Slide 1

Good morning everyone. I am Jia Ying, and today I’ll be presenting about my final year project, speech enhancement using a Kalman filter.

Slide 2

This is an outline of what I will cover today. I will first give a brief introduction of my project.

Next, I will cover some important background, followed by the performance testing mechanisms.

After that, I will present and evaluate each of my modifications. These affect the Kalman filter equations, linear prediction coefficients, and noise estimation.

Finally, I will deliver my conclusion, and briefly outline possible areas for future work.

Slide 3

Speech enhancement algorithms aim to reduce the background noise of a noise-corrupted speech input without distorting the original clean speech. Recent evidence has pointed to the use of the modulation domain for speech enhancement, where the modulation domain is defined as the temporal variations of the acoustic spectral components.

In this project, we propose to improve the performance of an existing Kalman filter-based speech enhancer, by incorporating additional information obtained from an ideal binary mask. The performance of these proposed modifications is assessed by measuring the quality and intelligibility of the enhanced speech.

Slide 4

To start off, some background information is needed. An ideal binary mask (IBM) is defined as a time-frequency (T-F) matrix of binary numbers, constructed by comparing the local Signal-to-Noise Ratio (SNR), ~~defined as the difference between the target signal energy and the masker or noise energy in each T-F unit~~, against a threshold known as the local criterion (LC).

T-F units are assigned 1 where the local SNR exceeds the LC, and 0 otherwise.

~~For speech enhancement, this mask can then be applied to the T-F representation of a noisy speech input, multiplying the T-F matrices of the mask and input element-wise. There is evidence to show that the 0-dB LC IBM is overall optimal for real-world implementations, and is hence what will be used in this project.~~

Slide 5

Next, I will describe the modulation-domain Kalman filter (MDKF).

Firstly, an input noisy time-domain signal undergoes the short-time Fourier Transform (STFT) to form a series of time-varying frequency components known as modulating signals, each of which undergoes processing by a separate Kalman filter. This diagram shows only one frequency component, but the process is identical for all of them. The Kalman filter tries to recursively estimate the underlying clean speech amplitude spectrum given the noisy amplitude spectrum and past estimates, which depend on the estimated noise power , the estimated excitation variance and the linear prediction coefficients of the underlying clean speech stored in the matrix . The output of the filter is combined with the original noisy phase spectrum to undergo the inverse STFT to produce the output enhanced signal. This project uses the original MDKF as a baseline algorithm, and proposes a few modifications on top of it.

To investigate the theoretical upper-bound performance, clean speech is used to generate the linear prediction coefficients that model speech.

Slide 6

Speech enhancers are typically assessed in terms of speech quality and intelligibility.

Speech quality measures typically involve a group of listeners rating how “good” a speech signal is, scoring the signal from 1 to 5, where 5 is the best rating.

We approximate this with the Perceptual Evaluation of Speech Quality (PESQ) measure.

Speech intelligibility, on the other hand, is the accuracy with which we can identify what is being spoken, and refers to the proportion of correctly identified words relative to the total number of words.

A widely-used method to evaluate intelligibility is the Short-Time Objective Intelligibility (STOI) measure.

Slide 7

The first proposed modification is to the Kalman filter equations. The thought process is as follows: if the IBM can give some indication of speech presence or absence, this will be useful as a 3rd source of information, along with the prediction and measurement, to form the updated estimate of clean speech. In the Kalman filter, the observation is used to “correct” the prediction; the accuracy of this correction can be improved if more knowledge is provided.

Firstly, this IBM information is obtained as separate statistical distributions for mask 1s and mask 0s, and averaged over a set of training inputs.

The MDKF equations are then decoupled, so that when the IBM information is incorporated, the previous states are not permanently affected.

The measurement, prediction and mask information are then combined to construct the updated state.

Slide 8

The block diagram of this proposed modification is shown here, and is termed the (binary mask-modified MDKF or BMMDKF).

As you can see, the IBM is inserted during the updating step.

In this modified algorithm, the Kalman filter equations are decoupled, by applying a transformation to the state vector. The transformed state vector is such that its first element, the current state, is uncorrelated with the rest of the vector i.e. all other previous states. In the original MDKF, this is combined with the observation. In the BMMDKF, this is instead combined with both the observation and the relevant mask statistics, by multiplying together their probability distributions. ~~This is straightforward due to the assumption of Gaussian distributions, whereby the product of two Gaussian PDFs is another 2 Gaussian probability distributions.~~ The mask statistics used will depend on what the mask indicates at that particular time-frequency unit i.e. if the mask has value 1 at that unit, the speech-dominated statistics are used, and vice versa if the mask indicates 0.

