ABSTRACT

Title of Thesis:	DEEP NEURAL NETWORK FOR DEREVERBERATION
	Yang Jiao, Master of Science, 2019
Thesis directed by:	Professor Ramani Duraiswami Electrical and Computer Engineering

Recently, deep neural networks have achieved incredible success in the area of computer vision and natural language processing. The research topics under the umbrella of speech enhancement have embraced this chance to revolutionize. Dereverberation as one such topics has gained less popularity compared to other tasks of speech enhancement such as cocktail party problem, speech enhancement in open area. Our aim is to extend method of deep learning into the domain of dereverberation.

We leverage a successful neural network method on a similar task of speech enhancement to dereverberation, specifically, time frequency mask supported GEV beamformer. This data driven approach introduces feature transferrable to related tasks compared to hand-engineered methods. Our experiments illustrate that the original framework arise from open area enhancement tasks is proved to be effective in our closure space tasks.

DEEP NEURAL NETWORK FOR DEREVERBERATION

by

Yang Jiao

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2019

Advisory Committee:
Professor Ramani Duraiswami, Chair
Professor Gang Qu
Professor Rama Chellappa

Table of Contents

Chapter 1: Introduction

1.1 Introduction

The technique of dereverberation is controversial [1]: on one hand, human ears are capable of making use of reverberation, perceiving space and direction from echoed copies of direct sound waves; on the other hand, reverberation has made it difficult for sound to be perceived by computers and machines, hurting the performance of carefully designed speech recognition algorithms and causing trouble for hearing-aid devices to be of help to hearing loss people. In a way, the value of dereverberation should be judged based on the specific application in concern, and in areas where it is suitable, it has proved to be an indispensable part of modern automatic speech recognition system, hearing-aid devices and so on. Before the advent of deep learning, state-of-the-art dereverberation algorithms are, essentially hand-designed by researchers. The model-based statistical algorithms require much effort and domain knowledge to start with. On the contrary, deep learning takes a data-driven approach, assumes the least about underlying model and learns a model directly from data. The availability of data and computational power has made this paradigm come to live. In a wide range of areas, this paradigm has achieved success and generated many state-of-art algorithms [2].

Recently, deep neural network has shown popularity in the wide area of speech enhancement, demonstrating better performance than their model-based counterparts. This has led to the invention of deep learning based beamforming [4], denoising [5], speech separation [3] and speech dereverberation [6, 7, 8]. These algorithms are not based directly on a stochastic model but learned from a set of data that is large enough to train a neural network. One such work that has made an import impact in this area is time frequency mask estimation using deep neural networks [9, 10, 11]. Important applications of neural network based time frequency mask are speech denoising and speech separation. This thesis is another important application of neural network based time frequency mask, i.e. speech dereverberation. This application has also been studied by a group of researchers, with a different focus as part of a bigger system of real-time distant automatic speech recognition system [8]. This thesis treats this as an isolated dereverberation module and evaluates it using a different metric from WER. We now describe the structure of this thesis.

1.2 Structure of This Thesis

Chapter 2 presents earlier works of time frequency mask and important variants along this line, ending with an introduction of recent development of them using neural networks. Chapter 3 goes over related background of dereverberation by presenting important concepts in reverberation and room acoustics. This is followed by an overview of three major categories of dereverberation algorithms and their limitations. In the end, recent neural network based approaches are introduced. Chapter 4 describes the system we considered in depth, with details about the dataset, reverberation simulation configuration, major components of the system, network architecture, training scheme, and evaluation metrics. Finally, Chapter 5 gives our concluding remarks and future works.

Chapter 2: Time Frequency Mask

2.1 Overview

2.1.0 Binary Masks

Time frequency masks work in the spectral domain. A waveform signal is windowed and taken spectral transform to be represented as time and frequency bins. Its use can be traced back to a 1983 article where a binary TF mask is used for sound separation [12, 13]. However, a time frequency mask doesn't have to be binary.

As an introductory example, a binary mask is shown in Figure 2.1 B. In this example, a mixture of speech is taken spectrogram [14] as shown Figure 2.1 A. A binary mask $BM(t, f)$ is to map every time frequency bin to a binary value of 1 or 0. A value of 1 indicates that this bin is dominated by the energy of target speech signal $s(t, f)$ and the value of 0 indicates that this bin is mostly energy of unwanted source of signal $n(t, f)$, such as interferer's speech or noise, and should be removed.

$$BM(t, f) = \begin{cases} 1 \; if \; s(t, f) > n(t, f) \\ 0 \; otherwise \end{cases}$$

To apply this time frequency mask, a Hadamard product, i.e. entry-wise product is applied between the spectral representation of the speech mixture and the binary time frequency mask. The unwanted component in the mixture is removed with the multiplication by 0. After the product, a synthesis step is applied to the spectral domain, so that the signal waveform is reconstructed.

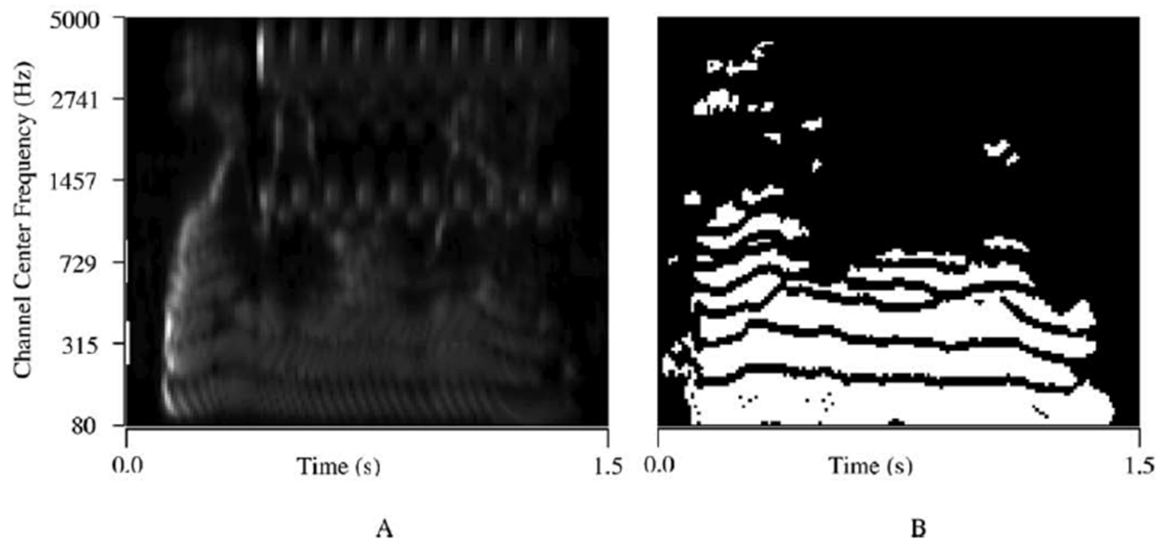

Figure 2.1 An example of spectrogram in A and binary mask of speech in B. Taken from [14]

2.1.1 Variants of Masks

The idea of time frequency mask has been intensively studied for decades. There are several important variants from the basic binary mask. To start with, an ideal binary mask is proposed as a major goal of CASA [12, 15, 16, 17] and has from then on being used as an important criterion for measuring the performance of a CASA system. In each time frequency unit of an ideal binary mask $IBM(t, f)$, a value of 1 is only assigned if and only if the target energy $s(t, f)$ is larger than interference energy $n(t, f)$ by a predetermined threshold in dB. The intuition is to guarantee that only bins with a local SNR greater than this threshold will be kept. A typical choice of this threshold is the $LC = 0\ dB$ criterion.

$$IBM(t, f) = \begin{cases} 1\ if\ s(t, f) > n(t, f) + LC \\ 0\ otherwise \end{cases}$$

