Tampere University

Anant Mital

Speech Enhancement for Automatic Analysis of Child-Centered Audio Recordings

Faculty of Information Technology and Communication Sciences
Master of Science Thesis
December 2019

Abstract

Anant Mital: Speech Enhancement for Automatic Analysis of Child-Centered Audio Recordings
Master of Science Thesis
Tampere University
Master's Degree Programme in Information Technology
December 2019

Analysis of child-centred daylong naturalist audio recordings has become a de-facto research protocol in the scientific study of child language development. The researchers are increasingly using these recordings to understand linguistic environment a child encounters in her routine interactions with the world. These audio recordings are captured by a microphone that a child wears throughout a day. The audio recordings, being naturalistic, contain a lot of unwanted sounds from everyday life which degrades the performance of speech analysis tasks. The purpose of this thesis is to investigate the utility of speech enhancement (SE) algorithms in the automatic analysis of such recordings. To this effect, several classical signal processing and modern machine learning-based SE methods were employed 1) as a denoiser for speech corrupted with additive noise sampled from real-life child-centred daylong recordings and 2) as front-end for downstream speech processing tasks of addressee classification (infant vs. adult-directed speech) and automatic syllable count estimation from the speech. The downstream tasks were conducted on data derived from a set of geographically, culturally, and linguistically diverse child-centred daylong audio recordings. The performance of denoising was evaluated through objective quality metrics (spectral distortion and instrumental intelligibility) and through the downstream task performance. Finally, the objective evaluation results were compared with downstream task performance results to find whether objective metrics can be used as a reasonable proxy to select SE front-end for a downstream task. The results obtained show that a recently proposed Long Short-Term Memory (LSTM)-based progressive learning architecture provides maximum performance gains in the downstream tasks in comparison with the other SE methods and baseline results. Classical signal processing-based SE methods also lead to competitive performance. From the comparison of objective assessment and downstream task performance results, no predictive relationship between task-independent objective metrics and performance of downstream tasks was found.

Keywords: Speech enhancement, deep learning, naturalistic audio, syllabification.

Acknowledgements

All the research work for this thesis was conducted in the Speech and Cognition group at Tampere University. This thesis contributes to an international collaborative project called Analyzing Child Language Experiences Around the World (ACLEW). The research is funded by Academy of Finland grants no. 312105 and 320053. I would like to convey my sincere thanks to my supervisor Prof. Okko Räsänen for his patience, guidance and supervision. I am also thankful to team ACLEW for providing the data for this work. A big thank goes to Prof. Tom Bäckström, Prof. John Hansen, Prof. Björn Schuller, Lei Sun and Shreyas Seshadri for kindly sharing their models and algorithms without which this work might not have been possible.

Finally, I want to thank my parents, my sister and friends for their support.

Tampere, Decmeber 8, 2019
Anant Mital

List of abbreviations

ACLEW Analyzing Child Language Experiences Around the World.

ADS Adult-directed Speech.

ASCE Automatic Syllable Count Estimation.

ASR Automatic Speech Recognition.

DNN Deep Neural Network.

DSP Digital Signal Processing.

DTFT Discrete-Time Fourier Transform.

FFT Fast Fourier Transform.

IDS Infant-directed Speech.

ILD Inter-Aural Level Difference.

IRM Ideal Ratio Mask.

ITD Inter-Aural Time Difference.

LPC Linear Predictive Coding.

LSTM Long Short-Term Memory.

MMSE Minimum Mean-squared Error.

MS Minimum Statistics.

SAD Speech Activity Detector.

SE Speech Enhancement.

SIIB Speech Intelligibility in Bits.

SNR Signal-to-Noise Ratio.

STFT Short-Time Fourier Transform.

UAR unweighted Average Recall.

VAD Voice Activity Detector.

WCE Word Count Estimation.

WER Word Error Rate.

WGN White Gaussian Noise.

Contents

1 Introduction

Speech is the principal carrier of the human language [1]. Though humans also express themselves through non-verbal ways (e.g. humming, nodding, dancing, winking, blushing, hand gestures etc) but communication in this manner is not determined by the rules of the language. There are cogent reasons, (e.g. speech communication can be carried out in tandem with other activities such as food gathering or hunting) which have their roots in evolution of our species, that makes us use speech naturally and frequently [2]. Speech is not only fundamental to communication but also plays an important part in facilitating interaction. It remains our principal way to communicate abstract thoughts [3].

Speech is the outcome of a complex psycho-acoustic process involving brain (organization of thought into spoken language; motor control), lungs (creating the airflow by expulsion of air through throat, oral and nasal cavities), larynx (modulation of airflow through vocal folds; altering length of vocal tract) and organs of vocal tract (such as nasal and oral cavity, tongue, lips, teeth etc) [4], [5]. Even though both conceptualisation and articulation of speech is highly complex and specialised task, it is a common observation that young children (who are otherwise normal) are able to acquire their mother tongue (or native language) rapidly and effortlessly [6]. Moreover, it is quite astonishing and interesting that young children across different cultures follow a similar developmental path as far acquisition of language and speech are concerned. The infants start babbling at 6 months of age, by their first birthday they might name few things, by their second birthday they might put few words together for a rudimentary sentence, by the age three they are able to make a full sentence and, by the time they are four, they are able to speak fluently in their mother tongue [2], [6].

The spontaneity, ease and speed with which young children are able to acquire spoken language and speech has often made linguists, psychologists, neuro-scientists and now even computer scientists (in context of designing intelligent machines which can acquire language from their environment) wonder how the young children acquire language? Early language acquisition in infants (and young children) is the research area that seeks to answer this fundamental question. The researchers in this area are grappling with the questions like how infants are able to acquire vocabulary and structure despite richness and complexity of natural language [7], [8] (for e.g. segregating word boundaries from continuous speech), how infants (and young children) are able to adapt to various speaking styles in varied acoustic environments and generalise the different acoustic patterns to a same external concept (for e.g. a word) [9], how caregiver-child interactions influence language development of infant (or young child) [10], how language directed at child can be quantified [11] and many more.

One of the major area of interest among the early language acquisition researchers is how the quantity and quality of language exposure to an infant (or young child) impacts the child's future language development (see for e.g. [12], [13]). The difference in the linguistic exposure to the child is often characterised by the environment in which

the child is being brought up in (cultural, socioeconomic and household background), the content of speech child hears, and how often child is addressed directly as compared to overhearing of adult conversations [14]–[16]. However, Tamis-LeMonda et al. [17] argues that the variability in the language exposure to a young child is not adequately captured by structured tasks in the lab (or at the home) and controlled studies (e.g. [18]) do not sufficiently capture the fluctuation in language input that an infant faces in a naturalistic environment. Moreover, the social context of most of the controlled studies is often educated American/European households, so-called WEIRD communities [19] (Western, Educated, Industrialized, Rich, Democratic), which accounts for limited diversity in terms of language, culture and socioeconomic environment and thus, restricts the generalizability of the results.

To better understand an infant's real language learning environment (for e.g. at home where infant encounters a large fluctuation in language input over a longer period of time) researchers are collecting audio (or audio-visual) data of children going about their routine activities (see for e.g. [20]). These naturalist long duration audio recordings require automated processing tools for doing useful analysis such as quantification of speech directed to the child, which might be impractical to be done by a human on massive corpora of recordings. Our ongoing collaborative project called ***Analyzing Child Language Experiences Around the World*** (ACLEW; [21]) is an attempt in this direction. It brings together naturalistic, geographically and culturally diverse set of child-centered daylong audio recordings to aid scientific study of child language acquisition. It also endeavours to build state-of-the-art speech processing tools which can help to bring out and measure the diversity, variability in the language environments which an infant encounters in its daily life. The goals of the present work are closely related to the tools development (for speech analysis) efforts of the project ACLEW.

1.1 Problem description

In the scientific study of child language acquisition, it is often required to quantify the amount of speech input that a child hears (e.g. [10]) and detect whether the speech was directed towards the child or not (e.g. [13], [18]). Such tasks, even though can be performed by a trained person, become impractical and cumbersome when a diverse set of long duration recordings are involved. Thus, it is imperative to use automatic speech processing tools for such analyses. Particularly challenging examples of such automatic speech processing tasks that relate to child language analysis are automatic syllable count estimation (syllable counts are useful in quantification of language input to the child) and speech addressee classification (whether speech was directed to child or adult; also known as Infant directed speech and Adult directed speech classification, IDS/ADS classification in short).

The child-centered day long recordings provides a unique case for the application of speech processing algorithms. These recordings are generally mono channel, unconstrained in terms of possible environments, contain near and far field speech at varying and very

low SNRs (signal-to-noise ratio), have signal artefacts (e.g. friction of child's clothes with microphone), and are corrupted with varied nature of acoustic events from ubiquitous television running in the background to honking vehicles. In other words, these naturalistic recordings as captured by child-worn microphones are extremely noisy. For instance, Räsänen et al. [11] reports an average speech SNR of approximate 0 dB for a range of child-centered audio recordings. The inherently low and varying SNR of these recordings suggest that application of speech enhancement (SE) algorithms as a front-end (pre-processor) might improve robustness of downstream speech processing tasks. A similar use case of SE front-end for downstream task of word count estimation (WCE) from long-duration naturalist personal log recordings from adult subjects was reported by Ziaei et al. [22] . In their study, they compared a couple of classical signal processing based SE algorithms and found that a SE front-end using spectral subtraction improved the robustness of their WCE pipeline. Following this, spectral subtraction was also used by Räsänen et al. [11] for WCE in child-centered recordings. Additionally, SE pre-processor has also been successfully applied in the tasks of speaker diarization [23], automatic speech recognition [24] etc. In contrast, submissions to the child-centered audio analysis challenges, held as a part of Computational Paralinguistic Challenges (ComParE) [25], [26], have relied on robust classifiers devoid of a SE front-end. This leaves it unclear whether SE comparisons as reported in [22] generalize to child-centered day long recordings or to downstream tasks other than WCE. In addition, the earlier studies [11], [22] have not explored modern machine learning-based methods for SE, which might provide additional performance gains in the downstream tasks beyond conventional DSP based approaches.

In this context, it might also be useful to have task-independent metrics which can point to the right enhancement algorithm by comparing a number of SE algorithms on standardised tests (e.g. objective quality evaluations). Such standardised tests would be computationally cheaper to execute on an array of SE methods then constructing number of SE front-ends for each downstream task to evaluate which front-end maximises performance of the downstream task. However, to our knowledge, such a study which demonstrates the application of standardised test results (e.g. results of objective quality evaluations) as predictive pointers to select a suitable SE front-end for a downstream task has not been conducted. This makes it unclear that whether performance on metrics (for e.g. spectral distortion, instrumental intelligibility) can be translated into a criterion for selecting a suitable SE algorithm for pre-processing, when the involved downstream tasks are essentially varied speech processing tasks.

1.2 Research goals of the thesis

The goal of this thesis is to compare a number of DSP-based classical SE algorithms and modern machine learning-based SE algorithms in the analysis of child-centered daylong audio recordings, with focus on downstream tasks of automatic syllable count estimation and IDS/ADS classification. Additionally, we compare the downstream task performance of SE methods with the objective metrics of spectral distortion and speech intelligibility

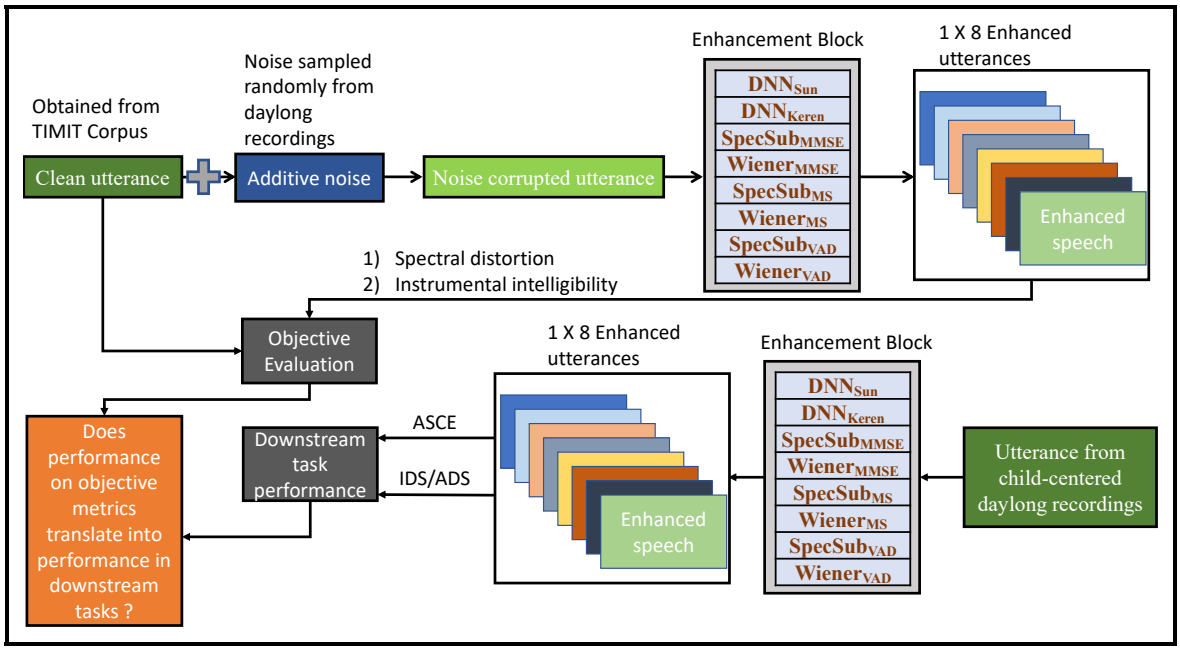

Figure 1.1 *Schematic diagram of the experiments conducted in this work. ASCE is automatic syllable count estimation task and IDS/ADS is Infant-directed speech and adult-directed speech classification task. In the enhancement block, DNN_{xxx} denotes deep neural network based SE methods, SpecSub is spectral subtraction, and Wiener is Wiener filter. The subscripts with spectral subtraction and Wiener filter denotes the noise estimation techniques. Please see chapter 3 for details.*

calculated in noise conditions typical to what is encountered in child-centered audio data. Through this effort, we want to analyse whether objective metrics could be used as a reasonable proxy in selection of a SE front-end for these two downstream tasks that focus on different signal characteristics. To summarise, following objectives can be delineated:

O1: Compare different SE methods by calculating objective metrics of spectral distortion and instrumental intelligibility in noise conditions reflective of those typical in child-centered long duration naturalistic recordings.

O2: Compare different SE front-ends by evaluating their performance with respect to the two downstream tasks of IDS/ADS classification and automatic syllable count estimation.

O3: Compare objective evaluation results with downstream task performance to assess whether performance on objective measurement translate into performance on downstream tasks.

1.3 Experimental scheme of the thesis

Figure 1.1 illustrates the basic experimental scheme utilized in this work. Broadly, the purpose of the present experimental scheme is to evaluate the performance of the SE

methods in additive noise conditions similar to what is observed in child-centered long-duration recordings through objective metrics and efficacy of SE methods as front-ends for performance improvement in downstream tasks. Further, the performance of SE methods through objective assessment is compared with performance in downstream tasks.

1.4 Organization of the thesis

The thesis is organised as follows. The chapter 2 following this introduction discusses theoretical preliminaries related to speech production and perception, noise, basic principles behind speech enhancement techniques and evaluation measures for speech enhancement algorithms. The chapter 3 discusses the speech enhancement methods (including noise estimation procedures), downstream speech analysis tasks and evaluation measures used in the experiments conducted as part of this work. The chapter 4 describes the data and experimental setup and the chapter 5 presents the results obtained from the experiments and a discussion of the results. The final chapter 6 reports the conclusions derived from this study.

2 Theoretical background

2.1 Speech production and perception

Speech at a physical level is variation in air pressure (an air pressure wave) produced as a result of coordination between various complex mechanisms in the human body. The physiology of the speech production involves air flow generation (by expulsion of air from the lungs), vibration induced in the air flow by oscillation of vocal folds for the voiced sounds or turbulence created in air flow due to constrictions for unvoiced sounds (e.g. modulating airflow by opening and closing of lips to produce /p/), movements and configuration of articulatory organs (e.g. jaws, lip, tongue, velum, teeth) and muscle control exercised through brain (e.g. regulating tension in vocal folds through larynx muscles).

Figure 2.1 illustrates the human vocal apparatus. In figure 2.2, the grey portion with black outline is initial position of passage from epiglottis to lips (correlate with the same section in the figure 2.1). When a sound is to be produced, the shape of passage and position of articulators (e.g. lips, tongue, teeth) change as demonstrated in figure 2.2 by the thick white outline (compare with initial position). The position of articulators as depicted by the thick white line shows tongue is rolled against teeth in upper jaw and lips are slightly open, like when producing sound /s/.

The opening between vocal folds is called *glottis*. Glottal opening takes place due to the accumulating pressure at the vocal folds when they are closed and steady pressure from the lungs is present. When the glottis opens, the vocal folds are soon sucked back together due to the Bernoulli effect. This change in pressure at the level of glottis gives rise to periodic opening and closing of vocal folds. When the vocal folds are closed no air flow reaches the vocal tract, and this phase is referred to as glottal *closed phase*. When the folds are open the period is known as glottal *open phase*. The temporal duration of glottal cycle is known as pitch-period, its reciprocal is *fundamental frequency* or $F0$. Typically, in males the fundamental frequency range is between $60 - 150$ Hz and $200 - 400$ Hz in females and children. The fundamental frequency is closely related to *pitch*, which is our perception of the fundamental frequency. Pitch is what our ears and brain interprets as periodicity of the signal [3].

The shape of the vocal tract is determined by the size and shape of various cavities (larynx, pharynx, oral). The structure of these cavities is plastic and is modified in response to movement and position of articulators. Thus, the shape of vocal tract changes during speech articulation. The shape of the vocal tract determines the spectral envelope (demonstrated in figure 2.3 with thick curved line) and in this context, it can be considered as a filter that spectrally shapes the acoustic signal. The shape of vocal tract is also responsible for timbre present in speech sound as it acts as a filter amplifying certain frequencies and attenuating others.

