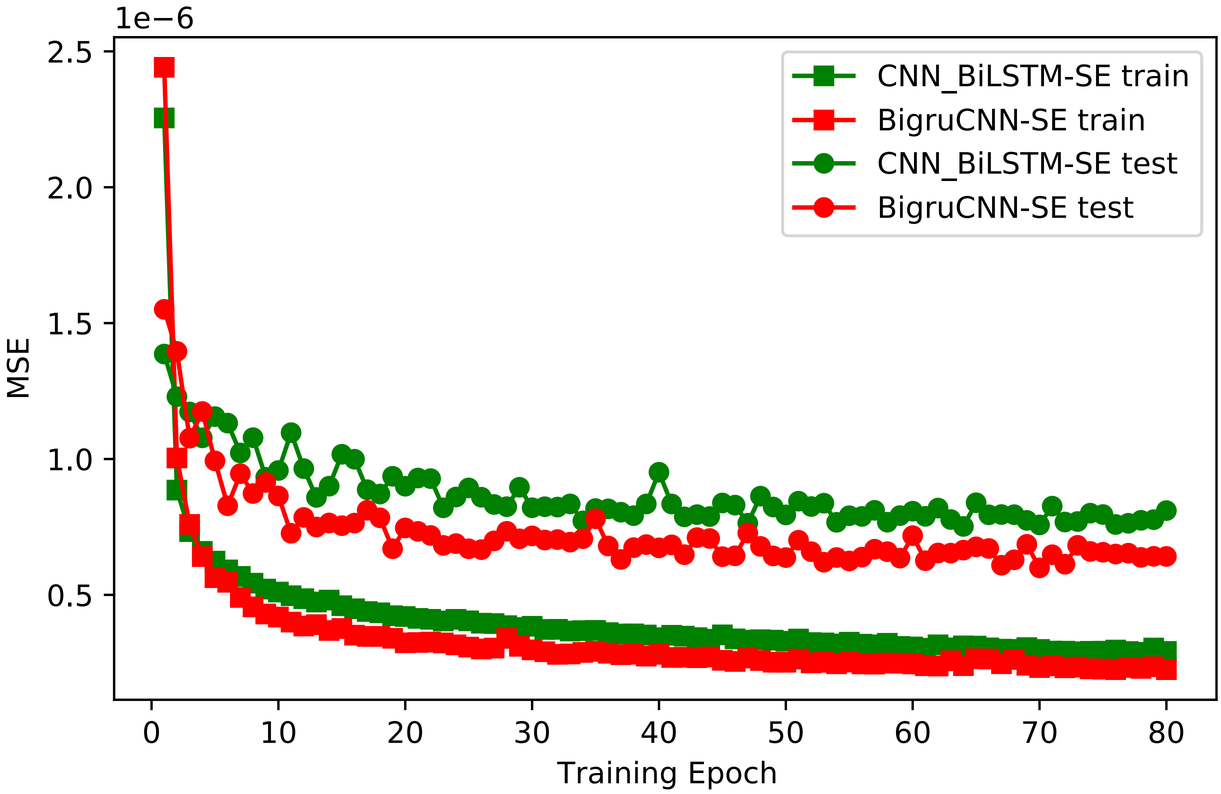


Figure 3.10 – Learning progress of the neural models on training and test data.

3.4.2 Key observations

Analysis of model convergence: Before delving into the numerical comparison of enhanced samples, it is informative to investigate how well both models have generalized to the enhancement task. Generalization of enhancement models is a factor of major concern for many practical applications where devices would operate in diverse noise conditions. The best and more direct way to quantify it is by comparing the convergence of loss function that has been used to train the model.

Often, the noise type that is encountered in deployment may not have been seen in training. This disparity in performance can only be quantified with the use of separate noise conditions for training and testing as has been followed in our experiment. A reliable model therefore must have stable convergence on the training and testing noise conditions. On this basis, we plotted the convergences of the loss function defined in Equation (3.25) on training and test data as a continuous plot with interpolation between the epochs. It is clear from Figure 3.10 that although the training of BigruCNN-SE had started at an higher point compared to CNN_BiLSTM-SE, the convergence of the loss function is steeper as training progresses. Besides, the minima of BigruCNN-SE curve (at epoch 80) is below to that of the CNN_BiLSTM-SE. These facts indicate that the BigruCNN-SE parameter optimization started in a higher manifold, and later is being confidently converged to an optimal final state.

Besides, the performance disparity between seen (training data) and unseen (test data with unseen noise) conditions is analysed. It is evident that the deviation of the two curves, for train and test conditions, is smaller in BigruCNN-SE model compared to CNN_BiLSTM-SE. When measured at 80^{th} epoch, the deviation was 5.182 X 10^{-7} for CNN_BiLSTM-SE and 4.160 X 10^{-7} for BigruCNN-SE – a 20 % relative reduction. This reduction in performance disparity must have been due to the better generalization of bidirectional gruCNN module to data distortions, as has been reported in the case of unidirectional gruCNN module with image data [Har18].

Objective evaluations: The improved generalizability of the proposed model should correspond to better quality of enhanced samples, where the quality is assessed by the mean objective scores averaged over the test set at different *SNR* levels. The results are displayed in Table 3.7. Both the seen and unseen noise conditions are categorized separately. The scores of unprocessed noisy speech are also included to better understand the relative gains/loses that are produced by different enhancement strategies.

First, we consider the performance of Wiener filtering approach. The perceptual quality in terms of PESQ, and intelligibility in terms of ESTOI & *SSNR* has been improved over unprocessed noisy speech across seen and unseen conditions. Since it is data independent, there is not much difference in performance between seen and unseen conditions. However, one must closely examine the CSIG and CBAK scores to identify the source of the gain, for they categorically count the gains from speech restoration and noise suppression. At very low *SNR*s, background noise has been suppressed largely as quantified by the higher CBAK scores over the noisy speech, subsequently resulted to the large improvements in *SSNR*. However, in the process, speech components have also been affected, leading to the lower CSIG scores. The lagging performance of the model at high *SNR*s must be attributed to this aggressive filtering leading to speech distortions.

Concerning the neural models, the performance varies considerably between seen and unseen noise conditions. At the lowest input *SNR* of 0 dB, the performance falls around 2 dB on *SSNR* scale in unseen case. Nonetheless, the performances of neural models are still far better than that of the Wiener approach across different noise levels. With the proposed network, intelligibility improvements of 3.5 to 7.1 dB in seen noise conditions and 2.1 to 6.6 dB in unseen conditions are achieved compared to the Wiener filter in 0 to 15 dB input adversities. A similar improvement pattern is observable in other metrics as well. This underlines the fact that the non-linearity modelling in neural networks is beneficial in enhancement task due to the complex nature of noise attributes. It must also have been the case that the temporal dependency modeling of speech in the Wiener filtering is relatively basic compared to the neural cells, like bidirectional LSTM (BiLSTM).

When compare the performances of two neural speech enhancement models, the modelling of BigruCNN module with recurrency modelled in the local spectral patches seems to have benefited for restoring speech from the mixture. Table 3.7 tells that the suggested BigruCNN-SE model produced better enhancements both in seen and unseen noise conditions across the input adversities. Segmental *SNR* gains of 0.5 to 0.8 dB in seen and 0.5 to 0.7 dB in unseen conditions are observed over the CNN_BiLSTM-SE. While relatively higher differences in intelligibility on ESTOI scale were observed between the models in unseen conditions compared to seen conditions. In unseen condition, about 5.0% relative gain in ESTOI at the lowest SNR of 0 dB, and 2.5% relative gain at the highest SNR of 15 dB are observed. This improved performance in unseen scenarios must be attributed to the speaker generalizability of the BiguCNN-SE model, for ESTOI is more robust metric than SSNR that takes into account the temporal modulation of speech in the measurement. Equally, the perceptual quality of processed samples (PESQ) has been improved by 8 – 10% in seen conditions and 12 – 15% in unseen conditions in reference to the baseline neural model. The reliability of these gains on PESQ scale is validated by the higher values of CBAK and CSIG. Although the performance of neural models declined in unseen conditions, BigruCNN-SE model have still retained the relative benefits across input adversity levels. This further points to the better generalization of BigruCNN module in modelling distorted data as part of the enhancement

Table 3.7 – Objective performance scores in seen and unseen noise conditions

		SEEN NOISE				**UNSEEN NOISE**			
Noise level	**Metric**	**Noisy**	**Wiener**	**CNN_BiLSTM-SE**	**BigruCNN-SE**	**Noisy**	**Wiener**	**CNN_BiLSTM-SE**	**BigruCNN-SE**
0 dB	PESQ	1.25	1.42	1.60	**1.76**	1.28	1.45	1.49	**1.67**
	CSIG	2.23	1.42	1.73	2.42	2.36	1.61	1.69	2.20
	CBAK	1.65	1.85	2.33	**2.49**	1.61	1.79	2.11	**2.27**
	COVL	1.68	1.32	1.60	**2.05**	1.76	1.46	1.53	**1.88**
	ESTOI	0.60	0.62	0.67	**0.69**	0.55	0.58	0.61	**0.64**
	SSNR	- 4.09	1.23	4.26	**4.73**	-4.60	0.78	2.32	**2.86**
5 dB	PESQ	1.46	1.68	1.91	**2.09**	1.50	1.69	1.73	**1.98**
	CSIG	2.63	1.77	2.14	2.90	2.70	1.85	1.92	2.67
	CBAK	1.99	2.10	2.62	**2.80**	1.94	2.05	2.43	**2.63**
	COVL	2.01	1.62	1.98	**2.46**	2.05	1.69	1.77	**2.28**
	ESTOI	0.69	0.71	0.74	**0.75**	0.65	0.68	0.69	**0.72**
	SSNR	-1.33	1.92	5.72	**6.53**	-1.99	1.65	4.48	**5.22**
10 dB	PESQ	1.77	2.03	2.24	**2.43**	1.81	2.01	2.03	**2.34**
	CSIG	3.07	2.18	2.48	**3.36**	3.11	2.18	2.24	**3.22**
	CBAK	2.40	2.37	2.92	**3.08**	2.35	2.32	2.75	**2.97**
	COVL	2.40	2.02	2.34	**2.87**	2.44	2.00	2.10	**2.75**
	ESTOI	0.76	0.77	0.78	**0.79**	0.74	0.76	0.75	**0.78**
	SSNR	2.07	2.45	7.30	**8.15**	1.26	2.33	6.56	**7.28**
15 dB	PESQ	2.17	2.42	2.54	**2.80**	2.22	2.36	2.35	**2.72**
	CSIG	3.55	2.65	2.71	**3.77**	3.57	2.56	2.53	**3.69**
	CBAK	2.89	2.64	3.19	**3.39**	2.84	2.58	3.07	**3.31**
	COVL	2.86	2.47	2.60	**3.27**	2.90	2.38	2.42	**3.19**
	ESTOI	0.82	0.82	0.81	**0.83**	0.80	0.81	0.79	**0.81**
	SSNR	6.17	2.83	9.14	**9.97**	5.28	2.79	8.69	**9.40**

network.

Subjective evaluations: The subjective evaluation was conducted at three *SNR* points: 0 dB, 5 dB and 15 dB. The evaluation was done in unseen noise conditions and the results are plotted in Figure 3.11. It is clear that the Wiener filtering has improved the quality of speech across the *SNR* levels. However, the relative gain is largest at the highest *SNR*, which must have been due to the large distortions of speech resulted from the inaccurate estimation of noise at low *SNR*s. In contrast, both the neural models have further improved the perceptual quality of samples than the signal processing approach. Besides, the quality of samples enhanced by suggested BigruCNN-SE model is consistently higher than that of the other two approaches. Especially in the low *SNR* range, where MOS improvements of 0.20 points at 0 dB and 0.42 points at 5 dB are observed over the CNN_BILSTM-SE model. This dominance at low *SNRs* must be attributed to the robust restoration of speech spectral components realized through the multi-level modelling of temporal dependency as different layers of abstraction in the BigruCNN-SE architecture.

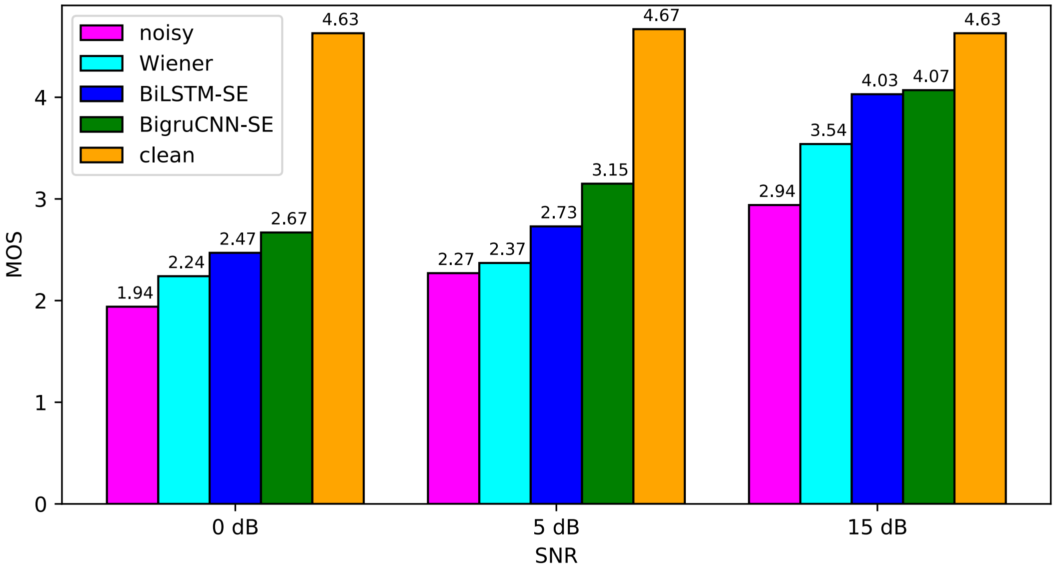
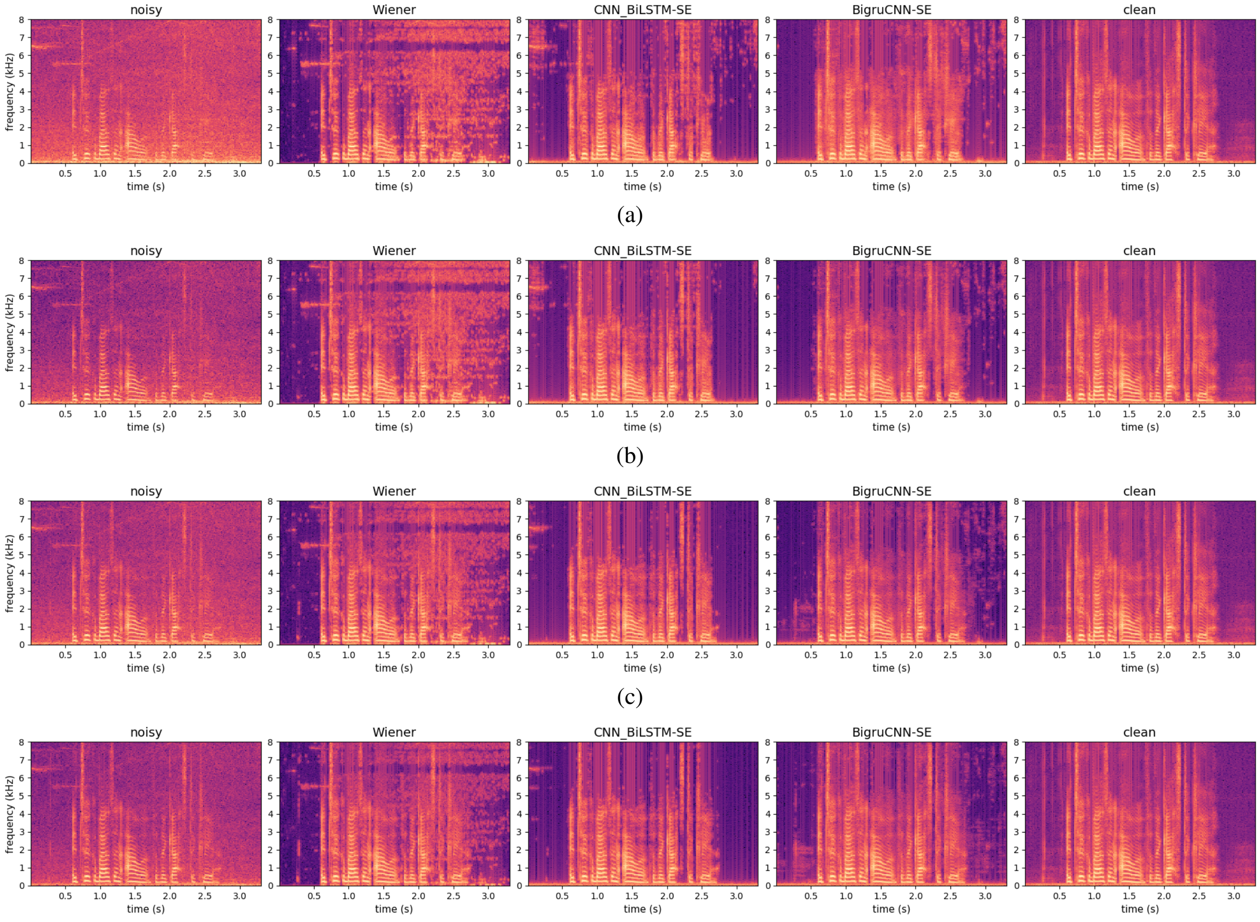


Figure 3.11 – Subjective Mean Opinion Scores at different SNR levels.

To further quantify the disparity in modeling the temporal dependency among the enhancement models, we masked a speech segment with a non-stationary noise, i.e., noise with a temporally varying character, at different *SNR*s and the enhancements are plotted in Figure 3.12. The noise levels were chosen to be the same as the test set *SNR*s, which are 0, 5, 10, 15 dB. By comparing the noisy spectrum (first column) against the clean spectrum (last column) in Figure 3.12, one can easily identify the spectral pattern of additive noise as with small strips of large intensity at high frequencies in the beginning of the spectrum. Isolating such localized activities has been a challenging exercise for speech enhancement models. For instance, consider the Wiener filtering wherein most portions of the strips have been removed, but, leaving behind a large channel of speech distortion around that frequency region in the entire duration of the spectrum. This is because of the fact that the estimation of noise activities in the initial speech absent region identified those frequency components as the noise source frequencies, therefore are suppressed until a new speech absence region appears to update the estimate.

When compare the enhancements by CNN_BiLSTM-SE and BigruCNN-SE, neither of them has introduced any large-scale distortion to the speech. However, in the case of CNN_BiLSTM-SE, nor the suppression of the noise strips has been satisfactory, mainly at the low *SNR* conditions of 0 dB and 5 dB. On the other hand, BigruCNN-SE was highly effective in detecting and suppressing the localized noise strips without affecting any active speech components. This performance is visible across the *SNR* levels. Besides, as the *SNR* increases, the model starts recovering more and more



(a)

(b)

(c)

(d)

Figure 3.12 – Model enhancements at different input signal to noise ratios (SNRs); (a) 0 dB SNR, (b) 5 dB SNR, (c) 10 dB SNR, (d) 15 dB SNR.

finer speech features in the masked spectrum, making the predictions closer to the clean spectrum. This further validates the argument that the recurrency over time modeled at different levels of abstraction in the BigruCNN-SE is beneficial to detect and isolate local activities in the spectrum.

Although the enhancements were carried out in the spectral domain, it might still be useful to see the resulting changes in the waveform domain. Thus, the reconstructed waveforms from the enhanced magnitude spectrograms using noisy phase information are plotted in Figure 3.13. The noisy speech is the clean speech masked at 0 dB *SNR* level by a random unseen noise from the test set. One could easily perceive the level of speech distortion that has been introduced by the Wiener filtering technique. While both the neural-based models preserved the speech components, BigruCNN-SE yielded the closest to the clean waveform enhancement.

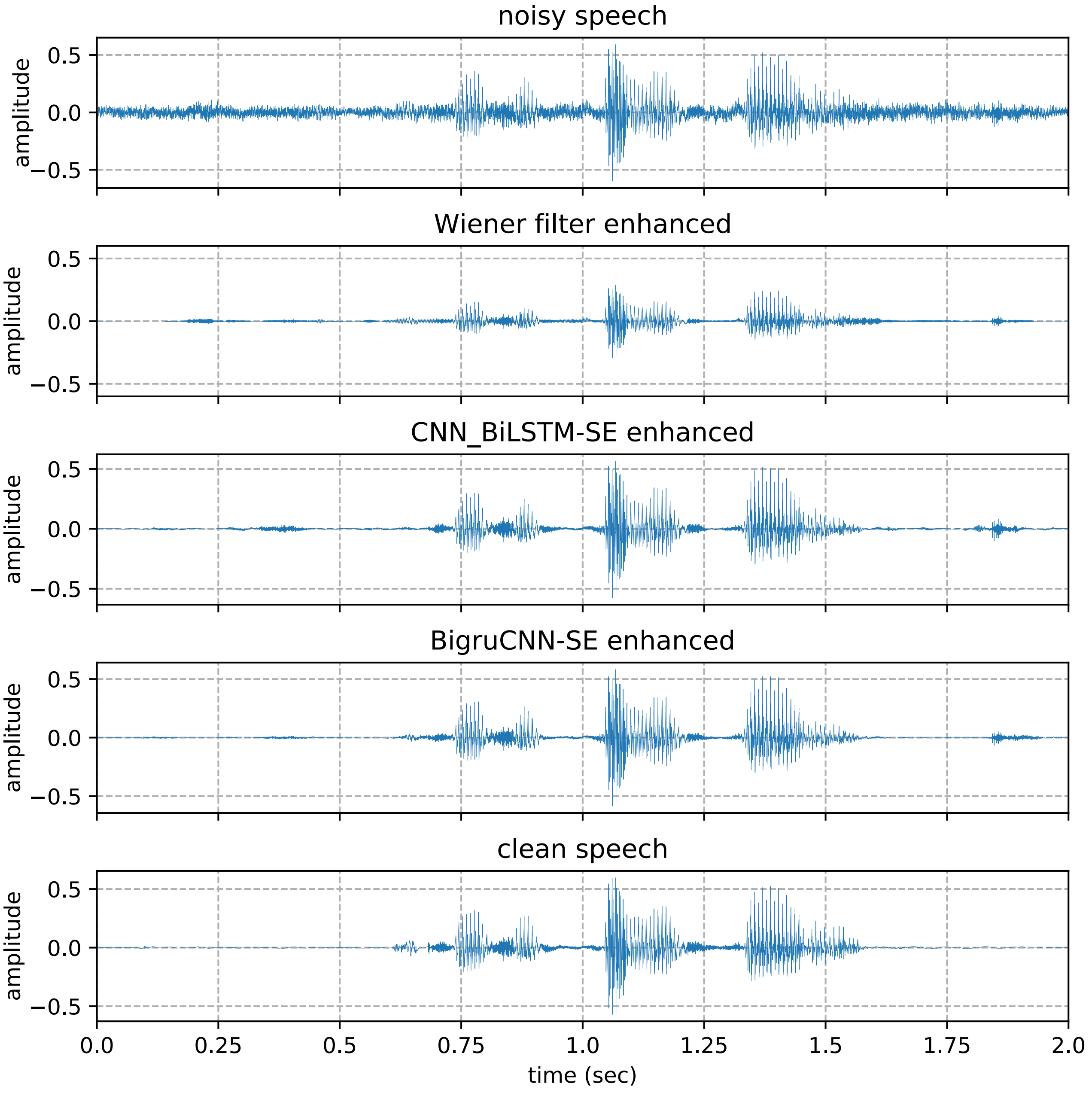


Figure 3.13 – Reconstructed waveforms from enhanced magnitude spectra with the noisy phase information at 5 dB SNR level.

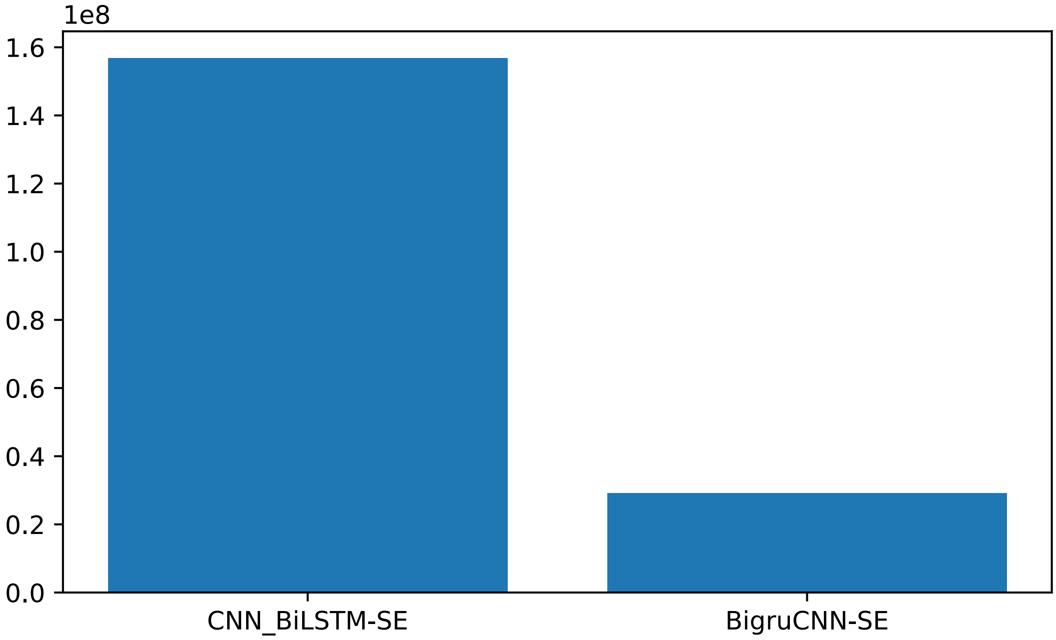


Figure 3.14 – Parameter counts of the neural models

Analysis of parameter complexity: Pragmatically, the performance gain of a neural model could be argued by the additional parameters that are floated into the modeling. To this end, the parameter counts of different models are presented in Figure 3.14. The Wiener filter was avoided for it only has very few statistical parameters. Although Wiener filtering is the least complex among the models, the performance of which is much weaker than the other two models as have seen above. While, the suggested BigruCNN-SE produces a far better quality enhancement with only 19% parameters than that of CNN_BiLSTM-SE model. This reduction in complexity is due to the replacement of the fully connected BiLSTM layer for modelling the temporal flow with the fixed size kernels of BigruCNN neural module. Such a large reduction in parameters is highly desirable for low-end devices, like hearing aids. A Tensorflow implementation of the network is provided at [3].

3.5 Conclusions

In this chapter, the concept of feature domain enhancement of speech was discussed and a few new inventions in that regard were presented. The speech was first transformed to short-time Fourier transform (STFT) representation, and only the magnitude of the spectra was enhanced while considering the noisy phase as a replacement for the clean phase. This was justified by the fact that phase does not contribute much to the intelligibility of speech. Subsequently, two new novel architectures, namely gruCNN-SE and BigruCNN-SE, were presented for improved enhancement of the magnitude spectra.

gruCNN is a feature extraction cell with recurrent connections to memorize the past, which was initially suggested for the image classification task showing robust detection of images in very low SNR conditions. gruCNN-SE network was built with gruCNN as layers by visualizing the sequence of 2D spectrograms as a time series of which the dependency has to be modeled. It enabled us to capture the correlations in spectral events at different levels of abstraction; fine details at the initial layers and coarse details towards the end layers of the network. Both the objective and subjective evaluations have confirmed the robustness of the proposed system to detect and isolate noise events in the spectrum. The

[3] https://github.com/shifaspv/BigruCNN-SE-tensorflow

effectiveness of the system is more prominent in non-stationary noises. When the noise spectrum has high variability, it is much more important to pay attention to the fine structures to isolate them effectively, which explains why the proposed model outperformed the standard CNN+LSTM model which models the temporal recurrency only on the smoothed representation of the input spectrum from the CNN layer. Besides, gruCNN cell requires only very few parameters than the standard LSTM to persist the memory, therefore, the gruCNN-SE architecture is far less complex than the traditional enhancement models.

Subsequently, a bidirectional extension of gruCNN cell (BigruCNN) is proposed. BigruCNN is presented as a new neural module to use in applications where we have the entire time series available to model the dependencies in forward and backward directions. In the context of speech, such a model will be useful to clean pre-recorded speech for applications like clean data set creation, or for no-real time applications in general. BiguCNN-SE network was built with BigrCNN as layers. The evaluation showed that the proposed BigruCNN layer can further improve the enhancement process. As in the case of gruCNN-SE, the model was able to generalize better to various noisy conditions. Besides, the model showed better generalization to unseen conditions than the traditional approaches.

The speech enhancement techniques considered so far have only enhanced the magnitude spectral information while keeping the phase untouched. In the next chapter, we will be introduced to the importance of the phase of speech, and will also research on how to develop neural models to jointly enhance the magnitude and phase of the signal.

Chapter 4

Neural Networks for Speech Enhancement in Time Domain

The enhancement models described in the last chapter explored the non-linearity modeling capability of neural architectures in the feature domain, specifically in the magnitude of the STFT representation of speech, therefore ignore the phase information. There has been mixed opinion about the enhancement of phase information for various speech applications. Although many old approaches totally ignore the phase information, some recent studies have reported the importance of phase in speech enhancement, especially to improve the quality of the enhanced signal [PWS11]. Similarly, the use of enhanced phase spectrum was observed to improve the recognition accuracy of automatic speech recognizers (ASRs) [FSS+16]. Approaches like the Griffin-Lim algorithm were also suggested to restore the phase spectrum from the magnitude spectrum of the signal through iteratively minimizing the error between the reconstructed spectrum and real spectrum [GL84]. Such techniques have been widely used in modern speech synthesizers to enhance the quality of synthetic voices [WSRS+17].

To analyze the importance of phase to the quality of reconstructed waveform in the context of speech enhancement, a reconstructed waveform from noisy phase spectrum (at 5 dB SNR) and clean magnitude spectrum (assuming an ideal enhancement model) is presented in Figure 4.1. It is evident that the replacement of the clean phase with a noisy corrupted version does not guarantee a smooth reconstruction of the waveform. The impact of phase mismatch is more prominent in consonant sounds. Although consonants contribute little to sentence intelligibility compared to vowels [KPBL07], such degradations could result to poor quality of the enhanced signal.

Therefore, having an enhancement strategy that can clean both the phase and magnitude of speech is essential for quality enhancement. Attempts to enhance the complex STFT spectrum by splitting into real and imaginary spectral streams were made in the past [FHTL17, HLL+20]. Although such approaches may improve the performance compared to older models, they would still suffer from the abnormality in spectral estimates in reconstruction. On the other hand, waveform architectures for speech enhancement have attracted considerable attention recently: WaveNet [RPS18] and Generative Adversarial Networks (GANs) [PBS17] architecture are being used commonly. They were built on the waveform domain operating with raw noisy samples as input and the corresponding clean samples as the target. However, there are some serious limitations for the present state-of-the-art enhancement models: (1) none of them have given enough attention to the patterns of speech and noise in a noisy mixture when designing their architectures; (2) the complexity in terms of model parameters and latency in the processing are huge, therefore, they are not suitable candidates for real-world applications.

As an alternative to alleviate such limitations, we suggest a new shallow time-domain architecture for speech enhancement. The model has dilated convolution layers with the dilation factor decreasing over the depth, which is found effective at exploiting the quasi-stationary nature of speech while suppressing the non-stationary noise in the mixture. Such a framework was suggested earlier for text-to-speech synthesis to generate speech from the text in an autoreg-

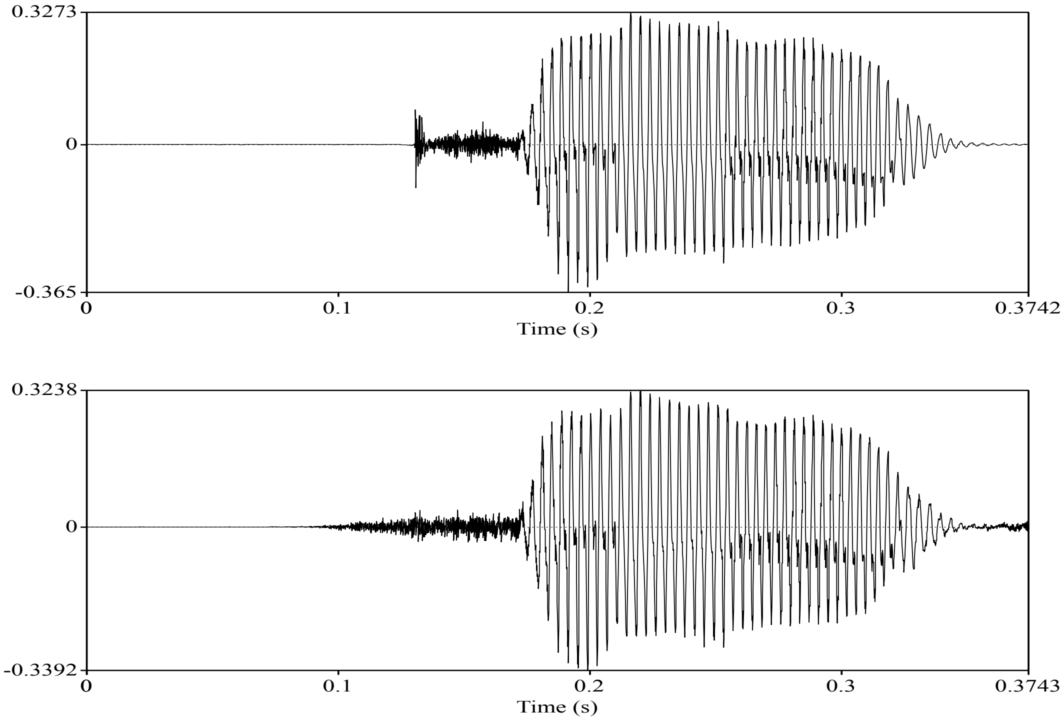


Figure 4.1 – (a) reconstructed speech from clean magnitude and phase STFT spectrums (b) reconstructed speech from the clean magnitude spectrum and the noisy phase spectrum (at 5dB noise level.)

gressive manner, called as FFTNet architecture [JFML18]. Therefore, the new speech enhancement model is called SE-FFTNet. In contrast to the original FFTNet auto-regressive structure, SE-FFTNet processes the entire input in parallel which significantly increases the prediction speed of the model. Furthermore, SE-FFTNet extends the architecture to a non-causal form for improved acoustic modeling.