The ideal binary mask is supported by a fundamental feature of auditory perception, i.e. the auditory masking phenomenon [12, 18]. This feature states that a louder source of sound renders a weaker one inaudible within a critical band. In the ideal binary mask, this corresponds to removing the weaker one from the spectrum when its energy is smaller than the louder one by a margin of the predetermined threshold. The binary mask is also supported by evidence from a fundamental feature of speech signal, which should be differentiated from the masking phenomenon. This feature states that speech signal has a sparse energy distribution [12, 19]. More specifically, the overlap between different speech signals is small given that the spectral transform has a high resolution. In the binary mask, this corresponds to set a time frequency bin's value to 1, assuming this bin is approximately exclusively occupied by one of the sound sources. This assumption is referred to as orthogonality of speech sources. Another important variant of binary mask is ideal ratio mask [20]. The ideal ratio mask $IRM(t, f)$ is closely related to the Wiener gain [12, 20]. The Wiener filter can be included into the general framework of time frequency mask, if each TF unit of the mask represents the ratio of target energy $s(t, f)$ to mixture energy $z(t, f) = s(t, f) + n(t, f)$ within the unit.

$$WienerGain(t, f) = IRM(t, f) = \frac{s(t, f)}{s(t, f) + n(t, f)}$$

With the advent of deep neural network based time frequency mask estimation, a plethora of variants has emerged, including phase sensitive TF mask [21] and complex ideal ratio mask [7, 11]. They are aimed to remove the limitation of ideal ratio mask. Specifically, the ideal ratio mask only changed the magnitude of the spectrum, while the phase remains the same. In contrast, binary mask doesn't need to take care of phase since the 0 value totally removes the inference and 1 value keeps all energy in that bin. To tackle the phase problem of ideal ratio mask, a group of researchers have proposed to estimate a phase mask $PSM(t,f)$ jointly with the magnitude mask as in the phase sensitive TF mask work [21]. In the $PSM(t,f)$, the phase difference $\theta_{t,f}^{XY}$ of speech signal $X(t,f)$ and mixture signal $Y(t,f)$ is preserved in the term $cos\theta_{t,f}^{XY}$.

$$PSM(t,f) = \sqrt{\frac{s(t,f)}{s(t,f)+n(t,f)}} cos\theta_{t,f}^{XY}$$

$$s(t,f) = X^H X$$

$$n(t,f) = Y^H Y$$

And they proposed to estimate a mask in the complex domain as in the complex ideal ratio mask $cIRM(t,f)$ work [7, 11], so that phase and magnitude are considered implicitly as a complex number.

$$cIRM(t,f) = \frac{X(t,f)}{Y(t,f)}$$

2.1.2 Estimation of Time Frequency Masks

While the estimation of time frequency mask can be developed monaurally, i.e. when only single microphone is considered [14], it is out of the scope of this thesis. Instead, binaural and microphone array based estimation will be briefly introduced in this section. To start with, the estimation of time frequency mask has been researched for decades. In early works, the proposed approaches use similar frameworks [12, 22, 23], where features are extracted in each time frequency bin and then clustered in the feature space. As an example, a group of researchers developed such a system in their work [22]. In their system, there are two microphones. For each microphone, a spectrogram is calculated, rendering two time-frequency bin representation of the signals. From these two spectrums, the feature of phase differences and magnitude differences are extracted. Specifically, in each time frequency bin, the phase value from one microphone spectrum is subtracted from phase value from another microphone. The same procedure is applied to magnitude. The pairs of features are then clustered using a 2D histogram. Figure 2.2 illustrated that, in the 2D histogram, there are 6 peaks corresponding to 6 different speech sources. Each peak is then used to build a binary TF mask for the recovery of that source.

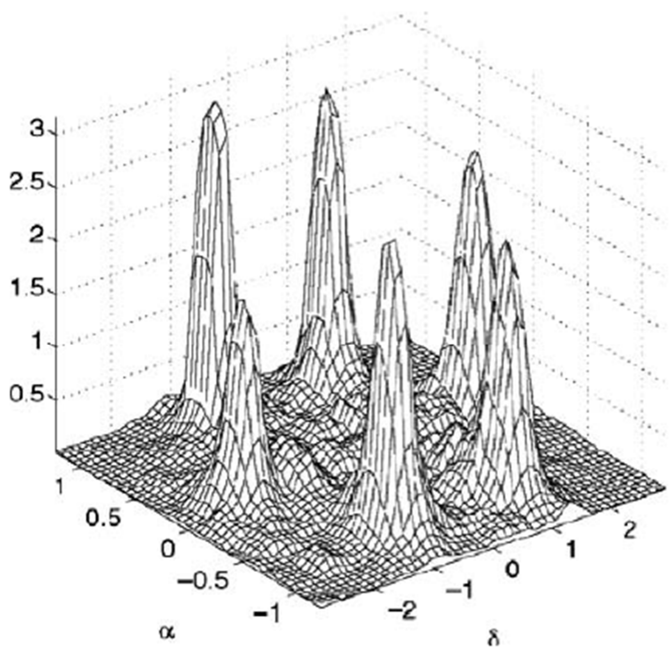

Figure 2.2 The 2D histogram presented in [22]. α indicates amplitude difference, δ indicates time difference. The histogram is generated from 6-source mixture. Each peak corresponds to one source.

A similar work at the same time period is proposed by another group of researchers [23]. In their work, they extract pairs of phase and magnitude differences in each time frequency bin. They formulate the estimation of IBM from the extracted feature as a binary classification problem. As a first step, they estimate the distribution of each class using a nonparametric density estimation method. And then, the classifier is built using a Maximum A Posterior (MAP) decision rule.

It is worth noting that, in both systems, the phase and magnitude differences are used as features. In terms of human hearing system or headphones, these features are called interaural intensity difference and interaural time difference. Theses features are widely used in similar system and related tasks, including direction of arrival estimation and source separation [14]. Illustrations of Interaural Intensity Difference (IID) and Interaural Time Difference (ITD) are found in Figure 2.3 and Figure 2.4. These features are inspired by human hearing system localization cues. It is found by researchers that, changes of spatial configuration trigger systematic changes of binaural cues. What features are used when the system has more than two microphones? In some systems, pairwise generalized cross correlation (GCC) are used among 8 microphones [24]. More spatial features are proposed recently [25], as compared to end-to-end systems where features are learned by neural networks in a data driven way.

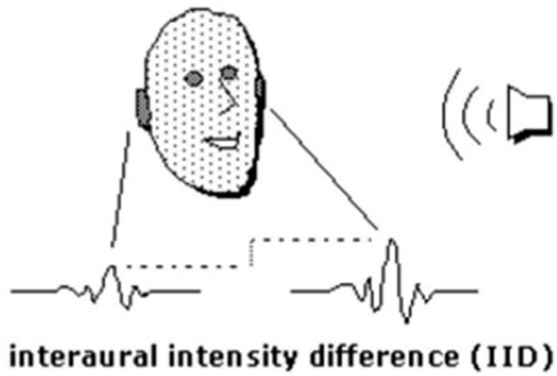

Figure 2.3 An illustration of IID. Taken from [50]

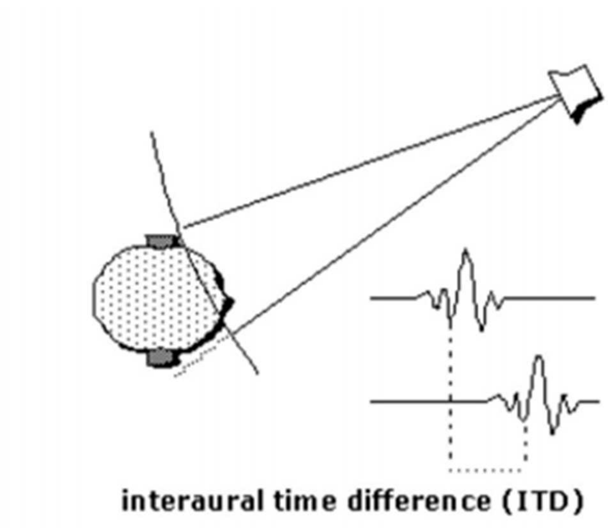

Figure 2.4 An illustration of ITD. Taken from [50]

2.2 Neural Time Frequency Mask

While there is a plethora of work developed on the idea of estimating time frequency mask from neural network, in this thesis, the sections below only introduce the one we applied to perform dereverberation.

