Imperial College London

Department of Electrical and Electronic Engineering

Final Year Project 2017: Final Report

Project Title:	**Quality-preserving Speech Intelligibility Enhancement using a Kalman Filter**
Student:	**Jia Ying Goh**
CID:	**00749529**
Course:	**4T**
Project Supervisor:	**Brookes, D.M.**
Second Marker:	**Evers, C.**

Contents

Nomenclature

LMS	Least Mean Squares
NLMS	Normalised Least Mean Squares
RLS	Recursive Least Squares
STFT	Short-Time Fourier Transform
MMSE	Minimum Mean Squared Error
VAD	Voice Activity Detector
IBM	Ideal Binary Mask
TBM	Target Binary Mask
AMS	Analysis-Modification-Synthesis
KF	Kalman Filter
TDKF	Time-Domain Kalman Filter
MDKF	Modulation-Domain Kalman Filter
PESQ	Perceptual Evaluation of Speech Quality
STOI	Short-Time Objective Intelligibility
SNR	Signal-to-Noise Ratio
SNRseg	Segmental Signal-to-Noise Ratio
fwSNRseg	Frequency-Weighted Signal-to-Noise Ratio
MOS	Mean Opinion Score

Chapter 1

Introduction

1.1 Motivation

In today's highly interconnected world, communication between people, as well as with the world around them, is a major and critical aspect of their lives. Among the methods of communication (including but not limited to speech, text, images and bodily cues), speech generally stands out as the most efficient. Other methods such as visual indicators are sometimes useful to communicate ideas and thoughts, but a complex message is often best brought across via speech.

Applications utilising speech are thus widespread and numerous, and are generally designed to make use of clean speech. In a real-world environment, however, when speech is recorded, the recording inherently picks up not just the speech signal of interest, but also undesired background noise and channel noise. This damages the quality and intelligibility of the recorded speech, which poses a major problem for these applications requiring undamaged speech. Speech enhancement is hence often needed, with the goal of restoring the desired speech signal from the noisy mix, ideally by eliminating this noise while retaining the quality and intelligibility of the original speech signal.

There are various types of noise, including but not limited to additive noise, convolution noise and transcoding noise [1]. Additive acoustic noise that is uncorrelated with the speech signal generally degrades the intelligibility and quality of the perceived speech, and in cases of large noise may dominate and mask out the original speech. Convolution noise, on the other hand, manifests as reverberation, which is introduced by acoustic reflection, degrading intelligibility. Unlike additive noise, reverberation is highly correlated with the speech signal. Finally, transcoding noise can occur due to amplitude clipping in a microphone and appears as distortion. This report is concerned with the removal of additive acoustic noise.

Speech enhancement methods can be broadly classified into two types. Single-channel methods consider a single signal source. On the other hand, multi-channel methods consider multiple speech signals obtained from multiple microphones, where additional noise reduction can be achieved using information unavailable to a method relying on a single source, such as phase alignment from multiple microphones, leading to better overall noise reduction. However, this introduces additional costs and complexity, and in many applications such as hearing aids and mobile phones, single-channel methods are necessary due to constraints such as size. This report focuses on single-channel speech enhancement methods.

However, speech enhancement is complex. Traditional speech enhancement techniques such as spectral subtraction have very successfully improved speech quality by attenuating noise, but they tend to introduce speech spectral distortion [2], thus damaging its intelligibility. This project therefore aims to modify existing techniques to improve both the quality and intelligibility of speech.

1.2 Project Objectives

In this project, the objective is to improve both speech quality and intelligibility by modifying an existing speech enhancement algorithm. Standard tests for quality and intelligibility will be used to quantify the enhanced speech, and these include the Perceptual Evaluation of Speech Quality (PESQ, [3]) and Short-Time Objective Intelligibility (STOI, [4]) respectively.

Specifically, this project aims to modify an existing speech enhancement algorithm based on a Kalman filter, by further including additional information obtained from a so-called “ideal binary mask”. The goal is to scale the predicted value in the Kalman filter and modify its variance by an amount pre-determined from training data. The desired outcome is that PESQ remains high and STOI increases.

1.3 Project Scope

This project assumes the binary mask is already provided, and how it is generated is out of scope of this project. This project focuses on incorporating a given estimated binary mask into an existing Kalman filter speech enhancement implementation.

This project makes use of MATLAB and signal processing techniques. In particular, the project utilises VOICEBOX, a speech processing toolbox for MATLAB [5], which is included in the Imperial College London Software Library.

1.4 Report Overview

CHANGE THIS ******************

This interim report is categorised into four main chapters. Chapter 1 focuses on introducing and providing context to the problem, as well as providing a high-level overview of the project objectives. Chapter 2 describes the background information required for this project, offering more detail regarding the algorithms used.

Chapter 3 describes the implementation plan, identifying the milestones and timeline for the remainder of the project. This includes a summary of completed project work and identifies a checklist of upcoming tasks. Finally, Chapter 4 details the expected measures of success for the project.

Chapter 2

Background

The world that we live in today contains a lot of noise, originating from sources such as vehicles and babble from other human speakers. In the numerous applications that utilise microphones, including telecommunications, speech recognition software and hands-free communications, the desired signal can be significantly degraded by background noise. This noise damages the signal's quality and intelligibility. In many cases, this noise degradation is undesirable and unavoidable. Therefore, the noisy signal needs to be processed before it is useful for transmission or storage [6].

Traditionally, speech enhancement algorithms for noise reduction can be grouped into three main categories: noise reduction via filtering techniques, noise reduction via spectral restoration, and speech-model-based noise reduction methods [7]. Overall, speech enhancement techniques aim to improve the speech using audio signal processing techniques; some widely-used methods are described in this section.

2.1 Enhancement Domains

Speech enhancement can be performed in one of several domains. The following sections briefly describe these domains.

2.1.1 Time Domain

In the time domain, speech is usually enhanced using fixed or adaptive filtering techniques [8]. Fixed filters require prior knowledge of both the clean signal and noise, while this is not required for adaptive filters, which are able to adjust their parameters according to an optimisation algorithm, with little to no knowledge of the signal or noise characteristics, thus being more practical.

There are different approaches to adaptive filtering, one of which is the Least-Mean-Squares (LMS) filter. LMS algorithms aim to mimic a desired filter by finding a set of filter coefficients to minimise the mean squared error, where the error is the difference between the desired and actual signal [9].

The basic idea is to iteratively update the filter coefficients to approach the optimum coefficients, using a certain step size at each iteration. The LMS is a stochastic gradient descent approach, meaning that it is adapted based on the current error. It is, however, sensitive to input scaling, making it difficult to find an optimum step size to guarantee convergence. This limitation motivated the development of a variant, the Normalised Least Mean Squares (NLMS) algorithm, which is a variant of LMS that solves this problem by normalising with the power of the input [10].

Another popular approach is the Recursive Least Squares (RLS) algorithm, which recursively finds the filter coefficients to minimize a weighted least squares cost function relating to the input signal. This is unlike LMS and NLMS, which aim to reduce the mean squared error. Compared to LMS and NLMS, the RLS exhibits very fast convergence, but at the cost of higher computational complexity.

2.1.2 Time-Frequency Domain

Speech enhancement can be performed in the time-frequency (T-F) domain, which analyses signals in both time and frequency domains simultaneously, using various T-F representations [11]. Assuming speech is quasi-stationary over sufficiently short periods [12], the noisy input speech signal is divided into overlapping short frames, typically using a Hamming window, where the frame length is a compromise between temporal and frequency resolution [13]. These frames will be called acoustic frames, and are separate from the modulation frames referred to in the modulation domain in Section 2.1.3. Performing the Fourier transform on these frames produces a T-F matrix, on which processing can then be done. This entire process is called the short-time Fourier Transform (STFT).

Generally, T-F enhancement methods apply a gain function to suppress T-F regions which are noise-dominated while preserving speech-dominated regions, typically on the magnitude spectrum only. Computing the gain function depends on the noisy power spectrum, which needs to be estimated separately. After processing, inverse STFT followed by overlap-add reconstruction is performed to produce the enhanced time-domain speech signal. This approach works because speech is relatively sparse, due to limitations of the human ability in terms of speaking and listening i.e. with reasonable levels of noise, the speech can be divided into speech-dominated and noise-dominated regions.