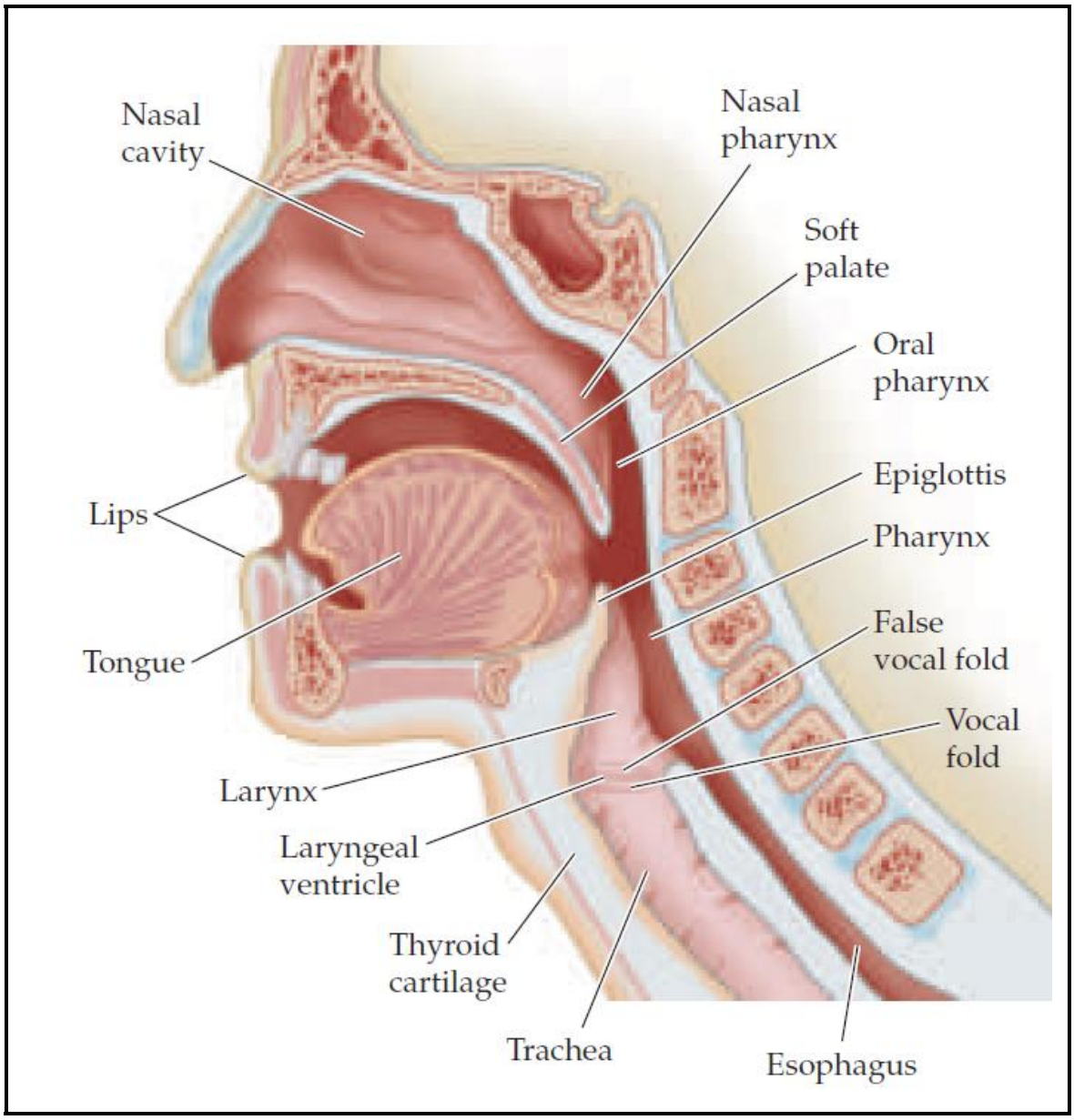

Figure 2.1 *Human vocal tract (reproduced from [27])*

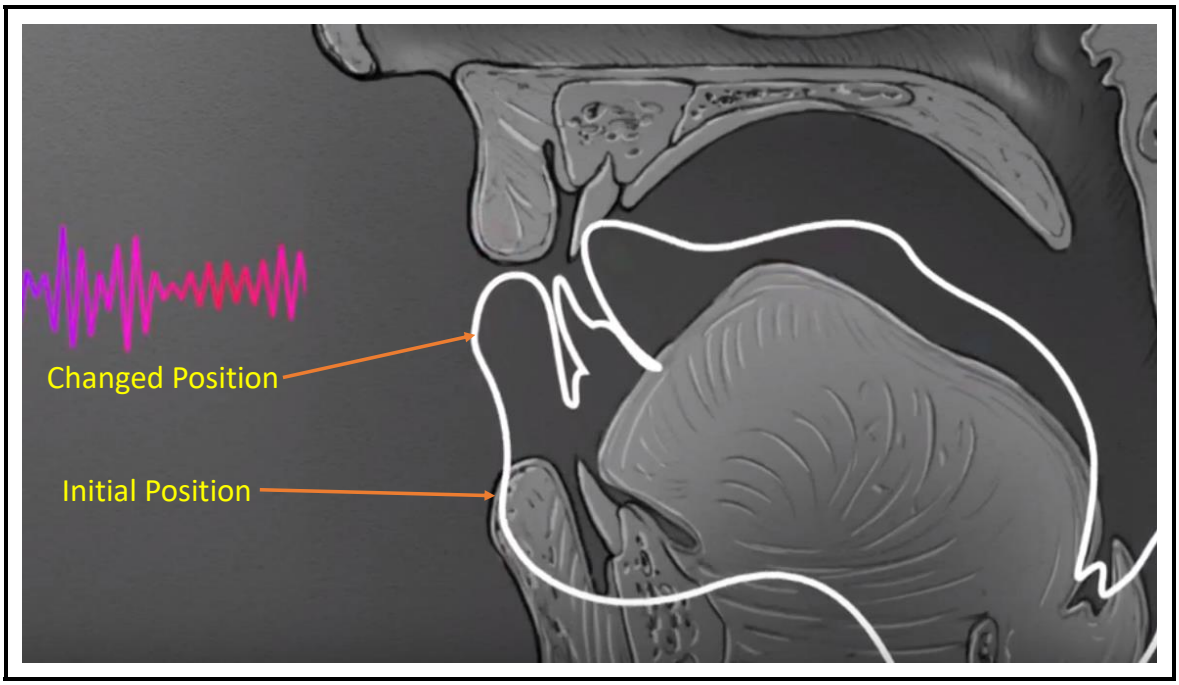

Figure 2.2 *Articulatory movement when speaking (adapted from the presentation video of [28]).*

The resonances of the vocal tract are known as *formants*. They can be observed as peaks in spectral envelope (see $F1$, $F2$ and $F3$ in the figure 2.3). In this context, it is to be noted that location of higher formants might not always be prominent and thus, would be difficult to resolve on the spectral envelope (e.g. $F4$ and $F5$ are difficult to locate in the figure 2.3).

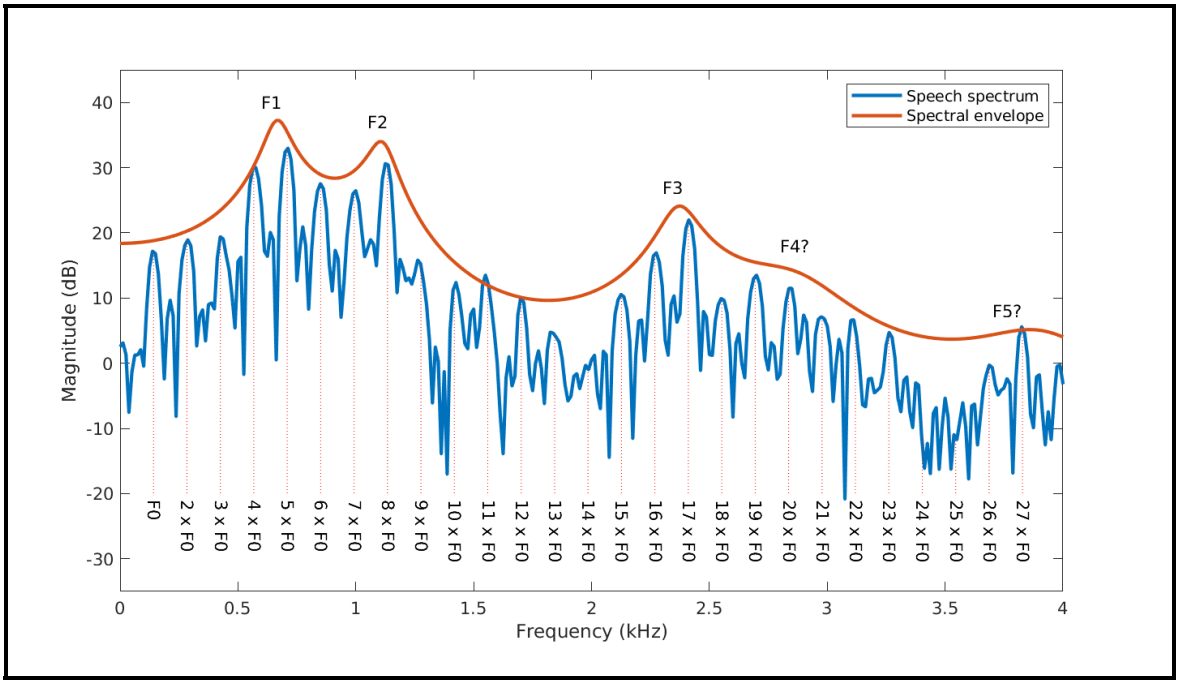

Figure 2.3 *Spectral envelope of speech with formants (reproduced from [3]).*

Speech perception is how we hear, understand and interpret a linguistic sound [29]. *Phonemes* are smallest unit of linguistic sound which may bring about change in meaning. For example, the /s/ in 'soar' distinguishes it from /r/ in 'roar' [30]. There are 40 phonemes in English language and are broadly classified into eight groups: vowels, diphthongs, semivowels, nasals, fricatives, affricates, stops, and whispers (see chapter 3 of [31] for details). The speech signal is processed by our auditory system to extract perceptual acoustic cues (e.g. harmonic structure of the signal, formant frequencies, duration). The acoustic cues help in perception of various classes of speech sounds (e.g. formant frequencies and duration are most important acoustic cues in perception of vowels [31]).

2.2 Speech Enhancement

Speech enhancement (SE) is concerned with improving speech signal quality, degraded by additive noise, for human listening (in terms of comfort and intelligibility) and (or) machine processing (e.g. pre-processor for other speech processing tasks). Some of the common applications areas of speech enhancement are:

- Voice communication over cellular telephony, which typically suffers from background noise in surroundings (honking vehicles, varied sounds in market place, multiple-talkers in restaurants etc.) at the transmitting end. Speech enhancement algorithms are used as a pre-processor in speech coding systems employed in cellular phones to improve speech quality at the receiving end [32].
- Improving recognition accuracy and robustness of automatic speech recognition (ASR) systems by enhancing noisy speech signal before feeding them into the ASR systems [24].
- Military communications, for instance, improving intelligibility of pilot's speech which is corrupted by high levels of cockpit noise [33].
- Cleaning noisy signal before amplification in hearing aids designed for hearing-impaired listeners [34].

So, there are wide variety of contexts in which it is desired to enhance speech. Depending upon the specific application the enhancement system may be directed to improve speech quality, intelligibility or both [35]. In the speech enhancement literature, process of enhancement is often referred to as "denoising" or "noise-reduction". The terminology captures the essence of speech enhancement systems i.e. suppression of noise.

2.2.1 Noise

Our environment is full of sound sources. We are usually interested in the sounds which carry useful information to us such as speech, music, doorbells, calling tone etc. These sounds are often labelled as "desired". Noises are unwanted signals which interfere with

the communication, processing, and measurement of an information-bearing desired signal [36]. For example, consider a conversation going in a crowded restaurant. Too many concurrent sound sources (different people speaking at the same time, background music, footsteps etc.), severely impacts the intelligibility of the speech in the conversation. Such environments are often referred to as noisy. In the presence of loud noise, it is not only difficult to hear (background noise heavily degrades speech) but it also becomes problematic to speak clearly. The speaker has to put extra vocal effort by speaking louder, improving intonation, changing fundamental frequency etc. (see Lombard effect [37]). Another example, in this context, is of machine processing of degraded speech. Speech processing systems generally show deterioration in performance when input speech is noise corrupted. For example, word error rate (WER), used in evaluating performance of ASR systems, increases rapidly when input speech is degraded (see experiments section of [38] and [39]).

Degradation and distortion of speech (or audio in general) can be attributed to the effects of *additive noise* (e.g. multi-talker babble), *echo* due to reflection of sound waves (e.g. by walls in a room, by mountains in nature) or due to coupling between loudspeakers and microphone (e.g. in a teleconference system), *reverberations* (distortion produced due to acoustics of enclosed space), and different transmission channel effects (e.g interference from nearby channels, faulty equipment etc). These different categories of degradation impact the speech signal in different manners. For example, additive noise, as the name suggests, is additive to the speech signal. Mathematically, if we represent noisy speech signal as $y(n)$, clean speech as $x(n)$ and noise signal as $d(n)$ then noisy speech signal can be modelled as

$$y(n) = x(n) + d(n) \tag{2.1}$$

Here, it is assumed that the additive noise source is statistically independent with the speech and locally stationary. In figure 2.4, we show clean speech signal, its spectrogram followed by additive WGN corrupted clean speech and its spectrogram. The noise is added at SNR = 10 dB. It is interesting to see in spectrogram (bottom), how noise has swarmed ≥ 4 kHz frequencies.

In contrast to additive noise, echo, reverberations, and transmission channel effects produce distortions in the speech through convolution with the original speech signal. In the presence of convolutive noise, mathematical representation of the noisy speech in 2.1 can be modified as

$$y(n) = x(n) * h(n) \tag{2.2}$$

Here, $h(n)$ is the impulse response of the convolutive noise source.

These different categories of degradation in speech are addressed by different signal processing techniques and are active area of research. Speech enhancement primarily deals with suppression of additive noise (e.g. [40], [41]). Echo suppression and cancellation techniques (e.g. [42], [43]) are used to handle distortions produced by echo. Speech dereverberation techniques (e.g. [44], [45]) aims to minimise reverberations in speech signal.

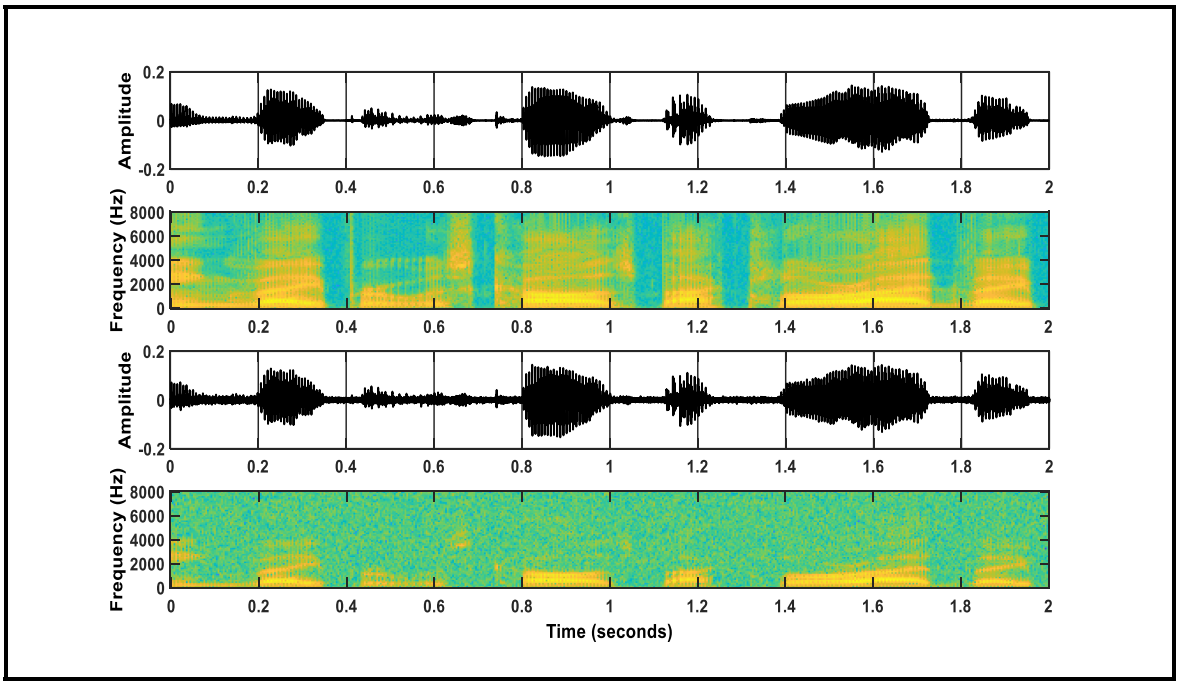

Figure 2.4 *Clean speech in time domain (top) followed by its spectrogram. Then, clean speech corrupted with additive white Gaussian noise in time domain followed by its spectrogram (bottom).*

2.2.2 Noise in context of speech enhancement

In general, noise can be typically viewed as a stochastic process with respect to the target signal of interest. It could be stationary (noise-characteristics constant with respect to time) or non-stationary (noise-statistics are changing with respect to time). A good example of stationary noise process is additive white Gaussian noise process commonly known as white Gaussian noise (WGN). The WGN is a standard noise model used in modelling of many random process in nature [46]. Theoretically, it has a constant power spectral density i.e. all frequencies has equal power. It has zero mean and finite variance. The samples of WGN are independent and identically distributed (samples belong to normal distribution or Gaussian distribution) and are thus, uncorrelated. Figure 2.5 illustrates the temporal waveform and spectral characteristics of WGN. The spectrogram and the shape of spectrum are particularly interesting as they relate to distribution of energy in the frequency domain. For example, in the spectrogram we can see that distribution of energy is nearly uniform across frequencies. On the other hand, ambience of an airport lobby where there are concurrent sound sources such as people talking, sound of footfalls, flight announcements etc. is a good example of non-stationary noise source. Noise sampled randomly from daylong recording is another example of non-stationary noise. Figures 2.6 and 2.8 illustrates the characteristics of noise from airport lobby and daylong recording respectively in time and frequency domains.

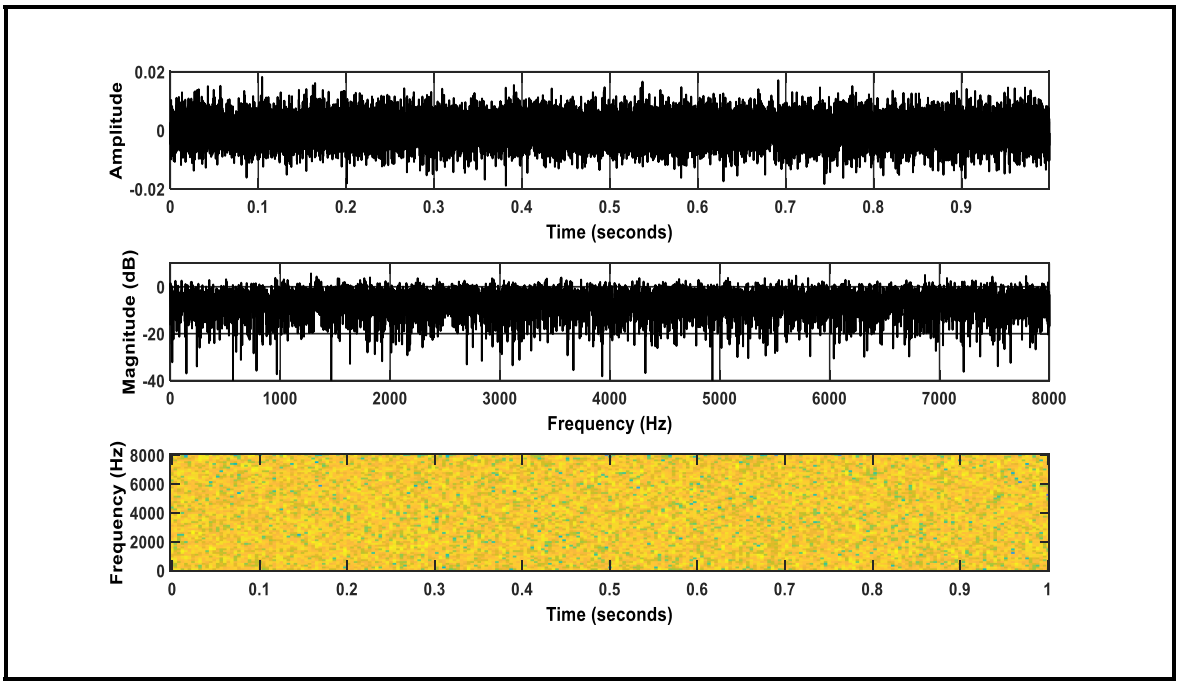

Figure 2.5 *White Gaussian noise in time domain (top), its long-term average spectrum (middle) and its spectrogram (bottom).*

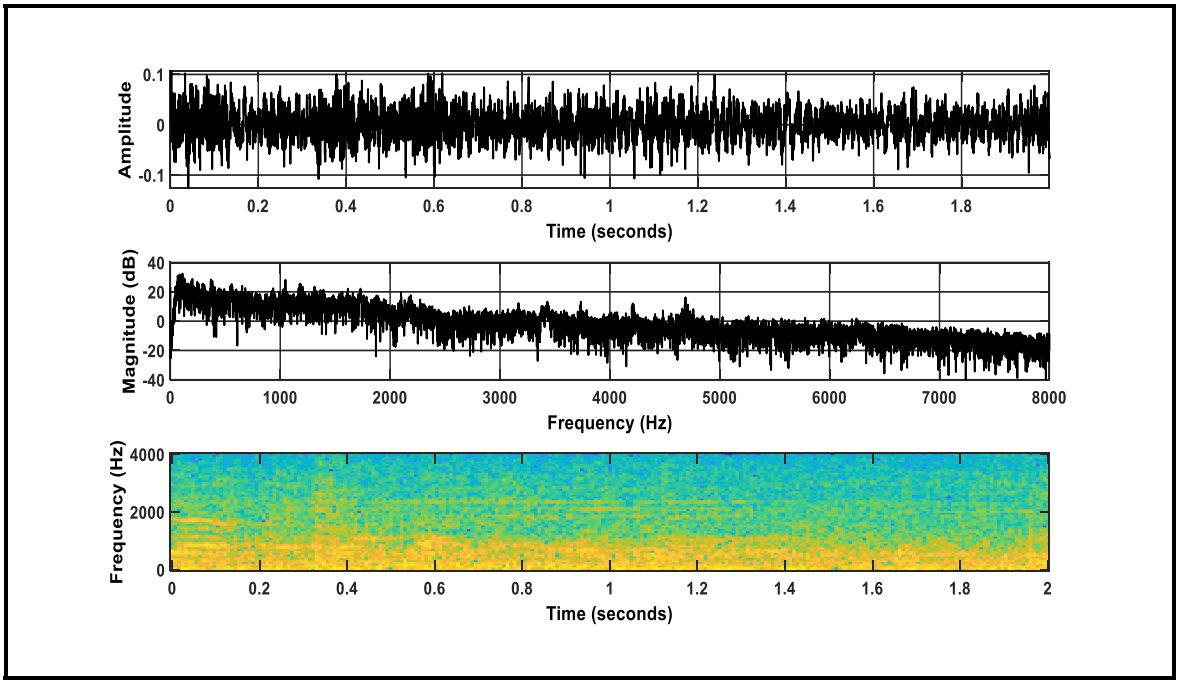

Figure 2.6 *Non-stationary noise from airport lobby in time domain (top), its long-term average spectrum (middle) and its spectrogram (bottom).*

