**Report Format Example**

**108061217鍾永桓**

**Design Specification**

Parameter(original value):

frame\_length = 512 # Frame length(samples)

frame\_step = 256 # Step length(samples)s

emphasis\_coeff = 0.95 # pre-emphasis para

num\_bands = 12

num\_bands2 = 64 # Filter number = band number

num\_FFT = frame\_length # FFT freq-quantization

freq\_min = 0

freq\_max = int(0.5 \* sr) # Nyquist

signal\_length = len(source\_signal) # Signal length

**Inverse discrete cosine transform:**

Input: MFCC (12, 552)

Output: inv\_DCT (552, 12)

**Magnitude restoration:**

Input: inv\_DCT (12, 552)

Output: inv\_features (12, 552)

**Invert the Mel-filter convolution:**

Input: inv\_features (12, 552)

Output: inv\_spectrogram(257,552)

**Restore time-domain signal with Fast-Griffin-Lim algorithm**

Input: inv\_spectrogram(257,552)

Output: inv\_audio(141056, )

**All System**

Input: MFCC (12, 552)

Output: inv\_audio(141056, )

**Design Implementation**

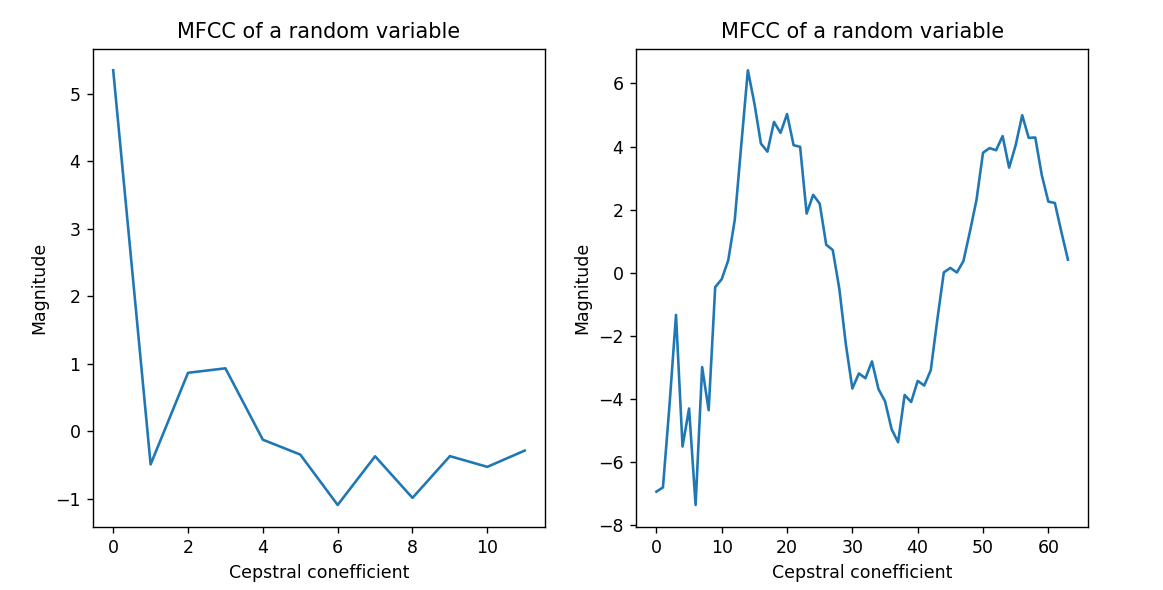
1. Use idct(MFCC.T,n=num\_bands, norm = 'ortho') to perform inverse DCT on MFCC.

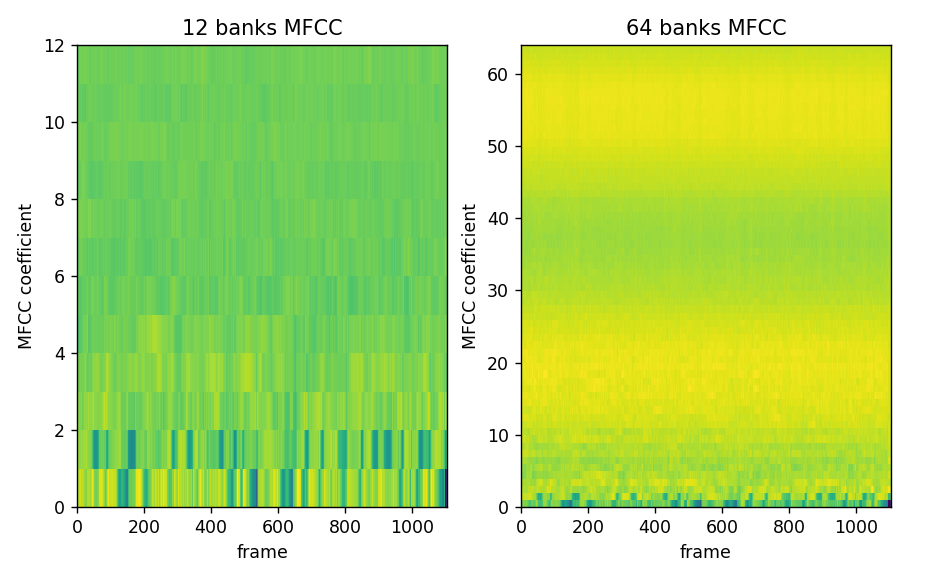
2. We can use np.exp(inv\_DCT) to restore magnitude of inv\_DCT(in logarithm scale).