2.2.0 Overview

In contrast to the two approaches of time frequency mask estimation described in the section 2.1.2, this approach adopts an end-to-end strategy. Specifically, no hand-crafted feature is extracted in each time frequency bin. Instead, the raw spectrogram is fed as input to the overall system, leaving the feature to be learnt by neural network directly.

Another major difference from previous systems is that this system takes a data-driven strategy. A different category of time frequency estimation approaches takes a model-based strategy [27, 28, 29, 30, 31, 32]. The detailed introduction of such approaches is beyond the scope of this thesis. In brief, model-based approaches employs feature space clustering based on an underlying model of the mixture distribution. For example, for Fourier transform coefficients of microphone signals, the underlying model has been proposed to be Watson mixture model. After the model is selected, an Expectation Maximization (EM) algorithm is employed to fit data onto the model. The results of [26] are presented in Figure 2.5. The model-based method recognizes high frequency noise as speech in the right part of subfigure (c). In contrast, the data-driven method jointly estimates the mask over all frequencies. As shown in (b), high frequency noise doesn't confuse the learned neural network.

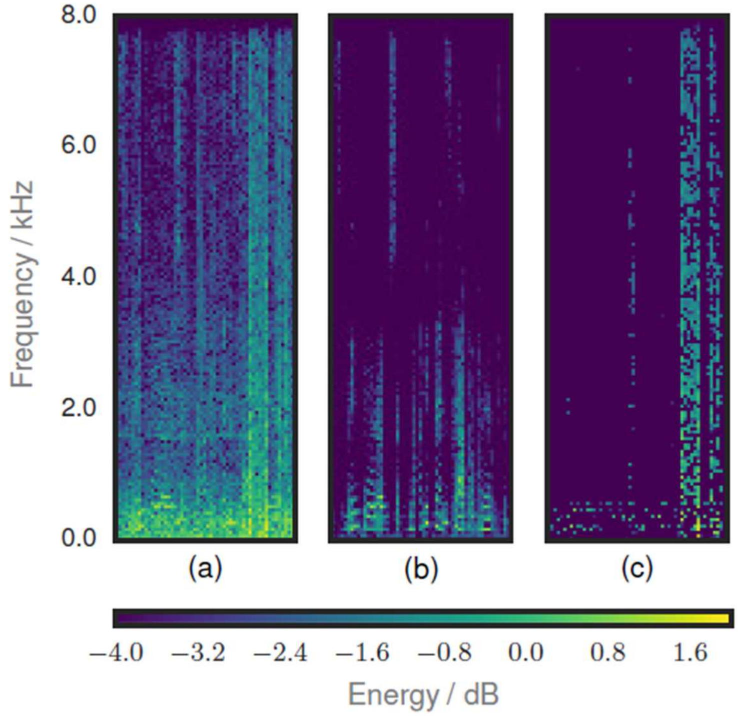

Figure 2.5 Spectrogram of mixture in (a). Mask obtained by deep neural network in (b). Mask obtained by model-based method in (c). Taken from [26].

Chapter 3: Related Works

3.1 Reverberation and room acoustics

The phenomenon of reverberation has been utilized by human auditory perception of space [33]. It is also well known that other species, such as bats, also utilizes reverberation for navigation. Humans utilizes reverberation by perceiving the difference of sound reaching at two ears. The difference of arrival time and magnitude has been used to estimate the direction of source, a phenomenon that has also been used by researchers to develop algorithms for machines to estimate direction of source as mentioned in chapter two. This spatial processing of sound perceived at two ears enhances the intelligibility of speech. Similarly, in music audio, consumers prefer stereo or surround sound reproduction, which creates vivid and enjoyable music experience. Reverberated music audio is added great realism by multiple copies of the direct path signal. In both examples, human spatial perception and stereo music reproduction, the existence of reverberation is beneficial and desirable. The question of dereverberation arises, "Why do we want to remove reverberation in the first place?".

To answer this question, we need to be aware of the situations where reverberation is a major obstacle for researchers to improve the performance of their algorithms. In several application contexts, such as automatic speech recognition, speaker identification, automatic speech-to-text conversion and car interior communication systems, the existence of reverberation leads to the degradation of performance of algorithms. In situations where speech signals are obtained from a distance, such as

Amazon Echo or hearing aids devices for hearing loss community, the removal of reverberation is critical for improving user's experience.

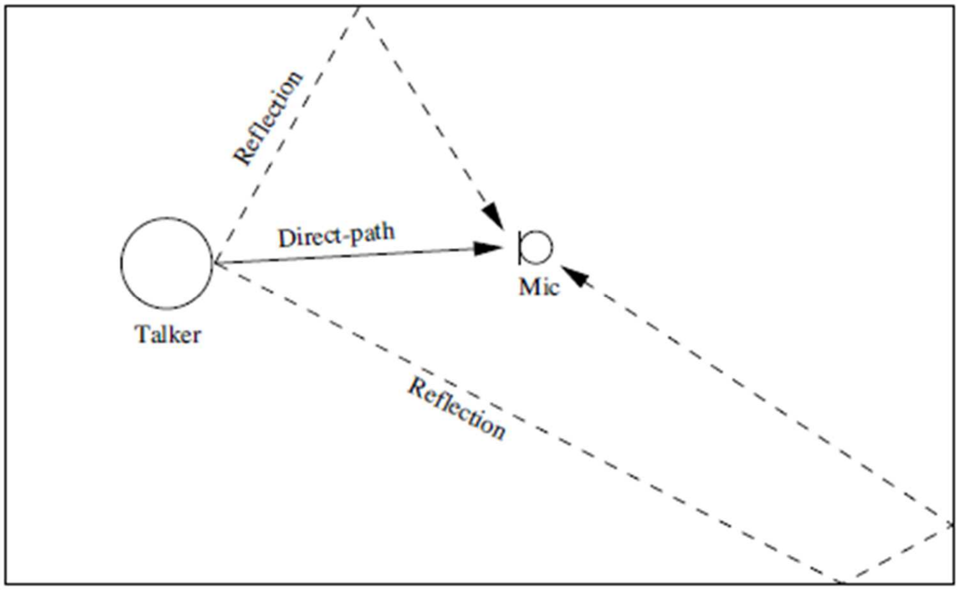

Figure 3.1 Schematic illustration of room reverberation taken from [1]

In Figure 3.1, reverberation is created at the M receiving microphones (only one is pictured), where talker speaks utterance in an enclosed space, such as a meeting room or car interior. The source speech $s(n)$ from talker travels towards all directions, some of which hits surrounding wall, floor or other objects (not shown in the figure) in the room, resulting in multiple delayed and attenuated copies of the source signal. The received signal $x_m(n)$ can be represented as

$$x_m(n) = \sum_{i=0}^{\infty} h_{m,i}(n)s(n-i)$$

where $h_{m,i}(n)$ represents the attenuation caused by reverberation. The purpose of dereverberation is to find a system that takes $x_m(n)$ as input and output $\hat{s}$ which is an estimate of $s(n)$.

