Imperial College London

Department of Electrical and Electronic Engineering

Final Year Project 2017: Interim 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.**

Contents

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 therefore widespread and numerous. These applications, such as mobile phones and hearing aids, are generally designed to make use of clean speech. However, in a real-world environment, 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 which require undamaged speech. The overall goal of speech enhancement is therefore to restore the desired speech signal from the noisy mix, by ideally eliminating this noise while fully retaining the quality and intelligibility of the original speech signal.

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 [1], 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, [2]) and Short-Time Objective Intelligibility (STOI, [3]) 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

In the interest of time, this project focuses on the implementation-side, assuming the binary mask is already provided provided; how the mask is generated is therefore 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 [4], which is included in the Imperial College London Software Library.

1.4 Report Overview

This report 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 [5]. Speech enhancement techniques, which vary in terms of the algorithms used, aim to improve the speech using audio signal processing techniques; some popular methods are described below.

2.1 Spectral Subtraction

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 [6].

A particularly common filtering technique is spectral subtraction, which operates in the frequency domain. In spectral subtraction, stationary or slowly-varying noise is attenuated from noisy speech by subtracting the magnitude noise spectrum, estimated during periods where speech is absent [7]. 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 [7], 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 [7]. This assumption means that the 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

This is true for the complex spectrum but it will not be true for the power spectrum (which has no phase information) 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 [9].

The local stationarity assumption means the processing must 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 [10], 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 frames, each of which have been multiplied by an input window. For the signal to remain undistorted, multiplying by the window should not change its magnitude. To achieve this, particular overlap factor/window pairs must be used; for example, if a Hamming window is chosen, the overlapped windows will approximately sum to unity for an overlap ratio of 2 i.e. each windowed frame overlaps each of its neighbours by 50%, ensuring the output signal remains undistorted.

you have to be a bit careful here. The window actually gets applied twice (during the analysis and then again during the synthesis). So the squares of the overlapping windows need to sum to unity. This is true for a Hamming with an overlap factor of 4 or for a square-root-Hamming with an overlap facor of 2

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 [11]. 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 [12]. However, it is generally evaluated that these modifications improve speech quality further but do not significantly affect the intelligibility of the input signals [11].

2.2 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) [13]. 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 human auditory system typically performs this signal separation process, known as auditory scene analysis, in two stages, to understand the message spoken by the target speaker. Firstly, the input sound is decomposed into a matrix of individual time-frequency (T-F) units, where each unit represents the signal occurring at a particular instance in time and 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 [14]. 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 [15]. CASA techniques aimed to pick out just the target signal from the noisy mix, and the computational method of choice was the ideal T-F binary mask [16].

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 0 dB 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 [17].

