MP3: Cepstrum and Mel-Frequency Cepstrum

Group: Pansi Heng, Yangge Li

1. Introduction

In this mp, we are using raw sound data, frequency cepstrum and MFCC as features to solve the problem. In our case, frequency cepstrum and MFCC is more useful than raw sound data because frequency cepstrum and MFCC can give us better results than raw data and can actually achieve that by using smaller amount of data.

In this mp, we are using the knn classifier to classify the sound data. The algorithm requires calculating the euclidean distance between the test data and each training data, and find the training data that have lowest distance to the the test data and classify the test data as the same class as the training data.

2. Methods

For raw data, we just reshape each sound data to a column and return the vector as the feature of each .wav file. This part is implemented at line 26 in run.m file

The part to calculate the frequency cepstrum is implemented in the cepstrum.m file. After pre-emphasizing raw data, the cepstrum was calculated by first set the signal to overlapping frames and windowing each frame. Then, do inverse fft to the log of fft of each frame as the following.

$$c[n] = F^{-1}(log(|F(x[n])|))$$

The last step is to choose the first Ncc = 12 coefficients and unroll the output to a single column vector.

The part ot calculate the MFCC is implement in the mfcc.c file. The first step is to pre-emphasize signal and convert resulting signal to overlapping frames. Then we need to compute the Mel Filterbank Weights, H using the melfilterbank(M,K,R,Fs) provided in the melfilterbank.m file. Then we need to compute the Average Magitude Spectrum for each Frame and save the result in matrix X. The next step is to apply H on X to get the Mel magnitudes Y. Then we need to do an inverse discrete cosine transform to the log of Y as

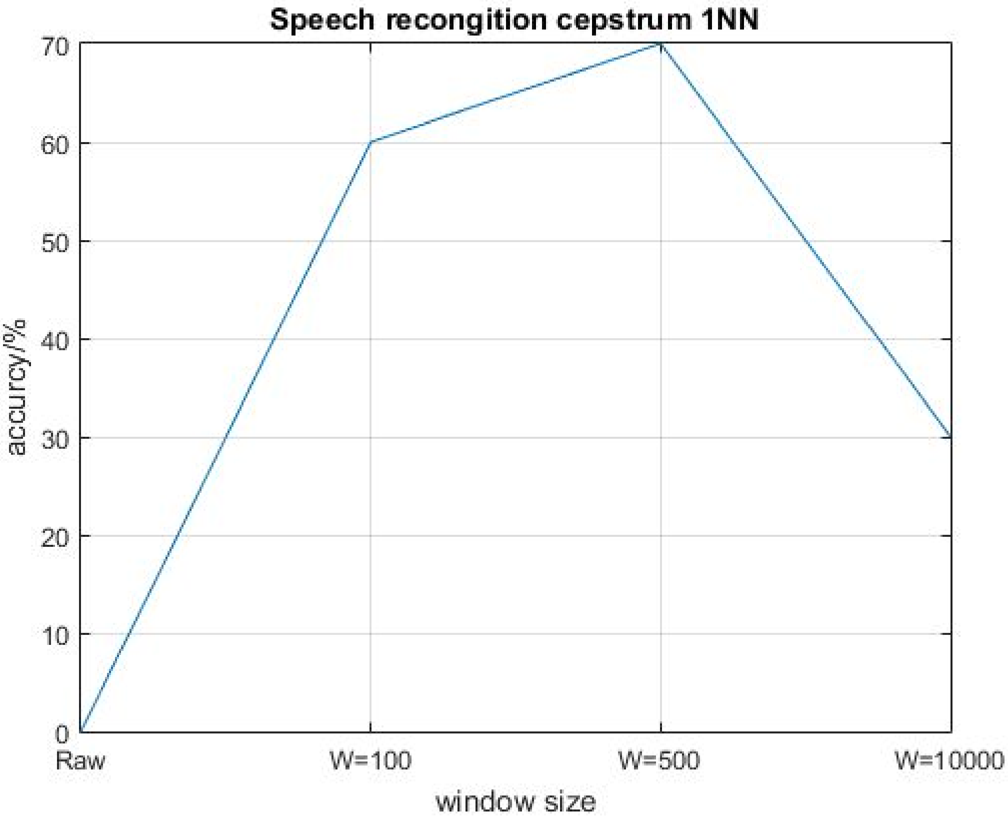
$$c[n] = dct^{-1}(log(Y[n]))$$

The final step is to choose the first Ncc = 12 coefficients and unroll the output to a single column vector.