The room impulse response (RIR) characterizes the reverberation phenomenon of a given room $\boldsymbol{h_m}(n) = \left[h_{m,0}(n), h_{m,1}(n), \dots, h_{m,L_h-1}(n)\right]^T$. Usually only first L_h taps are considered. Several models of RIR has been intensively studied in the literature. In this thesis, computer simulated room impulse response is used to create reverberated speech signal given a clean source speech signal as input. An important quantification of room impulse response is reverberation time. The reverberation time, T60 or T30, is the time taken for the reverberation energy to decay by 60 dB or 30 dB compared to its original sound source energy [1]. Surprisingly, the reverberation time for a room is only determined by the room geometry and reflectivity of the reflecting surfaces. In contrast, the room impulse response is a complex function of the location of sender and receiver. Within the same room, room impulse responses are different given each pair locations of sender and receiver, while reverberation time stays the same. A greater value of reverberation time indicates more reverberation and thus more challenging for the dereverberation algorithm to find a 'good' estimate of *s(n).*

Besides reverberation time, another useful clue to develop dereverberation algorithms is the difference between early and late reflections. An example of RIR marked with early and late parts is presented in Figure 3.2. Usually the time at which to separate early and late reflection is chosen by the researchers' preference. A commonly used value is 50 milliseconds. Early reflection features well-defined large magnitude impulse response. Early reflection is beneficial to intelligibility of speech but degrades the quality of recorded speech. In contrast, late reflection features diffuse and seemly randomly distributed impulse response [1]. Dereverberation algorithms can choose to treat early and late reverberation differently based on their properties. Reverberation simulation software also utilizes this difference to render them separately for faster speed.

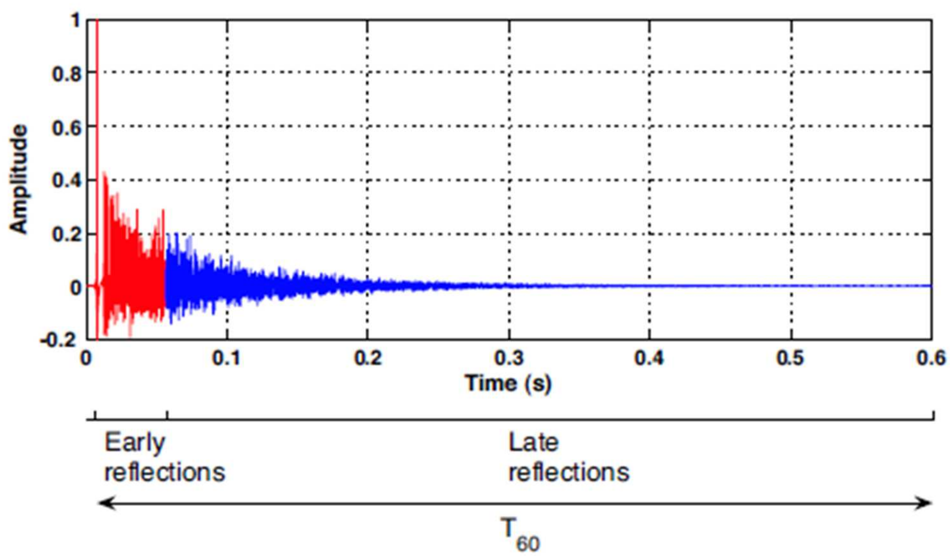

Figure 3.2 An example room impulse response taken from [1]

3.2 Dereverberation algorithms

Dereverberation algorithms fall into three categories, beamforming, speech enhancement and blind system identification and inversion [1]. A thorough review of them is beyond the scope of this thesis. In the following paragraphs, a brief introduction of them is presented.

Beamforming are fundamental techniques for microphone array signal processing. Beamforming can be viewed as spatial filtering as compared to its counterpart of temporal filtering. The most direct and straightforward technique in this category is delay and sum beamformer [34, 35], where microphone signals x_m's are delayed, to compensate for different propagation delay τ_{m}, and then weighted by w_m and summed. The output $\bar{x}(n)$ of a delay and sum beamformer can be represented as

$$\bar{x}(n) = \sum_{m=1}^{M} w_m x_m(n - \tau_m)$$

In this way, the target signal is enhanced, and reverberation component are attenuated. In other words, the microphone array is steered towards the direction of the direct path signal. The reverberation, which comes from all directions, are excluded from the aligned direction. The term beamforming comes from "forming a beam" of sensitivity in the chosen direction. In this way, the limitation of beamforming based dereverberation emerges. It works best for sources that is localized and worst for diffuse case. In cases where reverberation enters the look direction, it can only be partially suppressed.

Speech enhancement approaches covers several techniques. Weighted prediction error algorithm is a popular method in this category. WPE removes only the late reverberated speech $x_{t,f}^{late}$ from the reverberation mixture $y_{t,f}$.

$$y_{t,f} = x_{t,f}^{early} + x_{t,f}^{late}$$

The main idea of WPE is to estimate the late reverberated part as

$$x_{t,f,m}^{late} = \sum_{\tau=\delta}^{\delta+K-1} \sum_{m\prime} w_{\tau,f,m,m'}^{H} y_{t-\tau,f,m'}$$

and then subtract it from the mixture. In the summation, multi-channel inputs are weighted and summed over time and microphones.

The main assumption of WPE is that $x_{t,f}^{early}$ follows a zero-mean complex Gaussian distribution with unknown time-varying variance $\lambda_{t,f}$.

$$p\left(x_{t,f,m}^{early}\right) = CN(0, \lambda_{t,f})$$

$\lambda_{t,f}$ is estimated in a maximum likelihood way in WPE along with $w_{\tau,f,m,m'}$.

Previous two approaches don't assume full knowledge of the room impulse response. A blind system identification approach first estimates the impulse response $h_m(n)$ using second order statistics [37]. And then an inverse filter $g_m(n)$ is applied to the received signals [38] such that $\mathrm{h_m}(n)^T g_m(n) = K\delta(n-\tau)$ where K and τ are arbitrary scale and delay. There are several major challenges in this approach. One of them is caused by the long duration of room impulse response. Usually, room impulse response contains thousands of coefficients. The accurate estimation of them and the numerical stability of its reverse system remains a problem.

3.3 Neural network based dereverberation

The application of neural network in the area of speech enhancement has been a hot topic [3, 5]. Among them, dereverberation has received less attention compared with other enhancement tasks, such as speech separation where multiple speakers are present [3], and speech denoising without special concerns of reverberation [5]. Among those dereverberation works, where deep learning comes in varies. A straightforward formulation is to directly map reverberated speech to anechoic speech using stacked fully connected layers [6]. A schematic illustration of the architecture of the neural network is shown in Figure 3. 3. Another research line follows the estimation of time frequency mask using deep neural network and estimates a TF mask in the complex domain, which can then be applied to the reverberated speech directly to obtain anechoic speech [7]. Finally, the approach considered in this thesis uses a combined TF mask and beamforming approach [26]. A thorough treatment is presented in the next chapter.