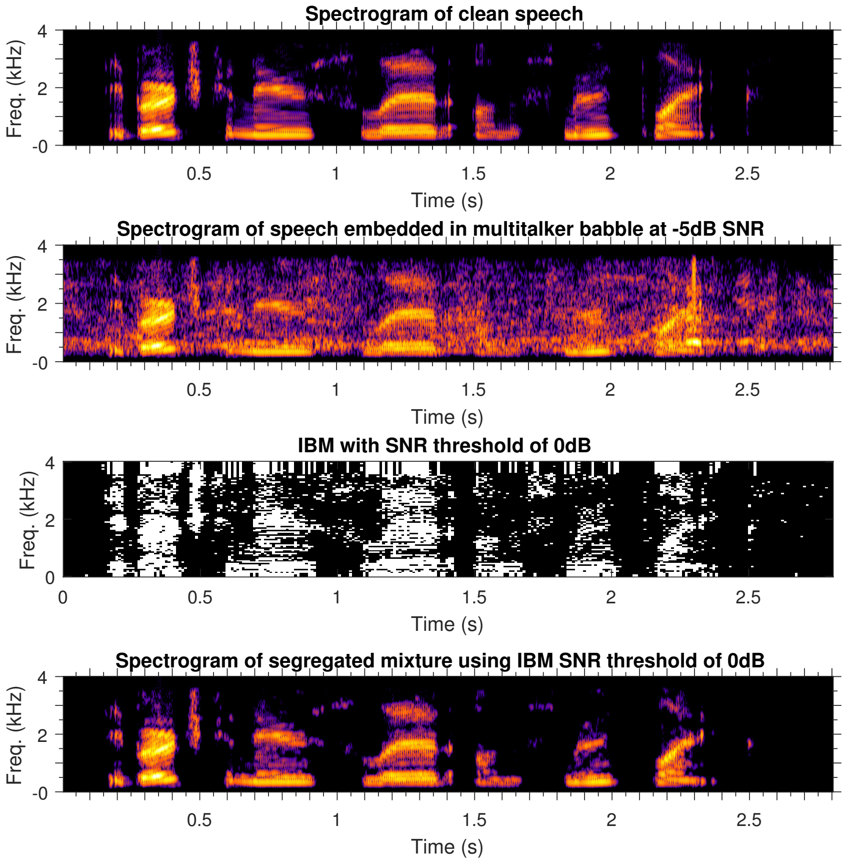
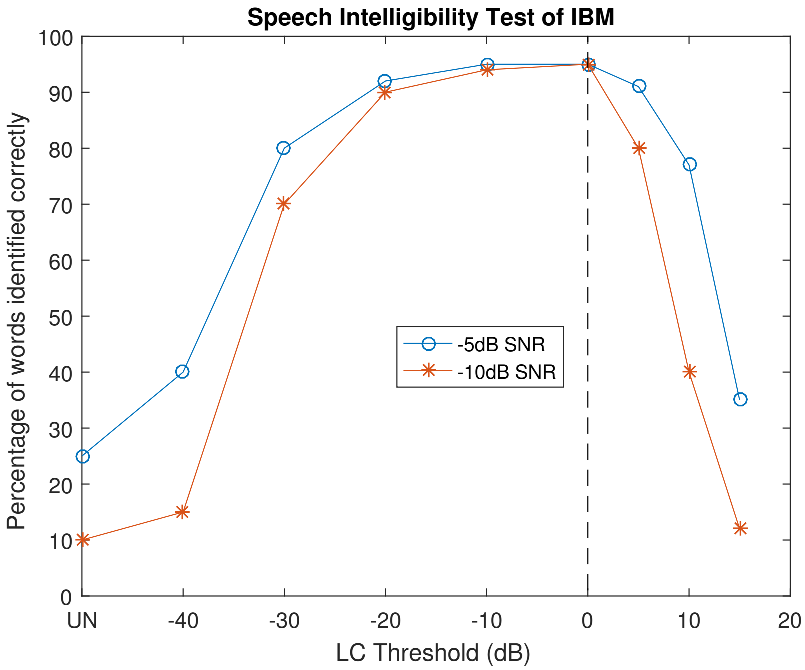


Figure 2.1: Top to bottom: clean speech, noisy speech, IBM and IBM-processed speech

An example of the IBM at work is shown in Figure 2.1, with a clean sentence obtained from the NOIZEUS corpus [18]. From top to bottom, the spectrograms shown are that of: a) clean speech; b) clean speech embedded in multitalker babble at -5 dB SNR; c) IBM constructed using LC threshold of 0 dB, where white pixels denote 1 (target stronger than interference masker) and black pixels denote 0 (target weaker than masker); d) segregated mixture obtained with the 0 dB LC IBM, obtained by multiplying the spectrograms in (b) and (c), one T-F unit at a time.

The 0 dB LC IBM, a particularly simple and intuitive comparison, is theoretically optimal in terms of SNR gain ([19], [20]); Figure 2.1 demonstrates this, 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 [21]. 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 [17]. 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 Figure 2.2, where UN represents the unprocessed noisy speech. In this example, the short-time Fourier Transform was used to process the input noisy signal, where multitalker babble was used as the masker [22]. As shown, the performance peaks out between approximately -20 dB and 5 dB for an input SNR of -5 dB, and the range is slightly smaller for an input SNR of -10 dB.



When you take a figure from somewhere else, you **MUST** credit its source in the caption, typically by adding something like this at the end "(taken from [22])".

