Blind Speech Separation in Distant Speech Recognition Front-end Processing

A Thesis submitted to the department of - Natural Science and Technology II -
in partial fulfillment of the requirements for the degree of Doctor of Engineering (Dr.-Ing.)
Saarland University
Germany

by
Rahil Mahdian Toroghi

Saarbrücken
2016
Tag des Kolloquiums: 10.11.2016

Dekanin/Dekan: Univ.-Prof. Dr. Guido Kickelbick

Mitglieder des Prüfungsausschusses:
Prof. Dr. Dietrich Klakow
Prof. Dr. Dyczij-Edlinger
Prof. Dr.-Ing. Chihao Xu
Frau Dr. Nadezhda Kukharchyk
## Contents

1 **Introduction**
   1.1 Distant Speech Recognition (DSR) Problem ........................................... 1
   1.2 ASR problem formulation: ASR front-end & back-end .................................. 2
   1.3 Scenarios in DSR Front-End Processing ................................................... 7
      1.3.1 Mixture models: Instantaneous, Anechoic, and Echoic .......................... 8
      1.3.2 Discriminant key points for the state-of-the-art approaches ............... 10
   1.4 Databases usable in DSR research ......................................................... 14
   1.5 A brief overview of the thesis contents ................................................. 14

2 **Acoustic Propagation: Analysis and Evaluation** ....................................... 17
   2.1 Physics of Distant Speech Wave Propagation .......................................... 17
   2.2 Entities in a Distant speech propagation scenario .................................... 19
      2.2.1 Speech ............................................................................................... 20
         2.2.1.1 Speech Models .............................................................................. 21
         2.2.1.2 Speech Representations .............................................................. 23
      2.2.2 Noise ................................................................................................. 27
      2.2.3 Reverberation ..................................................................................... 28
      2.2.4 Interference ......................................................................................... 31
   2.3 Human Auditory System versus ASR ......................................................... 33
   2.4 Evaluation Measures of Quality and Intelligibility ..................................... 35

3 **Speech Enhancement: Single Speaker DSR Front-End** ................................. 37
   3.1 Introduction ............................................................................................... 37
   3.2 Single-Microphone Denoising for speech Enhancement ............................... 38
      3.2.1 Spectral Subtraction ........................................................................... 40
      3.2.2 Wiener Filter ...................................................................................... 41
      3.2.3 Maximum Likelihood and Bayesian Methods (Nonlinear Methods) ....... 42
      3.2.4 Bayesian framework for Speech Enhancement ...................................... 43
         3.2.4.1 Estimating a-priori SNR ................................................................. 46
         3.2.4.2 Estimation of the Noise Variance .................................................. 47
      3.2.5 CASA-based enhancement (Masking Method) ...................................... 48
      3.2.6 Dictionary-based enhancement (NMF Method) .................................... 49
   3.3 Single-Microphone Reverberation Reduction ............................................. 51
      3.3.1 A quick survey ..................................................................................... 51
      3.3.2 Linear Prediction based dereverberation ............................................ 54
      3.3.3 Statistical Spectral Enhancement for dereverberation ......................... 57
      3.3.4 Harmonicity-based dERverBeration (HERB) method .......................... 60
### CONTENTS

3.3.5 Least-Square inverse filtering .............................................. 61
3.4 M-Channel Noise/Reverb Reduction for Speech Enhancement .......... 62
3.4.1 Introduction ........................................................................ 62
3.4.2 Beamforming - A General Solution ................................. 63
3.4.3 Multi-Channel NMF-/NTF-based enhancement ....................... 68
3.5 Experiments .......................................................................... 70
3.6 The Proposed Enhancement Structure ....................................... 75

4 Speech Separation: Multi-Speaker DSR Front-End .......................... 81
4.1 Introduction ........................................................................... 81
4.2 Beamforming: An extended view ........................................... 82
4.3 Independent Component Analysis and extensions for BSS ............. 86
4.3.1 ICA and measures of independence .................................... 89
4.3.2 ICA algorithm using the sparsity prior ............................... 91
4.4 Sparse Component Analysis .................................................. 94
4.5 CASA-based Speech Separation ........................................... 96
4.6 Proposed Separation Strategies ............................................ 99
4.6.1 Incorporating DOA with further residual removing filters ....... 99
4.6.2 Removing the coherent residuals, by designing a filter ............ 106

5 Conclusions and Future Works .................................................. 115
5.1 Future Works ...................................................................... 117

A Appendix ............................................................................... 119
A.1 Optimum MVDR and Super-directive Beamformer derivations ....... 119
A.2 Maximum Likelihood ICA estimation .................................... 122

Bibliography ............................................................................. 125
Chapter 3

Speech Enhancement: Single Speaker
DSR Front-End

3.1 Introduction

A human listener needs to put more mental effort to understand the contents of a noisy speech, and can easily lose attention if the signal-to-noise-ratio (SNR) is low.

There are scenarios related to the DSR problem, in which only one speaker is active at a time and the voice is recorded in a reverberant echoic enclosure. *Speech enhancement* is referred to a set of techniques which try to estimate the most likely clean speech, underlying the recorded signal(s) in these noisy/reverberant environments.

Speech enhancement could be performed in a *multi-channel* or a *single-channel* case, based on the number of the microphones used for the recordings. For a set of microphones, having a known geometry is an advantage which could be exploited as an extra prior, with which the localization of the source in the environment becomes feasible. This information is also exploited further, for speech enhancement.

The *anechoic* mixing condition occurs idealistically in an open area *without reflections*. However, that is not what practically occurs. In fact, what we observe in every single microphone of an array is the direct signal propagated from the speaker, along with the reflections from the unknown past frames of the speech, which are attenuated and combined with the direct signal with random delays. This type of mixing process represents an *echoic* environmental condition, which happens in real case.

In the subsequent sections of this chapter, it is assumed that the corrupted signal contains only *one speaker*. Thus, the enhancement algorithm elaborates to recover only one clean source out of a single microphone or a set of microphones in an array. The enhanced speech, later, will be fed into an ASR decoder for intelligibility assessment (e.g., Word Error Rate), and its quality will be evaluated by the corresponding measurements, e.g., PESQ, segmental-SNR, etc.
Noise and reverberation are both, corrupting factors of any application that takes in a recorded speech from microphones located in a closed area. Background noise is assumed uncorrelated with the signal of interest\(^1\) since they are originating from different sources. Reverberation, on the other hand, originates from the same source of interest, however, relates to past signal frames. Due to the reverberation, every frame of the signal (after some lag with respect to the direct frame) would have some level of correlation with the frames coming later, and this level depends on the time difference between the reference frame and the subsequent frame, distance between the speaker and the microphone, frequency of the signal, and conditions of the room (e.g., reflection coefficients, objects in the room, reverberation time constant, and so on).

While early reverberation has been proved to have a positive effect on the intelligibility of the speech, late reverberation deteriorates the signal intelligibility and it is behaved as the correlated noise.

There exist approaches which try to tackle both uncorrelated noise (e.g., background or ambient noise) and correlated noise (e.g., reverberation) simultaneously, such as beamforming\(^7\). However, speech enhancement methods have emerged somewhat chronologically through denoising applications\(^3, 5, 6, 72, 9, 10, 73, 8, 12, 4\), even though methods combining both denoising and dereverberation blocks are of more interest to the speech community\(^4\). Dereverberation methods will also be briefly explained in the subsequent sections of the current chapter.

3.2 Single-Microphone Denoising for speech Enhancement

Figure 3.1, provides a brief classification of all the state-of-the-art methods (single/multi-channel case) and algorithms for denoising process in speech enhancement applications.

Among the methods shown in this figure, the classical techniques are applicable in a single-channel scenario, as well as the multi-channel case. From the other categories, some are applicable only in the multi-channel scenario (e.g., beamforming, Neural networks, and BSS-based methods) and some in both cases (e.g., Dictionary-based methods).

Assuming that the acoustic propagation properties of the environment remains unchanged during the time evolution, the observed data illustrates a noisy convolutive mixing process, which can be formulated\(^2\) as the following equation:

\[
y(n) = x(n) + \nu(n)
\]

\[
y(n) = \sum_{\ell \in \Omega} h(n-\ell)s(\ell) + \nu(n) = h(n) * s(n) + \nu(n)
\]

where \(y(n)\) denotes the noisy observed speech signal in the microphone, \(s(n)\) is the desired

\(^1\)In fact, they are assumed independent which is a stronger condition

\(^2\)Of course for the single microphone speech enhancement scenario in an enclosure
3.2. SINGLE-MICROPHONE DENOISING FOR SPEECH ENHANCEMENT

Figure 3.1: State-of-the-art denoising methods for speech enhancement

Clean speech signal, $\nu(n)$ is the additive ambient noise which is assumed uncorrelated with $s(n)$ or the propagated version\(^1\) of it ($x(n) = h(n) * s(n)$), and $h(n - \ell), \ell \in \mathbb{L}$ denotes the direct and delayed attenuation coefficients for the direct clean signal and delayed versions with random time delays $\ell$, corresponding to the propagation FIR filter\(^2\) with length $\mathbb{L}$. The direct signal attenuation coefficient is usually normalized to one, since the real power of the signal originated from the speaker’s mouth is ambiguous.

There are two essential tasks to perform in order to achieve an effective enhancement. One, is to estimate the noise and second is to remove it from the corrupted observation signal to acquire the desired clean speech. Many of the classical denoising methods, aim at restoring the clean speech spectrum from the noisy microphone signal by applying a gain function to the magnitude spectrum of the noisy signal in each frequency bin to suppress the frequency components based on some criteria, such as the mean square of the error (MSE)\(^46\). In many of these methods, noise spectrum is estimated during the silence periods of the speaker using a voice activity detection (VAD) component. In this chapter, we take a glimpse to the core of the theory which governs denoising methods of speech enhancement, and let the reader to study the details in\(^2\).

---

\(^1\) which is recorded by the microphone

\(^2\) Also known as Room Impulse Response (RIR)
There are three major issues in any of the aforementioned enhancement problems:

1. Determining a specific domain (e.g., time, Fourier, Gabor or STFT\(^1\) Wavelet, etc.) in which the signal can best represent its properties.

2. Determining the optimization rule: e.g., ML, MMSE, MAP.

3. Using a spectral distance measure: e.g., Linear or logarithmic.

4. Determining the statistical model of the speech: e.g., Gaussian, Super-Gaussian, or HMM model.

A broad class of algorithms in speech enhancement choose the STFT domain to represent the data. What makes this transform domain reasonable is that the speech signal is represented sparsely in this domain. It means that, in every short time frame of a speech signal only few frequency components are active at the same time. Since the STFT representation, in general, is a complex value a specific representation which depicts the magnitude of the transform in each frequency bin versus the time evolution of the signal, and is called spectrogram, is preferred. Due to the core operation in STFT, which is Fourier transform, the convolution nature of the equation 3.1 changes to multiplication. Hence, the data model in STFT domain is shown as,

\[
Y(n, \omega) = X(n, \omega) + N(n, \omega) \\
Y(n, \omega) = H(\omega)S(n, \omega) + N(n, \omega)
\]

where the quantities of the signals are based on their magnitude of their spectrogram in every frame index \(n\), and for each frequency bin index \(\omega\). It is clear that, the Acoustic Impulse Response\(^2\) (AIC), \(H(\omega)\), is assumed stationary with respect to the time frame evolutions, and that requires the room condition to remain unchanged within the processing period.

### 3.2.1 Spectral Subtraction

The idea behind spectral subtraction is to estimate the noise magnitude spectrum during the noise only frames, and then to subtract it from the magnitude spectrum of the noisy observation to obtain the clean speech magnitude spectrum. To reconstruct the estimated clean speech signal, both the magnitude and phase are required. Thus, in the absence of an estimate for the clean speech phase, it has been proved that based upon some conditions the phase related to the noisy signal is the optimum surrogate to be assigned as the estimated clean speech phase \(\text{[6,2]}\). Therefore the spectral subtraction, in every frame index \(n\) can be succinctly illus-

\(^1\)Short Time Fourier Transform
\(^2\)which is the Fourier transformed version of RIR
3.2. SINGLE-MICROPHONE DENOISING FOR SPEECH ENHANCEMENT

trated in the following compact mathematical form:

\[
\hat{X}(\omega) \approx \max \left[ |Y(\omega)| - |\hat{N}(\omega)|, 0 \right] e^{j\phi_y(\omega)} \quad \text{for every frame } n
\]  

(3.3)

\[
|\hat{X}(\omega)| = G(\omega)_{SPS} |Y(\omega)| \quad \text{where: } G(\omega)_{SPS} = \sqrt{1 - \frac{|\hat{N}(\omega)|^2}{|Y(\omega)|^2}}
\]  

(3.4)

where \(\phi_y(n,\omega)\) denotes the noisy signal’s phase, and \(G(\omega)_{SPS}\) denotes the gain function associated to the spectral subtraction in order to retrieve the clean speech out of the noisy one. Derivation of the gain function in (3.4) is based on the assumption that clean speech (or propagated one) and additive noise of the environment are independent sources, and hence uncorrelated \(^1\). In contrast to the classical speech enhancement, in a DSR problem happening in a closed area this estimation is far from the desired clean speech, \(S(n,\omega)\). Since the AIR filter, \(H(\omega)\), is a vector of complex values which represents the random reflections of the direct signal for a length of thousands of milliseconds, then the desired clean speech is already distorted in both amplitude and phase for at least the length of this AIR filter. Overlapping the consecutive analysis frames, specially if those frames contain voiced phonemes even deteriorates the problem. In such problems there is always a need for estimating the inverse of the AIR propagation filter, \(H(\omega)\) in (3.2), even though it has been proved that this filter is mostly a non-minimum-phase system and therefore does not have a unique inverse \(^2\). The conclusion is that, spectral subtraction is not an appropriate method in a DSR problem which occurs in an enclosure, because

- In estimating of the clean speech magnitude or power spectrum the propagation filter, which in our DSR scenario is quite complex, is not considered.
- It has this intrinsic weakness of distorting the signal, by ignoring the clean speech phase estimate.
- Overestimation of the noise can make the amplitude estimate negative, and thus the magnitude estimate of the clean speech is set to zero, see (3.3). This nonlinear behavior against the negative values creates isolated random peaks in the spectrum in random frequency bins, which after being converted to the time domain lead to a significant musical noise in the reconstructed signal, and this severely impacts the speech intelligibility.

3.2.2 Wiener Filter

Since spectral subtraction was not derived in an optimal way, Wiener filter was derived as a linear FIR filter to achieve an optimal mean of squared error solution to the clean speech estimation problem of (3.1). The assumption is that the observed noisy signal, \(y(n)\), and the

\(^1\) As we already mentioned all the classical methods of speech enhancement lead to a gain function derivation.

\(^2\) with somewhat random attenuation and delay lags
described clean signal $x(n)$ are to be jointly stationary\(^1\) thus their cross-correlation will depend only on the time lag. Noise is also assumed as Gaussian with zero mean, and is uncorrelated with the clean signal. The frequency domain derivation of Wiener solution to the enhancement problem then, would be as follows:

$$|\hat{X}(n, \omega)| = G(n, \omega)_{\text{WF}} |Y(n, \omega)| \quad \text{where:} \quad G(n, \omega)_{\text{WF}} = \frac{P_X(n, \omega)}{P_X(n, \omega) + P_N(n, \omega)} \quad (3.5)$$

where $G(n, \omega)_{\text{WF}}$ denotes the Wiener gain function for the enhancement, and $P_X$ and $P_N$ denote the power spectral density (PSD) of the clean signal and the noise respectively, which both are unknown and should be estimated from the observed signal, $y(n)$. By rewriting (3.5) using the assumption that clean signal and noise are uncorrelated, and the superposition property of the clean signal and noise (3.2) and approximating the PSD values with a sort-term magnitude spectrum squared, we see that the Wiener gain function is just the square power of the spectral subtraction gain function,

$$G_{\text{WF}}(n, \omega) = \frac{P_Y - P_N}{P_Y} = \frac{|Y(n, \omega)|^2 - |N(n, \omega)|^2}{|Y(n, \omega)|^2} = G_{\text{SpS}}^2(n, \omega) \quad (3.6)$$

In doing so, we still suffer from an audible musical noise which is much less than the spectral subtraction case. Moreover, the AIR propagation filter estimation is still remained as an important problem for a DSR scenario. On the other hand, we need to estimate the unknown values for the clean speech and the noise power spectral density, anyway.

### 3.2.3 Maximum Likelihood and Bayesian Methods (Nonlinear Methods)

A popular statistical approach to estimate the clean speech out of the observed noisy signal would be to use the maximum likelihood (ML) method\(^8\). Regularly, in this method we consider a conditional probability of the observed signal vector given the latent parameters, $\theta$, to follow a distribution whose parameters are unknown but deterministic. These parameters are essentially taken to be as the clean speech power spectral density and this conditional probability is called the likelihood function. The goal of ML method, then would be to infer the optimum latent parameter\(^2\) which can maximize this conditional probability. To ease the calculations, the logarithm of the likelihood function is usually involved which does not affect the solution, since logarithm is a strictly monotonic function over its arguments. Thus, we have,

$$\hat{\theta}_{\text{ML}}(\omega) = \arg\max_{\theta} \log p(Y(\omega) \; | \; \theta(\omega)) \quad (3.7)$$
3.2. SINGLE-MICROPHONE DENOISING FOR SPEECH ENHANCEMENT

where $\theta$ denotes the set of latent parameters in each frame as a vector, underlying the observed noisy signal of that frame in Fourier domain, $Y(\omega)$. These parameters contain the magnitudes and phases corresponding to the complex valued spectrum of the noisy and clean speech sources, as well as the additive noise. A speech enhancement method then, aims at estimating the clean signal, $X(\omega)$, given the noisy signal and these latent parameters, as:

$$\hat{X}_{ML}(\omega) = G_{ML}(\omega)Y(\omega)$$ \hspace{1cm} (3.8)

where $G_{ML}(\omega)$ denotes the gain function of ML-estimate in every frame. The noise spectrum is assumed to follow a zero mean complex Gaussian probability distribution with a symmetric structure for real and imaginary parts, and assuming the clean signal as an unknown but deterministic value, we are implicitly assigning a Gaussian distribution to the noisy signal, too. There are two unknown values of magnitude and phase of the spectrum which are to be estimated. However, the phase parameter is considered as unimportant \cite{76}, and though is integrated out of the distribution.

$$p_L(Y(\omega_k);X_{\omega_k}) = \int_{0}^{2\pi} p_L(Y(\omega_k);X_{\omega_k},\theta_x)\, p(\theta_x)\, d\theta_x$$ \hspace{1cm} (3.9)

where the values are in a frequency bin $\omega_k$ of every time frame, and $\theta_x$ denotes the phase value for the clean signal, and is assumed to have a uniform distribution on $[0,2\pi]$. By placing the associated values of the probability distributions in \eqref{3.7}, and taking the derivatives as in \cite{77}, the corresponding gain function is obtained as,

$$\hat{X}(\omega_k) = \left[ \frac{1}{2} + \frac{1}{2} \sqrt{\frac{\gamma_{\omega_k} - 1}{\gamma_{\omega_k}}} \right] Y(\omega_k)$$ \hspace{1cm} (3.10)

where $\gamma_{\omega_k}$ denotes the a posteriori or measured signal-to-noise ratio (SNR) based on the observed data.

### 3.2.4 Bayesian framework for Speech Enhancement

In this approach, the latent parameters, contrary to the ML approach, are assumed as random variables yet unknown. Hence, the prior information about these random variables (unknown parameters) can be involved. The maximum a posteriori (MAP) objective function to infer the latent parameters, is as follows:

$$\hat{\theta}_{MAP}(\omega) = \arg\max_{\theta} p(Y(\omega)\mid \theta(\omega))\, p(\theta)$$ \hspace{1cm} (3.11)
where the conditional distribution represents the likelihood function, whereas the marginal distribution represents the prior knowledge about the parameters.

While the Wiener filter achieves the optimum linear estimation for the complex spectrum, it is not the optimum magnitude spectral estimator. Therefore, the optimum spectral amplitude estimator in a Bayesian framework can be achieved by solving the following problem [78]:

\[
\hat{X}_{\text{mmse}}(\omega_k) = \mathbb{E}[X(\omega_k)|Y] = \int X(\omega_k) p(X(\omega_k)|Y) \, dX(\omega_k)
\]

where \(Y\) denotes the spectral amplitude vector of every frame of the noisy speech for all the frequency bins, and \(X(\omega_k)\) denotes the clean speech spectral amplitude of the frequency bin with central frequency of \(\omega_k\). Figure 3.2 illustrates a conventional block diagram of the underlying tasks in an MMSE based Bayesian speech enhancement system.

Unlike the Wiener filter, the Bayesian MMSE estimator requires some knowledge about the probability distributions (pdf) of the clean speech and the noise. Obtaining the true pdf for the speech in the Fourier domain, is not easy. That is largely due to non-stationarity of speech. Speech signals are only quasi-stationary for short time frames and the true pdf can not be achieved using the information of a short time period.

---

1 in the minimum-mean-square error sense (MMSE)  
2 by ignoring the phase, due to the unimportance assumption
3.2. SINGLE-MICROPHONE DENOISING FOR SPEECH ENHANCEMENT

Ephraim and Malah [9] proposed a statistical model that circumvents these difficulties by utilizing the asymptotic statistical properties of the Fourier coefficients. The reasonable assumptions they made in their model, are:

1. The real and imaginary Fourier coefficients of the noisy speech have a Gaussian pdf with zero mean and time-varying variances due to non-stationarity of speech. That can be justified, since Discrete Fourier Transform (DFT) can be defined as a sum over the samples contained in the windowed frame of the time domain signal weighted by the exponential terms, as:

\[ X(\omega) = \sum x(n) e^{-j\omega n} = x(0) + e^{-j\omega} x(1) + \cdots + e^{-j\omega(N-1)} x(N-1). \]

Now, using the Central Limit Theorem (CLT), sum of the random variables that follow any type of distribution with a finite variance, tends toward the Gaussian pdf with a limited variance.

2. The Fourier coefficients of the noisy speech signals are statistically independent (real and imaginary parts) and therefore, uncorrelated, so:

\[ x(\omega_i) \perp x(\omega_j), \forall i \neq j, \]

and \( \perp \) is the sign for independence. It is worth noting that this assumption only holds, when the analysis time frame length tends toward infinity. Conversely, according to the Heisenberg uncertainty principle, the frequency resolution tends toward infinity and that entails the frequency components to become independent.