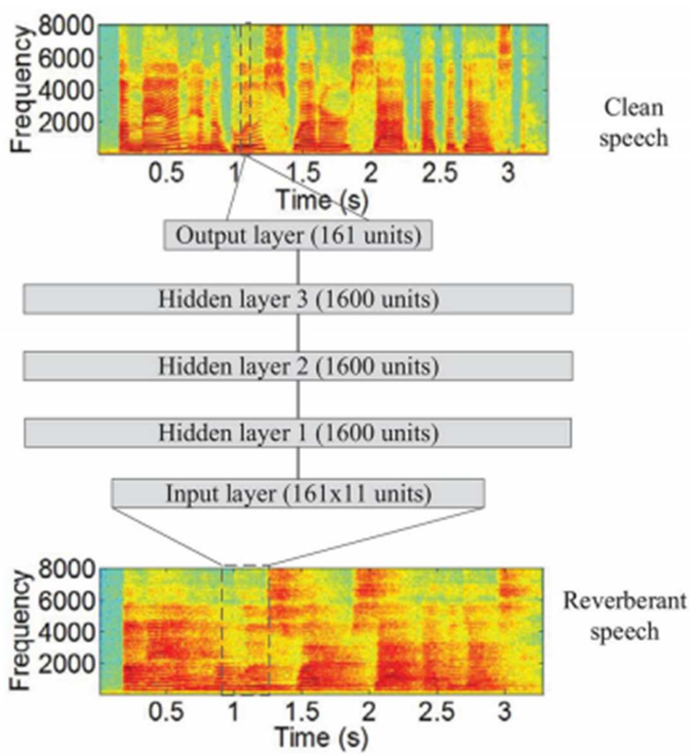

Figure 3.3 A schematic illustration of the architecture of the neural network taken from [6]

Chapter 4: Deep neural network for dereverberation

4.1 Dataset

The dataset used by [26] to train the neural network is the CHiME 3 Challenge dataset [39, 40]. The CHiME 3 Challenge is designed to test speech separation and recognition tasks for each team. The dataset is derived from the popular Wall Street Journal corpus. As shown in Figure 4.1 and 4.2, there is only one speaker talking that is recorded by a 6-channel tablet-based microphone array. Over the whole dataset, four different types of noises are added either as real data or simulated data. The noise types are bus, café, pedestrian area and street junction. Since these are mostly open areas, dereverberation is not a major concern and reported to be most effective in the bus environment [40].

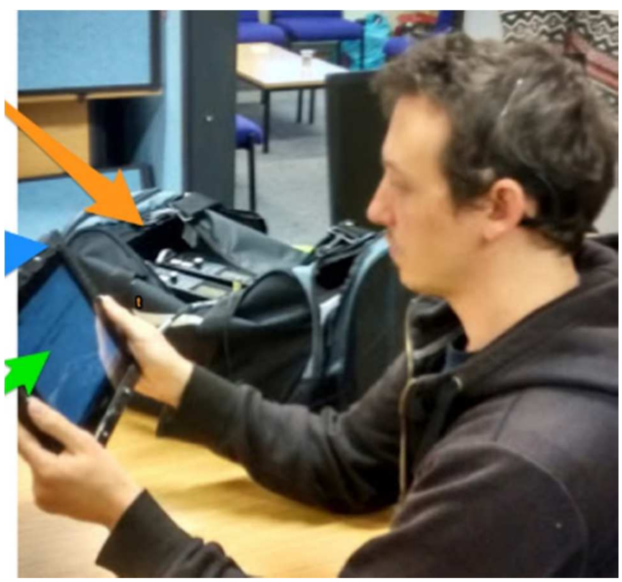

Figure 4.1 An example data recording scene taken from CHiME 3 website

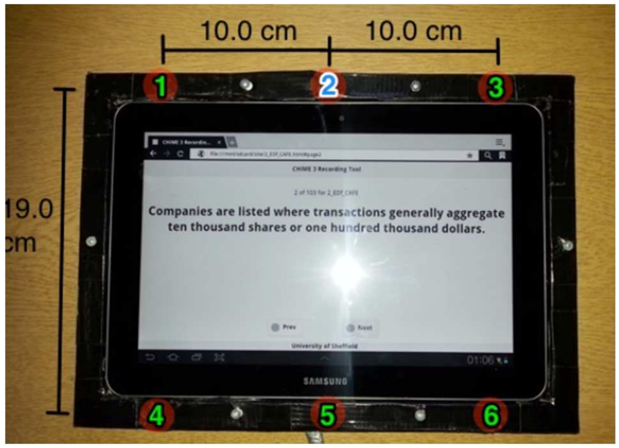

Figure 4.2 Tablet and microphone array taken from CHiME 3 website

To test the trained neural network on the task of dereverberation, we also derive the dataset from CHiME 3 Challenge dataset. However, we used a different microphone array configuration, which is shown if Figure 4.3 and 4.4. We used a circular microphone array with 8 microphones instead of 6. The circular array is of radius 50 cm. Note that a smaller size of head size will be considered as future work, aiming to serve the research topics on head-mounted microphone array. Instead of additive noise, room impulse response is simulated using a research software [41] for a small room of size [width, length, height] = [4m, 4m, 2.5m], under three different reverberation times T60 = {0.3s, 0.6s, 0.9s}. The distance between target and center of array is 1m. The array is located at the center of the room with a height of 1.7m. We treat the reverberated speech as a summation of two parts, one is clean speech convolved with early room impulse response, the other is convolved with late room impulse response. Since these two parts are additive, we treat the second part resulting from late room impulse response as the counterpart of noise in the original dataset. The aim of dereverberation under this formulation, is to remove this second part.

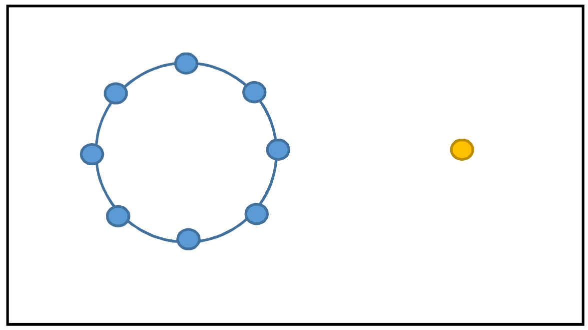

Figure 4.3 A schematic illustration of the microphone array (in blue) and target (in yellow)

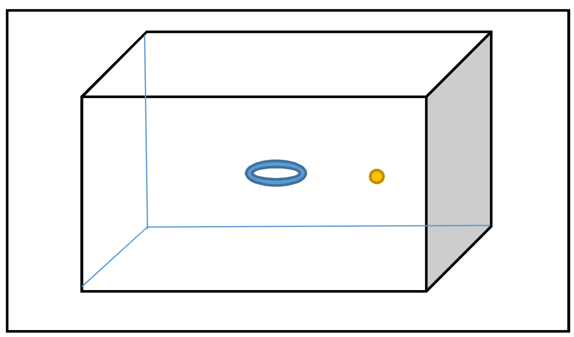

Figure 4.4 A schematic illustration of the room

The software used to simulate reverberation is critical in such research topics. The model of impulse response assumed by the software has direct impact on how the dereverberation algorithm works. In the software we used [41, 42], the early room impulse response is calculated using the classical imaging source method [43]. The idea of this method is to treat reverberation as generated by fake sources obtained by the mirroring or imaging of the receiving microphone w.r.t the wall. The late room impulse response, however, can also be calculated this way, requiring a high computational cost. As a fast computational approach, the software calculates the late response as a random distribution that has a logistic PDF.

We generate a 100 of reverberated speech signal using the software for each reverberation time. The clean speech signal is obtained from Wall Street Journal corpus that is a subset of the CHiME 3 dataset. In other words, CHiME 3 uses the same subset of WSJ to generate their own noisy speech data.