3. Results

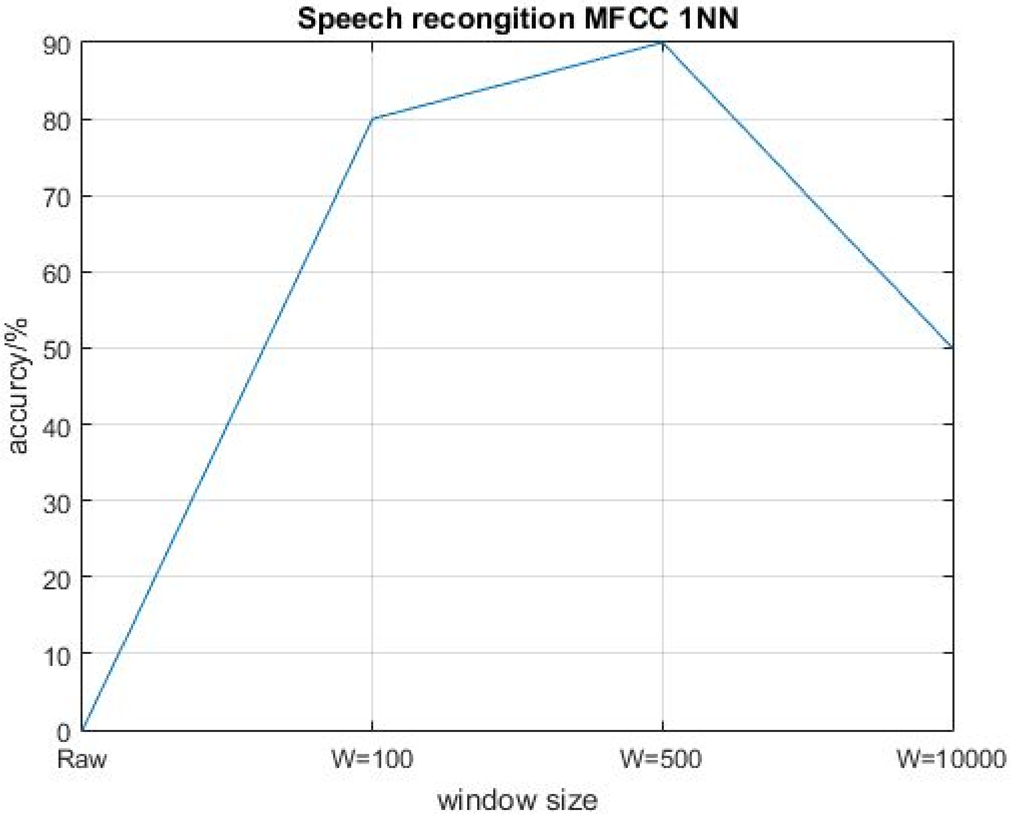
Cepstrum
Speech Recognition
1NN

	Raw	W=100	W=500	W=10000
D1	0	70.0000	65.0000	20.0000
D2	30.0000	45.0000	50.0000	70.0000
D3	15.0000	70.0000	65.0000	70.0000
D4	50.0000	70.0000	65.0000	65.0000
D5	0	60.0000	70.0000	30.0000
AVG	19.0000	63.0000	63.0000	51.0000