Apart from the above-mentioned assumptions, what really happens is that the analysis frame has 10 – 40msec length. Thus, the FFT coefficients are somewhat correlated. Moreover, the overlapping frames cause the correlation between time samples of the signal, too.

By applying the Bayes theorem in (3.12) and using the sum and product rules on the conditional pdf to include the phase information as well, the MMSE estimate of the clean speech spectral amplitude takes the form of:

\[
\hat{X}(\omega_k) = \frac{\int_0^{\infty} \int_0^{2\pi} X(\omega_k) p(Y(\omega_k)|X(\omega_k), \theta_\phi(\omega_k)) p[X(\omega_k), \theta_\phi(\omega_k)] d\theta_\phi dX(\omega_k)}{\int_0^{\infty} \int_0^{2\pi} p[Y(\omega_k)|X(\omega_k), \theta_\phi(\omega_k)] p[X(\omega_k), \theta_\phi(\omega_k)] d\theta_\phi dX(\omega_k)} \quad (3.13)
\]

The conditional pdf of the above equation is a Gaussian too, since \( Y(\omega_k) = X(\omega_k) + N(\omega_k), \) and the noise is assumed as a Gaussian. Hence, given the clean signal amplitude and phase spectrum, we have: \( p(Y(\omega_k)|X(\omega_k), \theta_\phi) = p_N(Y(\omega_k) - X(\omega_k)), \) which has again a Gaussian pdf. It is notable, that the complex spectrum of the noisy speech follows a Reileigh distribution, since it is a superposition of two Gaussian random processes. By assuming that the spectral phase information for the clean speech is independent from its amplitude spectrum, and is uniform in \((-\pi, \pi)],\) then the joint pdf of the clean signal amplitude spectrum and phase spectrum is a factorization of their individual pdf’s.

Now, by replacing the conditional and joint distributions inside the probabilities contained in (3.13), and following the same gain-function strategy as before, the MMSE estimate of the
spectral amplitude and the associated gain-function \( G_{\text{mmse}} \) is obtained \([2]\), as:

\[
\hat{X}(\omega_k) = \frac{\sqrt{\pi}}{2} \frac{\sqrt{\psi_\omega}}{\gamma_\omega} e^{-\psi_\omega/2} \left( (1 + \psi_\omega) I_0(\frac{\psi_\omega}{2}) + \psi_\omega I_1(\frac{\psi_\omega}{2}) \right) Y(\omega_k)
\]

where \( \gamma_\omega \) denotes the a-posteriori SNR, \( I_0 \) and \( I_1 \) are the Bessel functions of the zero and first order, respectively. The entity \( \psi_\omega \) is related to the a-posteriori SNR as well as a new defined entity named \( \xi_\omega \), a-priori SNR or true SNR. This relation is, as follows:

\[
\psi_\omega = \frac{\xi_\omega}{1 + \xi_\omega} \gamma_\omega
\]

The clean speech amplitude spectrum would be estimated by first approximating the a posteriori SNR as \( \gamma(n, \omega_k) = |Y(n, \omega_k)|^2 / P_X(n, \omega_k) \), and \( \xi(n, \omega_k) = P_X(n, \omega_k) / P_N(n, \omega_k) \) as the a priori SNR, with \( P_X \) and \( P_N \) referring to the PSD of the clean speech and noise, respectively.

Loizou showed that the MMSE gain-function during the large values of the a priori SNR performs exactly as the Wiener filter noise suppression \([77]\). When the a priori SNR is low, then the Wiener filter provides higher suppression than MMSE which also costs more musical noise distortions in the output, while MMSE compromises between suppression and distortion by the inherent trade-off between a priori SNR (\( \xi \)) and a-posteriori SNR (\( \gamma \)), and hence leads to much less audible distortions. Moreover, if the noise is assumed as a Gaussian, then the optimal phase estimate would be the noisy signal phase \([2]\).

The problems with the MMSE estimate is that \( \xi_\omega \), and the noise variance are to be calculated for every frequency bin in advance. However, only the noisy speech signal is available. Therefore, a Voice Activity Detector (VAD) system should be used to determine the frames in which the speech is active and therefore the remaining frames would belong to noise. The approximation of the noise variance would be feasible, if it is stationary. An alternative method would be to assign a speech presence probability to every frame \([79, 80]\). The following sections only briefly glimpse the issues of a priori SNR, and noise estimation.

A more sophisticated derivation of this Bayesian suppression rule was derived by Ephraim and Malah in the MMSE log-spectral amplitude sense, which mimics the logarithmic regime of humans auditory perception system \([6]\). Furthermore, there are extensions of these methods which consider different distributions than Gaussian \(^1\)pdf for the clean speech signal \([12]\).

### 3.2.4.1 Estimating a-priori SNR

The MMSE estimation method of the clean speech spectrum is sensitive to the inaccuracy of the a priori SNR estimate. Several methods have been proposed to overcome the inaccuracies. Among them, are:

\(^1\)Such as Laplace or Gamma pdfs
3.2. SINGLE-MICROPHONE DENOISING FOR SPEECH ENHANCEMENT

1. **Maximum-Likelihood method**: This method first estimates the clean speech variance (which is assumed deterministic, and unknown) and then using VAD system finds the noise variance estimate from non-speech frames. The clean speech variance is calculated by moving-averaging over the past $L$ frames of the noisy speech (in every frequency bin at every time frame), while the variance of the estimated noise is subtracted out. This is shown, as the following equation:

$$\hat{\lambda}_x(\omega, n) = \max \left\{ \frac{1}{L} \sum_{j=0}^{L-1} Y^2(\omega, n - j) - \sigma_n^2(\omega, n), 0 \right\}$$  \hspace{1cm} (3.16)

Now, by dividing both sides by the noise variance, $\sigma_n^2(\omega, n)$, we obtain the *a priori* SNR, as:

$$\hat{\xi}_\omega(n) = \max \left\{ \frac{1}{L} \sum_{j=0}^{L-1} \gamma^2(\omega, n - j) - 1, 0 \right\}$$  \hspace{1cm} (3.17)

where $\gamma$ is the *a posteriori* SNR which is obtained, by $\gamma(n, \omega_k) = |Y(n, \omega_k)|^2 / P_N(n, \omega_k)$.

2. **Decision Directed approach**: This approach is based on the relationship between a-priori and a-posteriori SNRs [9]. The idea behind this method is that, the speech signal changes more slowly than the normal framing period, 10-40 ms. Therefore, there is a high correlation between the neighboring frames of clean speech amplitudes. Therefore, we can use this correlation to approximate the clean speech amplitude of the current frame with the one from the previous frame. It actually combines the present a-priori SNR estimate from the Maximum Likelihood method (3.17), and the past a-priori SNR estimate using the definition, and gives weight to them to represent the correlation between the adjacent frames, as follows:

$$\hat{\xi}(n, \omega) = a \frac{\hat{X}^2(\omega, n - 1)}{\sigma_n^2(\omega, n - 1)} + (1 - a) \max\{\gamma(\omega(n) - 1, 0\} \hspace{1cm} (3.18)$$

where $0 < a < 1$, is the weighting factor and can be optimized based on a measure of intelligibility. The value of $a = 0.98$ is a reasonable value.

### 3.2.4.2 Estimation of the Noise Variance

Methods developed for noise variance estimation are mainly based on three facts:

- Clean speech and noise are independent random processes. the periodogram (square of the magnitude spectrum) of the noisy speech is approximated by the sum of the clean speech and noise periodograms. Hence, the frame with speech absence should have the minimum periodogram, since noise is assumed always active and stationary, or at least it has less variability than speech. This fact leads us to the famous method of **Minimum Statistics**, proposed by Rainer Martin in [11].
Noise has typically a non-uniform effect on the spectrum of speech, meaning that only few regions of the speech spectrum are affected by the noise more than the others. Therefore, the effective SNR for each spectral component of speech is different. This leads us toward the methods of **Time-Recursive Averaging**\(^\text{[72]}\).

The most frequent value of energy values in individual frequency bins correspond to the noise level of that specified bin. So, the noise level corresponds to the maximum energy values. This facts leads us toward the **Histogram-based methods**\(^\text{[81]}\).

For all the aforementioned classes of methods, the sequence of operations are the same. First, the short-time-Fourier-Transform (STFT) analyzes the signal into short time spectra with overlapping frames (e.g., 20–30 msec. windows, and 50% overlap). Secondly, the Computation of the noise spectrum is performed using several of these consecutive frames, and is called as the analysis segment. Typical time span of this analysis segment ranges from 400 msec to 1 sec. One assumption is that, speech varies more rapidly than the noise in the analysis segment and noise is more stationary than speech. Another assumption is that, the analysis segment has to be long enough to encompass speech pauses and low energy segments, but also short enough to track fast changes in the noise level. Therefore, there should be a trade-off between the adequate length that contains the speech variations and noise level changes, while choosing the duration of the analysis segment.

### 3.2.5 CASA-based enhancement (Masking Method)

Human has a remarkable ability to recognize speech under several harsh conditions, such as a closed room environment with noise and reverberation and even multiple concurrent speakers. This ability to pay a selective attention to a single speaker in the midst of noise and babble from several other speakers, motivated the masking methods. These methods extract a desired speaker from an observed noisy signal by selectively retaining the time-frequency components of a signal spectrum, which are dominated by the desired speaker and masking out other components. This spectrographic masking, requires a computational model of perceptual grouping of components in the signal spectrum, and the related approach is termed as **CASA**\(^1\). These methods try to mimic the human auditory perception mechanism by grouping together some acoustic cues that can exhibit a certain relationship in the time-frequency plane, by which human is enabled to form a reasonable picture of the acoustic events.

There are several methods of grouping the spectral components mentioned in the literature, such as:

- Based on a harmonic relationship to the fundamental frequency of the speaker\(^\text{[82]}\).
- Based on physiologically motivated factors\(^\text{[83]}\), e.g., onsets and offsets, temporal continuity, etc.

\(^1\)CASA ≡ Computational Auditory System Analysis
3.2. SINGLE-MICROPHONE DENOISING FOR SPEECH ENHANCEMENT

- Based on a data-driven approach called as spectral clustering [84].
- Based on statistical dependencies between spectral components [85, 86].

Ideal binary mask (IBM) is a method proposed in CASA, which tries to segregate a speech signal from the noise, by deciding whether a T-F unit in the spectrum of the noisy signal is dominated by the desired signal or the noise. A general definition of the binary mask is as,

\[ M(n, \omega) = \begin{cases} 1 & \text{SNR}(n, \omega) > \eta \\ 0 & \text{Otherwise} \end{cases} \]  

where \( \eta \) denotes a preset threshold. When the clean speech is also available, then this \( \eta \) value can be obtained accurately. In this case, the resulting mask is termed as ideal binary mask (IBM). True mask, yields the clean signal to be approximately reconstructed, by selecting the units in T-F spectrum which truly belong to the speech signal and masking out the T-F units which belong only to the noise. Due to its mechanism of grouping the speech associated channels, this approach is also referred to as channel selection in speech enhancement. It has been shown, that the ideal binary mask achieved a significant intelligibility improvements for both the normal hearing and hearing impaired listeners [87, 88, 89, 90].

It should be noted that IBM in itself is not a practical applicable enhancement method, since it requires both the clean signal and the noise, separately. However, a reasonable estimate of it could be the goal for any practical algorithm, in this regard. Thus, in a practical binary mask a speech dominant T-F bin is preserved, while the noise dominant unit is discarded according to a threshold. Hence, the problem would be to perform a binary classification, and this could be achieved using several machine learning methods (e.g., SVM\(^2\), GMM\(^3\), etc).

3.2.6 Dictionary-based enhancement (NMF Method)

So far, most of the speech enhancement methods were based on the statistical model-based approaches, in which the desired speech and noise have separate models with associated parametric distributions, and these parameters were estimated from input noisy signal. The advantage of these methods is that no a priori training is required. However, these (generally unsupervised) class of methods do not work effectively for a non-stationary noise case, since the noise model is constructed based on the stationarity assumption.

Another major class of speech enhancement methods (for single channel case) have been emerged later, which leverages a data-dependent or supervised enhancement. In these methods, a priori information is required to train the speech signal or noise bases (i.e., dictionary atoms), separately. A prevalent method of NMF(Non-negative Matrix Factorization) usually

\(^1\)Time-Frequency  
\(^2\)Support Vector Machine  
\(^3\)Gaussian Mixture Model
CHAPTER 3. SPEECH ENHANCEMENT: SINGLE SPEAKER DSR FRONT-END

trains an over-complete dictionary for speech signal and noise, independently [91, 41, 92, 93]. NMF projects a non-negative matrix onto a space which is spanned by a linear combination of a set of basis vectors, as follows:

\[ \mathbf{Y} \approx \mathbf{B} \mathbf{W} \]  \hspace{1cm} (3.20)

where \( \mathbf{B} \) is the matrix whose column vectors are the bases trained by the data, and \( \mathbf{W} \) is the gain (or activation) matrix, whose rows denote the set of weights or activation gains to be assigned to each of the corresponding bases, and both are non-negative matrices (see figure 3.3). Since,

there is no assumption about the nature of the noise these methods are more robust against the non-stationary noise. The solution to the problem of (3.20) is obtained through solving the following equation:

\[ (\mathbf{B}, \mathbf{W}) = \arg\min_{\mathbf{B}, \mathbf{W}} D(\mathbf{Y} | | \mathbf{B} \mathbf{W}) + \lambda g(\mathbf{B}, \mathbf{W}) \]  \hspace{1cm} (3.21)

where \( D(\mathbf{Y} | | \mathbf{B} \mathbf{W}) \) denotes the Kullback-Leibler divergence (KLD) measure of distance between the approximation and the input magnitude spectrum matrix \( \mathbf{Y} \), and the second term \( g(\mathbf{B}, \mathbf{W}) \) is the regularization term. There could be other cost functions than the KLD, such as Euclidean distance, Itakura-Saito divergence or Negative log-likelihood measure for probabilistic version of NMF. The regularization function \( g(\cdot, \cdot) \), also could be based on the sparsity of the \( \mathbf{W} \) weights or the temporal dependencies of the input data matrices across frames. It should be noted that (3.21) is not a convex problem, hence it should be solved using alternating minimization of a proper cost function such as multiplicative update, iterative gradient descent, or Expectation Maximization (EM) algorithms. Detailed solution to the NMF problem is avoided here, due to their variety based upon the regularization criteria or the deterministic or probabilistic
3.3 SINGLE-MICROPHONE REVERBERATION REDUCTION

Algorithm 3.1 An example of NMF-Based speech enhancement algorithm

training:
\[ B^{(n)} : \text{Noise Basis Matrix} \]
\[ B^{(s)} : \text{Clean Speech Basis Matrix} \]
\[ B = [B^{(n)}, B^{(s)}] : \text{Noisy Speech Basis Matrix} \]

model:
\[ Y = S + N \]

[loop]: \( \forall t = 1 \ldots T \) time frames

Hold Basis matrix \( B \) of noisy signal fixed, and then
Obtain \( W^y_t \), as:
\[ Y_t \approx B W^y_t \text{ using equation } (3.21) \]

[Extract clean signal]:
\[ \hat{S}_t = \frac{B^{(s)} W^{(s)}_t}{B^{(s)} W^{(s)}_t + B^{(n)} W^{(n)}_t} \odot Y_t \]

End Loop

approaches. Any required derivations, will be shown on the occasion.

The enhancement procedure using the NMF method is briefly explain in algorithm 3.1, in which the data model is assumed as a simple superposition of signal and noise. The algorithm, primarily learns the basis vectors for the clean speech and noise independently, and acquires the basis matrix for each. The noisy speech, logically should be constructed by a combination of these basis dictionaries. Then, for each time frame, the basis matrix is assumed to be fixed and the associated gain function is obtained based on the KLD divergence minimization, as in (3.21). The obtained weight (activation) matrix is then used in a Wiener-type filter which yields the clean speech portion of the observed spectrum. All operations in the Wiener-type filter are element-wise. We should notice, that in all the previously explained methods, the model of the microphone signal(s) is assumed to be a superposition of the clean signal and the noise. However, we already mentioned that even the signal part in the microphone which is assumed to be clean, is a heavily distorted version of the source originated by the talker, due to the reverberation or propagation filter. Thus, in many cases dealing with reverberation is a critical problem to achieve a reasonable intelligibility in the ASR system output.

3.3 Single-Microphone Reverberation Reduction

3.3.1 A quick survey

Reverberation reduction methods can be divided into several categories. Referring to (3.1), unless the propagation of the real source is identified, the previously mentioned noise reduction-based enhancement methods are incomplete and what they can achieve is only a distorted version of the clean signal. Thus, a class of methods try to identify or estimate the propagation acoustic impulse response (AIR) filter, by which the inverse filtering may yield the original speech signal. Other methods, try to deal with the late reverberation part of the AIR filter, ow-
CHAPTER 3. SPEECH ENHANCEMENT: SINGLE SPEAKER DSR FRONT-END

ing to the fact that the early reflections only colorize the dry speech and make it more intelligible. These methods can deal with the late reverberation using linear prediction based models \[94, 95\], or correlation of the reverberant speech with the desired part \[96\], or statistical modeling of the late reverberation and using cost functions to optimize the parameters of the chosen model \[33, 75, 97\].

Even though the domain of dereverberation methods is remarkably growing recently, in figure 3.4, we try to represent a set of the major single-versus multi-channel processing methods as the state-of-the-art dereverberation techniques. However, we are merely contented to briefly explain the major techniques of either side (single- vs. multi-channel case). The

methods known by far (to our knowledge), could be classified into various category of approaches. There are deterministic or probabilistic model-based methods which lead to LP residual \[96, 98, 99\], spectral subtraction \[100, 101\], and statistical methods \[102, 75\], and are applicable to both single-, and multi-microphone scenarios. Beamforming \[71\], and some other methods in inverse filtering, instead belong to the multi-microphone cases, in which they try to somehow estimate the AIR propagation filter primarily and then invert it, to achieve the clean sources. Some of the most important methods from either categories will be discussed briefly, later in this chapter.

The Room Impulse Response (RIR), which represents the propagation filter between the source of speech to the microphone(s) has some properties, such as:

(a) RIR is comprised of three major parts, as in figure 3.5: 1) An impulse which represents the direct sound attenuation level (also known the anechoic coefficient) after a certain propa-

Figure 3.4: Dereverberation methods based on the Single-/multi-channel processing, along with the approaches to which they belong to, e.g., Beamforming, inverse filtering, HERB, etc.
3.3. SINGLE-MICROPHONE REVERBERATION REDUCTION

3.3.1 Single-Microphone Reverberation Reduction

Figure 3.5: A typical Room Impulse Response

(b) Statistically, the early reflection part is assumed to be sparse, however, the late reflection part is assumed diffuse and the phase assigned to the reflections at each instantaneous time belonging to the late reflection part is random \[75\].

(c) Late reflections start about 50-100 millisecond after the direct sound. This time could be roughly estimated for every room based on an approximate formula, called mixing time, which shows the transition time between the early to late reflection as follows:

\[
t_{\text{mix}} = 1000 \sqrt{V} \quad \text{sec} \quad (3.22)
\]

where \(V\) denotes the volume of the room per cube meter, \(m^3\).

By increasing the distance between the speaker mouth and the microphone beyond the critical distance, the reverberation sound pressure level starts to get stronger than the direct sound pressure level. The critical distance could be estimated for a room, as follows:

\[
D_c = \sqrt{\frac{V}{100\pi T_{60}}} \quad (3.23)
\]

where \(T_{60}\) is the time when the reverberation energy reaches the \((-60)dB\) of its initial energy.

To conform with the topic of this chapter, we study the theory behind the single microphone based dereverberation methods, succinctly. There are several single-microphone dereverberation methods, which could be divided into the following classes:

- Linear Prediction-based methods
CHAPTER 3. SPEECH ENHANCEMENT: SINGLE SPEAKER DSR FRONT-END

Figure 3.6: Spectrogram of a Clean speech (top) vs. its Reverberated version (bottom) in a room with $T_{60} \approx 500$ msec. The spectral amplitude components are smeared over time, obviously.

- Spectral enhancement methods (e.g., Spectral subtraction, statistical method)
- HERB method, which requires some knowledge about the room AIR (Acoustic Transfer Function)

These methods, usually work better on the early reflection part of the Acoustic Transfer Function (ATF) than the late reflections portion. Therefore, we do not expect a huge effect on the speech recognition outcome. The idea behind these methods is to equalize the propagation channel to achieve dereverberation. Since, the direct inversion of the Acoustic Transfer Function (ATF) is not possible, these methods try to adapt a compensation filter instead, among which the LPC based methods are shown to be the most successful \[103, 45\].

### 3.3.2 Linear Prediction based dereverberation

The speech production mechanism can be modeled as an all-pole filter (a.k.a Auto-Regressive model) which is excited either by a glottal pulse to synthesize the voiced speech phonemes or
3.3. SINGLE-MICROPHONE REVERBERATION REDUCTION

A noise to synthesize the unvoiced speech. On the other hand, it is assumed that the reverberation channel has an all-zero filter model. Thus, the detrimental effect of reverberation will not affect the clean speech, but only the residual, due to forcing merely zeros to the overall system. This motivates the LP-residual based dereverberation method. In such a model, distortions which emanate from the additive noise only affect the excitation sequence and the all-pole filter coefficients (e.g., \( a_k \), \( \forall k \in \{1, \cdots, p\} \)) in Figure 3.8 are assumed to remain intact, since reverberation only forces zeros to the system rather than poles.

A block diagram of a typical LPC based dereverberation, and a typical all-pole model of speech production are presented in Figure 3.7 and 3.8 respectively. Speech dereverberation can be performed through computing the LP residual of the observed signal. There are peaks in the estimated LP residuals which are due to reverberation and noise and are uncorrelated.

The reverberation in the Z-domain reflects the delayed versions of the input signal being accumulated with the direct path signal, and the acoustic transfer function, as a result of it will be an all-zero system.

---

\[ s[n] = \sum_{k=1}^{p} a_k s[n-k] + Gu[n] \]

\[ A(x) = \frac{1}{1-\sum_{k=1}^{p} a_k x^{-k}} \]

---
with the speech related ones. By identifying these peaks and attenuating them using an inverse filter, the clean speech signal can be reconstructed using the estimated all-pole model.