Slide 9

This slide compares the performance results of the BMMDKF with the original MDKF, plotted over a range of input SNR and averaged over a test dataset. By both PESQ and STOI performance measures, we can see that the BMMDKF performs better than the MDKF. Particularly, the improvement increases as the input gets noisier. ~~At -20dB, the improvement is 17% for PESQ and 6% for STOI.~~

Slide 10

This figure here shows the differences between the BMMDKF and the MDKF. At a very noisy input of -20dB SNR, the BMMDKF is clearly superior in recovering the clean speech. Above the blue dotted line (about 1.9kHz), the spectral components are retained better in the BMMDKF. It also does not have the initial spike of the MDKF. Even though some of the high-amplitude areas are not as well retained, this only manifests as a barely-noticeable lower volume, and does not significantly affect the quality or intelligibility of the enhanced speech.

Slide 11

Overall, we can hence say that the BMMDKF performs better than the MDKF in terms of quality and intelligibility, **(CLICK)** and the improvement is greater for noisier input.

Slide 12

The next proposed enhancement is to the linear prediction coefficients (LPCs). Here, the IBM is applied to the test input directly.

Slide 13

The LPC estimation process involves minimising a sum of errors between the actual clean speech and predicted speech. In this modification, we propose to minimise a weighted sum of errors instead, with the weights determined by the applied IBM.

These weights are as shown, with the IBM 1s and 0s mapped to 2 different weights. The optimal threshold was found to be 0.15. Notice that this indicates that the error is given a greater weight when the mask shows noise i.e. mask value 0. A possible explanation is that speech generally has a much larger variation in amplitude as compared to random noise. It can then be argued that if the model needs to predict noise i.e. when speech is absent, it has to do so with greater accuracy than predicting speech presence. This lends itself to requiring a greater weight for speech absence.

Slide 14

This algorithm is termed the LMDKF, and the block diagram is as shown **(CLICK)** with the corresponding point where the IBM is inserted.

Slide 15

Results show that the LMDKF performs similarly to the MDKF overall, especially in terms of STOI. However, we do see a small improvement in PESQ across the board.

Slide 16

As a result, informal listening tests showed a slight preference for the LMDKF due to the increase in PESQ being more significant than the almost negligible variations in STOI.

A possible reason that the two algorithms perform so similarly is that clean speech was used to estimate the LPCs, and thus the original LPC estimation performs very well already. It is conceivable that the difference is more pronounced if non-clean speech LPCs were used.

Slide 17

The final modification uses an IBM applied to the input to tweak the noise estimation, which is used in the MDKF updating step.

Slide 18

The original noise estimator estimates the noise based on the noise periodogram ~~(a type of noise power spectrum)~~, which can be calculated as a function of the power of the noisy observation, the *a priori* SNR and the spectral noise power ~~of the previous frame~~.

These depend on the *a posteriori* speech presence probability, or the probability of speech presence given the noisy observation, ~~which depend on the predicted SNR~~. In the algorithm, certain assumptions were made that were adjusted based on IBM information.

Particularly, the best improvements were found when the *a priori* SNR and *a posteriori* speech presence probability (SPP) were tweaked with the IBM.

In the modified noise estimation, the *a posteriori* SPP was multiplied with a weighting function determined from the IBM as shown. In addition, the *a priori* SNR was scaled with a number dependent on the weighting frame. The optimal weights were found to be 1 (for mask 0s) and 11 (for mask 1s).

Slide 19

We call this the NMDKF, with its block diagram and **(CLICK)** IBM intervention shown here.

Slide 20

Overall, the NMDKF performs similarly to the MDKF. Particularly, it posts slightly higher PESQ scores at noisier input, while peaking off at higher input SNR.

Slide 21

The adjustments made to the noise estimation were on a frame-wise basis, independent of the rest of the signal. These local adjustments could have improved a very noisy input, but when the input was at high SNR then these could have negatively affected the overall enhanced signal and degraded the perceived quality.

However, the MDKF already posts very high scores for high input SNR, so even though the NMDKF performs worse at high input SNR, it still performs very well.

Slide 22

In closing, my project has involved modifying an existing MDKF using IBM information.

Three modifications were proposed; the first involves incorporating IBM information directly into the Kalman filter equations, termed the BMMDKF.

The second uses a weighted sum to estimate linear prediction coefficients, termed the LMDKF, **(CLICK)** and the final adjustment is to the noise estimation process, termed the NMDKF. In particular, the BMMDKF showed the greatest improvement over the original algorithm.

Slide 23

Finally, there are some possible areas for future work. This project is concerned with investigating the theoretical best performance, and therefore uses clean speech to estimate LPCs, which is not possible in reality. For practical use, an intermediate speech enhancer could be used to enhance the noisy input first, and the LPCs would be estimated from this enhanced input.

In addition, for simplicity of computation, the BMMDKF assumes that the measurement, prediction and mask values are all Gaussian distributed. It is possible that a different distribution represents these variables better, but also possible that this increases computational complexity.

Finally, PESQ and STOI were used to quantify the algorithms, but there is still merit to carrying out real human listening tests if possible.

Slide 24

And that is all! Thank you.