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, including its motivation and scope.

Next, I will cover some background that is important to the project, as well as the testing mechanisms.

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

Finally, I will deliver my conclusion, and briefly outline some 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. In real-world applications, this can be very challenging. Although many algorithms have been developed to improve the Signal-to-Noise Ratio of a noisy input, they also introduce speech distortion and artefacts such as musical noise, damaging speech quality and intelligibility. 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 speech enhancer based on a Kalman filter, by incorporating statistical information obtained by applying an Ideal Binary Mask (IBM) on noisy speech samples. The performance of these proposed modifications is assessed by measuring the quality and intelligibility of the enhanced speech.

In this project, a binary mask is assumed provided, and thus the focus is on implementing a baseline Kalman filter and its modifications.

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. An intuitive example is the 0 dB LC mask, where a T-F unit is assigned 1 if the local signal energy is greater than the local noise energy, and vice versa for 0.

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. It thus acts like a selective filter, allowing only some parts of the signal to pass through. There is evidence to show that the 0-dB LC IBM is overall optimal for real-world implementations, and hence this is what will be used in this project.

Slide 5

Next, I will describe the Kalman filter, which is a recursive algorithm. For speech enhancement, the modulation-domain Kalman filter (MDKF) is becoming increasingly widespread.

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, whereby the process is identical for all frequency components and can hence be done in parallel. The STFT produces a phase spectrum and amplitude spectrum . The Kalman filter tries to recursive estimate the underlying clean speech amplitude spectrum given the noisy amplitude spectrum and past estimates, which depend on the estimated noise power and the estimated excitation variance and 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.

Slide 6

Finally, I will briefly describe the performance metrics. Speech enhancement algorithms can be quantified in numerous ways, but typically are assessed in terms of speech quality and intelligibility.

Speech quality can be assessed in an objective manner, such as using Signal-to-Noise Ratio (SNR), which calculates the ratio between signal and noise power in a noisy signal. Subjective quality measures typically entail a group of listeners subjectively rating how “good” a speech signal is, scoring the signal from 1 to 5, where 5 is the best rating. This process is costly and time-consuming, and can be approximated by the Perceptual Evaluation of Speech Quality (PESQ), which was developed to model subjective quality tests with high correlation.

Speech intelligibility, on the other hand, is a different measure. It is the accuracy with which we can identify what is being spoken, and specifically 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, which has been found to be highly correlated with actual intelligibility scores.

Slide 7

This project proposes a few modifications to the modulation-domain Kalman filter. The baseline framework is hence the original MDKF, shown here again for reference. Particularly, to investigate the theoretical upper-bound performance of our modifications, clean speech is used to generate the linear prediction coefficients that model speech. In a real-world scenario, this is of course unavailable, and the LPCs must be estimated in a different manner.

Slide 8

The performance of the modified algorithms will be compared with the original MDKF, according to the speech quality and intelligibility metrics described earlier. One difference is that instead of using SNR for objective quality measurement, we instead use segmental SNR (segSNR), which calculates SNR in short frames and takes the average, to prevent silent frames or frames with excessively large or small speech magnitudes to dominate the SNR ratio, which better represents true speech quality.

Slide 9

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 obtain the updated estimate of clean speech. Firstly, this 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 such that the current state and all previous states are uncorrelated, so that when the IBM information is incorporated, the previous states are not permanently affected.

Slide 10

The block diagram of this proposed modification, termed the (binary mask-modified MDKF or BMMDKF), is shown here, along with the point where the IBM is included.

Slide 11

The next 2 slides show the performance results of the clean-LPCs BMMDKF compared against the original clean-LPCs MDKF, an MMSE speech enhancement algorithm, and the input noisy speech, plotted over a range of input SNR and averaged over a test dataset. This graph clearly shows that the BMMDKF and MDKF are similar in terms of segSNR, with the BMMDKF very slightly better at -5 and -10dB input.