Yegnanarayana et al. [94], proposed to use the Hilbert transformation for LP reconstruction. The Hilbert envelope represents large amplitudes at strong excitations in the temporal signal. Therefore, the Hilbert transform of the reverberant LP residual, causes the pulse train structure of voiced speech to be amplified and the reverberation effects to be attenuated, and that can identify the peaks in the residuals. Moreover, Gillespie [104] used the kurtosis, as a measure of peakedness of the LP residual. The clean speech signal follows a super-Gaussian distribution with high kurtosis. However, the speech distorted with reverberation depicts low kurtosis. Thus, in the LP residual of reverberant speech, the kurtosis decreases with increasing reverberation. Using an online adaptive gradient descent approach that maximizes the LP kurtosis, the reverberation effects can be mitigated and the clean speech estimate can be enhanced. The inverse LP filter gives the LP residual, which is a close approximation of the excitation signal. The clean speech model as an output of an all-pole process is, as follows:

\[
s(n) = - \sum_{k=1}^{p} a_k s(n-k) + u(n) \tag{3.24}
\]

where \(a_k\)'s are the filter coefficients, and \(u(n)\) is the glottal excitation signal. Assuming that the predicted clean speech be \(\tilde{s}(n)\), which can also be modeled as an output of an all-pole process,

\[
\tilde{s}(n) = - \sum_{k=1}^{p} b_k s(n-k) \tag{3.25}
\]

where \(b_k\)'s are the LP coefficients, if the speech were truly generated by an all-pole filter, this equation would precisely predict the speech signal except for the glottal excitation instants, i.e.,

\[
\text{For } a_k = b_k; \text{ error in prediction, } e(n) = s(n) - \tilde{s}(n) = u(n) \tag{3.26}
\]

which is referred to as LP residual. Equation (3.26) clearly shows, that the LP residual whitens the speech signal, and ideally speaking, represents the excitation signal. Similarly, the reverberant speech can be modeled as,

\[
x(n) = - \sum_{k=1}^{p} h_k x(n-k) + e_x(n) \tag{3.27}
\]

where \(e_x(n)\) is the LP residual of the reverberant speech. By modifying the LP residual in such a way that we can achieve \(e_x(n) = u(n)\), the clean speech signal can be synthesized from the filtered residual. Gillespie [104], used the idea that the LP residual of a speech signal increases as the reverberation in speech increases. He presented an adaptive algorithm to maximize the kurtosis of LP residuals. The adaptation filter is controlled by a cost function in a feedback structure (See figure 3.7). This filter (i.e., inverse filter), should affect the LP analyzed signal
such that the resulted output gets the highest kurtosis as a clean speech has. Therefore, the cost function should be applied on the residual such that the signal in the adaptive filter output, $\tilde{y}(n)$, results the maximum kurtosis (or normalized kurtosis of $\tilde{y}(n)$) based upon the filter coefficients, $h_k$, as:

$$J(n) = \frac{E\{\tilde{y}^4(n)\}}{E^2\{\tilde{y}^2(n)\}} - 3 \quad (3.28)$$

Thus, the gradient of the cost function with respect to the filter coefficients are to be zeroed:

$$\frac{\partial J}{\partial h} = 4 \tilde{y} \frac{E\{\tilde{y}^2\}^2 - E\{\tilde{y}^4\}}{E^3\{\tilde{y}^2\}} = g(n)\bar{x}(n) \quad (3.29)$$

where $g(n)$ is the desired feedback function. The filter coefficients, are then updated as:

$$h(n+1) = h(n) + \mu g(n)\bar{x}(n) \quad (3.30)$$

where $\mu$ is the step size. In addition, the expected values can also be calculated recursively,

$$E\{\tilde{y}^2(n)\} = \beta E\{\tilde{y}^2(n-1)\} + (1-\beta)\tilde{y}^2(n)$$

$$E\{\tilde{y}^4(n)\} = \beta E\{\tilde{y}^4(n-1)\} + (1-\beta)\tilde{y}^4(n) \quad (3.31)$$

where the parameter $\beta$ is a factor to control the smoothness of the moment estimates. In every update step, the output (enhanced speech) is calculated as follows:

$$y(n) = h^T x \quad (3.32)$$

### 3.3.3 Statistical Spectral Enhancement for dereverberation

Using deterministic models for reverberation usually comes along with a large number of unknown parameters which are difficult to estimate blindly, and are dependent on the exact spatial positions of the microphones and sources. On the other hand, objects in the room also change these parameters and this entails specific parameter calculations for every room, even with similar dimensions. Because of this difficulty in explicitly modeling the room acoustics, statistical room acoustic modeling has grabbed a significant attention of the researchers. This model provides a statistical description of the acoustic transfer function between the speaker and the specific microphone, which depends on few quantities, e.g., reverberation time. In such a modeling, it is implicitly assumed that the reverberation is classified into two major parts, namely the early reverberation and the late reverberation. Late reverberation needs to be addressed directly for the ASR improvement, and it is assumed to be the superposition of a sufficient handful of reflected individual waves from random preceded versions of the direct signal, with a random delay and attenuation. They also have less correlation with the direct signal than the early reflections. Moreover, based upon the Central Limit Theorem (CLT), the superposition of this handful of unknown speech waves with the same distribution

---

1The late reverberation part is very dense, and is non-sparse, in contrast to the early reverberation, see figure 3.5
and limited variance\(^1\) follows approximately a Gaussian distribution.

Polack developed a time-domain model for the late reverberation in a statistical framework\(^2\). In this model, the acoustic impulse response of the late reverberation part is describes as a realization of a non-stationary process, as

\[
h(n) = \begin{cases} 
  b(n)e^{-\zeta n} & \text{for } n \geq 0 \\
  0 & \text{O.W.}
\end{cases}
\]  

(3.33)

where \(b(n)\) is a zero-mean random stationary Gaussian sequence and \(\zeta\) is a decaying constant where, \(\zeta \overset{\Delta}{=} (3\ln 10)/(T_{60} f_s)\), and \(T_{60}\) is the reverberation time and \(f_s\) is the sampling frequency. This model depends on parameters which are nearly constant as long as the configuration of the room and objects inside the room remain stationary. Note that this model is only valid for distant speech case which implies that the source to microphone distance is farther than the critical distance, \(D_c\). Also notable that, not necessarily all the rooms present an exponential decay property for the late reverberation envelope. Nevertheless, for most room shapes this model holds, and thus the energy decay format of the AIR\(^2\) envelope takes the following form,

\[
\mathbb{E}\{h^2(n)\} = \sigma^2 e^{-2\zeta n}
\]

(3.34)

where \(\sigma^2\) denotes the variance of \(b(n)\) or reverberation energy density. The solution to dereverberation problem using the statistical model, will be analogous to that of noise suppression method. We assume that the speech in the microphone is a combination of the early and late reflections corrupted with an additive noise, as

\[
x(n) = \sum_{l=n-n_e+1}^{n} x_e(n) s(l)h_e(n-l) + \sum_{l=-\infty}^{n-n_e} x_l(n) s(l)h_l(n-l) + \nu(n)
\]

(3.35)

where \(h_e\) and \(h_l\) denote the early and late reflection of the AIR, respectively. Accordingly, \(x_e\) and \(x_l\) are the early and late speech components, as responses to their associated acoustic transfer functions. Also, \(\nu(n)\) denotes the additive background noise. \(n_e\) is a boundary sample from the acoustic impulse response \(h(n)\), which divides it into early and late reflections, i.e. \(h(n) = \{[h_e(n), \text{for } n = 1, \cdots, n_e - 1], [h_l(n), \text{for } n \geq n_e]\}\). Practically, \(n_e/f_s\) is between 30 to 60 ms. In practice, we desire to reduce the effect of late reverberation speech, as well as the background noise. However, the early reverberation part contributes to the intelligibility of speech by coloration, and is preferred to be existed\(^7\).

To solve the problem in such a statistical framework, similar to the noise suppression algorithms, we set up hypotheses which involve the speech presence and absence, as follows: Similar to the noise suppression method, the suppression is introduced as a gain function over

\(^1\) As it is defined for the late reverberation part
\(^2\) Acoustic Impulse Response
3.3. SINGLE-MICROPHONE REVERBERATION REDUCTION

\( H_{0\omega,n} \): Speech is absent; \( x(\omega, n) = x_l(\omega, n) + \nu(\omega, n) \)

\( H_{1\omega,n} \): Speech is present; \( x(\omega, n) = x_e(\omega, n) + x_l(\omega, n) + \nu(\omega, n) \)

the amplitude of the spectrum of the noisy-reverberant speech, as

\[ \hat{x}_e(\omega, n) = G_{LSA}(\omega, n) |x(\omega, n)| \]  \hspace{1cm} (3.36)

where \( G_{LSA} \) is defined, as in MMSE noise reduction case, with the following parameters:

\[ \zeta \triangleq \frac{\lambda_x}{\lambda_v} = \frac{\mathbb{E}\{s(\omega, n)^2\}}{\mathbb{E}\{\nu(\omega, n)^2\}} \]  \hspace{1cm} (3.37)

\[ \psi(\omega, n) = \frac{\zeta}{1 + \zeta} = \frac{\xi^2}{\lambda_v} \]  \hspace{1cm} (3.38)

where \( \zeta \) denotes the a-priori SNR, \( \psi \) is the integral lower bound as in 3.15 and \( \gamma \) is the a-posteriori SNR value, all depend on \((\omega, n)\). Now, based on the Cohen improved version \([75]\), we can constrain the lower bound of the gain function to avoid distortions, by introducing \( G_{min} \), as well as speech presence probability, \( p(\omega, n) \), which modifies the gain function, as

\[ G_{OM-LSA} = \left\{ G_{LSA} \right\}^p \left\{ G_{min} \right\}^{(1-p)} \]  \hspace{1cm} (3.39)

Habets \([102, 53]\), proposed the following formulation for the estimation of the late reverberation variance, conditioned on the analysis window being stationary over a short period of time (with a duration much less than \( T_{60} \)),

\[ \lambda_r(\omega, n) = e^{-2\zeta(\omega)R} \lambda_r(n-1, \omega) + \frac{\xi_r}{\xi_d} \left( 1 - e^{-2\zeta(\omega)R} \right) \lambda_d(n-1, \omega) \]  \hspace{1cm} (3.40)

where \( \lambda_r \) denotes the spectral variance of the reverberation component, the ratio \( \xi_r/\xi_d \) denotes the inverse of the direct-to-reverberation-ratio (DRR), and \( R \) denotes the number of samples shared between two adjacent frames. \( \zeta(\omega) \) is again the decaying coefficient, in which \( T_{60} \) depends on the frequency. \( \lambda_d \) is the spectral variance of the direct signal, which depends on the model definition parameters, as follows:

\[ \lambda_x(\omega, n) = \lambda_d(\omega, n) + \lambda_r(\omega, n) \]  \hspace{1cm} (3.41)

\[ \lambda_d(\omega, n) = \beta_d(\omega) \lambda_s(\omega, n) \]  \hspace{1cm} (3.42)

where, \( \lambda_i, \; i \in \{x, d, r, s\} \) denotes the spectral variance of any of the indexed signals, and the coefficient \( \beta \) is defined, based on the spectral variance of the acoustic impulse response filter \( \lambda_h(\omega, n) = \mathbb{E}\{|H(\omega, n)|^2\} \), as:

\[ \lambda_h(\omega, n) = \begin{cases} \beta_d(\omega), & \text{for } n = 0 \\ \beta_r(\omega)e^{-2\zeta(\omega)Rn} & \text{for } n \geq 1 \end{cases} \]  \hspace{1cm} (3.43)
By using these pre-calculations, Habets derived the spectral variance of the late reverberation component of the signal, $\lambda_l(\omega, n)$, to be calculated as:

$$\lambda_l(\omega, n) = e^{-2\zeta(\omega)(n_e-R)}\lambda_r \left( n - \frac{n_e}{R} + 1, \omega \right)$$ \hspace{1cm} (3.44)

Now, analogous to statistical denoising method described in the last section, to compute the gain function appropriate for the dereverberation process we only need to convert the a-priori and a-posteriori SNR values to a-priori and a-posteriori signal-to-interference-ratios (SIR), as

$$\xi(\omega, n) = \frac{\lambda_e(\omega, n)}{\lambda_l(\omega, n) + \lambda_v(\omega, n)}$$ \hspace{1cm} (3.45)

$$\gamma(\omega, n) = \frac{|x(\omega, n)|^2}{\lambda_l(\omega, n) + \lambda_v(\omega, n)}$$ \hspace{1cm} (3.46)

Habets\[106\] enhanced the lower bound gain function, $G_{\text{min}}$, to account for time and frequency variations of the spectral variances of both noise and reverberation, and based on his modified $G_{\text{min}}$, the gain function is computed, as

$$G_{\text{min}}(\omega, n) = \frac{G_{\text{min},x,}\hat{\lambda}_l(\omega, n) + G_{\text{min},v,}\hat{\lambda}_v(\omega, n)}{\hat{\lambda}_l(\omega, n) + \hat{\lambda}_v(\omega, n)}$$ \hspace{1cm} (3.47)

and finally, the equation (3.39) would be used to achieve the entire gain function which applying over the corrupted data, will give us the clean speech estimate. The complete algorithm of Habets is mentioned in\[107\]. We notice that, these calculations are based on the joint noise and reverberation presence, for which if we have no noise, which is rare to happen, we can simply remove the dependency for the noise in all derivations, since it has been assumed uncorrelated with the other signals.

### 3.3.4 Harmonicity-based dERverBeration (HERB) method

This algorithm was introduced by Nakatani et al.\[108\], which uses the fact that the reverberation mostly emerges from the voiced phoneme segments of speech, and unvoiced frames are less reverberated and, therefore less influence the subsequent frames. This motivates the idea that, the voiced frames are needed to be dereverberated and cleansed. On the other hand, the voiced frames are constructed based on a harmonic structure with a fundamental frequency specific for each speaker, while the unvoiced frames are almost noise-like. Thus, if the voiced segments are detected correctly and the harmonic structure is re-synthesized, the re-assembled frames will hopefully result a dereverberated speech signal.

This method has several problematic issues. First, is that it needs the fundamental frequency, $f_0$, to be estimated precisely, otherwise the harmonic model will not match the real signal and the rest of the process will be affected. Second, is the dereverberation operator which depends on the unvoiced segments. During an interference presence, the dereverberation operator confuses the voiced and unvoiced segments and the calculations encounter a
3.3. SINGLE-MICROPHONE REVERBERATION REDUCTION

high uncertainty. Moreover, this method is more suitable for offline processing.

3.3.5 Least-Square inverse filtering

One of the most recent works in single channel inverse filtering-based dereverberation is presented by Kodrasi et al. \[109\]. Traditionally, the noise-free time-domain observation signal is defined, as

\[ x(n) = s(n) * h(n) \]

where \( h(n) \) denotes the room impulse response (RIR). The inverse filter, \( g(n) \), with length \( L_g \) should perform so as to cancel the effect of the AIR transfer function, such that

\[ h(n) * g(n) = d(n) \]

where \( d(n) = 1 \), only if \( n = 1 \). Therefore, as a vector \( d = [10 \cdots 0]^T \), and also \( g = [g(0), g(1), \cdots, g(L_g - 1)]^T \). Then, the cost function which seeks the approximation of the AIR filter, will be as: \( J = ||Hg - d||_2^2 \), and the solution for this minimum norm solution would be, as

\[ g = (H^T H)^{-1} H^T d \quad (3.48) \]

However, approximating the RIR filter is not a trivial task. Kodrasi proposed that instead of direct inverting the acoustic transfer function in the frequency-domain, which generally yields instability and non-causality, to use a frequency-domain inverse filtering technique that incorporates regularization and uses a single-channel speech enhancement scheme. This means, assuming that the AIR filter is stationary in time, knowing the approximate subband AIR filter, \( \hat{H}(\omega), \omega \in \{0, 1, \cdots, \Omega - 1\} \), the inverse should be obtained directly from,

\[ G(\omega) = \frac{1}{\hat{H}(\omega)} \]

However, this inverse leads to instability and non-causality. It has been shown that, the poles of inverse filter on the unit circle will result instability and causes undesirable tones in the processed signal. Moreover, for a typical AIR with zeros outside the unit circle, the inverse filter is non-causal, and yields undesirable pre-echoes. Kodrasi decided to manipulate the unstable poles directly, by using the regularizer, \( \delta \), as

\[ G_{\delta}(\omega) = \frac{\hat{H}^*(\omega)}{||\hat{H}(\omega)||^2 + \delta} \quad (3.49) \]

where \( \hat{H}^*(\omega) \) is the conjugate of \( \hat{H}(\omega) \). While the regularization strongly reduces the tones in the processed microphone signal, the synthesized signal \( s_{\delta}(n) \) exhibits pre-echoes due to the remaining non-causality in \( G_{\delta}(\omega) \). In order to reduce the pre-echoes in \( s_{\delta}(n) \), Kodrasi applied a single-channel speech enhancement scheme which estimates the pre-echo power spectral density.
density (PSD) and employs this estimate to compute an enhancement gain function. The desired signal is assumed, as
\[ S_D(k, l) = S_D^d(k, l) + E(k, l) \]  
where \( S_D^d(k, l) \) denotes the direct clean signal after a complete inverse filtering, and \( E(k, l) \) the pre-echoes which are assumed as a non-stationary noise, uncorrelated with the desired signal. The indices, \((k, l)\), represent the frequency and time indices in which \( k \in \{0, 1, \cdots, K - 1\} \), and \( K \ll \Omega \). Now, the noise PSD estimator, \( \hat{\sigma}_E^2(k, l) = \mathcal{E}\{ |E(k, l)|^2 \} \), based on the speech presence probability with fixed priors,\(^{11}\) is employed to estimate the pre-echoes PSD. The penultimate stage would be to estimate the a-priori SNR as in\(^{111}\), using cepstral smoothing,
\[ \hat{\xi}(k, l) = \frac{\mathcal{E}\{|S_D^d(k, l)|^2\}}{\hat{\sigma}_E^2(k, l)} \]  
and the final stage, to apply these estimated values in a Wiener filter, as
\[ G_W(k, l) = \frac{\hat{\xi}(k, l)}{1 + \hat{\xi}(k, l)} \]  
which is applied to the \( S_D(k, l) \), to extract out the desired signal, \( S_D^d(k, l) \). Setting of the parameters in this method is crucial, which in practice the \( \Omega = 16384 \), \( K = 512 \), \( \delta = 10^{-2} \), and 50\% overlap of the frames has shown a reasonably good results.

### 3.4 M-Channel Noise/Reverb Reduction for Speech Enhancement

#### 3.4.1 Introduction

In certain scenarios, multiple recordings of a given source or multiple sources are available through a microphone array, which follows a known geometrical architecture, or a distributed set of microphones with an unknown geometry. One important aspect of multi-channel speech source processing is the ability to use the spatial information of the sources. When the known geometry of the array is informative, one can leverage the multiple recordings facility to reduce the noise or reverberation of the room from the desired signal, in a more sophisticated way. Contrary to the theoretical claim concerning the ability of the microphone array to perform distortionless denoising of the signal\(^{14}\), this never happens in reality. The reason is due to some limitations, such as the number of microphones to be used, the approximations usually considered for the noise pdf and its adverse effect on the phase of the signal, which is mostly ignored in the majority of the algorithms.

As already mentioned in the quick survey, there are early and late reflections of the direct signal from the surfaces of the room. While the early reflections make the resulted signal more intelligible, the late reflections introduce significant distortions, which degrade the speech and

---

\(^{1}\)It has been experimentally validated that this estimator exhibits a fast tracking performance for non-stationary noise, and therefore is appropriate for non-stationary noise estimation.
make the content of the speech difficult to be understood.

A strong motivation of using the microphone-array is the ability it gives us to find the location of the sources in the environment by calculating the direction of arrival from each source with respect to the array. The most widely used method in this regard is called the Generalized Cross Correlation (GCC) which calculates the Time Difference Of Arrival (TDOA) between the source and each of the microphones, and draws a geometric space which is most likely to possess the source. Steered Response Power (SRP) is a different approach, which uses spatial filtering and searches the entire space greedily to locate a zone which yields the highest output energy. While the former is fast enough to be employed in a real-time scenario, the latter can more accurately locate the acoustic sources, specially in a noisy and reverberant environment.

When the configuration of the microphones does not evoke any known geometry for us, the statistical properties of the sources being recorded by different sensors are used. The techniques which leverage all the priors and diversities other than geometrical ones, are classified under the category of Blind Source Separation (BSS) methods. BSS methods could also be used for enhancement purposes, when the interfering sources are considered as independent noise sources. The separation aspect of these methods will be more emphasized in the coming chapter of this thesis.

### 3.4.2 Beamforming - A General Solution

Beamforming is assigned to a set of techniques which try to emphasize a spatially propagated signal from a desired direction, while attenuating other directions. To do so, the beamformer takes the direction of arrival (DOA) of the desired speech source into account, and calculates a set of appropriate gains to be assigned to the microphones of the array. Therefore, a desired spatial gain pattern would be formed, which emphasizes the DOA and attenuates the rest of the angles. Based on whether the computations are done adaptively with respect to the input signal or not, they are classified into data-dependent (adaptive), and fixed beamformers. A "novel interpretation of the beamforming is presented in this section, which elaborates to derive the optimum beamformer weights through the inverse system framework. The general mixture model of the microphone array signal in time domain is modeled as,

\[
x(n) = h(s(n)) + n(n)
\]  

(3.53)

where \( x \) denotes the vector of \( M \) microphone channels, \( s \) is the clean source emitted from the speakers mouth, \( n \) is the multichannel noise which is assumed to be independent from the signal and its samples are independent and identically distributed (iid), and \( h \) denotes a 1 to \( M \), nonlinear function that maps the source data to the microphone channels and represents the propagation function. While the only known part is the observation vector \( x \), estimation of

---

1. Also known as spatial filtering
2. Which is novel, and it is for the first time, to our knowledge
3. Or, the Acoustic transfer function from source to the microphones
CHAPTER 3. SPEECH ENHANCEMENT: SINGLE SPEAKER DSR FRONT-END

the clean source, noise contribution and mapping function from the observation vector alone
is an *ill-posed* problem, which is called *inverse problem*.