4.1 Theoretical background on the design of SE-FFTNet

The neural networks modelling capacity is highly depended on the data set and task on which it is deployed. A model that performs well on images domain may not be the best promising model for speech application, as the speech has rapidly varying samples (16 K samples per second) over time in contrast to the image. This variation should be considered when implementing the neural architecture for speech applications. Even, among the speech applications, differences between the tasks should be taken into account, i.e., the task of Vocoder is very different from that of a speech enhancer. In the specified application of speech enhancement, often the noise in the recorded speech will be less correlated over time than the clean speech. Though many neural models have been suggested for speech enhancement task in recent years, very few of them had given enough attention to the correlation patterns of noisy speech.

In this context, we explored the long-term correlation of speech through an initial wide dilation pattern architecture. In contrast to the traditional waveform models which used the local neighbouring samples for extracting the features from the input mixture, the suggested model accounts the wide apart samples of input. By doing so we expect that the network could effectively discriminate the noise from clean speech. This idea was motivated by the recently proposed FFTNet architecture [JFML18]. In FFTNet, the input is split into two equal segments and the merged representation of the two segments is used as input on the next stage. It has been applied successfully in speech synthesis and has a reduced computational complexity compared to other neural-based vocoders. The novelty of this architecture is further important for speech enhancement on exploring the correlation structure of speech and noise.

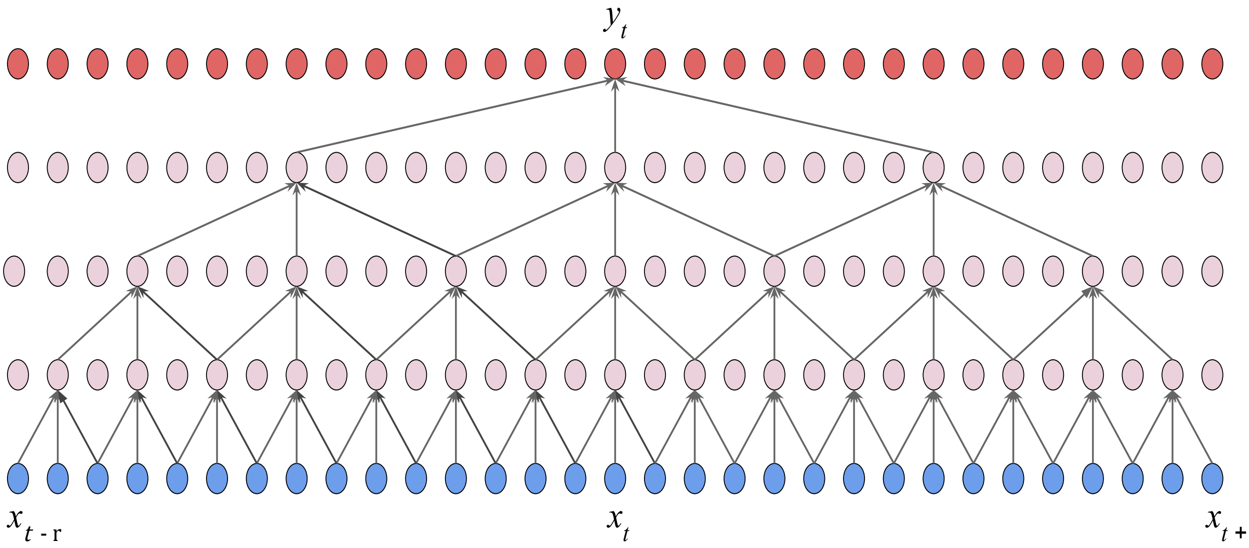


Figure 4.2 – Convolution pattern of SE-WaveNet/ SE-InvFFTNet

4.2 Speech enhancement FFTNet (SE-FFTNet)

The time domain models have the ability to capture high-level acoustic features. Their performance superiority has been proven for many speech applications [ODZ+16]. In the case of speech enhancement, the target is to estimate the clean speech samples from noisy speech samples. As it would be challenging to model the sample distribution of the clean speech from the noisy input, we modeled the denoising task as a regression problem: the model will be looking for the hidden function in the data which represents the mapping from noisy input speech x_t to the clean output speech y_t. This is mathematically formulated in Equation 4.1. Here, the objective of the model is to learn the hidden function f from the given data.

$$\hat{y}_t = f(x_{t-r1}, \dots, x_{t-1}, x_t, x_{t+1}, \dots, x_{t+r2}) \tag{4.1}$$

The model receptive fields enabled the dependency of past x_{t-r1} and future x_{t+r2} input samples. The model can be causal and non-causal depending on whether to consider the future samples or not, while performing the current sample prediction. This can be done by controlling the variable $r2$. We have compared the performance of the causal ($r2 = 0$) and non-causal model ($r2 \neq 0$) and it has been found that adding non-causality improves the model performance. Hence, in the rest of the paper, the discussion will be on the non-causal model.

In WaveNet [ODZ+16], sample dependency is introduced by a dilated convolution structure of increasing dilation rate, having the convolution pattern similar to Figure 4.2. This means the first layer of the network extracts the features by looking into the immediate behind and ahead samples. Since the speech and noise variation being equally negligible on these closer time instances, the model may not learn any good discriminating features in its initial layers. This will be rippled on the following layers. To account for this, one must look into the further apart samples of input where time domain correlation for speech is expected, in contrast to noise. To model this, inspired by FFTNet, the suggested SE-FFTNet has the dilation pattern as shown in Figure 4.3. We argue that such an architecture will enable SE-FFTNet to easier learn the weights which could discriminate the speech from noise. The similar convolution strategy has been repeated over the layers, until the final enhanced sample is obtained. In other words, SE-FFTNet enables coarser representation at initial layers and finer towards the end. Thus, helping to propagate much cleaner features from the bottom layer to the end.

In order to evaluate our hypothesis on the influence of initial versus later wide dilation pattern while keeping the internal blocks of the network the same, we suggest to investigate an FFTNet structure where a *later* (similar to WaveNet

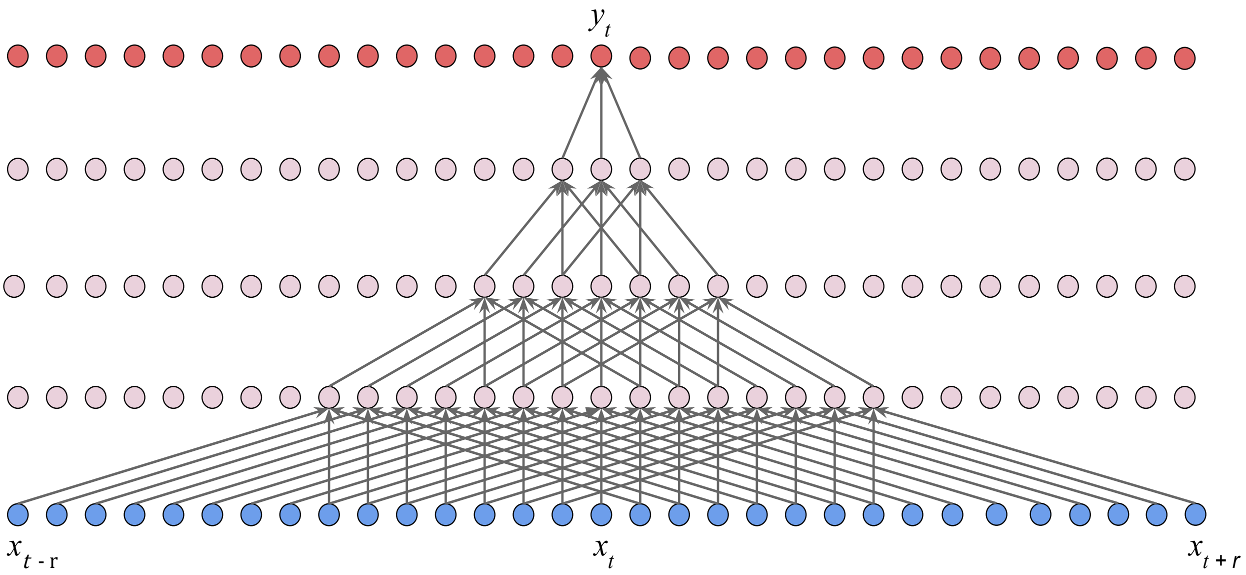


Figure 4.3 – Convolution pattern of the suggested SE-FFTNet

model) dilation pattern is used. We will refer to that model as SE-InvFFTNet and that is shown in Figure 4.2. Therefore, the dilation structure of the SE-FFTNet shown in Figure 4.3 has been inverted so that to have a local neighbouring representation of the input as shown in Figure 4.2. It has the dilation pattern similar to the WaveNet presented in SE-WaveNet, but with a difference: the block convolution is retained as shown in Figure 4.4 in contrast to the WaveNet residual block. This is needed as the actual WaveNet-SE model and the proposed SE-FFTNet has a different internal block convolution structure connecting to each layer.

As the denoising model has to compete with real-time computational constraints we have removed the temporal recurrence on the predicted samples. This means the sample generated at each time instances are totally disjoint, which was not the case in the initial FFTNet model [JFML18]. This significantly speeds up the generation process in contrast to the original model while retaining the acoustic modeling ability. The skip connections have been put in place between the layers to facilitate further information flow to the succeeding layer in each level. This is further helpful to restore the phase information which was lost/distorted on passing the signal through the block convolution operations and also, to facilitate gradient back-propagation on training [VOKK16].

The series of operations hidden between the layers are highlighted in Figure 4.4. The past, present and future samples being processed through an one-by-one convolution ($[1 \times 1]$) of specific channel size. It is followed with an ReLU activation before being sum up into a single representation. The summed output then passed through another set of one-by-one convolution ($[1 \times 1]$) followed by a ReLU activation, to have the final output from the block. This will be added onto the skip bypassing signal from the block input, to have the final input to the next layer. In the end, it is a fully connected layer merges the channel dimension into a speech sample.

4.2.1 Evaluation of SE-FFTNet

Data Set: To evaluate the proposed model, 30 speakers were selected from the Voice Bank corpus [VYK13]. Out of these, 28 speakers were used for training and each speaker data consists of around 400 sentences. To create the noisy mixture, each of these files has been chosen randomly and mixed at a specific SNR point from [0, 5, 10, 15] dB with a selected noise type from the noise set that contains 10 different real-life noises. The different type of noise is selected from the DEMAND database[TIV13]. The remained two speakers were used for testing with the same type of noises used in training but with 4 different SNR level falling in [2.5, 7.5, 12.5, 17.5] dB. To compare the performance, two recently proposed waveform domain speech enhancement models are considered, namely SEGAN & SE-WaveNet, which are described below.

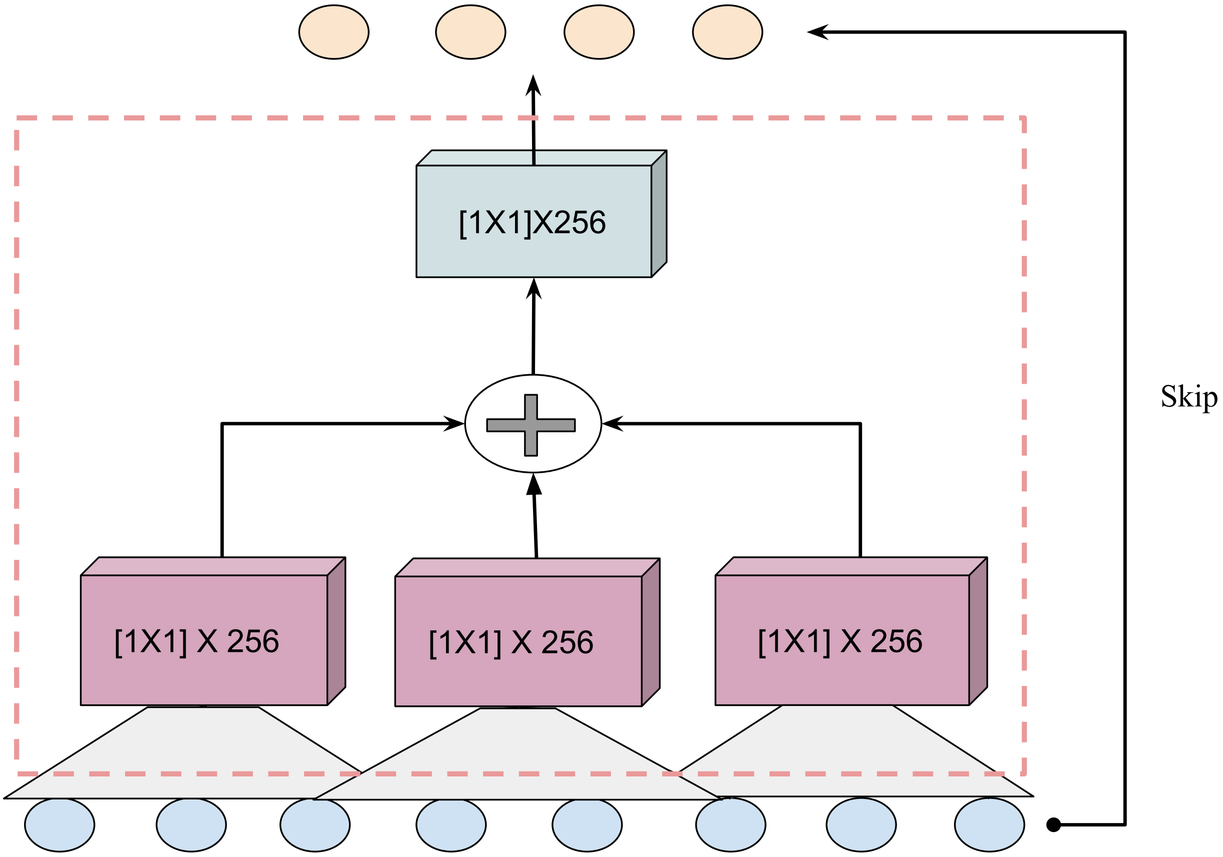


Figure 4.4 – Block insight of SE-FFTNet

Loss Function: Next we define an appropriate loss function. Since the enhancement task has been formulated as a regression problem, a solution is the mean of the absolute value between the predicted samples and the corresponding clean samples.The distance for the k^{th} training utterance is defined as:

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - r} \sum_{t=r/2}^{T^{(k)}-r/2} |y_t^{(k)} - \hat{y}_t^{(k)}|. \tag{4.2}$$

where, the symbols $y^{(k)}$ and $\hat{y}^{(k)}$ = SE-FFTNet($x^{(k)}$) correspond to clean signal and to the output of the network, respectively, while $x^{(k)}$ is the noisy signal. $T^{(k)}$ is the number of samples of the k-th utterance and r is the extend of the receptive field. The parameters of the model are tuned in the direction that minimize this loss. The model is trained with noisy speech as input and the corresponding clean speech as the target.

Baseline Models:: *Speech enhancement GAN (SEGAN):* Pascual et al. [PBS17] proposed speech enhancement generative adversarial network (SEGAN). The SEGAN consists of two neural networks, namely, Generator and Discriminator. The Generator network is inspired by Autoencoder architecture. The Generator encoder consists of 11 layers of stride-2 convolution with growing depth, resulting in a feature map at the bottle-neck of 8-time steps with depth 1024. This feature map is concatenated with latent vector ”z”, sampled randomly from uniform noise distribution. The resultant concatenated vector is input to an 11-layer up-sampling decoder, with skip connections from corresponding input feature maps. The least square based loss function is used to train SEGAN with additional $L1$ norm to preserve the structure of the enhanced signal.

Speech Enhancement WaveNet (SE-WaveNet): Rethage et al. [RPS18] modified the actual WaveNet Vocoder architecture to fit into the speech denoising task. It used a non-causal WaveNet architecture having a dilation pattern similar to Figure 4.2, by posing the denoising as a regression task. The model had a series of residual blocks plus the post-processing unit to process the skip outputs from each of these residual blocks. The model was trained to minimize the sample absolute difference objective function, same as in our proposed model. The model had in total 28 residual blocks, with a similar configuration as mentioned in the original paper [RPS18].

SE-FFTNet model configuration: Both these models have in total 30 layers made up by thrice repeating a block of depth 10 having the dilation factors: [512, 256, 128, 64, 32, 16, 8, 4, 2, 1] for SE-FFTNet and [1, 2, 4, 8, 16, 32, 64, 128, 256, 512] for SE-InvFFTNet. It sums up to a receptive field of size 6138 (3069 past & 3069 future samples), which means it considered 0.38 s of noisy input samples (for 16 kHz signal) when predicting a single clean sample. In all the layers, one-dimensional convolutions are used with the same number of 256 channels. As the final fully connected layer is being enrolled to merge this channel dimension into a single sample, that has a dimension of [256,1]. During training, the target samples predicted in a single traverse is a set of 4096 (training target field size). The model is fed with a single data point every time with a batch size of 1. In the testing phase, the target field size being varied depends on the test frame length. Just before feeding into the model, the wave files have been normalized to an RMS level of 0.06. This removed the loudness variations among the wave files. The model output loss is minimized with an Adam optimizer of the initial learning rate of 0.001.

As mentioned before, in order to evaluate the influence of dilation steps in the performance, we considered and inverted SE-FFTNet architecture as shown in Figure 4.2. We refer to this architecture as the SE-InvFFTNet model.

Procedure: All these models were being trained in a speaker independent fashion. The output speech quality is evaluated both on subjective and objective scale. The perceptual evaluation of speech quality (PESQ) is used as an objective measure of naturalness [HL07]. The Short-time objective intelligibility (STOI) score is used to measure the intelligibility gain by processing the noisy mixture, in reference to the clean [THHJ11]. The gain in SNR through the model processing is being evaluated by segmental SNR (SSNR) scale [HL07]. The speech distortion and the residual noise intrusion on the enhanced signal are measured with CSIG & CBAK along with the overall quality of the signal with COVL [HL08].

The subjective evaluation was done with non-native English listeners listened to the processed samples from different models. To cover the entire test set we have used both lower and high SNR samples while selecting the sentences for listening experiments. They were asked to rate the quality of the samples on a scale of 1-5. In total 15 responses were collected and averaged across all the participants to get the final mean opinion score (MOS).

4.2.2 Key observations

The models testing is done over 824 files from the test set comprised of different noises. Hence, the results displayed are an average performance on the test set. Table 4.1 included the objective performance gain of the proposed SE-FFTNet model along with its competitors. It is clear that the SE-FFTNet model outperforms both the waveform based SE-GAN or SE-WaveNet models. This improvement is reflected in all the subjective metrics in Table 4.2. The higher values in CBAK and CSIG is a clear indication of the model capability to suppress the noise components in the signal without distorting the target speech of interest. The same trend can be seen on the COVL score which is a reflection of the overall signal quality. This is even more clear when we look into the segmental SNR gain through the processing. SSNR has been increased by around 1 dB by processing with SE-FFTNet in comparison to the SE-WaveNet method.

The results from the MOS study is displayed in Table 4.2. Though the SE-FFTNet has got higher scores compared to all the other models, the model is slightly under scored compared to the SEGAN.

Table 4.1 – The Objective measurements comparing the performance among the models

Metric	**Noisy**	**SEGAN**	**SE-WaveNet**	**SE-FFTNet**	**SE-InvFFTNet**
PESQ	1.96	2.24	2.23	**2.37**	2.24
STOI	0.28	0.87	0.86	0.87	0.87
CSIG	3.35	3.34	3.33	**3.60**	3.31
CBAK	2.44	3.09	3.00	**3.20**	3.13
COVL	2.63	2.78	2.76	**2.98**	2.77
SSNR	1.63	9.18	8.12	**9.65**	9.61

Table 4.2 – MOS with standard error for different methods

Noisy	SEGAN	SE-WaveNet	SE-FFTNet	Inv-FFTNet
2.67±0.12	3.51±0.09	2.8±0.10	3.27±0.10	2.91±0.09

The reason behind SE-FFTNet performance improvement might be attributed to the initial hypothesis we have mentioned, where the initial wider dilation of the proposed SE-FFTNet model being enabled a better extraction of the features which could discriminate the noise on the input. By this assumption, the SE-FFTNet should outperform the SE-InvFFTNet. From Table 4.1, all the readings show an inline relation to our assumption. The CSIG gain from 3.31 to 3.60 is a strong sign of target speech restoration by the SE-FFTNet model compared to SE-InvFFTNet. At the same time, noise suppression (CBAK) has been improved from 3.13 to 3.20. Hence the overall quality of the output speech (COVL) is got improved by 0.21. A similar trend can be observed in the MOS test results displayed in Table 4.2. This is a clear indication that the model with decreasing dilation fields (SE-FFTNet) performs better than the one with increasing dilation (SE-InvFFTNet) for the speech enhancement task. This validated the hypothesis on which the model was built. The enhanced samples from all these models can listen from this link [1].

In the real-time application of these neural network based speech enhancement algorithms, the parameter complexity is the biggest constraint. In Table 4.3, we have listed the number of parameters used in SEGAN, SE-WaveNet, and SE-FFTNet models. The parameters displayed is the *testing* complexity of the model. Note that the training of model like SEGAN needs additional parameters for the discriminator network. From Table 4.3, it is clear that the suggested SE-FFTNet model has a far lesser number of parameters compared to others; 32% lesser than the WaveNet and 87% lesser than the SEGAN. This reduction in parameter further highlights the potential of the proposed model towards real-time enhancement applications. One must note that this reduction is accompanied by the performance equal or higher, compared to the existing models.

Table 4.3 – Total number of model parameters in Million (M)

SEGAN	SE-WaveNet	SE-FFTNet
193 M	34.3 M	23.5 M

[1] https://www.csd.uoc.gr/~shifaspv/

4.3 Perceptually trained SE-FFTNet

In the above formulation of the speech enhancement task, the training of SE-FFTNet was performed in the time domain with minimizing the deviation of predictions from the clean speech samples at the output of the network. However, speech sounds are perceived in a frequency selective manner at human ears where sound pressure variations are analyzed with cochlear filter banks. Therefore, training neural networks with optimizing a frequency selective loss would be more appropriate for enhancement models. To meet this requirement, we modified the training pipeline of SE-FTTNet architecture to extract the frequency information of predicted samples. The frequency information is extracted with the integration of the short-time Fourier transform (STFT) module as a post-processing stage to the time-domain architecture. A cost function is subsequently defined on the STFT representation. This extension to the new training pipeline for SE-FFTNet with frequency domain objective is illustrated in Figure 4.5. The resulting optimized model is called SE-FFTNet(f). Note that the SE-FFTNet(f) still generates samples in the time domain, while the spectral analysis is performed only to optimize model parameters while training.

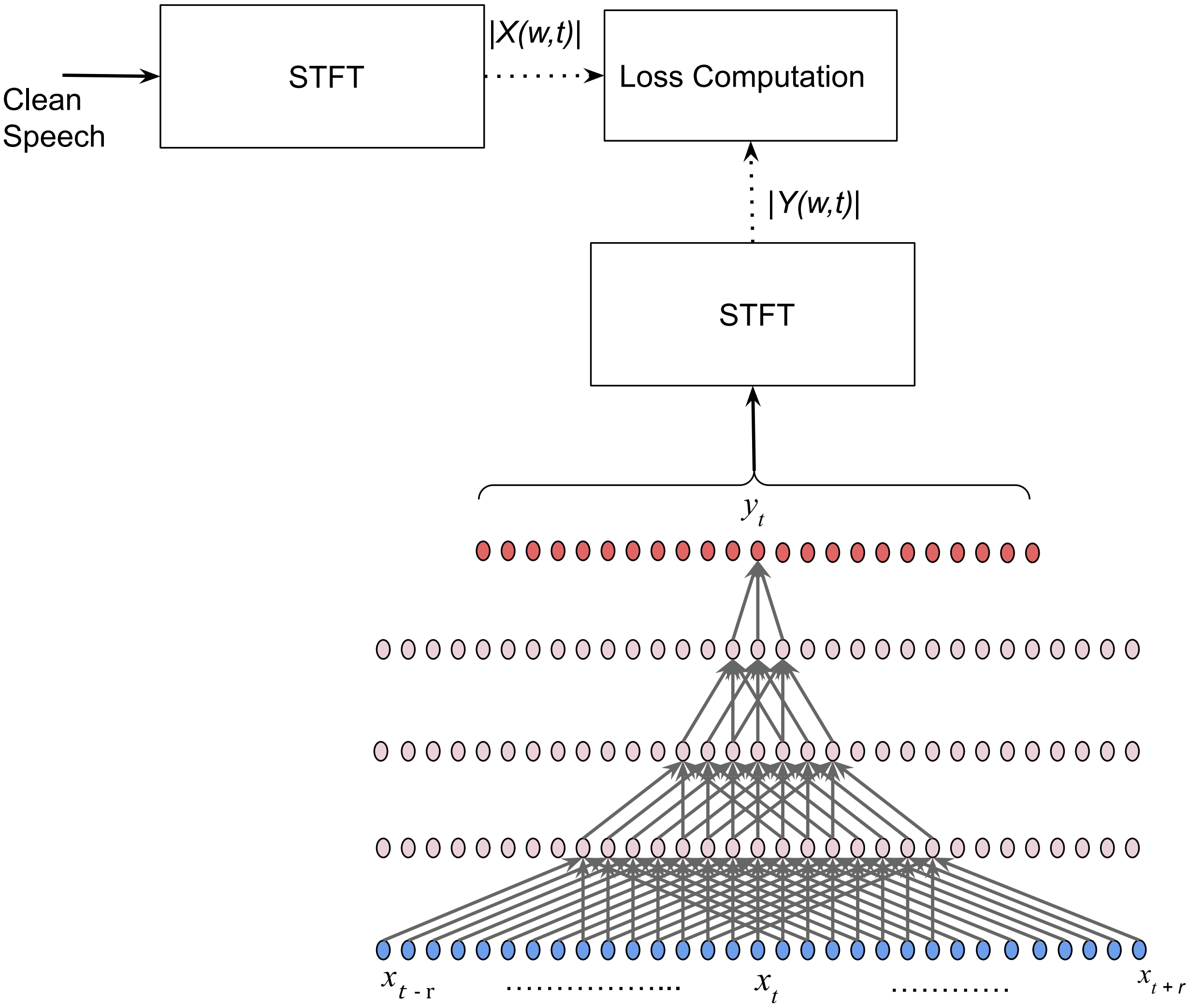


Figure 4.5 – Training framework of SE-FFTNet(f). Here, the STFT denotes the short-time Fourier transform.

Loss Function: Selecting an appropriate loss function in the spectral domain is important for having a robust model returned. Although the frequency spectrum contains both magnitude and phase information of speech, we only considered the magnitude of spectral representation for the obvious reason that the phase would not contribute much to speech intelligibility. As such, for the k- th clean training utterance y^k, we compute the STFT to extract the magnitude spectral $|Y^k(t,f)|$. The STFT of model prediction $\hat{Y}^k(t,f) =$ STFT(SE-FFTNet(x^k)) is computed once the entire speech samples are processed, where x^k is the corresponding noisy segment in time domain. Subsequently, the spectral domain objective function for the k- th training utterance is defined as the $l1-$ norm:

$$L^k(y^k, \hat{y}^k) = \sum_{t=1}^{T^k} \sum_{f=1}^{N} ||Y^{(k)}(f,f)| - |\hat{Y}^k(t,f)|| \tag{4.3}$$

where, T^k is the number of frames in the k-th utterance and N is the number of spectral bins. When training the gradient will back propagate through the STFT module, thereby, network parameters will be tuned to minimize the frequency distortions. Once trained properly, the STFT module is removed so that samples are generated in the time domain.

Data Set: To evaluate the proposed model, 30 speakers were selected from the Voice Bank corpus [VYK13]. Out of these, 28 speakers were used for training and each speaker data consists of around 400 sentences. To create the noisy mixture, each of these files has been chosen randomly and mixed at a specific SNR point from [0, 5, 10, 15] dB with a selected noise type from the noise set that contains 10 different real-life noises. The different type of noise is selected from DEMAND database[TIV13]. The remained two speakers were used for testing with the same type of noises used in training but with 4 different SNR level falling in [2.5, 7.5, 12.5, 17.5] dB. For performance comparison, two recently proposed waveform domain speech enhancement models, SEGAN & SE-WaveNet, together with SE-FTTNet trained on the time domain are considered. In the discussions follow, SE-FFTNet(t) denotes the FFTNet trained with time domain objective function.

Procedure: All these models were being trained in a speaker independent fashion. The output speech quality is evaluated objectively using common evaluation metrics. The perceptual evaluation of speech quality (PESQ) is used as an objective measure of naturalness [HL07]. The Short-time objective intelligibility (STOI) score is used to measure the intelligibility gain by processing the noisy mixture, in reference to the clean [THHJ11]. Finally, the spectral distortion between the model prediction and the clean samples is captured by the Log spectral distortion (LSD) metric [Loi13]. Since the samples' quality of the model prediction was optimised in spectral domain with paying attention only to the magnitude spectrum, there had been a phase mismatch between the true signal and predictions that is not perceptually notable. However, this has limited the use of other objective metrics for analysis that had been used in the previous experimentations.

4.3.1 Evaluation and results

The performance testing is done with 824 samples from the test set comprised of different noises. Hence, the results displayed in Table 4.4 are the average performance on the test set. We compare the objective performance gain of the proposed SE-FFTNet model along with its competitors. Perceptually (in PESQ scale), the SE-FFTNet model enhanced the speech quality about 14% over the existing SEGAN and SE-WaveNet architectures. Although there is not any noticeable change in the STOI intelligibility, the LSD which measures the spectral distortion has recorded minimum value for the sample generated by the SE-FFTNet. This indicates that the SE-FFTNet was able to restore maximal spectral information of speech with the new training pipeline even as compared to the SE-FFTNet model trained in time domain (SE-FFTNet(t)). This further highlights the advantage of optimizing parameters of time domain models with perceptually relevant objective functions.

To better understand the accuracy of signal recovery by different models, spectrograms of a test set sample enhanced by various models are plotted in Figure 4.6. First, when we look at the noise suppression of SE-WaveNet, it has not been as good as the SEGAN or SE-FFTNet(f). While SEGAN has suppressed the noise very well, some essential speech portions have also been removed in the process. Further, the SEGAN introduces high-frequency processing artifacts on the speech. In contrast, the SE-FFTNet(f) has well suppressed the noise events in the spectrum without affecting even minute structures of the target speech. This might be attributed to the new training framework with spectral objectives together with the wider dilated structure of SE-FFTNet neural model. Enhanced samples from different models are

Table 4.4 – Objective measurements comparing the performance between models.