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

Large intelligibility benefits were demonstrated in [22], 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 [17]. 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 [22].

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 [17], 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*, in a T-F auditory space, the target signal is [22]. 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. 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 [23].

An alternative ideal mask definition was proposed in [24], 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. The TBM is very similar to the IBM, but instead compares, in each T-F unit, the target energy to that of a speech-shaped noise (SSN) reference signal matching the long-term spectrum of the target speaker. 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. By definition, the TBM and IBM are identical when SSN is the masker used.

It is not quite true that TBM and IBM are identical with SSN for two reasons: (1) TBM depends on the average spectrum of that specific speaker whereas SSN is normally a more generic spectrum, (2) TBM compares with the average spectrum (i.e. a threshold that is time-invariant) whereas IBM compares with the actual noise in this time-frequency cell (which varies with time).

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 ([25], [26]).

Of course many other masks depend on oracle knowledge. The word "ideal" is, misleadingly, used only for one specific mask (which may not be the best possible mask).

2.2.1 Practical Considerations

The IBM is ”ideal” because it depends on oracle knowledge; the mask definition 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. In the presence of significant noise, this can be a difficult task, and it will be impossible to fully accurately compute the IBM for all T-F units. A noise-robust method based on target sound estimation to estimate the TBM was proposed in [26]. The estimation error is also an area of interest, and the effect of overall binary mask error was investigated further in [22]. 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 -5 dB input masked with multitalker babble to yield significant gains in intelligibility.

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

2.3 Kalman Filter

The Kalman filter [27] is a recursive optimal data processing algorithm. It is optimal with respect to any practical measure, under certain assumptions. The reason is that 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 [28].

The filter has a recursive ”predictor-corrector” structure [29]; 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 [30], where excellent noise reduction was achieved when linear prediction coefficients (LPCs) were estimated from clean speech. However, for practical use, these parameters are degraded as they have to be estimated from noisy speech, 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 [31] and a subband KF algorithm that achieves better performance and reduces computational complexity [28] 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 [32]. 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 [32], due to physiological limitations on how rapidly the vocal tract is able to change with time [33]. The slowly-varying modulation domain hence represents how the vocal tract changes over time [34].

There are STFT-based methods that estimate the phase as well as the magnitude of the STFT coefficient.

The KF is capable of handling non-stationary signals as well as estimating both magnitude and phase spectra [35], which puts it at an advantage over STFT-based methods for speech processing as phase information has been shown to be more important in the modulation domain than in the acoustic domain [36]. It was also noted in [34] 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 [34].

2.3.1 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 [34]. 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.

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, we obtain Equation 2.2. Assuming speech is quasi-stationary means that it can be analysed in frames using the STFT (analysis), thus obtaining Equation 2.3, 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.

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

$$Y(n,k) = X(n,k) + V(n,k) \tag{2.3}$$

Traditionally, AMS-based methods only modify the noisy acoustic magnitude spectrum $|Y(n,k)|$; the modified spectrum is thus obtained by combining the enhanced magnitude spectrum with the original noisy phase spectrum. 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 processed using a KF [34].

An additive noise model is assumed for each modulating signal, assuming white Gaussian noise, giving Equation 2.4.

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

In the KF autoregressive model, a p-order linear predictor is used to model the evolution of speech over time, as shown in Equation 2.5, where $a_{j,k}; j = 1, 2, ..., p$ are the LPCs and $W(n,k)$ is a random white excitation with a variance of $\sigma^2_{W(k)}$.

$$|X(n,k)| = -\sum_{j=1}^{p} a_{j,k}|X(n-j,k)| + W(n,k) \tag{2.5}$$

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

$$\mathbf{X}(n,k) = \mathbf{A}(k)\mathbf{X}(n-1,k) + \mathbf{d}W(n,k) \tag{2.6}$$