In order to find a solution for such an ill-posed problem, we need to consider some re-
laxations and utilize the existing diversities to be able to convert the problem to a well-posed
one and find a unique solution for it. One such relaxation would be to move into an approx-
imate linear operation, which is compatible with the physics of the problem. Assuming that
\( \mathbf{x} \in \{ \mathbf{x}^1, \ldots, \mathbf{x}^T \}; \forall \mathbf{x}^t \in \mathbb{R}^{M \times 1} \), is a *stationary temporal sequence* of the observed multi-channel
data with time length \( T \), the nonlinear mapping operation could be replaced with FIR filters
with proper dimensionality, and then (3.53) can be rewritten as,

\[
x(n) = \mathbf{H}(n) * \mathbf{s}(n) + \mathbf{n}(n)
\]  

(3.54)

where \( \mathbf{H} = [ \mathbf{h}_1, \ldots, \mathbf{h}_M ]; \mathbf{h}_i \in \mathbb{R}^{L \times 1} \), and the operator changes into a linear convolution. If the
length of the \( \mathbf{h}_i \) FIR propagation filters, \( L \), is smaller than the frame size \( T \), then this linear
convolution can accurately approximate the true model. Otherwise, since the computers do
circular convolutions, this operation will result in amplitude and phase distortions. Since, the
speech signal is only stationary for around 30msecs, and the propagation FIR filter length is
always much longer in practice, this operation always ends up with distortions, and this ap-
proximation is not a perfect one, nevertheless being used ubiquitously. This model, however
works decently in practice and is well suited to the speech processing applications. In Fourier
domain, we have

\[
X(n,k) = \mathbf{H}(k) \mathbf{S}(n,k) + \mathbf{N}(n,k)
\]

(3.55)

where \( n \) and \( k \) denote the time and frequency indices, and the propagation matrix, \( \mathbf{H}(k) \),is
assumed to be stationary within the short time period, and only depends on frequency.

Linear approximation of the nonlinear mapping in (3.53) introduces an error which can be
denoted as, \( \mathbf{e} = \mathbf{X} - \mathbf{H} \mathbf{S} \). Minimization of this error function, converts the source estimation into
an optimization problem. Assuming that the random noise \( \mathbf{N} \sim \mathcal{N}(0, \mathbf{R}_{nn}) \) follows a Gaussian
pdf, with zero mean and the covariance matrix, \( \mathbf{R}_{nn} \), then for every frequency bin \( (\forall \omega_k) \):

\[
\mathbf{X} \sim \mathcal{N}(\mathbf{H} \mathbf{S}, \mathbf{R}_{nn})
\]

\[
\mathbf{e} = \mathbf{X} - \mathbf{H} \mathbf{S} \sim \mathcal{N}(0, \mathbf{R}_{nn})
\]

(3.56)  
(3.57)

and minimizing the squared norm of the error\(^1\) is performed, due to the smoothness of the \( \ell^2 \)
norm and its differentiability. Moreover, the squared of the norm gives the same solution as
the norm itself, therefore we can find the *variance normalized* least squared solution of the \( \ell^2 \)

\(^1\)Which is called as *Least Squared* optimization problem

\(^2\)The resulting error function follows the unit Gaussian pdf, \( \mathcal{N}(0, \mathbf{I}) \)
3.4. M-CHANNEL NOISE/REVERB REDUCTION FOR SPEECH ENHANCEMENT

\[
J(S) = \|X - HS\|_{R_{nn}^{-1}}^2
= (X - HS)^T R_{nn}^{-1} (X - HS) \tag{3.58}
\]
\[
= \underbrace{X^T R_{nn}^{-1} X}_{\text{const.}} - 2\underbrace{X^T R_{nn}^{-1} HS}_{\text{Linear}} + \underbrace{S^T (H^T R_{nn}^{-1} H) S}_{\text{Quadratic}} \tag{3.59}
\]

By vanishing the gradient of (3.59) with respect to \(S\) and extracting the stationary points, we get the normal equation, as

\[
(H^T R_{nn}^{-1} H) S = H^T R_{nn}^{-1} X \tag{3.60}
\]

When matrix \(H\), is symmetric and positive definite (P.D.) then the solution of the normal equation would be unique. Since \(H^T H\) is symmetric, and matrix \(R_{nn}^{-1}\) is symmetric and P.D., then \((H^T R_{nn}^{-1} H)\) is also symmetric, and if \(H^T H\) is P.D., then \((H^T R_{nn}^{-1} H)\) would be P.D., as well. Therefore, the condition which makes the inverse problem well-posed and solvable is that the propagation matrix \(H\) is full-rank or the columns of it are linearly independent. Otherwise, there could be multiple minimum points for the optimization problem (3.59), and the answer would be ambiguous. Since (3.59) is convex, the solution, would be the global minimum.

The solution to this problem could be investigated in a general case, however since here we discuss the beamformer as a speech enhancement problem, the condition is over-determined, meaning that the number of observations is to be more than the sources. Assuming that \(H\) is full-rank, because there is only one source considered, \(\text{rank}(H) = \min(M, 1) = 1\). Therefore, \(H^T H\) is non-singular and has inverse\(^1\) and so does \(H^T R_{nn}^{-1} H\), due to the symmetric positive definiteness (SPD) property of \(R_{nn}^{-1}\). However, \(H H^T\) has singularity points, and may not be invertible. Thus, the estimated source is extracted as [112],

\[
\hat{S} = (H^T R_{nn}^{-1} H)^{-1} H^T R_{nn}^{-1} X \tag{3.61}
\]

The above equation, clearly indicates the gain (i.e. \(\hat{S} \approx W_{BF} X\)) which should be applied to the observation data vector in order to achieve the optimum source estimate.\(^3\) Since \((H^T R_{nn}^{-1} H)\) is a scalar, the best linear beamformer from the inverse system viewpoint will be, as

\[
W_{BF} = \frac{H^T R_{nn}^{-1}}{H^T H R_{nn}^{-1} H} \tag{3.62}
\]

Obtaining the optimum weights in (3.62) entails knowing of two parameters, namely \(H\), which is the propagation matrix, and \(R_{nn}^{-1}\), which is the covariance matrix of the noise. In conventional beamforming methods \(H\) is approximated, by the array manifold (or steering) vector. In fact, in beamforming it is assumed that the sound sources introduce time differences of arrivals (TDOAs) in relation to their position with respect to the array. Let us consider a plane wave

\(^1\)It is the property of every covariance matrix to be S.P.D.

\(^2\)In fact, in this case the inverse is a scalar.

\(^3\)In the existing literature exactly the same solution is extracted by the Minimum Variance Unbiased estimator (MVDR), as the Best Linear Unbiased Estimate (BLUE), and using the Lagrange multipliers
approaching the array aperture from direction
\[
a = \begin{bmatrix}
\cos \phi \sin \theta & \sin \phi \sin \theta & \cos \theta
\end{bmatrix}^T
\] (3.63)

with azimuth \( \phi \) and elevation \( \theta \). Then, using the far field assumption, the delay which is introduced at the \( i \)-th microphone position, \( m_i \), in relation to the array center is \( \tau_i = -a^T m_i / c \), where \( c \) denotes the speed of sound. Translating these delays to phase shifts in the frequency domain leads to the so-called \textit{array manifold vector} \( v \), which depends on the sampling rate:
\[
v(\omega) = \begin{bmatrix}
e^{-j\omega\tau_1} & \ldots & e^{-j\omega\tau_M}
\end{bmatrix}^T
\] (3.64)

Hence, the vector \( v(\omega) \) is assumed as a complete summary of the interaction of the array geometry with the propagating wave, and replaces \( H \) in (3.62). Therefore the optimum beamformer from inverse system viewpoint, is derived as,
\[
W_{\text{opt}} = \frac{v^T R_{nn}^{-1} v}{v^T R_{nn}^{-1} v} v
\] (3.65)

The \textit{Minimum Variance Distortionless Response} (MVDR) beamformer, minimizes the variance of the linear estimate of the source, i.e. \( \hat{S} = W^H X \), subject to the condition that this estimate is unbiased, i.e. \( W^H v = 1 \). The solution to the optimum MVDR beamformer is exactly as in (3.65) \cite{113}. The problem is defined in the convex optimization framework, as
\[
W_{\text{MVDR}} = \arg\min_W W^H R_{nn}^{-1} W \quad \text{subject to: } W^H v = 1 \quad (3.66)
\]

When the noise field is diffuse \cite{2}, then the covariance matrix of the noise is a factorization of a constant noise power, \( \Phi_{nn} \), and the noise coherence matrix in the microphones, \( R_{nn} = \Phi_{nn} \Gamma_{nn} \). On the other hand, the elements of the noise coherence matrix, \( \Gamma_{nn} \), in a spherically isotropic noise field (i.e. diffuse) can be derived from a \textit{sinc} function, as
\[
(\Gamma_{nn})_{ij}(\omega) = \text{sinc} \left( \frac{||m_i - m_j||}{c} \right) \omega
\] (3.67)

which these values could also be used in the optimum beamformer in (3.65), as well. Solution of the MVDR beamformer (or our inverse system solution), while the noise field is assumed to be diffuse is called as, \textit{super-directive} beamformer.

We can add a regularization term to the objective function in (3.66) so as to incorporate the sparsity condition \cite{3} of the estimated source to the objective function, using the \( \ell - 1 \) norm, as
\[
W_{\text{sp-MVDR}} = \arg\min_W \left\{ W^H R_{nn}^{-1} W + \lambda ||W^H X||_1 \right\} \quad \text{subject to: } W^H v = 1 \quad (3.68)
\]

\footnote{The straightforward derivations of the optimum MVDR and Superdirective beamformer are in the appendix-A.1}

\footnote{For which the definition and associated properties are already mentioned in chapter 2}

\footnote{One of the diversities that could be exploited for the speech signals, is their sparsity in the Fourier domain, as it was shown in chapter 2, and norm-1 of a vector represents its sparsity}
3.4. M-CHANNEL NOISE/REVERB REDUCTION FOR SPEECH ENHANCEMENT

There are other possible regularization terms to be replaced the sparsity or being used along-side. Speech signal statistically has a peaked probability distribution (pdf), and applying a beamformer (BF) weight which yields the highest Kurtosis\(^1\) leads us to the peakedness property. However, for the Kurtosis to be calculated we need to know the pdf of the speech signal. This would not be available in an online processing system. Therefore, we can only use a sample Kurtosis as an approximation. Furthermore, Random samples associated to a uniform pdf, tend to be the most informative data received at a sensor. Applying a BF weight that yields the minimum entropy, as a measure of information content of a signal, may provide an outcome which is closer to a meaningful speech signal, in terms of the information. It is worth mentioning that, the regularization term makes us solve the optimization problem in (3.68) by the convex optimizer tool\(^2\) and that leads to an exhausting runtime procedure. That is a good reason to resort the closed form solution of (3.65), and let the more enhancements being performed in further processing blocks.

There are practical problems regarding the beamforming derivations and the assumptions we made, which are enumerated as follows:

1. For the high frequencies, in which the distance between a pair of microphones is more than half of the input signal wavelength, the notion of spatial aliasing occurs which creates large spurious sidelobe peaks in the gain pattern of the beamformer, but in the directions which are not associated to the desired source. This can easily allow the undesirable noise, reverberation and interference signals to be accumulated with the signal of interest and have adverse effects on the output quality and intelligibility. The spatial aliasing could be interpreted analogously to the aliasing in the signal sampling theorem, except that it occurs for the spatially propagated signal in the space.

2. For the high frequencies, when the distance between a pair of microphones is more than half of the input signal wavelength, the estimation of the Time Difference Of Arrival (TDOA) from that pair encounters the phase ambiguity, since the Inter-channel Phase Difference (IPD) wraps around. Therefore, the GCC\(^3\) function may yield some spurious peaks, which cause finding the true peak (associated to the source location) difficult.

3. For the low frequencies, the noise signals at the microphones are highly correlated and the diffuse noise field assumption is not an appropriate model. Therefore, the beamforming (BF) weights which are derived based on the coherence matrix of the noise (e.g., super-directive BF) for low frequencies are erroneous.

4. For a low SNR\(^4\) scenario in an environment which is highly reverberated several spurious peaks occur in the GCC function, and that makes the source locating problem very challenging.

---

1 Which Kurtosis is the normalized fourth-order cumulant of the pdf of a random variable
2 Such as cvx for MATLAB users
3 Generalized Cross Correlation
4 Signal-to-Noise-Ratio
Apart from the aforementioned limitations of the beamformers, they are widely used in practice due to their potential power to increase the SNR, reduce the reverberation rate, and enhance the intelligibility of the speech. Needless to say, that for the far-field sources, as in the DSR problem, knowing the accurate DOA is very critical and achieving a good accuracy is not a trivial task in a practical harsh environment. Therefore, a beamforming problem inherently carries two sub-problems of DOA estimation, and source extraction. However, for the multi-speaker scenarios even extra problems pop up. These problems consist of source enumeration in the environment, and in case of moving sources their individual tracking over time.

3.4.3 Multi-Channel NMF-/NTF-based enhancement

The multi-channel NMF is an extension of the conventional NMF which elaborates using the redundant information of the multiple microphones in an optimal way, to extract and enhance the underlying signal [114, 115]. The Non-Negative Tensor Factorization (NTF) can also be leveraged as a straightforward paradigm of analyzing the data in a high dimensional case. While in an NMF the spectrogram, as a two-dimensional input data (i.e. time and frequency), is factorized into non-negative components (see 3.20), by including a new dimension which represents the extra available observations through different microphones (channels), the problem could be casted in an NTF paradigm [116]. In figure 3.9 a multichannel input spectrogram data $X$ is factorized into three matrices of $A$, $S$, and $D$, which point to the $R$ number of dictionary bases, corresponding basis activation weights over time $T$, and the $R$ number of activation vectors corresponding to the basis dictionary across the $k$ recording channels, respectively. This is only one demonstration of the tensor $X$, and there are several other ways to represent a tensor decomposition, such as the raw-wise, column-wise, rank-one, and so on [116, 91]. Tensor $E$ also represents the error of factorization.

Generalization of the NMF to a three-way tensor, enables us to incorporate a spatial cue associated to a speaker and adapt the NMF algorithm to a point source. Since a spatial cue indicates which frequency bins of the spectrogram are important, it would be possible to give higher weights to the specific bins whose associated target is more likely to exist [117].

While most of the extended-NMF algorithms are proposed for source separation, there are few techniques which were reported as for the enhancement purposes. An NTF-based dereverberation has been performed by Mirsamadi et al. [118] which uses a multi-channel convolutive NMF using an NTF structure. The convolutive NMF is in turn an extension of the normal NMF, which is capable of identifying components with temporal structure. In this type of NMF, decomposition is performed as [119],

$$Y \approx \sum_{t=0}^{T-1} B_t \cdot W$$  \hspace{1cm} (3.69)

where $Y$ is as usual the spectrogram of the observed signal wished to be decomposed, $B_t$ is the basis matrix at time $t$, and $W$ represents the weight matrices being involved in the convolution
3.4. M-CHANNEL NOISE/REVERB REDUCTION FOR SPEECH ENHANCEMENT

Figure 3.9: A Tensor representation of a multichannel signal spectrograms

sum after shifting $t$ spots to the right and inserting zero column vectors from leftmost columns to the matrix $W$. The complete derivation is quite similar to the standard NMF and could be followed from [119, 120].

As mentioned in [118, 121], the spectral smearness of reverberation\(^1\) can be modeled using the convolutive combination of the room impulse response (RIR) and the speech signal, as follows:

$$Y^{(i)}(n, k) = \sum_{p=0}^{L-1} H^{(i)}_k(p) S(n-p, k)$$  

(3.70)

where $Y^{(i)}$ and $S$ are the magnitude Short Time Frequency representation of the observation signal in microphone $i$ and the clean source, respectively, with indices $n$ and $k$ corresponding to the time and frequency bin, and $L$ is the length of the RIR filters $H^{(i)}_k(p)$. The subband envelope of the RIR from the source location to $i$’th microphone is represented by $H^{(i)}_k$. The goal of the algorithm is then, to find the nonnegative factors $\hat{H}^{(i)}_k(n)$ for all channels, together with the common nonnegative factor $\hat{S}(n, k)$, which jointly minimize the error criterion between the reverberant signals $Y^{(i)}$, and their approximations $Z^{(i)}(n, k) \triangleq \hat{H}^{(i)}_k(n) * \hat{S}(n, k)$. They used the Euclidean distance error criterion, as

$$E = \sum_{k,n} ||Y(n, k) - Z(n, k)||^2_F,$$

subject to:

$$H^{(i)}_k(n) > 0, \ S(n, k) > 0, \ \sum_{i=0}^{M} \sum_{p=0}^{L-1} H^{(i)}_k(p) = 1, \ k = 1, \cdots, K$$

\(^1\)Since the RIR length is longer than the frame length, by the convolution operation the spectral content of each frame influences the subsequent frames.
where the cost function is the Frobenius norm, and denotes the sum of all the error terms associated with each individual microphone channel. \( H_k^{(i)}(n) > 0 \) and \( S(n, k) > 0 \), are the nonnegativity constraints for all \( i, p, n, k \), together with another constraint for scale indeterminacy, where \( M \) is the number of microphones (channels). A variable step size iterative method is used in a multiplicative update framework and ensures the nonnegativity of the results \[122]\.

By constraining the base matrices, \( H_k^{(i)}(p) = \text{diag}(H_k^{(i)}(p), \cdots, H_K^{(i)}(p)) \), to be diagonal, and vanishing the gradient of the error function (3.71) with respect to the quantities \( H_k^{(i)}(p) \) and \( S(n, k) \), in a multiplicative update rule we have the estimates as \[122],

\[
\hat{H}_k^{(i)}(p) \leftarrow \hat{H}_k^{(i)}(p) \frac{\sum_n Y^{(i)}(n, k) \hat{S}(n-p, k)}{\sum_n Z^{(i)}(n, k) \hat{S}(n-p, k)} \quad (3.72)
\]

\[
\hat{S}(\ell, k) \leftarrow \hat{S}(\ell, k) \frac{\sum_i \sum_n Y^{(i)}(n, k) \hat{H}_k^{(i)}(n-\ell)}{\sum_i \sum_n Z^{(i)}(n, k) \hat{H}_k^{(i)}(n-\ell)} \quad (3.73)
\]

The experiments have shown a significant improvement of the Word Error Rate (WER) for a distant source compared to the case where no dereverberation was performed \[118\].

### 3.5 Experiments

The speech enhancement algorithms described in this chapter can improve speech quality but not necessarily the speech intelligibility \[77\], which is the goal of the automatic speech recognition (ASR) system. Therefore, we need to evaluate the previously mentioned algorithms specially in real conditions, in which the desired signal is combined with different noise and reverberation conditions with various SNR levels to figure out the feasible algorithms which can substantially improve the intelligibility of speech in practical scenarios.

When the background noise is stationary (e.g., car noise), then voice activity detection (VAD) or noise estimation can generally perform well, whereas in nonstationary type of noise (e.g., multitalker babble) they are erroneous. Furthermore, the majority of algorithms introduce distortions, which might be more harmful to the intelligibility than the background noise. Therefore, there is a trade-off between the amount of noise reduction and minimization of speech distortion.

we performed our experiments on the Multi-Channel Wall-Street-Journal Audio-Visual (MC-WSJ-AV) corpus \[123\]. This corpus consists of 352 utterances spoken by 10 speakers, with a total recording length of 40 minutes, and has been recorded in a real room condition. The performance measures, are both the ones used for the quality measurement, as well as some measures which are highly correlated with the intelligibility score. The best way to evaluate the validity of the algorithms for the case of intelligibility is to perform the Mean-Opinion-Score (MOS) for the human perception evaluation and the Word-Error-Rate (WER) for the ASR system evaluation, and both of these measures are very difficult to become available for every
situation or at least to be modified easily. There are some studies that describe the legitimate conditions for the MOS measures to become valid as mentioned in ITU-R BS.562-3 standard [124].

The measures which are used to evaluate the quality of speech has some degree of correlation with the intelligibility measures on ASR systems (measured by WER), which the most correlated ones are used in our study, and the complete list of them along with their correlation coefficient can be found in [125, 126]. The performance of the implemented speech enhancement algorithms was evaluated by calculating perceptual evaluation of speech quality (PESQ) [77], Short-Time Objective Intelligibility measure (STOI) [127, 128], Speech to Reverberation Modulation Energy Ratio (SRMR) [45, 70], frequency-weighted Segmental Signal-to-Noise-ratio (fwSegSNR) [126], Itakura-Saito distance measure (ISdist), Log-Likelihood ratio (LLR), and Cepstral Distance (CD) as mentioned in [125]. All the distance measures, reflect the naturalness of the speech signal compared to the clean speech recorded by a close-talking microphone, but from different aspects which their theoretical justification are explained in detail in [129].