Although this approach can improve the calculated signal-to-noise (SNR) of noisy speech, it can lead to undesired “musical noise” artifacts. These appear as isolated spectral components of noise and manifest as brief tones in the enhanced speech, which are generally deemed unnatural and disturbing [14]. This is because the amplitude of the short-time spectrum exhibits large fluctuations in noisy regions. After processing, the enhanced spectrogram consists of randomly located spectral peaks corresponding to the maxima of the original spectrogram, where the regions between these peaks have been suppressed as they are close to or below the averaged estimated noise spectrum. The result is residual noise comprising of sinusoids of random frequencies between each time frame.

This is used in the setup of the algorithm described in Section 2.6.1.

2.1.3 Modulation Domain

Modulation-domain processing starts off similarly to T-F processing, in that the noisy input signal undergoes STFT analysis to produce time-varying frequency components.

For speech enhancement, the amplitudes envelope of each frequency band is regarded as one modulation signal; the spectral amplitudes of each frequency band are windowed into overlapping modulation frames, with a separate modulation frame length and frame overlap compared to the acoustic frame length and overlap of STFT, where each acoustic frame provides one modulation-domain sample for each frequency bin. If each modulation frame contains M samples (i.e. M acoustic frames form one modulation frame) and each acoustic frame contains N time-domain samples, each modulation frame is constructed from MN time-domain samples. The modulation-domain signal has a frequency determined by the acoustic frame increment: since each acoustic frame provides one modulation sample, successive modulation samples are spaced apart by the acoustic frame shift. If the time-domain signal has a sampling frequency of f_s Hz and the acoustic frame shift is LH z, the modulation-domain sampling frequency is (f_s/L) Hz.

A processing algorithm then estimates the modulation frames of clean speech, which are then overlap-added to form the modified modulation signals. Combining this with the phase spectrum of the noisy input signal and performing the inverse STFT then produces the enhanced time-domain speech signal.

Even though the acoustic envelope directly contains the speech information, the temporal dynamics of the envelope better represent the information contained in speech [15]. These dynamics, which are at significantly lower frequencies than the speech signal itself, are provided in the modulation spectrum, suggesting that working in the modulation domain for speech processing can produce better results. More detailed description of modulation-domain-based speech processing is provided in Section 2.6.

2.2 Noise Estimation

Noise estimation is an important part of speech processing. In many algorithms including those described in this report, performance is heavily affected by the accuracy of the noise estimation.

One method to estimate the noise power spectral density is to use a minimum mean-squared-error (MMSE) optimal estimation method [16], which can be interpreted as a voice activity detector (VAD)-based noise power estimator [17]. A VAD detects when speech is present or absent, and can be used to update the noise estimate when speech is absent. To detect speech presence or absence, the algorithm must distinguish between speech and noise, which requires known or assumed information about how they differ. The estimator in [17], which improves on the estimator in [16], uses a fixed a priori SNR as a parameter of the likelihood of speech presence, using a value that is typical in speech presence. It was modified to be unbiased, while retaining similar performance and achieving a lower computational complexity compared to [16].

In this project, noise estimation is done using the MATLAB VOICEBOX routine `estnoiseg` [5], which implements the noise estimator in [17].

2.3 Spectral Subtraction

Spectral subtraction is a widely-used filtering technique which operates in the time-frequency domain. In this method, stationary or slowly-varying noise is attenuated from noisy speech by subtracting the magnitude noise spectrum, estimated during periods where speech is absent [18]. It is also possible to estimate the noise using a secondary sensor [8]. The estimated noise spectrum is then subtracted from the noisy spectrum to produce an approximated spectrum of the clean speech. The spectral error can then be computed and reduced separately. The algorithm can be further enhanced by incorporating residual noise reduction and non-speech signal attenuation [18], achieving even greater noise reduction.

Spectral subtraction works on the back of a few assumptions: firstly, that the background noise is additive to the clean signal [18]. This assumption means that the complex spectrum of the input noisy signal can be expressed as the sum of the speech spectrum and the noise spectrum. Next, it is assumed that the noise is a stationary or a slowly varying process (locally stationary). This allows the algorithm enough time to accurately formulate an updated estimate for the new noise magnitude spectrum before speech activity starts again. Lastly, the underlying assumption is that noise can be significantly reduced by removing its effect in the magnitude spectrum only i.e. phase spectrum is untouched, and the estimate of the clean speech magnitude spectrum is combined with the phase spectrum of the noisy input signal [19].

As mentioned in Section 2.1.2, the local stationarity assumption means the processing should be done on small-enough chunks of the input. Therefore, the input must first be split into overlapping frames using overlap-add processing. In the final step after processing, these frames are reassembled to form the continuous output signal.

To avoid signal distortion introduced by data segmentation [20], each frame is first multiplied by a windowing function before performing the Fourier Transform (typically using the Fast Fourier Transform or FFT). The output signal is then formed by the sum of these overlapping windowed frames. After processing, when the signal is being reassembled, the window is applied again.

For the signal to remain undistorted, multiplying by these windows should not change its magnitude. To achieve this, particular overlap factor/window pairs must be used; for example, if a Hamming window is chosen, applying the window twice requires that the overlapped windows approximately sum to unity for an overlap factor of 4 i.e. each windowed frame overlaps each of its neighbours by 50%, ensuring the output signal remains undistorted.

Spectral subtraction is popular largely because it is simple and easy to implement, requiring mainly the forward and inverse Fourier Transforms. However, this comes at a cost to performance. Subtracting the noise spectrum from the noisy input spectrum introduces distortion in the signal, known as musical noise [1], as mentioned in Section 2.1.2. Variations have been developed in attempts to mitigate this. A common variation involves over-subtraction and a noise floor. This method involves an over-subtraction factor, whereby an overestimate of the noise power spectrum is subtracted from that of the input, and using a noise spectral floor, which prevents the processed spectrum from going below a preset minimum value, to control both the amount of residual noise and musical noise [21]. However, it is generally evaluated that these modifications improve speech quality further but do not significantly affect the intelligibility of the input signals [1].

2.4 MMSE Speech Enhancement

2.5 Ideal Binary Mask

Sound is generated by acoustic sources, and these sources are typically complex, containing multiple frequency components. In a typical environment, multiple acoustic sources are simultaneously active, including undesired background noise, and a listener's ear will pick up only the sum of all these sources. There are various types of corrupting background noise, including but not limited to acoustic noise (e.g. vehicle vibration), speech-shaped noise, industrial noise and multi-talker babble (e.g. noisy cafeteria with other speakers) [22]. For the listener to distinguish between the different sounds in the incoming mix, such as picking out a particular speaker in a busy supermarket, the incoming audio signal has to be partitioned and categorised accurately into individual sounds.

Human beings have auditory systems that are remarkably capable at doing this; humans are thus generally able to understand speech in many of these noisy conditions. The signal separation process, known as auditory scene analysis, is typically performed in two stages, to understand the message spoken by the target speaker. Firstly, the input sound is decomposed into a matrix of time-frequency (T-F) units, where each unit represents the signal occurring at a particular instance in time with a particular frequency component. These T-F units are then analysed, and the auditory system utilises a combination of cues, learned patterns and other prior knowledge about the target to pick out the T-F units of the target signal, and group these individual components into a single recognisable "image" of the desired signal [23]. Essentially, the auditory system employs an analysis-synthesis strategy to organise the input into separate streams corresponding to different audio sources.

To model the human auditory system, computational auditory scene analysis (CASA) was proposed to approach sound separation in two stages: segregation and grouping [24]. The aim of using these CASA techniques was to pick out the target signal from the noisy mix, and the computational method of choice was the ideal T-F binary mask [25].

The ideal binary mask (IBM) is defined in the T-F domain as a matrix of binary numbers, and is constructed by comparing the local signal-to-noise ratio (SNR), defined as the difference between the target signal energy and the noise energy, in each T-F unit against a threshold known as a local criterion (LC). In the IBM, the T-F units with local SNR exceeding the LC (in decibels) are assigned 1, and 0 otherwise. If a 0dB SNR threshold is used to generate the mask, a T-F unit being assigned 1 indicates that the energy of the target signal is stronger than that of the interference (masker) within that particular T-F unit, which is a particularly intuitive implementation. Let $T(t,f)$ and $M(t,f)$ denote the target and masker signal power measured in dB respectively, at time t and frequency f; the IBM is then defined as

$$IBM(t,f) = \begin{cases} 1 & \text{if } T(t,f) - M(t,f) > LC \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

This mask can then be applied to the T-F representation of the incoming noisy signal; it acts as a selective filter, allowing some parts of the signal to pass through (those T-F units assigned to 1) while eliminating other parts (those assigned to 0). This means that at each T-F unit, the IBM either retains target energy or discards interference energy. The IBM therefore offers an indication of the T-F areas of audible target speech, and offers significant improvements in intelligibility [26].