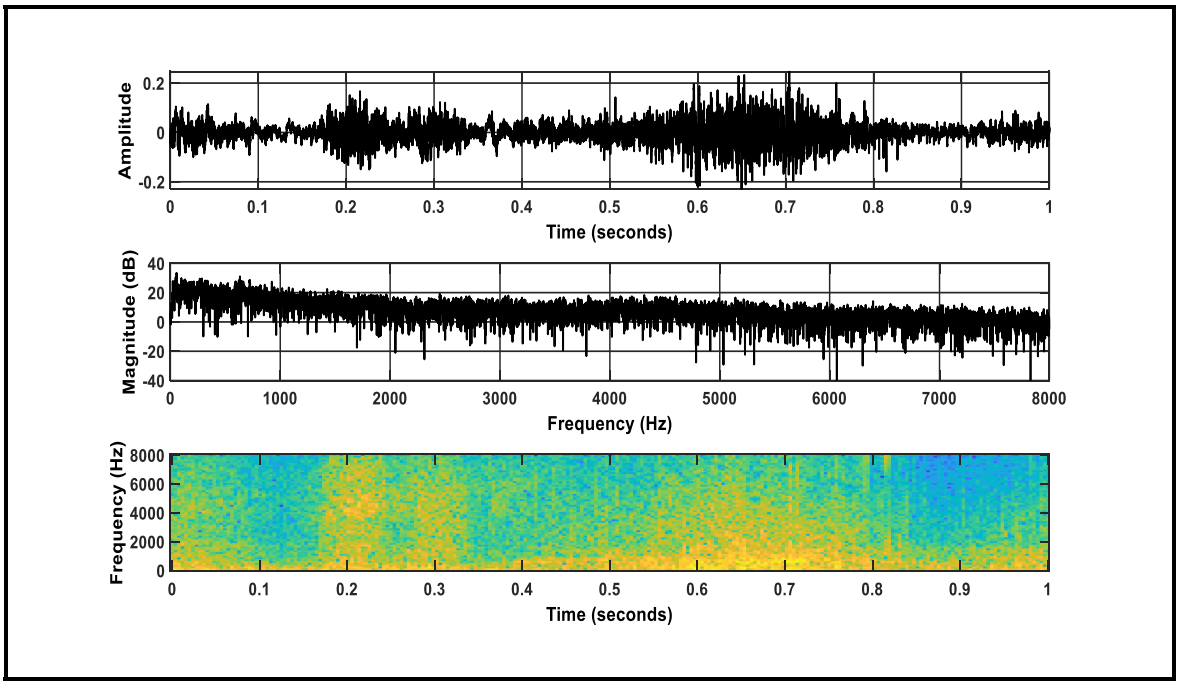

Figure 2.7 *Noise sampled randomly from daylong recordings in time domain (top), its long-term average spectrum (middle) and its spectrogram (bottom).*

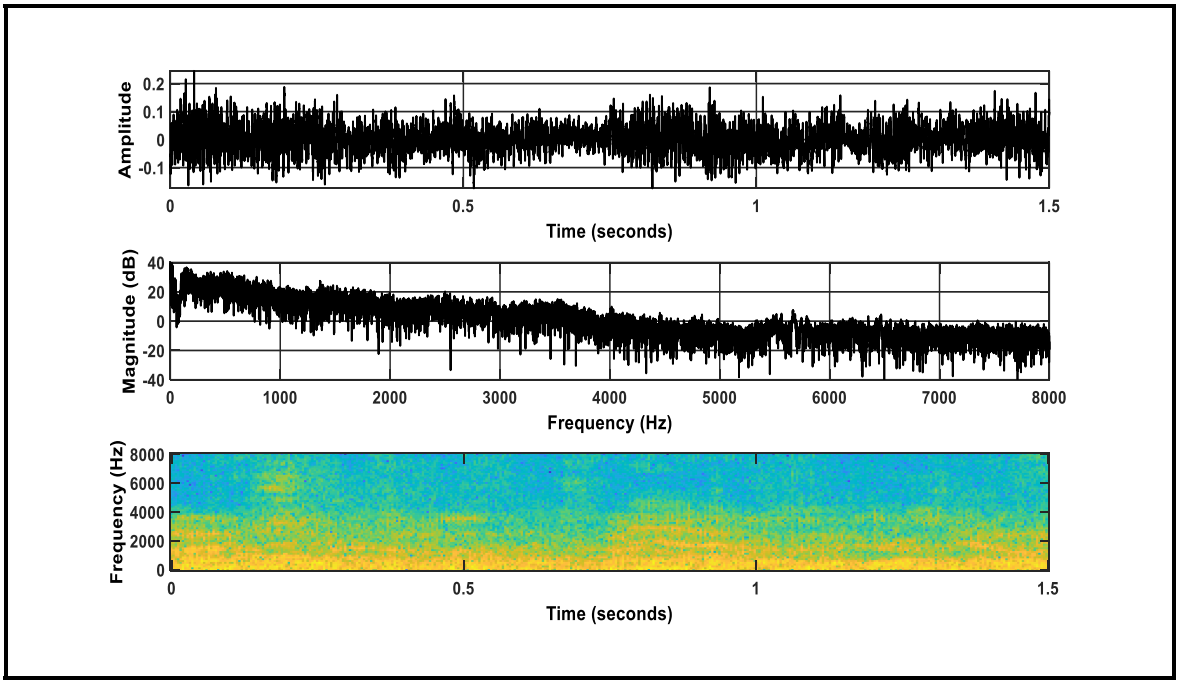

Figure 2.8 *Multi-talker babble noise in time domain (top), its long-term average spectrum (middle) and its spectrogram (bottom).*

In case of non-stationary noises, the statistical characteristics of noise signal varies with time and this makes them hard to model. From perspective of speech enhancement, additive non-stationary noises presents a complex case for noise suppression as their estimation

is hard due to difficulty in accurate modelling. In context of SE, it is also important to mention that speech-shaped noise and babble noise are two standard noise examples which are frequently encountered in speech processing. These two noises exhibit spectrum characteristics which are similar to speech and thus, are very effective in masking desired speech signal.

2.2.3 Human listening in noisy conditions

Humans with normal hearing functions are able to perceive speech even in adverse communication conditions [4]. For example, we are able to selectively attend to conversation happening at our table in a restaurant even though there are several people talking (on other tables, waiters attending to guests etc.) and other sources of sound present in the background (e.g. footfalls, people eating, music etc). This is possible due to special abilities of human cognition which are able to recognise acoustic cues even in extremely degraded acoustic environments. In such environments, human cognition system exploits the binaural capabilities (in terms of inter-aural level difference (ILD) and time difference (ITD) to localize and selectively attend to a sound source. It tries to reinforce the observed information from acoustic cues by correlating it with observed visual cues (e.g. lip movement and facial expressions) and context of the conversation. Further, in complex multi-speaker scenarios human cognition system utilizes auditory grouping mechanisms to segregate desired speech from other concurrent sound sources by grouping sound components according to their sources. It also uses fundamental frequency differences (for e.g. between male and female speaker) in competing speech to recognise speech from target speaker (for more details see Chapter 4 in [31] and references there in). In this context, it is also important to note that there are some inherent acoustic features of speech which protects the information from external degradation and distortion. For instance, Parikh & Loizou [47] report that low-frequency spectral peaks (formant resonances) of speech are preserved in noise to a greater degree. This is due to energy dominance of lower formants in speech, so they are last to be masked by noise. This implies that listener has reliable access, even in noisy conditions, to lower formants ($F1, F2, F3$) which provide critical cues for vowel and stop-consonant perception.

2.2.4 Basic principles in SE methods

The problem of speech enhancement can be addressed by classical digital signal processing (DSP) methods or modern machine learning techniques. Both categories of method approach the problem in fundamentally different manner. DSP based methods assume that noise is additive, source of noise and speech are uncorrelated, and noise is stationary in the analysis time window. Based on these assumptions the noisy speech signal $y(n)$ can be considered as superimposition of clean speech signal $x(n)$ and noise signal $d(n)$. The noisy speech signal can be modelled as equation 2.1. Taking discrete-time Fourier transform (DTFT) of both sides in equation 2.1 (Fourier transform is linear) we get

$$Y(\omega) = X(\omega) + D(\omega) \tag{2.3}$$

In polar form,

$$|Y(\omega)|e^{j\phi_y(\omega)} = |X(\omega)|e^{j\phi_x(\omega)} + |D(\omega)|e^{j\phi_d(\omega)} \tag{2.4}$$

Here, ϕ is phase of the signal and $|.|$ represents the magnitude of the signal.

It is preferred to execute enhancement related operations in spectral domain as spectral analysis can be easily correlated with knowledge about speech production and perception. For example, phase of noisy speech signal can be used in reconstruction of clean speech signal, as long as the phase difference is not perceptible to human auditory cognition. Design process of SE methods can benefit from such information. In real world situations, the noise signal $d(n)$ is not known. It has to be estimated from observed noisy speech signal. DSP based SE methods utilise different noise-estimation methods (e.g. minimum statistics algorithm [48]; discussed in chapter 3) to estimate noise magnitude spectrum (or noise power-spectrum). Let us consider that with the help of a noise estimation method, the noise magnitude is estimated as $|\hat{D}(\omega)|$ (symbol with $\wedge$ on top represents estimated parameter) and phase-difference with noisy speech is imperceptible so the noisy speech phase can be used instead of estimated phase. Now, clean speech $|\hat{X}(\omega)|$ can be estimated as

$$\hat{X}(\omega) = [|Y(\omega)| - |\hat{D}(\omega)|]e^{j\phi_y(\omega)} \tag{2.5}$$

Rearranging the equation 2.5,

$$\hat{X}(\omega) = [1 - \frac{|\hat{D}(\omega)|}{|Y(\omega)|}]|Y(\omega)|e^{j\phi_y(\omega)} \tag{2.6}$$

Let $H(\omega) = 1 - \frac{|\hat{D}(\omega)|}{|Y(\omega)|}$, then equation 2.6 can be written as

$$\hat{X}(\omega) = H(\omega)|Y(\omega)|e^{j\phi_y(\omega)} \tag{2.7}$$

Here, $H(\omega)$ is referred to as a suppression function [31] as it attenuates the magnitude of noisy speech. Different DSP-based SE methods, use different techniques to estimate the suppression function. It is to be noted here that the suppression function is dependent on observed noisy speech spectrum and estimated noise spectrum. Thus, using the same SE method but changing the noise estimation procedure will modify the suppression function and change the estimate of clean speech.

Machine learning-based SE methods, on the other hand, approach SE as a supervised learning problem. Supervised learning techniques attempts to infer a model from labelled training data. The inferred model then can be utilised for mapping unseen data sample to known labels. In context of speech enhancement, a supervised learning task attempts to learn clean speech representation from a training set of clean and noisy speech signal pairs

[49]. Let us consider

$$Z_{train} = \{(x_i(n), y_i(n)) \mid (x_i(n), y_i(n)) \text{ is a pair of clean and noisy speech signal and i is an integer} \in [1, card(Z_{train})]\} \tag{2.8}$$

Here, $card(Z_{train})$[1] represents cardinality of the set i.e. number of elements in the set. Similarly, testing set Z_{test} can also be defined. Let us consider, a pair of clean speech and noisy speech signals $(x(n), y(n)) \in Z_{train}$. Now, the supervised learning problem in context of speech enhancement can be formulated as (equation taken from [49])

$$\hat{o_x} = F(v(y(n)), \theta) \tag{2.9}$$

Here, $F(\cdot, \theta)$ is a parametrized model with parameters θ, $v(\cdot)$ is vector-valued function that applies feature transformation to noisy speech signal $y(n)$ and $\hat{o_x}$ is output from model $F(\cdot, \theta)$ which is representation of estimated clean speech (e.g. it could be magnitude spectrum). The $\hat{o_x}$ is such that on applying a transformation function $u(\cdot)$ (e.g. if $\hat{o_x}$ is magnitude spectrum then transformation function could be inverse Fourier transform) it will give estimated clean speech,

$$\hat{x}(n) = u(\hat{o}_x) \tag{2.10}$$

The transformation function and output from model is dependent upon specific application [49]. For example, DNN based approach for SE proposed by Sun et al. [50] predicts an IRM (Ideal Ratio Mask) which can be applied to the input noisy speech to get enhanced speech. Now, the task of supervised algorithm is to estimate optimal parameters θ^* by optimizing $F(\cdot, \theta)$ with respect an objective function (e.g. a cost function like mean squared error) on training data set. With the incorporation of optimal parameters in model $F(\cdot, \theta)$, the model can be used to estimate clean speech from an unknown sample of noisy speech.

2.3 Evaluation measures for speech enhancement

The purpose of evaluation methods for speech enhancement is to quantify the quality of enhanced speech. The quantification helps in comparing the speech enhancement algorithms performance. The performance of speech enhancement algorithms can be measured through either subjective listening tests or objective evaluation measures [31]. Subjective listening tests use human subjects to determine performance. Listening tests are often in nature of exercising preference for one sample over other (e.g. Which one is better, A or B ?) or rating the samples over a predefined scale (e.g. rank A, B and C from best to worst). The setup of listening tests are carefully designed to extract information which aids in performance measurement and is reliable and accurate [3]. The listening tests

[1]Equation 2.8 represents a training set through parametric notation for sets. The set members are parmetrised by i.

can be conducted in controlled settings in a laboratory for example following recommendation from ITU-T P.800-series (International Telecommunication Union-Telecommunication Standardization Sector) or can be crowd-sourced following recommendations from ITU-T P.808. The listening tests are followed by statistical analysis of results to determine whether the obtained results are statistically significant or not.

Speech perception by humans is a highly complex process, involving not only spoken language understanding but also the utterance context as well as the emotional, psychological, gender and social attributes of the speaker [51]. After decades of research, still there is no single evaluation measure which can replace human listener conclusively to assess the speech quality [31]. The subjective listening tests, though still the most reliable instruments for speech quality evaluation, are prohibitive in terms of cost and time consumption. For these reasons, objective evaluation measures have been developed by the researchers to complement and in some cases replace the subjective evaluations.

The objective assessment techniques involve use of algorithms to predict the output of subjective listening tests. The objective evaluation of speech quality is usually classified into intrusive and non-intrusive methods, based on the type of input they require [52]. The intrusive methods need the input (clean) and output (processed or in our case enhanced) signal to assess speech quality whereas non-intrusive methods assess speech quality using only the output (processed) signals. The intrusive methods tend to calculate the distortion measure between the original and processed signal and then correlate the distortion measure with the speech quality. The intrusive quality measures can be mathematically oriented (see for e.g. segmental SNR [53], Linear Predictive Coding (LPC) based spectral distance measures such as Itakura-Saito [54], Speech Intelligibility in Bits (SIIB) [55], [56]) or psychoacoustically motivated (see for e.g weighted spectral slope distance [57] or Bark distortion measure [58]). Essential difference between mathematical and psychoacoustic approaches is that, mathematical measurement techniques operate on simple mathematical models where as psychoacoustic motivated approaches in their design involve the characteristics of human auditory perception such as non-uniform frequency resolution in the ear and subjective perception to loudness [52]. In this manner psychoacoustic-based approaches tries to emulate the human auditory system.

Non-intrusive objective measures, on the other hand, compute quality metric through analysis of processed signal only. In this context, consider the case of VoIP [59] applications where it is of utmost importance to continously monitor the performance of telecommunication networks in terms of speech quality [31]. In such a case, where the system has access to only transmitted (or processed) signal to evaluate quality a non-intrusive system is desired. A good example of non-intrusive objective measure is recommendation ITU-T P.563.

3 Methods

In this chapter, we describe the different methods and techniques which were used during the experiments reported in this thesis. In sections 3.1 and 3.2, we provide technical, mathematical (wherever necessary), and implementation level details of SE methods utilised in this work. The section 3.1 also includes a brief description of noise statistics estimation techniques utilised in the DSP-based methods for SE. We also describe the used objective evaluation metrics for quality and intelligibility assessment of SE methods in section 3.3 and, finally, the downstream tasks on which SE front-end was evaluated are described in section 3.4.

3.1 Digital signal processing based methods for SE

In this section, we describe the core principles behind two algorithms, spectral subtraction and Wiener filtering, which are traditionally used as DSP based methods in SE.

1. **Spectral subtraction**

   Spectral Subtraction algorithm was first proposed by Boll in 1979 [40] for Fourier domain. It is based on a simple principle that if noise is additive and relatively stationary, clean speech spectrum can be obtained by subtracting noise spectrum from noisy speech spectrum. While discussing basic principles behind DSP-based SE methods in section 2.2.4, we have derived the basic equation utilized in spectral subtraction. Rewriting the basic equation 2.5 here

   $$\hat{X}(\omega) = [|Y(\omega)| - |\hat{D}(\omega)|]e^{j\phi_y(\omega)} \tag{3.1}$$

   Here, it is to be noted that assumptions related to noise being additive and uncorrelated with speech stated in section 2.2.4 holds. The noise signal $d(n)$ is not known, so the noise spectrum is estimated through different noise estimation techniques discussed in section 3.1.1. After plugging in the estimated $|\hat{D}(\omega)|$ in equation 3.1, the Fourier representation of clean speech signal is obtained. As the magnitude spectra ($|\hat{X}(\omega)|$) can not be negative, so care is taken while carrying on subtraction in equation 3.1. If the magnitude spectra, comes out as negative it is made non-negative (or zero) by different methods such as half-wave rectification [31]. Now, the clean speech waveform can be reconstructed by taking inverse Fourier transform of $|\hat{X}(\omega)|$ using the phase of noisy speech signal.

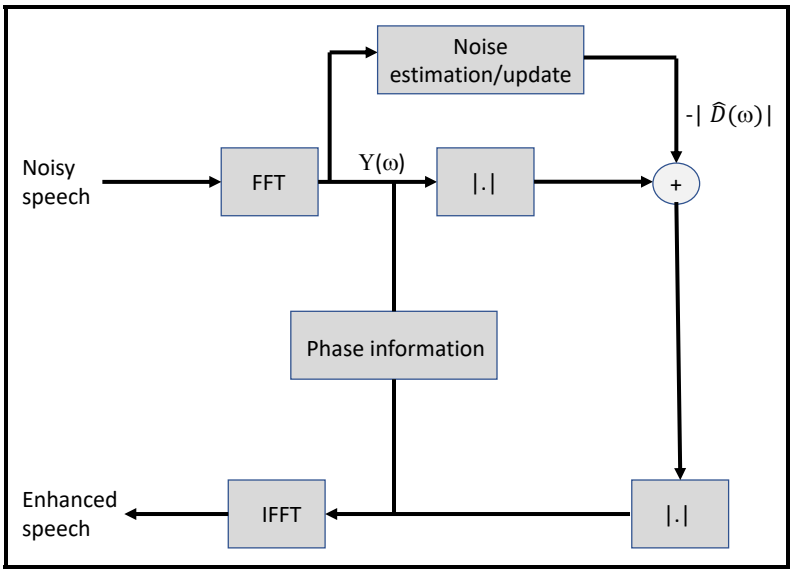

Figure 3.1 *Block diagram depicting process of spectral subtraction (adapted from [31]). $|\cdot|$ represents magnitude operation.*

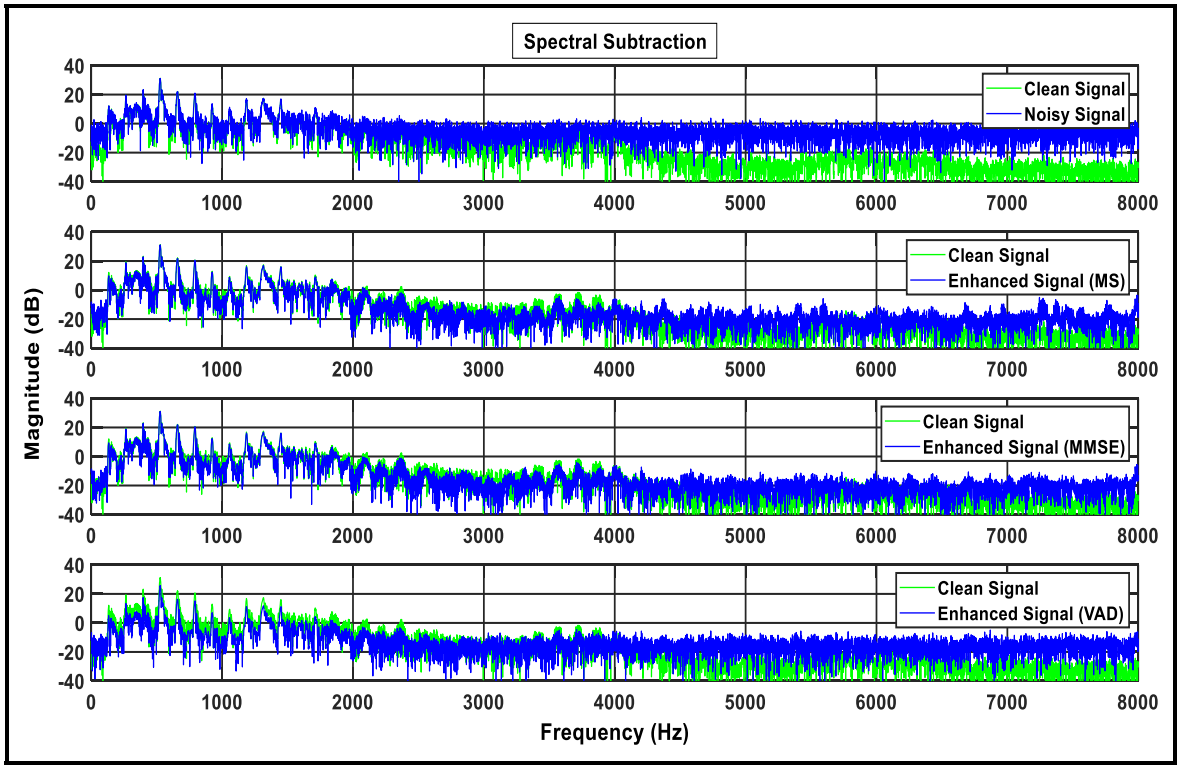

Figure 3.2 *The figure demonstrates long term average spectrum (LTAS) of clean and noisy signal in the top plot and LTAS of clean and enhanced signals in subsequent plots. Enhancement is done through spectral subtraction using different procedures for noise estimation. MS, MMSE and VAD are noise estimation techniques described in section 3.1.1. The clean speech signal is obtained from TIMIT [60] corpus and noisy speech signal is corrupted version of clean speech obtained by adding white Gaussian noise at SNR = 10 dB. In the plots, frequency (Hz) is plotted on X-axis and magnitude (dB) is plotted on Y-axis.*

2. **Wiener filtering**

The spectral subtraction algorithm discussed in the previous section was based on what Loizou [31] calls "intuitive and heuristically based principles". The enhanced signal was not derived in an optimal manner. In the Wiener filtering (named after mathematician Norbert Wiener's contribution to this problem in the continuous domain [41]) approach, the enhanced signal is derived by optimizing a mathematical error criterion, the minimum mean square error between the output signal $\hat{x}(n)$ and desired signal $x(n)$.