The most important algorithms from each category, whose theoretical background has been briefly explained in this chapter have been chosen, implemented on the above-mentioned corpus, and the results are compared with the stated measures of quality and intelligibility. These algorithms consists of:

1. **Single channel** enhancement algorithms, including:
   - Statistical log-mmse algorithm, section 3.2.3
   - log-mmse algorithm with Speech Presence Uncertainty (SPU), section 3.2.4
   - Spectral Subtraction, with minimum statistics-based noise estimation, section 3.2.4
   - Spectral Subtraction with iMCRA noise estimation 3.2.4
   - multiband Spectral Subtraction algorithm, section 3.2.1
   - Wiener Filter, section 3.2.2
   - Binary Mask channel selection algorithm, section 3.2.5
   - Single channel weighted Linear Prediction based dereverberation [96, 99, 130]
   - Single channel unsupervised NMF based denoising, section 3.2.6
   - Single channel unsupervised convolutive NMF dereverberation algorithm [119]

2. **Multi-channel** enhancement algorithms, including:
   - Super-directive Beamforming, section 4.2
   - Multi-channel Linear Prediction based dereverberation [96]
   - Dual channel Coherent-to-Diffuse-Ratio based dereverberation algorithm [131]
   - Multi-channel NMF based enhancement, section 3.4.3
   - Multi-channel Wiener-filter [132]
The first measure which has been compared for the listed algorithms is the Short-Time-Objective-Intelligibility measure (STOI), proposed in [127, 128] recently and has a high correlation with the ASR intelligibility measure which is obtained by the Word Error Rate (WER). The higher this value, the better an ASR system can recognize the speech. Figure 3.10 shows this measure for the single-channel based algorithms, whereas figure 3.11 reveals the results of STOI on multi-channel enhancement algorithms. It is obvious from these two figures, that

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>STOI Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>log-mmse</td>
<td>53</td>
</tr>
<tr>
<td>log-mmse-SPU</td>
<td>64</td>
</tr>
<tr>
<td>Min-Statistics Spectral Subtraction (R. Martin)</td>
<td>73</td>
</tr>
<tr>
<td>iMCRA (I. Cohen)</td>
<td>72</td>
</tr>
<tr>
<td>mBand (Spectral Subtraction)</td>
<td>71</td>
</tr>
<tr>
<td>BinaryMask</td>
<td>72</td>
</tr>
<tr>
<td>Wiener Filter</td>
<td>73</td>
</tr>
<tr>
<td>NMF-denoise</td>
<td>69</td>
</tr>
<tr>
<td>NMF-dereverb.</td>
<td>23</td>
</tr>
<tr>
<td>WPE (Linear Prediction)</td>
<td>84</td>
</tr>
<tr>
<td>Input Data- no process</td>
<td>73</td>
</tr>
</tbody>
</table>

Figure 3.10: STOI measure for single-channel enhancement algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>STOI Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beamforming (SDB)</td>
<td>81</td>
</tr>
<tr>
<td>MC-Wiener Filter (SDB+PostFilter)</td>
<td>79</td>
</tr>
<tr>
<td>MC-Linear Prediction (WPE)</td>
<td>84</td>
</tr>
<tr>
<td>2-Ch CDR (Coherence-to-Diffuse) dereverb.</td>
<td>78</td>
</tr>
<tr>
<td>MC- NMF</td>
<td>73</td>
</tr>
</tbody>
</table>

Figure 3.11: STOI measure for multi-channel enhancement algorithms.

the single channel algorithms have not been very successful in improving this intelligibility measure, compared to the baseline system. In fact, many of the algorithms of single-channel
classes, degrade the intelligibility due to a distortion they produce for the speech signal. In many cases, this distortion is even more deteriorating for the signal than the contained noise. The only successful system in this class, is the weighted linear-prediction based enhancement, which linearly removes the reverberation based on long-term prediction filtering. The only drawback of this system is that, it needs to have access to several blocks of the data to predict the true signal in the current frame, in order to achieve a good result. This experiment has been accomplished based on 30 blocks of previous data, for the prediction.

On the other hand, from figure 3.11, it can be seen that all the multi-channel algorithms have been achieved a higher intelligibility score than the baseline. The super-directive beamformer (SDB), as the one with one of the highest scores has actually one of the lowest computational costs among all of the others which makes it a perfect choice for real-time systems, yet achieving a high intelligibility score only based on the spatial diversity of the sources in the space. This conforms the fruitfulness of using the spatial location information of the sources and the array geometry. Multi-channel linear prediction actually does not gain too much compared to its single channel case, yet has the highest score. This may hinder us from imposing the high computational cost for a multichannel prediction, and justifies to have the single channel linear prediction, instead when it is required. The multi-channel Wiener filter, even though is composed of a SDB beamformer and a further post filter to denoise the signal of the beamformer output, achieves a lower STOI score. The output signal sounds more comfortable than the SDB output alone, but the score is not revealing an improvement. The WER results of the ASR system (i.e., WER) showed an improvement of about 7% in the overlapping speaker case, which has been shown in next chapter. This might be interpreted that the small difference between the scores might night be considered as serious, until the ASR decoding experiment is accomplished.

The next measure, is the perceptual estimate measure for the speech quality (PESQ). This is one of the most validated measures of the speech quality, in the community. This measure truly holds the quality performance, however has been investigated to have about 0.79% correlation with the ASR system score, WER. Figure 3.12 obviously shows that the multichannel enhancement algorithms outperform the single channel algorithms. Combination of the SDB beamformer, and post filter again does not show any quality improvements, however as already mentioned will improve the WER. Multichannel linear prediction still outperforms the others. The next score is SRMR (Speech to Reverberation Modulation Ratio), which uses the idea of the temporal envelope modulation of the signal as a cue for the objective quality and intelligibility estimation. It is known, that for clean speech, temporal envelopes contain frequencies ranging from 2 to 20 Hz with spectral peaks at approximately 4 Hz which corresponds to the syllabic rate of spoken speech. The diffuse reverberation tail, in a reverberant signal is often modeled as an exponentially damped Gaussian white noise process. When the reverberation level increases, the signal would attain the properties more like a white Gaussian noise. It can be expected that reverberant signals exhibit higher-frequency temporal envelopes due to the
"whitening" effect of the reverberation tail. This idea has been incorporated in SRMR, to investigate the level of reverberation contained in the signal. Figure 3.13 shows the responses of the algorithms to the SRMR measure. The scores show the equal ability of the two classes of algorithms to reduce the reverberation level. The multi-band spectral subtraction (mBand) method deploys the spectral subtraction over the sub-channels of the signal which has been decomposed using a Mel-filter bank, and separately weighted based on the properties of the human audio perception system [77]. This algorithm, shows a competitive ability in reverberation reduction. CDR method is a binaural method which estimates the ratio between the coherent and diffuse noise and then uses a subsequent minima tracking in order to increase the estimation accuracy [131]. This method has been proven itself successful in both achieving a high quality (i.e., PESQ score), as well as high reverberation reduction performance (i.e., SRMR score).

One of the most relevant measures to the ASR system intelligibility score (i.e., WER) is the frequency-weighted-Segmental SNR (fwSegSNR). This measure has a very high correlation with ASR score (e.g., about 0.81%). Figure 3.14 shows the results of the algorithms of the two classes in reducing the local noise level, over the full band of speech which is demonstrated by fwSegSNR measure. The multi-channel algorithms clearly outperform the single-channel ones. Here, the DOA-based methods (i.e., SDB beamforming and its counterpart multi-channel...
3.6. THE PROPOSED ENHANCEMENT STRUCTURE

Wiener filter which has a combined SDB-PF[structure] show their power of noise reduction, in spite of their implementation simplicity. Cepstral distance (CD) mainly reflects the naturalness of speech based on its similarity in properties with a human auditory system, which performs in a logarithmic domain. For the distance measures the lower value reveals the better performance of the algorithm, and in this case the closer the signal to a natural sound. Again, the multi-channel methods outperform their single-channel counterpart in all cases, as demonstrated in figure 3.15.

The log-likelihood ratio (llR) also performs as a measure of distance between the signal and its clean version. Therefore, the lower ratio reflects a better performance. Figure 3.16 again shows the superior performance of the multi-channel algorithms compared to their single-channel counterpart.

3.6 The Proposed Enhancement Structure

Considering that the goal of our enhancement system is to achieve a higher intelligibility score with an ASR system, rather than obtaining a high quality signal, and by observing the results of different measures, which already explained in the previous section, we proposed a new
Figure 3.14: frequency-weighted Segmental SNR, which has a high correlation with intelligibility ($r \approx 0.81$).

Figure 3.15: Cepstral Distance as a measure of speech naturalness. The less the value the closer to a natural speech.
3.6. THE PROPOSED ENHANCEMENT STRUCTURE

Figure 3.16: log-likelihood Ratio (llR), a measure of speech naturalness. The less the merrier.

structure which can employ the best characteristics of different methods to achieve the higher intelligibility score than individuals. Figure 3.17 shows the conceptual reasoning, as well as the functional blocks which are to be used in our proposed system. In the proposed system, the initial stage performs as to convert the echoic condition to anechoic, by shortening the room impulse response, through linear prediction filtering. Assuming that the initial signals are a combination of the anechoic and the late reverberation parts, the linear prediction dereverberation system will linearly reduce the signal late reverberation, which represents the long tail of the room impulse response. Therefore, the outputs of the first stage will contain approximately the anechoic signal. However, there are two other advantages which are gained by using the single channel linear prediction based dereverberation for each input channel:

- Since the system is linear, further processing of the signal will introduce less distortion to the signal, than the nonlinear processors. That this method (i.e., WPE) can achieve a high STOI score, implies that the distortion level should be low enough to preserve the intelligibility. In addition, achieving almost low distance scores, also confirms our idea.

- Using the single-channel WPE dereverberation, preserves the directivity information of the input data (i.e., DOA of the sources), since the linear prediction performs a linear process on the late reverberation part. Meanwhile, a similar process should be performed for all the channels, which preserves the DOA information for further exploitations.

Since the DOA information is preserved, we can employ a multi-channel Wiener filter, for which the score graphs clearly show its superiority in noise reduction compared to the other algo-
Figure 3.17: The proposed enhancement system, as a combination of Dereverberation-Beamforming-Denoising system. Multi-channel Wiener filter is actually a combination of an MVDR beamformer (e.g., Super-directive BF) plus a Postfilter. Modified Zelinski postfilter have been shown to perform perfectly, with noise-overestimation factor (A simple modification to the Zelinski postfilter using this parameter, has shown a significant improvement of ineligibility score, WER), $\beta = 0.5$ for WSJ-AV-corpus. The above graph shows the functionality of the system, which has been developed as in the bellow graph.

Thus, we would be able to achieve a good reverberation reduction, as well as noise suppression. Multi-channel Wiener filter is a combination of an MVDR beamformer cascaded by a Wiener filter [132]. We have used a SDB beamformer which is a special case of MVDR solution, when the noise field is diffuse. However, this is not an optimum solution, since the anechoic condition changes the noise field and the true solution would be achieved by replacing the covariance matrix of the noise, instead of the noise coherence matrix, as in (3.62). The results of the best algorithms, as well as our new structures are presented in table 3.1. The bottom line, relates to the input data, without any processing being applied. WPE-BF-PF, denotes the proposed structure of the figure 3.17 which combines a network of single-channel weighted linear prediction based dereverberator to achieve a high reverberation reduction, as well as converting the echoic impulse response of the room to anechoic, followed by a multi-channel Wiener filter to achieve a high noise suppression.

---

1 Minimum Variance Distortionless Response, which is the Best Linear Unbiased Estimate to the enhancement problem, as derived in appendix. A1.
3.6. THE PROPOSED ENHANCEMENT STRUCTURE

Table 3.1: Results of the best enhancement methods, compared with the most relevant measures to the speech intelligibility. The best results are in bold. The above section of the table, shows the results of the single channel algorithms, whereas the second and third sections show the multi-channel case results, as well as the proposed systems. The bottom section associates the input unprocessed data. Spectrograms of the best methods are depicted in figure 3.19.

<table>
<thead>
<tr>
<th>measure vs. algorithm</th>
<th>STOI</th>
<th>PESQ</th>
<th>fwSegSNR (dB)</th>
<th>SRMR (dB)</th>
<th>IIR</th>
<th>CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBand (specsub)</td>
<td>0.71</td>
<td>1.99</td>
<td>5.73</td>
<td>6.22</td>
<td>0.87</td>
<td>4.57</td>
</tr>
<tr>
<td>Min. Statistics</td>
<td>0.72</td>
<td>2.02</td>
<td>5.19</td>
<td>4.32</td>
<td>0.91</td>
<td>4.43</td>
</tr>
<tr>
<td>Binary Mask</td>
<td>0.73</td>
<td>2.02</td>
<td>5.24</td>
<td>4.92</td>
<td>1.00</td>
<td>4.91</td>
</tr>
<tr>
<td>1-ch WPE</td>
<td><strong>0.84</strong></td>
<td>2.38</td>
<td>6.09</td>
<td>5.64</td>
<td>0.72</td>
<td>4.27</td>
</tr>
<tr>
<td>SDB - beamform</td>
<td>0.81</td>
<td>2.33</td>
<td>4.92</td>
<td>4.66</td>
<td>1.05</td>
<td>4.28</td>
</tr>
<tr>
<td>SDB - PF</td>
<td>0.79</td>
<td>2.34</td>
<td><strong>6.35</strong></td>
<td>4.97</td>
<td>0.67</td>
<td>4.16</td>
</tr>
<tr>
<td>MC-WPE</td>
<td><strong>0.84</strong></td>
<td>2.39</td>
<td>6.12</td>
<td>5.65</td>
<td>0.71</td>
<td>4.26</td>
</tr>
<tr>
<td>2-Ch CDR</td>
<td>0.78</td>
<td>2.24</td>
<td>6.27</td>
<td><strong>7.86</strong></td>
<td>0.71</td>
<td>3.97</td>
</tr>
<tr>
<td>WPE-BF</td>
<td>0.79</td>
<td>2.42</td>
<td>4.68</td>
<td>5.10</td>
<td>1.06</td>
<td>4.33</td>
</tr>
<tr>
<td>WPE-BF-PF</td>
<td><strong>0.80</strong></td>
<td>2.47</td>
<td>6.23</td>
<td><strong>5.95</strong></td>
<td><strong>0.64</strong></td>
<td><strong>3.93</strong></td>
</tr>
<tr>
<td>BF-PF-WPE</td>
<td><strong>0.83</strong></td>
<td>2.53</td>
<td><strong>6.92</strong></td>
<td><strong>6.18</strong></td>
<td><strong>0.63</strong></td>
<td><strong>4.02</strong></td>
</tr>
<tr>
<td>no-processed data</td>
<td>0.72</td>
<td>1.96</td>
<td>5.37</td>
<td>3.70</td>
<td>0.82</td>
<td>4.53</td>
</tr>
</tbody>
</table>

Another proposal would be as in figure 3.18, which employs a multi-channel Wiener filter, followed by a single channel weighted linear prediction dereverberator. What is important is that, the multi-channel Wiener filter is the optimum noise reduction system in the diffuse noise field. This system has been derived as a cascade of an MVDR beamformer (BF), with a post Wiener filter and as we stated before the postfilter (PF) can be designed as Zelinski proposed. Once the optimum system is employed in its right place, we can expect the noise reduction to be the optimum compared to the other cases. On the other hand, the diffuseness and statistics of the noise, after it has been processed will change significantly. When a dereverberation system comes first, the output noise field will not be diffuse anymore, and the optimality of the BF-PF structure will be violated. This is the reason that justifies the use of the final structure in figure 3.18. The result of the last line (penultimate line, i.e., BF-PF-WPE) clearly shows that the outcome of this structure has superiority to the other methods, although still competitive with the system in figure 3.17. We can obviously see the superior results of this structure, which outperforms the other methods in PESQ and the naturalness measures, while having very high scores of noise and reverberation suppression and intelligibility measures. The spectrogram of the processed files, clearly shows that the proposed system, enhances the speech file while preserving the speech components. Our proposed systems show to have the same level of output naturalness, however the second system has more level of noise and reverberation suppression, and therefore the quality measure (PESQ) is higher.

---

1As, it is derived in appendix-A 1.
2Minimum Variance Distortionless Response
Figure 3.18: A proposed multichannel Wiener filter, which is a cascaded MVDR/Wiener filter, combined with single channel weighted linear prediction based dereverberator.

Figure 3.19: Spectrogram of the processed speech files. From top to bottom: The proposed system from figure [3.17] Multi-channel WPE dereverberation method, Single-channel WPE dereverberation, m-Band Spectral Subtraction, Binary Mask. The vertical axis denotes the frequency bins, whereas the horizontal axis denotes the time. The proposed system, has best preserved the speech components while reducing the noise and reverberation.
Chapter 4

Speech Separation: Multi-Speaker DSR
Front-End

4.1 Introduction

Early sound capturing systems for hands-free speech recognition \[133, 134\] aimed at acquiring a high-quality speech signal from a single, distant speaker. When the research focus shifted towards meeting recognition \[135\], it turned out, however, that overlapping speech is far more difficult to handle. In the presence of an overlapping speaker, the conditions of the separation problem in a room environment gets more difficult. In hostile environments, in order to deploy ASR systems it is necessary to cope with multiple speech and noise sources, whereas the state-of-the-art ASR systems are only trained on clean and single talk condition. Such ASR models will confront inevitably serious problems in noisy multi-talker cases.

When the number of overlapping talker signals exceeds a threshold, the interference is called the \textit{babble noise} (or Cafeteria noise), whereas in case of only few interfering utterances it is named the \textit{cocktail-party} problem. There are different frameworks employed to deal with each of these two conditions. This type of noise/interference is uniquely challenging because of its time evolving statistical properties and its similarity to the desired target speech. Li and Lutman \[136\] modeled the changing of the kurtosis as a function of speakers in babble. Their study explored the characteristics of babble, and the impact of source number on the speech recognition score. Considering the notion of babble as the noise created by simultaneous crowd talking together, the lower the number of background speakers, the easier speech recognition is, and more gaps could be found in the speech signals. Conversely, by increasing the number of speakers, there are fewer gaps in the spectrogram which makes the identification of each individual speech, more difficult \[77\].

Identification score of multiple uttered consonants with respect to the number of simultaneous individual active speakers is studied in \[137\]. This study clearly shows that, identification score will not change when the number of active individuals exceeds about 7. Therefore, for
more than seven overlapping speakers the characteristics of the babble noise become more stationary and sounds like a diffused background rumble. For the range of 4 to 7 concurrent active speakers, even though individual talkers are barely identifiable, some individual words associated to the speakers could be heard, occasionally. For less than 4 speakers, the problem is more like competing talkers. Therefore, for more than 7 simultaneous active sources, it is almost impossible to achieve the speaker enumeration or recognize the individual words associated to them. When the babble is a crowd of people contributing the interference, the most recent achievements (to our knowledge) has been reported through learning a dictionary of babble noise bases and utilizing them in an NMF framework \[138, 139\]. Our focus in this thesis would be more on the competing talker interference than the crowd generated babble noise.

In this chapter we aim at different multi-channel speech separation techniques by which multiple source recognition originated from different speakers in a noisy enclosure would be feasible. Single channel speech separation is not our focus, since the previously explained Non-negative Matrix Factorization(NMF), is the only known method which can deal with this degenerate mixing system.

Depending upon the known geometry of the microphones, and the locations of the sources the BSS techniques in a distant speech recognition (DSR) front-end problem could be divided into different classes of techniques:

- Beamforming (BF) methods: (Fixed BFs, adaptive BFs)
- ICA-based methods
- SCA\(^2\)-based methods
- NMF-based methods
- CASA\(^3\)-based methods

### 4.2 Beamforming: An extended view

In chapter 3, we explained the theory of beamforming from linear algebra viewpoint and how this multidimensional signal processing\(^4\) extracts the desired speech which is contaminated by noise and interference, based on directivity (i.e. spatial signal selectivity). In many real conditions, position of the target, which was assumed to be constant so far, changes even with a trivial movement of the speaker. Therefore, robustness against the steering vector errors and other microphone array imperfections would be of an interest. To deal with the movements of the speaker and other adverse effects, one way is to constantly compute the DOA of the desired

---

1. And essentially the dictionary based single channel decomposition techniques, such as NMF, NTF, KSVD, etc.
2. Sparse Component Analysis
3. Computational Auditory Scene Analysis
4. Since the processing is performed in space and time over multiple observations
speaker in every meaningful time period. This task is out of the scope of our thesis and has a high computational demand. The second solution would be to change the directivity pattern adaptively, in such a way that less noise and interference be allowed in the output. Superdirective BF, as in \[3.66\] tries to minimize the output signal noise energy while preserving the gain toward the direction of the desired signal. A simpler delay-and-sum beamformer (DSB), only takes the average over the multiple observation channels after the alignment of the signals. This is done by the array manifold vector that compensates for the signal lags in different microphones, which occur due to the different positions of the microphones with respect to the source. As derived in \[3.64\], since the noise components are noncoherent while the signal components are coherent, this averaging is very likely to decrease the noise level and increase the output signal level, simultaneously. Therefore, the DSB gain would be simply the average over the multiple aligned signals as \( W_{DSB} = v/M \), where \( M \) is the number of the microphones.

In speech separation scenario, however we would actually like to “listen” to one speaker while suppressing the other. This can be achieved by using the general Linearly Constrained Minimum Variance (LCMV) solution beamforming \[27\] in order to apply a one in the direction of the desired speaker and a null in the direction of the interference, i.e. \( w^Hv_1 = 1 \) and \( w^Hv_2 = 0 \) with \( v_1 \) and \( v_2 \) associated to the location of the desired and interfering speaker, respectively. This leads to the following weight vector:

\[
\mathbf{w}_{lcmv} = \Gamma^{-1}nn \mathbf{C} \mathbf{C}^H \Gamma^{-1}nn f
\]

(4.1)

with \( \mathbf{C} = [v_1 \ v_2] \), \( f = [1 \ 0]^T \) and with \( \Gamma_{nn} \) again denoting the noise coherence matrix of the spherically isotropic noise field. LCMV-beamformer in fact generalizes the superdirective idea by imposing constraints on the desired subspace and on the noise/interference subspaces. The LCMV solution is a standard constrained optimization problem and the optimal weights are sought to minimized the objective function in the entire input dimension space. By dividing the input domain into the signal subspace (which holds the signal constraints), and the noise subspace (which hold an orthogonal subspace with respect to the signal) a more delicate and efficient solution, namely Generalized Sidelobe Canceler (GSC) beamforming could be achieved, which is depicted in figure\[4.1\].

A GSC beamformer actually employs an adaptive filter structure, which separates the input noisy signal space into two subspaces of signal and noise, and then tries to extract the noise residuals, adaptively from the signal to improve the signal to noise (SNR) and signal to interference (SIR) ratios. Since the signal extraction (or estimation of the signal subspace) is performed only geometrically (i.e., using a beamformer) and this is just an approximation for the true sig-

\[\text{1}^\text{due to different signal arrival lags}\]
\[\text{2}^\text{due to the diffuse noise field, the phase difference between channels are random, and so they are noncoherent.}\]
\[\text{3}^\text{The distortionless signal constraint for the signal subspace, and null constraint for the interference subspace.}\]
\[\text{4}^\text{If the input dimension is assumed N, the optimization solution is sought in an N dimension space.}\]
Figure 4.1: A general adaptive beamformer system

The residual components would still remain in the direct path signal, \( b \), that justifies the above-mentioned parallel structure to achieve a more enhanced output signal, \( y \). One advantage of utilizing this structure is that, it converts the constrained optimization problem in (4.1) to an unconstrained problem and the search space for the solution is reduced only to the signal subspace, instead of the entire space \([22]\). The parallel structure, in figure 4.1, causes the main difference between the objective function in an adaptive beamformer and the previously mentioned fixed LCMV beamformer version. The duty of this parallel path is to adaptively suppress the noise (i.e., any unwanted signal other than the desired speech, such as interference or uncorrelated background noise), by tracking it from the fixed beamformer output, using a speech blocker processor\(^1\) following by multiple parallel adaptive noise cancellation filters, which are applied to the noise only channels. A practical structure is shown in figure 4.2. While the output signals \( z_0, \cdots, z_{N-1} \) contain the unwanted signals from the input observation channels, the multiple input canceler block (MC) which contains a set of noise cancelers, extracts these unwanted signals out of the fixed beamformer signal \( b(k) \). Therefore, the output \( y(k) \) should contain less noise and interference components than \( y(k) \) itself. Delays are used to align the desired signal in all the outputs, based on the required processing time of each block. Weights of the adaptive filters could be obtained using the update rules in a Least Mean Square (LMS) or Recursive Least Squares (RLS) regime\([140]\). The complete derivations of the GSC-structured adaptive beamformer solution is presented in \([22, 71]\).