Metric	Noisy	SEGAN	SE-WaveNet	SE-FFTNet(t)	SE-FFTNet(f)
PESQ	1.96	2.24	2.23	**2.37**	**2.54**
STOI	0.28	0.87	0.86	0.87	0.87
LSD	1.48	1.17	1.22	**1.13**	**1.04**

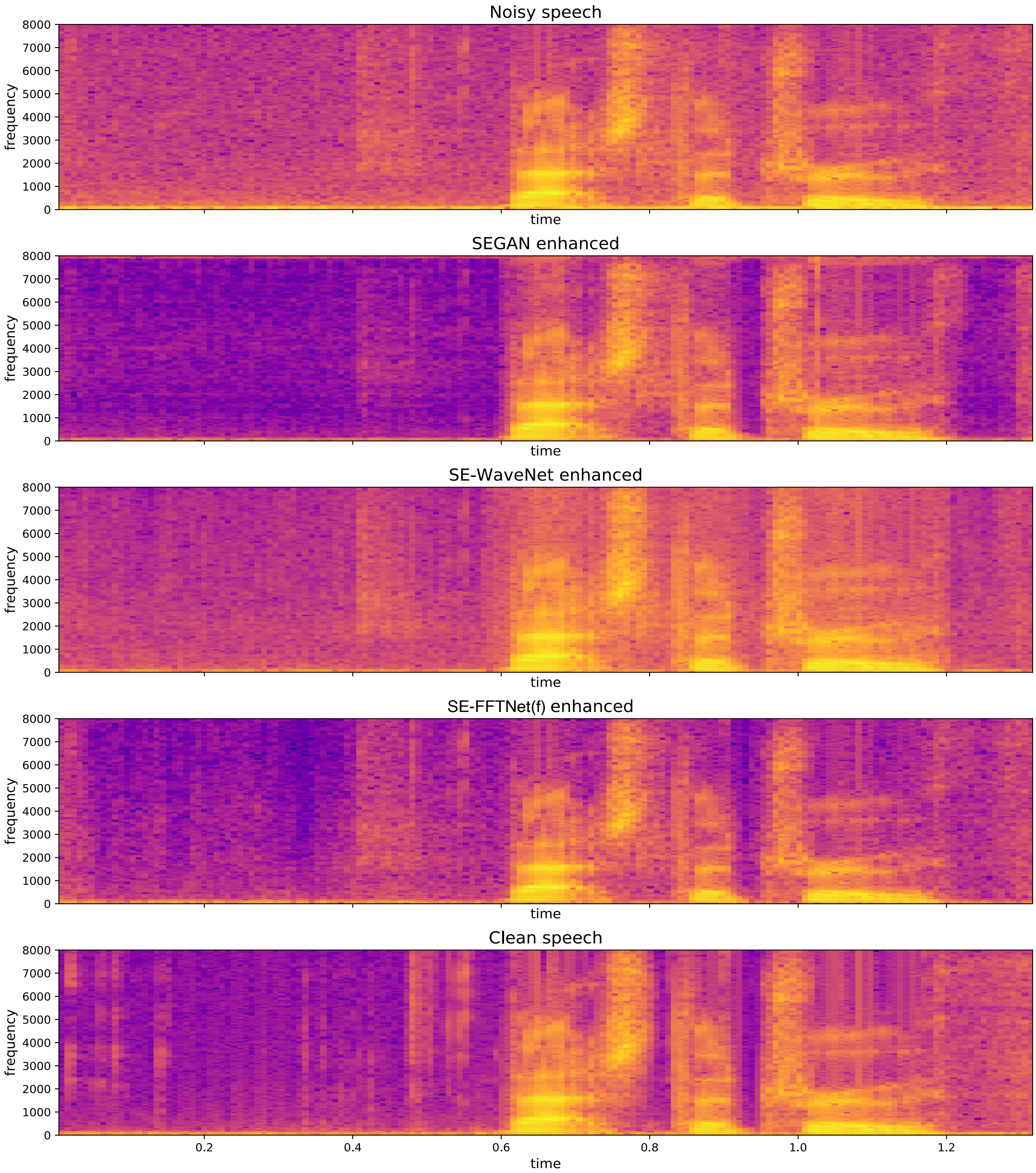


Figure 4.6 – Magnitude spectrograms of the enhanced speech by various models.

presented at this link [2].

[2] https://www.csd.uoc.gr/~shifaspv/

4.4 Conclusions

In this chapter, we have discussed the importance of phase information of speech. The quality of reconstructed speech can be enhanced by stitching a correct phase estimation in enhancement models. However, the disparity in the estimation of phase in STFT domain can deteriorate the reconstruction quality. Therefore, a novel architecture to enhance the speech in its raw representation is proposed. The proposed SE-FFTNet model takes noisy speech at a specific sample rate and returns the same resolution output. The dilated convolution layers were used in SE-FFTNet architecture to reduce the parameter complexity in modeling the long-term dependency in the time domain. In contrast to the standard dilation patterns that have been followed in building models like WaveNet, SE-FFTNet followed an alternative pattern. In the experimental evaluation, the SE-FFTNet was observed to produce better quality enhancement compared to other waveform architectures for speech enhancement. Follow-up experimentation on the new dilation pattern brings the conclusion that the new dilation pattern (dilation rate decreasing over the layers) helps better discriminate noise from speech.

Subsequently, a perceptual weightage was introduced to the training of SE-FFTNet. It was motivated by the fact that speech is perceived in a frequency selective manner at the cochlear level of the human ear. Such a frequency selective loss function have found to further improve the quality of enhanced signal. However, using the magnitude of STFT spectrum as the loss was found to result to phase reversed speech at the output of the network. This does not damage the perceptual quality as it was consistent over time.

So far we have discussed the techniques to suppress the noise on speech in order to restore its quality and intelligibility to the standard of human speech production. However, naturally articulated speech may not be very intelligible in many real-world listening scenarios like in a noisy environment. Therefore, modification of natural speech structure to enhance the listening experience in adverse conditions is required in practice. Approaches to enhance the listening in sub-optimal acoustic conditions are discussed next. Findings from this chapter, especially the noise robustness of SE-FFTNet architecture, are being revisited in the subsequent chapters.

Part II

Listening Enhancement

Chapter 5

Introduction

The intelligibility of a speaker is determined by factors such as the strength of articulators and the acoustics in which the speech is produced. For instance, it is commonly observed that female voices are more intelligible than males, therefore are preferred in most applications [YBBW19]. Similarly, speech presented in noise, e.g. announcement at airports, communication in military combat, is harder to understand and requires extra cognitive effort due to the noise masking at the auditory periphery. In face-to-face communication, human speakers were observed to adopt different articulatory strategies in self-perception that the interlocutor's comprehension is being affected (by noise or hearing disability) in order to ease the listening task – a phenomenon often referred to as *the listening speaker*. For instance, speakers were observed to produce clear articulations – with increased separation between vowel categories and more poses – when speaking to people with hearing disabilities in comparison with casual speech produced in conversations. Studies have shown that clear speech (speech articulated with deliberation) is more intelligible to normal and hearing-impaired listeners than casual speech in noise [PDB86, GKS14].

In addition to interlocutor-induced modifications, humans also adjust articulations adaptive to environmental factors resulting in Lombard speech [Jun96] which is generally produced in the realization that the perception is being affected by noise in the background or distortions in the channel. Perceptually, Lombard speech can be described as 'tense' and 'loud' with exaggerated articulation compared to the normal speech produced in quiet. Even after removing the loudness difference of the signals, Lombard speech is still observed to be more intelligible in noise for both native and non-native listeners [Lu09, CL12], indicating the existence of powerful acoustic cues that are linked to the intelligibility of speech. Many studies in the past examined the acoustic and phonetic features of Lombard speech. It is found that the Lombard speech exhibits higher fundamental frequency (F0), a shift in vowel space, slower speech rate and flatter spectral tilt compared to the plain speech [SPB$^+$88, GKS14, GBD$^+$06]. Investigating the individual contribution of these factors to the Lombard intelligibility has produced no clear conclusion for F0, vowel space alteration and duration increase [LC09, CMV14], however, noticeable effects on intelligibility were observed when the inclusive formant region is boosted, i.e., flattening the spectral tilt [GKS14]. This increase in intelligibility is explained by the increase in speech glimpses (the region of speech dominance over noise) in the mid-frequency by the boosting together with the high sensitivity of human ears to the 500 Hz to 4 kHz frequency band where most formants reside. In a study by Bosker and Cooke [BC20], it is observed that the Lombard style exhibits enhanced amplitude modulations, and therefore transplantation of Lombard modulation to plain speech is found to improve the intelligibility in noise.

Both the clear and Lombard styles are elicited to communicate the message intelligibly to a listener, wherein the speaker simultaneously acts as a listener analyzing the auditory scene of the interlocutor. However, in many real-world scenarios such as in mobile communication, the speaker would not have reliable feedback about the listener's acoustics. Therefore, machine intervention would be essential to facilitate communication and deliver the message effectively to listeners in such scenarios. Several targeted approaches were suggested in the past with inspiration from human articulatory adjustment to improve the intelligibility in noise. Both spectral and temporal manipulations of speech have been explored in the past [Gri68, NG76, SV06, HF10, RVD09, ZKS12, SRD15]. Since the high-frequency components are more vulnerable to noise and they play an important role in speech understanding, the earliest enhancement systems were simple high-pass filters [Gri68], which were later improved using amplitude compression [NG76]. Energy reallocation over time from sonorant speech segments to those less energetic by means of dynamic range compression (DRC) was proven to work well in many noisy scenarios [RVD09, ZKS12, SRD15]. A thorough subjective evaluation of most promising intelligibility modification approaches aimed for listening enhancement in noise has been performed during the Hurricane Challenge [CMVB13a, CMVB$^+$13c]. The evaluation was conducted under the constraint that the overall signal energy is the same before and after the signal modification. It was found that the best performing modification models produce an equivalent intensity change – the amount in decibels that unmodified speech would have to be adjusted to achieve the same intelligibility of modified speech – of up to 5 dB.

However, conventional intelligibility enhancement techniques are statistically based on the detection and modification of acoustic features of speech with the use of statistical tools. Besides, they assume that the speech to be modified is noise-free, which means they are acquired in acoustic isolation or studio condition. This is a wide generalization that can not be guaranteed in most practical scenarios as speech in the real world is often produced in a background of noise. For instance, in a mobile communication scene, one could easily imagine a situation where both the speaker and listener are immersed in noise. When the speech is not clean and is being enhanced for intelligibility with methods designed to work with clean signals, the modified signal may have worse intelligibility than the unprocessed one. Therefore, noise-robust speech modification techniques for intelligibility are very important in practice. So far, this problem of *listening enhancement (LE)*, modification of speech for enhancing the intelligibility of listeners in noise, was looked at mainly from a statistical framework. Whereas, the advent of neural networks in recent years has wide opened the arena to solve the challenges faced by current LE models.

As such, this chapter is dedicated to investigating the feasibility of using neural networks for intelligibility modification. This is experimented with building deep neural architectures and tuning its parameters to learn the feature modifications which are reported to be critical for improved intelligibility. Since this is the first attempt towards building neural-based speech modification techniques, the current chapter only addresses the intelligibility enhancement of clean (noise-free) speech, such that the findings from this chapter will set the stage for the next chapter where the noise robustness of such a system is tested.

5.1 Factors defining speech intelligibility

Speech understanding is conditioned on numerous factors such as the clarity of articulation, sensitivity of the listener's ear, the language proficiency of both speaker and the listener, the quality of communication scene. In self-perception of the intelligibility loss, human speakers adjust their articulation to ease the communication. Two major categories of speech production changes are identified (1) based on the interlocutor, e.g, foreign directed speech (FDS) or machine directed speech (MDS), (2) based on the communication environment, e.g, Lombard speech (LS) or speech addressed to a distant listener. There are some unique features associated with each of these categories. For instance, the LS is produced with an intense vocal effort to maximize the audibility of speech in noise, in which the clarity would have been lost, whereas the FDS is often seen as with reduced lexical variability and more clear intonations [UKB07, FAFZ12]. However, the categorical division between them is not absolute as such they share many acoustic and phonetic features.

Although there would be intra-speaker variability in each speaking style, studies in the past have reported clear acoustic and phonetic changes associated with individual production change. Therefore, this section presents a brief summary of the acoustic-phonetic changes observed in both interlocutor-induced and environment-induced modifications and their effects on the listener's intelligibility.

Lombard and Clear styles exhibit decreased speaking rate, therefore enlarged durations of phones over their casual / normal counterparts [PDB86], [Lu09]. The decrease in speaking rate is also associated with more number pauses and an increase in the duration of different sound segments. In the work of Bradlow et.al [BKH00], an overall increase in sentence duration of 51% and 116% respectively for male and female speakers in casual to clear was observed. To test this feature, Cooke et.al [CMV14] has artificially modified the durations of speech and evaluated their impact on the listener's intelligibility; no clear benefits were observed with durational modification. Even so, it is understandable that speaking slowly with more pauses will offer the time to process and comprehend the message effectively for at least some listeners.

Analysis of the energy difference between consonants and vowels reported an increased consonant-to-vowel energy ratio (CVR) in clear speech over casual [BKH03]. Hazan and Markham [HM04] experimented with the correlation of CVR and intelligibility. They have not seen any significant correlation between word intelligibility and CVR for nasals, fricatives, and stop consonants in naturally-produced speech. However, it has been shown that enhancement of the consonant energy in words, consonant-vowel syllables (CV) and vowel-consonant-vowel (VCV) syllables improve consonant identification for normal hearing listeners [HS98, GS86] and hearing-impaired listeners [GS87, ME88]. Motivated from this, Skowronski and Harris [SH06] performed an energy redistribution from vowels to consonants which were shown to improve the intelligibility of speech. Similarly, Godoy and Stylianou [GS12] evaluated the contributions of voiced and unvoiced regions to the Lombard intelligibility, and showed that the Lombard increase in intelligibility is primarily attributed to the vowel segments of the signal.

The short-term spectra analysis (STSA) unveils the information of speech in various frequencies at different time points. STSA revealed higher spectral prominences associated with vowel sounds in clear speech. On this assumption, Krause [Kra01] amplified magnitudes of the first and second formants of casual voice segments to match the spectral characteristics of clear speech. This formant sharpening was found to enhance the intelligibility of casual speech for normal-hearing listeners, but not for the hearing impaired [Kra01]. These differences in spectral information between casual and clear speech were further analyzed in [KS14], and presented a mixed-filtering technique to isolate the information from clear speech and add it to the casual speech to improve the intelligibility, which was found beneficial in improving the intelligibility in noise. The analysis of long-term average spectrum (LTAS) – the spectral information averaged over time – revealed an increase of energy in the frequency region spanning formants for the Lombard style compared to normal articulation, resulting in a reduction in spectral tilt. A similar, if not that prominent, tendency was also observed in the case of clear speech compared to the casual conversational speech [HM04]. Godoy et.al [GKS14] investigated further the influence of the relative spectral amplitude difference of different styles on speech intelligibility. Fig. 5.1 shows the LTAS difference between Lombard to normal and clear to casual styles. As observable, the Lombard speech exhibits a clear boost in average energy in the 500 – 4500 Hz band over its normal counterpart, while the exaggeration of spectral content is mild in clear speech compared to the casual style. This migration of spectral energy from the low and high bands to the mid-frequency range must be attributed to the increased intelligibility of the two styles. Investigating this, Lu and Cooke [LC09] artificially redistributed the spectral energy of normal speech to match that of Lombard, thereby reducing the tilt of the overall spectrum. This reduction of spectral tilt was found to improve speech intelligibility in noise.

Vowel sounds can be classified based on the position of articulation. The positioning of articulators is reflected as the formants of each vowel sound, mainly as the first (F1) and second (F2) formants, therefore constituting the two-dimensional vowel space. Analysis of correlation between vowel space and intelligibility revealed that the speakers with larger vowel space are more intelligible than the speakers with reduced vowel space [HM04, BTP96]. Specifically, speakers with a wide F1 range appeared to produce high intelligible speech. Whereas the F2 range was found to be significantly correlated with the sentence intelligibility [HM04] and little to the word intelligibility [BTP96]. Studies by Godoy et.al [GKS14] have also revealed a vowel space expansion in clear style compared to the casual, while this expansion was not observable in the case of Lombard style. However, the Lombard style seems to produce a consistent increase in F1 frequency resulting to a shift in vowel space. Motivated from the observations of vowel space expansion in clear speech, frequency wrapping techniques to achieve vowel space expansion were tested in [MKS12, GKS14], however, there were no benefits to intelligibility. On the other hand, since the formants and their transition play an important role in perceiving and classifying different sound segments, sharpening of formants with statistical approaches has been found helpful to improve the intelligibility in noise [ZKS12].

A change in the speaker's fundamental frequency (f0) is also observed in the case of Lombard speech [SB06]. This variation may be contributing to the improved intelligibility of the style. However, the modification of f0 characteristics of normal speech to match that of Lombard was not found to improve the word recognition intelligibility in noise for

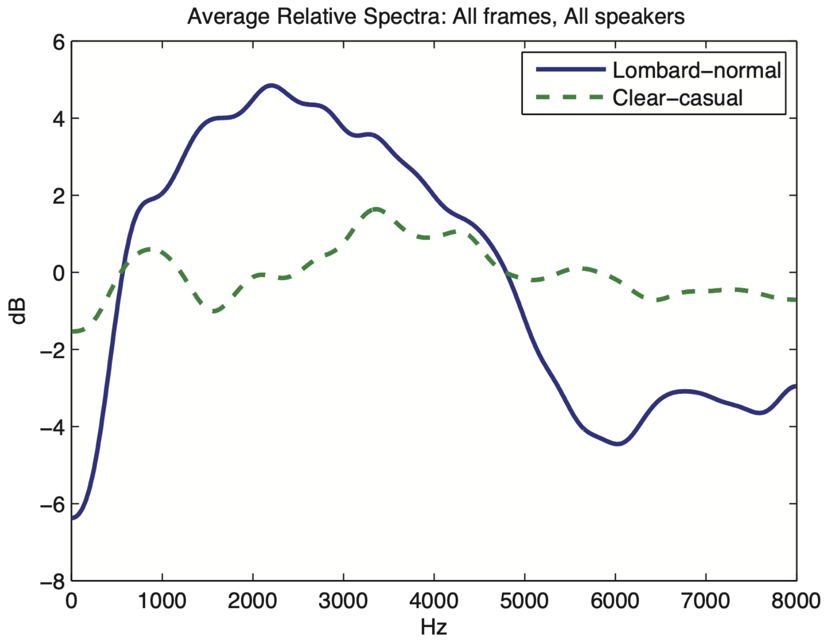


Figure 5.1 – Average relative spectra for all frames (from Godoy et.al [GKS14]).

normal listeners [LC09]. Besides, artificial flattening of f0 was reported to degrade the intelligibility [LB03, WS08], therefore, the real impact of F0 on speech intelligibility is still unclear.

Speech as a real-valued signal can be decomposed into a set of amplitude modulated (AM) signals with the carrier frequencies falling in the signal bandwidth [DFP94]. The temporal variation of such AM components is called the modulation of speech over time. Studies on clear speech revealed a higher modulation depth for the temporal envelopes [KB04, LZ06], therefore, appears to correlate with the clear speech intelligibility advantage. The study in [DFP94] showed intelligibility degradation of speech after smearing low-frequency modulations, showing the modulation frequencies falling in 4 to 16 Hz is the most relevant for intelligibility. Modulation index metric is used to quantify the modulation depth of the temporal envelopes [HS85] and therefore traditionally has been used as a benchmark measure of speech intelligibility in noise and reverberant conditions. This argument is even supported by the studies in neuroscience showing that speech is decomposed at the auditory cortex as spectro-temporal modulation content and the perception is driven by the sounds that best combine both the temporal and spectral modulations [MS05, SZ09, KDS96]. Consequently, modulation domain processing of speech for applications like noise reduction and echo cancellation was proposed and found to better isolate the masking components [WL12, SJS18, JSS16]. The transplantation of enhanced amplitude modulation in a clear speech to casual speech has been shown to contribute to improving intelligibility in noise [KS16]. In a recent study, Bosker and Cooke [BC20] observed the same trend in Lombard speech – enhanced modulations in the frequency range 1 – 8 Hz. Subsequently, they have shown that the transplantation of Lombard amplitude modulation onto plain speech produces additional intelligibility benefits, underlining the importance of amplitude modulation for speech intelligibility.

5.2 Spectral shaping and dynamic range compression (SSDRC)

Inspired by the intelligibility benefits of various speaking adaptations, artificial modification of speech to improve its intelligibility by altering the acoustic features has been recommended. Among the multitude of features contributing to intelligibility, spectral energy redistribution and increasing consonant-to-vowel ratio with dynamic time-domain energy reallocation were found to contribute largely to the intelligibility benefits in noise [RVD09, GKS14]. A combination of spectral shaping (SS) and dynamic range compression (DRC) was proposed in the work of Zorila et. al [ZKS12] as the SSDRC algorithm. SSDRC was tested in various listening settings on different languages since its introduction and has been found to produce the best intelligibility benefit in noise for normal and hearing-impaired listeners [CMVB13a, CMVB+13c, ZSFM17, SSCS20]. Therefore, we consider the SSDRC style over many natural styles as a reference for our research for the same reason that it produces the highest intelligibility. Since the feature modifications elicited by SSDRC are used in the neural network architectures in the following chapters, a brief description of the SSDRC algorithm must be informative at this stage. SSDRC performs a two-stage processing of speech to increase its intelligibility; 1) spectral shaping in the frequency domain, 2) dynamic range compression in the time domain.

5.2.1 Spectral shaping (SS):

As the first stage of the enhancement framework, the SS module is an adaptive spectral shaper in the Fourier domain. The main purpose of which is to provide the 'crisp' and 'clean' quality to speech with sharpening formants as they are the most important acoustic cue for speech perception; therefore, increasing the intelligibility even in quiet listening conditions. The whole process is done in adaptive to the voicing probability.

The module takes plain speech $x(t)$ as input, in frame-based processing (frames are of fixed duration), performs Discrete Fourier Transform (DFT) on each frame to obtain the magnitude spectral components $X(w, t)$, The shaping is done in adaptive to the voicing probability in order to avoid processing artifacts in fewer sonorant areas such as fricatives. The probability of voicing is computed with the equation

$$P_v(t) = \alpha \frac{rms(t)}{z(t)} \tag{5.1}$$

where $\alpha = 1/\max(P_v(t))$ is a normalization constant, and $rms(t)$ and $z(t)$ is the RMS value and zero crossings of the segment, respectively, for a window centered around the instant t with the length of 2.5 times the fundamental period (8.3 ms and 4.5 ms for male and female voices, respectively).

For each DFT frame $X(\omega, t)$, the SEEVOC spectral envelope estimator [Pau81] is used on the magnitude spectrum to get the envelope estimate $E(\omega_k)$. Then, the tilt $T(w)$ of the spectral envelope is computed as

$$\log T(\omega) = c_0 + 2c_1 \cos(\omega), \tag{5.2}$$

where the variable c_m denotes the m^{th} cepstrum coefficient computed as

$$c_m = \frac{1}{N/2+1} \sum_{k=0}^{N/2} \log E(\omega_k) \cos(m\omega_k). \tag{5.3}$$

Therefore, the final adaptive spectral shaper has the transfer function function (over frame instance t)

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)} \right)^{\beta P_v(t)}. \tag{5.4}$$

In this way, the formant inclusive regions of voiced spectra are sharpened by selectively isolating the unvoiced segments with parameters $P_v(t)$. The variable β was set to 0.25 in most cases.

Since studies in the past have shown that pre-emphasizing of spectrum above 1100 Hz contributes to the intelligibility advance in noise [NG76], an adaptive pre-emphasize filter is used as the second spectral shaping filter. Since doing such filtering over all the segments of speech would introduce noisy quality to speech, an adaptive pre-emphasis adapted to the probability of voicing is used with the transfer function

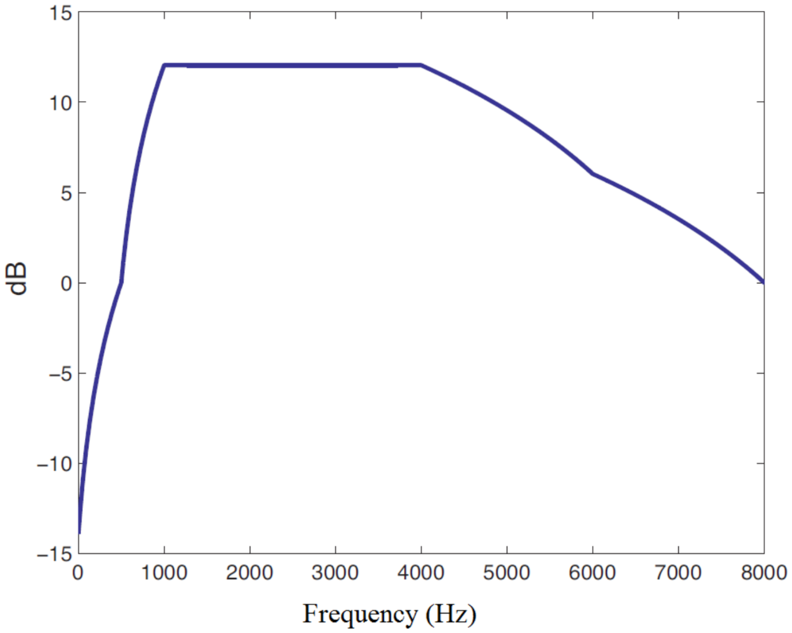


Figure 5.2 – Spectral shaping fixed filter

$$H_p(\omega,t) = \begin{cases} 1 & \omega \leq \omega_0 \\ 1 + \dfrac{\omega-\omega_0}{\pi-\omega_0} g P_v(t) & \omega > \omega_0 \end{cases} \tag{5.5}$$

where $\omega_0 = 0.125\pi$ for 16 kHz sampled speech, and the variable g is selected to 0.3.

Therefore, the adaptive cascaded spectral filtering can be expressed as

$$Y_{aSS}(\omega,t) = H_s(\omega,t)H_p(\omega,t)X(\omega,t). \tag{5.6}$$

Motivated from the spectral characteristics of Lombard Speech, a final fixed spectral gain filter to boost energy in the mid-frequency range of the spectrum was used. This non-adaptive, or time-invariant, filter $H_r(w)$ boosts frequencies in the range 1000– 40000 Hz by 12 dB while reducing the components below 500 Hz by 6dB/octave. This Lombard-inspired filter has the transfer function shown in Figure. 5.2, which matches with the average spectral distribution of Lombard style as shown in Figure.5.1. Hence, the final spectral shaped signal is

$$Y_{SS}(\omega,t) = H_r(\omega,t)Y_{aSS}(\omega,t). \tag{5.7}$$

Inverse Fourier transform with overlap and add technique reconstructs the spectral shaped waveform.

5.2.2 Dynamic range compression (DRC):

Speech signal from the spectral shaping module is amplitude compressed using a dynamic range compressor (DRC). The DRC's objective is to reduce the envelope variations of the signal. This gain of DRC is derived from a desired input/output envelope characteristic (IOEC) curve. The IOEC used in the SSRC algorithm is shown in Figure. 5.3, which has three characteristic zones: unity gain, expansion, and compression.

First, the envelope of the speech signal is computed using the analytic signal with the use of Hilbert transform. The estimated envelope, $e(n)$, is dynamically compressed with 2 ms release time constant and almost instantaneous attack time constant. Specifically using the expression

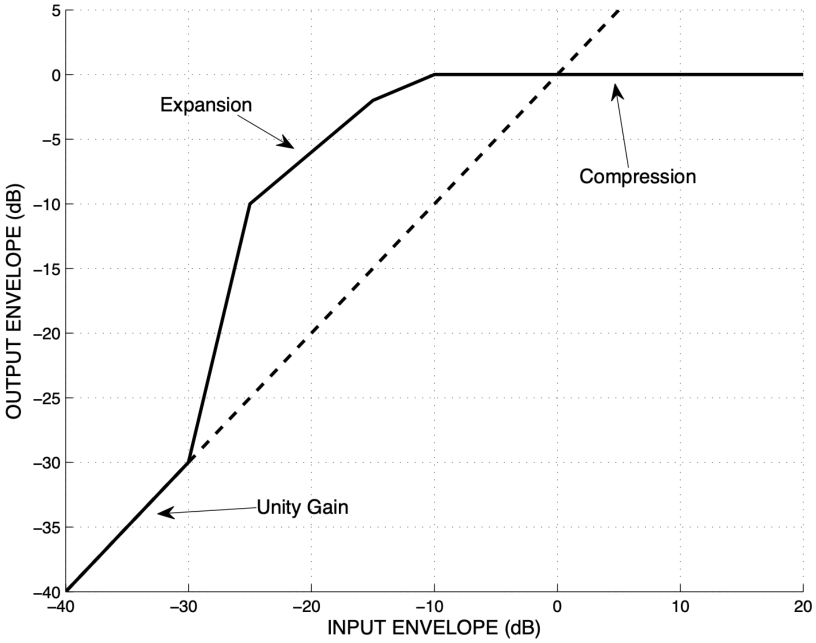


Figure 5.3 – Input-Output Envelope Characteristic (IOEC) Curve

$$\hat{e}(n) = \begin{cases} a_r \hat{e}(n-1) + (1-a_r)\, e(n), & \text{if } e(n) < \hat{e}(n-1) \\ a_a \hat{e}(n-1) + (1-a_a)\, e(n), & \text{if } e(n) \geq \hat{e}(n-1) \end{cases} \tag{5.8}$$

where the time constants are set to be $a_r = 0.15$ and $a_a = 0.0001$.

The 0 dB reference level of the envelope e_0 were set to the 30% of the maximum of the input signal envelope. With this reference value, the input envelope is computed as

$$e_{in}(n) = 20 \log_{10} \left(\hat{e}(n)/e_0 \right). \tag{5.9}$$

The corresponding output level $e_{\text{out}}(n)$ is obtained by projecting $e_{in}(n)$ onto the IOEC curve in Figure 5.3 and the equivalent gain is computed as:

$$g(n) = 10^{(e_{\text{out}}(n) - e_{in}(n))/20}.$$

Therefore, the dynamic range compressed signal would be

$$s_g(n) = g(n)s(n).$$

At the output, the global energy of $s_g(n)$ is rescaled to that of the original unmodified speech to ensure that the loudness is unchanged.

The alteration to a speech segment induced by the spectral shaping and dynamic range compression (SSDRC) algorithm is shown in Figure. 5.4 . It is evident that the dynamic range (the difference between the largest and smallest values) of SSDRC output is lower compared to the original signal. This helps to amplify the low-intensity phones like /p /, /k /, with the cost of high sonorant segments, which makes speech more intelligible as the transient sounds contribute largely to intelligibility while can be easily masked by noise. Note also that both the signals have an equal root mean square (RMS) energy, ensuring the gain in intelligibility is not a result of direct amplification of the signal.

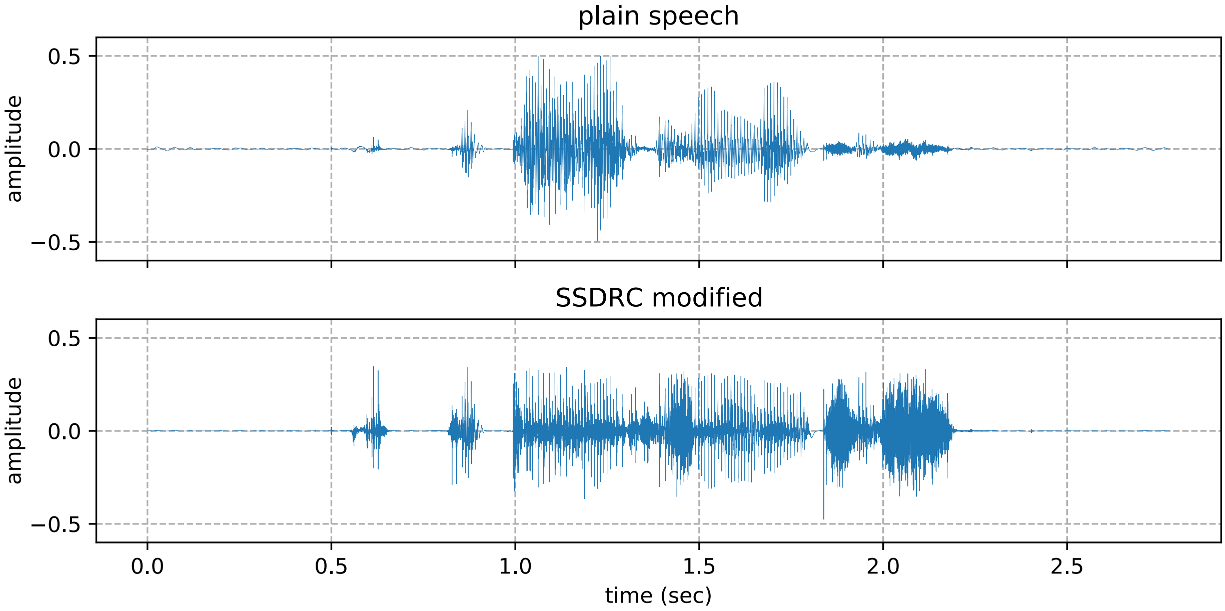


Figure 5.4 – Speech waveform modified for intelligibility with SSDRC algorithm.

Although controlled modifications of speech can help improve the intelligibility for listeners (in noise or at distance), doing so alters the natural modulations of the signal and therefore would degrade the quality (or naturalness) of the signal. As such, parameters such as β in the spectral shaping and the attack (a_a) and release (a_r) time constants in the dynamic range compression of SSDRC have to be selected with care while using it for different applications. Similarly, the variations in pitch-period between the male and female voices will also increase/decrease the quantities such as the window size for computing the voice probability $P_v(t)$, which can affects the quality and intelligibility of the output signal. Choosing the right trade-off between the quality and intelligibility is more of an application specific assignment – effortless message understanding is important at high noise conditions while high quality is preferred in quiet. Such a debate goes beyond the scope of this thesis. Therefore, we limit ourselves at extremely low SNR listening scenarios, where the intelligibility (or message understanding) is the most requisite with little focus on preserving the naturalness.

5.3 Conclusions and perspective

In this chapter, speech intelligibility was defined in detail. Natural speech lacks many characters which are important for high intelligibility, like flatter spectral tilt or more sharpened formants. Humans have been observed to adapt their articulation in various challenging conditions, producing Lombard voice in noise, or Clear speech when speaking to listeners with disability. We also adjust the lexical variability in communicating the message based on the interlocutor's language proficiency. An analysis of various natural speech styles has been conducted and their characteristic features were identified at the beginning of this chapter. Subsequently, modification of speech to artificially boost the intelligibility of plain (clean) speech was also discussed. Although multiple algorithms have been proposed and tested successfully to improve the intelligibility, a combination of spectral shaping (SS) and dynamic range compression (DRC) have consistently performed better. This must have been because of the DRC which redistributes energy from vowel segments to consonants, thereby increasing the consonant-to-vowel (CVR) ratio. The SSDRC method in [ZKS12], which has a phone adaptive spectral shaper and dynamic range compressor, was reviewed extensively. The SSDRC is still the most successful model for intelligibility enhancement. However, it has (as with all other models) limitations while operating in noisy (at input) acoustic conditions. To this end, we investigate the usability of neural networks (NNs) to perform the intelligibility enhancement of speech in the following chapters. If successful, we can later integrate the noise robustness of NNs with intelligibility enhancement.