Let us assume input signal $y(n)$ to the system and desired signal $x(n)$ are wide-sense stationary random processes. The estimation error can be computed as [31],

$$e(n) = x(n) - \hat{x}(n) \tag{3.2}$$

We know that the output signal $\hat{x}(n)$, can be obtained by convolution of filter's impulse response $h(n)$ with input signal $y(n)$,

$$\hat{x}(n) = h(n) * y(n) \tag{3.3}$$

Taking discrete time Fourier transform of equation 3.3, we get,

$$\hat{X}(\omega) = H(\omega)Y(\omega) \tag{3.4}$$

Now, we can define the estimation error at frequency ω_k as [31],

$$\begin{aligned} E(\omega_k) &= X(\omega_k) - \hat{X}(\omega_k) \\ &= X(\omega_k) - H(\omega_k)Y(\omega_k) \end{aligned} \tag{3.5}$$

The mean square error is given by [31],

$$\begin{aligned} \mathbb{E}[|E(\omega_k)|^2] &= \mathbb{E}\{[X(\omega_k) - H(\omega_k)Y(\omega_k)] * [X(\omega_k) - H(\omega_k)Y(\omega_k)]\} \\ &= \mathbb{E}[|X(\omega_k)|^2] - H(\omega_k)\,\mathbb{E}[X^*(\omega_k)Y(\omega_k)] - H^*(\omega_k)\,\mathbb{E}[Y^*(\omega_k)X(\omega_k)] \\ &+ |H(\omega_k)|^2\,\mathbb{E}[|Y(\omega_k)|^2] \end{aligned} \tag{3.6}$$

It can be noted from 3.6, that $P_{yy}(\omega_k) = E|Y(\omega_k)|^2$ is the power spectrum of $y(n)$, and $P_{yx}(\omega_k) = E[X^*(\omega_k)Y(\omega_k)]$ is the cross-power spectrum of $y(n)$ and $x(n)$, with this information we can express 3.6 as [31],

$$\begin{aligned} J &= \mathbb{E}[|E(\omega_k)|^2] \\ &= \mathbb{E}[|X(\omega_k)|^2] - H(\omega_k)P_{yx}(\omega_k) - H^*(\omega_k)P_{xy}(\omega_k) + |H(\omega_k)|^2 P_{yy}(\omega_k) \end{aligned} \tag{3.7}$$

To obtain optimum filter $H(\omega_k)$, we take derivative of 3.7 with respect to $H(\omega_k)$ and set it to zero (mmse error minimisation criterion),

$$\begin{aligned}\frac{\partial J}{\partial H(\omega_k)} &= H^*(\omega_k)P_{yy}(\omega_k) - P_{yx}(\omega_k) \\ &= [H(\omega_k)P_{yy}(\omega_k) - P_{xy}(\omega_k)]^* = 0\end{aligned} \tag{3.8}$$

Solving for $H(\omega_k)$, we get general form of Wiener Filter in Frequency domain as [31]

$$H(\omega_k) = \frac{P_{xy}(\omega_k)}{P_{yy}(\omega_k)} \tag{3.9}$$

For speech enhancement applications, equation 2.3 holds (see section 2.2.4). Rewriting the equation 2.3 at frequency ω_k

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \tag{3.10}$$

The desired signal $x(n)$ is the clean speech signal, whose estimation is the objective of the Wiener filter. The equation 3.9 gives the response of the Wiener filter in the frequency domain. To calculate $H(\omega_k)$ we need to compute $P_{xy}(\omega_k)$ and $P_{yy}(\omega_k)$ which can be done as [31]

$$\begin{aligned}P_{xy}(\omega_k) &= \mathbb{E}[X(\omega_k)Y(\omega_k)^*] \\ &= \mathbb{E}[X(\omega_k)(X(\omega_k) + D(\omega_k))^*] \text{ (substituting } Y(\omega_k) \text{ from equation 3.10)} \\ &= \mathbb{E}[X(\omega_k)X^*(\omega_k)] + \mathbb{E}[X(\omega_k)D^*(\omega_k)] \\ &= P_{xx}(\omega_k)\end{aligned} \tag{3.11}$$

Here, $\mathbb{E}$ is the expectation operator and $\mathbb{E}[X(\omega_k)D^*(\omega_k)] = 0$ as noise is assumed to be zero mean and uncorrelated with clean speech [31], [35]. Similarly, terms $\mathbb{E}[X(\omega_k)D^*(\omega_k)] = 0$ and $\mathbb{E}[D(\omega_k)X^*(\omega_k)] = 0$ in equation 3.12.

$$\begin{aligned}P_{yy}(\omega_k) &= \mathbb{E}[Y(\omega_k)Y(\omega_k)^*] \\ &= \mathbb{E}(X(\omega_k) + D(\omega_k))(X(\omega_k) + D(\omega_k))^*] \\ &= \mathbb{E}[X(\omega_k)X^*(\omega_k)] + \mathbb{E}[X(\omega_k)D^*(\omega_k)] \\ &+ \mathbb{E}[D(\omega_k)D^*(\omega_k)] + \mathbb{E}[D(\omega_k)X^*(\omega_k)] \\ &= P_{xx}(\omega_k) + P_{dd}(\omega_k)\end{aligned} \tag{3.12}$$

After substituting $P_{xy}(\omega_k)$ and $P_{yy}(\omega_k)$ from equations 3.11 and 3.12 in equation

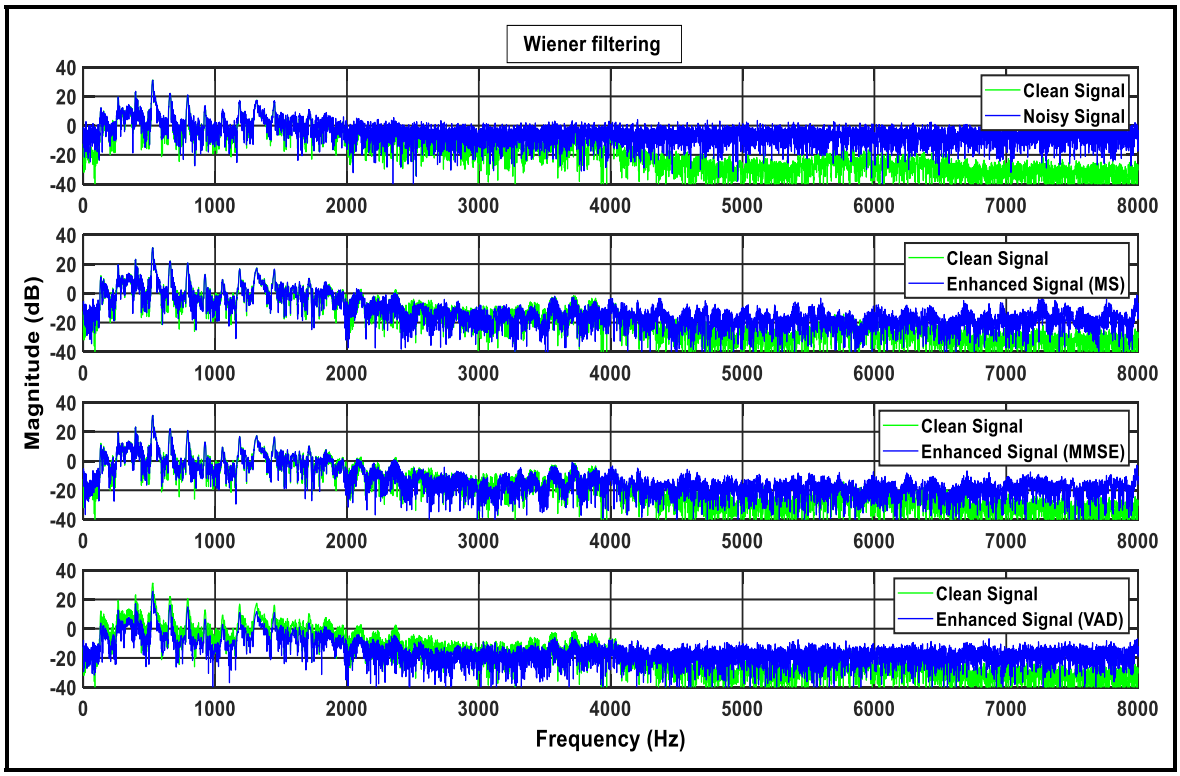

Figure 3.3 *The figure demonstrates long term average spectrum (LTAS) of clean and noisy signal in the top plot and LTAS of clean and enhanced signals in subsequent plots. Enhancement is done through Wiener filtering using different procedures for noise estimation. MS, MMSE and VAD are noise estimation techniques described in section 3.1.1. The clean speech signal is obtained from TIMIT [60] corpus and noisy speech signal is corrupted version of clean speech obtained by adding white Gaussian noise at SNR = 10 dB. In the plots, frequency (Hz) is plotted on X-axis and magnitude (dB) is plotted on Y-axis.*

3.9, we get,

$$H(\omega_k) = \frac{P_{xx}(\omega_k)}{P_{xx}(\omega_k) + P_{dd}(\omega_k)} \tag{3.13}$$

If we define ξ_k, *a priori* SNR at frequency ω_k as $\xi_k = \frac{P_{xx}(\omega_k)}{P_{dd}(\omega_k)}$, we can express Wiener filter in equation 3.13 as,

$$H(\omega_k) = \frac{\xi_k}{\xi_k + 1} \tag{3.14}$$

From the equation 3.13, it can be concluded that to estimate the Wiener filter the power spectrum of clean speech signal should be known which is a non-realistic assumption in the real applications. To mitigate this problem different techniques were utilised to estimate the suppression function of Wiener filter (equation 3.13). For example, [61] estimates *a priori* SNR to obtain suppression function (see equation 3.14). Once suppression function is obtained, the equation 3.4 can be used to estimate clean speech.

3.1.1 Noise estimation methods

Noise statistics estimation is critical for DSP based speech enhancement methods. The DSP based methods for spectral-subtractive algorithms (e.g. [40]), Wiener filtering (e.g. [41], [61]) approaches use noise estimation techniques for estimating noise spectrum from the noisy speech signal. Based on the SE technique utilized and estimated noise spectrum a suppression function is calculated (see section 2.2.4 for details). This suppression function is used in attenuating the spectral magnitude of noisy speech signal. The suppression function is dependent on estimated noise spectrum, so, it is important to have accurate noise estimates, otherwise overestimation of noise spectrum will result in speech distortion and possible loss of intelligibility and an underestimate will result in unpleasant residual noise in the enhanced signal [31]. Moreover, noise estimation becomes a very important component of a SE system, if the system needs to handle non-stationary noise [48]. In this section, we will describe three noise estimation techniques. They have been used with the DSP based SE algorithms used in the present work. The below discussions are brief and outlines the fundamentals of the procedures. For a thorough understanding of these techniques the provided references can be consulted.

1. **Noise Estimation using SAD** Speech activity detection (SAD also known as Voice activity detection (VAD)) is one of the fundamental techniques utilised in many speech processing tasks for identifying presence or absence of speech. In the context of speech enhancement methods, non-speech segments of signal identified by SAD can be utilised for the noise spectrum estimation. In the present work, we have used the SAD system (TO-Combo-SAD) developed by John Hansen's group at UT Dallas [62], [63].

   TO-Combo-SAD (Threshold-Optimized Combo SAD) system extracts five features (harmonicity, clarity, prediction gain and perceptual spectral flux; see [62], [63] for details) at a frame level from audio recordings. From these features, a 5-dimensional combo feature vector f_λ (λ is the frame index) is formed . The combo feature vector is then normalised to obtain normalised feature vector $\bar{f}_\lambda$ as [62]

   $$\bar{f}_\lambda = \frac{(f_\lambda - \mu)}{\sigma} \tag{3.15}$$

   Here, μ and σ are mean and standard deviation of features computed over the entire waveform. The normalised feature vector $\bar{f}_\lambda$ is further projected to X, the principle eigenvector corresponding to the largest eigenvalue of feature covariance matrix, to obtain projection p_λ as [62]

   $$p_\lambda = X^T \bar{f}_\lambda \tag{3.16}$$

   The features of the To-Combo-SAD system are designed in such a manner that p_λ values for speech is higher than non-speech. The speech/non-speech decision is made

with the help of this information by fitting a two mixture Gaussian Mixture Model (GMM) to the features and then estimating the detection threshold τ from weighted average of mixture means [62]

$$\tau = w\mu_{hs} + (1 - w)\mu_{hp} \tag{3.17}$$

Here, w is the weight factor such that $0 \leq w \leq 1$ and μ_{hs} and μ_{hp} are hypothesized speech and non-speech mixture means of the GMM. As the GMM model forces a bi-modal distribution, therefore it increases false alarms and misses in speech sparse and speech dense regions. To mitigate this problem of poor detection threshold, first a large mixture GMM model is trained on annotated speech only corpora and then the means of this GMM are projected into the single-dimension decision-making space

$$\hat{m_j} = X^T m_j \tag{3.18}$$

Here m_j is the j^{th} mixture mean of the M-mixture GMM and $\hat{m_j}$ is the projected value. Let, μ_{ts} be the mean of projected values $\hat{m_j}$. The μ_{ts} can be viewed as a prior model of speech (obtained from annotated speech corpora) and μ_{hs} can be considered as a posterior model of speech (constructed with current data) [62]. Now, the threshold decision can be improved by using $\max(\mu_{ts}, \mu_{hs})$. If the value of μ_{ts} is greater, it means the system trusts posterior model more and vice versa.

$$\tau = w\max(\mu_{ts}, \mu_{hs}) + (1 - w)\mu_{hp} \tag{3.19}$$

In the present work, we have used non-speech sections, identified by TO-Combo-SAD, to estimate average noise spectrum which is then employed in the estimation of suppression function.

2. **Minimum Statistics**

   Minimum Statistics (MS) approach with optimal signal power spectral density smoothing for noise estimation was proposed by Rainer Martin [48]. This method assumes that speech and the degrading noise are statistically independent and power of the noisy speech signal frequently decays to the power level of the degrading noise [48]. In other words, during speech pauses or intervals between words and syllables speech energy is close to zero. So, if minimum power is tracked during this period it will provide an estimation of noise floor. To capture, the very brief periods of zero speech energy the analysis window should be wide enough to bridge successive high power speech segments [48].

   As speech and noise are assumed to be independent, so it can be derived [31] that periodogram of noisy speech is approximately equal to the sum of the periodograms of clean speech and noise

$$|Y(\lambda,k)|^2 \approx |X(\lambda,k)|^2 + |D(\lambda,k)|^2 \tag{3.20}$$

Here, $|Y(\lambda,k)|^2$, $|X(\lambda,k)|^2$ and $|D(\lambda,k)|^2$ are periodograms of noisy speech, clean speech and, noise respectively. $|\cdot|$ is magnitude operation carried over short-time Fourier transform (STFT) of the signal, λ indicates the frame index, and k indicates the frequency bin index. The noise power spectrum is estimated by tracking the minimum of the periodogram $|Y(\lambda,k)|^2$. In practice, a recursively smoothed version of noisy speech periodogram is used with α as smoothing parameter [31], [48]

$$P(\lambda,k) = \alpha P(\lambda-1,k) + (1-\alpha)|Y(\lambda,k)|^2 \tag{3.21}$$

Here, $P(\lambda,k)$ represents smoothed periodogram. The original paper [48] further improves the equation 3.21 by employing a time varying smoothing parameter. For further details on optimal calculation of smoothing parameter and bias compensation strategy, the original paper [48] can be consulted. In the present study, the implementation of minimum statistics algorithm from VOICEBOX [64] is used which follows the approach from original paper [48].

3. **Unbiased minimum mean-squared error (MMSE)**

   A low complexity algorithm based on minimum mean-squared error (MMSE) estimator of the noise magnitude-squared DFT coefficients was proposed by Hendriks et al. [65]. The technique uses a limited maximum-likelihood (ML) estimate of *a priori* SNR to calculate MMSE estimate of the noise periodogram. However, estimation of *a priori* SNR introduces a bias in the MMSE-based estimator [66]. In order to compensate for the bias, the authors use a second estimate of *a priori* SNR calculated through a decision-directed [67] approach.

   The method assumes noise is additive and uncorrelated with speech source. So, equation 2.3 holds. Rewriting the equation 2.3 with frame index (λ) and frequency index (k)

$$Y(\lambda,k) = X(\lambda,k) + D(\lambda,k) \tag{3.22}$$

   Further, the discrete Fourier Transform (DFT) coefficients of noise and speech are assumed to have complex Gaussian distribution [66]. The spectral speech and noise power are defined by $\mathbb{E}(|X|^2) = \sigma_X^2$ and $\mathbb{E}(|D|^2) = \sigma_D^2$ respectively. The *a posteriori* SNR is defined by $\gamma = \frac{|Y|^2}{\sigma_D^2}$ and the *a priori* SNR by $\xi = \frac{\sigma_X^2}{\sigma_D^2}$ [65], [66].
   The noise periodogram is given by equation 3.23 [66]

$$|\hat{D}|^2 = (\frac{1}{1+\hat{\xi}})^2|Y|^2 + (\frac{\hat{\xi}}{1+\hat{\xi}})\hat{\sigma_D^2} \tag{3.23}$$

After estimating the noise periodogram from equation 3.23, the noise power spectral density can be updated via recursive smoothing using equation 3.24.

$$\hat{\sigma_D^2}(\lambda) = \alpha\hat{\sigma_D^2}(\lambda - 1) + (1 - \alpha)(|\hat{D}(\lambda)|^2) \tag{3.24}$$

Here, α is smoothing parameter and is used as $\alpha = 0.8$ in [66].

Gerkmann et al. [66] demonstrated that MMSE based noise power estimator (see equations 3.23, 3.24) is only updated when *a posteriori* SNR is less than 1. This threshold on *a posteriori* SNR can be interpreted as a hard decision by a speech activity detector (SAD) (SAD gives a binary response, speech is either present or absent). As an improvement on the approach proposed by [65], Gerkmann et al. [66] proposed to replace the hard decision of the SAD by a soft decision of Speech Presence Probability (SPP) with fixed priors, making bias compensation redundant and resulting in an unbiased MMSE estimator. In the present study, the implementation of unbiased MMSE algorithm from VOICEBOX [64] is used which follows the approach from original papers [65], [66].

3.2 Machine learning-based methods for SE

This section describes two recent approaches to SE proposed by Keren et al. [39] and Sun et al. [50] using deep neural networks. In contrast to the DSP based methods, they do not require any explicit noise estimation procedure.

1. **DNN based approach proposed by Keren et al.**

   In their recent work, Keren et al. [39] ($\text{DNN}_{\text{Keren}}$) described a deep neural network based speech enhancement system which has been designed to generalize to unseen noisy environments. This has been achieved by training the model with noises from a large number of different environments ($16,784$), mixed with clean speech at different SNR's (signal-to-noise ratio). The authors expect that this large set of noises from different environments should share some properties with unseen noisy signal and thus, exposing the model to such a large set during training should assist in generalisation of enhancement process to unseen noisy environments. Moreover, the model can accept additional non-speech recordings (providing non-speech recording to model is optional) from the noisy speech signal environment to generate a noise embedding which conditions the residual layers of enhancement subnetwork (see figure 3.4). The authors have hypothesized that such an approach (injecting noise embeddings to residual layers) might assist the model in identifying which frequency components are to be discarded and which are to be enhanced [39]. The authors state that their best model has demonstrated considerable reduction in word error rate (WER) when used as a front-end for a pretrained speech recognition system, reducing WER from 34.04% on noisy speech to 15.46% on enhanced speech.