4.2 System Description

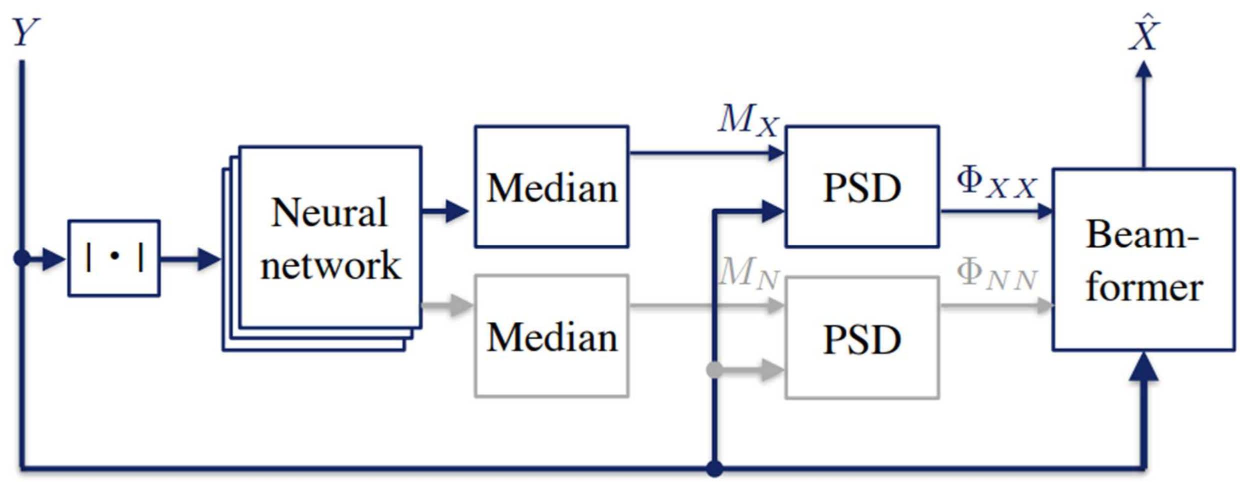

Figure 4.5 System diagram of neural network enhanced beamformer taken from [26]. Parts in grey estimate the noise mask and can be replaced by alternatives. The alternative method can be found in [26] and is omitted here.

4.2.1 Overview

As mentioned in the previous chapter, the adopted approach is a combination of beamforming and time frequency mask based dereverberation scheme. The overall system is shown in Figure 4.5. As a first step, a time frequency mask is estimated using neural network. As a second step, this time frequency mask is applied to the spectrum of microphone array using a Hadamard product. As a third step, PSD of the spectrum is estimated and then used to calculate the weights of beamformers. As a final step, beamforming is performed on the microphone array. In this process, the microphone array is steered toward the direct path signal and early reverberated signal *X*. The late reverberated signal *N* is partially suppressed. In the following description, *X* represents target signal and *N* represents unwanted signal.

4.2.2 Generalized Eigenvector (GEV) beamformer

Let $\mathrm{Y}(t,f), X(t,f), N(t,f)$ be the spectrogram of received signals, speech signals and noise signals respectively.

$$Y(t,f) = X(t,f) + N(t,f)$$

Dropping the time index for clarity, a beamformer is to find $\mathrm{F(f)} = [F(f)_1, F(f)_2, \dots, F(f)_M]$ such that the output $\mathrm{Z(f)} = F(f)^H Y(f)$ is steered towards the look direction.

The objective function of GEV is to maximize Signal-Noise-Ratio (SNR) at the receivers for each frequency bin.

$$\mathrm{F}_{\mathrm{GEV}} = argmax_F \frac{F^H \phi_{XX} F}{F^H \phi_{NN} F}$$

In this objective function, F represents the beamformer weights at given frequency. Since this formula is the same for each frequency bin, the notation of frequency is dropped for clarity. The Power Spectral Density (PSD) of clean speech is represented by $\phi_{\mathrm{XX}} = XX^H$ and of noise is represented by $\phi_{\mathrm{NN}} = NN^H$. In the dereverberation case, clean speech corresponds to speech convolved with early impulse response, and noise corresponds to late reverberation. This object function has a closed formed solution,

$$\phi_{XX} F = \lambda \phi_{NN} F$$

,where F_{GEV} is generalized eigenvector and λ is the eigenvalue.

Usually GEV beamformer is followed by a single channel post filter to obtain a response without distortion in the look direction [45].

$$g_{BAN} = \sqrt{\frac{F_{GEV}^{H} \phi_{NN} \phi_{NN} F_{GEV} / M}{F_{GEV}^{H} \phi_{NN} F_{GEV}}}$$

Finally, the source signal is obtained as

$$Z = g_{BAN} F_{GEV}^{H} Y$$

4.2.3 Estimation of TF mask using neural network

The overall system consists of multiple neural networks, one for each microphone channel. The multiple output of networks will take median operation to get a condensed mask, so that outliers caused by broken microphone can be removed. Specifically, two masks are obtained instead of one: one mask is for speech and the other is for noise. The same condensed masks are then entry-wise multiplied with the microphone array. Then the Power Spectral Density (PSD) of speech and noise are estimated.

The input for each network is a single frame of the spectral magnitude of one channel. The output are two masks, M_X and M_N. The ground truth of these two masks are obtained as below, where $th_N(f)$ and $th_X(f)$ are chosen separately by hand.

$$IBM_N(t,f) = \begin{cases} 1, \frac{||X||}{||N||} < 10^{th_N(f)} \\ 0, otherwise \end{cases}$$

$$IBM_X(t,f) = \begin{cases} 1, \frac{||X||}{||N||} > 10^{th_X(f)} \\ 0, otherwise \end{cases}$$

Note that these two are ideal binary mask which are introduced in Chapter 2. In order to obtain the ground truth of these two masks, it is assumed that both clean speech and noise are known. However, this is often only possible when the data is simulated; when they are recorded in real environment, such information are not always available. Thus, this has been pointed out to be a limitation of this approach [26].

4.2.4 Neural network configuration

The network configuration is summarized in Table 4.1.

	Units	Type	Non-Linearity	p_{dropout}
L1	256	BLSTM	Tanh	0.5
L2	513	FF	ReLU	0.5
L3	513	FF	ReLU	0.5
L4	513/1026	FF	Sigmoid	0.0

Table 4.1 BLSTM network configuration for mask estimation

In the network, the first layer is a Bi-directional Long Short-Time Memory (BLSTM) layer. An illustration of BLSTM is presented in Figure 4.6. BLSTM models the inter-dependency of the time frequency bins both forward and backward in time. The input x_t , representing a single time column in the spectrogram, is mapped to an intermediate representation y_t, that is both depending on forward and backward time columns of $[x_1, x_2, \ldots, x_{t-1}]$ and $[x_{t+1}, x_2, \ldots, x_T]$.

The fully connected layers following BLSTM are omitted from the diagram.

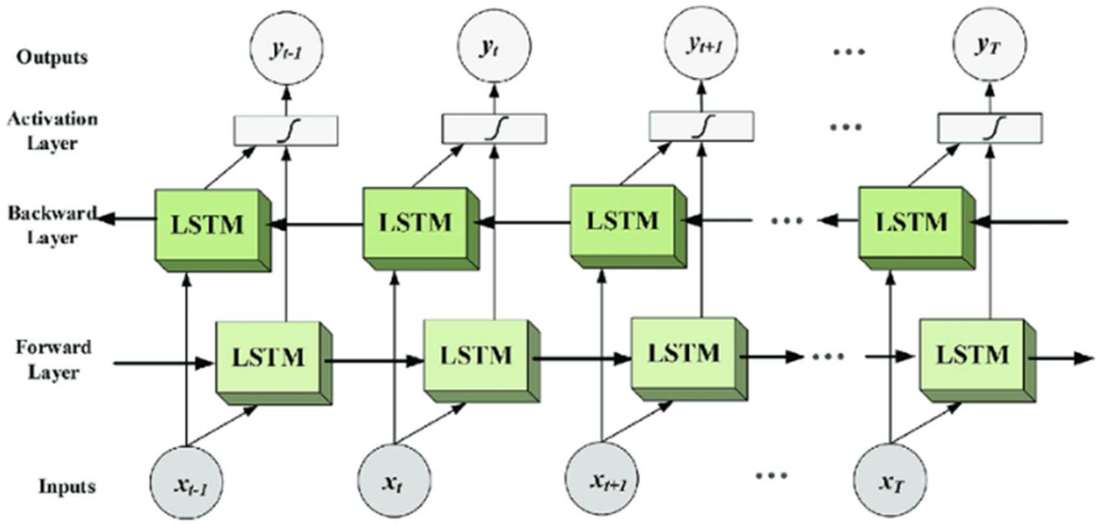

Figure 4.6 An illustration of BLSTM layer taken from [51]

4.2.5 Training scheme

The neural network is initialized with uniform distribution. An RMSProp optimizer [48] is used for training. The optimizer is configured to use a fixed learning rate of 0.001 and a momentum of 0.9. To prevent overfitting, dropout [46] of rate 0.5 is used for training. Batch normalization [47] is applied to all but the output layers. Early termination is applied after 10 epochs when the loss doesn't decrease any more.