Chapter 6

Neural Based Intelligibility Modification

So far, the intelligibility modification of speech was addressed in the framework of statistical modeling. However, with the recent advent of neural networks together with the availability of large-scale recordings, we are provided with the unique opportunity to learn the representative features of different styles from data. This style conversion of speech in speaking devices such as TTS was tried in the past where the Lombard style was learned from data with transfer learning technique [BJA19b]. However, since a multitude of factors is linked to the perceived intelligibility of a speech for a listener, designing individual models for each style to promote intelligibility would not be optimal in reality. Instead, we must utilize the progress made in the intelligibility research so far when developing optimal modification strategies, as it is a combination rather than an individual style that contributes to the intelligibility. It is because of this fact that many artificial styles, like SSDRC, are observed to produce higher intelligibility benefits compared to natural styles like Lombard or Clear articulation [CMVB+13c, GKS14].

Moreover, our motivation to use neural networks for speech modification has another dimension as well, which is to use the noise robustness of the network to enhance the intelligibility of corrupted speech. The existing statistical modification techniques, including SSDRC, can execute the intelligibility enhancement task only when the speech is clean of noise. The presence of noise at the input can severely degrade the performance of these systems. Therefore, a neural processing perspective to the problem would give us a more generalized framework as neural models have been proven to be more robust to input noise compared to the traditional signal processing approaches. Besides, having a neural speech enrichment module would also ease the effort to integrate the intelligibility factor into state-of-the-art neural-based devices, like neural text-to-speech (TTS) systems.

With those targets in mind, I started designing neural models for intelligibility modification. There are two factors that defines the performance of a neural intelligibility enhancement system; 1) the powerfulness of the model to capture speech acoustic variations, 2) the target intelligibility style to be learned by the model while training. Although smaller feature domain networks, either CNN or LSTM, can be used as the model architecture, the initial research experiments had showed that such models were not competitive enough to learn a target intelligibility style compared to deep waveform models like WaveNet. When come to the selection of target intelligibility style, there are a list of natural (Clear/Lombard) and artificial (dynamic range compression) intelligible styles that have been reported in the literature. It is often observed that natural styles are not always very intelligible to listen to [CMVB13a], therefore, may not be an ideal style to be learned with a neural network. Learning a style from dark is not an easy task to perform as we are unsure about the ideal acoustic features for an intelligible voice, without which the gradients can not be computed to optimise the model parameters.

Therefore, the only feasible solution was to follow the findings from the past intelligibility studies. Although there are many combinations of acoustic feature modification suggested in the literature, it was a combination of spectral shaping and dynamic range compression that has been found to produce the best gain in intelligibility [GKS14]. As such, the target style to be learned by the proposed models is also set to such a feature combination. Setting labels from a classical digital signal processing (DSP) model would limit the performance in the sense that the neural model can't outpass the intelligibility score, instead enable the intelligibility enhancement of noisy samples.

6.1 WaveNet based SSDRC (wSSDRC)

WaveNet is a powerful generative neural model to synthesize speech/audio samples through a non-linear autoregressive approach [ODZ+16]. A regression extension of the WaveNet was suggested for speech denoising in [RPS18], where the network was trained to learn the mapping from noisy to clean speech. Inspired by the work in [RPS18], with the objective of learning intelligibility features, we design a WaveNet like approach to map plain speech to SSDRC generated signal using a non-causal WaveNet-like architecture. In short, we are looking for the deterministic function that is apt to map samples of plain speech to those (time-domain) samples generated by SSDRC algorithm.

Our motivation for such a sample-based non-linear mapping can be also applied to noisy speech. Then we expect at some higher layers a representation of a cleaner version of the input noisy speech will be available to the subsequent higher layers, which will target their output to be the same as SSDRC-based signals. These target signals have been computed by simply applying SSDRC to the clean version of the input to the network, noisy speech. This might also lead to a better quality of modified speech while still intelligibility is maintained. More specifically we will work with a non-causal WaveNet-like architecture exploring, therefore, the conditional dependencies of the sample generated at the current time step to the future and past samples of the model input. This modeling of sample dependencies is being implemented through dilated convolution structures. We will refer to this new model as WaveNet-based SSDRC, or shortly wSSDRC.

WaveNet [ODZ+16] is a powerful generative approach for the probabilistic modelling of raw audio, which is based on the assumption that speech/audio is a Markov process where the conditional probability for a sample, x_t, given the r previous samples is given by:

$$P\left(x_t \mid x_{t-1}, \ldots, x_{t-r}\right) \tag{6.1}$$

The WaveNet generates samples in a way to maximize these conditional probability terms. This conditional mapping has been implemented as an autoregressive network with a stack of residual blocks, Figure. 6.1, where each block contains expert and gate followed the one-dimensional dilated causal convolution. The output of the expert and the gate are being combined via element-wise multiplication. Block, i, computes hidden state vector $h(i)$, Eq. (4), which then being added (due to the residual connections between layers) to the input after a one dimensional convolution, to generate its output $z(i)$.

$$h^{(i)}=\tanh \left(W_f^{(i)} * z^{(i-1)}\right) \odot \sigma\left(W_g^{(i)} * z^{(i-1)}\right) \tag{6.2}$$

$$z^{(i)}=\operatorname{Conv} 1D\left(h^{(i)}\right)+z^{(i-1)} \tag{6.3}$$

where symbol $*$ denotes convolution and symbol $\odot$ denotes element-wise multiplication.

wSSDRC has two major architectural changes compared to the WaveNet. First, we use the network as a deterministic mapping, f, from input speech $x=\left[x_1, \ldots, x_T\right]$ to an enhanced signal $\hat{y}=\left[\hat{y}_r, \ldots, \hat{y}_{T-r}\right]$. Technically, this is done by removing the final softmax layer and adding a layer which projects the output of the post-processing layers to an one-dimensional signal. Also, the compression of the input signal and its 8 -bit quantization which are important preprocessing steps in the original WaveNet [ODZ+16] are not used in this architecture. Second, instead of considering only the previous r samples of x (receptive of size r), to predict a sample of y at time t, we also consider the next r samples of x, which in essence increased the receptive field size to $2r-1$. Therefore, the enhanced sample at time $t \in\{r+1, \ldots, T-r\}$ is predicted as:

$$\hat{y}_t=f\left(x_{t-r}, \ldots, x_{t-1}, x_t, x_{t+1}, \ldots, x_{t+r}\right) \tag{6.4}$$

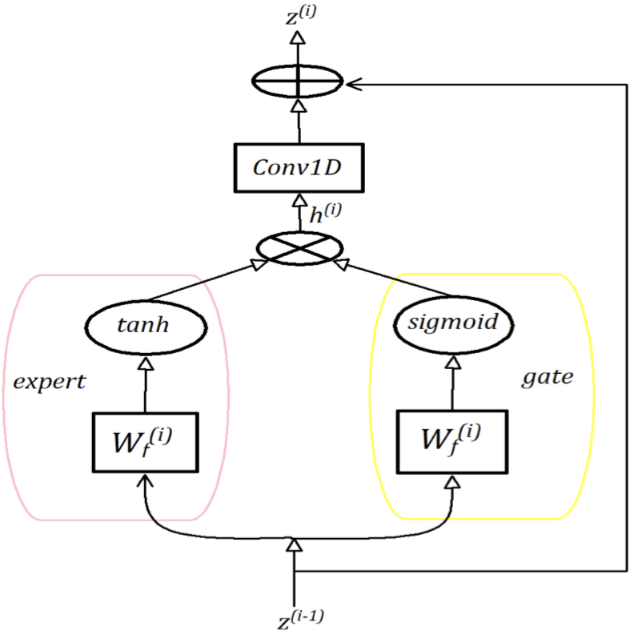


Figure 6.1 – Residual block of wSSDRC / WaveNet.

Figure. 6.2 shows the dependence of the output sample $\hat{y}_t$ on the input samples. As shown in the figure, the dilated convolution structure being used to calculate the activations of the nodes in each block. Which means that the nodes on the i^{th} level in a block ignores the $2^i - 1$ in between samples on the layer below while calculating the response, which is usually been known as the dilation factor of the WaveNet. The skip connections from each blocks are being summed up and processed through a post-processing unit to get the final enhanced samples y_t. The post processing includes two layers of non-causal convolutions having filter width equal to 3 whose output pass through a corresponding ReLU non-linear function, and a one-dimensional convolution, without non-linearity function, which projects to the output one dimensional signal. This model architecture facilitates the generation of set of samples in a single traverse through the structure. When the whole input sequence is available, then all output samples can be computed in parallel.

6.1.1 Teacher-student topology for training wSSDRC

The main factor to customize is the target function (f) and the kind of modification the network is expected to learn, which reflects the articulatory style to be mimicked by the network. One could set the model to mimic natural intelligibility modifications, like in Lombard speech, as long as they satisfy the time alignment constraint stated in Eq. 6.4. However, since multiple studies have shown considerable intelligibility gain of SSDRC-processed speech over Lombard speech [18][9], we decided that the model should learn an SSDRC-style modification. This has been accomplished by setting the SSDRC (signal processing approach) as the teacher-network to expose the neural model (wSSDRC) to the modification style to be learned. Figure 2 depicts the aforementioned teacher-student framework. While training the SSDRC feeds the target labels which the wSSDRC learns to produce alone over time without the assistance from the teacher SSDRC.

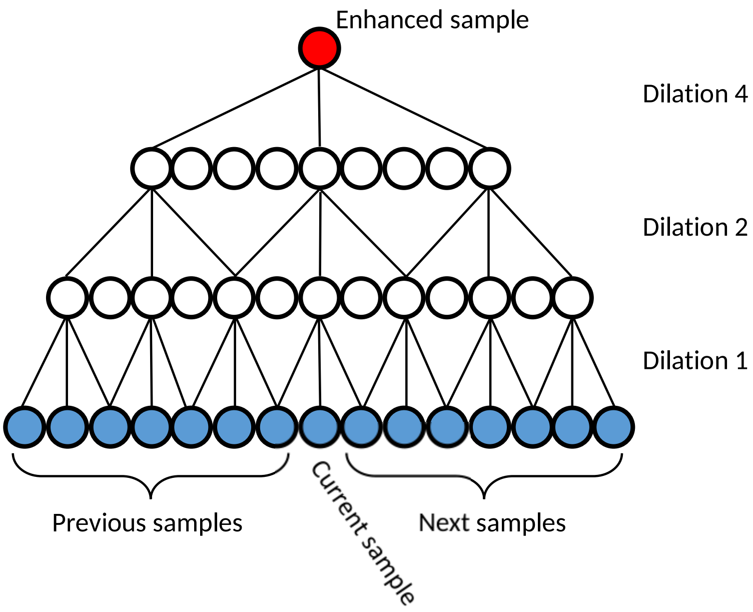


Figure 6.2 – Dilation pattern of the wSSDRC architecture.

Since the model is operating in an end-to-end fashion on the waveform domain, the deviation of the prediction from the target is calculated as the average absolute difference between the predicted sample $\hat{y}_t$ and the target sample y_t. For an input-target wave pair $\left(x^{(k)}, y^{(k)}\right)$, the loss function is computed as:

$$L\left(x^{(k)}, y^{(k)}\right) = \frac{1}{T^{(k)} - 2r} \sum_{t=r}^{T^{(k)}-r} \left| y_t^{(k)} - \hat{y}_t^{(k)} \right| \tag{6.5}$$

where $T^{(k)}$ is the length of signals $x^{(k)}$ and $y^{(k)}$. Therefore, the loss term differs from the actual WaveNet model which had a probability loss function. This is because by removing the final softmax layer from the post-processing stage, we turned the network task to estimate sample error instead of the distribution. The model learns its weights during training by minimizing the above loss. Unlike the actual WaveNet architecture the proposed model doesn't intend to learn the distribution of the output, which makes the conditioning insignificant in the context of this model architecture. Since we have the parallel data samples in hand, the model has specifically designed to generate a set of samples in a shot, rather than individual samples. This gives more momentum for the generation process than the actual WaveNet model and will be practically quite significant for nearly real-time applications.

6.1.2 Database selection

Unlike statistical techniques, neural networks are data centric, therefore, the performance would largely depend to the data on which they are trained. Similarly, the features learned by the network depend on the linguistic variety of the used corpus, like dialects of the speakers or phonological differences of languages. Because of which, the evaluation of wSSDRC model is done in two different languages – English and Greek. For this, two separate models were trained from scratch on the corresponding database – the details of each data set are provided in the following part. On the evaluation side, since the intelligibility of speech varies for different listening groups, the processed samples are evaluated with native and non-native listeners with normal and hearing impairment.

Before going to the subjective evaluation of the model, a multitude of experiments were conducted to find the best model hyperparameters with the help of objective metrics and informal listening tests. Therefore, the final model have had the specification follows. The wSSDRC model has in total 30 layers made up by thrice repeating a block of depth

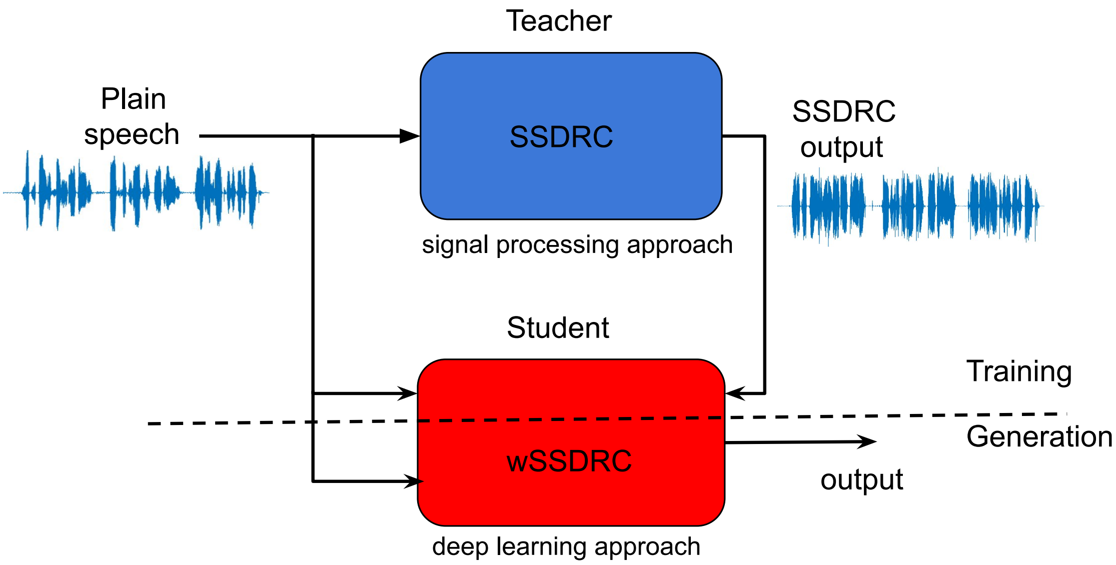


Figure 6.3 – Teacher-Student framework followed to train the wSSSDRC model.

10 that has the dilation factors $[1, 2, 4, 8, 16, 32, 64, 128, 256, 512]$, starting from the beginning. It sums up to a receptive field of size 6138 (3069 past & 3069 future samples), which means it considered 0.38 s of input samples (for 16 kHz signal) when predicting a single clean sample. In all the layers, convolutions of 256 channels are used. During training, the target samples predicted in a single traverse is a set of 4096 (training target field size). The model is fed with a single data point every time with a batch size of 1. In the testing phase, the target field size being varied depends on the test frame length. Just before feeding into the model, the wave files have been normalized to an RMS level of 0.06. This removed the loudness variations among the wave files. The loss function in Equation 6.5 was optimized with the Adam optimization algorithm, with an exponential decaying learning rate method [KB14]. The hyper parameters of the exponential decay method are learning rate $= 0.001$, decay steps $= 20000$, and decay rate = 0.99.

wSSDRC is evaluated against the SSDRC which was used to train the neural model. The plain speech is also included into the evaluation for understanding the relative intelligibility gain/loss produced by these approaches. The evaluation has been performed both in objective and subjective domains.

6.1.3 Objective comparison of SSDRC and wSSDRC

The objective evaluation is done only on the English database as it is just a preliminary study before final subjective testing of the system. The Speech Intelligibility Index (SII) has been used as the objective metric. SII captures the intelligibility of a speech signal in noise by looking at the long term average spectral distributions of the energy [RV05]. Here, we used an extension of the conventional SII by incorporating the temporal characteristics of the noise as well, known as the exSII [RV05]. The speech set is a selection from the Voice Bank corpus [VYK13]. It contains 48kHz recorded samples from 28 native English speakers of both genders speaking 400 different sentences. Two speakers recordings were kept for testing, which were not seen during the neural network training. To being fit with our WaveNet-like model the data has been down sampled to 16 kHz. Individual audio files were processed by the SSDRC and wSSDRC. All the signals have been normalized in root mean square (RMS) energy so that they have the same loudness before and after modification. The signals have been mixed with SSN and SWN type of noises with Signal to Noise Ratios (SNR) in the range of -5 to 5 dB. This tuning of SNR is being done in reference to the plain speech signals.

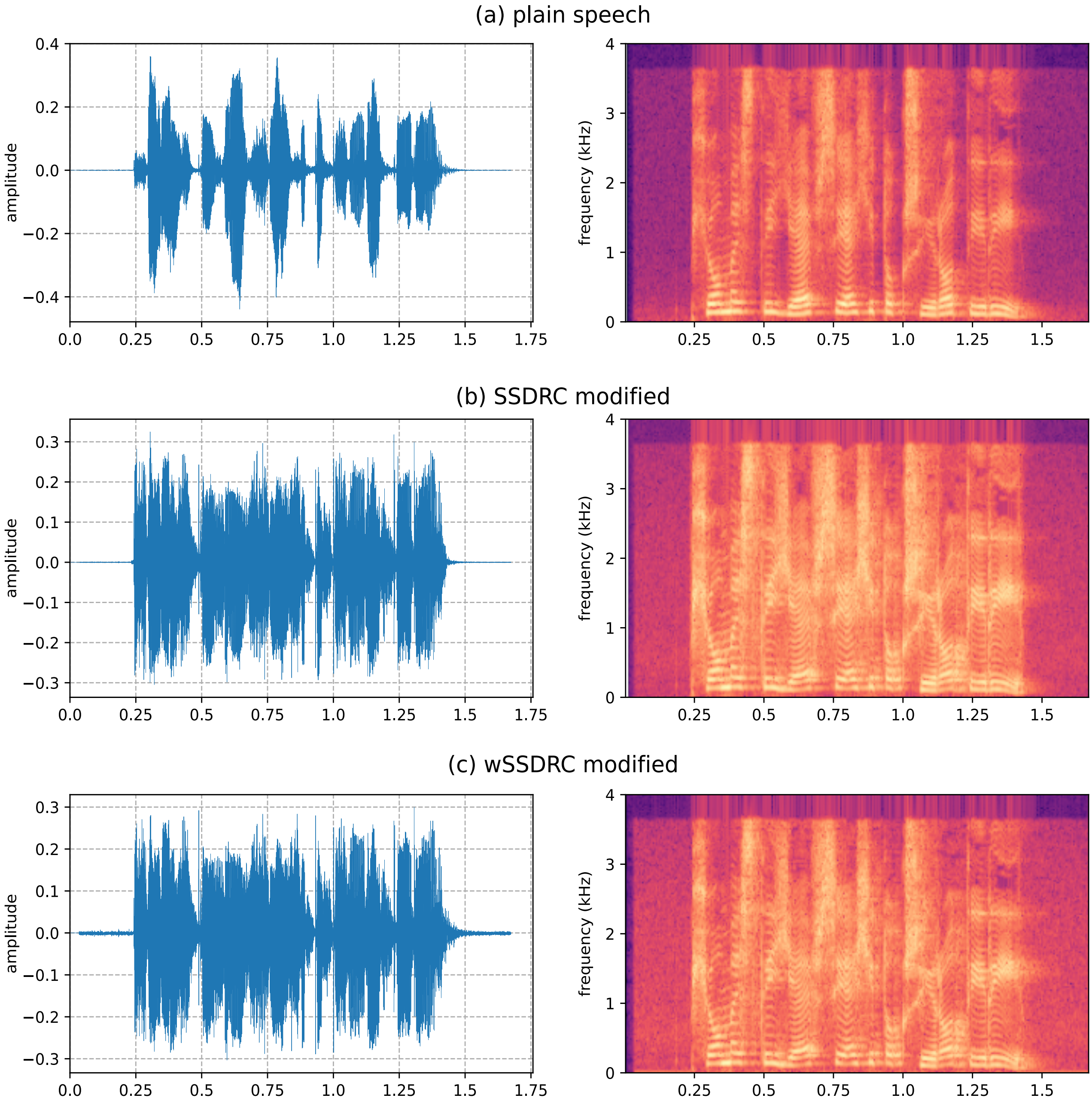
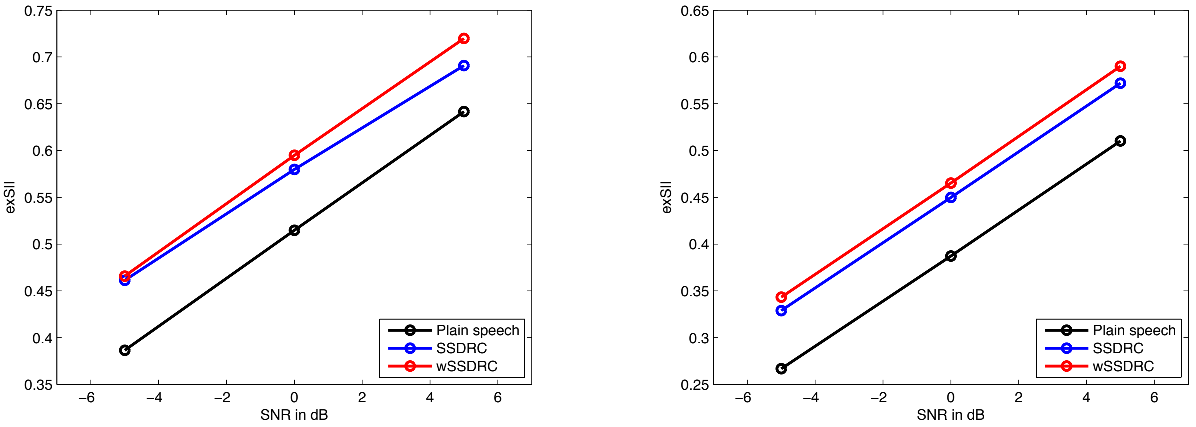


Figure 6.4 – Sample wave file processed for intelligibility by wSSDRC and SSDRC.

The time and frequency domain representations of a sample speech processed by SSDRC and wSSDRC are provided in Figure 6.4. We can clearly observe that the wSSDRC is producing the same level of modification as the SSDRC which was our target style. The main characteristic is that both produces much lower peak to root mean square (RMS) ratio signals compared to the original plain speech (upper panel of Figure. 6.4), which is an effect of dynamic range compression. Similarly, an abundance of energy in the high frequency regions of the modified signals. Therefore, this analysis confirms that the speech modification can be learned by a network through effective training.

The results observed from the objective experiments on both SSDRC and proposed wSSDRC are presented next. The gain in terms of intelligibility score as measured by exSII is shown in the Figure. 6.5a and Figure. 6.5b for the speech shaped noise (SSN) and stationary white noise (SWN), respectively. Higher the exSII score better the performance in intelligibility. As anticipated, the intelligibility of plain unprocessed speech improves as the listening SNR increases. From the exSII score it is clear that the intelligibility of speech processed through both the SSDRC and the suggested wSSDRC systems have significantly improved compared to that of the unprocessed plain speech. This improvement on intelligibility retained across the SNR range. As in plain speech, the intelligibility improves steadily as the SNR increases, indicating that the processing has not produced any damage to the signal. It is worth mentioning that in both types of noise the suggested wSSDRC system can maintain the intelligibility gain at the same level as that of SSDRC. Furthermore, in some cases intelligibility prediction of wSSDRC is even slightly higher to that of SSDRC, when SNR is increasing. Since this type of variations in objective intelligibility can not be fully related on, we had to have a wide scale subjective evaluation of both the models.



(a) exSII in speech shaped noise

(b) exSII in stationary white noise

Figure 6.5 – Objective intelligibility in extended SII (exSII) score.

6.2 Subjective intelligibility evaluations of wSSDRC

We have performed two sets of subjective evaluation of wSSDRC model. In the first evaluation, the intelligibility gain produced by the wSSSDRC to native and non-native listeners is analysed. Second, the performance benefits for normal and hearing impaired groups are tested. Since the subjective intelligibility would be influenced the sentence structure of database, e.g. the long sentences would be hard to remember, we had to use phonetically balanced corpus to perform these studies. Therefore, the Harvard/IEEE sentence style corpus [RCG+69] designed following the IEEE recommended practice for measuring speech processing models – in English and Greek languages – are selected for the experimentation. The details of English and Greek Harvard sentences are mentioned in the following sections.

6.2.1 Native and non-native listeners

The evaluation of differences in intelligibility benefits between native and non-native listeners is performed by considering English language as the base language. The speech corpus was from the Hurricane Challenge Natural Speech Corpus, which features a recording of 720 Harvard Sentences spoken by British male actor [CMVB13a]. The Harvard corpus contains 5 keyword sentences such as ”the salt breeeze came across from the sea” arranged into phonetically-balanced subsets. Subsequently, the wSSDRC model was retrained on the Harvard corpus, while SSDRC configuration unchanged.

For evaluation, three speech types are considered. *Plain:* Unprocessed plain speech selected from the above mentioned databases, 2) *SSDRC:* Speech stimuli enhanced with the SSDRC algorithm, 3) *wSSDRC:* Speech stimuli produced by the proposed neural-based wSSDRC model. The wSSDRC neural model was trained on SSDRC processed speech (as target output). Out of the 720 speech wave stimuli, 500 was used for training and the remaining 220 was reserved for testing.

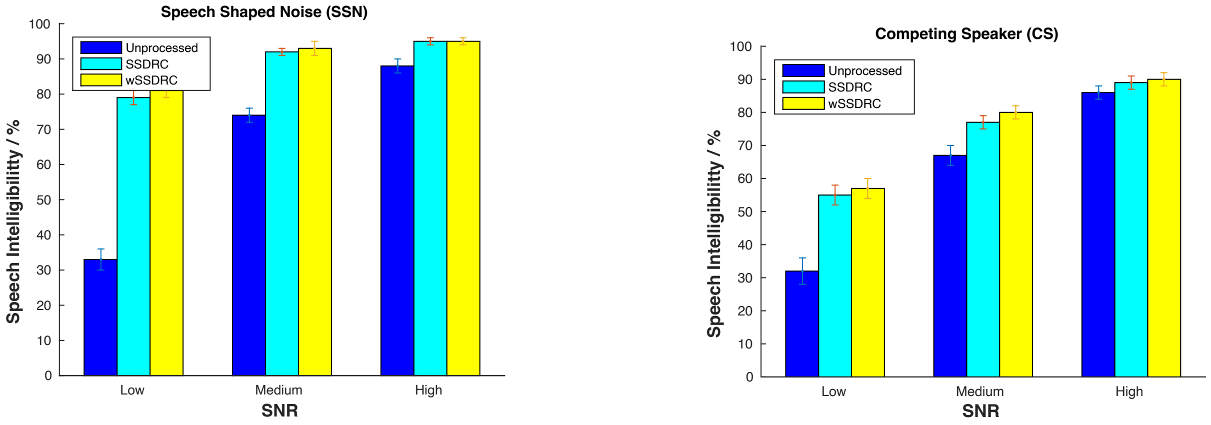
Speech intelligibility is measured as the ability to understand speech despite the communication barrier. If the spectrum of the masker is not fully overlap with the spectrum of the speaker, speech remains intelligible since only relatively a small band of frequencies is required for speech recognition [Moo12]. Here, we artificially create the communication barrier by simulating noise in the background while listening. Hence, the intelligibility can vary based on the spectral/temporal characteristics of masking noise. Noises in real life can broadly be categorised as the steady and fluctuating noise. Steady noise has the spectral characteristics that are relatively constant over time, e.g., noise produced by the room exhaust. This type of masker may mask some words more than others because of its spectral content. One common type of steady state masker used in speech-in-noise studies is the speech shaped noise (SSN). The SSN is synthesised in a way that to match the noise spectrum to the long-term average spectrum of clean speech. Whereas the second category of noise – called fluctuating noise – has changing spectral content over time. Therefore, they mask the same word differently based on their appearance on the time axis. Examples are competing speakers in the background or traffic noise.

Besides the spectral masking at the auditory periphery, the speech perception is also affected by the content in the masking signal. For instance, If the masker is speech from a competing speaker in the background it would be harder for the listener to discriminate the acoustic events of the speaker from the masker. Such types of maskers are called informational masker. Therefore, we have considered both informational and steady state noises. Speech shaped noise (SSN) is considered as the steady state masker. The stationary SSN was generated by passing white noise through a filter whose frequency response matches to the long-term average spectrum of the clean speech database. Thus, SSN has a steady energy in the same frequency band of speech. As informational masker, a competing speaker (CS) recording is selected. The non-stationary CS maker was a female voice from the Hurricane challenge corpus. The selection of different listening groups are discussed next.

Native listeners: For this experiment, we recruited N = 30 British English native speakers (age range = 18-34; mean age = 25 years). Participants were screened for hearing loss via a Pure Tone Audiometry (PTA), at frequencies of 0.5, 1, 2, 4 kHz. Subjects passed the test with a hearing threshold equal or less than 25 dB HL (averaged across frequencies) in both ears. The listening tests were performed in sound treated booths at University of Edinburgh. Stimuli were presented via Beyerdynamic 770 headphones and participants had to type onto a keyboard what they had heard. Stimuli were heard once and the test was balanced with a Latin square design. The study was self-paced with an average durations of 25 minutes. The intelligibility is tested at three SNR points for each noise type. For this study, the SNR levels were CS Low = -21dB, Mid = -14 dB, High = -7 dB; SSN Low = -9 dB, Mid = -4 dB, High = +1dB. These SNR levels roughly correspond to 25%, 50% and 75% subjective intelligibility scores of the native listeners [CMVB$^+$13c]. There were 8 sentences in each masking noise condition that subject had to respond.

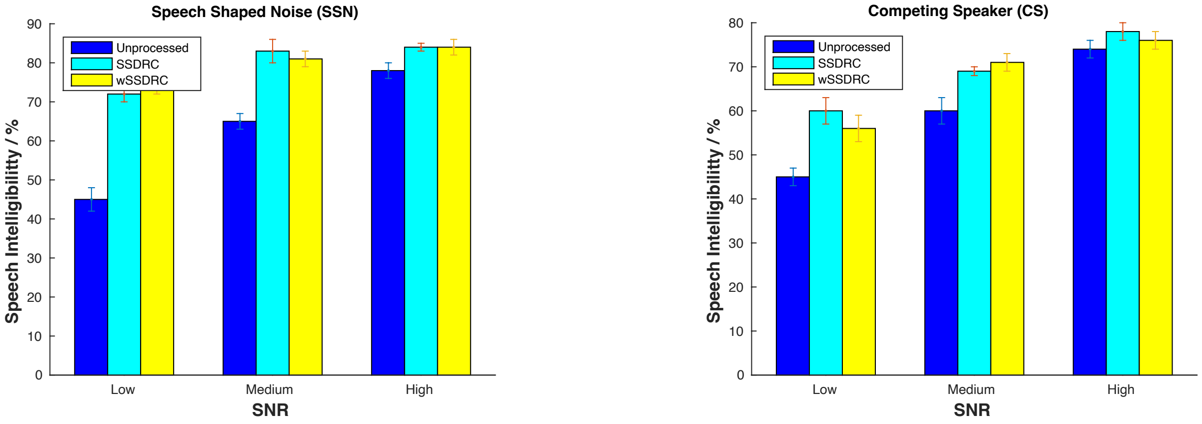
Non-native listeners: This study was conducted with the participation of students from the University of Crete (UoC). As the subjects were non-native English speakers, they were required to hold a B1 English language qualification. The listening tests were conducted in sound treated booths at the University, and stimuli were administered via high quality headphones. Participants were screened for hearing loss via a Pure Tone Audiometry (PTA), at frequencies of 0.5, 1, 2, 4 kHz. Subjects passed the test with a hearing threshold equal or less than 25 dB HL (averaged across frequencies) in both ears. The study was self-paced with an average duration of 25 minutes. Stimuli were heard once and the test was balanced with a Latin square design. In total we had N = 30 NH participants (average age = 24 years) who were screened for hearing loss in the same way as the native English speakers. An initial pilot experiment were conducted to identify the 25%, 50% and 75% intelligible SNR levels for the non-native group. It was found to be: CS Low = -13 dB, Mid = -6 dB, High = +1 dB; SSN Low = -4 dB, Mid = +1 dB, High = +6 dB. There were 8 sentences in each masking noise condition that subject had to respond as in the case of native experiment.