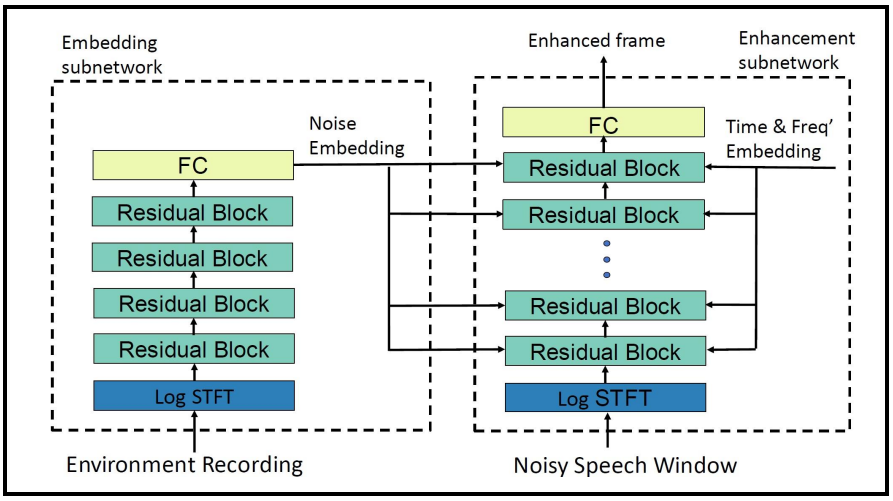

Figure 3.4 *The enhancement model architecture proposed by Keren et al. [39] (reproduced from [39]).*

The model consist of two distinct but linked processing blocks (see figure 3.4): the embedding and the enhancement subnetwork. Both the processing blocks take log absolute value of the STFT as input [39]. The embedding subnetwork generates noise embeddings from an additional recording by processing the non-speech input signal with a sequence of 4 residual blocks (see [68] for details on residual blocks). Each residual block comprises of two 2D-convolutional layers, each layer having same number of feature maps. The number of feature maps in each residual block is twice the previous block. It begins as 64 for the first residual block and ends at 512 for the fourth residual block. After the processing of the additional recording is finished through the 4 residual blocks (for details see original paper [39]), a single 512-dimensional embedding vector is obtained by averaging the feature maps from all the 4 blocks through the fully connected layer (denoted as FC in the figure 3.4). The enhancement subnetwork has 8 residual blocks. The structure of each of these residual blocks in enhancement subnetwork is identical to those in the embedding subnetwork in terms of convolutional layers and number of feature maps in convolutional layers. Each residual block is repeated twice in enhancement subnetwork (e.g. residual block 1 in embedding subnetwork is identical to residual blocks 1 and 2 in enhancement subnetwork). In case an additional noise recording is supplied with noisy speech signal then each residual layer of enhancement subnetwork is conditioned with noise embedding generated by the embedding subnetwork. Additionally, the time steps and frequency components indices are added at appropriate locations in the output map of convolutional layers to enable them to process different time steps and frequency components differently [39]. The output of the final residual block of enhancement subnetwork is flattened and fed into a fully connected (denoted by FC in the figure 3.4) layer to obtain an enhancement mask which is then applied to the noisy speech segment to obtain an enhanced frame.

In the present work, we have used the pretrained model and python code files pro-

vided by the original authors for conducting speech enhancement.

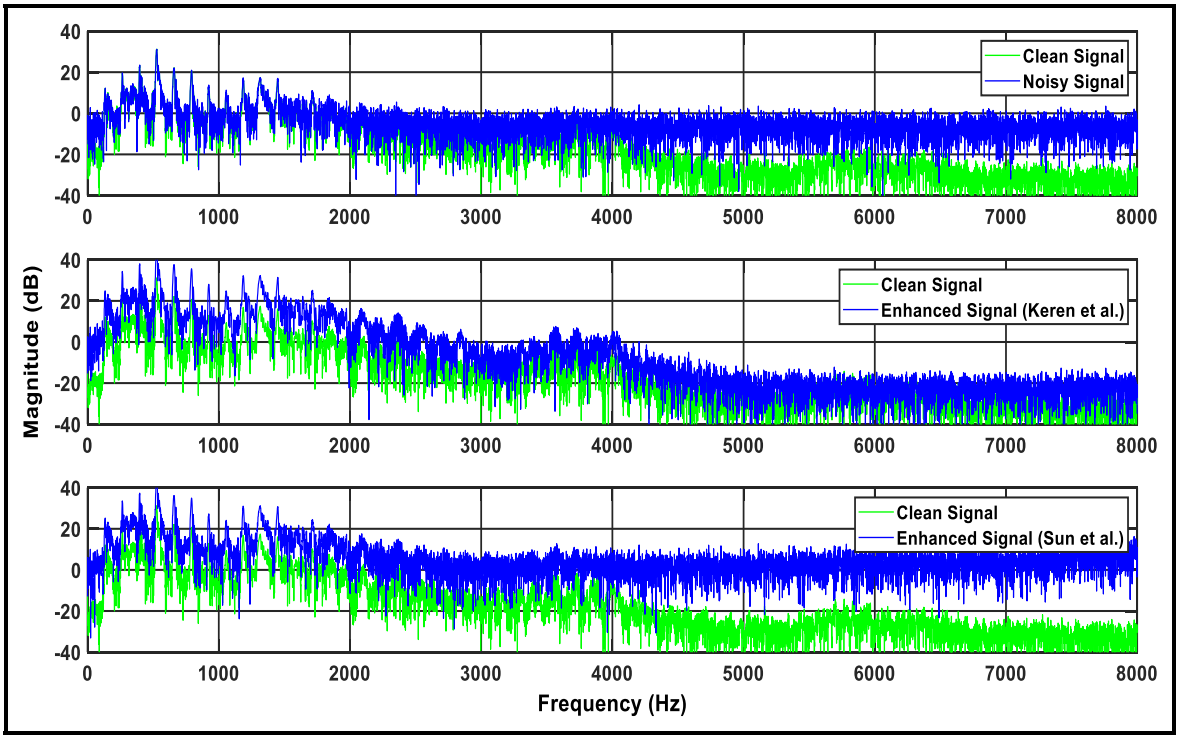

Figure 3.5 *The figure demonstrates long term average spectrum (LTAS) of clean and noisy signal in the top plot and LTAS of clean and enhanced signals in subsequent plots. Middle plot has enhancement done by method DNN_{Keren} [39] and bottom plot has enhancement done by method DNN_{Sun} [50]. The clean speech signal is obtained from TIMIT [60] corpus and noisy speech signal is corrupted version of clean speech obtained by adding white Gaussian noise at SNR = 10 dB. In the plots, frequency (Hz) is plotted on X-axis and magnitude (dB) is plotted on Y-axis.*

2. **DNN based approach proposed by Sun et al.**

   The speech enhancement system proposed by Sun et al. [23], [50], [69] (DNN_{Sun}) consists of a stack of LSTM layers where each layer is designed to progressively learn intermediate speech at a higher SNR then the previous layer. The concept of progressive learning is illustrated in figure 3.6 and details can be obtained from the original paper [70]. Each LSTM layer contributes to multistage processing of the noisy signal by accepting the original log-power spectrum features and target from previous layer as input and producing output speech target at a higher SNR. This process is demonstrated in figure 3.7, where target 1, 2 are intermediate speech targets and target 3 is the final output from the network. At each stage the estimated target speech (except the final stage) is spliced together with the original input and fed to the successive LSTM layer to learn the next target. This SE system acted as a front-end to the challenging problem of Speaker Diarization for the first DIHARD challenge [50]. The authors have reported that a SE front-end constructed with this model has conclusively boosted the performance of the diarizaion system, including

the increased performance of intervening task of speech activity detection can be directly attributed to the enhanced speech [23], [50].

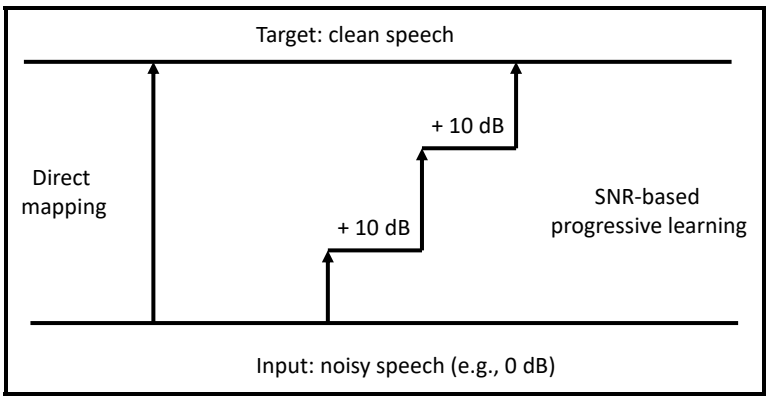

Figure 3.6 *Progressive learning for speech enhancement (adapted from [70]). The direct mapping from noisy speech to clean speech is decomposed into multiple intermediate stages with each stage learning the speech at a higher SNR than the previous stage.*

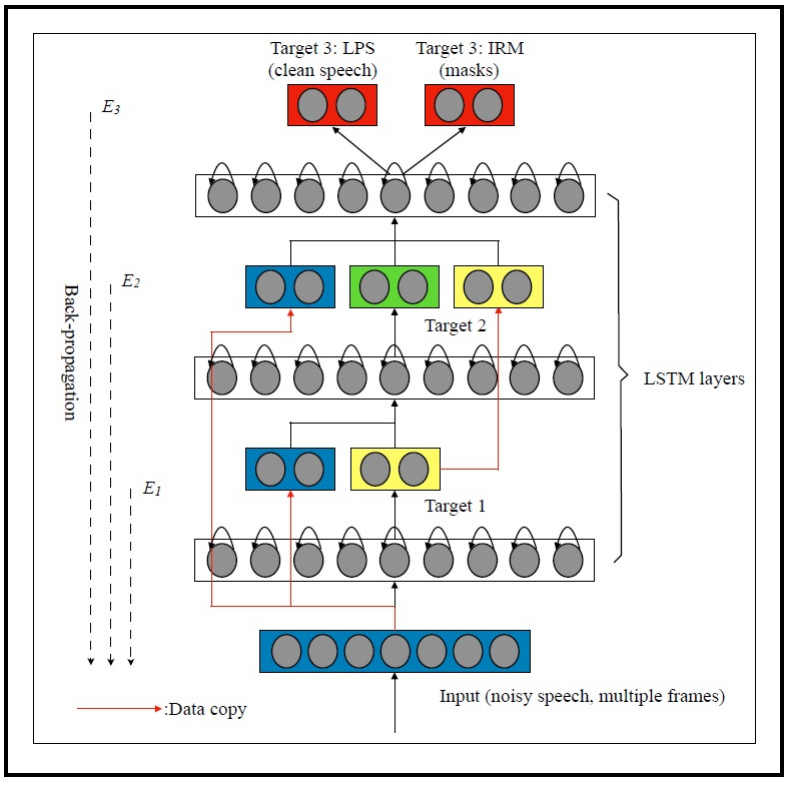

Figure 3.7 *Architecture of the model proposed by Sun et al. [50][23] (reproduced from [50]).*

The network is trained on 400 hours of simulated clean/noisy pairs of speech signals. The clean speech was obtained from WSJ0 corpus [71] (read English) and 863 Program corpus [72] (spoken Mandarin). The clean speech was corrupted with 115 noise types at different SNR's [23], [50]. The output of the network is the clean log-power spectral features and IRM (see figure 3.7). The predicted ratio mask is

used to construct an enhanced waveform. For further details on architecture and training of this model [23], [69], [73] can be consulted.

In the present work, we have used the pretrained model and python code files provided by the original authors for conducting speech enhancement.

3.3 Objective measures for quality and intelligibility of the enhanced speech

In this section, we describe the two objective measures utilised in the present work for the assessment of the quality and intelligibility of the speech enhanced by the aforementioned SE methods. Both of these, methods are intrusive in nature as they both require pair of input clean signal and processed output signal for calculation of objective metrics.

3.3.1 Spectral distortion for the quality assessment

The measurement of the distortion between a pair of input and output signal to a speech processing system involves the assignment of a non-negative number to such a pair [74]. A meaningful distortion measure will subjectively correlate well with speech quality i.e. small distortion translates to good speech quality and large distortion indicates bad speech quality [51]. In the context of speech enhancement, distortion measure calculated between clean speech and corrupted speech when compared with distortion between clean speech and enhanced speech should give us an idea about efficacy of the SE method. In other words, if distortion between clean speech and corrupted speech is d then after application of SE method the distortion between clean speech and enhanced speech should be $< d$.

In order to calculate spectral distortion between a pair of signals we first obtain the log-Mel representation of the signals. Incorporation of the Mel-scale in the transformation of signal in the spectral domain ensures that we inject auditory perception sensitivities in the spectral representation of the signal. The process of calculating log-Mel representation of the signal begins with the calculation of the absolute value of FFT of the windowed signal. The absolute value of FFT is then multiplied with the Mel-filter bank to get the signal representation in the Mel-spectrum and finally, the logarithm of the output is taken to get log-Mel representation of the signal. After log-Mel representation of the signal pair is obtained the distortion is calculated as root mean square (RMS) distance between the two log-Mel representations.

3.3.2 SIIB for the intelligibility assessment

SIIB, which stands for Speech intelligibility in bits [55], is an intrusive measure for estimating the instrumental intelligibility of speech. The SIIB metric is based on the information theoretic principles. It tries to capture mutual information between clean and degraded speech and hypothesises intelligibility as a function of the mutual information. Below derivation is taken from the original papers [55], [56].

Let us consider, a message $\{M_t\}$, speech signal $\{X_t\}$ and noisy (or corrupted) speech $\{Y_t\}$ represented by stationary ergodic discrete-time real-valued random processes (t is the time index). The message $\{M_t\}$ is encoded into the speech signal $\{X_t\}$ by the transmitter. While transmitting the speech signal, the signal can be distorted by noise, reverberation, or even speech processing algorithms such as coding, enhancement etc. This results in receiver getting the degraded speech $\{Y_t\}$. The authors represents the clean and degraded speech as auditory log-spectra, obtained by the application of auditory filter bank to the squared magnitude of Short-time Fourier transform (STFT) of the speech signal and followed by a logarithm operation.

The communication process from message to speech and to transmission and receiving of speech can be represented as a Markov chain,

$$\{M_t\} \rightarrow \{X_t\} \rightarrow \{Y_t\} \tag{3.25}$$

When three random variables form a markov chain, we know that, joint probability mass function can be denoted as (see [75]),

$$p(m_t, x_t, y_t) = p(m_t)p(x_t|m_t)p(y_t|x_t) \tag{3.26}$$

If we incorporate the concept of data processing inequality, we can represent the equation 3.26 in terms of mutual information (represented by I) as,

$$I(\{M_t\}; \{Y_t\}) \leq I(\{M_t\}; \{X_t\}) \tag{3.27}$$

and

$$I(\{M_t\}; \{Y_t\}) \leq I(\{X_t\}; \{Y_t\}) \tag{3.28}$$

Combining equations 3.27 and 3.28 , we get

$$I(\{M_t\}; \{Y_t\}) \leq min(I(\{M_t\}; \{X_t\}), I(\{X_t\}; \{Y_t\})) \tag{3.29}$$

The goal of SIIB is to capture the mutual information between the message and the degraded speech and relate the estimated mutual information to the intelligibility. The message to be transmitted can be thought of as sequence of latent variables that represent, for example, a sequence of sentences.

Let $M^K = [(M_1)^T, (M_2)^T, ..., (M_K)^T]$ (here T denotes transpose) represent a stack of K consecutive message vectors, similarly we can define degraded speech Y^K. Now, the mutual information between two random variables (see Markov chain illustration 3.25), is given as,

$$I(M^K; Y^K) = \int_{M^K, Y^K} p(M^K, Y^K) log(\frac{p(M^K, Y^K)}{p(M^K)p(Y^K)}) dM^K dY^K \tag{3.30}$$

We can now define, mutual information rate between $\{M_t\}$ and $\{Y_t\}$ as,

$$I(\{M_t\};\{Y_t\}) = \lim_{K\to\infty} \frac{1}{K} I(M^K; Y^K) \tag{3.31}$$

To estimate $I(\{M_t\};\{Y_t\})$, realizations of $\{M_t\}$ and $\{Y_t\}$ will be needed which might not be possible in realistic applications. So, we consider equation 3.29, to estimate the mutual information between $\{M_t\}$ and $\{Y_t\}$.

In the equation 3.29, we need the mutual information rate between $\{M_t\}$ and $\{X_t\}$ as well as between $\{X_t\}$ and $\{Y_t\}$. It is beyond the scope of present work to delve into the details of estimation of these mutual information rates, so we will only provide the equations, explanation of notations and implementation specific details here. The readers can consult original papers [55] and [56] for further details.

The mutual information rate between $\{X_t\}$ and $\{Y_t\}$ is given by,

$$I(\{X_t\};\{Y_t\}) = \lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{KJ} I(\tilde{X}_j^K; \tilde{Y}_j^K) \tag{3.32}$$

Here, j denotes elements index in the vector, $\tilde{X}^K$ and $\tilde{Y}^K$ are KLT (Karhunen-Loeve Transform) transforms for X^K and Y^K.

The mutual information rate between $\{M_t\}$ and $\{X_t\}$ is given by,

$$\begin{aligned} I(\{M_t\};\{X_t\}) &= \lim_{K\to\infty} \frac{1}{K} \sum_{j=1}^{KJ} I(\tilde{M}_j^K; \tilde{X}_j^K) \\ &= \lim_{K\to\infty} -\frac{1}{K} \sum_{j=1}^{KJ} \frac{1}{2} \log_2(1 - r_j^2) \end{aligned} \tag{3.33}$$

Here, j denotes elements index in the vector, $\tilde{M}^K$ and $\tilde{X}^K$ are KLT (Karhunen-Loeve Transform) transforms for M^K and X^K, r_j is production noise correlation coefficient which describes the efficiency of encoding a message to speech signal and is measured to be 0.75 for all j.

Now, if we combine equations 3.32 and 3.33, as per equation 3.29, we get,

$$SIIB = \frac{F}{K} \sum_{j=1}^{KJ} min(-\frac{1}{2} \log_2(1 - r_j^2), I(\tilde{X}_j^K; \tilde{Y}_j^K)) \tag{3.34}$$

Here F is the frame rate in Hz.

At the level of implementation, the information rate $I(\tilde{X}_j^K; \tilde{Y}_j^K)$ is estimated by applying a k-nearest neighbour mutual information estimator [76] to the sample sequence $\tilde{X}_{j,t}^K$ and $\tilde{Y}_{j,t}^K$.

To further simplify formulation in 3.34, the authors in the following paper [56] have quantified the mutual information between clean ($\{X_t\}$) and degraded ($\{Y_t\}$) speech by assuming clean and degraded signals are jointly Gaussian and the mutual information can be estimated using information capacity of a Gaussian channel.

$$SIIB^{Gauss} = -\frac{F}{2K}\sum_{j}\log_2(1 - r^2\rho_j^2) \tag{3.35}$$

Here, ρ_j is the correlation coefficient between j^{th} clean eigenchannel and j^{th} degraded eigenchannel. The parameters in equation 3.35 are defined as $F = 80$ Hz, $K = 15$ and $r = 0.75$ (see [55], [56] for further explananation).

In the present work, we have used $SIIB^{Gauss}$ as the metric for the intelligibility measurement.

3.4 Downstream tasks

In this section we will describe the two downstream speech analysis tasks used in the experiments of this work: 1) IDS/ADS classification, i.e., speech produced by an adult has to be classified whether it is directed towards an infant (or child) or adult (infant-directed speech (IDS) or adult-directed speech (ADS)), and 2) automatic syllable count estimation, whose goal is to provide an estimate of the number of syllables in a speech signal.

3.4.1 IDS/ADS classification

The task of automatic classification of speech either directed towards an infant or an adult is useful in the scientific study of language exposure to the child [13], [18], [77]. The infant directed speech is a speaking style often used when addressing to an infant and is understood to engage an infant's attention more productively [77]. IDS has also been found to possess distinct phonetic and linguistic characteristics (e.g. raised pitch, wider pitch range, exaggerated prosody, hyper-articulation of vowels, slower speech rate, reduced linguistic complexity [78]) that makes it distinguishable from ADS.

In the present work, we have used a Support Vector Machine (SVM) [79] with a linear kernel function to perform the binary classification between IDS/ADS. This approach is similar to the baseline systems used in Computational Paralinguistics challenges (ComParE) held at INTERSPEECH conferences (e.g., [25] where addressee sub-challenge was identification of infant-directed or adult-directed speech). The feature set supplied to the SVM was INTERSPEECH-2013 ComParE challenge feature set [80] extracted using openSMILE toolkit [81]. The set includes a total of 6373 features corresponding to utterance-level statistical descriptors (mean, variance etc.) of low-level signal features such as MFCCs, PLPs (similiar to MFCCs, cube root suppression instead of log suppression), F0, and zero-crossing rate. The SVM was always trained and tested on features from signals enhanced with the same SE method.