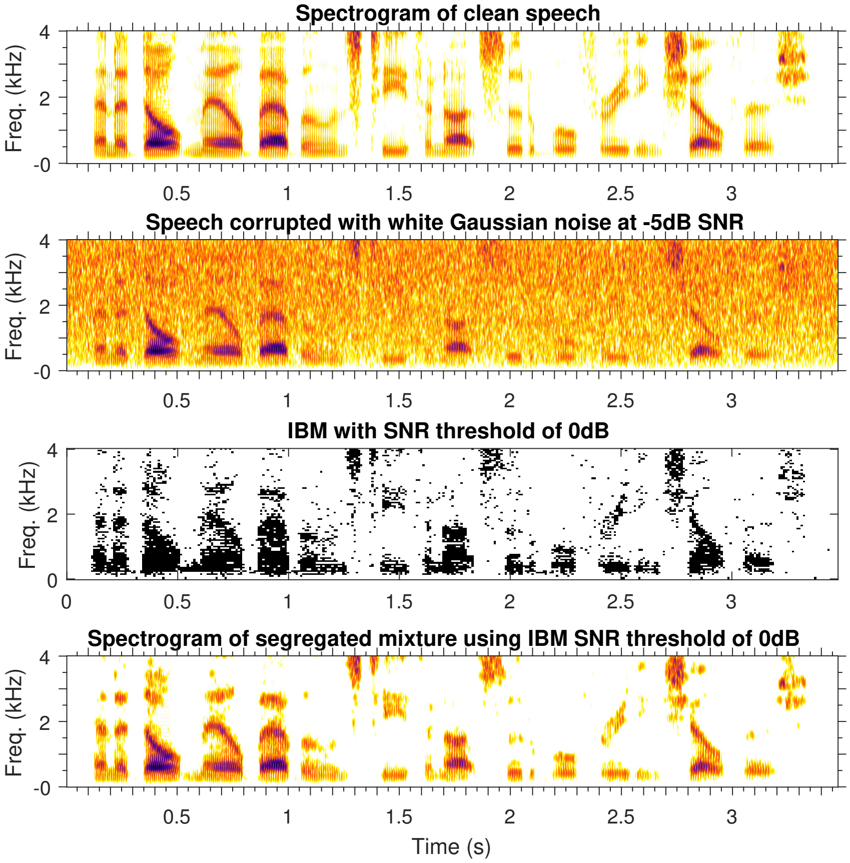

Figure 2.1: Top to bottom: clean speech, noisy speech, IBM and IBM-processed speech

An example of the IBM at work is shown in Fig. 2.1, with a clean sentence obtained from the TIMIT database [27]. From top to bottom, the spectrograms shown are that of: a) clean speech; b) clean speech corrupted with white Gaussian noise at -5dB SNR; c) IBM constructed using LC threshold of 0dB, where black pixels denote 1 (target stronger than interference masker) and white pixels denote 0 (target weaker than masker); d) segregated mixture obtained with the 0dB LC IBM, obtained by multiplying the spectrograms in (b) and (c), one T-F unit at a time.

The 0dB LC IBM, a particularly simple and intuitive comparison, is theoretically optimal in terms of SNR gain ([28], [29]); Fig. 2.1 shows its good performance, whereby the spectrogram of the processed speech is nearly identical to that of clean speech. It was later shown that while it is not optimal due to certain constraints, it performs almost as well as the proposed alternative, and is in fact more practical for real-world implementation [30]. Multiple studies have examined further the effects of the LC, input SNR level and masker type on the performance of the IBM. For example, a technique called ideal T-F segregation (ITFS) has been effective in making use of the IBM to improve the intelligibility of human speech masked by competing voices [26]. It is argued that the ITFS removes informational masking caused by the IBM-eliminated T-F units with large masker energy, where informational masking refers to the inability to accurately distinguish the target signal from the noisy mixture.

To demonstrate the benefits of IBM processing, various studies carried out intelligibility tests, in which listeners listen to a set of IBM-processed sentences and write down the words they hear; results produced are in terms of the percentage of words identified correctly.

A typical test result looks like Fig. 2.2, where UN represents the unprocessed noisy speech (replicated from [31]). In this example, the short-time Fourier Transform was used to process the input noisy signal, where multitalker babble was used as the masker [31]. As shown, the performance peaks out between approximately −20dB and 5dB for an input SNR of −5dB, and the range is slightly smaller for an input SNR of −10dB.

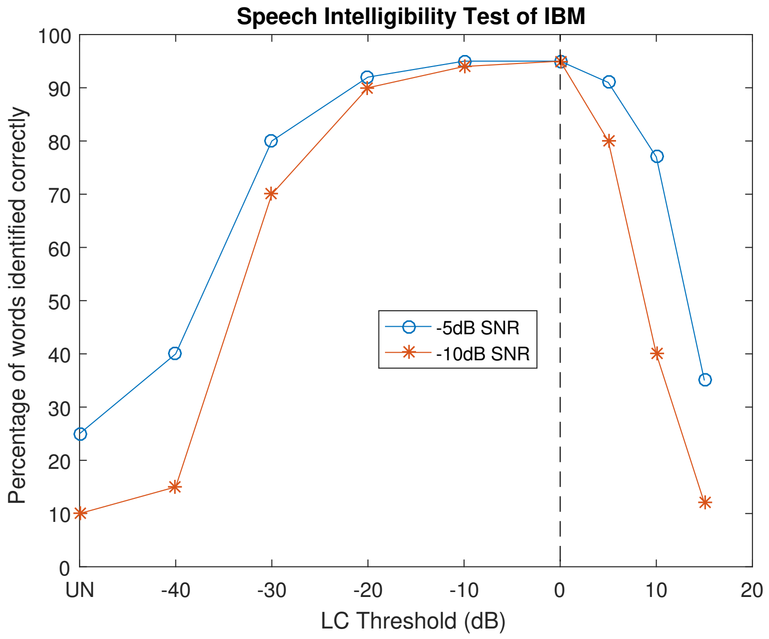

Figure 2.2: Performance (percentage of words identified accurately) as a function of LC (dB) for two input SNR levels, masked in multitalker babble (replicated from [31])

Large intelligibility benefits were demonstrated in [31], but they came up with a range of LC values for near-perfect intelligibility (performance plateaus of near 100% accuracy) that were different to that in [26]. Attributing this to differences in the setup and signals used, it was suggested that the pattern of the IBM was the critical factor for intelligibility, rather than the local SNR of individual

T-F units [31].

The significant improvements to intelligibility made IBM a notable candidate for speech enhancement applications such as hearing aids, provided the IBM could be approximated to a high degree of accuracy. However, to apply it, it is important to understand how IBM enhances intelligibility. In [26], it is argued that the IBM suppresses informational masking by directing the listener's attention to the T-F units containing target information i.e. *where* the target signal is, in a T-F auditory space [31]. This led to the conclusion that listeners need not extract specific knowledge from individual T-F units, but rather the overall pattern of the IBM, i.e. pattern of target-dominated and masker-dominated T-F units, was the most important factor for intelligibility, which was also concluded in [31]. However, this interpretation is limited to the range of LCs where the IBM pattern represents the T-F units that are audible to normal human listeners i.e. LCs close to 0 dB [32].

An alternative ideal mask definition was proposed in [33], which also produced large intelligibility improvements. This alternative mask was named the target binary mask (TBM), as the mask was calculated based on the target signal only. TBM depends on the long-term spectrum of the target speaker, and compares with an average spectrum i.e. a time-invariant threshold. The mask pattern naturally resembles the target signal and is unaffected by the masker specifically. Instead, the TBM generated in this manner can be applied to a mixture of the target signal and a different masker. On the other hand, the IBM pattern depends on the masking signal; IBM compares with the actual noise in the T-F units, which is time-dependent.

In certain applications, it may be easier to estimate the TBM than the IBM, and so it was of interest to investigate the intelligibility performance of the TBM: it was shown that the TBM has comparable performance to the IBM [34]. A noise-robust method based on target sound estimation to estimate the TBM was proposed in [35].

2.5.1 Musical Noise

The largest calculated SNR gain is achieved by using a fully-binary mask of 1s and 0s. However, doing so generally degrades the quality of the enhanced speech due to the introduction of musical noise, which refers to random, short tone-like bursts that, in some situations, can be more bothersome than the original noise. In [33], the mask values were instead 1 and 0.2: if the mask indicates that speech is present, the gain is 1, otherwise the gain is set to 0.2 instead of cutting completely. By doing so, the attenuation is limited to less than 20dB, which reduced the overall amount of musical noise introduced by the mask.