The keyword identification scores averaged over the participants in each condition are plotted in Figure. 6.6 for native and Figure. 6.7 for non-native groups. The speech intelligibility is measured in terms of correct keywords identified to



(a) (b)

Figure 6.6 – The key word identification scores of native listeners both in speech shaped and competing speaker noise.



(a) (b)

Figure 6.7 – The keyword identification scores of non-native listeners both in the competing and speech shaped noise.

the total keywords presented in each condition. First, the plain speech intelligibility increases with increase in maker SNR in the listening background. It is easy observable that both the modifications (SSDRC and wSSDRC) boosted the intelligibility for both native and non-native listeners, in comparison to unprocessed plain speech. However, the relative intelligibility gain varies largely on the type of noise under which the speech is listened. For instance, both for native and non-native at the low SNR, the intelligibility benefits by the processing in speech shaped noise were higher than in competing speaker masker.

Specifically, for native listeners in SSN, the wSSDRC modification produces 150% relative improvement in intelligibility over the plain speech at the lowest SNR. This gain is reduced to 24% as the SNR improves to the Mid point, which was further reduced to 10% in the highest SNR. However, the relative intelligibility benefits were lesser in non-native group. In the same maker, only 60% relative improvement were observed at Low SNR, which then reduced to 21% at Mid point, then to 5% at the highest measured SNR. This disparity in identifying words in sentences might have been partly due to the less language proficiency of non-native listeners.

Similarly, in the case of competing speaker (CS), an 83% relative performance gain is observed for native listeners at the lowest SNR over the unprocessed plain speech. This gain is reduced to 15% at Mid point, which further come down to 10% at the highest SNR. While in the case of non-natives, the relative gains were lower. At the Low SNR, only 22% improvement was observed, which was then increased to 16% and this further reduces to 3% at the High SNR point.

Across all conditions, no statistically relevant difference in intelligibility gain were observed between wSSDRC and

SSDRC. This was expected from the fact that the wSSDRC has been trained to mimic the feature modifications of SSDRC algorithm. Therefore, the findings further underline the fact that the proposed network has been fitted very veal into the problem of intelligibility modification.

6.2.2 Normal and hearing impaired listeners

In the section above we observed a clear benefits by the wSSDRC processing both for native and non-native listeners. However, several studies have pointed out that older listeners have difficulty in understanding speech, particularly in the presence of noise [Plo86, DP80]. This hearing loss has been found to be largely linked to the weaker sensitivity of the cochlear filters. Since the cochlear filters are band selective, reallocation of energy in the spectral domain must improve the intelligibility for damaged ears. In the past, spectral contrast enhancement with enhancing the peak to valley contrast of the speech spectrum was found to contribute to the intelligibility for listeners with sensorineural hearing impairment [BMG93]. Since wSSDRC does perform the spectral redistribution of energy, it is relevant to see its effects on hearing impaired group.

This evaluation is done within the Greek community due to the easy accessibility of impaired patients. We used the Greek Harvard corpus as the corpus to train and evaluate the wSSDRC model.

The Greek Harvard corpus: The Greek Harvard (GrHarvard) Corpus was recently designed to address a lack of Greek sentence corpora developed for intelligibility testing [Sfass]. It comprises 720 sentences in the format of the Harvard/IEEE material [RCG$^+$69] with the necessary accommodations for the Greek language. The original Harvard material has been used extensively in speech intelligibility experiments (e.g. [CMVB$^+$13c], [HL10]) and has also been adapted for the Spanish language [ALC14]. Each sentence of the GrHarvard Corpus includes five keywords consisting of one, two or three syllables, with the total number of words per sentence varying strictly from five to nine. Sentence content was inspired in part by the original Harvard sentences; a translation of the original material was not possible in most cases, because grammatical differences between the English and the Greek language rendered many of the keywords unsuitable candidates for the GrHarvard Corpus. The majority of keywords have been selected from GreekLex 2 [KvHPL17] so that the resulting sentences are meaningful, semi-predictable and resemble everyday language. For example, Το ξύλο είναι άριστο υλικό για παιχνίδια και κύβους' [to.”ksilo.”ine.”aristo.ili”ko.jja.pe”xniDja.ce.”civus] (Wood is an excellent material for toys and cubes), “Καυτός ατμός ξέφυγε από τη σπασμένη βαλβίδα” [ka”ftos.a”tmos.” ksefijje.a ”po.ti.spa”zmeni.val”viDa] (Hot steam escaped from the broken valve). The GrHarvard Corpus is freely available to the research community for non-commercial purposes. The 720 sentences in Greek orthography and phonetic transcription as well as metadata information are provided [1].

The 720 utterances of the GrHarvard Corpus were divided into two groups, 600 for training and the remaining 120 for validating and testing the model. We used the same samples as the validation and test set. Sentences with a maximum of 7 words in total were selected for testing / validating. Although the dataset was recorded at 44.1 kHz, it was downsampled to 16 kHz, as feeding high-resolution samples into the model would limit the phone context covered by the receptive fields. The corresponding target pairs were generated by running the SSDRC algorithm over the samples. In the process of finding the optimal configuration, the model trained with British Harvard was tested on the Greek test set. It performed well, except for some occasional clicks in the generated samples that would make listening less comfortable. Therefore, the Greek training set was ultimately selected to fully train the network. As such, the final evaluating model is purely trained on the Greek Harvard corpus.

Two groups of listeners were recruited: individuals with normal hearing (NH) and hearing impairment (HI). The participants with HI were screened for hearing loss via Pure Tone Audiometry (PTA) at frequencies of 0.5, 1, 2, 4 kHz in both ears. The group with HI was characterized by an average hearing loss of 62 dBHL. Most of the participants wore hearing aids which were removed during the test.

This experiment has considered masking based on stationary speech shaped noise (SSN) only. SSN was selected from the Hurricane Challenge [9]. Since intelligibility level varies from subject to subject, intelligibility gains should

[1] https://www.csd.uoc.gr/~asfakianaki/GrH.html

be observed from a common reference point. This was achieved by designing subject-specific Signal-to-Noise Ratio (SNR) sets to match the speech reception threshold (SRT), i.e. the point at which 50% of speech is intelligible for each individual listener. For this, an initial pilot study was carried out, during which each participant was asked to listen to an initial set of samples, masked with SSN at SNR points in the range of -7 dB to -1 dB for NH and -3 dB to +9 dB for HI individuals. After analysing the responses, subject-specific SNRs were selected that matched each listener's SRT. The masking noise level for the final test was set on this SNR value.

The percentage of correct words recalled in each condition from the 13 participants with normal hearing and 11 with hearing impairment are plotted in Figures 3 and 4, respectively. The lower and upper sides of the boxes are the lower and upper quartiles. The boxes cover 50% of data distribution. Median of the distribution is represented by the horizontal line inside the box. The whiskers are the two lines outside the box that extends up to 1.5 times the interquartile range (or box width) from the lower and upper quartiles. Samples beyond that are labelled outliers in each conditions.

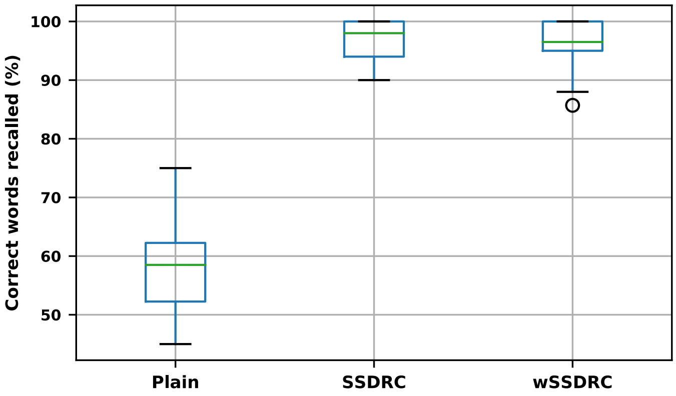


Figure 6.8 – Words recalled by participants with Normal Hearing (NH) in different conditions; boxes represent data dispersion

The intelligibility score of plain, unmodified speech for both groups, with NH and HI, is on median 58% and 45%, respectively. The values confirm that participants in each group on average listened to the plain test at the SRT points.

Looking at the group with NH, we observe that the neural enrichment model (wSSDRC) has induced a median intelligibility of 97%, a rise of 39% from the plain unprocessed speech. SSDRC has produced a median gain of 98%, a value. closely matching that of the wSSDRC model. The difference between the two results is not statistically significant. Regarding the group with HI, the median intelligibility of the samples from the neural model (wSSDRC) was 83%, which is an improvement of 38% over the Plain condition. SSDRC produced a slightly higher gain of 88%. This might be due to the few outliers in the wSSDRC condition, as can be seen in Figure 4, which have caused the larger median deviation between SSDRC and wSSDRC, in contrast to the group with NH.

To statistically account for this variability among the groups, and observe its influence on the between group variability, an one-way analysis of variance (ANOVA) has been conducted.

ANOVA is a comparative measure of variance among and between groups. If within-group variability is more significant than between-group variability, the dominance of one group over the other should not be appraised as a reliable gain. ANOVA examines these variations in a more absolute statistical way. In the present study, this is important in order to capture the real gain, if any, as different processing types vs. unprocessed speech are being compared, and more importantly, in order to match the performance of SSDRC with that of wSSDRC, and investigate how close the two models are.

ANOVA computes F-statistics, which is the ratio of inter- group to intra-group variability. Higher F-value indicates higher inter-group variability, which in turn means one group is dominant over the other. The p-value accompanying the F-value indicates that the probability of the predicted F-value could be random. Lower p value indicates higher

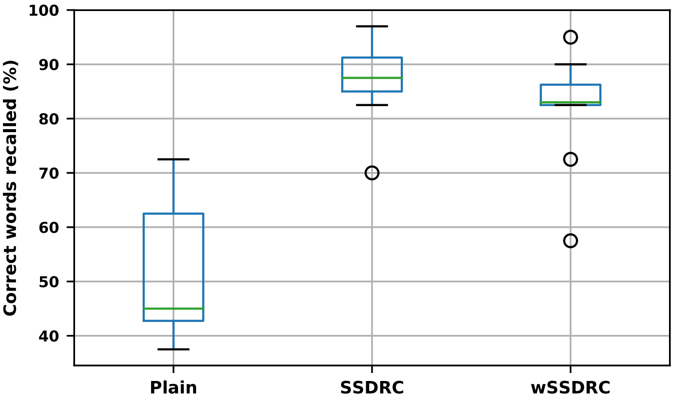


Figure 6.9 – Words recalled by participants Hearing Impairment (HI) in different conditions; boxes represent data dispersion

confidence of the returned F-value.

Firstly, let us consider the NH group. On the null hypothesis that the three modifications - Plain, SSDRC and wSSDRC - produce the same intelligibility gain, we ran the one way ANOVA over the three methods. It rendered the result $\left(F = 163.6, p = 7.4 \times 10^{-18}\right)$, the very high F and very low p indicates that at least one of the compared groups is significantly different. Though it is obvious from Figure 3 which group falls behind, we have computed an additional series of ANOVA; dividing the three pair groups into sub groups of two pairs. The Plain - SSDRC produces $\left(F = 211.2, p = 9.36 \times 10^{-13}\right)$, Plain $-$ wSSDRC produces $\left(F = 184.5, p = 3.56 \times 10^{-12}\right)$, and SSDRC-wSSDRC produces $(F = 0.192, p = 0.66)$. The picture is clearer now that Plain class is significantly farther from the other two categories. More importantly, when comparing the SSDRC with wSSDRC the F-value is 0.192, which is very close to the ideal case, $F = 0$, the case where the two categories would be exactly equal. This confirms that the wSSDRC produces an equivalent statistical intelligibility gain as the SSDRC for NH.

In the case of the HI group, when performed the statistical test between SSDRC - Plain categories, the statistics shows $\left(F = 65.3, p = 1.02X10^{-7}\right)$, while the neural enrichment (wSSDRC)$-$ Plain gives $\left(F = 39.28, p = 4.04X10^{-6}\right)$ Though the F-values are not as large as the NH, here also, the higher F values indicate the obvious fact that the processing has resulted in substantial intelligibility gain. Though the two F values differ significantly, when computing the same test between SSDRC - wSSDRC the F score $(F = 1.94, p = 0.178$) was close to the matching point, which again manifests that both models are rendering relatively similar gain.

The ANOVA tests further confirm the fact that the neural enrichment model (wSSDRC) produces an equivalent intelligibility gain with the signal processing model (SSDRC) that was used as the target style. Hearing impaired people benefits equally as the normal hearing group with the wSSDRC processing. Besides, the study confirms that a carefully designed neural model could learn the speech features for intelligibility even on a language like Greek which differs from languages of Latin origin. Though it may not be attractive at this point, the same neural model could have been robust against noise if it were trained with noise perturbations at the input, in contrast to the statistical model.

6.3 Conclusions

In this chapter, the usability of neural networks (NNs) for intelligibility enhancement of speech was experiments. A network was trained to mimic the spectral shaping and dynamic range compression of famous SSDRC algorithm. The resulting model is called wSSDRC. The objective analysis showed that if trained sufficiently wSSDRC can learn the intelligibility patterns from data. Multiple subjective intelligibility evaluations with phonetically balanced English and Greek Harvard corpuses were conducted. The results show that the wSSDRC samples are as intelligible as the SSDRC in all listening conditions for both normal and hearing impaired listeners. This was confirmed with statistical ANOVA test in some evaluations. However, wSSDRC is not language independent as SSDRC, therefore had to be retrained for Greek and English voices. Once trained adequately, the model was found to produce stable intelligibility gain which is important in practice. The findings that neural networks can generate intelligible speech with training on an intelligible speech style lies the foundation for the following chapters.

Chapter 7

Intelligible Text-to-Speech Synthesis (TTS)

With the finding that intelligibility can be learned on the weights of neural networks, we extended the same concept to text-to-speech synthesis (TTS) domain to generate intelligible speech from text. Over the years, text-to-speech (TTS) systems have become more prevalent with a substantial range of applications including personal voice assistants, public address systems and navigation devices. In a quiet environment, the intelligibility of synthetic speech corresponds to that of natural speech. However, the intelligibility is typically fallen below the level of natural speech in noisy conditions [CMVB+13c]. Listeners in real-world scenarios often hear speech in noisy surroundings where the intelligibility of synthetic speech is also compromised. Therefore, highly efficient TTS systems which are able to simulate Lombard effect and make the speech more intelligible are essential for the end listeners. Such speaking style conversion retains the linguistic and speaker-specific information of the original speech.

Few studies have explicitly adapted Lombard speech onto speech synthesis models by focusing on articulatory effort changes [RSVA11, PDD14]. Previously, the majority of such studies were conducted using hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) due to its superior adaptation abilities and flexibility. The HMM model trained on normal speech was then adapted using a small amount of Lombard speech and improvements were shown under different noisy conditions [CMVB+13c]. Yet, these approaches were limited to poor acoustic modeling and inability to synthesize high-fidelity speech samples. To overcome this, deep neural network approaches were implemented where the robustness of acoustic modeling is improved by efficient mapping between linguistic and acoustic features. Inspired by the success of adversarial generative models, Cycle-consistent adversarial networks (Cycle-GANs) showed promising results in terms of speech quality and the magnitude of the perceptual change between speech styles [SJY+19, SJA+19]. An extension to recurrent neural networks and particularly long short-term memory networks (LSTMs) were proposed that it successfully adapted normal speaking style to Lombard style [BAA17]. In [BJA+19c], the authors demonstrated results with sequence to sequence (seq2seq) TTS models along with the recently-proposed Wavenet vocoder where the audio samples are generated through a non-linear autoregressive manner. Along with different adaptation approaches, various TTS vocoders are compared in the context of style transfer and assessment was performed in terms of speaking style similarity and speech intelligibility [SJRA19, BJA+19a].

To train a TTS system with Lombard style, a sizable amount of training data is required. However, the collection of a large portion of Lombard speech is difficult. Such data sparsity limits the usage of typical data-driven approaches similar to the recent end-to-end TTS systems. Our work takes into account the use of speaking style adaptation techniques leveraging on large quantities of widely available normal speech data referred to as transfer learning. It assumes the prior knowledge from a previously model trained with large variations in linguistic and acoustic information and adapts to the target styles even with limited amount of data. In the literature, most of the vocoders for style transfer in TTS systems are either source-filter based models or convolutional models [SJRA19, BJA+19c]. However, such techniques are limited by their inefficiency both in modeling proper acoustic parameters and in computational complexity of sample generation. Inspired by the performance and computational aspects of recurrent neural networks, in this work, we employ WaveRNN as a vocoder [KES+18] which generates speech samples from acoustic features, i.e., mel-spectrograms. Experimental

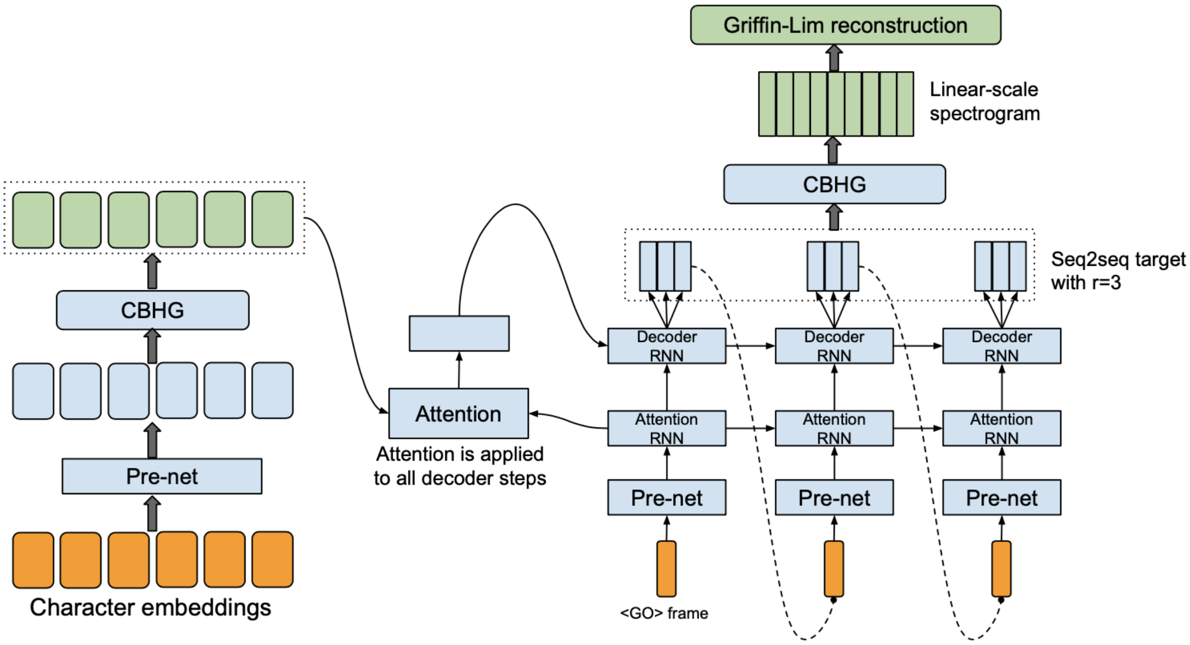


Figure 7.1 – Block diagram of Tacotron architecture (from [WSRS+17]).

results indicate that WaveRNN is capable of adapting appropriate target speech style and able to provide more stable high-quality speech samples. To generate the mel-spectrograms from text, we utilize a popular architecture Tacotron, a seq2seq encoder–decoder neural network with attention mechanism [WSRS+17].

Improvement of speech intelligibility in noise can also be achieved by signal processing techniques such as amplitude compression [NG76], changes in spectral tilts [LC09], formant sharpening and dynamic range compression [ZS14].The method Spectral Shaping and Dynamic Range Compression (SSDRC) has been shown to provide high intelligibility gains in various noisy conditions by redistributing signal energy on time-frequency information [ZKS12]. In [VBYKS13], the best performing method was achieved by applying additional processing, i.e., dynamic range compression after generating Lombard style adapted TTS. The results, however, failed to increase the intelligibility under competing-speaker noise. In order to develop a highly intelligible communication system and restrict the latency imposed by additional processing after the TTS synthesis. Here, we implement Lombard-SSDRC TTS where the TTS is trained with Lombard speech processed through the SSDRC algorithm. Hence, we combine the advantages of naturally-modified Lombardness with speech enhancement strategies in frequency-domain (spectral shaping) and in time-domain (dynamic range compression) into an intelligibility-enhanced TTS synthesis system.

7.1 Neural TTS architecture

The proposed TTS system is composed of two separately trained neural networks: (a) Tacotron, which predicts mel-spectrograms from text and (b) WaveRNN vocoder, which converts the mel-spectrograms into time-domain waveforms.

7.1.1 Tacotron

Tacotron [WSRS+17] (Figure 7.1) is a seq2seq architecture with attention mechanism and it is heavily inspired by the encoder-decoder neural network framework. The system has two main components: (a) an encoder and (b) an attention decoder. The encoder consists of 1-D convolutional filters, followed by fully-connected (FC) layers and a bidirectional gated recurrent unit (GRU). It takes text as input and extracts sequential representations of text. The attention decoder is a set of recurrent layers which produces the attention query at each decoder time-step. The input to the decoder RNN can

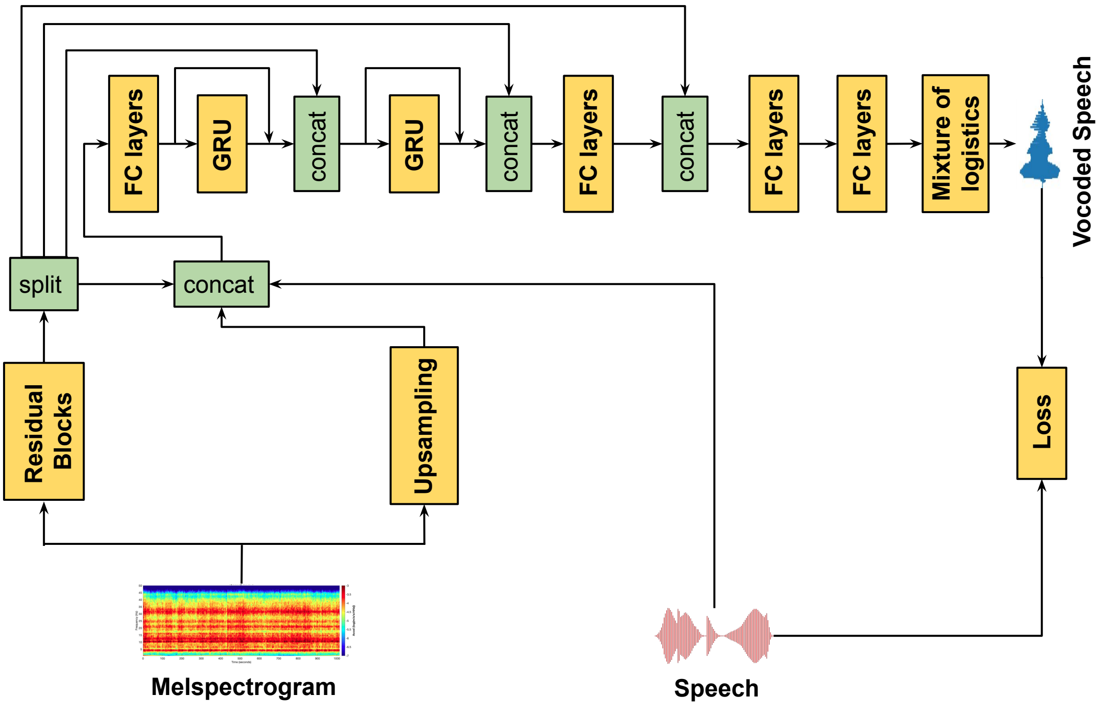


Figure 7.2 – Block diagram of WaveRNN architecture.

be produced by concatenating context vector and output of the attention RNN. The decoder RNN is basically a 2-layer residual GRU whereas the attention RNN has a single GRU layer. The output of the attention decoder is a sequence of mel-spectrograms which is then passed to the vocoding stage.

7.1.2 WaveRNN

The implemented WaveRNN vocoder is based on the repository[1] which in turn is heavily inspired by WaveRNN training [KES+18]. This architecture is a combination of residual blocks and upsampling network, followed by GRU and FC layers as depicted in Figure 7.2.

The architecture can be divided into two major networks: the conditional network and the recurrent network. The conditional network consists of a pair of a residual network and an upsampling network with three scaling factors. At the input, we first map the acoustic features, i.e., the mel-spectrograms to a latent representation with the help of multiple residual blocks. The latent representation is then split into four parts which are later used as input to the subsequent recurrent network. The upsampling network is implemented to match the desired temporal size of the input signal. The outputs of these two convolutional networks i.e., residual and upsampling networks along with speech are fed into the recurrent network. As part of the recurrent network, two uni-directional GRUs are employed with a few FC layers. By designing, such network not only reduces the overhead complexity with less parameters but also it takes advantage of temporal context resulting in better prediction.

In addition, we apply continuous univariate distribution to be a mixture of logistic distributions [OLB+18] which allows us to easily calculate the probability on the observed discretized value. Finally, discretized mix logistic loss is applied on the discretized speech samples.

[1] https://github.com/fatchord/WaveRNN

7.2 Transfer learning

The majority of deep learning methods perform well under the standard assumption that the training and inference data are drawn from similar feature space and data distribution. When the distribution changes, models need to be trained from scratch using new training data. Under the condition of data scarcity such as in our case for Lombard data, training a new model on such a limited sample size might lead to poor execution. In such cases, transfer learning (TL) offers a desirable and extremely important adaptation framework [PY09]. Assuming that there are two tasks, source task and target task, TL tries to boost the performance of the target task by utilizing knowledge learned from the source task via fine-tuning prior distributions of the hyper-parameters.

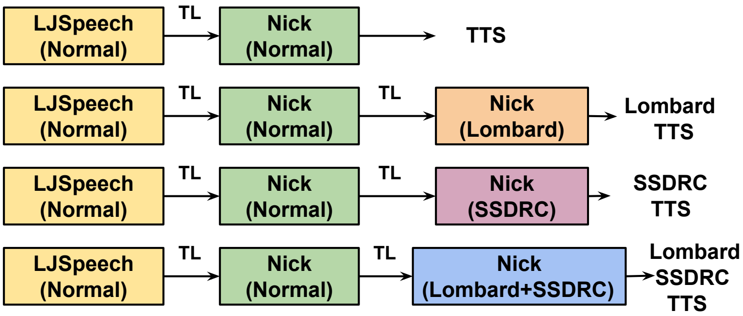


Figure 7.3 – A functional block diagram of the proposed adaptation techniques used in this study. Each block represents a TTS system (Tacotron + WaveRNN) which takes text as input and generates speech samples.

We develop four TTS systems based on the speaking styles: normal TTS, Lombard TTS, SSDRC TTS and Lombard-SSDRC TTS. To effectively transfer the prior knowledge, we initially train the TTS system with normal speech (single female speaker from LJSpeech corpora) which has a large amount of linguistic variability. Then, we adapt the learned model with normal speech from a male speaker (Nick). This normal TTS serves as the baseline system for our experiments. Lombard TTS system is then fine-tuned using again the TL approach on the limited Lombard data from the same male speaker (Nick). Whereas, SSDRC TTS uses training data processed with SSDRC algorithm applied on Nick normal speech. The last TTS system is fine-tuned on data that is prepared by applying SSDRC algorithm on Nick's Lombard speech, referred to as Lombard-SSDRC TTS. Please note that all proposed TTS systems comprise of Tacotron and WaveRNN modules [PPSed] and each module is trained separately using data from the corresponding target speech style.

7.3 Database and Hyperparameters Selection

The proposed TTS systems are trained using two publicly available database, i.e., LJSpeech corpus [Kei17] and Nick Hurricane Challenge speech data [CMVB+13b]. LJspeech consists of 13,100 short audio clips of a single female professional speaker reading passages. The Nick data has both normal and Lombard styles of British male voice professional speech. The normal speech consists of 2592 utterances (∼2 hours) whereas the Lombard speech data has 720 utterances (∼30 minutes). During training, we always consider 2400 utterances for normal and 500 utterances for Lombard speech. We additionally compare with the baseline Lombard TTS system which is built on Tacotron and WaveNet architecture [BJA+19c]. The WaveNet configuration used in their system consists of three repetitions of a 10-layer convolution stack with exponentially growing dilations, 64 residual channels and 128 skip channels whereas the Tacotron architecture is similar to ours. The proposed Tacotron and WaveRNN models use 80 dimensional normalized mel-spectrograms, extracted from audio frames of width 50ms, hop length of 12.5ms and 2048-point Fourier transform. In Tacotron, character embeddings are set to 256 and a progressive training schedule is employed with reducing batch size from 32 to 8. WaveRNN architecture is based on a set of 10-layer convolution stack inside residual blocks followed by 2 GRUs. Each GRU has 512 hidden units. Code and audio samples can be found in [2].

[2] https://dipjyoti92.github.io/TTS-Style-Transfer/

Table 7.1 – $SIIB^{Gauss}$ intelligibility measure at different SNR levels under speech-shaped and competing-speaker noise.

Systems	SSN			CSN		
	-10 dB	-5 dB	0 dB	-21 dB	-14 dB	-7 dB
TTS	15.03	26.80	42.43	13.3	17.86	28.27
Lombard TTS [BJA+19c]	17.89	33.89	54.53	9.91	18.1	36.21
Lombard TTS (ours)	20.02	37.43	58.65	13.52	22.51	41.65
SSDRC TTS	29.90	51.02	77.97	16.73	29.75	55.56
Lombard-SSDRC TTS	**35.04**	**58.68**	**88.35**	**19.13**	**35.84**	**68.35**

7.4 Observations and discussion

Objective intelligibility scores are computed first for the five style adapted methods (TTS, Lombard TTS [BJA+19c], proposed Lombard TTS, also refer to as Lombard TTS (ours), SSDRC TTS and Lombard-SSDRC TTS) under two different noisy conditions. A recently developed intelligibility metric called 'speech intelligibility in bits' ($SIIB^{Gauss}$) [VKKH18] is implemented as an objective evaluation metric. It takes into account the information capacity of a Gaussian channel between clean and noisy signals. Higher values refer to better intelligibility. The scores are evaluated from 250 utterances and each adaptation approach has 50 distinct utterances. Table 7.1 presents $SIIB^{Gauss}$ intelligibility scores. We consider three different Signal-to-Noise Ratio (SNR) levels masked with two types of noise: speech-shaped noise (0 , -5 and -10 dB) and competing-speaker (-7, -14 and -21 dB). Since we are focusing in the context of TTS, we omitted the scores for natural speech in our experiments.

It can be observed that the standard synthesis system trained with normal speech, referred to here as the speech type 'TTS', is the worst performer when compared to the rest of the methods under any condition as expected. To enhance the intelligibility, TTS is re-trained with limited Lombard style data. We observe that the proposed Lombard TTS i.e., Lombard TTS (ours) is able to successfully mimic the Lombardness and outperforms baseline Lombard TTS from [BJA+19c] with a relative improvement between 8% and 12% in SSN and 15% to 36% in CSN conditions across different SNR levels: from low to high SNRs. The results also show high performance gain of 18% and 36% in Low SNR i.e., -10 dB for SSN and -21 dB in CSN conditions, respectively. The use of WaveRNN instead of WaveNet vocoder as in the baseline Lombard TTS, demonstrates how the choice of vocoder affects the intelligibility of synthesized speech. WaveRNN effectively adapts to the new style while trained with limited amount of target style data. Furthermore, taking into account the SSDRC approach, we aim towards additional intelligibility gains under adverse noise conditions. Our results reveal that SSDRC TTS archives further improvement compared to the Lombard TTS. Motivating by the boosting effect of Lombard style, along with the enhancement by SSDRC data in terms of speech intelligibility, the proposed Lombard-SSDRC TTS shows significant intelligibility gains between 110% and 130% in SSN, and 47% to 140% in CSN against TTS. Those results can be attributed by the fact that the combined model exploits efficiently both Lombardness and spectral shaping with range compression by modifying time-frequency regions.

To assess the performance on subjective evaluation, metric scores were computed based on the number of keywords correctly identified in each sentence. The short common words 'a', 'the', 'in', 'to', 'on', 'of', and 'for' were excluded. The listening test was conducted via a web-based interface and ten native listeners participated in the test. No listener heard the same sentence twice, and each condition was heard by the same number of listeners. Since intelligibility level varies from one listener to another and large variability in scores can be possible when listeners use different hearing devices or backgrounds, intelligibility gains should be observed from a common reference point. This was achieved by designing an initial pilot study where subject-specific SNR levels are matched with the speech reception threshold (SRT) at which 40% of normal speech is intelligible for each individual listener. In the final listening test, we choose SNR levels based on the values obtained from the pilot study for each listener individually.