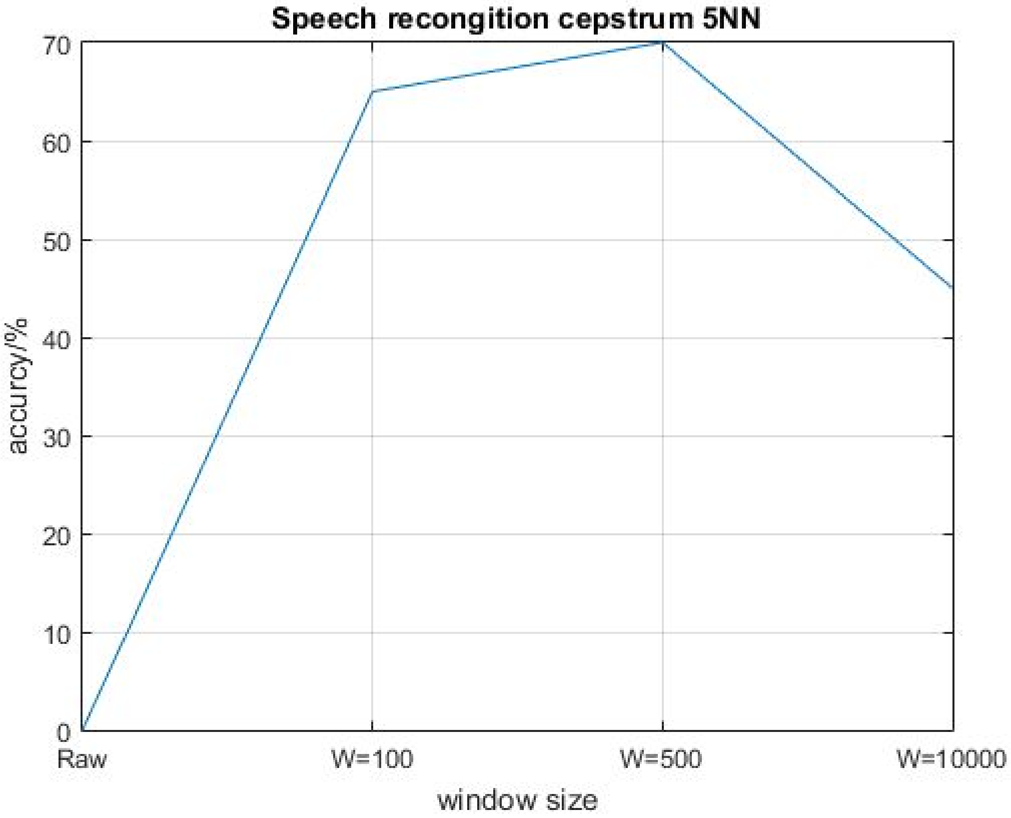
MFCC
Speech Recognition
1NN

	Raw	W =100	W=500	W=10000
D1	0	55.0000	60.0000	35.0000
D2	30.0000	65.0000	90.0000	90.0000
D3	15.0000	95.0000	100.0000	90.0000
D4	50.0000	75.0000	75.0000	60.0000
D5	0	80.0000	90.0000	50.0000
AVG	19.0000	74.0000	83.0000	65.0000



Cepstrum
Speech Recognition
5NN

	Raw	W=100	W=500	W=10000
D1	0	85.0000	65.0000	25.0000
D2	45.0000	45.0000	55.0000	55.0000
D3	5.0000	60.0000	65.0000	60.0000
D4	30.0000	55.0000	55.0000	55.0000
D5	0	65.0000	70.0000	45.0000
AVG	16.0000	62.0000	62.0000	48.0000