It is notable that, this noise suppression processing could be also performed as an individual enhancement block after the fixed beamformer is applied. The noise suppressor block which usually comes after the fixed beamformer is called as postfilter\([141, 142]\). Postfilter

\(^1\)Which allows all but the desired signal to pass through
4.2. BEAMFORMING: AN EXTENDED VIEW

Figure 4.2: A GSC-structured adaptive beamformer example, containing a fixed beamformer (FBF), a signal (speech) blocking section using adaptive filter set with alignment delays (BM), and a multiple noise cancellation structure (MC). The system is also called as Griffiths-Jim beamformer (GJBF). This figure is taken from the reference [132], page 92.

can employ any of the denoising techniques already mentioned in the previous chapter (e.g., Wiener filter, logmmse processor, dereverberators, etc.). A simple yet efficient postfilter is a modified Wiener filter (called, Zelinski post filter after R. Zelinski [143]) which can be derived, as follows

\[
H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \Phi_{nn}(\omega)}
\]  

(4.2)

Here, \(\Phi_{ss}(\omega)\) and \(\Phi_{nn}(\omega)\) denote the speech and noise power at the output of the array. Following Zelinski [143], these two parameters are here estimated as follows:

\[
\Phi_{ss} \approx \frac{2}{L(L-1)} \Re \left\{ \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} v_i^* \Phi_{x_i x_j} v_j \right\} 
\]  

(4.3)

\[
\Phi_{nn} \approx \frac{1}{L} \sum_{i=1}^{L} \Phi_{x_i x_i} - \Phi_{ss} 
\]  

(4.4)

where \(\Phi_{x_i x_j}\) and \(\Phi_{x_i x_i}\) denote the cross and power spectral densities of the sensor signals and where \(v_i\) denotes the \(i\)-th coefficient of the array manifold vector. The dependency on \(\omega\) has again been dropped for the sake of readability.

Another noticeable point is that, an adaptive beamformer is only adaptive in the sense that it tracks the unwanted interference/noise and cancels them in an adaptive way. Therefore, the
advantage of it over the fixed beamformer is the parallel processing block which is employed at the same time with the fixed beamformer. Therefore, a combined system of a fixed beamformer cascaded by a postfilter and an adaptive beamforming system both have almost the same applicability with different implementation strategies. The only shortcoming of the parallel adaptive system is that, by erroneous direction of arrival fed into the beamformer, the output of it contains more noise and the desired signal will instead leak through the orthogonal blocking matrix, because it looks for a subspace which is orthogonal to the beamformer output. Therefore, the parallel structure will try not only to cancel the noise, but also some signal components. This signal cancellation is inevitable, specially in a reverberant environment, in which source localization is very challenging and erroneous. Therefore, in a closed area we can expect the cascaded beamformer-postfilter system to outperform the adaptive structure or at least be comparable due to its inherent robustness against signal cancellation. A complete discussion could be found in [132, 144].

4.3 Independent Component Analysis and extensions for BSS

Independent Component Analysis (ICA) is a generative model for the multivariate observed data (e.g., a signal), which leverages the statistical and computational techniques in an information theoretic framework, to extract the underlying independent components of the data [39, 145, 146, 56]. ICA emerged primarily as a solution to the Blind Source Separation (BSS) problem, in conjunction with Independent Factor Analysis (IFA) or as a generalization of the PCA technique. While beamforming could encounter serious shortcomings, due to its sensitivity to the array geometry and the source directional-of-arrival accuracy, and contained inherent problems across higher frequencies in terms of spatial aliasing and in lower frequencies in terms of diffuseness field violation, ICA has none of these problems [132]. Moreover, contrary to the beamforming methods, the spatial noise in ICA is not necessarily assumed to be white.

Independent Component Analysis (ICA) utilizes a representation of the data in a statistical domain rather than a time or frequency domain. That is, the data are projected onto a new set of axes that fulfill some statistical criterion, which implies independence [147]. An important difference between ICA and Fourier-based techniques is that the Fourier components on which a data segment is projected are fixed, whereas PCA- or ICA-based transformations depend on the structure of the data. The axes onto which the data are projected are therefore discovered, adaptively. If the structure of the data (or rather the statistics of the underlying sources) changes over time, then the axes onto which the data are projected will change, too. A general model illustrating blind source separation system is depicted in figure 4.3. Any projection onto another set of axes is essentially a method for separating the data out into individual

---

1Which is called multi-channel Wiener filter if the beamformer is an MVDR.
2Principal Component Analysis, also known as Karhunen-Loève Transform (KLT), or Hotelling Transform
4.3. INDEPENDENT COMPONENT ANALYSIS AND EXTENSIONS FOR BSS

Figure 4.3: General Blind Source Separation system, with $Q$ sources and $m$ mixtures. ICA utilizes the independence measure to adaptively extract the outputs $y_q(n), q \in 1, \ldots, Q$. Microphone signals $x_i(n), i \in 1, \ldots, m$ are the observations. The unknown part is a model assumed for the observations. The demixing (inverse) system requires apriori knowledge to uniquely estimate the original source signals. It can be linear or nonlinear, like Neural Networks.

components which will hopefully allow us to see important structure more clearly. Therefore, the direction of projection increases the SNR for a particular signal source. In PCA and ICA we attempt to find a set of axes which are independent of one another in some sense. We assume that, there are a set of independent sources in the data, but do not assume their exact properties. Therefore, they may overlap in the frequency domain in contrast to Fourier techniques.

We then define some measure of independence and maximize this measure for projections onto each axis of the new space. The sources are in fact the data projected onto each of the new axes. Since we discover, rather than define the new axes, this process is known as blind source separation, because we do not look for specific predefined components, such as the energy at a specific frequency $f$ but rather, we allow the data to determine the components itself. These projections are then performed using the transformation matrices, which map the data onto the desired axes, and are also called separation matrices.

By setting columns of the ICA separation matrix that correspond to unwanted sources to zero, we produce a non-invertible matrix. If we then force the inversion of the separation matrix and transform the data back into the original observation space, we can remove the unwanted source from the original signal.

For PCA the measure we use to discover the axes is variance and leads to a set of orthogonal axes. For ICA this measure is based on nonGaussianity, therefore the axes are not necessarily orthogonal. The idea is that if we maximize the nonGaussianity of a set of signals, then they are maximally independent. This is according to the central limit theorem (CLT). If we

---

1Such as STFT transform, or Power Spectrum
keep adding independent signals together, we will eventually arrive at a Gaussian distribution. Conversely, if we break a Gaussian-like observation down into a set of non-Gaussian mixtures, each with distributions that are as non-Gaussian as possible, then the individual signals will be independent. ICA algorithms consider different criteria in their optimization goal, to achieve independence:

- ICA by maximizing the Non-Gaussianity (e.g., Kurtosis or Neg-entropy)
- ICA by minimizing the Mutual Information
- ICA by maximizing the Likelihood criterion (as in)
- ICA by minimizing the nonlinear decorrelation
- ICA by using higher-order moments or cumulants (see)
- ICA by maximization of the information transfer

There are two unavoidable uncertainties coming along with the ICA solutions. 1) We cannot determine the variance of the independent components, 2) we cannot determine the order of the independent components in the output of the ICA-based separation system. These ambiguities naturally arise due to the ill-posed problem in (4.5), in which only the left hand side is known and observed. By representing the relationship between the observations (e.g. microphone signals) and the sources linearly, the following equation in vector form appears:

\[ X(n) = A S(n) \]  

(4.5)

where \( X \in \mathbb{R}^{M \times T} \), and \( S \in \mathbb{R}^{N \times T} \) are the observations and sources matrices in the time domain respectively, and \( A_{M \times N} \) represents the linear mixing matrix, which is assumed stationary over time. \( M \) is the number of the observation vectors and \( N \) is the source number, and \( T \) denotes the frame length. When the linear system of equations in (4.5) is even-determined (i.e. \( M = N \)), or over-determined (i.e. \( M > N \)), ICA may find a solution, but for the under-determined case (i.e. \( M < N \)), ICA does not work. Sparse Component Analysis (SCA) is a framework which can play an important role in under-determined case, which would be explained in the next section.

ICA essentially requires two assumptions. The first, is that the sources are independent of each other and identically distributed (iid). The second assumption is that the source densities should not be Gaussian distributed.

The noise in an ICA problem can be regarded as an individual source to be separated or as a measurement error. If the former condition is assumed, then the constraint of iid assumption of the sources in ICA would be violated. Because, noise is usually best modeled as a Gaussian

\[ ^1 \text{which have highly non-Gaussian PDFs} \]
\[ ^2 \text{Which is not the case in real scenarios. In reality, their relation is convolutive, and noisy as in} \]
\[ ^3 \text{In fact, not more than one source density can be Gaussian.} \]
\[ ^4 \text{Independent and identically distributed.} \]
distribution, whereas the speech signal has to be non-Gaussian and in our specific application super-Gaussian distributed. Therefore, the latter assumption (i.e. noise as a measurement error) would be more logical and is handled in many noisy ICA versions [39,158].

The general mixing model in ICA is usually represented in an instantaneous form, as follows [159]:

\[ X = A S + N \]  \hspace{1cm} (4.6)

where \( N \) denotes the multivariate Gaussian noise, \( S \) is the matrix of the sources and \( A \) denotes the mixing matrix. As in [160,161,34], all methods taking noise explicitly into account assume it to be Gaussian distributed. Gaussian distribution contrary to the super-Gaussian one (like the pdf of a speech signal) contains only the first and second cumulants \(^2\). Thus, it would make sense to utilize higher order cumulants of the signal (e.g. Kurtosis) which are unaffected by the Gaussian noise, for the separation purpose.

H"{y}va"{r}inen in [160] uses the Maximum Likelihood method to extract the original source approximation in a mixing model as in (4.6) for an even-determined linear mixing model. Combining the solution given by H"{y}va"{r}inen, with the sparse signal coding presented by Zibulevsky et al. in [34], we can briefly develop a solution for a sparse-ICA based solution to the speech separation and denoising problem in a multi-channel scenario. In the following, we briefly explain the idea of the two well-known ICA algorithms, namely sparse-ICA and fast-ICA.

### 4.3.1 ICA and measures of independence

As we already mentioned, the key principle to estimate the ICA model is nongaussianity. Actually, without nongaussianity the estimation is not possible at all. By leveraging the CLT \(^3\) in the linear ICA model we are looking for a separation matrix \( W \), such that applying it on the known observations, as \( W^T X \approx \hat{S} \), yields the maximum nongaussianity in the output signals \( \hat{S} = [S_1, \ldots, S_N]^T \). To use the nongaussianity in ICA we need a quantitative measure. One measure would be Kurtosis, which is the fourth-order normalized cumulant of the random variable (i.e., signal), as

\[ \text{kurt}(y) = \text{E}\{y^4\} - 3(\text{E}\{y^2\})^2 \]  \hspace{1cm} (4.7)

where \( \text{E} \) denotes the expectation operation. If we normalize the signal energy, then the variance (the negative term of (4.7)) is one, and kurtosis becomes a normalized fourth-order moment, \( \text{kurt}(y) = \text{E}\{y^4\} - 3 \). Therefore a Gaussian random variable would have a zero kurtosis. On the other hand, the speech signal distribution is super-Gaussian \(^4\) and therefore has a high kurtosis. This measure is easy to compute and could be estimated from the data samples, but has the

---

1. After dropping the dependencies on time or frequency, depending on the representation domain.
2. Which are the mean and variance.
3. Central Limit Theorem
4. As it was mentioned in chapter 2, figure 2.5 and figure 2.8
drawback when its value is to be estimated from the measured samples it is very sensitive to outliers. Therefore, it is not a robust measure of nongaussianity.

Negentropy is a robust measure of nongaussianity. The more random, unpredictable and unstructured the variable is, the larger the entropy. Entropy as a measure of information buried in a stochastic signal, signifies that a Gaussian random variable must have the largest entropy among all random variables of equal variance. Our intention is to achieve separation by maximizing the nongaussianity of the output components. It implies that we are looking for a solution that makes the outputs to have maximum entropy distance from a Gaussian equivalent random variable which has the same covariance matrix as the outputs. This distance measure is named as negentropy and is defined based on what it was stated:

$$J(y) = H(y_{\text{gauss}}) - H(y)$$

(4.8)

where $H$ denotes he entropy measure, and is defined as

$$H(y) = -\int f(y) \log f(y) \, dy$$

(4.9)

where $y$ is a random vector (i.e. signal samples), and $f(y)$ is the probability density function of that random vector $y$. Obviously, the negentropy estimation would be difficult, as the density function is required which is not available since we do not have access to the entire data. Therefore, an approximate negentropy should be obtained. A reasonable approximation is by using the contrast functions which are some nonquadratic functions, in the following form:

$$J(y) \approx \sum_{i=1}^{p} k_i [E\{G_i(y)\} - E\{G_i(\nu)\}]^2$$

(4.10)

where $k_i$ are some positive constants, $\nu$ is a Gaussian random variable with zero mean and unit variance similar to the variable $y$, and the functions $G_i$ are some nonquadratic functions. If $G(x) = x^4$ then (4.10) becomes equal to kurtosis. There are functions that can be used for $G$ that give a good approximation to negentropy and are less sensitive to outliers than kurtosis. Two commonly used contrast functions are:

$$G_1(u) = \frac{1}{a_1} \log \cosh a_1 u, \quad G_2(u) = -\exp(-u^2/2)$$

(4.11)

where $1 \leq a_1 \leq 2$ provides a suitable constant. This way a good approximation of negentropy would be achievable, by replacing the contrast functions of (4.11) into (4.10), which is more robust against outliers than kurtosis.

For all ICA algorithms some prepossessings are necessary. Separation is performed based on the higher-order moments of the distribution of the sources, such as kurtosis. Therefore, we can remove all the linear dependencies (i.e. second order correlations) in the data set
and normalize the variance along all dimensions. This operation is named whitening or sphering. This operation maps the data into a spherically symmetric distribution to emphasize that no preferred directions exist. In other words, the covariance of the transformed data is diagonalized. Therefore, performing PCA is a preprocessing for most ICA algorithms.

Another measure which is used for many ICA algorithms is the mutual information rate. NonGaussianity implicates the independence of the output signals, however independence may also be realized, if the rate of the mutual information among the outputs are minimized. Assuming that $W$ is a separation matrix (i.e. $Y = W^TX \approx \hat{S}$), independence of the output signals entails that,

$$ p_Y(y) = \prod_{i=1}^{n} p_i(y_i) \quad (4.12) $$

where the left hand side is the joint density versus the product of the marginal densities of the output signals in the right hand side of the equation. Investigating this independence condition is not a trivial task in practice. Therefore, we need a measure which is more pragmatic. We can seek for the case where the equality in $(4.12)$ is realizable through Kullback-Leibler divergence (KLD) between the joint and the product densities,

$$ I(y) \triangleq D_{KL}[p_Y(y) \| \prod_{i=1}^{n} p_i(y_i)] $$

$$ = \int p_Y(y) \ln \frac{p_Y(y)}{\prod_{i=1}^{n} p_i(y_i)} \, dy \quad (4.13) $$

Minimization of the mutual information now boils down to the problem of finding the separation (or demixing) matrix $W$, such that $I(y)$ is minimized. This could be performed using the natural gradient descent algorithm, as in $[39, 145]$.

### 4.3.2 ICA algorithm using the sparsity prior

In the following, we briefly explain the sparse ICA method based on $[24, 160, 34, 164]$:

- Practically, speech sources are sparse in some system of basis functions (e.g., wavelet packets, STFT, etc.). Thus, they could be synthesized based on an overcomplete dictionary of these bases with a sparse matrix of coefficients:

$$ s(t) = \sum_{k=1}^{K} C_k \psi_k(t) \quad (4.14) $$

where $\psi_k(t)$ are the atoms of the dictionary matrix $\Phi$, and $C_k$ are the vector of coefficients.

---

1 Principal Component Analysis
In a matrix notation form we can rewrite the equations, as follows:

\[
\begin{align*}
X &= AS + N \\
S &= C\Phi
\end{align*}
\] (4.15)

Now, assuming the Gaussian noise with zero mean and variance \(\sigma^2\), we set a goal which can be formulated as follows: Given the sensor signal matrix \(X\) and the dictionary \(\Phi\), find a mixing matrix \(A\) and matrix of coefficients \(C\) such that \(X \approx AC\Phi\) and \(C\) is as sparse as possible. The mathematical form of this statement would be expressed, as

\[
\min_{A,C} \frac{1}{2\sigma^2} ||AC\Phi - X||_F^2 + \sum_k h(C_k) \quad \text{subject to: } ||A|| \leq 1 \tag{4.16}
\]

where \(F\) denotes the Frobenious norm. Assuming that the source follows a sparse distribution (from the exponential family, e.g. a Laplacian pdf), we have

\[
p(C_k) = e^{-h(C_k)}, \quad h(C_k) = |C|^\gamma, \quad \gamma \leq 1 \tag{4.17}
\]

where \(h(\cdot)\) is a function such as an absolute value, for a Laplacian distribution. There are some smoothed substitution versions of the \(h(\cdot)\), which could be replaced. More details are mentioned in [34]. Matrix \(A\) has to be restricted, as in (4.17), because otherwise it can tend to infinity and the coefficients \(C\) can also tend toward zero, which yields the minimum of the cost function without achieving a reasonable solution. Therefore, we restrict the matrix \(A\) by some norm value. An iterative hard-thresholding method could be applied to solve (4.17) as in [164, 165, 166, 167].

The above log-likelihood cost function (4.16), is rather better to be solved directly for the inverse mixing matrix \((W \approx A^{-1})\), so that it could be directly applied to the multichannel signal \(X\) to extract the original signal, \(\hat{S} \approx WX\). If the dictionary of the source system of the basis functions is non-overcomplete and invertible \(\Psi = \Phi^{-1}\), then a simple modification of the cost function would be necessary, as follows:

\[
\min_W \mathcal{L}(W; Y = X\Psi) = -K \log|\det(W)| + 1/K \sum_k h((WX)_k) \tag{4.18}
\]

where \(C = S\Psi = WY\), and the negative sign of the log-determinant appears, since the determinant of \(\det(W) = |\det(A)|^{-1}\). By finding the inverse of the mixing matrix \(W\), then the separated source estimates would be \(\hat{S} \approx WX\).

The explained solutions to the ICA problem, based on the independence, nongaussianity measure, or the sparsity prior all have some assumptions in common:

\[\text{The negative log-likelihood function, } \mathcal{L}(X; A, C).\]
4.3. INDEPENDENT COMPONENT ANALYSIS AND EXTENSIONS FOR BSS

- The input system model is assumed to be instantaneous, as in (4.6).
- There are two inevitable indeterminacies, permutation and scaling.
- The linear system of equations in (4.6) is over-/even-determined.

Except for the indeterminacies, the two other assumptions are many times violated in the real experiments. Considering specifically a distant speech recognition in a closed area with reverberation and noise, the mixture model will be convolutive. The only preprocessing which makes the previous ICA algorithms usable again, would be a simple transformation into a Fourier domain\(^1\) which has the property that any convolution operation in the time domain, now converts to a multiplication. With frequency-domain ICA, if we use a sufficiently long frame window, the convolutive mixture can be approximated with multiple instantaneous mixtures, each of which is defined for a frequency bin\(^1\). However, we should bear in mind that, the frame size should be long enough to cover the main part of the room impulse response, in order for the conversion from the time domain convolution to the frequency domain multiplication to come true. Otherwise, this conversion is just an approximation.

In frequency-domain ICA (FD-ICA), the permutation uncertainty of the ICA solutions becomes serious. The ambiguities should be handled properly so that the separated frequency domain components that originate from the same source be grouped together. There are many solutions discovered to cope with the permutation alignment in FD-ICA, in which some prior knowledge about the source locations, or the fact about the dependencies between frequency channels of an individual source while independence (or at least uncorrelatedness) between the frequency channels of different sources are employed\(^{27, 28, 169}\), see figure 4.4. There are many algorithms to solve the ICA problem with all different objective functions mentioned in

\(^1\)e.g., by applying a Short Time Fourier Transform (STFT)

Figure 4.4: Frequency-domain ICA system with permutation alignment, from Reju VG \(^{170}\).
this section, which employ the measures of independence as described here, along with some a priori knowledge of the mixing system or the sources, such as fastICA. A maximum likelihood solution to the ICA problem is also explained in Appendix-A, as in [149]. Recently, an extension of independent component analysis from one to multiple datasets, termed independent vector analysis (IVA) [171, 172], has been emerged which utilizes the dependency between the source components across different datasets. Therefore, IVA can bypass the permutation problem using this dependency.

4.4 Sparse Component Analysis

Contrary to ICA where the mixing matrix and source signals are estimated simultaneously the sparse component analysis (SCA) is usually a multi stage procedure. The first stage is to find an appropriate linear transform in which the transformed sources are sufficiently sparse (e.g., STFT domain, or wavelet packet transform). The second stage, is to estimate the mixing matrix, usually by using a type of clustering technique (e.g., k-means, fuzzy C-means, etc.). The last stage, then would be to employ an optimization algorithm tailored for recovery of a sparse signal, such as linear programming (LP), quadratic programming (QP), semi-definite programming (SDP), or smoothed $\ell_0$-norm solver [173]. The power of SCA appears in solving the degenerate problems, in which the number of observations is less than the number of sources [1]. In fact, by orchestrating the stages as mentioned above, we do not need to be worried about the type of the problem. There are two major algorithms in this category, DUET [36] and LOST [174].

**DUET algorithm** assumes the convolutive or instantaneous mixing model with only two microphones. The convolutive case works for Anechoic environments (as in (1.12)), and not for the Echoic (reverberated) case (1.13). One of the two microphones is assumed to be the origin and the delay and attenuation of the signal receiving by the microphones are normalized based on this microphone. The DUET algorithm is briefly stated in the following steps:

- Transform the microphone signals $x_1(t)$, and $x_2(t)$ into the STFT domain.