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

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, $\mathbf{d} = [1,0,...,0]^T$ and $\mathbf{c} = [1,0,...,0]^T$ are the measurement vectors for the excitation noise $W(n,k)$ and observation respectively, and $\mathbf{A}(k)$ is the state transition matrix utilising the LPCs:

$$\mathbf{A}(k) = \begin{bmatrix} -a_{1,k} & -a_{2,k} & \dots & -a_{p-1,k} & -a_{p,k} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{2.8}$$

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 $(|Y(1,k)|, |Y(2,k)|, ...|Y(n,k)|)$ using the following equations:

$$\mathbf{P}(n|n-1,k) = \mathbf{A}(k)\mathbf{P}(n-1|n-1,k)\mathbf{A}(k)^T + \sigma^2_{W(k)}\mathbf{d}\mathbf{d}^T \tag{2.9}$$

$$\hat{\mathbf{X}}(n|n-1,k) = \mathbf{A}(k)\hat{\mathbf{X}}(n-1|n-1,k) \tag{2.10}$$

$$\mathbf{K}(n,k) = \mathbf{P}(n|n-1,k)\mathbf{c}[\sigma^2_{V(k)} + \mathbf{c}^T\mathbf{P}(n|n-1,k)\mathbf{c}]^{-1} \tag{2.11}$$

$$\mathbf{P}(n|n,k) = [\mathbf{I} - \mathbf{K}(n,k)\mathbf{c}^T]\mathbf{P}(n|n-1,k) \tag{2.12}$$

$$\hat{\mathbf{X}}(n|n,k) = \hat{\mathbf{X}}(n|n-1,k) + \mathbf{K}(n,k)[|Y(n,k)| - \mathbf{c}^T\hat{\mathbf{X}}(n|n-1,k)] \tag{2.13}$$

where $\sigma^2_{V(k)}$ represents the variance of the corrupting noise. These equations can be categorised into two main steps: prediction and updating. Equations 2.9 and 2.10 predict the error covariance and state respectively based on past samples, while the other equations update the Kalman gain, error covariance and state based on the predicted values. In particular, Equation 2.13 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 $|\hat{X}(n|n,k)|$.

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.

Experimental results from the NOIZEUS corpus [18] demonstrate 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 modulation domain KF (MDKF) was by far the best performing algorithm, doing better than all acoustic and time-domain methods tested, including the time-domain KF (TDKF), for both white and coloured noise [34]. This was despite both algorithms having access to clean speech LPCs.

However, similarly to the IBM, clean speech is not available in reality; the presence of noise generally degrades the LPC estimates, thus worsening the performance of the MDKF algorithm. In [34], a practical MDKF algorithm was evaluated, which used an acoustic-domain pre-processor for LPC estimation to reduce the effect of noise degradation.

A nice description of modulation-domain Kalman filtering

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 [22].

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 [4] 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.

should use a square-root Hamming window if using 50% overlap

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 Figure 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 Figure 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.

How were you measuring intelligibility?

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 (Figure 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

I am more interested in using the TBM actually because it is independent of the noise.

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.

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

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 [22] and [24]. Once this has been determined, the binary mask is then available for use, and can be set aside for the time being.

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 [34], 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

Evaluation Plan

4.1 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.

4.2 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.

A detailed evaluation plan needs to decide how many speech samples you will use for evaluation, what types of noise you will add and the range of SNRs that you will evaluate performance over.

Bibliography

[1] 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).

[2] *ITU-T Recommendation 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.* URL: `http://www.itu.int/rec/T-REC-P.862/en`.

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

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

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

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

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

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

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

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

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

[12] 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).

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

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

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

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

[17] 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).

[18] Y. Hu and P. Loizou. "Subjective evaluation and comparison of speech enhancement algorithms". In: *Speech Communication* (2007).

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

[20] 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.

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

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

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

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

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

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

[27] 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).

[28] Wen-Rong Wu and Po-Cheng Chen. "Subband Kalman Filtering for Speech Enhancement". In: *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS* 45 (1998).

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

[30] 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).

[31] Jerry D. Gibson. "Filtering of Colored Noise for Speech Enhancement and Coding". In: *IEEE TRANSACTIONS ON SIGNAL PROCESSING* 39 (1991).

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

[33] 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).

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

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

[36] 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).