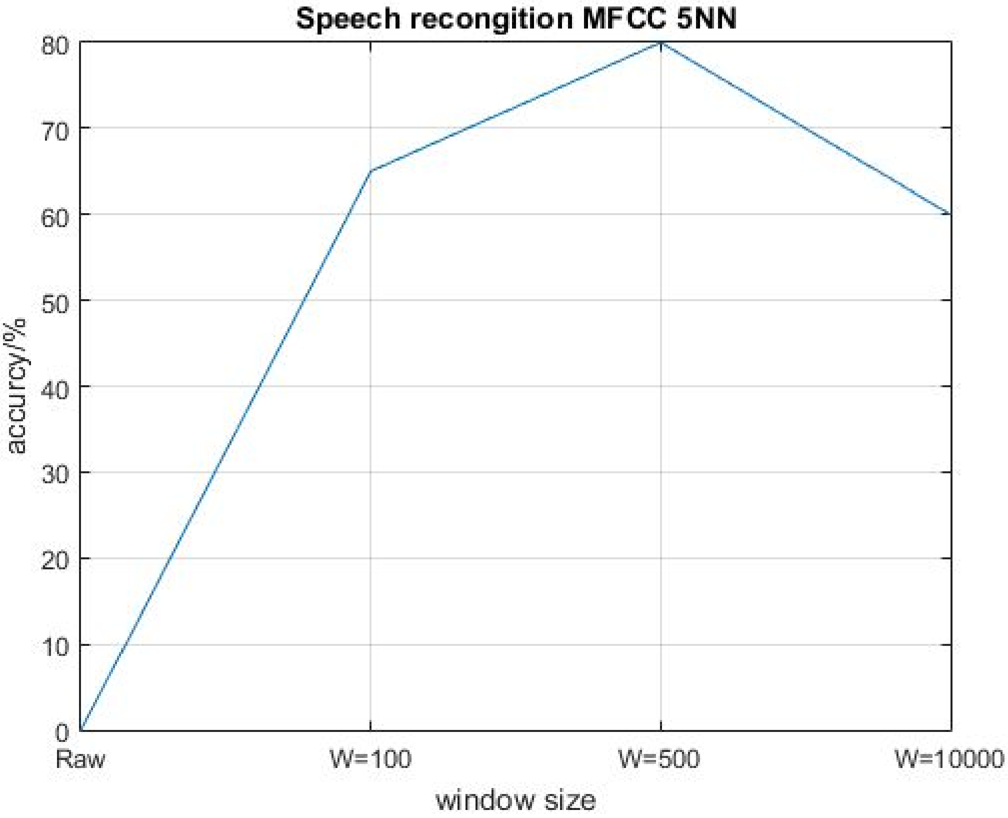


MFCC

Speech Recognition

5NN

	Raw	W=100	W=500	W=10000
D1	0	85.0000	70.0000	30.0000
D2	45.0000	65.0000	90.0000	80.0000
D3	5.0000	90.0000	95.0000	85.0000
D4	30.0000	50.0000	65.0000	60.0000
D5	0	65.0000	80.0000	60.0000
AVG	16.0000	71.0000	80.0000	63.0000