2.5.2 Practical Considerations

By definition, the IBM depends on oracle knowledge, as the mask is constructed based on the target and interfering signals before mixing. In a real-world situation, the target signal is of course unavailable, meaning the IBM has to be estimated from noisy data only. In the presence of significant noise, this can be a difficult task, and it is impossible to compute the IBM for all T-F units with complete accuracy. The effect of overall binary mask estimation error was investigated further in [31], and it was demonstrated that the estimation needs to be very accurate overall. As an example, $> 90\%$ accuracy is required to estimate the IBM for the case of -5dB input masked with multitalker babble to yield significant gains in intelligibility.

While it is of interest to further investigate the effects of estimation uncertainty and error on speech intelligibility improvements, this project focuses on the Kalman filter algorithm, and assumes that an ideal or estimated binary mask has already been computed and is available.

2.6 Kalman Filter

The Kalman filter [36] is a recursive optimal data processing algorithm. Under certain assumptions, it is optimal with respect to any practical measure. This is because the Kalman filter (KF) makes use of all data available to it, processing all available information to estimate the current value of the desired variables. In the context of speech enhancement, speech signals are modelled as autoregressive processes using the state space method, where the processed speech is recursively estimated, one sample at a time [37].

The filter has a recursive "predictor-corrector" structure [38]; firstly, a prediction of the desired variable at the next measurement time is made, based on all previously available data, producing a prediction value and its associated uncertainty. When the next measurement is actually taken, the difference between the measurement and the predicted value is used to "correct" the prediction, to produce the new estimate. Note that this recorded measurement comes with its associated uncertainty, arising from imperfections of measuring instruments. The new estimate is thus updated using a linear combination of the prediction and the measurement, with more weight given to estimates with lower uncertainty.

The KF was initially proposed for speech enhancement by Paliwal and Basu in 1987 [39], where excellent noise reduction was achieved when linear prediction coefficients (LPCs) were estimated from clean speech. The KF is of particular interest for speech enhancement, as the speech model is inbuilt into the KF recursion equations, and the enhanced speech contains no musical noise, assuming clean LPCs are available [40]; the performance of the KF is highly dependent on the accurate estimation of LPCs. However, for practical use, these parameters have to be estimated from noisy speech since the clean speech is not known a priori, causing a significant drop in performance. Better performance has been demonstrated in variations of the original KF algorithm, including a cascaded estimator/encoder structure which improves LPC estimates [41] and a subband KF algorithm that achieves better performance and reduces computational complexity [37] than the original KF method.

In recent years, the focus has shifted away from the traditional KF methods which utilise the acoustic domain, defined as the short-time Fourier Transform (STFT) of the signal. Instead, there has been growing interest in the modulation domain, defined as the variation over time of the magnitude spectrum at all acoustic frequencies [42]. Studies have increasingly shown the importance of the modulation domain for speech analysis; for example, very low frequency modulations of sound have been shown to be the fundamental carriers of information in speech [42], due to physiological limitations on how rapidly the vocal tract is able to change with time [43]. The slowly-varying modulation domain hence represents how the vocal tract changes over time [44].

The KF is capable of handling non-stationary signals as well as estimating both magnitude and phase spectra [45], which puts it at an advantage over STFT-based, acoustic domain-based methods for speech processing, as phase information has been shown to be more important in the modulation domain than in the acoustic domain [46]. It was also noted in [44] that the low order linear predictor KF was more appropriate for enhancing slower-varying modulating signals than for enhancing time-domain speech, as the time-domain signals contain long-term correlation which the low order linear predictor cannot capture. This is important for the KF, as its optimality works on the basis of incorporating and using all data available to the algorithm. These results suggest the use of the KF in the modulation domain as an improved method of speech enhancement [44].

2.6.1 Modulation-domain Kalman filter

The modulation-domain KF (MDKF) is an adaptive minimum mean-squared error (MMSE) estimator that uses the statistics of time-varying changes in the magnitude spectrum of both speech and noise [44]. In the MDKF, an analysis-modification-synthesis (AMS) framework is used to obtain the modulation domain in three steps. In the analysis stage, the input speech signal is processed using STFT; next, the noisy input spectrum undergoes some modification or processing; and lastly, the output processed signal is synthesised by inverse STFT followed by the overlap-add method.

Analysis-modification-synthesis framework in the acoustic domain

Considering an additive noise model, where $y(n)$, $x(n)$ and $v(n)$ represent zero-mean signals of noisy speech, clean speech and noise respectively:

$$y(n) = x(n) + v(n) \tag{2.2}$$

Assuming speech is quasi-stationary means that it can be analysed in frames using the STFT (analysis), thus obtaining the STFT of the noisy signal $y(n)$:

$$Y(n,k) = \sum_{l=-\infty}^{\infty} y(l)w(n-l)e^{-j\frac{2\pi kl}{N}} \tag{2.3}$$

which can be represented using STFT analysis as Equation 2.4:

$$Y(n,k) = X(n,k) + V(n,k) \tag{2.4}$$

where $Y(n,k)$, $X(n,k)$ and $V(n,k)$ denote the STFTs of noisy speech, clean speech and noise respectively and k refers to the discrete acoustic frequency index, N is the acoustic frame duration in number of samples and $w(n)$ is a window analysis function. For speech enhancement, a Hamming window is typically used. Note that this model is noise-additive in the complex STFT domain.

Each one of $Y(n,k)$, $X(n,k)$ and $V(n,k)$ is a complex spectrum, and can be expressed in terms of their acoustic magnitude and acoustic phase spectra. For example, $Y(n,k)$ can be represented as:

$$Y(n,k) = |Y(n,k)|e^{j\angle Y(n,k)} \tag{2.5}$$

where $|Y(n,k)|$ is the acoustic magnitude spectrum and $\angle Y(n,k)$ is the acoustic phase spectrum.

Traditionally, AMS-based methods only modify the noisy acoustic magnitude spectrum $|Y(n,k)|$ to obtain a processed magnitude spectrum $|\hat{X}(n,k)|$; the modified spectrum is thus obtained by combining the enhanced magnitude spectrum with the original noisy phase spectrum $\angle Y(n,k)$:

$$\hat{X}(n,k) = |\hat{X}(n,k)|e^{j\angle Y(n,k)} \tag{2.6}$$

The enhanced speech $\hat{x}(n)$ is then reconstructed by performing the inverse STFT of the enhanced acoustic spectrum $\hat{X}(n,k)$ followed by synthesis windowing and overlap-add [47].

Kalman filter model in the modulation domain

In the modulation domain, the acoustic magnitude spectrum of noisy speech is interpreted as a series of modulating signals spanning across time, where each modulating signal $|Y(n,k)|$ represents the variation of one frequency component over time, with $k = 1, 2, ..., N$ where N is the number of frequency bins. Each modulating signal is individually processed with a separate KF [44].

To visualise this, imagine that a time-domain noisy speech signal is windowed with a 64ms frame (window) length and 4ms frame shift. Taking the STFT, each window is analysed individually: the samples within a 64ms window are viewed as a frequency-domain signal with (for example) 256 frequency bins. When the next window is taken (original window shifted by 4ms), the samples are again analysed into a set of 256 frequency bins. Doing this for the entire signal produces 256 time-varying signals (modulating signals), one for each frequency component and processed with its own KF, where the samples in each signal are 4ms apart. Within each KF, the modulating signal is further windowed, but the signal now has a much lower frequency: in this case, $\frac{1}{0.004} = 250$Hz. Assuming a modulating window of 64ms, each window only contains $\frac{64}{4} = 16$ samples, compared to 512 samples for a 64ms window of a 8kHz time-domain signal.