3.4.2 Automatic syllable count estimation

The task of automatic syllable count estimation from speech signal has usage in variety of applications, for example speaking rate estimation [82], daily activity analysis from long duration personal audio recordings [22], [83], quantification of linguistic input a child hears

from its natural environment [11], [21] etc. For a child language development researchers, data on quantity of language child hears helps them to correlate language development in children with the amount of language input children get from their daily environment (e.g. [18]).

In the present work, we evaluate a recently proposed deep neural network based algorithm SylNet [84] which performs automatic syllable count estimation from input speech signal. Architecture of the SylNet is motivated from the WaveNet [85] model. WaveNet is a deep neural network based model that can generate realistic sounding audio from text. The authors have used WaveNet inspired architecture in combination with an additional LSTM layer to estimate syllable count from an utterance.

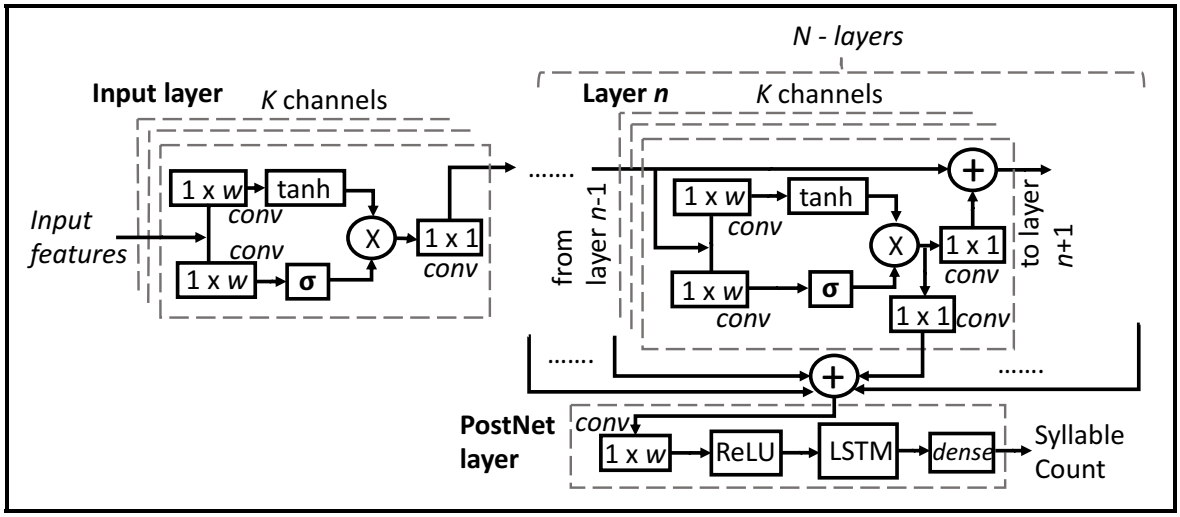

Figure 3.8 *Architecture of the SylNet model (reproduced from [84]).*

Figure 3.8 illustrates the architecture of SylNet. The model takes an input of log-Mel spectrum of input speech. The input is fed to a stack of convolutional layers, each layer having gated non-linearities and skip connections to the PostNet layer. The PostNet layer acts as an integrator for all the information coming from the residual connections from the N-layers (see figure 3.8). The LSTM layer combines syllable count information coming over time from PostNet layer and feeds it to the dense layer. The dense layer outputs the estimate of syllable count in the utterance.

The SylNet model is quite promising for speech analysis tasks as the results presented by authors indicate that it generalises well for different languages and performs better if given additional data to adapt to the novel language. It also performed better when compared to other tested supervised and unsupervised syllabification algorithms on a range of held-out languages as presented in the original paper [84].

Here, we use a pretrained SylNet model from the original study [84] and adapt it to the section of enhanced data. The rest of the data is used as test data for getting syllable count estimate from the SylNet model and in this manner, we also maintain the consistency of SE method between the adaptation and the test set.

4 Experiments

This chapter describes the data and experimental setup used in the present work.

4.1 Data

4.1.1 Data for objective evaluation of SE methods

For the objective assessment, the clean speech files provided by the TIMIT Acoustic-Phonetic Continuous Speech Corpus [60] were used. The TIMIT corpus has broadband recordings of phonetically rich read speech from 630 speakers of eight major dialects of American English. The corpus has been delivered in two neatly divided test and train sets for easier application to tasks such as ASR etc. which require model training. For the objective evaluation, separate training and test sets were not required. The two sets (train and test) were combined to get a consolidated set containing 6300 speech waveforms from 630 speakers. From this set, 29 speech waveforms were combined together at a time to get approximately 90 seconds long speech waveform. This process reduced the size of consolidated set to total 217 waveforms. The concatenation of shorter waveforms was done as SIIB [55], [56] metric requires long signals to operate. This set of speech waveforms hereafter is referred as $TIMIT_{All}$.

4.1.2 Data for downstream evaluation of SE methods

The downstream speech data utilised for the experiments has been collected as part of ACLEW project corpus [21]. This data has been collected in the form of six different corpora of child centered day long recordings. These include the Bergelson corpus ("BER") from US English speaking families from Rochester(US) [86], the LuCiD Language 0-5 corpus ("L05") consisting of English-speaking families from NorthWest England [87], the Casillas corpus ("CAS") of Tseltal-speaking families form a rural Mayan community in Southern Mexico [88], the McDivitt and Winnipeg corpora ("MCD") of Canadian English families [89], [90], the Warlaumont corpus ("WAR") of US English speaking families from Merced, California [91] and the Rosemberg corpus ("ROS") of Argentinian Spanish families from Buenos Aires metropolitan area [92]. The recordings were collected with a device young child wore in a breastpocket throughout a normal day. BER, MCD, L05, and WAR recordings were collected with the LENA[1] recorder, while CAS was recorded with Olympus WS-382 or WS-852, and ROS was recorded with a mix of Olympus, Panasonic, Sony, and LENA recorders. Each corpus consists of daylong (4–16 hour) at-home recordings, with spoken language varying across corpora. Due to the unconstrained nature of the recordings, they contain both near and far-field speech in acoustically varied environments and at highly varying SNRs. The approximate average speech SNRs for different corpora

[1] `https://www.lena.org/`

are BER 2.1 dB, CAS -0.5 dB, L05 3.6 dB, ROS -2.6 dB, MCD 0.8 dB, and WAR 2.4 dB as reported by Räsänen et al. [11].

Table 4.1 *Details of corpora and speech recordings used in downstream experiments.* ***Audio total*** *= total amount of audio annotated for verbal activity;* ***Speech total*** *= duration of all utterances in the annotated audio;* ***Adult speech total*** *= total duration of utterances from male or female adults that contain at least one unambiguously transcribed word. Table is adapted from [11].*

ID	Corpus name	Region	Language	Subjects (N)	Audio total (hours)	Speech total(mins)	Adult Speech total(mins)	Audio per subject(mins)
BER	Bergelson	Northeast US	US English	10	5.0	116.7	50.7	30.0
CAS	Casillas	Northern Bolovia	Tseltal	10	7.5	212.0	100.8	45.0
L05	Language 0-5	Northwest England	UK English	10	5.0	95.9	39.1	30.0
ROS	Rosemberg	Buenos Aires Metropolitian area Argentina	Argentinian Spanish	10	5.0	149.3	70.3	30.0
MCD	McDivitt+	Western Canada	Canadian English	8	4.5	80.9	44.0	33.8
WAR	Warlaumont	Western US	Us English	10	5.0	100.3	39.6	30.0
			Total	**58**	**32.0**	**755.1**	**344.5**	

From each these corpora, 8–10 children were sampled for manual annotation, selected children were carefully chosen to represent diversity of infant ages (0-36 months) and socio-economic contexts (also see Table 4.1). For each child, 15 x 2-minute randomly sampled audio segments (9 x 5 min for Tseltal, see [88]) were manually annotated for hearable utterance boundaries, who is the addressee (child vs. adult), who is the speaker, vocal maturity of child vocalizations, and all adult speech was transcribed. The manual annotation procedure followed a comprehensive protocol documented in [93], [94].

Reference syllable counts of the daylong audio were obtained by automatic syllabification of the hand-annotated transcripts. First, the transcripts were cleaned up of all non-lexical entries such as incomprehensible speech, non-linguistic communicatives (e.g., <hmm>) and other non-speech sounds (e.g., <yawn>), and paralinguistic markers (e.g., <singing> to denote singing speaking style). The resulting word strings were then converted into sequences of phonemes using Phonemizer tool (`https://github.com/bootphon/phonemizer`), and finally syllabified using the maximum onset principle (see, e.g., [95]). In short, maximum onset principle creates syllabic boundaries such that the algorithm operates backwards along the string of words word-by-word, and for each vocalic nucleus within a word, assigns the maximum number of preceding consonants to the syllable so that the resulting consonant cluster is still a valid syllable onset in the given language. Note that while the procedure does provide relatively systematic syllable counts for the given transcripts, the resulting phonology-based syllable counts should not be taken as an error-free gold standard of the syllabic structure of what was actually said (phonetic syllables). This is since some of the information regarding the detailed style of speech (rhythm & timing, pronunciation) is lost in the speech-to-transcript and transcript-to-phonemes conversions of the present pipeline.

For our experiments, we extracted all the annotated utterances by adult speakers. We also included a "*collared*"-version of the dataset where 500-ms of non-speech signal leading

and trailing each utterance was concatenated to the utterance (e.g. if utterance boundary in a speech segment, begins at t_0 second and ends at t_1 second, the non-collared set has the utterance exactly $(t_1 - t_0)$ seconds long whereas the collared set had the utterance $(t_1 - t_0) + 1$ seconds long beginning at $(t_0 - 0.5)$ second and ending at $(t_1 + 0.5)$ second. This was done to provide additional temporal signal context to the SE algorithms, with the intention that the additional context would improve enhancement performance due to more opportunities for estimating the noise statistics. The collars were removed after SE was completed and before the downstream tasks were evaluated.

4.2 Setup for SE methods

The SE methods consist of classical signal processing-based methods and modern machine learning-based methods described in chapter 3 and listed in Table 4.2. The process of SE was conducted in MATLAB environment except for machine learning-based methods. The machine learning-based methods used python scripts and pre-trained models shared by the original authors. MS and MMSE -based spectral subtraction and Wiener filtering were performed with VOICEBOX toolbox [64], whereas VAD -based variants used TO-Combo-SAD implementation from original authors [62], [63]. Speech enhancement setup was same for both objectve assessment and downstream evaluation.

Table 4.2 *SE methods compared in the present study. Sun et al. [23] and Keren et al. [39] are the recently proposed machine learning-based methods (represented as Deep Neural Network (DNN)) for speech enhancement. The subscript notations with method name indicates noise estimation technique except in case of machine learning-based methods where they indicate name of the author. The third column of table expands the acronym of noise estimation technique.*

SE Method	Representation	Noise Estimation
LSTM layers with progressive learning (Sun et al.)	DNN_{Sun}	NA
CNNs with ResNet like Architecture (Keren et al.)	DNN_{Keren}	NA
Spectral Subtraction	$SpecSub_{MMSE}$	Minimum mean square estimation
Wiener Filtering	$Wiener_{MMSE}$	Minimum mean square estimation
Spectral Subtraction	$SpecSub_{MS}$	Minimum statistics
Wiener Filtering	$Wiener_{MS}$	Minimum statistics
Spectral Subtraction	$SpecSub_{VAD}$	Voice activity detection
Wiener Filtering	$Wiener_{VAD}$	Voice activity detection

4.3 Setup for objective assessment

The purpose of this setup is to evaluate the performance of SE methods (described in chapter 3 and listed in Table 4.2) with respect to objective measures of speech quality in controlled and constrained conditions. The data utilised for this experiment has been described in section 4.1.1. The objective measures of speech quality calculated in this experiment are instrumental intelligibility (estimated using SIIB [55], [56]) and spectral

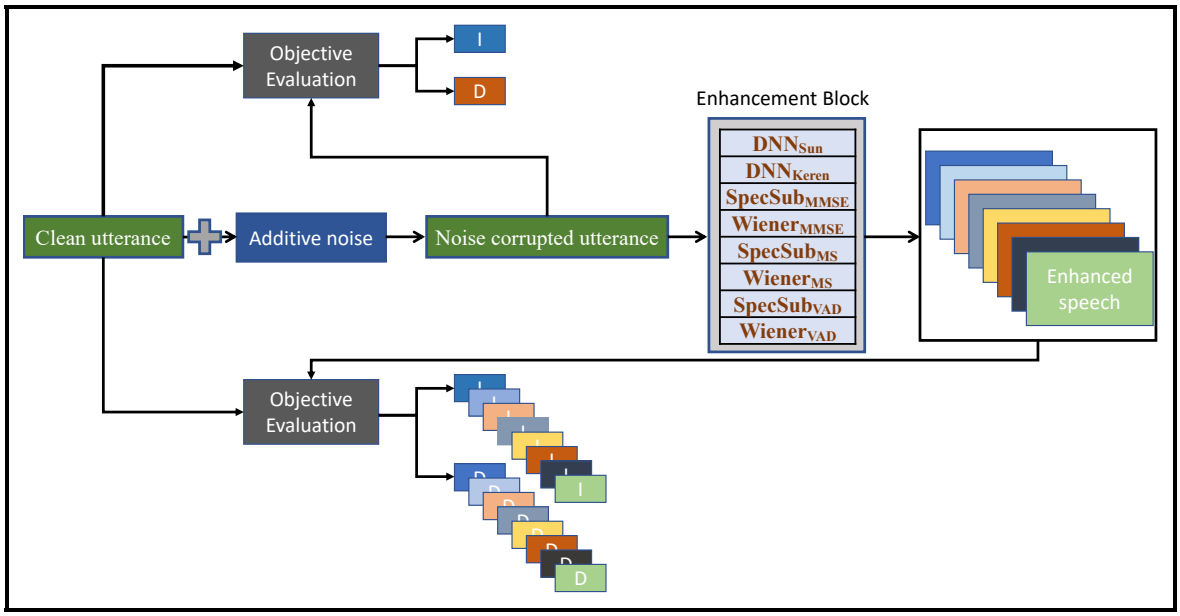

Figure 4.1 *Diagram depicts the process of objective evaluation of enhancement methods on noise corrupted utterance.* ***I*** *indicates intelligibility calculated through SIIB and* ***D*** *is spectral distortion.*

distortion. Both measures are described in chapter 3. The figure 4.1 illustrates the whole process through block schematic diagram.

For the objective assessment, all waveforms in $TIMIT_{All}$ dataset were degraded with 3 different additive noises with different statistical characteristics to get a set of corrupted speech signals. The noises added were white Gaussian noise (WGN), babble noise and noise sampled randomly from child-centered daylong recordings. These noises were added at varying SNRs beginning from -20 dB to $+40$ dB in steps of 5 dB. This process yielded 39 different degraded waveforms for a single clean speech waveform (3 different noises, and each noise added at 13 different SNRs, so $3 \times 13 = 39$). The SE methods illustrated in the enhancement block (see figure 4.1) act parallelly on a single noisy speech signal to produce a set of enhanced speech signals. In other words, when a single noisy utterance is processed by the enhancement block eight distinct enhanced waveforms are obtained each corresponding to the SE method used for denoising. The baseline results for objective assessment are obtained by comparing clean speech and degraded speech. Similarly, for the enhanced waveforms, the objective metrics are calculated by comparing clean speech with enhanced speech. This process is repeated for the entire dataset. The experiment was conducted in the MATLAB environment. MATLAB scripts for SIIB calculation was provided by the original authors.

4.4 Setup for downstream tasks evaluation

The purpose of this setup is to evaluate the performance of SE front-ends constructed with SE methods (listed in Table 4.2) with respect to downstream speech processing task performance (described in chapter 3). The data utilized in this setup is described in section

4.1.2. For this experiment, the enhanced signals corresponding to each SE method on both datasets (non-collared and collared) were obtained. Thus, from all SE methods 16 enhanced sets of signals were obtained ($8 \times 2 = 16$). Including the signal from original non-collared dataset, in total 17 sets of signals were used for evaluating performance of the downstream tasks. In the subsections below, individual downstream task setup is presented.

4.4.1 Automatic syllabification

For the task of automatic syllable count estimation, each of the datasets were divided into adaptation and testing sets randomly in ratio 1:10. The automatic syllabification was performed using a deep learning-based model called SylNet[84] (described in chapter 3). The adaption set was used for SylNet adaptation and testing set was used for performance evaluation. SylNet model and associated Python scripts were kindly provided by the original authors. Performance of the task was measured in terms of mean absolute relative error at the utterance level, i.e., relative % that estimated syllable count and ground-truth syllable count differ for each individual utterance independent of the sign.

4.4.2 IDS/ADS classification

For the IDS/ADS classification task, the datasets were subjected to feature extraction procedure (described in chapter 3). Features with associated labels (addressee information) was bundled up for the classification task. The bundle of features and labels was cleared of any other cases of addressee (meta data associated with utterances had other addressee information as well, e.g. speech directed to a composite group) except infant directed and adult directed and then features were normalised. The bundle was randomly split in training and testing sets with 2:1 (7922 for training, 3961 for testing). Since there were more ADS (N = 4515) than IDS (N = 3407) utterances in the training set, the training class distribution was balanced using sampling with replacement for IDS utterances so that every IDS utterance was used at least once in the training. The performance of classification task was measured using unweighted average recall (UAR), as it is the preferred metric in paralinguistic tasks with biased class distributions (e.g., [25], [26]). The experiment was performed in the MATLAB environment.

5 Results

In this chapter, the results of experiments conducted as part of this work are presented. The description of results is followed by a discussion of the obtained results.

5.1 Results for objective evaluation of SE methods

The results for objective assessment are presented in the figures 5.1, 5.2 and 5.3. Each figure presents result for a different additive noise added at varying SNRs. The objective metrics were calculated for SNR range −20 to 40 dB, but in this reporting a subset of SNR range from −15 to 20 dB is presented to maintain legibility in the figures.