4.3 Result

The trained neural networks are obtained from CHiME 3 dataset with noisy background. As mentioned before, we applied this same trained neural network on our simulated reverberated speech. Perceptual evaluation of speech quality (PESQ) score is used to evaluate the effectiveness of this approach.

PESQ score is an objective measure that can be used to predict subjective listening tests results [49]. It is an analogy to the subjective test of Mean Opinion Score (MOS). In a mean opinion score test, human subjects are asked to rate the quality of speech with a score from 5 to 1, mapping to Excellent, Good, Fair, Poor and Bad. To obtain MOS, the arithmetic mean is taken over single ratings from human subjects. This is a subjective test and involves human listeners to participate. PESQ score is, on the contrary, an object score, developed to model a subjective test such as MOS. It takes a reference signal and the signal under evaluation as inputs and output a score that predict the speech quality perceived by a human listener. The structure of PESQ is given below. After alignment of input signals in time, an auditory transform is applied to them. This follows a psychoacoustic model that transforms signal into a representation of perceived loudness in time and frequency. Then these auditory representations are compared and their difference are quantified to predict the audible error perceivable by human listeners. The range of PESQ score ranges from -0.5 to 4.5 and rarely falls below 1 in practice. A standard mapping function from PESQ to MOS-LQO is available, which maps to range between 1.02 to 4.56. A thorough treatment of PESQ score is beyond the scope of this thesis.

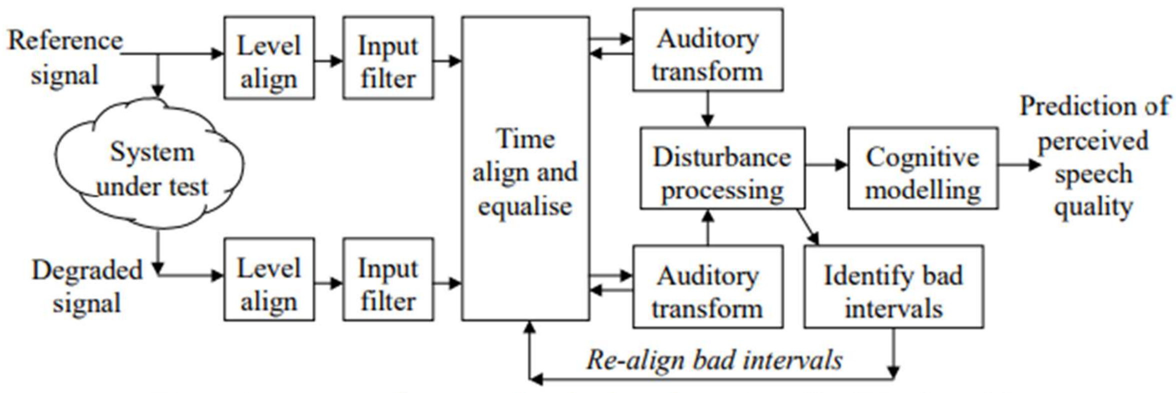

Figure 4.7 Structure of PESQ model taken from [49]

The PESQ score is applied to the reverberated speech before and after dereverberation using the proposed approach. For each reverberation time, 100 utterances are tested using the trained neural network without re-training and the average PESQ score is given in Table 4.2. In all three cases, the PESQ score is improved after the dereverberation. The effectiveness of the approach can be reasoned using the idea of transfer learning, where features extracted from one task can be used to another task and still proved to be beneficial. In our case, this means that the features obtained by the neural network to tell between a speech dominated bin and noise dominated bin, can be transferred to tell between an early reverberated speech dominated bin and late reverberated speech dominated bin.

	PESQ score		
Reverberation Time	0.3 s	0.6 s	0.9 s
Reverberated	2.344	1.858	1.659
Enhanced	2.713	2.199	1.886

Table 4.2

In addition, we also conducted a set of experiments with varying radius of microphone array, reverberation time and sender-receiver distance. A comprehensive summary of results is presented in Table 4.3. There are three major conclusions that we can obtain from those results. First, by varying the sender-receiver distance from 1 meter to 3 meters, the degradation of the algorithm is not significant, and overall improvement ranges from 0.1 to 0.3 in PESQ score. Two, for reverberation times T20 = [0.3, 0.6, 0.9] seconds, the reverberation is significantly more severe than in listed T60 times and the dereverberation is not so effective with only marginal PESQ improvements. Three, overall, a larger microphone array with radius 50 cm is slightly more robust to a head-size array of radius 10 cm. The improvement in PESQ is observed to be larger on the larger array.

	Radius = 10 cm, Head size				Radius = 50 cm			
T60	Before		After		Before		After	
Dist. 1m	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0.3	2.214	0.084	2.573	0.110	2.344	0.082	2.713	0.126
0.6	1.756	0.096	1.931	0.112	1.858	0.102	2.199	0.115
0.9	1.578	0.098	1.711	0.136	1.659	0.101	1.886	0.112
Dist. 3m	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0.3	2.208	0.091	2.394	0.116	2.288	0.094	2.697	0.123
0.6	1.788	0.087	1.975	0.104	1.843	0.087	2.154	0.110
0.9	1.615	0.095	1.744	0.114	1.663	0.089	1.923	0.096
T20	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0.3	1.477	0.096	1.554	0.149	1.567	0.101	1.758	0.111
0.6	1.241	0.135	1.285	0.165	1.332	0.112	1.416	0.130
0.9	1.147	0.200	1.175	0.161	1.228	0.165	1.280	0.143

Table 4.3

Chapter 5: Conclusion

In this thesis, we leverage a successful neural network method on a similar task of speech enhancement to dereverberation, specifically, time frequency mask supported Generalized Eigenvector (GEV) beamformer. This data driven approach introduces feature transferrable to related tasks compared to hand-engineered methods. Our experiments illustrate that the original framework arise from open area enhancement tasks is proved to be effective in our closure space tasks.

The main future direction is whether this framework can be extended to cases where human head is considered to impact the pattern of reverberation. The gap between a circular array and a one mounted on head can be filled with advanced reverberation simulation tools, where the Head Related Transfer Function (HRTF) can be incorporated into the room impulse response. The current simulation tool in concern, however, couldn't implement this feature.

Another question to ask is whether strong initial reflections can also be picked out in this framework. In a similar task, neural network based time frequency masks have been used to separate speech from multiple speakers. Instead of different speakers, can "fake speakers" resulting from initial early reflections be separated in a similar style? This borrows the idea of image-source method in the area of reverberation simulation and hasn't been considered much in the literature.