Going back to the model, recall that the phase spectrum is left untouched, and an additive noise model is assumed for each modulating signal, assuming white Gaussian noise, giving Equation 2.7. Note that this assumes that the speech and noise are additive in the STFT magnitude domain rather than the STFT complex domain from Equation 2.4.

$$|Y(n,k)| = |X(n,k)| + |V(n,k)| \tag{2.7}$$

In the KF autoregressive model, a p-order linear predictor is used to model the evolution of speech over time, as shown in Equation 2.8, where $a_{j,k}; j = 1, 2, ..., p$ are the LPCs and $W(n,k)$ is a random white excitation with a variance of $\sigma^2_{W(k)}$.

$$|X(n,k)| = -\sum_{j=1}^{p} a_{j,k}|X(n-j,k)| + W(n,k) \tag{2.8}$$

Including the noise signal, the overall state space representation for noisy speech can be written as:

$$\mathbf{X}(n,k)=\mathbf{A}(k)\mathbf{X}(n-1,k)+\mathbf{d}W(n,k) \tag{2.9}$$

$$|Y(n,k)|=\mathbf{d}^T\mathbf{X}(n,k)+|V(n,k)| \tag{2.10}$$

where $\mathbf{X}(n,k)=[|X(n,k)|,|X(n-1,k)|,...|X(n-p+1,k)|]^T$ is the clean speech modulation state vector, $\mathbf{d}=[1,0,...,0]^T$ is the measurement vector for both the excitation noise $W(n,k)$ and observation, and $\mathbf{A}(k)$ is the state transition matrix utilising the LPCs:

$$\mathbf{A}(k)=\begin{bmatrix}-a_{1,k} & -a_{2,k} & \dots & -a_{p-1,k} & -a_{p,k}\\ 1 & 0 & \dots & 0 & 0\\ 0 & 1 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \dots & 1 & 0\end{bmatrix} \tag{2.11}$$

The Kalman filter recursively calculates a linear unbiased MMSE estimate $\hat{\mathbf{X}}(n|n,k)$ of the k-th modulation state vector at time n, given the noisy modulating signal up to and including time n (i.e. $|Y(1,k)|,|Y(2,k)|,...|Y(n,k)|$) using the following equations:

$$\mathbf{P}(n|n-1,k)=\mathbf{A}(k)\mathbf{P}(n-1|n-1,k)\mathbf{A}(k)^T+\sigma^2_{W(k)}\mathbf{d}\mathbf{d}^T \tag{2.12}$$

$$\hat{\mathbf{X}}(n|n-1,k)=\mathbf{A}(k)\hat{\mathbf{X}}(n-1|n-1,k) \tag{2.13}$$

$$\mathbf{K}(n,k)=\mathbf{P}(n|n-1,k)\mathbf{d}[\sigma^2_{V(k)}+\mathbf{d}^T\mathbf{P}(n|n-1,k)\mathbf{d}]^{-1} \tag{2.14}$$

$$\mathbf{P}(n|n,k)=[\mathbf{I}-\mathbf{K}(n,k)\mathbf{d}^T]\mathbf{P}(n|n-1,k) \tag{2.15}$$

$$\hat{\mathbf{X}}(n|n,k)=\hat{\mathbf{X}}(n|n-1,k)+\mathbf{K}(n,k)[|Y(n,k)|-\mathbf{d}^T\hat{\mathbf{X}}(n|n-1,k)] \tag{2.16}$$

where $\sigma^2_{V(k)}$ is the variance of the corrupting noise and $\mathbf{P}(n|n,k)$ is the error covariance matrix. These equations can be categorised into two main steps: prediction and updating. Equations 2.12 and 2.13 predict the error covariance and state based on past samples respectively, while the other equations update the Kalman gain, error covariance and state based on the predicted values.

In particular, Equation 2.16 is the main updating step, whereby a linear combination of the estimate based on previous samples $|\hat{X}(n|n-1,k)|$ and the current measurement $|Y(n,k)|$ is used to compute the current estimate $|\hat{X}(n|n,k)|$. To view this more clearly, we can rewrite Equation 2.16 as:

$$\hat{\mathbf{X}}(n|n,k) = [\mathbf{I} - \mathbf{K}(n,k)\mathbf{d}^T]\hat{\mathbf{X}}(n|n-1,k) + \mathbf{K}(n,k)|Y(n,k)| \tag{2.17}$$

The accuracy of the weighted sum producing the updated state is critical in determining the correctness of the algorithm. It is therefore imperative that the error variances are estimated as accurately as possible. Particularly, the noise variance can be estimated in a number of ways.

**********UNFINISHED DESCRIPTION AND EXPLANATION

As the algorithm is running, each modulating signal $|Y(n,k)|$ is windowed into modulation frames, and the LPCs and excitation variance $\sigma^2_{W(k)}$ are estimated. Within each frame, the LPCs are kept constant, whereas the Kalman gain $\mathbf{K}(n,k)$, error covariance matrix $\mathbf{P}(n|n,k)$ and estimated state vector $\hat{\mathbf{X}}(n|n,k)$ are updated every sample, regardless of frame.

2.6.2 Comparison with time-domain Kalman filter

For comparison purposes, the time-domain Kalman filter (TDKF) equations are shown below:

$$\mathbf{P}(n|n-1) = \mathbf{A}\mathbf{P}(n-1|n-1)\mathbf{A}^T + \sigma^2_w\mathbf{d}\mathbf{d}^T \tag{2.18}$$

$$\hat{\mathbf{x}}(n|n-1) = \mathbf{A}\hat{\mathbf{x}}(n-1|n-1) \tag{2.19}$$

$$\mathbf{K}(n) = \mathbf{P}(n|n-1)\mathbf{d}[\sigma^2_v + \mathbf{d}^T\mathbf{P}(n|n-1)\mathbf{d}]^{-1} \tag{2.20}$$

$$\mathbf{P}(n|n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{d}^T]\mathbf{P}(n|n-1) \tag{2.21}$$

$$\hat{\mathbf{x}}(n|n) = \hat{\mathbf{x}}(n|n-1) + \mathbf{K}(n)[y(n) - \mathbf{d}^T\hat{\mathbf{x}}(n|n-1)] \tag{2.22}$$

where $\hat{\mathbf{x}}(n|n-1)$ and $\hat{\mathbf{x}}(n|n)$ are the *a priori* (predicted) and *a posteriori* (updated) state vectors respectively, $\mathbf{P}(n|n-1)$ and $\mathbf{P}(n|n)$ are the *a priori* and *a posteriori* error covariance matrices respectively, $\mathbf{K}(n)$ is the Kalman gain, and σ^2_v, σ^2_w are the noise and excitation variances respectively.

Fig. 2.3 compares the spectrograms of TDKF and MDKF applied on speech from the TIMIT database [27], corrupted by white Gaussian noise and sampled at 8kHz. For the purposes of comparing performance limits, clean speech LPCs were used in the filters. Generally, both algorithms perform well in removing noise, especially when speech is absent. However, there is visibly some noise in the TDKF-enhanced speech; particularly, frequency components above 1.8kHz have been noticeably degraded by noise. A listening test confirmed this, detecting the presence of high-frequency artifacts.

USE A CLEANER GRAPH HERE!!!!!!!

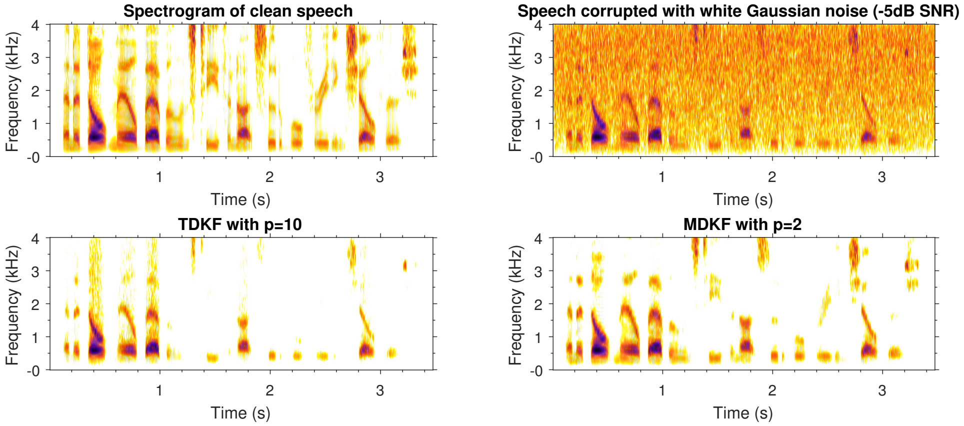

Figure 2.3: Top row: clean speech (left), speech corrupted with white Gaussian noise (right); bottom row: TDKF-enhanced speech (left), MDKF-enhanced speech (right)

This TDKF noise is a limitation of using the KF for speech enhancement. Similarly to how we rearranged Equation 2.16, Equation 2.22 can be rewritten to show that the enhanced output is a weighted combination of the estimated speech and measured speech, where the relative weight depends on the Kalman gain $\mathbf{K}(n)$. When speech is absent, $\mathbf{P}(n|n-1) = \mathbf{0}$, meaning that $\mathbf{K}(n) = \mathbf{0}$ and the estimated state is equal to the predicted state, being completely unaffected by the noisy measurement.