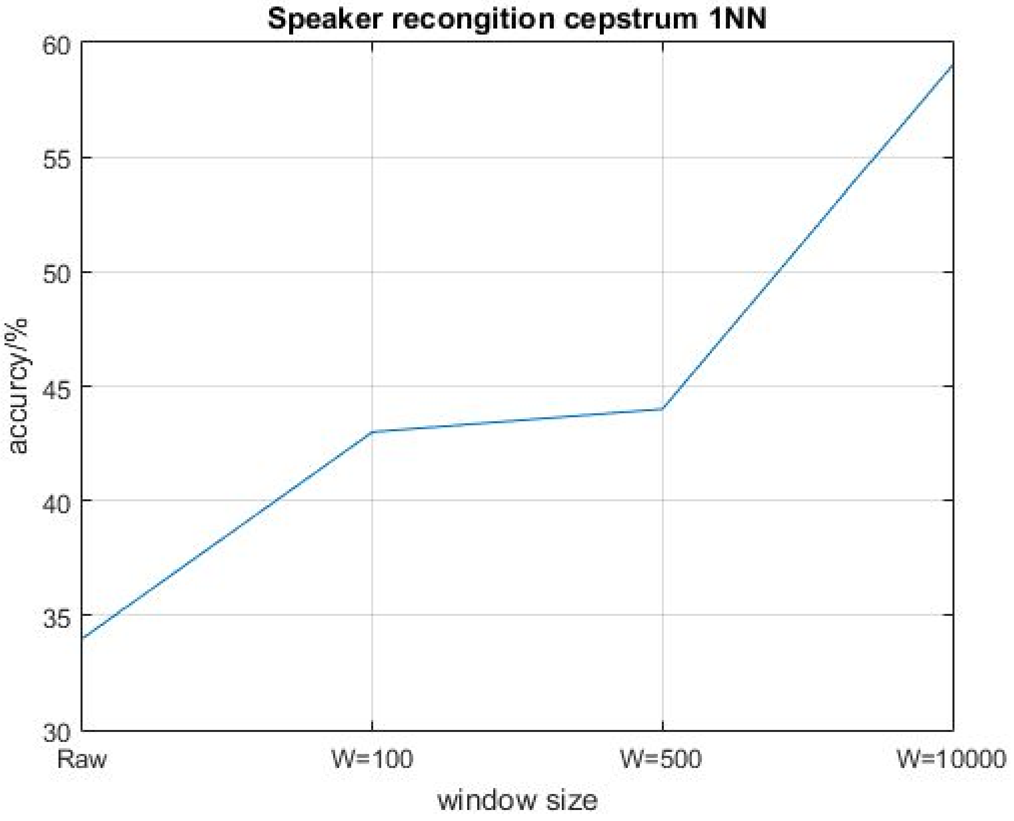


Cepstrum

Speaker Recognition

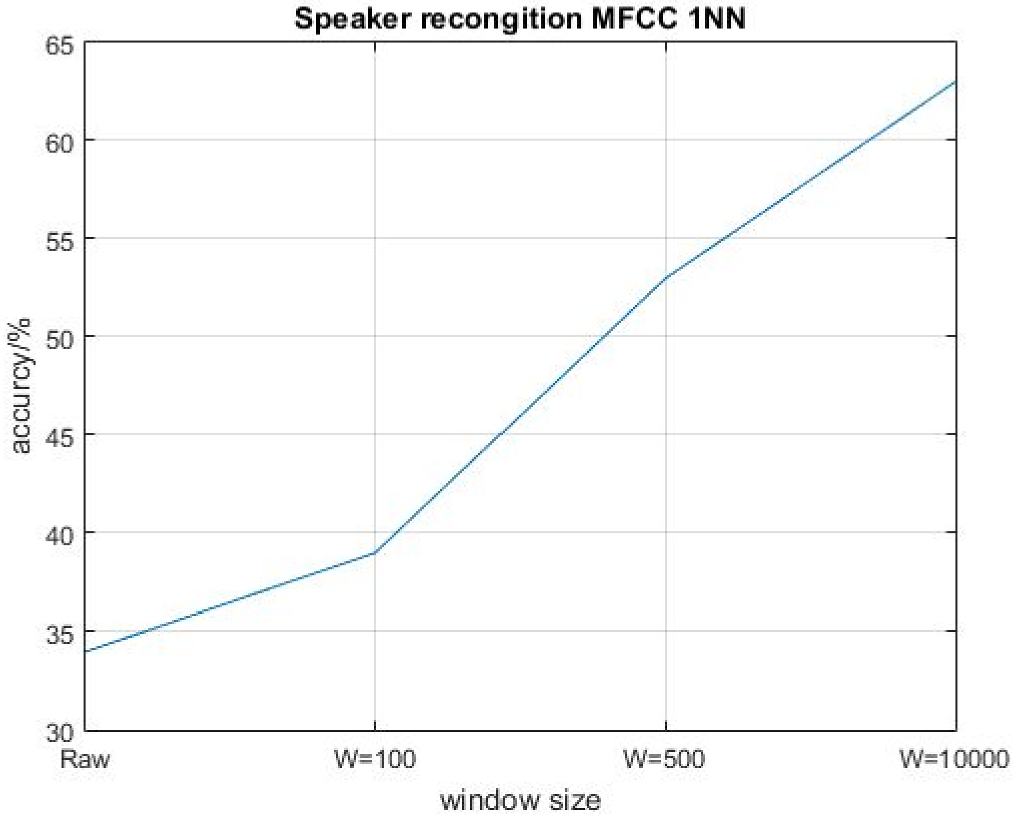
1NN

	Raw	W=100	W=500	W=10000
P1	8.0000	52.0000	60.0000	56.0000
P2	76.0000	60.0000	52.0000	88.0000
P3	36.0000	20.0000	32.0000	52.0000
P4	16.0000	40.0000	32.0000	40.0000