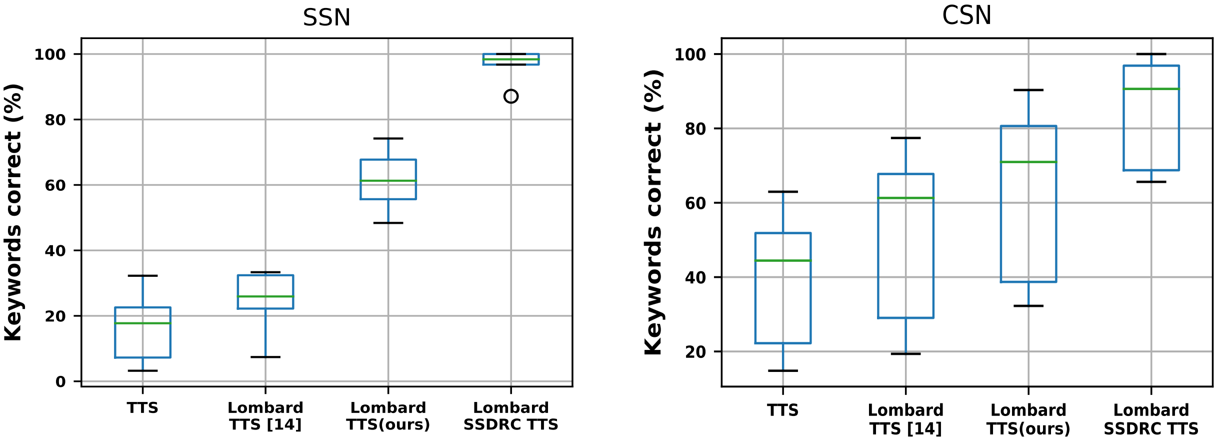


Figure 7.4 – Box plot results for listeners' keyword scores across of methods for SSN and CSN.

Box plots reported in Figure 7.4 allow comparison between different TTS modification algorithms. The subjective results reveal a similar pattern to the objective metrics. The proposed Lombard-SSDRC TTS outperforms all other methods with a remarkable margin under all noisy conditions. Lombard-SSDRC TTS shows superior performance by achieving a remarkable relative improvement of 455% for SSN and 104% for CSN in median keyword correction rate compared to TTS method. It is worth noting that the performance gains are immensely higher in SSN condition, although we observe outstanding performance gains in both noisy conditions. Moreover, the comparison between Lombard TTS [BJA+19c] and Lombard TTS (ours) adaptation methods highlights that Lombard TTS (ours) method achieves significantly better performance in terms of keyword correction rate. This confirms the adaptability of WaveRNN for limited data scenarios, and shows its effectiveness in the transfer learning approach. The results indicate a relative improvement of 136% in SSN and 16% in CSN compared to Lombard TTS [BJA+19c] in terms of median keyword correction rate.

7.5 Conclusions and perspective

In summary, we built and evaluated a set of intelligible TTS systems for various speaking styles with the help of transfer learning in Tacotron + WaveRNN architecture. The synthesized voice was adapted to two strategies: Lombard style recordings and SSDRC algorithm. First, we showed that the Lombard-adapted TTS system (ours) is able to successfully learn Lombard style under limited training data and outperforms the baseline Lombard TTS system [BJA+19c] by a significant margin when masked either with SSN or CSN noise. This shows the advantage of applying neural-based WaveRNN vocoder and its importance in achieving highly-intelligible Lombard synthetic speech.

The SSDRC adaptation of TTS was found to further improve the intelligibility substantially compared to both the normal and Lombard-adapted TTS systems in objective metric. Furthermore, to enjoy larger intelligibility gains, we combined the benefits of Lombardness with the SSDRC modification strategy. Experiments on both objective and subjective intelligibility scores confirmed that the combined system contributed to significant gains under all noisy conditions. In conclusion, these observations further underline the fact that neural networks can be optimized to learn various speaking styles so that to generate intelligible speech in adverse conditions; an observation that was reported in the previous chapter with speech input.

Now that it has been shown that neural models can learn the intelligibility patterns from data, we will investigate their robustness to noise (at the input) in the next part.

Part III

Joint End-to-End Speech and Listening Enhancement

Chapter 8

Joint Far- and Near-End Enhancement

Speech acquisition in the real world is detrimented by the presence of background noise. Speech enhancement (SE) models are designed to restore the clean message before its delivery to the targeted listener. SE systems clean the signal by suppressing the noise, thereby improving the perceived quality and/or intelligibility of speech. Whereas the listening enhancement (LE) systems, such as wSSDRC, are designed with the supreme objective of improving speech intelligibility through modifying speech spectral and temporal structures, as the naturally produced speech is not always very intelligible for listeners in noise or at distance. They have been reported to produce substantial intelligibility boosts both for normal and hearing impaired listeners in multitude of acoustic settings [CMVB$^+$13c, RSVBC20, ZSIA16].

However, most of the previous LE models were designed under the assumption that clean speech is readily available to be modified (noise-free far-end in Fig. 8.1). This assumption does not hold in the majority of practical cases. For instance, one could easily imagine a situation where both the speakers in a mobile communication are being subjected to noise. Therefore, integration of SE module to address the noise at the far-end before going to the intelligibility modification for near-end is essential in practice. Attempts have been made in the past to integrate the statistical SE and LE modules for joint far-end near-end enhancement. In [KHK17] and [GZS15] (Fig. 8.2), the authors have tried to optimise the gains of far-end and near-end filters for a stable communication setup. In such a framework with SE as front-end for LE models, the intelligibility gains depend on the speech restoration capability of the SE module. For instance, statistical-based SE front-ends have been reported to introduce large speech distortions in low SNR far-ends [EMLF06, EMLF05]. To this end, a multi-band SSDRC (MBSSDRC), which combines SE and LE stages, was suggested in [ZS17] (Fig. 8.3). For MBSSDRC, speech modifications at near-end were adapted to the level of the input (far-end) noise. Although this approach has helped to lower the distortions of enhanced speech, its intelligibility gains for the near-end listener were reduced compared to baseline SSDRC.

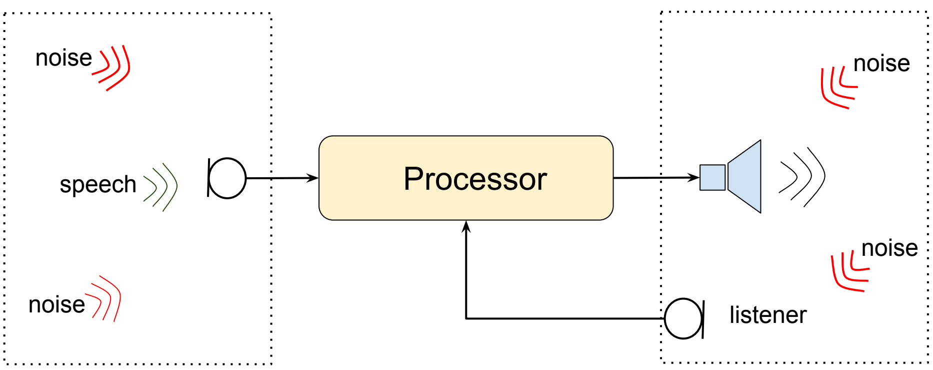


Figure 8.1 – Typical end-to-end communication scenario.

So far, the problem of near-end intelligibility enhancement with noisy far-end speech has been addressed mainly with classical signal processing techniques. However, the advent of neural networks has wide-open new opportunities. Inspired by the results observed in the previous chapter with neural networks for enhancing the intelligibility of clean speech, this chapter investigates the robustness of that system with noisy input. To the best of our knowledge, this is the first study using neural networks for improving near-end intelligibility in noisy far-end settings.

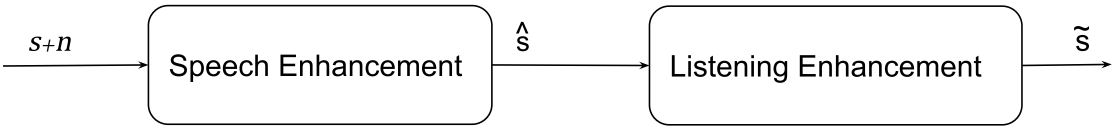


Figure 8.2 – Modular setup for joint speech and listening enhancement [GZS15].

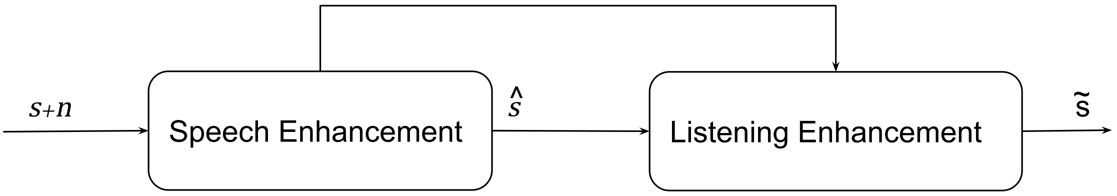


Figure 8.3 – Multi-band SSDRC method [ZS17].

We present a method that replaces the SE and LE stages with a single neural network. The proposed system operates on raw speech samples and is based on a CNN architecture with decreasing dilation factor for kernels over the layers. Previous framework was first introduced for speech synthesis as FFTNet [JFML18], and was extended for speech enhancement in the first chapter of this thesis showing promising results over other architectures. In this paper, we perform experiments with both the causal and non-causal extensions of FFTNet. A Teacher–Student strategy is followed for network training, where the Teacher is a well established expert-driven intelligibility enhancement method (SSDRC), and the Student is FFTNet.

8.1 The Problem Definition

In this paper we assume an additive model for the noise. Given the end-to-end communication scenario in Fig. 8.1, denoting $m(t)$ as the clean far-end speech and $n_f(t)$ as the background noise, the mixture signal at the speaker's side is defined as,

$$x(t) = m(t) + n_f(t). \tag{8.1}$$

Suppressing the noise before encoding the mixed signal for transmission is critical because speech codecs have poor robustness with noisy data [JMV$^+$00]. Therefore, speech enhancement strategies are often used before speech encoding [VSL02]. As depicted in Fig. 8.1, the near-end listener is also affected by noise in the surrounding, $n_n(t)$. Assuming no additional noise from the transmission channel, and in the absence of any processing, the signal perceived by the listener is

$$m_r(t) = x(t) + n_n(t), \tag{8.2}$$

which can be rewritten as

$$m_r(t) = m(t) + n_f(t) + n_n(t). \tag{8.3}$$

As such, the design of the processor in Fig. 8.1 should be concerned with the effects of noise in both far and near-ends. Indeed, as the far-end noise $n_f(t)$ is accessible at the processor's input, the noise statistics can be measured and later used to suppress the interference. In contrast, and assuming that the listener does not carry any on ear device for denoising, the near-end noise $n_n(t)$, acting as an auditory masker, can only be alleviated by modifying the speech components such that its intelligibility is promoted in this environment. Hence, an optimal processor should meet two requirements: (i) restore the speech that has been masked by the far-end noise, and (ii) modify the restored speech features to be more intelligible under noise for the near-end listener.

A trivial solution can be a two-stage neural processing, namely a neural denoising (e.g., FFTNet [MSATS19]) network, at far-end to restore speech from the noisy recording, followed by a neural intelligibility enhancement module (e.g., wSSDRC [MSTS18]), at near-end to boost the listener's intelligibility. However, such a modular approach would be prone to propagation of distortions from the front-end to back-end module. Informal listening tests have confirmed the propagation and amplification of these distortions. Therefore, we suggest an end-to-end solution which jointly solves both problems using a single deep neural network. As such, this model takes the noisy speech $x(t)$ in Eq. (8.1) as input and transforms it to a higher intelligible space $y(t)$ which meets the near-end listening requirements. Subsequently, the far-field noise is removed and the generated samples are more intelligible at the listener's side.

Transformation from a noisy to an intelligibility improved feature (or just speech) space can be formally defined as

$$\hat{y}_t = f(x_{t-r1}, \dots, x_{t-1}, x_t, x_{t+1}, \dots, x_{t+r2}; \boldsymbol{W}). \tag{8.4}$$

Given r_1 left and r_2 right input frames, the NN model f outputs the enhanced frame at current time index t. $\boldsymbol{W}$ represents the model's weights, and it is learned in a supervised fashion during network training. The conditional dependency on the past $[x_{t-r1}, ..., x_{t-1}]$ and future $[x_{t+1}, ..., x_{t+r2}]$ input samples is attained through the wide receptive field of the network. Model f is either causal ($r_2 = 0$) or non-causal ($r_2 > 0$), and it can be of various topologies. Generally, the mapping function described in Eq. (8.4) can be learned through many neural network architectures. However, our experiments have shown that end-to-end models based on CNNs with raw speech input outperform other architectures employing standard acoustic features as input. Both the causal and the non-causal variants of the proposed CNN topology were investigated and they are described below.

8.2 Proposed neural models

8.2.1 Causal convolutional FFTNet neural network

Convolutional neural networks are popular for their ability to isolate temporal patterns of speech with kernels of fixed size. For this work, convolution kernels with variable dilation factors are used as the basic building blocks [YK15]. In a causal dilated convolution with factor d, the activation of present sample is determined by the present and the $(d)^{th}$ past sample instances. The dilation helps to reduce the computational complexity of the final model by removing redundant computations that would have been caused by normal convolution kernels.

In the proposed architecture, the dilation rate reduces by a factor of 2 as we pass from one layer to the next (Fig. 8.4). Since the resulting dilation pattern resembles to the Butterfly structure of classical Fast Fourier Transform's (FFT) coefficient computation, the network is denoted FFTNet. A large receptive field helps to capture long-term temporal variations of input speech.

A similar architecture was originally proposed for speech synthesis as an auto-regressive model in [JFML18]. However, since the autoregressive models require heavy computations in the generation stage, our model is designed as a usual feed-forward network. While this modification has removed the stochasticity in sample generation, the acoustic modelling ability is still preserved. Furthermore, the speech samples are generated in parallel, which is highly desirable in real-life applications.

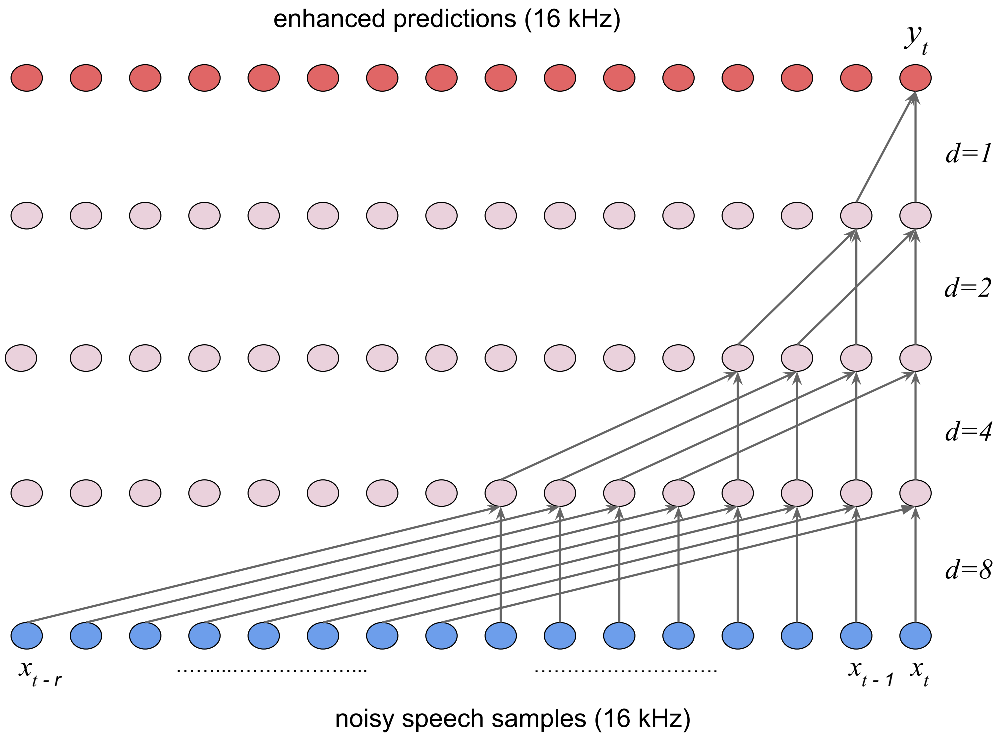


Figure 8.4 – Convolution pattern of causal FFTNet with dilation rate (d) decreasing from bottom to top layers.

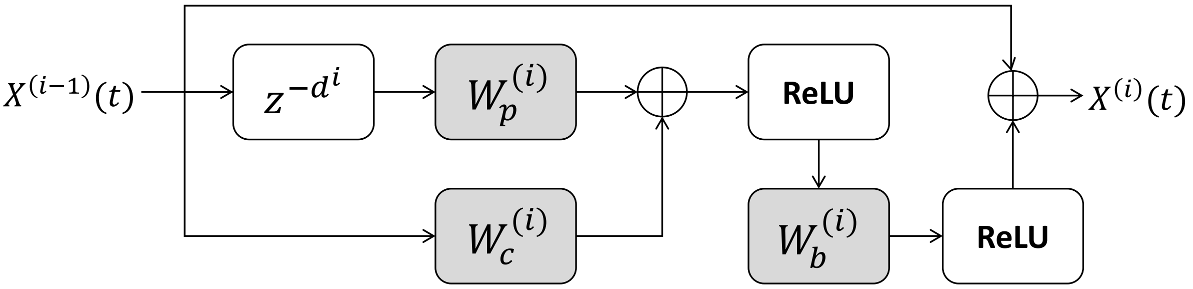


Figure 8.5 – Layer structure of causal FFTNet.

The structure of an FFTNet layer is shown in Fig. 8.5. Two convolution operations are cascaded. The first one has dilation factor d corresponding to the current layer, which is then followed by a [1 × 1] convolution. A dilated kernel with rate d has two active coefficients separated by $d-1$ zeros, and its operation can be expressed in time domain as

$$h^i(t) = W_p^i \times X^{i-1}(t-d^i) + W_c^i \times X^{i-1}(t), \tag{8.5}$$

where W_p^i and W_c^i are the kernel coefficients at layer i, and the variables $X^{i-1}(t)$ and $X^{i-1}(t-d^i)$ are the current and the $(d^i)^{th}$ past samples of the input, respectively. The dilation rate d^i is an integer power of 2. The operation $\times$ denotes the vector product as the variables involved are vectors with a fixed channel dimension.

Subsequently, multiple feature streams can be extracted by independent sets of kernels with coefficients $(W_p^{(i)}, W_c^{(i)})$ as shown in Eq. (8.5). Hence, the channel state vector over time is

$$H^i(t) = \begin{pmatrix} h_1(t) \\ h_2(t) \\ \vdots \\ h_n(t) \end{pmatrix} = \begin{pmatrix} W^i_{p1} \times X^{i-1}(t-d^i) + W^i_{c1} \times X^{i-1}(t) \\ W^i_{p2} \times X^{i-1}(t-d^i) + W^i_{c2} \times X^{i-1}(t) \\ \vdots \\ W^i_{pn} \times X^{i-1}(t-d^i) + W^i_{cn} \times X^{i-1}(t) \end{pmatrix}, \quad (8.6)$$

where n is the channel dimension. The activation function used in our network is the rectified linear unit ($ReLU$), which is applied on the feature map $H^i(t)$, resulting

$$H^{(i)}(t) = ReLU(H^i(t)). \quad (8.7)$$

The $[1 \times 1]$ convolution with kernel $W_b^{(i)}$ operates over the depth of the input channels with unit time resolution. This creates a new feature stream with $ReLU$ activation that is then mixed with the skip connection from the layer input $X^{(i-1)}(t)$,

$$X^{(i)}(t) = ReLU(W_b^{(i)} \times H^{(i)}(t)) + X^{(i-1)}(t). \quad (8.8)$$

The skip connection facilitates bottom-up phase information flow and the top-down gradient back propagation. Besides, the skip connection alleviates the gradient vanishing problem associated with deep CNNs.

The output of the final layer $X^{i=I}(t)$ is transformed to the speech sample $\hat{y}(t)$ by an n node fully connected layer having unit time resolution

$$\hat{y}(t) = W_{Pr} \times X^{(i=I)}(t), \quad (8.9)$$

where I is the index of the final network layer.

8.2.2 Non-causal convolutional FFTNet neural network

In contrast to the causal model, better acoustic modelling is possible with the inclusion of future samples ($r_2 > 0$ in Eq. (8.4)). Hence, a non-causal extension of the suggested neural architecture is presented in this part. In a non-causal dilated convolution of factor d, the estimate of the current sample depends on the present and the d-th backward and forward samples of the input signal. Reducing the dilation rate by a factor of 2 from bottom to top layers, the non-causal network has the structure shown in Fig. 8.6. The expanded context is expected to improve performance, especially in terms of noise suppression since the statistics at both sides of the target sample are considered. Same as for the causal version, this system is also designed as a regression model which processes the entire input segment in a single forward pass.

The structure of a non-causal FFTNet layer is depicted in Fig. 8.7. An additional coefficient is employed in the dilated convolution block to count the contribution of future samples. Therefore, the dilated convolution kernel generates the feature stream

$$h^i(t) = W^i_p \times X^{i-1}(t-d^i) + W^i_c \times X^{i-1}(t) + W^i_f \times X^{i-1}(t+d^i), \quad (8.10)$$

where $X^{i-1}(t)$, $X^{i-1}(t-d^i)$ and $X^{i-1}(t+d^i)$ are the current and the $(d^i)^{th}$ past and future samples of the input, respectively.

The output of n-channel convolution is

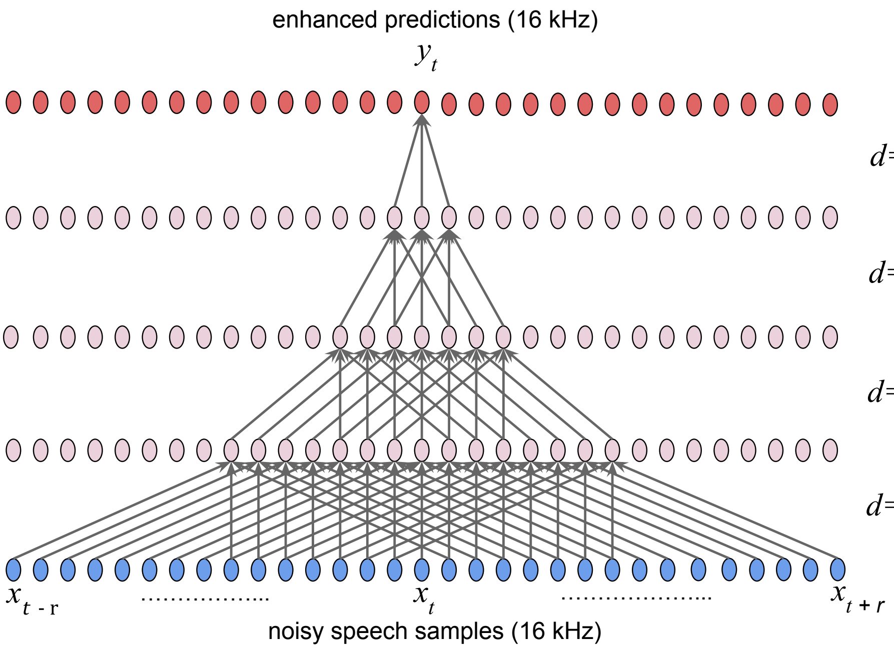


Figure 8.6 – Convolution pattern of non-causal FFTNet with dilation rate (d) decreasing from bottom to top layers.

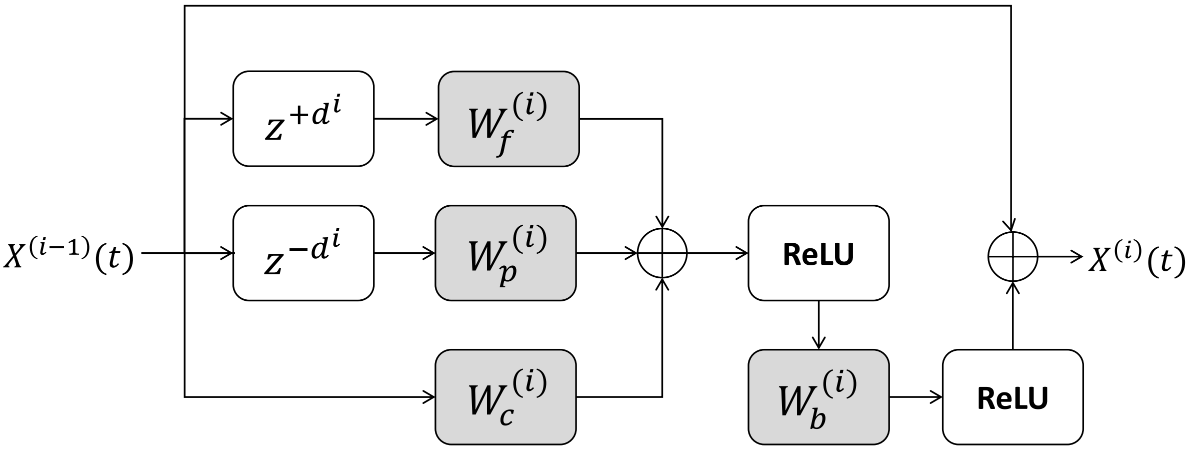


Figure 8.7 – Layer structure of non-causal FFTNet.

$$H^i(t) = \begin{pmatrix} W^i_{p1} \times X^{i-1}(t-d^i) + W^i_{c1} \times X^{i-1}(t) + W^i_{f1} \times X^{i-1}(t+d^i) \\ W^i_{p2} \times X^{i-1}(t-d^i) + W^i_{c2} \times X^{i-1}(t) + W^i_{f2} \times X^{i-1}(t+d^i) \\ \vdots \\ W^i_{pn} \times X^{i-1}(t-d^i) + W^i_{cn} \times X^{i-1}(t) + W^i_{fn} \times X^{i-1}(t+d^i) \end{pmatrix}, \tag{8.11}$$

which is followed by the same operations as in Eqs. (8.7) – (8.9) to get the filtered representations at the output block.

8.3 Model Training Strategy

A Teacher–Student approach was selected for training the suggested system, as depicted in Fig. 8.8. The Teacher is a well-established signal processing approach for near-end listening enhancement designed to work with clean speech input, and the Student is FFTNet. Although the Teacher is not a pre-trained network as in other similar frameworks, such as in Parallel WaveNet[**?**], it still dictates the quality of samples generated by the Student network. While training, the

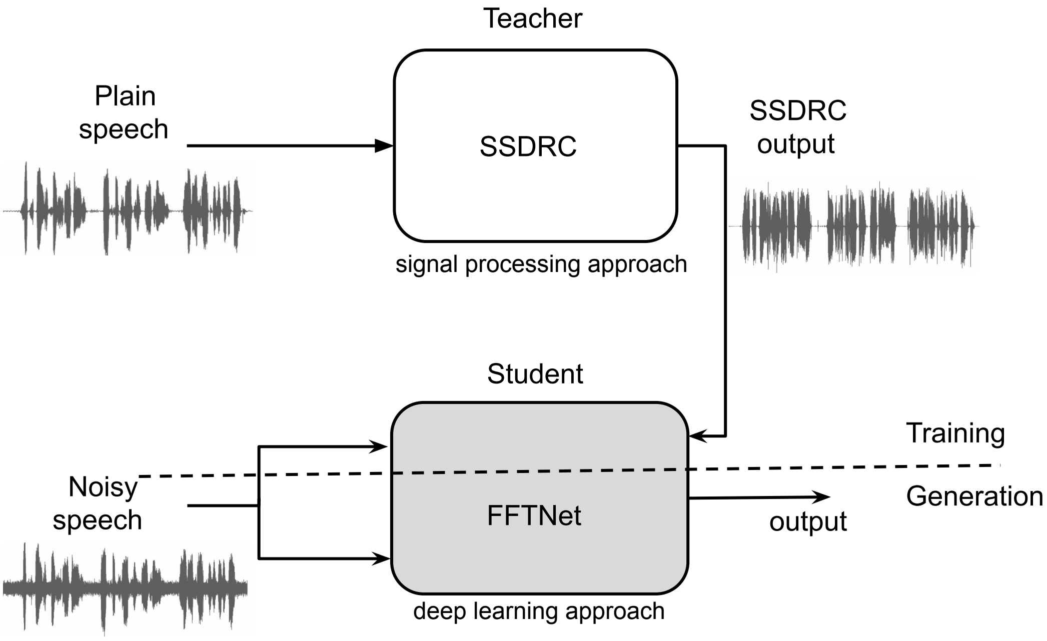


Figure 8.8 – Training strategy of proposed method.

knowledge about intelligibility defining feature modifications is transferred to the Student. The training was performed using parallel data with noisy speech as input and SSDRC enhanced clean samples as target. Therefore, the FFTNet Student is performing both noise suppression of input speech, and near-end intelligibility enhancement by mimicking the Teacher's behaviour.

Concerning the Teacher, although speech intelligibility in noise is determined by both acoustic and linguistic features, only the contribution of acoustic features was considered in this work. There are many expert-driven approaches in the literature that aim to achieve improved speech-in-noise intelligibility through spectral/temporal modifications, such as the maximization of the glimpse proportion count [TC11] or the maximization of the mutual information [KHK17]. However, we have chosen the SSDRC method presented in [ZKS12] as the Teacher module because it achieved state-of-the-art performance in large scale human intelligibility evaluations [CMVB13a] [CMVB$^+$13c].

SSDRC aims to improve the intelligibility of clean/plain speech in noise [ZKS12]. It has two cascaded systems that perform modifications over frequency (spectral shaping) and time (dynamic range compression) domains. In the spectral shaping stage, voiced speech frames are enhanced for formant sharpening, and the spectral tilt is reduced. In the dynamic range compression stage, the temporal envelope of the reconstructed waveform from the spectral shaping module is dynamically scaled to reduce its variations, which promotes reallocation of energy from the most sonorant to the less sonorant (or soft) speech regions. Fig. 8.9 illustrates the magnitude spectrum of a speech segment processed by SSDRC for improved intelligibility in noise. A reduced spectral tilt is noticeable due to the reallocation of spectral energy from the low to the mid and high frequency regions. This helps to preserve important speech cues in noisy conditions. More implementation details of SSDRC can be found in [ZKS12].

The loss function to be optimized during FFTNet training is defined in the waveform domain as the average absolute difference between the predicted and the target samples. Therefore, the cost function for causal FFTNet is defined as

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - r} \sum_{t=r}^{T^{(k)}} |y_t^{(k)} - \hat{y}_t^{(k)}|, \tag{8.12}$$

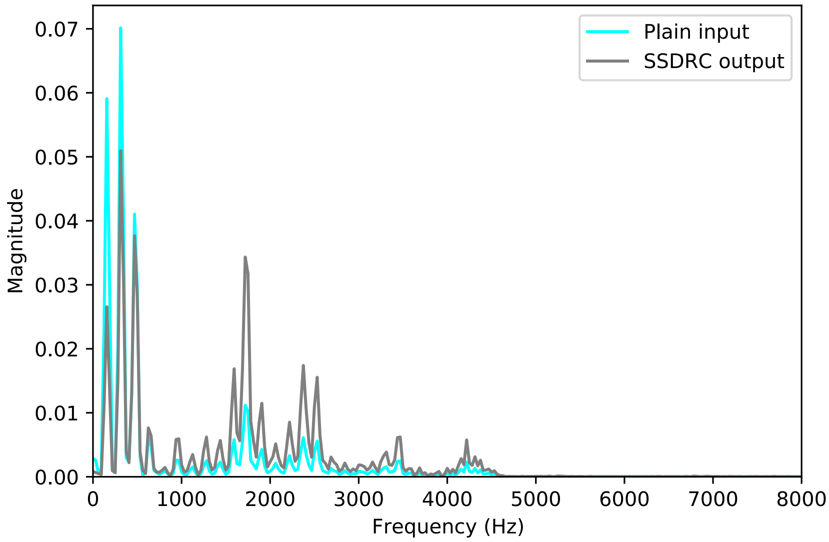


Figure 8.9 – Magnitude spectrum of a speech segment modified by SSDRC for improved intelligibility.

and the objective function for non-causal FFTNet is defined as

$$L(x^{(k)}, y^{(k)}) = \frac{1}{T^{(k)} - r} \sum_{t=r/2}^{T^{(k)}-r/2} |y_t^{(k)} - \hat{y}_t^{(k)}|. \tag{8.13}$$

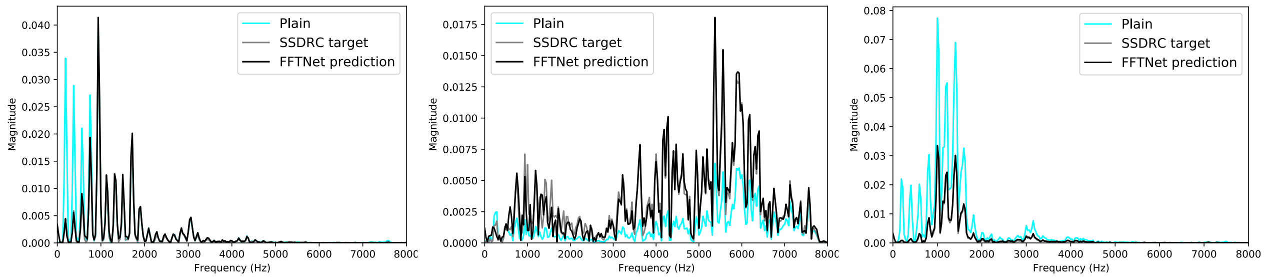
The variables in above equations are as follows:
$x^{(k)}$: input noisy speech segment at k^{th} training instance,
$y^{(k)}$: SSDRC modified samples of the clean speech corresponding to the noisy $x^{(k)}$,
$\hat{y}^{(k)}$: model predictions for the input $x^{(k)}$,
$T^{(k)}$: length of the input noisy segment $x^{(k)}$, and
r : length of the network's receptive field.