When speech is present, however, the algorithm does not work quite so perfectly. A typical TDKF model order is $p = 2 + f_s(inkHz)$, which in this case is 10, meaning that the model only uses short-term correlation information up to 10 adjacent samples. This TDKF linear predictor model is unable to fully replicate the harmonic structure of speech, requiring autocorrelation lags in the order of the number of samples in a pitch period [44]. The prediction thus has unvoiced and noise-like characteristics, and the result is that the updated output only preserves the speech component below 1.8kHz [44]. The resultant noise will be especially prevalent in regions of low SNR, where the prediction is weighted more heavily due to Equation 2.20 producing a smaller $\mathbf{K}(n)$.

INCOMPLETE

2.6.3 Performance of MDKF

Overall, experimental results from the TIMIT corpus (Fig. 2.3) show that under ideal conditions where clean speech LPCs can be obtained accurately, the linear predictor is sufficient to model the modulating signals of clean speech. As described earlier, the vocal tract tends to change slowly due to physiological constraints, and thus low LPC orders ($p = 2$) were found to be sufficient. Using this, the MDKF was by far the best performing algorithm, doing better than all acoustic and time-domain methods tested, including the TDKF, for both white and coloured noise [44]. This was despite both algorithms having access to clean speech LPCs.

However, clean speech is not available in reality; the presence of noise generally degrades the LPC estimates, worsening the performance of the MDKF algorithm. In [44], a practical MDKF algorithm was evaluated, which used an acoustic-domain pre-processor for LPC estimation to reduce the effect of noise degradation.

Chapter 3

Implementation Plan

The overall goal of this project is to modify an existing Kalman Filter (KF) speech enhancement algorithm by incorporating data obtained from an ideal binary mask (IBM) or target binary mask (TBM), by scaling its predicted value and variance by amounts determined from training data.

The overall implementation plan of this project can therefore be split into a few main parts: 1) implement an IBM/TBM algorithm; 2) calculate the IBM/TBM and note the parameters required for the most intelligibility gain; 3) implement an existing KF enhancement algorithm, and evaluate it using PESQ and STOI; and finally 4) modify the KF implementation to incorporate information from the IBM/TBM.

3.1 Completed Work

At this early stage, an IBM algorithm has been implemented, based on oracle data providing both the target and masker signals; the algorithm and its demonstration is based on [31].

To synthesise the mask, the target signal (clean) and noisy signal (mixture) were used. Both signals were processed using a Fast Fourier Transform (FFT) applied to 20ms segments of the signal, which were Hamming-windowed with 50% overlap between adjacent segments. The windowing and FFT were performed using algorithms from [5] and done in MATLAB. In IBM implementation, the masker signal is required; the masker spectrum was obtained by subtracting the clean spectrum from the noisy spectrum.

Using Equation 2.1, the energy of the target signal was compared to that of the masker. The resultant local SNR of each T-F unit was compared against a pre-determined LC threshold (in Fig. 2.1, 0 dB was used) to determine whether to retain the noisy mixture's T-F unit (binary mask value of 1) or not (mask value of 0). This unit-wise comparison produced a pattern of binary mask values consisting of 0s and 1s, and this mask was applied to the magnitude spectrum of the noisy signal using a simple unit-wise matrix multiplication.

Inverse-FFT was then applied to the resultant processed spectrum, with the phase spectrum of the original noisy spectrum being used. This was the exact inverse of the initial FFT processing, thus producing 20ms segments. The resultant time-domain waveform of this processed spectrum was thus generated using the overlap-add method, performed on these segments.

The results have been shown in Fig. 2.1 to be very good, and previously-discussed studies have illustrated that optimal performance depends on parameters such as the local criterion (LC) threshold, masker type and input signal-to-noise ratio (SNR). As discussed earlier, estimating the binary mask is out of scope of this project, and so an existing mask implementation will simply be selected and implemented. The choice of mask and its corresponding parameters will be critical in determining the eventual effectiveness of the modified KF algorithm.

Significant intelligibility gains were observed with IBM processing for a range of LC threshold values: the intelligibility of the -10 dB input mixture dramatically rose from 10% for the original noisy mixture to 95% (almost perfect intelligibility score) when processed using an IBM with an LC threshold of 0 dB. Similarly, the intelligibility of the -5 dB input signal increased from 25% for the original noisy mixture to 95% when processed using an IBM with an LC threshold of 0 dB (Fig. 2.2). Unsurprisingly, the plateau region for near-perfect performance was wider for the -5 dB input signal as compared to the -10 dB input signal.

3.2 Milestones

The remainder of the project can be set as the following overall milestones:

1) Calculate the IBM or TBM and note the parameters required for optimal intelligibility improvements

2) Implement an existing KF enhancement algorithm, and evaluate it using PESQ and STOI standards

3) Modify the KF implementation to incorporate information from the IBM/TBM, to provide a third piece of information for the KF predictor

3.3 Timeline

Given the milestones above, the next steps will involve tweaking the IBM to find its optimal performance parameters in a variety of situations, choosing between the IBM or TBM, implementing an existing KF speech enhancement algorithm, and modifying the KF algorithm. A table of achievables, along with their associated risks and expected dates, is shown below.

The first major next step is to implement the TBM, which requires a minor modification from the IBM. Parameters will need to be varied for both masks to find the optimal mask under specific conditions such as input SNR level, LC and masker type. To avoid unnecessary repeats, preliminary results can be obtained from previous studies, such as [31] and [33]. Once this has been determined, the binary mask is then available for use, and can be set aside for the time being.

Date	Objective
2/2/2017	Implement TBM and compare to IBM
	No associated risks; completed IBM requires minor tweaking to get TBM
10/2/2017	Determine optimal parameters and associated assumptions/conditions
	Process can be sped up by starting with known results
17/2/2017	Complete readings about KF
	Papers based on modified MDKF algorithms
24/2/2017	Implement existing ideal KF algorithm (TDKF, MDKF)
	This has been started, but progress has been slow
3/3/2017	Implement KF algorithm based on noisy LPC estimates
	When ideal KF algorithms have been implemented, should only require minor changes to incorporate noisy data estimates
10/3/2017	Determine optimal algorithm to use
	As with IBM, start off using known results
24/3/2017	Generate training data from IBM
	Risks currently unknown
28/4/2017	Incorporate training data into KF
	Use training data to generate scaling/shifting of KF-generated estimates
12/5/2017	Evaluate enhanced algorithm using PESQ and STOI
	Proper procedures (PESQ, STOI) required to evaluate modified algorithm

Table 3.1: Timeline of deliverables and associated dates

Next, the primary step is to implement an existing KF algorithm. Work is still in progress regarding background reading for this section, and a variety of KF algorithms for speech enhancement need to be implemented and compared with one another. Their advantages and disadvantages need to be assessed and a final algorithm should then be chosen. Based on [44], the modulation-domain KF is a good place to start; the paper compares a variety of different KF-based methods, and the MDKF was demonstrated to be the best-performing algorithm.

After that, the useful data needs to be generated from the IBM to be included into the KF algorithm. Currently, this step has not been evaluated in much detail, and what information gets incorporated into the KF may change slightly as the project progresses. Based on the ideal timeline, the final step would be to evaluate the modified algorithm using PESQ and STOI standards.

Chapter 4

Problem Analysis

Chapter 5

Improved Observation

From the MDKF iteration equations (Equations 2.12 to 2.17), we see that there are two variables used in forming the updated state: the prediction of the current state and the observation of the current (noisy) state. This observation is noisy; if the observation can be modified in some way to better represent the inherent speech (or silence in absence of speech), the algorithm can perform better. One way to do so is to include information from an IBM or TBM.

5.1 Incorporating IBM into observation

From Section 2.5, we know that the 0dB-threshold IBM produces a mask of 1s and 0s, where 1s represent T-F units where the signal has higher energy than noise, and vice versa for 0s. The noisy observation used in the KF equations [48]

Chapter 6

Incorporate Binary Mask into Kalman filter

The IBM can be used in a slightly different way; instead of tweaking the observation itself, information from the mask can be directly used in the KF equations. Now, the KF equations use three pieces of information (prediction, observation, binary mask) to estimate the current state, rather than just the former two as in the original KF equations.

6.1 Transformed Kalman Filter Equations

RMB TO CITE PAPER FROM MIKE BROOKES

For the scalar output case i.e. $|Y(n,k)|$ is a scalar, with $\mathbf{d} = [1, 0, ..., 0]^T$, the observation can be decorrelated from the rest of the state vector.