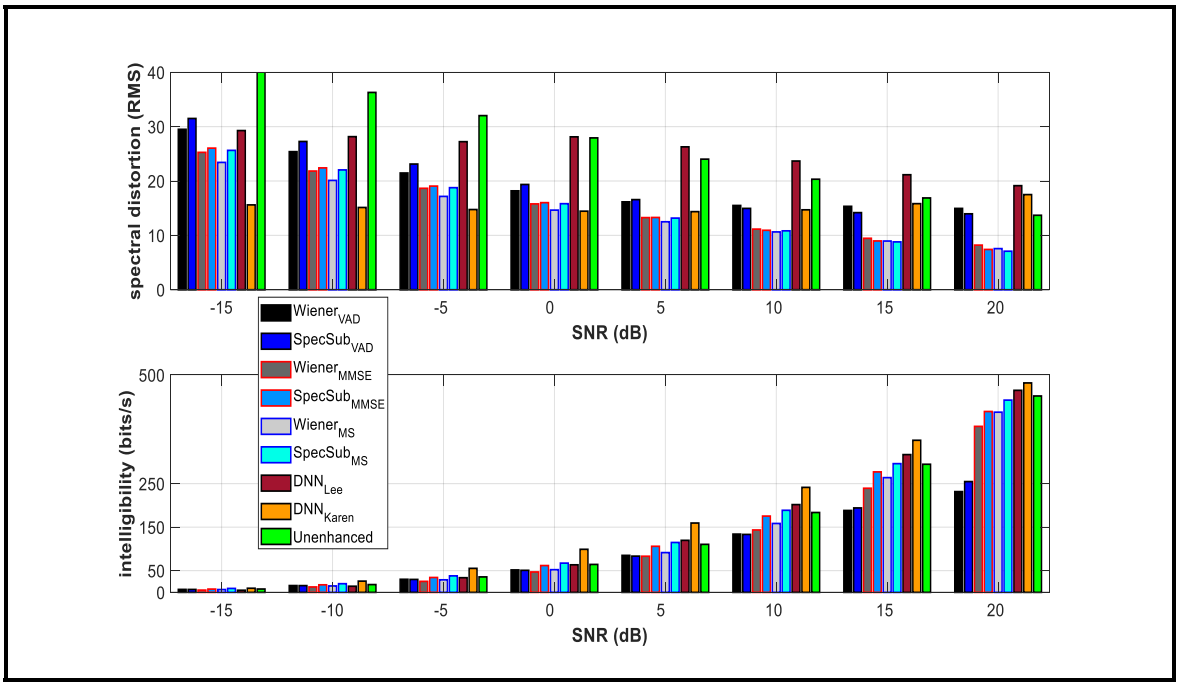

Figure 5.1 *Results from objective evaluation of the SE methods. The additive noise is white Gaussian. In the figure, the plot at the top illustrates spectral distortion and the bottom plot presents instrumental intelligibility.*

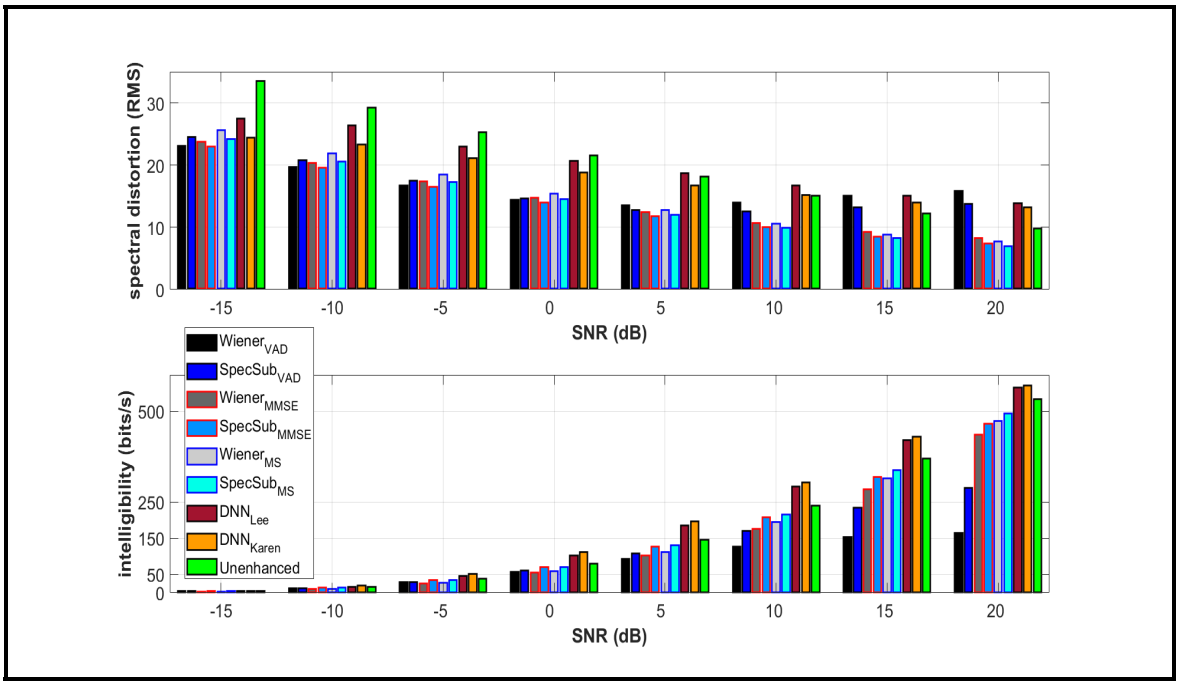

Figure 5.2 *Results from objective evaluation of the SE methods. The additive noise is multi-talker babble. In the figure, the plot at the top illustrates spectral distortion and the bottom plot presents instrumental intelligibility.*

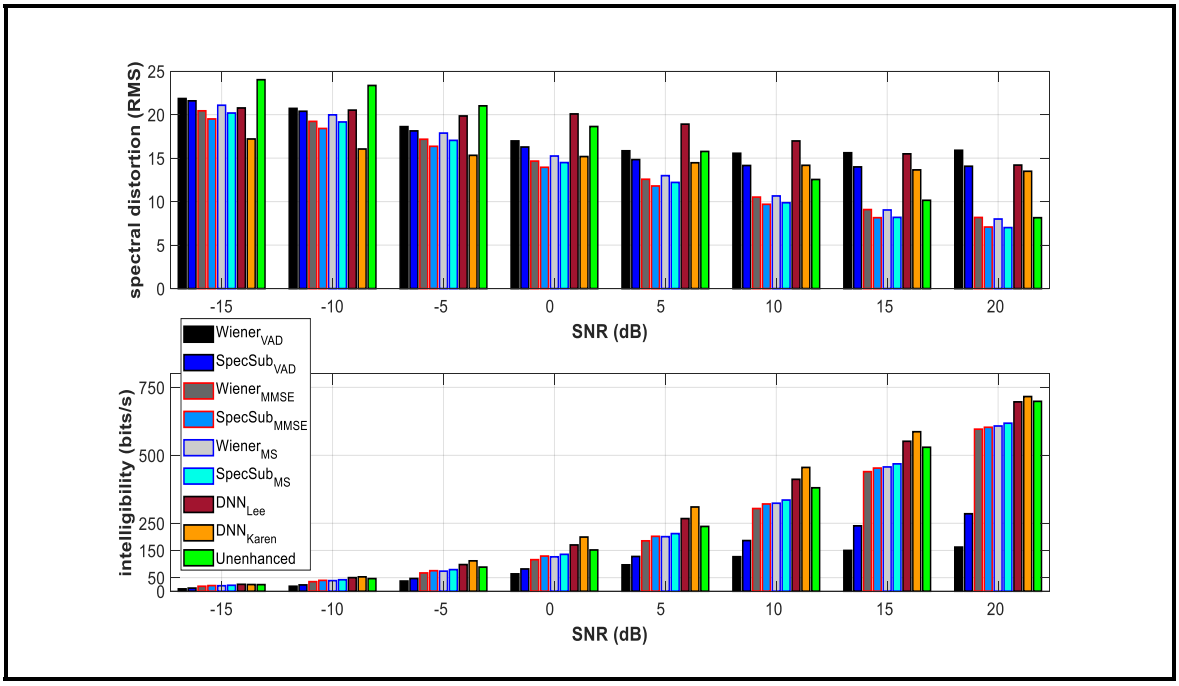

Figure 5.3 *Results from objective evaluation of the SE methods. The additive noise is sampled randomly from child-centered daylong recordings. In the figure, the plot at the top illustrates spectral distortion and the bottom plot presents instrumental intelligibility.*

5.2 Results for downstream task evaluation of SE methods

The results obtained from the performance of the downstream task with the application of SE front-ends are presented in the Table 5.1.

Table 5.1 *The table presents downstream task performance with the application of SE front-end. For IDS/ADS classification, higher value of unweighted average recall in comparison to baseline results predicts better performance and for automatic syllabification, lower value of means absolute relative error with respect to baseline result translates into better performance. The results which have better performance than baseline are indicated with green colour.*

	IDS/ADS Classification (Unweighted Average Recall %)		Automatic Syllabification (Mean Abs. Relative Error %)	
SE Methods	**Non-collared**	**Collared**	**Non-collared**	**Collared**
$\text{Baseline}_{\text{Unenhanced}}$	66.79		47.94	
DNN_{Sun}	67.98	67.76	45.79	46.41
$\text{DNN}_{\text{Keren}}$	65.75	65.03	48.92	47.96
$\text{SpecSub}_{\text{MMSE}}$	67.88	67.68	53.06	46.13
$\text{Wiener}_{\text{MMSE}}$	66.41	67.18	47.79	46.79
$\text{SpecSub}_{\text{MS}}$	66.94	67.26	46.69	46.24
$\text{Wiener}_{\text{MS}}$	66.36	66.16	53.22	46.61
$\text{SpecSub}_{\text{VAD}}$	66.16	66.63	48.06	48.12
$\text{Wiener}_{\text{VAD}}$	66.39	66.56	48.48	49.09

5.3 Comparison of objective evaluation metrics and downstream task performance

In this section, results from objective analyses and downstream task performance are compared statistically. The purpose of this comparison is to evaluate whether a relationship exists between the two sets of results. Statatistical comparison is performed using linear correlation.

To visualize the data to be compared, the downstream task performance values and the objective metrics values are illustrated through scatter plots. In the scatter plots, a best fit line is also plotted using least-squares fitting. In all the scatter plots, left column has downstream task performance over collared data set and in the right column over the non-collared dataset. On x-axis, downstream task performance value is depicted and on the y-axis objective metric value. Each row in a scatter plot depicts objective metrics calculated with respect to clean speech and enhanced speech obtained by denoising additive noise degraded speech. In the top row degrading noise is additive noise sampled randomly from child-centered recordings, in the middle row the additive noise is white Gaussian noise, and in the bottom row the additive noise is babble noise. The reporting of objective

metrics in plots is done for additive noises added at SNR = 0 dB as [11] has reported the average speech SNR of comparable child-centered daylong recordings to be approximately 0 dB. Each plot also depicts the linear correlation (denoted as r) between the plotted variables and p-value.

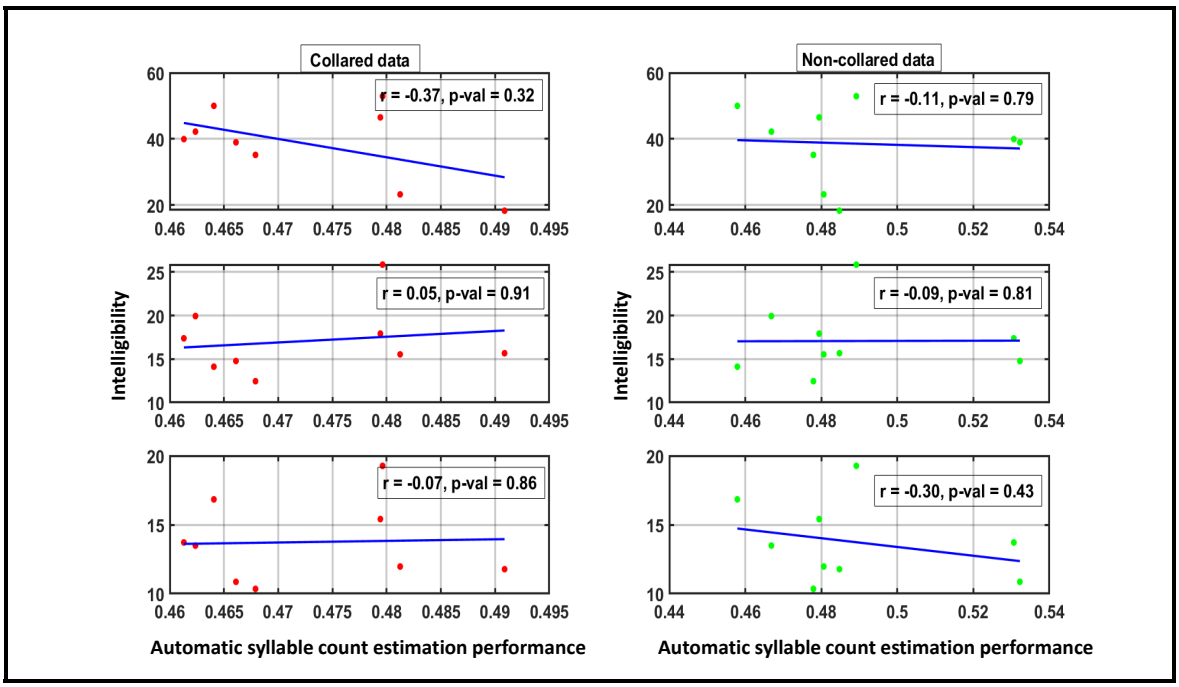

Figure 5.4 *The figure illustrates the objective metric values and downstream task performance values through a scatter plot. A best-fit line is also plotted using least-squares fitting. X-axis is automatic syllable count estimation (ASCE) performance measured through mean absolute relative error and Y-axis is instrumental intelligibility estimated through SIIB (bits/s).*

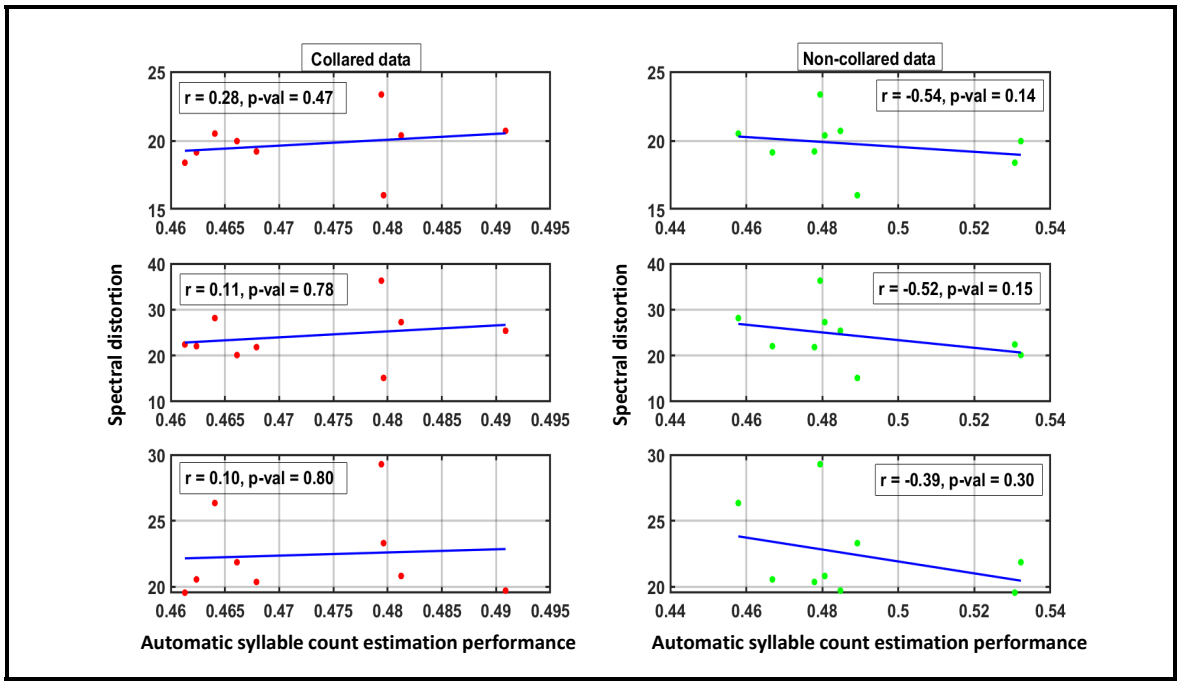

Figure 5.5 *The figure illustrates the objective metric values and downstream task performance values through a scatter plot. A best-fit line is also plotted using least-squares fitting. X-axis is automatic syllable count estimation (ASCE) performance measured through mean absolute relative error and Y-axis is spectral distortion (calculated as RMS).*

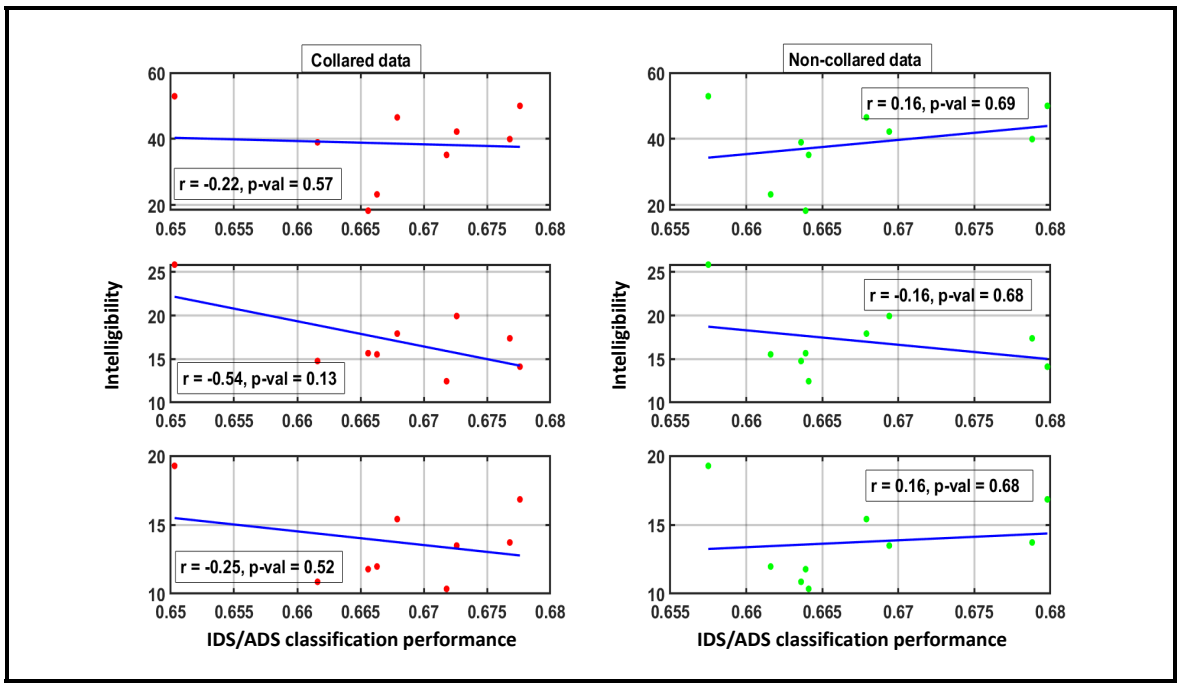

Figure 5.6 *The figure illustrates the objective metric values and downstream task performance values through a scatter plot. A best-fit line is also plotted using least-squares fitting. X-axis is IDS/ADS classification performance measured through unweighted average recall and Y-axis is instrumental intelligibility estimated through SIIB (bits/s).*

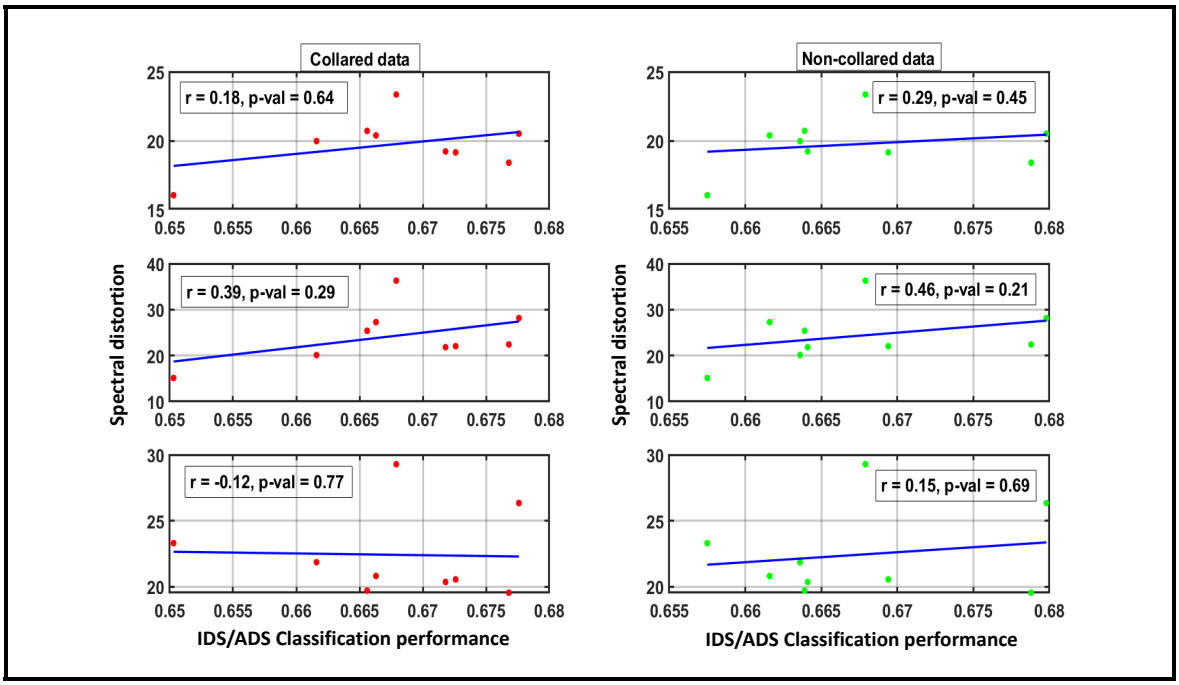

Figure 5.7 *The figure illustrates the objective metric values and downstream task performance values through a scatter plot. A best-fit line is also plotted using least-squares fitting. X-axis is IDS/ADS classification performance measured through unweighted average recall and Y-axis is spectral distortion (calculated as RMS).*

5.4 Discussion of the results

The results of objective assessment indicate that all SE methods reduce spectral distortion at SNR levels $\leq$ 5 dB in comparison to baseline unenhanced case in all 3 additive noise conditions. The exception to this observation is SE method DNN_{Sun} which produces additional distortion (in comparison to baseline) for $\geq$ 0 dB SNRs. Interestingly, the other DNN-based method $\text{DNN}_{\text{Keren}}$ also produces additional distortion in comparison to baseline at SNR levels $\geq$ 10 dB in case of additive noise sampled from daylong recordings and babble noise and at SNR = 20 dB in case of additive white Gaussian noise. In comparison with DNN-based approaches, classical signal processing SE methods except VAD based variants reduce spectral distortion at all SNR levels. VAD based variants produce additional distortion in comparison to baseline for SNR $\geq$ 15 dB in case of additive noise sampled from daylong recordings and babble noise and at SNR = 20 dB in case of white Gaussian noise. The results from instrumental intelligibility assessment show that only DNN-based SE methods DNN_{Sun} and $\text{DNN}_{\text{Keren}}$ improve the instrumental intelligibility at all SNR levels and in all the additive noise conditions in comparison to unenhanced baseline results as well as results obtained from enhancement through DSP-based methods. The superior performance of DNN-based methods on instrumental intelligibility assessment indicates that these methods generalize well to signal conditions that were not exposed to them during model training phase in the original studies [39], [50]. In contrast to DNN-based methods, the DSP-based SE methods reduce instrumental intelligibility of the enhanced

signals with respect to baseline at all SNRs. The degradation of intelligibility is known characteristic of classical DSP-based methods and is confirmed in literature ([96], chapter 13 of the textbook [31]). From the objective results, it emerges that SAD based variants of DSP-based SE methods demonstrate poor performance in comparison to baseline results as well as other SE methods at all SNR levels and noise conditions. Their performance is particularly degraded at low speech SNRs and this phenomenon is more pronounced in the naturalistic noise conditions. A possible explanation could be that underlying speech activity detector's (TO-Combo-SAD) performance degrades at low speech SNRs, resulting in faulty noise estimation (see [11] for SAD performance analysis on comparable dataset).