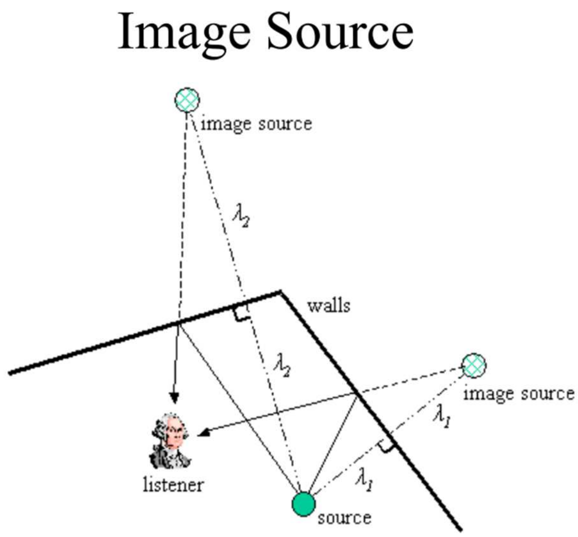

Figure 5.1 Illustration of image source method. Taken from [52]

Bibliography

[1] Naylor, Patrick A., and Nikolay D. Gaubitch, eds. *Speech dereverberation*. Springer Science & Business Media, 2010.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 2015.

[3] Wang, DeLiang, and Jitong Chen. "Supervised speech separation based on deep learning: An overview." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018): 1702-1726.

[4] Xiao, Xiong, et al. "Deep beamforming networks for multi-channel speech recognition." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[5] Lu, Xugang, et al. "Speech enhancement based on deep denoising autoencoder." *Interspeech*. 2013.

[6] Han, Kun, Yuxuan Wang, and DeLiang Wang. "Learning spectral mapping for speech dereverberation." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[7] Williamson, Donald S., and DeLiang Wang. "Speech dereverberation and denoising using complex ratio masks." *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017.

[8] Heymann, Jahn, et al. "Frame-online DNN-WPE dereverberation." *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018.

[9] Narayanan, Arun, and DeLiang Wang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.

[10] Huang, Po-Sen, et al. "Deep learning for monaural speech separation." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[11] Williamson, Donald S., Yuxuan Wang, and DeLiang Wang. "Complex ratio masking for joint enhancement of magnitude and phase." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[12] Wang, DeLiang. "Time-frequency masking for speech separation and its potential for hearing aid design." *Trends in amplification* 12.4 (2008): 332-353.

[13] Lyon, Richard. "A computational model of binaural localization and separation." *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 8. IEEE, 1983.

[14] Wang, Deliang, Guy J. Brown, and Chris Darwin. "Computational auditory scene analysis: Principles, algorithms and applications." *Acoustical Society of America Journal* 124 (2008): 13.

[15] Hu, Guoning, and DeLiang Wang. "Speech segregation based on pitch tracking and amplitude modulation." *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001.

[16] Roman, Nicoleta, DeLiang Wang, and Guy J. Brown. "Speech segregation based on sound localization." *The Journal of the Acoustical Society of America* 114.4 (2003): 2236-2252.

[17] Wang, DeLiang. "On ideal binary mask as the computational goal of auditory scene analysis." *Speech separation by humans and machines*. Springer, Boston, MA, 2005. 181-197.

[18] Moore, B. C. J. "An introduction to the psychology of hearing, academic." *San Diego* (2003).

[19] Jourjine, Alexander, Scott Rickard, and Ozgur Yilmaz. "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures." *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*. Vol. 5. IEEE, 2000.

[20] Narayanan, Arun, and DeLiang Wang. "Ideal ratio mask estimation using deep neural networks for robust speech recognition." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.

[21] Erdogan, Hakan, et al. "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[22] Yilmaz, Ozgur, and Scott Rickard. "Blind separation of speech mixtures via time-frequency masking." *IEEE Transactions on signal processing* 52.7 (2004): 1830-1847.

[23] Roman, Nicoleta, DeLiang Wang, and Guy J. Brown. "Speech segregation based on sound localization." *The Journal of the Acoustical Society of America* 114.4 (2003): 2236-2252.

[24] Xiao, Xiong, et al. "A learning-based approach to direction of arrival estimation in noisy and reverberant environments." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[25] Wang, Zhong-Qiu, and DeLiang Wang. "On spatial features for supervised speech separation and its application to beamforming and robust ASR." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[26] Heymann, Jahn, Lukas Drude, and Reinhold Haeb-Umbach. "Neural network based spectral mask estimation for acoustic beamforming." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[27] Mandel, Michael I., Daniel P. Ellis, and Tony Jebara. "An EM algorithm for localizing multiple sound sources in reverberant environments." *Advances in neural information processing systems*. 2007.

[28] Araki, Shoko, et al. "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior." *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009.

[29] Sawada, Hiroshi, Shoko Araki, and Shoji Makino. "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment." *IEEE Transactions on Audio, Speech, and Language Processing* 19.3 (2010): 516-527.

[30] Ito, Nobutaka, et al. "Relaxed disjointness based clustering for joint blind source separation and dereverberation." *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014.

[31] Vu, Dang Hai Tran, and Reinhold Haeb-Umbach. "Blind speech separation employing directional statistics in an expectation maximization framework." *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[32] Ito, Nobutaka, Shoko Araki, and Tomohiro Nakatani. "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.

[33] Benesty, Jacob, Shoji Makino, and Jingdong Chen. *Speech enhancement*. Springer Science & Business Media, 2005.

[34] Elko, Gary W. "Microphone array systems for hands-free telecommunication." *Speech communication* 20.3-4 (1996): 229-240.

[35] Van Veen, Barry D., and Kevin M. Buckley. "Beamforming: A versatile approach to spatial filtering." *IEEE assp magazine* 5.2 (1988): 4-24.

[36] Nakatani, Tomohiro, et al. "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model." *IEEE transactions on audio, speech, and language processing* 16.8 (2008): 1512-1527.

[37] Huang, Yiteng Arden, Jacob Benesty, and Jingdong Chen. "Adaptive blind multichannel identification." *Springer Handbook of Speech Processing*. Springer, Berlin, Heidelberg, 2008. 259-280.

[38] Miyoshi, Masato, and Yutaka Kaneda. "Inverse filtering of room acoustics." *IEEE Transactions on acoustics, speech, and signal processing* 36.2 (1988): 145-152.

[39] Barker, Jon, et al. "The third 'CHiME'speech separation and recognition challenge: Dataset, task and baselines." *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.

[40] Barker, Jon, et al. "The third 'CHiME'speech separation and recognition challenge: Analysis and outcomes." *Computer Speech & Language* 46 (2017): 605-626.

[41] Lehmann, Eric A., and Anders M. Johansson. "Diffuse reverberation model for efficient image-source simulation of room impulse responses." *IEEE Transactions on Audio, Speech, and Language Processing* 18.6 (2009): 1429-1439.

[42] Lehmann, Eric A., and Anders M. Johansson. "Prediction of energy decay in room impulse responses simulated with an image-source model." *The Journal of the Acoustical Society of America* 124.1 (2008): 269-277.

[43] Allen, Jont B., and David A. Berkley. "Image method for efficiently simulating small-room acoustics." *The Journal of the Acoustical Society of America* 65.4 (1979): 943-950.

[44] Frost, Otis Lamont. "An algorithm for linearly constrained adaptive array processing." *Proceedings of the IEEE* 60.8 (1972): 926-935.

[45] Warsitz, Ernst, and Reinhold Haeb-Umbach. "Blind acoustic beamforming based on generalized eigenvalue decomposition." *IEEE Transactions on audio, speech, and language processing* 15.5 (2007): 1529-1539.

[46] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15.1 (2014): 1929-1958.

[47] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167* (2015).

[48] T. Tielman and G. Hinton, "Lecture 6.5 - RMSProp," COURSERA: Neural Networks for Machine Learning, 2012.

[49] Rix, Antony W., et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE, 2001.

[50] Faculty of Architecture, Design and Planning. "Lecture: Virtual Surround Sound on Headphone." Course DESC 9115. 2015.

[51] Yildirim, Özal. "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification." *Computers in biology and medicine* 96 (2018): 189-202.

[52] Ramani, Course CMSC828D, Spring 2006.