Chapter 7

Implementation

For the MDKF, an acoustic frame length of 32ms was used with a 4ms frame shift, giving a 250Hz sampling frequency in the modulation domain. For each frequency bin, a modulation frame of

$$mswasusedwitha$$

ms frame increment to determine the LPC coefficients. A model order of $p = 2$ was used. However, it was shown in [49] that short modulation frame durations of $10 - 20$ms retain good intelligibility overall as compared to longer frame lengths, and thus a modulation frame length of

$$mswasusedinthisproject, witha$$

ms frame shift.

white noise - overall estimate of noise better? so used an overall noise estimate for entire speech duration

Chapter 8

Testing

There are two aspects to speech quality: the overall perceived speech quality, and the speech intelligibility [50].

The perceived overall speech quality is how “good” the quality of the speech is. The definition of “good” is typically left to the listener, who then gives a score to the speech. On the other hand, speech intelligibility is the accuracy with which we can hear what is being said. Specifically, it is measured as the percentage of correctly identified words relative to the number of words. Instead of words, one may also use phonemes or syllables as the test unit. If words or complete sentences are used, they typically encompass linguistically meaningful units, and thus the choice of test words is important to ensure a fair assessment.

Although there does not exist a completely clear relationship between speech quality and intelligibility, there exists some correlation between the two. Generally, “good” quality speech also gives high intelligibility and vice versa. However, this generalisation does not always hold; there are some samples that are highly intelligible yet are perceived as “poor” quality and vice versa.

8.1 Speech Quality

Methods to assess speech quality can be grouped into subjective and objective measures.

8.1.1 Subjective Speech Quality Measures

Subjective quality measures typically compare the original and processed speech by a listener or a group of listeners. The listeners subjectively rate or rank the speech according to a predetermined scale. Since every listener is unique, their ratings will vary; this variation in results can be reduced by averaging the scores from a group of listeners.

Mean Opinion Score

A widely used subjective quality measure is the Mean Opinion Score (MOS) [51]. Each listener gives a numeric MOS score, typically in the range $1-5$, where 1 is the lowest perceived quality and 5 is the highest perceived quality. The "Absolute Category Rating" scale is commonly used, as shown in Table 8.1 [52]. The overall score is obtained by averaging the ratings from all listeners, representing an overall perceived quality of the speech. With a large number of speech files, this test can be costly and time consuming.

Rating	**Label**
1	Excellent
2	Good
3	Fair
4	Poor
5	Bad

Table 8.1: Categories of MOS: Absolute Category Rating

8.1.2 Objective Speech Quality Measures

On the other hand, objective speech quality use physical measurements and some calculated values from these measurements. Typically, these calculations compare objective measurements for the reference clean speech and the distorted speech.

Many of the objective measures are highly correlated with subjective measures; it is thus common for a test to use objective measures to estimate subjective methods, which are usually more time-consuming and costly as they involve human listeners. However, as noted previously, there are situations in which high objective scores do not produce high subjective scores and vice versa.

SNR

Signal-to-Noise Ratio (SNR) is one of the oldest and most widely used objective quality methods. It has low computational complexity, but requires both clean and distorted speech. The classic formula is calculated (in dB) as:

$$SNR = 10\log_{10}\frac{\sum\limits_{n=1}^{N} x^2(n)}{\sum\limits_{n=1}^{N}\{x(n)-\hat{x}(n)\}^2} \tag{8.1}$$

where $x(n)$ is the clean speech, $\hat{x}(n)$ is the distorted speech and N is the number of time-domain samples.

This classic SNR formula, however, is not well related to speech quality as it averages over the entire signal even though speech is non-stationary. Speech energy fluctuates over time, and this formula is dominated by parts where speech energy is large and noise energy is small, which is not representative of the entire signal.

Variations of SNR have thus been proposed. To better represent the temporal variation of speech, segmental SNR (SNRseg) was proposed to calculate SNR in short frames and take the average:

$$SNR_{seg} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \frac{\sum\limits_{n=Lm}^{Lm+L-1} x^2(n)}{\sum\limits_{n=Lm}^{Lm+L-1} \{x(n) - \hat{x}(n)\}^2} \tag{8.2}$$

where L is the frame length in number of samples and M is the number of frames in the signal $(N = ML)$. The logarithm of the ratio is computed before averaging; frames with unusually large ratios are hence weighted less while frames with lower ratios are weighted higher. This matches the perceptual quality better i.e. frames with large speech and low noise do not dominate the overall ratio.

However, if the speech contains too much silence, the overall SNRseg value decreases significantly since silent frames usually give large negative SNRseg values. In this case, silent frames should be excluded from the averaging by using speech activity detectors. In the same manner, excluding frames which exhibit excessively large or small speech values from the averaging produces SNRseg values that match the subjective quality better. Thus, SNRseg often has upper and lower bounds of 35dB and −10dB respectively [53].

A separate variation of SNR is the frequency-weighed SNR (fwSNRseg), which weights the contribution of the different frequency bands. The fwSNRseg can be defined as:

$$fwSNR_{seg} = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum\limits_{j=0}^{K-1} W(j,m) \log_{10} \frac{X(j,m)^2}{\{X(j,m) - \hat{X}(j,m)\}^2}}{\sum\limits_{j=0}^{K-1} W(j,m)} \tag{8.3}$$

where $W(j,m)$ is the weight of the j^{th} frequency band in the m^{th} frame, K is the number of frequency bands and $X(j,m)$, $\hat{X}(j,m)$ are the spectral amplitude of the clean and distorted speech respectively. The weights can be chosen in many ways, one of which is the ANSI SII Standard [54].

Perceptual Evaluation of Speech Quality

One of the most popular objective speech quality measures is the ITU-T P.862: Perceptual Evaluation of Speech Quality (PESQ) [3].

PESQ was developed to model subjective tests commonly used to assess the voice quality by human beings (e.g. MOS), using true voice samples as test signals. It is designed for use over a wide range of conditions. A mapping from PESQ to MOS scores was standardised, allowing PESQ results to model MOS scores that range from 1 (Bad) to 5 (Excellent) (typical of Table 8.1). The average correlation between PESQ-mapped MOS scores and subjective MOS for a number of tests was a high score of 0.935 [55]. The block diagram of PESQ is shown in Fig. 8.1 (taken from [56]).

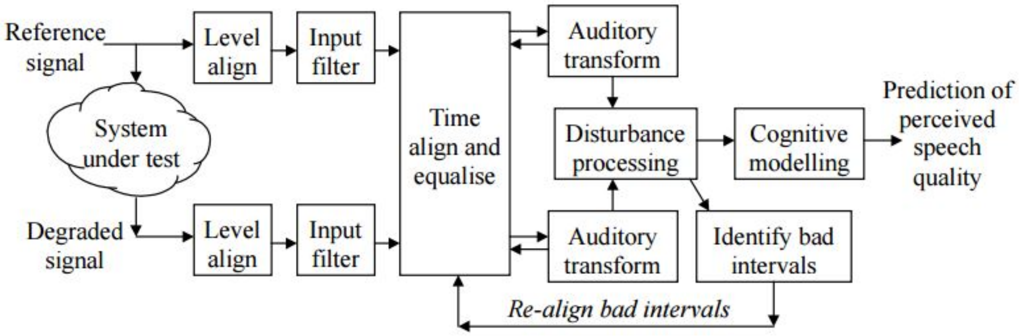

Figure 8.1: Structure of PESQ model (taken from [56])

In this report, the quality of the enhancement algorithms will be assessed using SNRseg and PESQ. SNRseg will be used to assess the effect of the enhancement on noise level while PESQ will be used to evaluate speech quality.

8.2 Speech Intelligibility

8.3 Speech Database

8.3.1 TIMIT

The TIMIT Acoustic-Phonetic Continuous Speech Corpus of read speech was designed to provide speech data for acoustic-phonetic studies as well as the development and evaluation of automatic speech recognition systems [27]. It is widely used in the research and testing of speech enhancement algorithms. A combined effort between the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI), the TIMIT database contains recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences, each of which is a few seconds long. The recordings were done with a microphone to create 16-bit resolution, 16kHz rate speech waveform files, and the database also includes time-aligned phonetic and word transcriptions.

REWRITE

For evaluating the algorithms proposed in this report, the core test set of the TIMIT database will be used which contains 16 male and 8 female speakers each reading 8 sentences for a total of 192 sentences all with distinct texts. This test set is the abridged version of the complete TIMIT test set which consists of 1344 sentences from 168 speakers. Also, in order to optimize the parameters of the algorithms, a development set is formed which consists of 200 speech sentences randomly selected from the test set of the TIMIT database and does not have any overlap with the core test set. The speech sentences in the development set are corrupted by white noise, car noise, factory noise, F16 noise and babble noise at SNRs between -10 and 15 dB at a 10

In this project, all speech sentences used were downsampled to 8kHz.