AVG	34.0000	43.0000	44.0000	59.0000



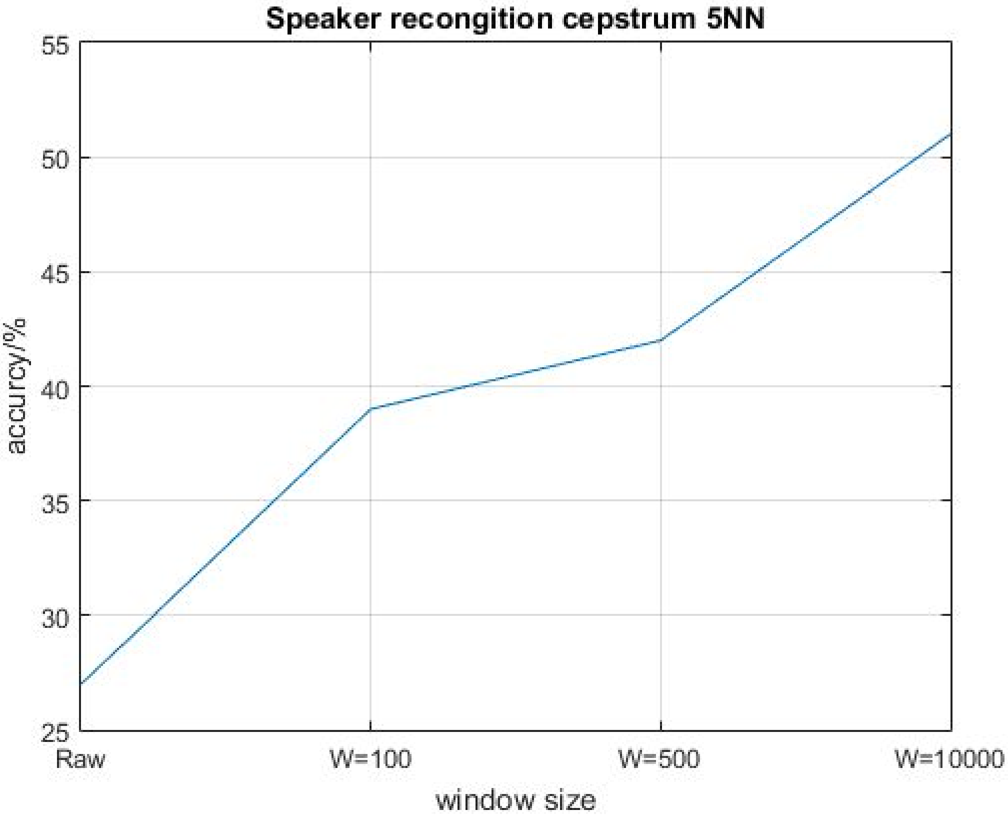
MFCC
Speaker Recognition
1NN

	Raw	W=100	W=500	W=10000
P1	8.0000	40.0000	60.0000	56.0000
P2	76.0000	60.0000	84.0000	76.0000
P3	36.0000	40.0000	36.0000	64.0000
P4	16.0000	16.0000	32.0000	56.0000
AVG	34.0000	39.0000	53.0000	63.0000