8.4 Experimental Evaluation

8.4.1 Database preparation for training and testing:

The far-end noisy data were simulated using clean speech and real noise recordings. Recordings from two speakers (one male and one female) uttering phonetically-balanced Greek Harvard (GrHarvard) sentences were used as clean speech. GrHarvard sentence corpus [Sfass] consists of 720 5-keyword sentences designed in the format of Harvard/IEEE sentences. The average duration of a recording is 3 seconds. A test set comprising of 240 recordings (120 samples for each speaker) was chosen for evaluation, and it remained unseen during neural network training. The model was trained on the remaining 1200 utterances of GrHarvard corpus. The noise data are from the DEMAND corpus [**?**], which contains 6 categories of perturbation (domestic, nature, office, public, street and transportation). There were multiple recordings in each of these categories, adding up to 15 total audio files. Five samples selected from each noise category were kept for testing and were not seen during neural network training. Since the original DEMAND recordings are multi-channel, only the first channel was used for our experiments. All speech and noise data had 16 kHz sampling frequency.

For the near-end environment, speech shaped noise (SSN) and competing speaker (CS) type of noise were considered. The stationary SSN was computed by passing white noise through a filter whose frequency response matches the long-



(a) Voiced segment

(b) Unvoiced segment

(c) Voiced segment

Figure 8.10 – Magnitude spectrum of FFTNet predictions contrasted against SSDRC labels that were used as target.

term average spectrum of the clean speech recordings. The non-stationary CS noise was another Greek female speaker form the ILSAP database[1]. The SNRs for the near-end mixtures were set at -9 dB, -4 dB and +1 dB for SSN, and at -21 dB, -14 dB, and -7 dB for CS. These SNR levels, for both conditions (SSN and CS), roughly correspond to 25%, 50% and 75% subjective intelligibility scores with native listeners, and they match with similar experiments in the literature [CMVB+13c].

8.4.2 Methods

The performance of proposed system was contrasted against a strong signal processing approach denoted Multiband SSDRC (MBSSDRC) [ZS17]. Being an extension of SSDRC, MBSSDRC was designed to work with noisy input. It consists of a far-end noise reduction stage, followed by adaptive modifications to improve intelligibility in adverse near-end conditions. The latter transformations are made adaptive to far-end noise statistics estimated in the former stage. MBSSDRC performs enhancement only on the magnitude spectra and keeps the noisy phase information for signal reconstruction.

Regarding the proposed end-to-end neural systems, both the causal (Fig. 8.4) and non-causal (Fig. 8.6) models have had in total 30 layers made by thrice repeating a convolution block whose dilation factors were [512, 256, 128, 64, 32, 16, 8, 4, 2, 1] with unit strides from the input. These yield a receptive field (r in Eqs. (8.12) and (8.13)) of size 3070 (3069 past samples plus the present one) for causal and 6139 (3069 past, one present and 3069 future samples) for non-causal networks. Hence, the context of 0.38s (for 16 kHz sampled signal) is captured while predicting every sample in the non-causal model, which is halved in the causal network. 256 convolution kernels were used in all layers. The top fully connected layer has a shape of [256,1] and is employed to merge the channel dimension to speech samples. Individual audio files were fed during model training, hence the batch size was one.

A more powerful noise reduction stage based on non-causal FFTNet (Fig. 8.6) was also cascaded with vanilla SSDRC to test the performance of the modular (non-end-to-end) design. This system is denoted DnsFFTNet+SSDRC, and it employs the non-causal FFTNet architecture described above.

All audio data were normalized to an RMS level of -24 dB to remove any loudness variations among the stimuli. The loss was minimized using Adam optimizer [KB14] with an exponentially decaying learning rate method, learning rate = 0.001, decay steps = 20000, and decay rate = 0.99. The training stops after 60 epochs and the last returned model is considered as the final model.

8.4.3 Procedure

The intelligibility gains of proposed and reference methods were assessed both objectively and subjectively. For the objective evaluation, the speech intelligibility index in bites (SIIB) [VKKH17] is used. SIIB is an intrusive intelligibility

[1] http://speech.ilsp.gr/

metric that counts the amount of information shared between a speaker and a listener in bits per second. In contrast to the standard speech intelligibility index (SII) [Ame97] or speech transmission index (STI) [HS71] metrics, SIIB partially accounts for the time-frequency dependencies in the signal and the speaker variability in the measurements. SIIB has the upper limit at 150 b/s and a higher score means better intelligibility.

The subjective intelligibility evaluation was conducted with 25 native Greek speakers aged between 20-35 years. All were normal hearing. They were asked to listen to speech stimuli and type what they heard using a computer interface. The stimuli were played only once following a Latin square design. The test was self-paced and lasted approximately 30 minutes to finish. To reduce listeners' fatigue, only the SSN condition was consider as the near-end masker. During testing, the SNRs for the far-end noise were 0 dB and +5 dB, while the SNRs for the SSN were -9 dB, -4 dB and +1 dB. For each condition there were 8 stimuli consisting of both male and female recordings from the test set. All signals were normalized in RMS before and after the modifications.

To statistically account for the intelligibility gains produced by different methods, an one-way analysis of variance (ANOVA) test was conducted. ANOVA is a comparative measure of variance among and between groups. ANOVA computes F-statistics, which is the ratio of inter-group to intra-group variability. Higher F-value indicates higher inter-group variability, which means that one group is dominant over the other. The p-value accompanying the F-value indicates that the probability of the predicted F-value could be random. Lower p value indicates higher confidence of the returned F-value. In the context of present study it is important to capture the intelligibility gains/losses produced by different enhancement techniques and also to quantify the statistical significance of these results.

8.5 Results and Discussion

First, we see how well FFTNet is trained to generate intelligible speech. Since SSDRC was chosen as the target intelligibility to be attained, the predictions of FFTNet are contrasted against it. The intelligibility modifications of FFTNet (non-causal architecture) in spectral domain for various segments of speech (from test set) are plotted in Fig. 8.10. In both voiced and unvoiced utterances, the modifications of FFTNet very closely match with that of SSDRC. It is observable that, through the modifications, some segments have gained energy over plain spectrum while others have lost. This is because energy was also redistributed over time among segments by dynamic range compression under an equal RMS constrain before and after processing. Examples in Fig. 8.10 show that FFTNet Student has learned the signal modifications critical for intelligibility from the SSDRC Teacher. Next, the intelligibility gains are measured both objectively and subjectively.

8.5.1 Objective evaluations

SIIB scores averaged over the test set at different levels of far-end and near-end adversities are displayed in Tables I and II for SSN and CS, respectively. Under the masking of both noise types, the far-end speech played without any pre-processing (Unprocessed) has severely been impaired by the noise. However, the speech intelligibility improves as the noise intensity reduces (or the SNR increases) in either end of the communication channel. While different speech modifiers have helped alleviate the information masking by noise at either end, their gains differ largely on operating SNR.

Table 8.1 – Objective SIIB scores for near-end SSN condition.

SIIB		Near-end noise level		
Far-end noise level	**Methods**	**-9 dB**	**-4 dB**	**+1 dB**
0 dB	Unprocessed	12.81	24.86	42.98
	MBSSDRC	18.65	24.21	29.73
	DnsFFTNet+SSDRC	23.48	38.21	54.70
	causal FFTNet	26.56	40.35	56.22
	non-causal FFTNet	34.51	61.40	99.96
5 dB	Unprocessed	16.97	33.85	58.05
	MBSSDRC	23.85	33.89	43.00
	DnsFFTNet+SSDRC	28.23	48.66	70.49
	causal FFTNet	30.93	49.88	71.25
	non-causal FFTNet	35.72	63.98	102.24

Table 8.2 – Objective SIIB scores for near-end CS condition.

SIIB		Near-end noise level		
Far-end noise level	**Methods**	**-21 dB**	**-14 dB**	**-7 dB**
0 dB	Unprocessed	9.92	16.86	32.98
	MBSSDRC	9.68	13.16	19.34
	DnsFFTNet+SSDRC	13.86	23.76	37.86
	causal FFTNet	15.53	24.93	40.82
	non-causal FFTNet	25.77	50.46	92.38
5 dB	Unprocessed	12.80	24.61	47.31
	MBSSDRC	12.64	19.11	28.14
	DnsFFTNet+SSDRC	18.70	32.59	53.21
	causal FFTNet	19.40	33.30	53.42
	non-causal FFTNet	26.61	51.11	92.51

The objective results suggest that MBSSDRC produces intelligibility benefits in low SNRs for the stationary background, and no gains for the fluctuating one. For both conditions, the SIIB intelligibility of MBSSDRC speech becomes worse as the SNR increases, suggesting that the processing artifacts have a significant effect as the ambient noise is weaker at the listener's side. This is however expected since the MBSSDRC noise reduction stage is blind and assumes that the far-end background is fairly stationary, which is not the case with the real-world recordings presented in DEMAND set. The stronger noise reduction front-end provided by non-causal FFTNet for DnsFFTNet+SSDRC produces more intelligible speech than MBSSDRC in terms of SIIBs in all conditions. However, due to the weak coupling between the denoising stage and SSDRC, informal listening tests have indicated that the processing artifacts produced by DnsFFTNet+SSDRC are stronger than those introduced by MBSSDRC, therefore the latter system was chosen as a baseline for the subjective evaluation. Concerning the samples processed by the proposed end-to-end neural models, they are more intelligible than those from both reference systems across all SNRs under both speech shaped and competing speaker type of noises.

In the case of SSN, the non-causal FFTNet architecture has produced a 170% relative intelligibility gain at the lowest far-end near-end SNR combination over unprocessed speech. The relative gain reduces to 75% at the highest SNR of +5 dB far-end and +1 dB near-end. The same network produces relative gains of around 180% and 100% in CS across the SNRs at 0 dB and 5 dB far-end, respectively, over the unprocessed speech. When compared against MBSSDRC, the non-causal network produced maximum relative improvements of 230% and 380% in SSN and CS conditions, respectively. The high SIIB scores of neural models even at high SNRs are clear evidence that the far-end noise has been suppressed largely without affecting the underline speech message. This high and consistent performance of the neural networks in unseen noise conditions is very promising for real-world applications where a system deals with diverse acoustic conditions.

Comparing the causal and non-causal neural network models, intuitively, the non-causal model should generate better quality samples as the future context has also been included in the modelling. This hypothesis has been confirmed by results in Tables 8.1 and 8.2, which show that the non-causal FFTNet clearly outperforms the causal version. The large relative improvements at higher near-end SNRs indicate that the far-end signal has been successfully cleaned. Given these results, only the non-causal FFTNet model was considered for the subjective evaluation.

8.5.2 Subjective evaluations

The subjective evaluation was conducted for 0 dB and 5 dB far-end ambient levels, and the results are presented in Fig. 8.11 and Fig. 8.12, respectively. The lower and upper sides of the boxes are the lower and upper quartiles. The boxes cover 50% of data distribution. Median of the distribution is represented by the horizontal line inside the box. The whiskers are the two lines outside the box that extends up to 1.5 times the interquartile range (or box width) from the lower and upper quartiles. Samples beyond that are labelled outliers in each conditions.

The boxes represent data dispersion at each condition while the line inside is the median point. The outliers in each condition are marked with circles. As such, in the discussion which follows, the gains are presented at the median point.

Firstly, the effect of far-end noise adversity on the processor's performance is analysed. On the null hypothesis that individual method produces equal intelligibility boost both at 0 dB and 5 dB far-end SNR, we ran independent ANOVA tests at -9 dB, -4 dB and +1 dB near-end SNR. The median intelligibility differences at these points together with the confidence of ANOVA scores are as follows. For a listener at -9 dB near-end SNR, the intelligibility gain by MBSSDRC is reduced by 26.6% points (F = 69.10, p = 1.01 $\times$ 10^{-10}) as the far-end acoustic deteriorates from 5 dB to 0 dB, while the non-causal FFTNet produces a relatively stable gain with only 15.4% difference (F = 39.16, p = 1.01 $\times$ 10^{-7}). At -4 dB near-end SNR, a similar difference of 27.3% (F = 27.6, p = 3.35 $\times$ 10^{-6}) is observed for MBSSDRC, while the disparity is substantially lower to 6.1% (F =23.6, p = 1.3 $\times$ 10^{-5}) for non-causal FFTNet. At the highest near-end SNR of +1 dB, the performance difference of MBSSDRC is still high at 19.4% (F = 16.75, p = 1.25 $\times$ 10^{-3}), while it is lower to 7.3% (F = 11.39, p = 1.90 $\times$ 10^{-4}) for non-causal FFTNet. These large and statistically significant differences in the intelligibility of MBSSDRC may be explained by the processing artifacts introduced by the denoising module as the input SNR deteriorates. On the other hand, the smaller differences for the neural network is a clear indication that FFTNet produces an almost equal intelligibility boost irrespective of the input noise's level.

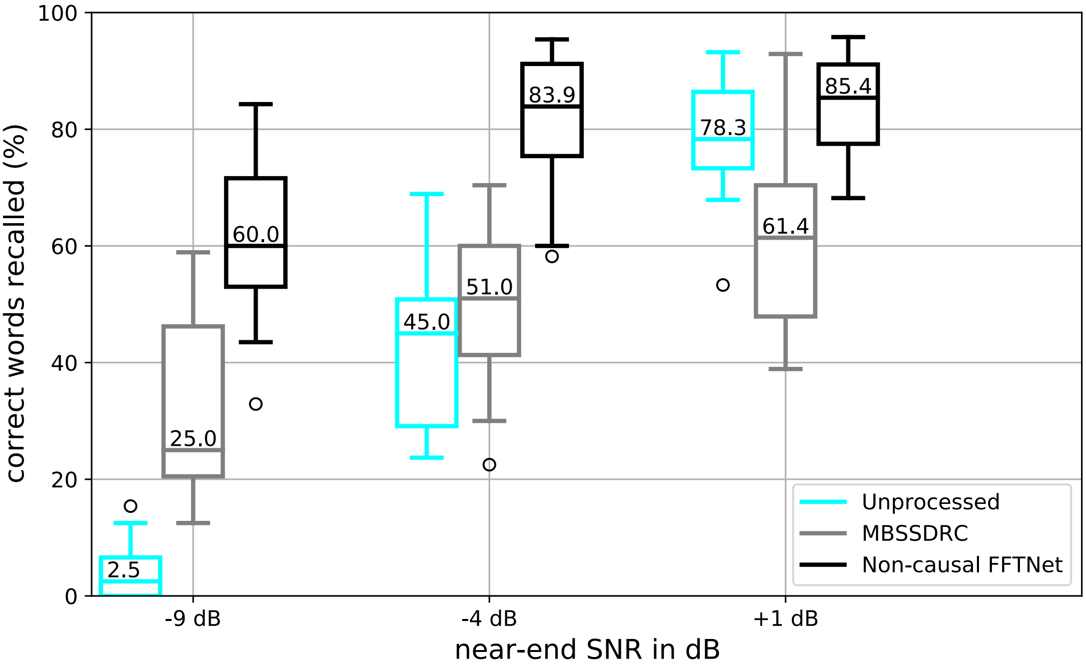


Figure 8.11 – Noisy far-end speech (0 dB SNR) in near-end SSN at various SNRs.

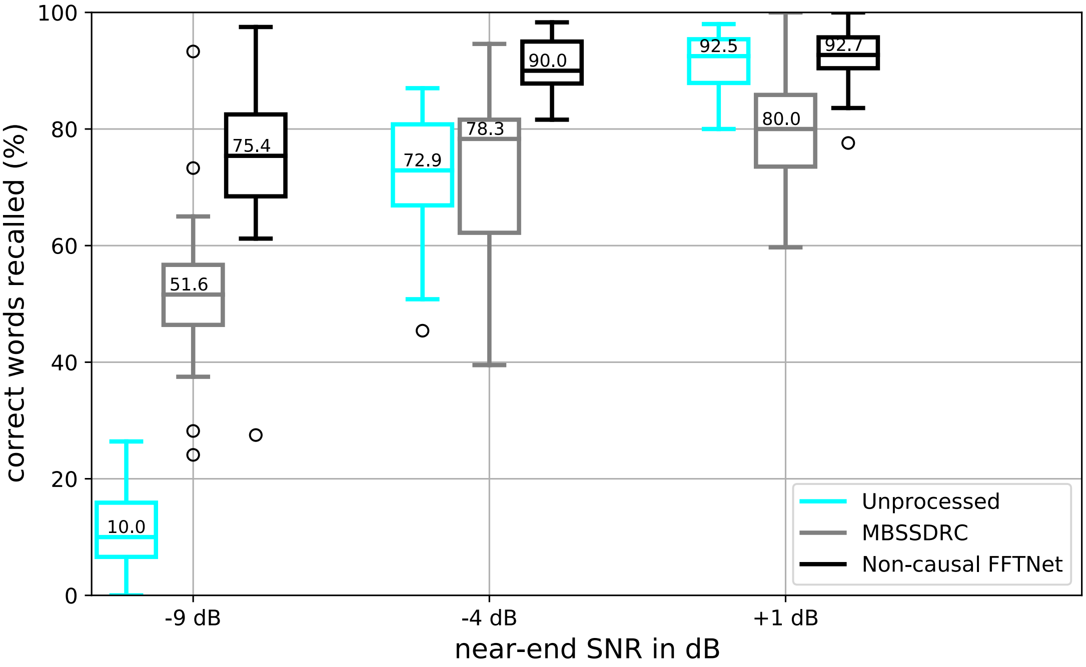


Figure 8.12 – Noisy far-end speech (5 dB SNR) in near-end SSN at various SNRs.

The intelligibility of processed samples at various near-end masking conditions are analysed next. Given the finding that the level of far-end noise affects the performance of MBSSDRC, this analysis is done separately for 0 dB and 5 dB far-end SNRs. When the far-end was at 0 dB (Fig. 8.11), only 2.5% of the unprocessed message was intelligible for the listener under -9 dB near-end masking. However, with the processing of MBSSDRC, the intelligibility was increased by

22.5% (F = 89.0, p = 1.64 $\times$ 10^{-12}). This gap is reduced to 6% (F= 5.18, p=0.027) when the near-end noise intensity reduces to -4 dB SNR. However, MBSSDRC has lowered the intelligibility for a listener at +1 dB SNR, causing an intelligibility loss of 17% (F = 24.6, p = 9.15 $\times$ 10^{-6}) with reference to unprocessed speech. On the other hand, the non-causal FFTNet produces intelligibility improvements of 57.5% (F = 417.8, p = 2.47 $\times$ 10^{-25}) and 38.9% (F = 135.9, p = 1.3 $\times$ 10^{-15}) in -9 dB and -4 dB SNRs, respectively, over the unprocessed speech. More promising is the performance at the highest near-end SNR of +1 dB, where non-causal FFTNet has improved the intelligibility by 7.10% (F= 5.23, p = 0.026).

When the SNR of far-end increased to 5 dB (Fig. 8.12), the median intelligibility of unprocessed speech is raised to 10% at the lowest near-end SNR. Similarly, the performance of MBSSDRC has also improved by producing the highest relative intelligibility boost of 41.6% (F = 153.6, p = 1.43 $\times$ 10^{-16}) over unprocessed speech. However, as the near-end SNR improves to -4 dB, the relative gain reduces to 5.40% (F = 0.0001, p = 0.99), then, at +1 dB near-end SNR, an intelligibility loss of 12.5% (F = 21.5, p = 2.8 $\times$ 10^{-5}) compared to the unprocessed speech is observed. On the other hand, the non-causal FFTNet model shows a stable performance across the near-end conditions. With reference to the unprocessed speech, absolute improvements of 65.4% (F = 387.5, p = 4.8 $\times$ 10^{-24}) and 17.1% (F = 54.5, p = 1.89 $\times$ 10^{-9}) were observed respectively at -9 dB and -4 dB SNRs. Their median gains almost overlap at the highest SNR of +1 dB.

Contrasting the performances of MBSSDRC and non-causal FFTNet, the neural model has shown relative improvements of 140% (F = 54.8, p = 1.75 $\times$ 10^{-9}), 64% (F = 92.89, p = 8.34 $\times$ 10^{-13}) and 39% (F = 45.3, p = 1.87 $\times$ 10^{-8}) at -9 dB, -4 dB and +1 dB near-end SNRs, respectively, for the 0 dB far-end. However, the relative gains were smaller when the input (far-end) SNR improved to 5 dB, with 46% (F = 30.7, p = 1.39 $\times$ 10^{-6}), 15% (F = 33.2, p = 5.6 $\times$ 10^{-7}) and 16% (F = 22.5, p = 1.99 $\times$ 10^{-5}) at -9 dB, -4 dB, +1 dB near-end SNRs, respectively.

8.5.3 Analysis of intelligibility enhanced samples in time and frequency domains

This disparity in performance between models can be visually observed in the enhanced samples, as shown in Fig. 8.13 for far-end noise intensity of 5 dB SNR. As visible from the plot, MBSSDRC introduces significant distortions in the signal by amplifying noise regions. In contrast, FFTNet has been more effective in identifying the active speech components in the noisy recording, which resulted in cleaner speech features to enhance for improved near-end intelligibility. Comparing the non-causal and causal models, the non-causal FFTNet architecture suppressed the noise more effectively.

A more in-depth analysis can be achieved by plotting the spectrogram (Fig. 8.14). The far-end speaker was exposed to 0 dB ambient noise, while the noise at the listener's end (near-end) varied from -8 dB to 30 dB from the top to the bottom panels. The near-end noise was speech shaped. The clean far-end message signal is also plotted for visual comparison purposes. As it can be seen in the top panel, for the unprocessed noisy signal only few glimpses of speech are available to the listener. However, by using enhancement more time-frequency speech regions become visible to the listener under the same RMS constraint. The signal artifacts introduced by MBSSDRC are more visible in quiet (Fig. 8.14c), where extended speech portions are lost due to an aggressive far-end noise reduction stage. On the other hand, the proposed FFTNet approaches have restored most of the speech components from the noisy input, which promote near-end intelligibility gains (Figs. 8.14a and 8.14b).

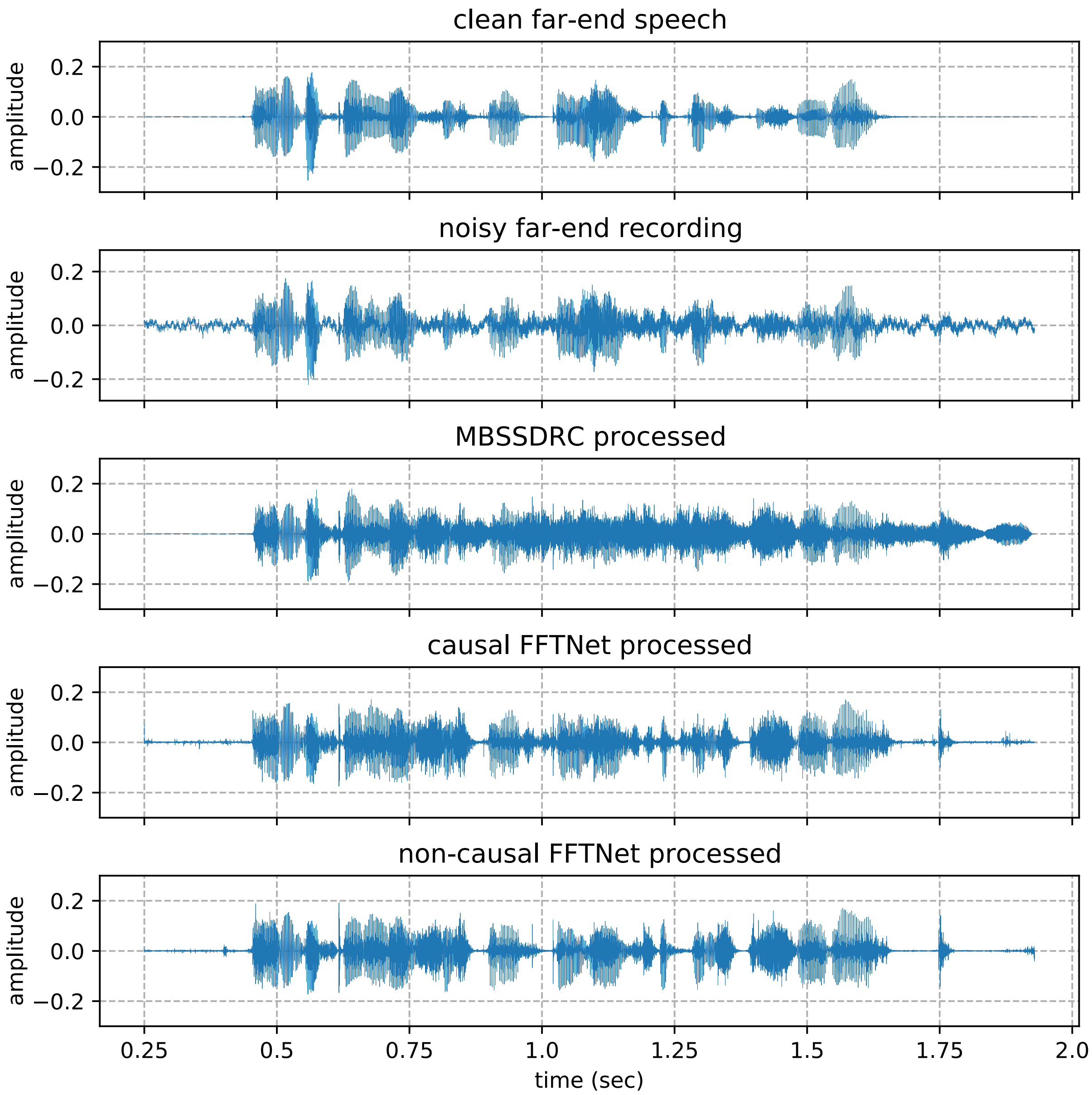


Figure 8.13 – Example of a speech sample enhanced using various methods (5 dB far-end noise).

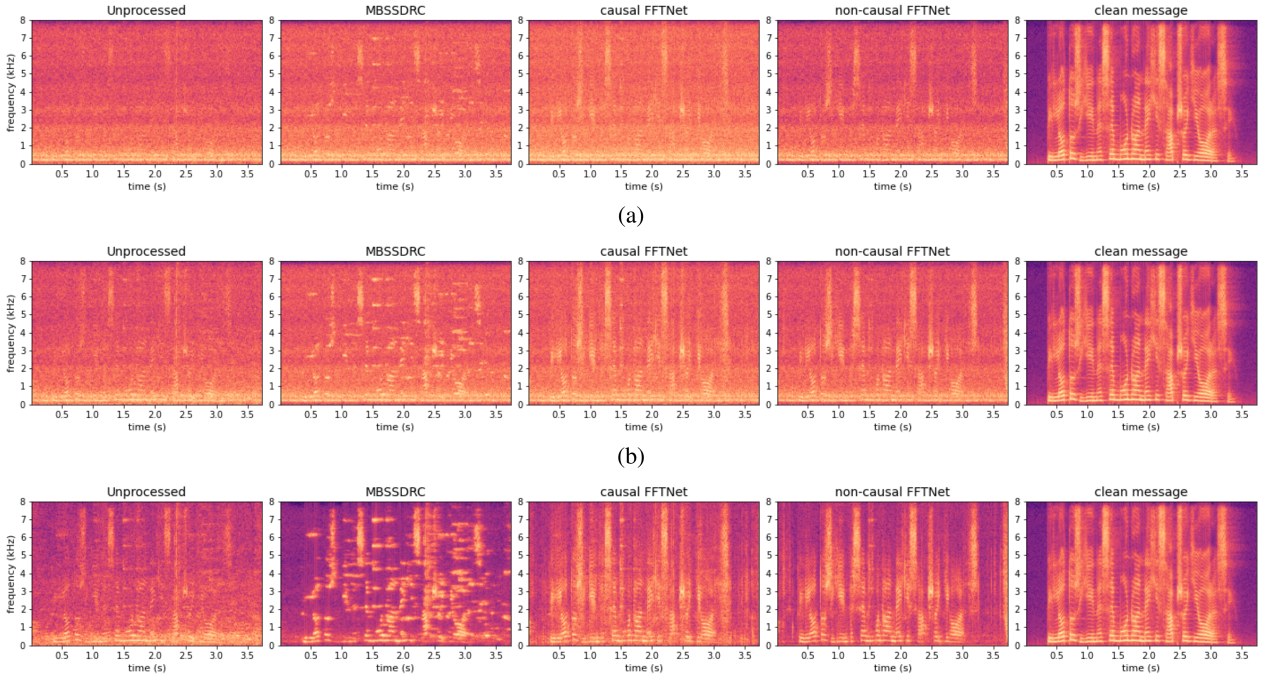
8.5.4 Real-time factor (RTF) analysis

The real-time factor (RTF) of proposed methods was assessed on an Intel 2.2 GHz CPU device. RTF is computed as the ratio between the time taken to process an input and the input's duration, therefore a lower RTF score indicates a faster algorithm. MBSSDRC is real-time at about 0.11 RTF score, however, the FFTNet models yielded 1.69 and 2.17 RTF scores for the causal and non-causal networks, respectively. Such a level of latency might be affordable in one-sided communication such as public address systems, e.g., announcements at airports or train stations. However, further refinement of the FFTNet architecture is required to improve its efficiency towards real-time and with low-latency requirement applications.

Audio samples[2] and a Tensorflow implementation of proposed model[3] are being provided.

[2] https://www.csd.uoc.gr/~shiaspv/IS2020-demo

[3] https://github.com/shifaspv/wSSDRC-tesnorflow-implementation



(a)

(b)

(c)

Figure 8.14 – Example of noisy far-end speech (0 dB SNR) enhanced for intelligibility in near-end SSN at: (a) -8 dB SNR, (b) 0 dB SNR, (c) 30 dB SNR.

8.6 Conclusions

In this chapter, an end-to-end approach for joint noise reduction and intelligibility enhancement was discussed. Such a system is required to fully meet the real-world requirement where both the speaker at near-and and listener at far-end are being immersed in noise. Traditional approaches tune the coefficients of the speech enhancement module at far-end and the listening enhancement module at near-end, either jointly or independently, to come to an optimal listening experience. However, such a framework suffers from the error propagation from one system to the other, as was observed in the case of multi-band SSDRC (MBSSDRC) in the above evaluation.

To address this, a joint formulation of the problem was required. As such, an end-to-end formulation of SE and LE problems was initially framed. Subsequently, an optimal architecture to learn such a formulation was to be identified. Since the dilation pattern of SE-FFTNet (presented in Section I of this thesis) was observed to help learn better noise discriminative features, the same dilation pattern was followed to construct the end-to-end system. To learn the intelligibility style at output, the spectral shaping dynamic range compression (SSDRC) algorithm was employed as a teacher module to drag the predictions towards the intelligible space. Once trained adequately, the end-to-end system has learned to suppress the far-end noise (at input) and to boost the intelligibility of underlying speech. The model has improved the intelligibility in keyword correct identification from 2.5% to 60% at the worst far- and near-end combination. Compared to the MBSSDRC model, the end-to-end system has always topped, with the highest absolute intelligibility gap of 35% between the two systems. More interestingly, the MBSSDRC system was ranked lower in intelligibility than the unprocessed speech in quiet near-end conditions, indicating the poor coupling of the SE and LE systems. Whereas, the end-to-end system performed consistently in all test conditions. This was due to the noise robustness of neural networks compared to statistical models.

The findings that a neural network can learn the intelligibility modifications can be exploited to build real-world speech intelligibility enhancers on top of existing neural network front-ends. These developments encourage a new research direction for designing the next generation of techniques to boost speech-in-noise intelligibility of normal and hearing-impaired listeners, though more future work is required to make the current system real-time. Similarly, such an approach of projecting the noisy observations to an acoustically more intelligible space could also improve recognition accuracy for automatic speech recognizers (ASR).

Chapter 9

Conclusions and Future Work

9.1 Overview

In this thesis, we investigated the prospect of using neural networks for noise removal and the intelligibility enhancement of speech. Speech in the real world can be significantly disturbed by noise at signal acquisition or perception stages, which can substantially reduce the quality and intelligibility of the signal. Neural networking has become the dominant approach to model data structures in recent years. As such, we approached the aforementioned challenges of noise on speech from the neural networking perspective.

To remove the noise on speech while signal acquisition by device, three new architectures have been proposed – gruCNN-SE, BigruCNN-SE, and SE-FFTNet. The first two models are feature domain models which take the short-time Fourier transformed (STFT) 2D representation of the noisy signal as the input and predict the corresponding clean spectrum. Both gruCNN-SE and BigruCNN-SE were observed to produce better quality spectrum restoration of speech compared to the traditional models. However, such techniques ignore the phase information of speech while performing the enhancement with only enhancing the magnitude spectrum, which creates quality degradations at the output. Therefore, a new time-domain architecture – SE-FFTNet is proposed, which performs the enhancement in the samples of raw speech. Besides, SE-FFTNet has a new dilation pattern differing from the traditional waveform models, which is found robust in learning the statistical dissimilarity of speech and noise in the noisy observation. Further experimentations on SE-FFTNet demonstrate that the optimization of model weights with a loss function computed in the frequency domain rather than time domain on which the model returns its predictions would produce better quality speech with minimal perceivable artifacts.