The results obtained on the downstream performance analysis (see Table 5.1) are much varied for both the downstream tasks. The SAD based variants of spectral subtraction and Wiener filtering methods gave all the results below baseline performance. But, a closer inspection reveals that in comparison with some DSP (e.g. compare $\text{Wiener}_{\text{VAD}}$ and $\text{Wiener}_{\text{MS}}$ for IDS/ADS classification) and DNN –based (e.g. compare $\text{Wiener}_{\text{VAD}}$ and $\text{DNN}_{\text{Keren}}$ for IDS/ADS classification) they gave nearly matching downstream task performance or even out-perform some of the other SE methods. This observation is in divergence with the objective assessment results for SAD based variants ($\text{SpecSub}_{\text{VAD}}$ and $\text{Wiener}_{\text{VAD}}$), where their performance, was particularly poor in comparison to all the other SE methods. Out of the two DNN-based SE methods, only DNN_{Sun} was able to beat baseline performance. From the DSP- based methods, $\text{SpecSub}_{\text{MMSE}}$, $\text{Wiener}_{\text{MMSE}}$, $\text{SpecSub}_{\text{MS}}$ performed the best including better performance than baseline results. With regard to the impact of collars, the best results for both the downstream tasks were obtained by DNN_{Sun} with non-collared data. But, the best performing SE methods among DSP-based approaches had better results with the collared data.

Both the downstream tasks, on application of DNN_{Sun} SE front-end, improved performance irrespective of collared or non-collared dataset. For the task of IDS/ADS classification, the UAR was improved 1.2 percentage points over the baseline result and for automatic syllable count estimation, mean absolute relative error was reduced by 2.15 percentage points with respect to baseline results.

The comparisons of objective metrics and downstream task performance do not reveal any correlation between the two measurements (p-value > 0.05 in all comparisons, Pearson's method for calculating linear correlation). As an illustration, consider $\text{DNN}_{\text{Keren}}$'s performance on the objective assessment and downstream task evaluations. Its application leads to the largest instrumental intelligibility gains in the objective analysis but it under-performs in the downstream task evaluations in our setup. The observations from the results of statistical comparisons indicate that task independent objective metrics (obtained from clean speech corrupted with additive noise) do not predict downstream task performance (on child-centered naturalistic audio).

6 Conclusions

The main aim of the present work was evaluation of a SE front-end for downstream speech processing analysis in real-world noisy child-centered daylong recordings. This overall aim was divided into three distinct objectives 1) Objective evaluation of SE methods in additive noise sampled from daylong recordings, 2) Application of SE front-end to downstream speech processing tasks to investigate efficacy of SE front-end on performance, and 3) Compare results of objective evaluation and downstream task performance, in order to assess can objective measurement predict downstream task performance. The downstream tasks evaluated in this work were Infant-directed speech and adult-directed speech (IDS/ADS) classification and automatic syllable count estimation. Both of these tasks are important in child language development analysis.

The results from conducted experiments indicate that a recently proposed LSTM-based architecture by Sun et al. [50] provided the best SE front-end among the compared methods with respect to improvement on baseline (unenhanced) results on downstream task performance. However, the results from the experiments do not show a substantial gain in downstream task performance on application of a SE front-end in the present setup. The classical signal processing based methods spectral subtraction and Wiener filtering, when using noise-estimation through MS or MMSE with additional temporal noise-context, also provided competitive downstream task performance. The comparison of downstream task results with the objective metrics of spectral distortion and instrumental intelligibility obtained through evaluation of SE methods in additive noise conditions do not indicate any predictive relationship between objective measurement and potential downstream task performance. However, the present setup compared the results for only two downstream tasks, whether the pattern holds for other downstream tasks is unknown.

The results from this study indicate that SE front-ends have not improved performance of downstream tasks significantly. So, the basic issue of improving downstream task performance persists. As part of this study we have compared the classical signal processing based SE methods and modern machine learning based SE methods. For classical DSP-based methods, it is a known characteristic that their performance depends on accuracy of noise-estimation techniques. It is probable that compared noise-estimation techniques were not able to capture the statistics of the underlying noise. In such a case, future efforts could be directed towards developing (or evaluating) advanced techniques for noise-estimation. The chosen machine learning based SE methods have demonstrated considerable gains in downstream tasks which were tested as part of their original studies. In this study, they did not deliver high performance gains. Another avenue of future effort could be evaluating the two machine learning based methods after optimising them with the domain data i.e. child-centered daylong audio data.

References

[1] M. A. Redford and M. E. Beckman, *The Handbook of Speech Production*, 1st ed. Wiley Blackwell, 2015.

[2] L. Raphael, G. Borden, and K. Harris, *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*, 6th ed. Lippincott Williams & Wilkins, 2012.

[3] T. Bäckström and O. Räsänen, *Introduction to speech processing*, `https://wiki.aalto.fi/display/ITSP/Introduction+to+Speech+Processing`, Accessed: 2019-09-15.

[4] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed. IEEE Press, 1999.

[5] W. J. M. Levelt, *Speaking: From Intention to Articulation*. MIT Press, 1989.

[6] P. K. Kuhl, "Early language acquisition: cracking the speech code", *Nature Reviews.Neuroscience*, vol. 5, no. 11, pp. 831–43, 2004.

[7] L. Krogh, H. A. Vlach, and S. P. Johnson, "Statistical learning across development: Flexible yet constrained", *Frontiers in Psychology*, vol. 3, 2013. DOI: `10.3389/fpsyg.2012.00598`.

[8] E. Bergelson and R. N. Aslin, "Nature and origins of the lexicon in 6-mo-olds", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 49, pp. 12 916–12 921, 2017. DOI: `10.1073/pnas.1712966114`.

[9] O. Räsänen, "Context induced merging of synonymous word models in computational modeling of early language acquisition", in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5037–5040. DOI: `10.1109/ICASSP.2012.6289052`.

[10] E. Bergelson, M. Casillas, M. Soderstrom, A. Seidl, A. S. Warlaumont, and A. Amatuni, "What Do North American Babies Hear? A large-scale cross-corpus analysis", *Developmental Science*, vol. 22, no. 1, 2019. DOI: `10.1111/desc.12724`.

[11] O. Räsänen *et al.*, "Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech", *Speech Communication*, vol. 113, pp. 63–80, 2019. DOI: `10.1016/j.specom.2019.08.005`.

[12] B. Hart and T. Risley, *Meaningful differences in the everyday experience of young American children*. Brookes Publishing Company, Inc, 1995.

[13] M. L. Rowe, "A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development", *Child Development*, vol. 83, no. 5, pp. 1762–1774, 2012. DOI: 10.1111/j.1467-8624.2012.01805.x.

[14] E. V. M. Lieven, "Crosslinguistic and crosscultural aspects of language addressed to children", in *Input and Interaction in Language Acquisition*. Cambridge University Press, 1994, pp. 56–73. DOI: 10.1017/CBO9780511620690.005.

[15] L. A. Shneidman and S. Goldin-Meadow, "Language input and acquisition in a mayan village: How important is directed speech?", *Developmental Science*, vol. 15, no. 5, pp. 659–673, 2012. DOI: 10.1111/j.1467-7687.2012.01168.x.

[16] A. Cristia, E. Dupoux, M. Gurven, and J. Stieglitz, "Child-directed speech is infrequent in a forager-farmer population: A time allocation study.", *Child Development*, vol. 90, no. 3, pp. 759–773, 2017. DOI: 10.1111/cdev.12974.

[17] C. S. Tamis-LeMonda, Y. Kuchirko, R. Luo, K. Escobar, and M. H. Bornstein, "Power in methods: Language to infants in structured and naturalistic contexts", *Developmental Science*, vol. 20, no. 6, 2017. DOI: 10.1111/desc.12456.

[18] A. Weisleder and A. Fernald, "Talking to children matters: Early language experience strengthens processing and builds vocabulary", *Psychological Science*, vol. 24, no. 11, pp. 2143–2152, 2013. DOI: 10.1177/0956797613488145.

[19] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?", *Behavioral and Brain Sciences*, vol. 33, no. 2–3, pp. 61–83, 2010. DOI: 10.1017/S0140525X0999152X.

[20] E. Bergelson, A. Amatuni, S. Dailey, S. Koorathota, and S. Tor, "Day by day, hour by hour: Naturalistic language input to infants", *Developmental Science*, vol. 22, no. 1, 2019. DOI: 10.1111/desc.12715.

[21] M. Soderstrom *et al.*, *Analyzing the child language experiences around the world project*, https://sites.google.com/view/aclewdid/home, Accessed: 2019-08-28.

[22] A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Effective word count estimation for long duration daily naturalistic audio recordings", *Speech Communication*, vol. 84, pp. 15–23, 2016. DOI: 10.1016/j.specom.2016.07.007.

[23] L. Sun *et al.*, "Speaker diarization with enhancing speech for the first DIHARD challenge", in *Proc. Interspeech-2018*, 2018, pp. 2793–2797. DOI: 10.21437/Interspeech.2018-1742.

[24] F. Xiong *et al.*, "Front-end technologies for robust ASR in reverberant environments–spectral enhancement-based dereverberation and auditory modulation filterbank features", *EURASIP Journal on Advances in Signal Processing*, 2015. DOI: 10.1186/s13634-015-0256-4.

[25] B. W. Schuller *et al.*, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring", in *Proc. Interspeech-2017*, 2017, pp. 3442–3446. DOI: 10.21437/Interspeech.2017-43.

[26] B. W. Schuller *et al.*, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity", in *Proc. Interspeech 2019*, 2019, pp. 2378–2382. DOI: 10.21437/Interspeech.2019-1122.

[27] In, *Neuroscience*, D. Purves, Ed., 3rd ed. Sinauer Associates, 2004, ch. Language and Speech.

[28] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex", *Neuron*, vol. 98, no. 5, pp. 1042–1054, 2018. DOI: 10.1016/j.neuron.2018.04.031.

[29] Wikipedia contributors, *Speech perception — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Speech_perception&oldid=918237958, Accessed: 2019-11-15, 2019.

[30] LiteraryDevices Editors, *Phoneme*, https://literarydevices.net/phoneme/, [Online; accessed 01-December-2019].

[31] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC Press, 2013. DOI: 10.1201/b14529.

[32] T. V. Ramabadran, J. P. Ashley, and M. J. McLaughlin, "Background noise suppression for speech enhancement and coding", in *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings. Back to Basics: Attacking Fundamental Problems in Speech Coding*, 1997, pp. 43–44. DOI: 10.1109/SCFT.1997.623887.

[33] R. Niederjohn and R. A. Curtis, "The development of a computer speech processing system and its use for the study and development of processing methods for enhancing the intelligibility of speech in noise", Rome Air Development Center, 1977.

[34] M. S. Kavalekalam, J. K. Nielsen, J. B. Boldt, and M. G. Christensen, "Model-based speech enhancement for intelligibility improvement in binaural hearing aids", *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 27, no. 1, pp. 99–113, 2019. DOI: 10.1109/TASLP.2018.2872128.

[35] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979. DOI: 10.1109/PROC.1979.11540.

[36] J. Benesty, *Noise Reduction in Speech Processing: Springer topics in signal procesing 2*. Springer Verlag, 2009.

[37] Wikipedia contributors, *Lombard effect — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Lombard_effect&oldid=917701738, Accessed: 2019-11-20, 2019.

[38] A. H. Moore, P. Peso Parada, and P. A. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures", *Computer Speech & Language*, vol. 46, pp. 574–584, 2017. DOI: 10.1016/j.csl.2016.11.003.

[39] G. Keren, J. Han, and B. Schuller, "Scaling speech enhancement in unseen environments with noise embeddings", in *Proc. CHiME 2018 Workshop on Speech Processing in Everyday Environments*, 2018, pp. 25–29. DOI: 10.21437/CHiME.2018-6.

[40] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979. DOI: 10.1109/TASSP.1979.1163209.

[41] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. MIT Press, 1950.

[42] S. M. Kuo and H. Zhao, "Adaptive acoustic echo cancellation algorithms in teleconferencing systems", *The Journal of the Acoustical Society of America*, 1989. DOI: https://doi.org/10.1121/1.2027564.

[43] J. Franzen and T. Fingscheidt, "An efficient residual echo suppression for multi-channel acoustic echo cancellation based on the frequency-domain adaptive kalman filter", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 226–230. DOI: 10.1109/ICASSP.2018.8462488.

[44] D. Giacobello and T. L. Jensen, "Speech Dereverberation Based on Convex Optimization Algorithms for Group Sparse Linear Prediction", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 446–450. DOI: 10.1109/ICASSP.2018.8462560.

[45] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation", *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 903–907, 2019. DOI: 10.1109/LSP.2019.2911179.

[46] Wikipedia contributors, *Additive white gaussian noise — Wikipedia, the free encyclopedia*, https://en.wikipedia.org/w/index.php?title=Additive_white_Gaussian_noise&oldid=923685673, [Online; accessed 23-November-2019], 2019.

[47] G. Parikh and P. C. Loizou, "The influence of noise on vowel and consonant cues", *Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3874–3888, 2005.

[48] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001. DOI: 10.1109/89.928915.

[49] M. Kolbæk, "Single-Microphone Speech Enhancement and Separation Using Deep Learning", PhD thesis, Aalborg University, Denmark, 2018.

[50] L. Sun *et al.*, "A Novel LSTM-Based Speech Preprocessor for Speaker Diarization in Realistic Mismatch Conditions", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. DOI: 10.1109/ICASSP.2018.8462311.

[51] S. Dimolitsas, "Objective speech distortion measures and their relevance to speech quality assessments", *IEE Proceedings I - Communications, Speech and Vision*, vol. 136, no. 5, pp. 317–324, 1989. DOI: 10.1049/ip-i-2.1989.0045.

[52] V. Grancharov and W. B. Kleijn, in *Springer Handbook of Speech Processing*. Springer,Berlin,Heidelberg, 2008, ch. Speech Quality Assessment, pp. 83–100. DOI: 10.1007/978-3-540-49127-9_5.

[53] P. Mermelstein, "Evaluation of a Segmental SNR Measure as an Indicator of the Quality of ADPCM Coded Speech", *Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1664–1667, 1979. DOI: 10.1121/1.383638.

[54] F. Itakura and S. Saito, "Analysis synthesis telephony based upon the maximum likelihood method", in *6th International Congress on Acoustics, Tokyo, Japan*, 1968.

[55] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory", *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018. DOI: 10.1109/LSP.2017.2774250.

[56] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Evaluation of Intrusive Instrumental Intelligibility Metrics", *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 11, pp. 2153–2166, 2018. DOI: 10.1109/TASLP.2018.2856374.

[57] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step", in *1982 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1982. DOI: 10.1109/ICASSP.1982.1171512.

[58] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders", *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992. DOI: 10.1109/49.138987.

[59] J. H. James, Bing Chen, and L. Garrison, "Implementing VoIP: a voice transmission performance progress report", *IEEE Communications Magazine*, vol. 42, no. 7, pp. 36–41, 2004. DOI: 10.1109/MCOM.2004.1316528.

[60] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, https://catalog.ldc.upenn.edu/LDC93S1, Accessed: 2019-03-01, 1993.

[61] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation", in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP)*, 1996. DOI: 10.1109/ICASSP.1996.543199.

[62] A. Ziaei, L. Kaushik, A. Sangwan, J. H. L. Hansen, and D. W. Oard, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions", in *Proc. Interspeech-2014*, 2014, pp. 1544–1548.

[63] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux", *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013. DOI: 10.1109/LSP.2013.2237903.

[64] M. Brookes, *Voicebox: Speech processing toolbox for matlab*, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, Accessed: 2019-08-29.

[65] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity", in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4266–4269. DOI: 10.1109/ICASSP.2010.5495680.

[66] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012. DOI: 10.1109/TASL.2011.2180896.

[67] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[69] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement", in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136–140. DOI: 10.1109/HSCMA.2017.7895577.

[70] T. Gao, J. Du, L. Dai, and C. Lee, "SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement", in *Proc. Interspeech-2016*, 2016, pp. 3713–3717. DOI: 10.21437/Interspeech.2016-224.

[71] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus", in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT '91, 1992, pp. 357–362. DOI: 10.3115/1075527.1075614.

[72] S. Lin, Y. Zhang, Y. Liu, H. Liu, and Q. Liu, "An introduction to corpora resources of 863 program for chinese language processing and human-machine interaction", in *In Proceedings of ALR-04 affiliated to IJCNLP 2004*, 2004.

[73] T. Gao, J. Du, L. Dai, and C. Lee, "Densely Connected Progressive Learning for LSTM-Based Speech Enhancement", in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5054–5058. DOI: 10.1109/ICASSP.2018.8461861.

[74] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 367–376, 1980. DOI: 10.1109/TASSP.1980.1163421.

[75] Wikipedia, *Data processing inequality–Wikipedia, the free encyclopedia*, http://en.wikipedia.org/w/index.php?title=Data%20processing%20inequality&oldid=919002407, Accessed: 2019-09-10.

[76] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information", *Physical Review E*, vol. 69, no. 6, 2004. DOI: 10.1103/PhysRevE.69.066138.

[77] M. Soderstrom, "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants", *Developmental Review*, vol. 27, no. 4, pp. 501–532, 2007. DOI: 10.1016/j.dr.2007.06.002.

[78] A. Batliner, B. Schuller, S. Schaeffler, and S. Steidl, "Mothers, adults, children, pets –towards the acoustics of intimacy", in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4497–4500. DOI: 10.1109/ICASSP.2008.4518655.

[79] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152. DOI: 10.1145/130385.130401.

[80] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism", in *Proc. Interspeech-2013*, 2013, pp. 148–152.

[81] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor", in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 835–838. DOI: 10.1145/2502081.2502224.

[82] D. Wang and S. S. Narayanan, "Robust Speech Rate Estimation for Spontaneous Speech", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007. DOI: 10.1109/TASL.2007.905178.

[83] A. Ziaei, A. Sangwan, L. Kaushik, and J. H. L. Hansen, "Prof-Life-Log: Analysis and classification of activities in daily audio streams", in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4719–4723. DOI: 10.1109/ICASSP.2015.7178866.

[84] S. Seshadri and O. Räsänen, "SylNet: An adaptable end-to-end syllable count estimator for speech", *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1359–1363, 2019.

[85] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio", in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[86] E. Bergelson, *Bergelson Seedlings HomeBank Corpus*, Accessed: 2019-08-28, 2017. DOI: 10.21415/T5PK6D.

[87] C. F. Rowland, S. Durrant, M. Peter, A. Bidgood, J. Pine, and L. S. Jago, *The Language 0-5 Project*, Accessed: 2019-08-28, 2019. DOI: 10.17605/OSF.IO/KAU5F.

[88] M. Casillas, P. Brown, and S. C. Levinson, *Casillas HomeBank Corpus*, Accessed: 2019-08-28, 2017. DOI: 10.21415/T51X12.

[89] K. McDivitt and M. Soderstrom, *McDivitt HomeBank Corpus*, Accessed: 2019-08-28, 2016. DOI: 10.21415/T5KK6G.

[90] M. Soderstrom, *Soderstrom HomeBank Corpus*, Accessed: 2019-08-28, 2017. DOI: 10.21415/T5KK6G.

[91] A. S. Warlaumont, G. M. Pretzer, S. Mendoza, and E. Walle, *Warlaumont HomeBank Corpus*, Accessed: 2019-08-28, 2016. DOI: 10.21415/T54S3C.

[92] C. Rosemberg, F. Alam, A. Stein, M. Migdalek, A. Menti, and G. Ojea, *Language Environments of Young Argentinean children*, Accessed: 2019-08-28, 2015.

[93] M. Casillas *et al.*, *Introduction: The ACLEW DAS template*, https://osf.io/aknjv, 2019.

[94] M. Casillas *et al.*, *DARCLE Annotation Scheme*, https://osf.io/4532e, 2018.

[95] D. Kahn, "Syllable-based generalizations in english phonology.", PhD thesis, Massachusetts Institute of Technology. Dept. of Foreign Literatures and Linguistics, 1976.

[96] P. C. Loizou and G. Kim, "Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2011. DOI: 10.1109/TASL.2010.2045180.