Chapter 9

Results

Chapter 10

Evaluation Plan

10.1 Deliverables

The primary deliverable is a modified Kalman filter-based speech enhancement algorithm, which takes into account information provided by an ideal/target binary mask. This adjustment should involve scaling the predicted value of the Kalman filter algorithm and tweaking its associated variance, and these adjustments should be based on values determined from training data from the binary mask. The results should be evaluated using PESQ and STOI.

10.2 Measures of Success

The measures of success and risks associated with each mini-goal are displayed in Table 3.1. Primarily, the goal of implementing and replicating known algorithms for IBM and MDKF is to verify the algorithm and its success in terms of PESQ and STOI, so these quality and intelligibility tests should produce similar results to that described in their papers.

Finally, the overall goal of a modified KF algorithm is to improve both the quality and intelligibility of speech. Using internationally-recognised standards, the desire is that PESQ remains high and STOI increases.

Chapter 11

Conclusion and Future Work

11.1 Future Work

11.1.1

for coloured noise/non-stationary noise assumption etc. better estimate of noise?

Bibliography

[1] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[2] Anuradha R. Fukane and Shashikant L. Sahare. "Different Approaches of Spectral Subtraction method for Enhancing the Speech Signal in Noisy Environments". In: *International Journal of Scientific & Engineering Research* 2 (2011).

[3] ITU-T P.862. *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.* 2001.

[4] Cees H. Taal, Richard C. Hendriks, and Richard Heusdens. "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech". In: *IEEE* (2010).

[5] Mike Brookes. *VOICEBOX: Speech Processing Toolbox for MATLAB.* 2012. URL: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[6] J. Benesty, S. Makino, and J. Chen (Eds.) *Speech Enhancement.* Springer, 2005.

[7] J. Benesty, M. M. Sondhi, and Y. Huang (ed). *Springer Handbook of Speech Processing.* Springer, 2007.

[8] Bernard Widrow et al. "Adaptive Noise Cancelling: Principles and Applications". In: *Proceedings of the IEEE* 63.12 (Dec. 1975).

[9] B. Widrow and M. E. Hoff. "Adaptive Switching Circuits". In: *IRE WESCON Convention Record* (1960).

[10] Jyoti Dhiman, Shadab Ahmad, and Kuldeep Gulia. "Comparison Between Adaptive Filter Algorithms (LMS, NLMS and RLS)". In: *International Journal of Science, Engineering and Technology Research (IJSETR)* 2 (2013).

[11] Leon Cohen. *Time Frequency Analysis: Theory and Applications.* Prentice-Hall, 1994.

[12] Michael D. Riley. *Speech Time-Frequency Representations.* Springer, 1989.

[13] Daniel W. Griffin and Jae S. Lim. "Signal Estimation from Modified Short-Time Fourier Transform". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1984).

[14] Olivier Cappe. "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor". In: *IEEE Transactions on Speech and Audio Processing* 2 (1994).

[15] Hynek Hermansky. *Modulation Spectrum in Speech Processing.* Springer, 1998.

[16] Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. "MMSE based noise PSD tracking with low complexity". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2010).

[17] Timo Gerkmann and Richard C. Hendriks. "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay". In: *IEEE Transactions on Audio, Speech and Language Processing* 20 (2012).

[18] Steven F. Boll. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". In: *IEEE Transactions On Acoustics, Speech and Signal Processing* 27 (1979).

[19] Saeed V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction*. Wiley, 2009.

[20] Nizamettin Aydin and Hugh S. Markus. "Optimization of processing parameters for the analysis and detection of embolic signals". In: *European Journal of Ultrasound* (2000).

[21] M. Berouti, R. Schwartz, and J. Makhoul. "Enhancement of speech corrupted by acoustic noise". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79* (1979).

[22] H. Kozou et al. "The effect of different noise types on the speech and non-speech elicited mismatch negativity". In: *Hearing Research* 199 (2005).

[23] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, Cambridge, MA, 1990.

[24] DeLiang Wang and Guy J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, 2006.

[25] DeLiang Wang. *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Speech Separation by Humans and Machines. Springer, 2005.

[26] Douglas S. Brungart et al. "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation". In: *Acoustical Society of America* (2006).

[27] John S. Garofolo et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. 1993. URL: `https://catalog.ldc.upenn.edu/ldc93s1`.

[28] Guoning Hu and DeLiang Wang. "Monaural speech segregation based on pitch tracking and amplitude modulation". In: *IEEE Transactions On Neural Networks* 15 (2004).

[29] Daniel P. W. Ellis. "Model-Based Scene Analysis". In: *Computational Auditory Scene Analysis: Principles, Algorithms, and Application* (2006). Edited by DeLiang Wang and Guy J. Brown.

[30] Yipeng Li and DeLiang Wang. "On the optimality of ideal binary time–frequency masks". In: *Speech Communication* 51 (2009).

[31] N. Li and P. C. Loizou. "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction". In: *Acoustical Society of America* (2008).

[32] Ulrik Kjems et al. "Speech Intelligibility of Ideal Binary Masked Mixtures". In: *European Signal Processing Conference* (2010).

[33] M. C. Anzalone et al. "Determination of the potential benefit of time-frequency gain manipulation". In: *Ear Hear* (2006).

[34] Ulrik Kjems et al. "Role of mask pattern in intelligibility of ideal binary-masked noisy speech". In: *Acoustical Society of America* (2009).

[35] Seliz Gulsen Karado et al. "Robust Isolated Speech Recognition Using Binary Masks". In: *European Signal Processing Conference* (2010).

[36] Rudolph Emil Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Transactions of the ASME–Journal of Basic Engineering* 82.Series D (1960).

[37] Wen-Rong Wu and Po-Cheng Chen. "Subband Kalman Filtering for Speech Enhancement". In: *IEEE Transactions on Circuits and Systems* 45 (1998).

[38] Peter S. Maybeck. *Stochastic Models, Estimation, and Control.* Vol. 1. Academic press, Inc., 1979.

[39] K.K. Paliwal and A. Basu. "A speech enhancement method based on Kalman filtering". In: *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* 12 (1987).

[40] Ning Ma, M. Bouchard, and R. A. Goubran. "Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14 (2006).

[41] Jerry D. Gibson. "Filtering of Colored Noise for Speech Enhancement and Coding". In: *IEEE Transactions on Signal Processing* 39 (1991).

[42] Les Atlas and Shihab A. Shamma. "Joint Acoustic and Modulation Frequency". In: *EURASIP Journal on Applied Signal Processing* (2003).

[43] Kuldip Paliwal, Kamil Wojcicki, and Belinda Schwerin. "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain". In: *Speech Communication* 52 (2010).

[44] Stephen So and Kuldip K. Paliwal. "Modulation-domain Kalman filtering for single-channel speech enhancement". In: *Speech Communication* 53 (2011).

[45] C. J. Li. "Non-Gaussian, Non-stationary, and Nonlinear Signal Processing Methods". PhD thesis. Aalborg University, Denmark, 2006.

[46] Steven Greenberg and Takayuki Arai. "The Relation Between Speech Intelligibility and the Complex Modulation Spectrum". In: *Proceedings of the 7th European Conference on Speech Communication and Technology* (2001).

[47] T. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice-Hall, 2002.

[48] M. Brookes. *The Matrix Reference Manual.* 1998-2011. URL: `http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html`.

[49] Kuldip Paliwal, Belinda Schwerin, and Kamil Wojcicki. "Role of modulation magnitude and phase spectrum towards speech intelligibility". In: *Speech Communication* 53 (2011).

[50] Kondo Kazuhiro. *Subjective Quality Measurement of Speech.* Springer, 2012.

[51] ITU-T P.830. *Subjective performance evaluation of telephone band and wideband codecs.* 1996.

[52] ITU-T P.800. *Methods for subjective determination of transmission quality.* 1996.

[53] John H. L. Hansen and Bryan L. Pellom. "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms". In: *Proceedings of the International Conference on Speech and Language Processing* (1998).

[54] American National Standards Institute (ANSI). *Methods for calculation of the speech intelligibility index.* 1997.

[55] A. W. Rix et al. "Objective Assessment of Speech and Audio Quality: Technology and Applications". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14 (2006).

[56] Antony W. Rix et al. "Perceptual Evaluation of Speech Quality (PESQ) - A New Method For Speech Quality Assessment of Telephone Networks and Codecs". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2001).

Appendix A

MATLAB