- Calculate the ratio of the signal amplitude values, and phase:

  \[
  \alpha = \left| \frac{x_1(\tau, \omega)}{x_2(\tau, \omega)} \right| \left| \frac{x_2(\tau, \omega)}{x_1(\tau, \omega)} \right| \\
  \delta = \frac{1}{\omega} \angle \left( \frac{x_2(\tau, \omega)}{x_1(\tau, \omega)} \right)
  \] (4.19) (4.20)

- Construct 2D histogram which clusters the sources (e.g., using k-means method) based on the estimated parameters of delay and attenuation. The number of the peaks reveals

---

1 For which the mixture model is named under-determined.
2 Degenerate Unmixing Estimation Technique.
3 Line Orientation Separation Technique.
4.4. SPARSE COMPONENT ANALYSIS

the number of the sources, while the location of the peaks reveals the mixing parameter of the associated source.

- Construct time-frequency binary masks for each of the peak centers, \((\check{a}, \check{\delta})\), in the 2D histogram:

\[
M_j(\tau, \omega) = 1, \quad \text{if } \left( \check{a}, \check{\delta} \right) = (a_j, \delta_j)
\] (4.21)

- Recover the underlying sources by applying the mask to the mixtures, and transforming them back to the time domain.

When a proper sparse transform is applied on the sources the scatter plot of the observation data will look like figure 2.15. It is obvious that a proper sparse transform domain will make the scatter plot more oriented along the lines which represent the data orientation. **LOST algorithm** tries to leverage this line orientations to achieve the mixing parameters and by assigning the data points to their most appropriate line orientation will cluster the data points. The final stage would be only a sparse recovery algorithm, due to the degenerate system of linear equations in under-determined case. A brief of the **LOST algorithm** is mentioned, below:

- Randomly initialize the \(M\) line orientation vectors, \(v_i\):

- Partially assign each data point to each line orientation vector using a soft or hard data assignment, \([29, 175]\).

- Determine the new line orientation estimate by calculating the principal eigenvector of the covariance matrix, and repeat the process till here until convergence.

- After convergence, adjoint the line orientations estimates to form the estimated mixing matrix, \(A = [v_1 | \cdots | v_M]\).

- For the even-determined case data points are assigned to line orientations using \(s(t) = A^{-1}x(t)\). For the under-determined case, calculate coefficients using a sparse optimization solver such as linear programming, for each data point:

\[
\min_{S} ||S||_\ell_1 \quad \text{subject to: } AS = X
\] (4.22)

where the \(\ell 1\) norm could be minimized by linear programming. A better solution for recovery is proposed in smoothed-\(\ell 0\)-norm solution in \([30]\).

The shortcoming of the SCA methods is that they work for Anechoic environments. Therefore, in a real reverberated room they demonstrate a poor quality. One contribution of this thesis would be to orchestrate a mechanism which can convert the reverberated environment into an approximately Anechoic case, and then utilizes the SCA methods with a higher performance.

\(^1\)The representation of one observation channel with respect to another.
4.5 CASA-based Speech Separation

Rickard and Yilmaz in [18], stated that different speakers tend to excite different frequency bands at a time. In other words,

\[ S_1(\omega, t)S_2(\omega, t) = 0 \quad \forall \omega, t. \]  

(4.23)

In particular, it has been shown that perfect demixing via binary T/F masks is possible if the time-frequency representations of the sources do not overlap. This fact is named W-Disjoint Orthogonality (WDO) and shows the percentage of the frequency bins in a time frame from the spectrogram of a mixture which is dedicated only to a single source.

In a research work Araki et al. [176], showed that the WDO quantity is a considerable factor to be utilized in BSS problems. As in figure 4.5, in an Anechoic environment the WDO factor is very dependent on the frame size, but it trivially degrades by increasing the number of the speakers.

![Figure 4.5: Approximate WDO for various frame size T, in an Anechoic environment.](image)

On the other hand, they have shown that, as in figure 4.6, by increasing the reverberation rate the WDO factor does not degrade substantially. By looking at these two figures, we can conclude that the number of simultaneous active talkers would have a more adverse effect on the WDO and consequently on the separation performance than the length of the reverberation. The next important sign is that the distance between the sources and the sensors does not significantly impact the WDO factor. Therefore, WDO-based methods could be expected to be robust against the distance from the speaker to the sensors. Altogether, the investigations ensures us about the WDO as a valid factor to be employed by the separation algorithms.

One more conclusion of the aforementioned study is that, while performing the WDO based masking in the initial stage would be erroneous, however if a preprocessing has already been able to partially separate the sources, then the WDO factor can be more efficiently applied to suppress the interference residuals. It has become highly questionable that why the algorithms
which are developed to enhance speech and improve its quality cannot improve the intelligibility of speech, too. For the ASR system, as the target, improving the intelligibility is the goal. Many studies has been performed\footnote{By Fletcher et al. in the 1920th onward.}, as mentioned in \cite{77}, and some important reasons have been found:

- The background noise spectrum is not often estimated accurately. While the enhancement algorithms can perform well in steady background noise (e.g., car noise), they cannot generally perform acceptable in nonstationary noise cases.

- The majority of the enhancement algorithms introduce distortions, which might be more damaging than the background noise.

- Nonrelevant stochastic modulations arise from the nonlinear noise-speech interaction. That would be one main reason why the spectral subtractive algorithms by shifting the envelope lift introduce the most detrimental effect on speech intelligibility.

There are two main distortions that perceptually do not equally impact the intelligibility degradation. Since the desired signal and the masker signal can seriously overlap in their spectrum, the frequency-specific gain functions (which are usually applied in enhancement algorithms) which depend on the estimated SNR and the estimated noise spectrum might introduce amplification distortion, as well as attenuation distortion.

In general, in order to make sure that our algorithm can improve the intelligibility or not, we need to establish a quantitative relationship between the distortion and intelligibility, or alternatively develop an appropriate intelligibly measure. Frequency-based segmental SNR
(fwSNRseg) is a measure that has been proved to have a high correlation with the intelligibility of the noise-suppressed speech.

The reasons of quality and intelligibility mismatch discussed above, motivated us to control the distortion level in order to achieve a higher intelligibility. We should notice that, all the enhancement algorithms discussed in chapter 3, perform as a gain function, \( G \), applied to the observed noisy signal. Assuming the background noise to be additive, the magnitude-squared spectrum of the enhanced speech at frequency bin \( k \) would be

\[
\hat{X}^2_k = G_k^2 Y_k^2 = G_k^2 (X_k^2 + N_k^2) \\
= G_k^2 X_k^2 + G_k^2 N_k^2 
\]

where \( X_k \) and \( N_k \) denote the clean and noise signals, respectively. Therefore, the output SNR at the specified frequency bin \( k \), is

\[
\text{SNR}_{\text{out}}(k) = \frac{(G_k X_k)^2}{(G_k N_k)^2} = \frac{X_k^2}{N_k^2} = \text{SNR}(k) 
\]

which indicates that the output band SNR cannot be improved by any choice of \( G_k \). Based on the Articulation Index (AI) theory, the intelligibility is proportional to the weighted sum of SNRs across all bands and not the output band SNR. If we can develop an algorithm that can improve the overall SNR (across all bands), we can achieve a higher intelligibility rate.

All the enhancement algorithms mentioned already, need to have an accurate SNR estimate, which is not attainable in low SNR levels (e.g. below \( \text{SNR} = 0 \text{dB} \)). Therefore, we need to adopt a new strategy to develop algorithms which do not require accurate estimate of SNR.

Channel (i.e., frequency bin) selection based algorithms introduce a new paradigm for denoising the corrupted signal which enables us to control the type of speech distortion we allow through the processing. It suggests that if we are not able to compute the soft gain functions accurately due to the low-SNR level, make them binary and discard the unfavorable channels. This paradigm has been successfully employed in CASA-based applications, and due to their ability to improve the overall SNR level, has been successfully utilized in speech applications with ASR target systems.

---

1. Correlation rate of fwSNRseg with WER as the ultimate intelligibility measure is about 0.81.
2. Computational Auditory Scene Analysis
4.6 Proposed Separation Strategies

There are several strategies utilized in this work\textsuperscript{1} to improve the speech intelligibility in noisy-reverberant multi-talk scenario cases:

- Incorporating DOA or source location prior along with the statistical information existing for the sources in order to partially separate the sources primarily, and then enhancing the intelligibility of the extracted sources by removing the residual noise, and interference in an incremental procedure.

- Converting the Echoic (reverberant) environment condition into an approximately Anechoic condition, to provide an appropriate condition for BSS\textsuperscript{2} algorithms to perform more efficiently.

- Using a combination of BSS and beamforming techniques to leverage all the possible priors about the ingredients of the problem.

- Incorporating the coherency of the signal, and noncoherency of the noise or alternative speech signals.

This point forward, most of the structures and formulations are presented for the case of two overlapping talker case, for which the AMI-WSJ-OLAP dataset is available. However, the algorithms are easily extensible for more source conditions.

4.6.1 Incorporating DOA with further residual removing filters

The idea is to separate the speech sources, by incorporating the DOA\textsuperscript{3} information of the speakers, and removing the residuals by subsequent processes in an incremental procedure. The spatial diversity of the sources could be effectively employed by a fixed or adaptive beamformer. It is very unlikely that two or more speakers are located alongside the same line with respect to a microphone array in an environment. If the angular distance of the interfering sources are larger than the beamformer mainlobe size, then the gain pattern will sufficiently attenuate the interference signals and the desired signal will be efficiently extracted. However, there is always some residuals, due to the sidelobes in the gain pattern and errors in the localization which deviates the main beam from steering the true angle. The noise and reverberation are also assumed to be more diffuse-like, and therefore no matter of how much error the true steering has, they will leak through the mainlobe and degrade the signal. This motivates us and entails further incremental processing.

\textsuperscript{1}There are other strategies also applied, but not achieved a better result, so far. For example, utilizing a data-driven approach to estimate the state of each frame and to decide the filter parameters adaptively, based on the frame state. This can go deeper in the signal and deals with feature domain parameters rather than the signal domain ones. Incorporating the phase of the signal, after a true estimation of it is also considered. They will be published later, as soon as some better results could be achieved.

\textsuperscript{2}Blind Source Separation

\textsuperscript{3}Direction Of Arrival, which relates to the source location
It is important to notice that, beamforming would be feasible provided that the microphone set are located as an array, with a known geometry. After the beamforming, there would be no access to the location information anymore, unless we memorize the sensor signals until the end of the procedure. Therefore, in realtime applications this block should come first. The structure of figure 4.7 is developed to achieve source separation with high intelligibility. In figure 4.7, the beamforming block exploits the spatial diversity of the sources.

The postfiltering block performs the noise suppression as already stated in the beamforming section of this chapter. Equations (4.2), (4.3) and (4.4) are based on the assumption that the noise is incoherent. But this may not really be the case in practice. Hence, we use a noise overestimation factor \( \beta \) in order to compensate for possible systematic errors. This is achieved by changing the frequency response \( H(\omega) \) in (4.2) to:

\[
H(\omega) = \frac{\Phi_{ss}(\omega)}{\Phi_{ss}(\omega) + \beta \Phi_{nn}(\omega)}
\]

Early speech recognition experiments indicated that a value \( \beta \) of 0.5 gives reasonable results – at least on the MC-WSJ-AV corpus [123] that has been used in the speech separation challenge. This value compares to a theoretical optimum of \( 1/L = 0.125 \) for delay-and-sum beamforming with an 8-sensor array [142] (under assumption of incoherent noise). Since the postfilter stage only deals with the uncorrelated noise and it cannot perform efficiently on the reverberation (as a correlated noise) and interference which is highly nonstationary, an additional block is used that performs the suppression of the interference residuals based on CASA, and will be explained in the following.

Maganti et al [177], made use of the \textit{W-disjoint orthogonality} of speech [18], assuming that different speakers tend to excite different frequency bins at a time. Consequently, he used a binary mask, which selects the stronger speaker in each time-frequency unit (based on the relative power at the beamformer-postfilter output), and then sets the weaker speaker to zero. This is what essentially happens in the T/F masking stage in Figure 4.7. Deviating from [177,178],...
we here however explore the use of soft masks that account for uncertainty of one speaker being stronger than the other. If one speaker is essentially stronger than the other, then masking will in fact reduce the crosstalk noise over the full band of the signal and this will lead to a higher intelligibility, based on the former discussion about the intelligibility improvement reasons. Consequently, applying the mask to the postfilter outputs \( Y_i, i \in \{1, 2\} \), will extract the estimated clean speech spectra \( \hat{S}_i(\omega, t) \) at time \( t \):

\[
\hat{S}_i(\omega, t) = M_i(\omega, t) \cdot Y_i(\omega, t)
\]  

(4.27)

In this equation, \( M_i(\omega, t) \) denotes the mask, which would optimally be chosen to be 1 if the T/F unit \((\omega, t)\) is used by that speaker (the \( i \)-th) and which is set to 0 otherwise. In particular, it has been shown that perfect demixing via binary T/F masks is possible if the time-frequency representations of the sources do not overlap (i.e. the WDO condition holds for all T-F points) \cite{179,180}. This, however, requires the masks \( M_i(\omega, t) \) to be known. Maganti et al. \cite{177} proposed the use of

\[
M_i(\omega, t) = \begin{cases} 
1, & |Y_i(\omega, t)| \geq |Y_j(\omega, t)| \quad \forall j \\
0, & \text{otherwise}
\end{cases}
\]  

(4.28)

which is based on the assumption that the spatial filtering stage has already suppressed the interfering speaker, such that \( |Y_i(\omega, t)| > |Y_j(\omega, t)| \) if the \( i \)-th speaker is using the \((\omega, t)\)-th frequency unit while the \( j \)-th speaker is not. The same approach has been used in \cite{181}.

Although binary masking is optimal in theory, it has certain deficiencies in practice. First of all, the mask estimates may be erroneous if the interfering speaker is not sufficiently suppressed through spatial filtering. Secondly, the approach may not be optimal in reverberant environments, such as the SSC\footnote{Speech Separation Challenge} task \cite{123}, where the spectral energy is smeared in time. Hence, we propose the use of soft masks with the aim of more appropriately treating the arising uncertainties. The use of sigmoid masks is motivated by

1. the work of Barker et al. \cite{182} where it has been shown that sigmoid masks give better results in the presence of mask uncertainties.

2. the work of Araki et al. \cite{183} where it has been shown (a) that soft-masks can perform better in convolutive mixtures and (b) that a simple sigmoid mask can perform comparably to other sophisticated soft masks or even better.

In case of the speech separation scenario, we may use sigmoid masks in order to apply a weight to each of the sources, based on the difference of their magnitudes:

\[
M_i(\omega, t) = \frac{1}{1 + \exp[-\alpha |Y_i(\omega, t)| - |Y_j(\omega, t)|] + \beta}
\]  

(4.29)
with \( i \in 1, 2, j = 3 - i \) and with \( \alpha \) being a scale parameter, which specifies the sharpness of the mask and represents our uncertainty about the initial intensity of the sources originated from the speaker’s mouth, and \( \beta \) is the bias parameter which represents our possible source estimation errors from the previous stages that might have been occurred due to the erroneous steering angle, for example. Instead of directly applying this mask now, we exploit the fact that the human auditory system perceives the intensity of sound in a logarithmic scale. This can be incorporated into (4.29) by replacing the magnitudes \(|Y_i(\omega, t)|\) by logarithmic magnitudes \(\log|Y_i(\omega, t)|\):

\[
M_{i, \alpha}(\omega, t) = \frac{1}{1 + \gamma \left( \frac{|Y_j(\omega, t)|}{|Y_i(\omega, t)|} \right)^\alpha} \tag{4.30}
\]

where the logarithms have been pulled out of the exponential function. Although the scale parameter \( \alpha \) and over-/under-estimation parameter \( \gamma \) may be chosen individually for each frequency bin, they are here jointly optimized once for each utterance. In the particular case where \( \alpha = 2 \), and \( \gamma = 1 \) the log-sigmoid mask is identical to a Wiener filter. Therefore, it would be interpreted that the soft mask considers interference as a noise source and tries to suppress it in MMSE sense\(^1\).

\[
M_{i, 2}(\omega, t) = \frac{|Y_i(\omega, t)|^2}{|Y_i(\omega, t)|^2 + |Y_j(\omega, t)|^2} \tag{4.31}
\]

with \(|Y_i(\omega, t)|\) being the magnitude of clean speech and \(|Y_j(\omega, t)|\) being the magnitude of noise.

As stated in the ICA section \(^2\) some measures of nongaussianity will guarantee the separation performance of super-Gaussian signals, such as speech. Even though kurtosis is sensitive to the outliers, it is computationally very efficient and the sample kurtosis can be implemented very fast. Motivated by Li and Lutman’s work\(^2\) we here use the subband-kurtosis as a measure for judging the separation quality of concurrent speech. Consequently, the quality of a separated utterance \(\hat{S}_{i, \alpha}\) is determined as the average kurtosis over all frequencies:

\[
\text{kurt}\{\hat{S}_{i, \alpha, \gamma}\} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \frac{\frac{1}{T} \sum_{t=1}^{T} |\hat{S}_{i, \alpha, \gamma}(\omega, t)|^4}{\frac{1}{T} \sum_{t=1}^{T} |\hat{S}_{i, \alpha, \gamma}(\omega, t)|^2} \tag{4.32}
\]

where \( \omega \in \Omega \) and \( t \in \{1, \ldots, T\} \) denote the angular frequency and the discrete time index, respectively. The \(\hat{S}_{i, \alpha, \gamma}(\omega, t)\) are the separated subband samples after time-frequency masking,

\[
\hat{S}_{i, \alpha, \gamma}(\omega, t) = M_{i, \alpha, \gamma}(\omega, t) Y_i(\omega, t).
\]

with scale parameter \( \alpha \), and estimation parameter \( \gamma \). Now, \( \alpha \) and \( \gamma \) may be optimized by running a grid search over a range \( R_\alpha \) of possible values and then selecting that \( \alpha \) and \( \gamma \) for which

\(^1\) As Wiener solution is an optimal solution in Minimum Mean Square of Error (MMSE) sense.
\(^2\) which clearly shows that the subband kurtosis decreases with the number of simultaneous speakers [184]
kurt\{\hat{S}_{i,a,\gamma}\} is maximized. To get a good coverage of different mask shapes with few parameters to test, we use:

\[
\mathcal{R}_a \triangleq \{\exp(a/10), \ a = -50, \ldots, 50. \ \gamma \in (0,1], \text{step} = 0.01\}. \tag{4.33}
\]

This optimization is done once for each utterance and individually for each speaker. Note that soft-masking can easily be extended to more than two speakers by using \(j = \arg\max_{k \neq i} |Y_k(\omega, t)|\) instead of \(j = 3 - i\) in (4.30).

The performance of the proposed system has been evaluated on the two speaker condition of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus\[123\]. This condition was used in the PASCAL Speech Separation Challenge II \[178, 181\] and it consists of two concurrent speakers who are simultaneously reading sentences from the wall street journal. This data set contains recordings of five pairs of speakers with total number of 356 utterances (or 178, respectively, if we consider the fact that two sentences are read at a time \[178\]) talking simultaneously to an array of 8 microphones planted symmetrically in a circle of 10cm radius. The speech recognition system used in the experiments is called "Millenium", which is identical to the one in \[181\], except that we use three passes only (instead of four): a first, unadapted pass; a second pass with unsupervised MLLR feature space adaptation; and a third pass with full MLLR adaptation. The estimated speaker positions are the same ones used in \[181\].

<table>
<thead>
<tr>
<th>Mask</th>
<th>PF</th>
<th>DSB</th>
<th>SDB</th>
<th>LCMV</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>no</td>
<td>77.87</td>
<td>68.73</td>
<td>58.56</td>
<td>57.58</td>
</tr>
<tr>
<td>Binary</td>
<td>no</td>
<td>58.89</td>
<td>57.20</td>
<td>49.97</td>
<td>52.15</td>
</tr>
<tr>
<td>log-sigm.</td>
<td>no</td>
<td>58.65</td>
<td>44.63</td>
<td>48.56</td>
<td>52.39</td>
</tr>
<tr>
<td>None</td>
<td>yes</td>
<td>69.07</td>
<td>61.65</td>
<td>56.77</td>
<td>56.98</td>
</tr>
<tr>
<td>Binary</td>
<td>yes</td>
<td>51.03</td>
<td>45.33</td>
<td>51.06</td>
<td>49.99</td>
</tr>
<tr>
<td>log-sigm.</td>
<td>yes</td>
<td>48.09</td>
<td>42.73</td>
<td>43.47</td>
<td>46.83</td>
</tr>
<tr>
<td>Headset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23.44</td>
</tr>
</tbody>
</table>

Table 4.1: Word error rates for different beamformers with and without postfiltering (PF) and time-frequency masking. The last row gives a comparison to the headset data which has been recorded in parallel to the array \[123\].

Table 4.1 shows the word error rates (WERs) we obtained with different configurations of the speech separation system from Figure 4.7. For spatial filtering, we used either a delay-and-sum beamformer (DSB), a superdirective beamformer (SDB), the LCMV beamformer from Section 4.2 or the minimum mutual information (MMI) beamformer from \[185, 181\].

MMI beamformer, in a very brief, was proposed by Kumatani et al. in \[185\]. He proposed to solve the speech separation problem by using two adaptive beamformers (with GSC\[\] struc-
ture) whose weights are jointly optimized to minimize the mutual information at the outputs. He modified the optimization criterion used for calculating the weights of the adaptive noise canceler, to minimize the following cost function:

$$I(Y_1, Y_2) = d \left( \log \frac{p(Y_1, Y_2)}{p(Y_1)p(Y_2)} \right)$$

(4.34)

where $I(Y_1, Y_2)$ is the mutual information between two beamformer outputs. This optimization is performed individually in each frequency bin. This MMI beamformer was reported to have achieved the state-of-the-art intelligibility score, before we proposed our method.

The first row of table 4.1 reveals that the WER of the plain SDB is 9% lower than that of the DSB. LCMV and MMI beamforming give a further reduction of 10% and therewith perform at the same level (58.6% versus 57.6%). The second row of table 4.1 shows the combination of spatial filtering with binary masking. This combination gives a significant improvement over the plain beamformers: almost 20% for the DSB, 10% for the SDB and still 8% and 5% for the LCMV and MMI beamformers. The use of kurtosis optimized log-sigmoid masks (row 3) results in similar improvements, except for the SDB where we have a further reduction of 13% compared to binary masking.