To improve the perception of speech for listeners in adverse listening conditions, a WaveNet like neural model (wSSDRC) has been proposed. The wSSDRC is perhaps the first neural network to cite to improve the intelligibility of speech. The network performs the spectral shaping and dynamic range compression of input through their learned weights rather than with statistical computations in the standard SSDRC algorithm. When tested, wSSDRC is found to produce substantial intelligibility boosts for normal as well as hearing-impaired listeners in various noisy listening conditions. Compared to the SSDRC algorithm, wSSDRC produces the equal intelligibility gain which was expected as the weights of wSSDRC is optimized for SSDRC style distribution. With the finding that intelligibility can be learned at the weights of a network, we experimented the same concept in text-to-speech (TTS) synthesis engines. Subsequently, a set of new intelligible TTS models were emerged with optimizing the network on different speaking styles like Lombard. Experimentations revealed that a combination of Lombard and SSDRC styles is producing the highest intelligibility for TTS in noise, a relative keyword identification improvements of up to 455% in speech-shaped noise over standard synthesize engines.

Finally, with the observations that the neural networks can be effective in suppressing noise on the speech and improving the intelligibility for listening in noise, we proposed a combined system that does the noise removal as well as the intelligibility enhancement in a single pass. Such a system would avoid the need for having a speech enhancement front-end to assist listening enhancement models when operating in many practical scenarios. This new network had a similar dilation pattern as the SE-FFTNet which was observed to better discriminate noise at the input. The concept from wSSDRC system is used to optimize the parameters of the network. When tested with human listeners, the proposed network was found to produce a substantial intelligibility boost at different input (acquisition) and output (perception) noise combinations. Whereas, the traditional modular setup had produced intelligibility degradations due to the poor coupling between the SE and LE modules.

9.2 Future research directions

There are a few directions to explore in the future. First, although we had proposed multiple architectures to learn different aspects of speech, a more detailed analysis of the networks' internal learning attributes is missing. This was partly because there were not any advanced tools back then, it is not an easy task even now to analyze the representations in neural layers, but techniques such as the singular value decomposition (SVD) on hidden layers can be explored to dive deep into the learning analysis.

Second, the recurrent feature extraction technique with gruCNN cells can be extended to other speech processing systems to combine the feature extraction and temporal attention modeling, like in the Tacotron-2 text-to-speech synthesizer where those were modeled as two independent tasks. Similarly, although the current wSSDRC model is trained on the SSDRC style, the same network can be extended to various speaking styles. In such cases, the current loss function of the network may not be optimal and would have to be redefined, especially when the input and output signals are not time aligned as is the case with many natural styles like Lombard. Therefore, a new loss function with a distribution learning framework like Kullback–Leibler (KL) divergence must be tried. Besides, in many intelligibility studies, it was observed that modified speech is not always pleasant or natural to listen to especially in quiet listening conditions. Therefore, setting a better loss function that can account for the naturalness of speech would further make the system more preferable in the community.

Third, the combined system for SE and LE is computationally complex to be deployed in many real-world applications like hearing aids where the intelligibility of noisy recordings has to be enhanced. Techniques like weight pruning or sparsification must be explored in the future to make it operational in low-power devices. Besides, the system is found to improve the intelligibility for human listeners, but such an approach may also improve the recognition accuracy for machines, which has to be tested. Similarly, the evaluation of the benefits of the system for hearing-impaired listeners is still pending.

Finally, all these systems were tested on simulated scenarios with manual mixing of speech recorded in isolated studio settings and noise recorded separately at outdoors. However, humans would articulate speech differently when speaking in noise compared to in a studio setting. Such articulatory variations would range from the normal (low vocal effort) to Lombard (high vocal effort) styles based on the intensity of background masking. Besides, neural models are highly data-dependent therefore may face difficulty for out-of-training samples. As such, the current systems would have to be fine-tuned to those variations to make them fully operational in practice.

Appendix A

Publications

During this work, the following publications took place (in chronological order):

1. **Conference and online publications**

   (a) **MPV Shifas**, V Tsiaras, Y Stylianou,
   Speech intelligibility enhancement based on a non-causal WaveNet-like model,
   In Proc. Interspeech, pp. 1868–1872, 2018.

   (b) **MPV Shifas**, N Adiga ,V Tsiaras, Y Stylianou,
   A Non-Causal FFTNet Architecture for Speech Enhancement,
   In Proc. Interspeech, pp. 1826–1830, 2019.

   (c) **MPV Shifas**, C Chermaz, T Chimona, V Tsiaras, Y Stylianou,
   Benefits of the WaveNet-Based Speech Intelligibility Enhancement for Normal and Hearing Impaired Listeners,
   In Proc. International Congress on Acoustics (ICA), pp. 5721–5725, 2019.

   (d) **MPV Shifas**, C Santelli, Y Stylianou,
   Towards Neural-Based Single Channel Speech Enhancement for Hearing Aids,
   In Proc. International Congress on Acoustics (ICA), pp. 5745–5748, 2019.

   (e) **MPV Shifas**, N Adiga ,V Tsiaras, Y Stylianou,
   Perceptually trained end-to-end neural model for single-channel speech enhancement,
   In Speech in Noise workshop 2020 (SpiN 2020), 9-10 January 2020, Toulouse,France.

   (f) D Paul, **MPV Shifas**, Y Pantazis, Y Stylianou,
   Enhancing Speech Intelligibility in Text-To-Speech Synthesis using Speaking Style Conversion,
   In Interspeech, pp. 1361–1365, 2020.

   (g) **MPV Shifas**, C Santelli, Y Stylianou, V Tsiaras, *A fully recurrent feature extraction for single channel speech enhancement*,

[Online]. Available: arXiv:2006.05233, 2020.

(h) **MPV Shifas**, A Sfakianaki, T Chimona, Y Stylianou,

Evaluating the Intelligibility Benefits of Neural Speech Enrichment for Listeners with Normal Hearing and Hearing Impairment using the Greek Harvard Corpus,

[Online]. Available: arXiv:2011.06548, 2020.

(i) C Chermaz, **MPV Shifas**, S Raman, A Govender, D Paul, O Simantiraki,

Enriched Speech for Effortless Listening,

In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Show & Tell, ST-P4.2., 2020.

(j) T Raitio, P Petkov, J Li, **M Shifas**, A Davis, Y Stylianou

Vocal effort modeling in neural TTS for improving the intelligibility of synthetic speech in noise ,

In interspeech 2022 (submitted).

2. **Journals**

(a) **MPV. Shifas**, C. Zorilă and Y. Stylianou,

End-to-End Neural Based Modification of Noisy Speech for Speech-in-Noise Intelligibility Improvement,

In IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 162-173, 2022, doi: 10.1109/TASLP.2021.3126947.

(b) **MPV. Shifas**, V. Tsiaras and Y. Stylianou,

On Bidirectional Recurrent Convolution and its Application to Speech Enhancement,

In IEEE/ACM Transactions on Audio, Speech, and Language Processing, to be submitted.

Bibliography

[AEFDA$^+$14] Marwa A Abd El-Fattah, Moawad I Dessouky, Alaa M Abbas, Salaheldin M Diab, El-Sayed M El-Rabaie, Waleed Al-Nuaimy, Saleh A Alshebeili, and Fathi E Abd El-Samie. Speech enhancement with an adaptive wiener filter. *International Journal of Speech Technology*, 17(1):53–64, 2014.

[Aga18] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[ALC14] Vincent Aubanel, Maria Luisa Garcia Lecumberri, and Martin Cooke. The sharvard corpus: A phonemically-balanced spanish sentence resource for audiology. *International journal of audiology*, 53(9):633–638, 2014.

[Ame97] American National Standards Institute. *American National Standard: Methods for Calculation of the Speech Intelligibility Index, ANSI S3.5-1997.* Acoustical Society of America, 1997.

[BAA17] Bajibabu Bollepalli, Manu Airaksinen, and Paavo Alku. Lombard speech synthesis using long short-term memory recurrent neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5505–5509, 2017.

[BBM15] Melissa M Baese-Berk and Tuuli H Morrill. Speaking rate consistency in native and non-native speakers of english. *The Journal of the Acoustical Society of America*, 138(3):EL223–EL228, 2015.

[BC94] Guy J Brown and Martin Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994.

[BC20] Hans Rutger Bosker and Martin Cooke. Enhanced amplitude modulations contribute to the lombard intelligibility benefit: Evidence from the nijmegen corpus of lombard speech. *The Journal of the Acoustical Society of America*, 147(2):721–730, 2020.

[Bir95] Martin Birgmeier. A fully kalman-trained radial basis function network for nonlinear speech modeling. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 1, pages 259–264. IEEE, 1995.

[BJA$^+$19a] Bajibabu Bollepalli, Lauri Juvela, Manu Airaksinen, Cassia Valentini-Botinhao, and Paavo Alku. Normal-to-Lombard adaptation of speech synthesis using long short-term memory recurrent neural networks. *Speech Communication*, 110:64–75, 2019.

[BJA19b] Bajibabu Bollepalli, Lauri Juvela, and Paavo Alku. Lombard speech synthesis using transfer learning in a tacotron text-to-speech system. In *INTERSPEECH*, 2019.

[BJA$^+$19c] Bajibabu Bollepalli, Lauri Juvela, Paavo Alku, et al. Lombard speech synthesis using transfer learning in a Tacotron text-to-speech system. *in Proc. Interspeech*, pages 2833–2837, 2019.

[BK03] Jounghoon Beh and Hanseok Ko. A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 3, pages III–633. IEEE, 2003.

[BKH00] Ann R Bradlow, Nina Kraus, and Erin Hayes. Speaking clearly for learning-disabled children: Sentence perception in noise. *The Journal of the Acoustical Society of America*, 108(5):2603–2603, 2000.

[BKH03] AR Bradlow, N Kraus, and E Hayes. Speaking clearly for learning-impaired children: Sentence perception in noise. *Journal of Speech, Language, and Hearing Research*, 46(1):80–97, 2003.

[BMG93] Thomas Baer, Brian CJ Moore, and Stuart Gatehouse. Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times. *Journal of rehabilitation research and development*, 30:49–49, 1993.

[BMY19] Michael W Berry, Azlinah Mohamed, and Bee Wah Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

[Bol79] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979.

[BTP96] Ann R Bradlow, Gina M Torretta, and David B Pisoni. Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, 20(3):255, 1996.

[BYPC15] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.

[CCHM93] PM Crozier, BMG Cheetham, C Holt, and E Munday. Speech enhancement employing spectral subtraction and linear predictive analysis. *Electronics Letters*, 29(12):1094–1095, 1993.

[CGCB14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[CL12] Martin Cooke and Maria Luisa García Lecumberri. The intelligibility of lombard speech for non-native listeners. *The Journal of the Acoustical Society of America*, 132(2):1120–1129, 2012.

[CMV14] Martin Cooke, Catherine Mayo, and Julián Villegas. The contribution of durational and spectral changes to the lombard speech intelligibility benefit. *The Journal of the Acoustical Society of America*, 135(2):874–883, 2014.

[CMVB13a] Martin Cooke, Catherine Mayo, and Cassia Valentini-Botinhao. Intelligibility-enhancing speech modifications: the hurricane challenge. In *Interspeech*, pages 3552–3556, 2013.

[CMVB$^+$13b] Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, et al. Hurricane natural speech corpus. *LISTA Consortium, Language and Speech Laboratory, Universidad del Pais.*, 2013.

[CMVB$^+$13c] Martin Cooke, Catherine Mayo, Cassia Valentini-Botinhao, Yannis Stylianou, Bastian Sauert, and Yan Tang. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585, 2013.

[Com92] Dirk Van Compernolle. Dsp techniques for speech enhancement. In *Speech Processing in Adverse Conditions*, 1992.

[CVMG$^+$14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[DFP94] Rob Drullman, Joost M Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95(2):1053–1064, 1994.

[DP80] AJ Duquesnoy and Reinier Plomp. Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis. *The Journal of the Acoustical Society of America*, 68(2):537–544, 1980.

[EM85] Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE transactions on acoustics, speech, and signal processing*, 33(2):443–445, 1985.

[EMLF05] Nicholas WD Evans, John S Mason, Wei M Liu, and Benoît Fauve. On the fundamental limitations of spectral subtraction: An assessment by automatic speech recognition. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE, 2005.

[EMLF06] Nicholas WD Evans, John SD Mason, Wei-Ming Liu, and Benoıt Fauve. An assessment on the fundamental limitations of spectral subtraction. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

[EVT95] Yariv Ephraim and Harry L Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing*, 3(4):251–266, 1995.

[EW95] Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

[FAFZ12] Thibaut Fux, Véronique Aubergé, Gang Feng, and Véronique Zimpfer. Speaker's prosodic strategy for a large physical distance communication task. *Acoust. Soc. Am*, 45(1):47–53, 2012.

[Fan81] Gunnar Fant. The source filter concept in voice production. *STL-QPSR*, 1(1981):21–37, 1981.

[FHTL17] Szu-Wei Fu, Ting-yao Hu, Yu Tsao, and Xugang Lu. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2017.

[FSS$^+$16] Johannes Fahringer, Tobias Schrank, Johannes Stahl, Pejman Mowlaee, and Franz Pernkopf. Phase-aware signal processing for automatic speech recognition. In *INTERSPEECH*, pages 3374–3378, 2016.

[FZME$^+$02] M Faundez-Zanuy, S McLaughlin, Arianna Esposito, A Hussain, Jean Schoentgen, G Kubin, WB Kleijn, and Petros Maragos. Nonlinear speech processing: overview and applications. *Control and intelligent systems*, 30(1):1–10, 2002.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[GBD$^+$06] Maëva Garnier, Lucie Bailly, Marion Dohen, Pauline Welby, and Hélène Lœvenbruck. An acoustic and articulatory study of lombard speech: Global effects on the utterance. In *Ninth International Conference on Spoken Language Processing*, 2006.

[GKS14] Elizabeth Godoy, Maria Koutsogiannaki, and Yannis Stylianou. Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles. *Computer Speech & Language*, 28(2):629–647, 2014.

[GL84] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

[GM86] Brian R Glasberg and Brian CJ Moore. Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *The Journal of the Acoustical Society of America*, 79(4):1020–1033, 1986.

[GM88] Brian R Glasberg and Brian C Moore. Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech. *Scandinavian Audiology*, 1988.

[Gri68] John D Griffiths. Optimum linear filter for speech transmission. *The Journal of the Acoustical Society of America*, 43(1):81–86, 1968.

[GS86] Sandra Gordon-Salant. Recognition of natural and time/intensity altered cvs by young and elderly subjects with normal hearing. *The Journal of the Acoustical Society of America*, 80(6):1599–1607, 1986.

[GS87] Sandra Gordon-Salant. Effects of acoustic modification on consonant recognition by elderly hearing-impaired subjects. *The Journal of the Acoustical Society of America*, 81(4):1199–1202, 1987.

[GS12] Elizabeth Godoy and Yannis Stylianou. Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[GSC99] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[GSD$^+$18] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018.

[GWK$^+$18] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[GZS15] Anthony Griffin, Tudor-Cătălin Zorilă, and Yannis Stylianou. Improved face-to-face communication using noise reduction and speech intelligibility enhancement. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5103–5107. IEEE, 2015.

[Har18] Till S Hartmann. Seeing in the dark with recurrent convolutional neural networks. *arXiv preprint arXiv:1811.08537*, 2018.

[HC05] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[HF10] Joseph L Hall and James L Flanagan. Intelligibility and listener preference of telephone speech in the presence of babble noise. *The Journal of the Acoustical Society of America*, 127(1):280–285, 2010.

[HL07] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.

[HL08] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

[HL10] Yi Hu and Philipos C Loizou. On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, 127(1):427–434, 2010.

[HLL$^+$20] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv preprint arXiv:2008.00264*, 2020.

[HM04] Valerie Hazan and Duncan Markham. Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116(5):3108–3118, 2004.

[HOZC20] Mojtaba Hasannezhad, Zhiheng Ouyang, Wei-Ping Zhu, and Benoit Champagne. An integrated cnn-gru framework for complex ratio mask estimation in speech enhancement. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 764–768. IEEE, 2020.

[HS71] Tammo Houtgast and Herman JM Steeneken. Evaluation of speech transmission channels by using artificial signals. *Acta Acustica united with Acustica*, 25(6):355–367, 1971.

[HS85] Tammo Houtgast and Herman JM Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America*, 77(3):1069–1077, 1985.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[HS98] Valerie Hazan and Andrew Simpson. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24(3):211–226, 1998.

[HXY15] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JFML18] Zeyu Jin, Adam Finkelstein, Gautham J Mysore, and Jingwan Lu. Fftnet: A real-time speaker-dependent neural vocoder. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2251–2255. IEEE, 2018.

[JMV$^+$00] Peter Jax, Rainer Martin, Peter Vary, Marc Adrat, Imre Varga, Walter Frank, Marc Ihle, and AG Siemens. A noise suppression system for the amr speech codec. In *KONVENS*, pages 43–46. Citeseer, 2000.

[Joy03] James Joyce. Bayes' theorem. 2003.

[JSS16] EP Jayakumar, PV Muhammed Shifas, and PS Sathidevi. Integrated acoustic echo and noise suppression in modulation domain. *International Journal of Speech Technology*, 19(3):611–621, 2016.

[JT16] Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

[Jun96] Jean-Claude Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech communication*, 20(1-2):13–22, 1996.

[Kay93] Steven M Kay. *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[KB04] Jean C Krause and Louis D Braida. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1):362–378, 2004.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KDS96] Nina Kowalski, Didier A Depireux, and Shihab A Shamma. Analysis of dynamic spectra in ferret primary auditory cortex. ii. prediction of unit responses to arbitrary dynamic spectra. *Journal of Neurophysiology*, 76(5):3524–3534, 1996.

[Kei17] Keithito. The LJspeech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[KES+18] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419, 2018.

[KG97] Arun Kumar and Allen Gersho. Ld-celp speech coding with nonlinear prediction. *IEEE Signal Processing Letters*, 4(4):89–91, 1997.

[KHK17] Seyran Khademi, Richard C Hendriks, and W Bastiaan Kleijn. Intelligibility enhancement based on mutual information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(8):1694–1708, 2017.

[KLHL09] Gibak Kim, Yang Lu, Yi Hu, and Philipos C Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 126(3):1486–1494, 2009.

[KPBL07] Diane Kewley-Port, T Zachary Burkle, and Jae Hee Lee. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 122(4):2365–2375, 2007.

[Kra01] Jean Christine Krause. Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement. 2001.

[KS14] M. Koutsogiannaki and Y. Stylianou. Simple and artefact-free spectral modifications for enhancing the intelligibility of casual speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4648–4652, 2014.

[KS16] Maria Koutsogiannaki and Yannis Stylianou. Modulation enhancement of temporal envelopes for increasing speech intelligibility in noise. 2016.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[KvHPL17] Antonios Kyparissiadis, Walter JB van Heuven, Nicola J Pitchford, and Timothy Ledgeway. Greeklex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PloS one*, 12(2), 2017.

[LB03] Jacqueline S Laures and Kate Bunton. Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of communication disorders*, 36(6):449–464, 2003.

[LC09] Youyi Lu and Martin Cooke. The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51(12):1253–1262, 2009.

[LO79] Jae Soo Lim and Alan V Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

[Loi13] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2013.

[LTMH13] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440, 2013.

[Lu09] Y Lu. Production and perceptual analysis of lombard effect. *Department of Computer Science, The University of Sheffield (Ph. D. thesis)*, 2009.

[LV05] Thomas Lotter and Peter Vary. Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Advances in Signal Processing*, 2005(7):354850, 2005.

[LZ06] Sheng Liu and Fan-Gang Zeng. Temporal properties in clear speech perception. *The Journal of the Acoustical Society of America*, 120(1):424–432, 2006.

[Mar02] Rainer Martin. Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–253. IEEE, 2002.

[ME88] Allen A Montgomery and Rodney A Edge. Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults. *Journal of Speech, Language, and Hearing Research*, 31(3):386–393, 1988.

[MKS12] Seyed Hamidreza Mohammadi, Alexander Kain, and Jan PH van Santen. Making conversational vowels more clear. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[MM80] Robert McAulay and Marilyn Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2):137–145, 1980.

[Moo12] Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[MS05] Nima Mesgarani and Shihab Shamma. Speech enhancement based on filtering the spectrotemporal modulations. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–1105. IEEE, 2005.

[MSATS19] PV Muhammed Shifas, Nagaraj Adiga, Vassilis Tsiaras, and Yannis Stylianou. A non-causal fftnet architecture for speech enhancement. *Proc. Interspeech 2019*, pages 1826–1830, 2019.

[MSTS18] PV Muhammed Shifas, Vassilis Tsiaras, and Yannis Stylianou. Speech intelligibility enhancement based on a non-causal WaveNet-like model. In *Proc. Interspeech*, volume 2018, pages 1868–1872, 2018.

[MTKI96] Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech synthesis using hmms with dynamic features. In *1996 ieee international conference on acoustics, speech, and signal processing conference proceedings*, volume 1, pages 389–392. IEEE, 1996.

[Mur91] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197, 1991.

[Myu03] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

[NBP$^+$17] Gaurav Naithani, Tom Barker, Giambattista Parascandolo, Lars Bramsl, Niels Henrik Pontoppidan, Tuomas Virtanen, et al. Low latency sound source separation using convolutional recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 71–75. IEEE, 2017.

[NG76] R Niederjohn and J Grotelueschen. The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):277–282, 1976.

[NIGM18] Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*, 2018.

[NM93] Chrysostomos L Nikias and Jerry M Mendel. Signal processing with higher-order spectra. *IEEE Signal processing magazine*, 10(3):10–37, 1993.

[ODZ+16] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[OL81] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[OLB+18] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In *International Conference on Machine Learning*, pages 3918–3926, 2018.

[OLKP79] A Oppenheim, Jae Lim, Gary Kopec, and SC Pohlig. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 632–637. IEEE, 1979.

[ORC18] Robert Ighodaro Ogie, Juan Castilla Rho, and Rodney J Clarke. Artificial intelligence in disaster risk communication: A systematic literature review. In *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–8. IEEE, 2018.

[Pat95] Roy D Patterson. The auditory filterbank: Fletcher's functional model of hearing. *The Journal of the Acoustical Society of America*, 97(5):3378–3378, 1995.

[Pau81] D Paul. The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4):786–794, 1981.

[PB87] K Paliwal and Anjan Basu. A speech enhancement method based on kalman filtering. In *ICASSP'87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 177–180. IEEE, 1987.

[PBS17] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[PDB86] Michael A Picheny, Nathaniel I Durlach, and Louis D Braida. Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 29(4):434–446, 1986.

[PDD14] Benjamin Picart, Thomas Drugman, and Thierry Dutoit. Analysis and HMM-based synthesis of hypo and hyperarticulated speech. *Computer Speech & Language*, 28(2):687–707, 2014.

[PL16] Se Rim Park and Jinwon Lee. A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*, 2016.

[Plo86] Reinier Plomp. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech, Language, and Hearing Research*, 29(2):146–154, 1986.

[PMS06] Cyril Plapous, Claude Marro, and Pascal Scalart. Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2098–2108, 2006.

[Por80] M Portnoff. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):55–69, 1980.

[PPG+16] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for ASR based on lattice-free mmi. In *Interspeech*, pages 2751–2755, 2016.

[PPSed] Dipjyoti Paul, Yannis Pantazis, and Yannis Stylianou. Speaker conditional WaveRNN: Towards universal neural vocoder for unseen speaker and recording conditions. In *Proc. Interspeech*, 2020 (accepted).

[PS03] Lucas Parra and Paul Sajda. Blind source separation via generalized eigenvalue decomposition. *The Journal of Machine Learning Research*, 4:1261–1269, 2003.

[PWS10] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech communication*, 52(5):450–475, 2010.

[PWS11] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. The importance of phase in speech enhancement. *speech communication*, 53(4):465–494, 2011.

[PY09] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

[RCG$^+$69] EH Rothauser, WD Chapman, N Guttman, H Hecker, K Nordby, H Silbiger, G Urbanek, and M Weinstock. IEEE recommended practice for speech quality measurements. *IEEE Trans. on Audio and Electroacoustics*, 17:225–246, 1969.

[RG75] Lawrence R Rabiner and Bernard Gold. Theory and application of digital signal processing. *Englewood Cliffs: Prentice-Hall*, 1975.

[RN02] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

[RNP20] Sujan Kumar Roy, Aaron Nicolson, and Kuldip K Paliwal. A deep learning-based kalman filter for speech enhancement. In *INTERSPEECH*, pages 2692–2696, 2020.

[RPS18] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[RSVA11] Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku. Analysis of HMM-based Lombard speech synthesis. In *Proc. Interspeech*, 2011.

[RSVBC20] Jan Rennies, Henning F Schepker, Cassia Valentini-Botinhao, and Martin Cooke. Intelligibility-enhancing speech modifications-the hurricane challenge 2.0. In *INTERSPEECH*, pages 1341–1345, 2020.

[Rud16] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[RV05] Koenraad S Rhebergen and Niek J Versfeld. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117(4):2181–2192, 2005.

[RVD09] Koenraad S Rhebergen, Niek J Versfeld, and Wouter A Dreschler. The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise. *The Journal of the Acoustical Society of America*, 126(6):3236–3245, 2009.

[SAK$^+$17] Muhammad Salman, Abdul Wahab Ahmed, Omair Ahmad Khan, Basit Raza, and Khalid Latif. Artificial intelligence in biomedical domain. *Artificial Intelligence*, 8(8), 2017.

[SB06] Mitchell S Sommers and Joe Barcroft. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, 119(4):2406–2416, 2006.

[SCS$^+$20] Muhammed PV Shifas, Santelli Claudio, Yannis Stylianou, et al. A fully recurrent feature extraction for single channel speech enhancement. *arXiv preprint arXiv:2006.05233*, 2020.

[SCW$^+$15] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[Sfass] A Sfakianaki. Designing a modern greek sentence corpus for audiological and speech technology research. *In Proc. 14th International Conference on Greek Linguistics (ICGL14)*, 2019 (in press).

[SH06] Mark D Skowronski and John G Harris. Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48(5):549–558, 2006.

[Sic92] Giovanni L Sicuranza. Quadratic filters for signal processing. *Proceedings of the IEEE*, 80(8):1263–1285, 1992.

[SJA$^+$19] Shreyas Seshadri, Lauri Juvela, Paavo Alku, Okko Räsänen, et al. Augmented CycleGANs for continuous scale normal-to-lombard speaking style conversion. *Proc. Interspeech 2019*, pages 2838–2842, 2019.

[SJRA19] Shreyas Seshadri, Lauri Juvela, Okko Räsänen, and Paavo Alku. Vocal effort based speaking style conversion using vocoder features and parallel learning. *IEEE Access*, 7:17230–17246, 2019.

[SJS18] PV Muhammed Shifas, EP Jayakumar, and PS Sathidevi. Robust acoustic echo suppression in modulation domain. In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pages 527–537. Springer, 2018.

[SJY$^+$19] Shreyas Seshadri, Lauri Juvela, Junichi Yamagishi, Okko Räsänen, and Paavo Alku. Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6835–6839, 2019.

[SM92] Michael A Stone and Brian CJ Moore. Spectral feature enhancement for people with sensorineural hearing impairment: effects on speech intelligibility and quality. *Journal of rehabilitation research and development*, 29(2):39–56, 1992.

[SMG90] AM Simpson, BCJ Moore, and BR Glasberg. Spectral enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners. *Acta Oto-Laryngologica*, 109(sup469):101–107, 1990.

[Sov95] P Sovka. Extended spectral subtraction-description and preliminary results. *Research report*, 1995.

[SPB$^+$88] W Van Summers, David B Pisoni, Robert H Bernacki, Robert I Pedlow, and Michael A Stokes. Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928, 1988.

[SPK96] Pavel Sovka, Petr Pollak, and Jan Kybic. Extended spectral subtraction. In *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, pages 1–4. IEEE, 1996.

[SRD15] Henning Schepker, Jan Rennies, and Simon Doclo. Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index. *The Journal of the Acoustical Society of America*, 138(5):2692–2706, 2015.

[SSB14] Hasim Sak, Andrew W Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

[SSCS20] Muhammed PV Shifas, Anna Sfakianaki, Theognosia Chimona, and Yannis Stylianou. Evaluating the intelligibility benefits of neural speech enrichment for listeners with normal hearing and hearing impairment using the greek harvard corpus. *arXiv preprint arXiv:2011.06548*, 2020.

[Ste85] Samuel D Stearns. of aldapfive signal processing. 1985.

[SV06] Bastian Sauert and Peter Vary. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

[SZ09] Marc Schönwiesner and Robert J Zatorre. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fmri. *Proceedings of the National Academy of Sciences*, 106(34):14611–14616, 2009.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[TC11] Yan Tang and Martin Cooke. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[THHJ11] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[TIV13] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. Demand: a collection of multi-channel recordings of acoustic noise in diverse environments. In *Proc. Meetings Acoust*, pages 1–6, 2013.

[TNH94] Jes Thyssen, Henrik Nielsen, and Steffen Duus Hansen. Non-linear short-term prediction in speech coding. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–185. IEEE, 1994.

[TW18] Ke Tan and DeLiang Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, pages 3229–3233, 2018.

[Tyl86] RS Tyler. Frequency resolution in hearing-impaired listeners. *Frequency selectivity in hearing*, pages 309–371, 1986.

[UKB07] Maria Uther, Monja A Knoll, and Denis Burnham. Do you speak e-ng-li-sh? a comparison of foreigner- and infant-directed speech. *Speech communication*, 49(1):2–7, 2007.

[VBYKS13] Cassia Valentini-Botinhao, Junichi Yamagishi, Simon King, and Yannis Stylianou. Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of hmm-based synthetic speech in noise. In *Proc. Interspeech*, pages 3567–3571, 2013.

[VKKH17] Steven Van Kuyk, W Bastiaan Kleijn, and Richard C Hendriks. An instrumental intelligibility metric based on information theory. *IEEE Signal Processing Letters*, 25(1):115–119, 2017.

[VKKH18] Steven Van Kuyk, W Bastiaan Kleijn, and Richard Christian Hendriks. An evaluation of intrusive instrumental intelligibility metrics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2153–2166, 2018.

[VLL$^+$10] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[VO$^+$13] Pritesh Vora, Bhavesh Oza, et al. A survey on k-mean clustering and particle swarm optimization. *International Journal of Science and Modern Engineering*, 1(3):24–26, 2013.

[VOKK16] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016.

[VSL02] David Virette, Pascal Scalart, and Claude Lamblin. Analysis of background noise reduction techniques for robust speech coding. In *2002 11th European Signal Processing Conference*, pages 1–4. IEEE, 2002.

[VYK13] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013.

[WEW$^+$15] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 91–99. Springer, 2015.

[Whi82] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

[WL12] Kamil K Wójcicki and Philipos C Loizou. Channel selection in the modulation domain for improved speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 131(4):2904–2913, 2012.

[WNF94] Lizhong Wu, Mahesan Niranjan, and Frank Fallside. Fully vector-quantized neural network-based code-excited nonlinear predictive speech coding. *IEEE transactions on speech and audio processing*, 2(4):482–489, 1994.

[WNK$^+$99] Eric A Wan, Alex T Nelson, Shigeru Katagiri, et al. Networks for speech enhancement. *Handbook of neural networks for speech processing. Artech House, Boston, USA*, 139(1):7, 1999.

[WNW14] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 22(12):1849–1858, 2014.

[WRR03] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

[WS08] Peter J Watson and Robert S Schlauch. The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *American Journal of Speech-Language Pathology*, 2008.

[WSRS$^+$17] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint:1703.10135*, 2017.

[XDDL14] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1):65–68, 2014.

[XSWN20] Yang Xian, Yang Sun, Wenwu Wang, and Syed Mohsen Naqvi. A multi-scale feature recalibration network for end-to-end single channel speech enhancement. *IEEE Journal of Selected Topics in Signal Processing*, 15(1):143–155, 2020.

[Xu19] Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, 2019.

[YBBW19] Sarah E Yoho, Stephanie A Borrie, Tyson S Barrett, and Dane B Whittaker. Are there sex effects for speech intelligibility in american english? examining the influence of talker, listener, and methodology. *Attention, Perception, & Psychophysics*, 81(2):558–570, 2019.

[Yeg09] Bayya Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[YK15] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[ZKS12] Tudor-Catalin Zorila, Varvara Kandia, and Yannis Stylianou. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[ZS14] Tudor Cătălin Zorilă and Yannis Stylianou. On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement. In *Proc. Interspeech*, 2014.

[ZS17] Tudor-Catalin Zorila and Yannis Stylianou. On the quality and intelligibility of noisy speech processed for near-end listening enhancement. In *Interspeech*, pages 2023–2027, 2017.

[ZSFM17] Tudor-Cătălin Zorilă, Yannis Stylianou, Sheila Flanagan, and Brian CJ Moore. Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss. *The Journal of the Acoustical Society of America*, 141(1):189–196, 2017.

[ZSIA16] Tudor-Cătălin Zorilă, Yannis Stylianou, Tatsuma Ishihara, and Masami Akamine. Near and far field speech-in-noise intelligibility improvements based on a time–frequency energy reallocation approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1808–1818, 2016.

[ZZTL18] Han Zhao, Shuayb Zarar, Ivan Tashev, and Chin-Hui Lee. Convolutional-recurrent neural networks for speech enhancement. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2401–2405. IEEE, 2018.