Cepstrum
Speaker Recognition
5NN

	Raw	W=100	W=500	W=10000
P1	8.0000	52.0000	68.0000	52.0000
P2	92.0000	56.0000	72.0000	76.0000
P3	8.0000	16.0000	8.0000	44.0000
P4	0	32.0000	20.0000	32.0000
AVG	27.0000	39.0000	42.0000	51.0000

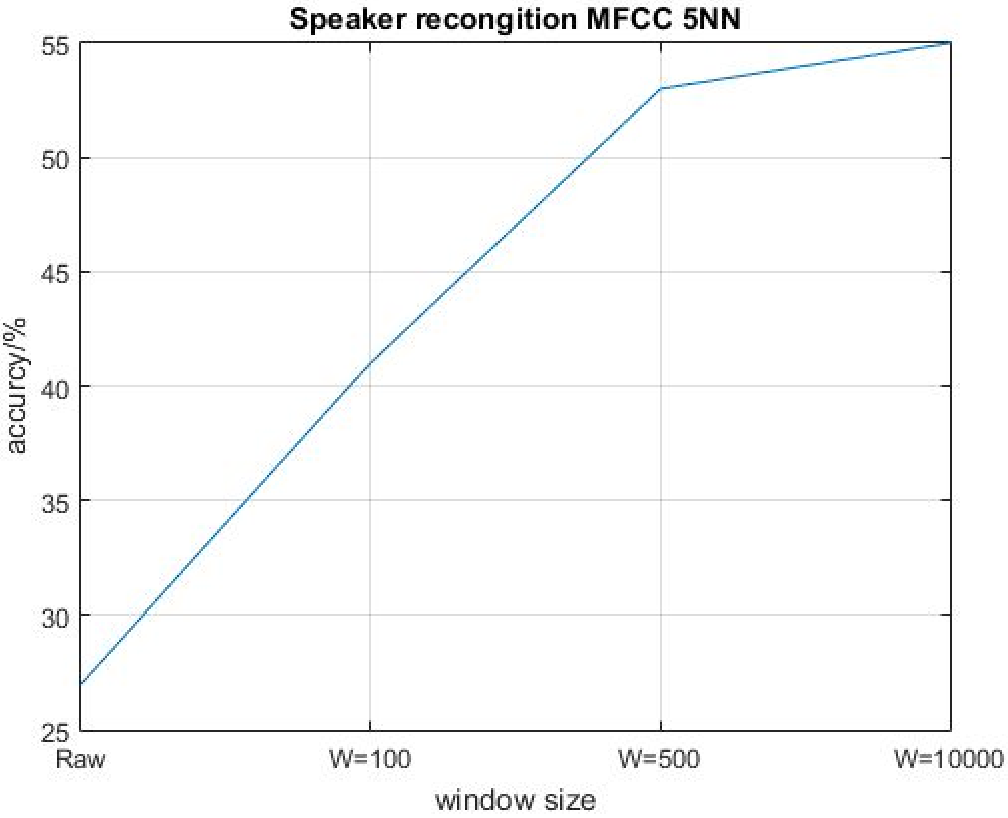


MFCC

Speaker Recognition

5NN

	Raw	W=100	W=500	W=10000
P1	8.0000	60.0000	68.0000	60.0000
P2	92.0000	56.0000	80.0000	80.0000
P3	8.0000	32.0000	44.0000	40.0000
P4	0	16.0000	20.0000	40.0000
AVG	27.0000	41.0000	53.0000	55.0000



4. Discussion

An interesting result from the experiment data above is that, contrasting to what people think that raw feature can always provide best classify result, in this case the raw data actually give us the worst classify result. In general, on one hand, the MFCC can give us better result than cepstrum and the 5nn classifier always offers us worse result than 1nn classifier in any features extraction methods in either speaker recognition and speech recognition. On the other hand, in speech recognition, window size of 500 samples will create the best estimating result among three but in speaker recognition, it is the window size of 10000 samples that did it.