These results changed dramatically when a postfilter was applied between spatial filtering and masking. In this case, the combination with log-sigmoid masks gave the best speech recognition results obtained so far, with a word error rate of approximately 43% for the SDB and LCMV beamformers. The MMI and DSB beamformer were only slightly inferior, with a performance of 47% and 48%. Binary masking was between 3% and 6% worse. These results demonstrate that the right choice of post-processing can have a tremendous effect. The best WER is not necessarily achieved with the most sophisticated beamformer.

Due to the large improvements obtained with log-sigmoid masks, we thought it might be worth investigating how the kurtosis optimization affects the mask shape. For this purpose, we first selected some utterances which seemed to be well separated (after processing with the SDB) and then plotted their kurtosis (after T/F-masking) in dependency of the scale parameter. An example of such a plot is shown in Figure 4.8, along with a plot for an utterance where the separation was poor. Motivated by the strong differences in these plots, we divided the corpus into a set of utterances for which the speech seemed to be well-separated and a set of utterances for which the separation seemed to be poor. Subsequently, we plotted the average mask shape for each of the sets, as shown in Figure 4.9, which revealed that the kurtosis maximization selects harder (closer to binary) masks when the separation through spatial filtering is poor. It selects softer (i.e. less strongly reacting) masks when the separation quality is good. This automatic selection of parameters confirms the idea that when the noise or interference level is highly competitive with respect to the signals, the best idea to get a more intelligible source is to chose the binary

\[ \text{the one from Section 4.2} \]
4.6. PROPOSED SEPARATION STRATEGIES

mask, as it was concluded in [180]. In this study we have shown that (1) kurtosis optimized log-sigmoid masks can significantly outperform binary masks and (2) their mask shape is chosen in dependence of the separation quality after spatial filtering. This led to a reduction of over 25% in WER over a plain SDB, in combination with post-filtering. Apart from the above, we have shown that an interference canceling LCMV with a diffuse noise field design (as used in the SDB) can give almost the same performance as a minimum mutual information beamformer.

So far the entire processes were done in a frame-wise manner and we did not take the temporal correlation of the adjacent frames of speech into account. The former process could be completed by a further smoothing factor which accounts for this temporal correlations, in sub-bands. Since, the auditory perception system of human performs the speech perception in a nonuniform frequency channels, we exploited the ability of the Nyquist filterbank\footnote{STFT has a uniform division structure over the spectrum and cannot perform as good as Nyquist filterbank, in this case.} used in [186], to facilitate this condition. we divided the spectrum of signals to 512 bands while using 64 as the decimation ratio of the filterbank structure. Then, we changed the masking process as follows to average over the adjacent frames:

\[
\overline{M}_i(\omega, t) = \beta \cdot \overline{M}_i(\omega, t - 1) + (1 - \beta) \cdot \overline{M}_i(\omega, t)
\]  

(4.35)

We tried for various \( \beta \) values changing from \( 0.1 < \beta < 1 \) and what we obtained was that for

Figure 4.8: Kurtosis for well (upper plot) and poorly (lower plot) separated speech, in dependence of \( \alpha \).

![Kurtosis plots](image_url)
Figure 4.9: Mask shape for well and poorly separated mixtures, with the x-axis representing the ratio of $|Y_2|/|Y_1|$.

$\beta = 0.85$ for both subband binary mask and subband log-sigmoid mask we can achieve the best results\(^1\) See table 4.2 for results.

<table>
<thead>
<tr>
<th>All for SDB-PostFilter</th>
<th>$\gamma$ factor</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubBand Binary Mask</td>
<td>0.85</td>
<td>41.00</td>
</tr>
<tr>
<td>SubBand log-sigm Mask</td>
<td>0.85</td>
<td>41.99</td>
</tr>
</tbody>
</table>

Table 4.2: Word error rates for SDB beamformer followed by post-filter and frame-based smoothing in Subband

### 4.6.2 Removing the coherent residuals, by designing a filter

The signals of the array microphones in Fourier domain can be written as a linear combination of the atoms taken from a spatial basis dictionary, in which the base associated with the direct signal propagation is $v_i$, $i = \{1, 2\}$, and the rest of the bases which are concatenated in a matrix $\Lambda_{\text{ref}}$ reflections of the direct signal and other interferences. Formalization of this assumption for a two speaker scenario in a closed room while the microphone signals are corrupted with an additive ambient noise becomes, as:

$$X = [v_1 | \Lambda_{\text{ref}}] \begin{bmatrix} S_1 \\ S_{1R} \end{bmatrix} + [v_2 | \Lambda_{\text{ref}}] \begin{bmatrix} S_2 \\ S_{2R} \end{bmatrix} + N$$ (4.36)

Where $v_i \in \mathbb{C}^M, (i = 1, 2)$ are the array steering vectors toward the desired sources $S_1$ and $S_2$ with $M$ as the number of microphones, $\Lambda_{\text{ref}}, (i = 1, 2)$ are the array steering matrices for all angles of

---

\(^1\)In general, this $\beta$ parameter can be frequency dependent. We chose a simple parameter for all channels.
the space but the desired sources, and \( S_{iR}, (i = 1, 2) \) are the reverberated versions of the desired sources that contain the lagged version of the desired (current-time) sources with random delays. \( N \) is multichannel ambient noise. All entities are transformed into short time frequency domain (STFT) and \((\omega, t)\) is dropped for simplicity.

Applying the weight vectors of the beamformers (corresponding to the sources) to \( 4.36 \), and assuming \( w_i^H v_i = 1, w_i^H v_j \approx 0; i, j = \{1, 2\}, i \neq j \), based on \( 3.65, 3.66 \) and MVDR distortion-less constraints, we get the following signals:

\[
Z_1 = w_1^H X \approx S_1 + a_1 S_{1R} + a_2 S_{2R} + n_1 \\
Z_2 = w_2^H X \approx S_2 + b_1 S_{1R}' + b_2 S_{2R}' + n_2
\]

(4.37)

where \( a_i, b_i, i \in \{1, 2\} \) are the gain vectors related to the reverberation terms of the desired sources \((S_1, S_2)\) and interference parts \((S_{1R}, S_{2R}, S_{1R}', S_{2R}')\). Notice that, \( S_{1R}, S_{1R}' \) are not necessarily the same, since they are originated from \( S_1 \) but with different lags (randomly combined), and so does \( S_{2R}, S_{2R}' \). Consequently, there are subterms in \( Z_1 \) and \( Z_2 \) that are coherent and subterms that are incoherent. The coherent terms included in \( Z_1 \) and \( Z_2 \) outputs, make them dependent. \( n_1 \) and \( n_2 \) are the residual noise terms after the BF. Since the noise terms are assumed independent of the signals and reverberation parts, we apply our postfiltering (PF) to the outputs of BF \((Z_1 \) and \( Z_2)\) to increase the SNR level, however, we assume that the structure of the equation \( 4.37 \) is preserved with a lower noise level.

Following the line of thought of the previous work \[187\], utilizing the mask directly after BF-PF stage could be erroneous. The reason comes from the deficiencies in beamforming which does not allow the interfering signals to be sufficiently suppressed. In addition, reverberation causes the signal spectral energy to smear in time and affect the mask estimation. Here we use our previously proposed logSigmoid mask and modify it to be applied as a coherency removal filter. The logSigmoid mask was defined, as

\[
\bar{M}_{i,\sigma,\xi}(\omega, t) = \frac{1}{1 + \xi \left( \frac{|Z_j(\omega, t)|}{|Z_i(\omega, t)|} \right) \sigma} \quad i \neq j
\]

(4.38)

where \( \xi \) and \( \sigma \) are parameters to control the sharpness and scale matching of the mask, respectively for the uncertainty reasons already discussed. In our experiments with logSigmoid mask, we noticed that \( \sigma = 2 \) is mostly the optimized value. Interestingly, this value corresponds to the power of the signals in spectral domain (i.e., power spectrum). Hence, we choose \( \sigma = 2 \), change the name of the mask to \( G_z(\omega, t) \) filter, and approximate \( 4.38 \) with the binomial ex-
pansion with first two terms of the series that represent this function. Therefore, we have:

\[ G_{z_i}(\omega, t) = 1 - \xi \frac{|Z_j(\omega, t)|^2}{|Z_i(\omega, t)|} \]

\[ = 1 - \xi \left( \frac{P_{z_j}}{P_{z_i}} \right) \quad i \neq j \quad (4.39) \]

where \( P_{z_i}, P_{z_j}, (i, j) \in \{1, 2\} \) denote the power spectrum of the contrasting signals. The coherence between two signals describes the strength of association between them, and is defined as:

\[ \gamma_{z_i z_j} = \left( \frac{P_{z_i z_j}}{P_{z_i} P_{z_j}} \right)^2 \quad i \neq j \quad (4.40) \]

where \( P_{z_i z_j} \) denotes the cross power spectral density of the two signals. logSigmoid mask \((4.38)\) is a parametric design that (roughly saying) improves the separation ability of the mask by choosing the parameters, so that the resulting signal in the output statistically resembles a clean speech signal (i.e., particularly Kurtosis maximization in subbands to follow the supergaussianity of the clean speech). These parameters we should learn in each frequency band, before employing the logSigmoid mask. Here, we only resort to one parameter, namely \( \xi \), and we let it be proportional to the coherence of the signals \( Z_i \) and \( Z_j \) in power domain, since we intend to remove the coherency between \( Z_i \) and \( Z_j \) using these filters. Thus, in (4.39), we set \( \xi = \lambda \gamma_{z_i z_j} \), and the filters are shown as:

\[ G_{z_i}(\omega) = 1 - \lambda \gamma_{z_i z_j} \left( \frac{P_{z_j}}{P_{z_i}} \right) \quad i \neq j \quad (4.41) \]

The value of the \( \lambda \) can also be viewed as the parameter that compensates for the approximation (expansion) error of (4.39). \( \lambda \) can be optimized for a measure that is related to the speech intelligibility \([184]\), such as maximum Kurtosis.

Notice that all the parameters introduced here depend on frequency \( \omega \). The final filter equation to be applied to the outputs is:

\[ G_{z_i}(\omega) = \max \left\{ (1 - \lambda \gamma_{z_i z_j} \frac{P_{z_j}}{P_{z_i}}), 0 \right\} \quad i \neq j \quad (4.42) \]

The new structure which uses this coherence removal filter is as figure4.10 shows. The temporal correlation of the frames is again preserved by the smoothing mask, as mentioned earlier. The result of the proposed method in Fig.4.10 has been compared to some of the well-known BSS\(^1\) methods for which we already discussed about the required theory in this chapter, succinctly. These algorithms are convolutive ICA (cICA) \([55]\) of Ikeda et al., convolutive BSS (cBSS) method of L. Parra \([188]\), as well as our previous method of logSigmoid masking (lgSigMsk) \([187]\). In our previous work, \([187]\), we have already shown that superdirective beamformer out-

\(^1\) Blind Source Separation
4.6. PROPOSED SEPARATION STRATEGIES

Figure 4.10: Block diagram of the multi-channel speech separation system of two sources in each frequency bin, using Beamformer, coherence removal filter and soft mask separator.

performs other advanced beamformers such as LCMV when the multistage structure, including BF combined with PF are followed by a masking stage. In addition, Table 4.3 shows that our new proposed method significantly outperforms the compared ones in case of Perceptual Speech Quality (PESQ), noise and reverberation enhancement measures shown by Segmental SNR, Signal to Reverberation ratio, respectively, and the measure of intelligibility (STOI).

<table>
<thead>
<tr>
<th>Method</th>
<th>SegmentalSNR</th>
<th>CD</th>
<th>LLR</th>
<th>SRMR</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BF/PF+cICA</td>
<td>-0.72</td>
<td>5.41</td>
<td>1.04</td>
<td>3.72</td>
<td>1.59</td>
<td>0.43</td>
</tr>
<tr>
<td>~ + cBSS</td>
<td>-0.31</td>
<td>5.31</td>
<td>0.96</td>
<td>5.56</td>
<td>1.53</td>
<td>0.45</td>
</tr>
<tr>
<td>~ + logSigmoid Mask</td>
<td>-0.30</td>
<td>5.96</td>
<td>1.11</td>
<td>7.17</td>
<td>1.45</td>
<td>0.33</td>
</tr>
<tr>
<td>Proposed system</td>
<td>2.05</td>
<td>4.89</td>
<td>0.95</td>
<td>8.30</td>
<td>2.10</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 4.3: Comparison of our proposed method, figure 4.10, with some known methods in BSS, applied to the outputs of BF/PE, based on the measures: Segmental-SNR (in dB), Cepstral Distance (CD), LPC based LogLikelihood Ratio Distance (LLR), Source to Reverberation Modulation Ratio (SRMR in dB), Perceptual Evaluation Quality (1 ≤ PESQ ≤ 5), and Short Time Objective Intelligibility (STOI) measure. The values are the averages over the two overlapped speaker sounds.

The cepstral and LPC based distances show that the features of our method are closer to the natural speech. Evaluation measures are found online in REVERB Challenge workshop [45, 103]. Looking at Fig. 4.11 clearly shows that the spectrum of the proposed method is more enhanced than the other compared methods. Comparing the spectrograms with the clean one, we see that the remained interference is well removed in our method, whereas in other BSS methods the interference components are still remained (a sample is marked in the figure). Noise is effectively removed however, it seems that there are parts of the clean signal that are
also removed. This can be due to overestimation of the noise that has been removed in consecutive stages (postfilter and masking). Moreover, we are using a fixed smoothing value for the binary mask than can make mistakes in frames that voiced phonemes are located between low energy frames. In these cases the mask is mostly dominated by the value of the low energy frames and the voiced information might be omitted. In general, noise, reverberation and interference are significantly removed. As a result, we see in Table 4.3 that in addition to the improvement in PESQ, the segmental SNR and SRMR which are highly correlated with the intelligibility score are also highly improved. Moreover, the distances from the natural speech are also less than the compared BSS methods which again emphasize the higher intelligibility of the outcome of the proposed method. STOI measure which has been recently developed to measure the short-time intelligibility is also showing a significant improvement.

Based on our proposed algorithms in chapter 3 and our separation filter (i.e., coherence removal filter, or smoothed masking system), we put a new system in our separation and enhancement experiment, which employs the previous structure whose performance was satisfying for one speaker scenario. The system has been designed and implemented as in figure 4.12 in which the dereverberation prior to our coherency removing filter is considered to reduce the auto-correlation of the signal with its past values, followed by the coherence removal filter which reduces the dependency level of different outputs. Moreover, based on our proposed enhancement system in chapter 3, section 3.6, we employed the dereverberation network with the same reasons that we discussed in the previous chapter as the initial stage of our proposed system. The new structure is depicted in figure 4.14. This new structure is expected to have the following advantages:

- The reverberant environment makes the localization process erroneous. By using the dereverberation network in the initial stage, beamforming part achieves a more precise source localization result, and therefore steering toward the sources would be more accurate.

- The dereverberation network, converts the mixing condition to convert from echoic to an approximately anechoic case. Therefore, even without using the spatial filtering we expect to be able to incorporate Sparse Component Analysis (SCA) based algorithms from the BSS\textsuperscript{1} field to separate even more sources than the microphones (i.e., the degenerate Under-determined case.). Such a design can be performed as in figure 4.16. The results of this structure are later shown in table 4.4.

- Masking mechanism works based on the WDO\textsuperscript{2} fact, which mostly fulfills for anechoic conditions, see figure 4.5. This condition in echoic environments is violated. By using the dereverberation network as the initial stage, we are implicitly providing a better WDO condition for the subsequent masking processor.

\textsuperscript{1} Blind Source Separation
\textsuperscript{2} W-Disjoint Orthogonality
4.6. PROPOSED SEPARATION STRATEGIES

Figure 4.11: Spectrogram of a sample output from the compared methods. From the top: convolutive ICA (Ikeda et al.), convolutive BSS (Parra et al.), Logsigmoid mask (Mahdian), proposed system (Mahdian), and clean speech. Black circle depicts a residual interference.

- Dereverberation network, helps us to include ICA based algorithms, and the further systems in lieu of DOA based spatial filtering. That could be useful specially when the microphones are spread around the room with an unknown geometry.

- Source enumeration has always been a serious problem in separation task and it is mostly assumed to be a priori known. By converting the echoic condition into anechoic, the necessary condition for the DUEIT\(^1\) algorithm to work is respected. Thus, by counting

\(^1\)Degenerate Unmixing Estimation Technique
Figure 4.12: Including the linear prediction based dereverberation network inside the previously proposed system. While the noise and reverberation would be suppressed in the spatial filtering, postfiltering and dereverberation stages, the final stage (i.e., coherency removal filter) separates the residual interference.

The peaks of the histogram in the delay-attenuation 2D histogram, source numbers could be counted, as in figure 4.13. This can be used later for both, DOA finding, beamforming stage and BSS problems such as ICA, which is highly dependent on the number of the sources.

Figure 4.13: Histogram in DUET algorithm. Left) histogram for the case where the DUET algorithm is applied after WPE dereverberation, Right) DUET directly applied on the input microphones. WPE clearly modifies the echoic condition into anechoic case. Peaks denote the sources, with their relative delay and attenuation, which for the left one two peaks are obvious.
4.6. PROPOSED SEPARATION STRATEGIES

Figure 4.14: New system, by including the weighted linear prediction based dereverberation network, as the input stage to our previously proposed system.

Based on the discussion we made about optimality of the multi-channel Wiener filter\textsuperscript{1} to reduce the background noise in diffuse noise fields, this block could be used in the initial stage and then we can employ the previously mentioned dereverberation systems for each speaker signal, see figure 4.15. The result of the systems pictured in figures 4.12 and 4.14 and 4.15, with the specified measures on the same corpus data (i.e., AMI-WSJ-OLAP) are presented in table 4.4.

Figure 4.15: Employing the Multi-channel Wiener filter as the input stage, followed by WPE dereverberation system.

The results of the table\textsuperscript{4.4} clearly show that our coherency removal system outperforms in all cases of noise/reverberation suppression, interference reduction and preserving the naturalness level of speech. As it was expected, system of figure 4.14 can not effectively perform, since the initial stages have already distorted the signal. Therefore, even though it tries to reduce the reverberation successfully due to the previous processing stages the distortion level is high and that degrades the quality and intelligibility, to some extent. In contrast, when the WPE dereverberation is employed as the initial stage, it can perform more effectively, as it can be seen from the table. As we can see from the table, the only degraded results belongs to the complete

\textsuperscript{1}As it is derived in appendix A.1
Table 4.4: A comparison of the proposed systems for intelligibility and quality measures. Figure 4.7 is the softmask based system, figure 4.10 denotes the coherency removal filter system, figure 4.12 denotes the WPE based dereverberation being inserted into the previous system and after the multichannel Wiener-filter, figure 4.14 denotes the dereverberation network in the initial stage of the coherence removal filter system, and figure 4.15 denotes the multi-channel Wiener-filter cascaded by a 1-channel WPE dereverberator. The last line follows figure 4.16.

<table>
<thead>
<tr>
<th>measure vs. algorithm</th>
<th>STOI</th>
<th>PESQ</th>
<th>fwsSegSNR</th>
<th>SRMR</th>
<th>LIR</th>
<th>CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Fig. [4.7] (logSigmoid)</td>
<td>0.33</td>
<td>1.45</td>
<td>3.37</td>
<td>7.17</td>
<td>1.11</td>
<td>5.96</td>
</tr>
<tr>
<td>System Fig. [4.10]</td>
<td>0.60</td>
<td><strong>2.10</strong></td>
<td>4.08</td>
<td><strong>8.30</strong></td>
<td>0.95</td>
<td><strong>4.89</strong></td>
</tr>
<tr>
<td>System Fig. [4.12]</td>
<td>0.24</td>
<td>1.12</td>
<td>3.21</td>
<td>8.29</td>
<td>1.20</td>
<td>6.10</td>
</tr>
<tr>
<td>System Fig. [4.15]</td>
<td>0.51</td>
<td>1.47</td>
<td>4.87</td>
<td>6.28</td>
<td>0.93</td>
<td>5.25</td>
</tr>
<tr>
<td>WPE-netw-&gt;BF</td>
<td>0.53</td>
<td>1.63</td>
<td><strong>4.95</strong></td>
<td>4.44</td>
<td><strong>0.93</strong></td>
<td>5.16</td>
</tr>
<tr>
<td>WPE-netw-&gt;MC-WienerFilt.</td>
<td>0.26</td>
<td>1.20</td>
<td>3.24</td>
<td>7.38</td>
<td>1.19</td>
<td>6.06</td>
</tr>
<tr>
<td>1ch-WPE(2 inputs)-&gt;DUET</td>
<td><strong>0.59</strong></td>
<td><strong>1.99</strong></td>
<td>4.59</td>
<td>5.40</td>
<td>1.05</td>
<td><strong>5.12</strong></td>
</tr>
</tbody>
</table>

Figure 4.16: A system that converts the echoic condition into an approximate anechoic condition, using WPE dereverberation for two channels, and then exploits the DUET algorithm for separation. Further noise reduction algorithms can be applied to the outputs.

a system of figure 4.14 which cascades several stages together and that can cause distortions. Instead, a simple two stage system of dereverberation and beamforming can outperform. More interestingly, what we expected about the echoic to anechoic condition changing, actually occurs and we see that the DUET algorithm, even without further noise reduction processes can effectively separate the signals with a considerable noise and reverberation reduction and preserving the speech naturalness.

More studies has been accomplished, to achieve the separation based on the deeper level of speech contents, such as phoneme detections and discriminant features for different purposes which are not included due to incompleteness of the results.
Bibliography


[102] E. Habets, “Multi-channel speech dereverberation based on a statistical model of late re-


[104] B. W. Gillespie, H. S. Malvar, and D. A. Florêncio, “Speech dereverberation via maximum-

[105] J.-D. Polack, “Playing billiards in the concert hall: The mathematical foundations of ge-


echo suppression of speech signals in noisy environments,” Audio, Speech, and Language

based on quasi-periodicity of speech signals,” in Advances in Neural Information Pro-

[109] I. Kodrasi, T. Gerkmann, and S. Doclo, “Frequency-domain single-channel inverse filter-
ing for speech dereverberation: Theory and practice,” in Acoustics, Speech and Signal

[110] T. Gerkmann and R. C. Hendriks, “Noise power estimation based on the probability of
speech presence,” in Applications of Signal Processing to Audio and Acoustics (WASPA),

based on selective cepstro-temporal smoothing,” in Acoustics, Speech and Signal Process-