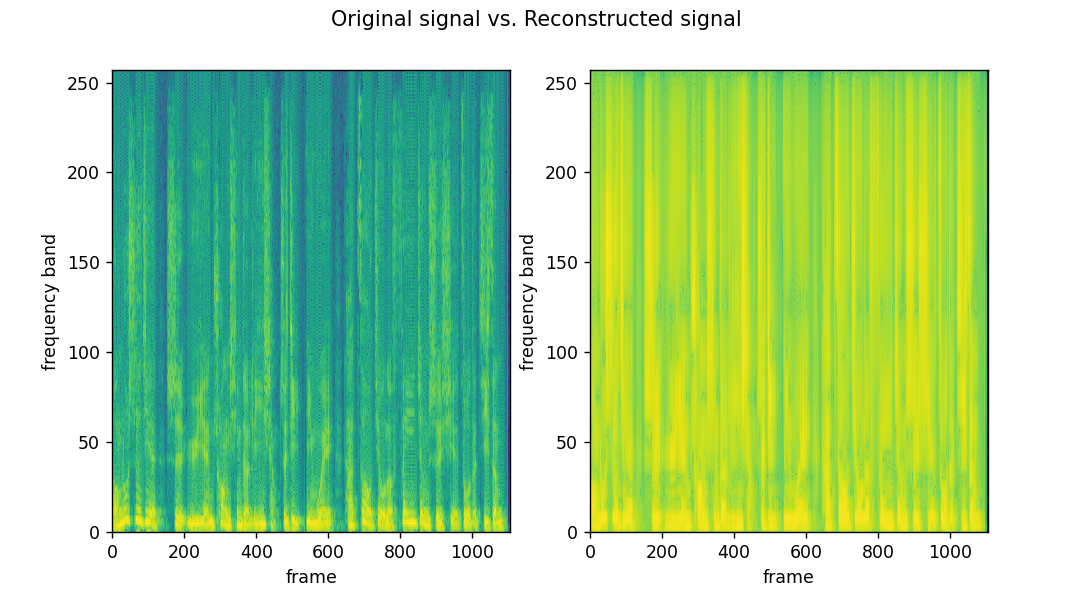
3.Do inverse convolution for inv\_features and we can get the spectrogram. We use np.dot( np.dot(fbanks1.T,pinv(np.dot(fbanks1,fbanks1.T))),inv\_features.T) to perform inverse convolution.

4. Use griffinlim() function to restore time domain signal from the inv\_spectrogram.

**Discussion**







五個問題

Why not using rectangular filters for “energy” calculation?

First, we use the triangular filter instead of rectangular becuase we want to mimic the human ear perception. With triangular filter, the features extracted may be more similar to features received by human. Besides, triangular filter can make our output smoother. The sudden change of rectangular may make our result not smooth.

Why should mel-filters be overlapping?

It can mimics human ear perception and make our information consistent. As we hear the sound, there may be more than one part of our ears can perceive it.

Why are high-quefrency MFCCs usually abandoned when doing speech recognition?

Human is not sensitive to high frequency signal and preserving high frequency needs more space. That is, preserveing high frequency information is not very helpful for speech recognition ,so high-quefrency MFCCs can be abandoned without seriously affecting performance.

Do you think MFCC is good for speaker identification purpose?

Yes, it use Mel-filter which can mimic the non-linear human ear perception and extract some important information for speech. Therefore, we can identify the speaker through the information. MFCC can also be used for baby cry detection. Although the frequency of baby cry is higher, MFCC can extract features to detect it.

If two sounds have similar MFCCs, does that imply they sound similar to our ears?

If these sounds are speech, it may be true. However, if these sounds do not belong to speech, MFCCs may not preserver information in high frequency well. It is possible that these sound are not similar ,but MFCCs cannot preserve the different features.

Question 1: What are the artifacts and distortions in the reconstructed audio? Suggest what the causes of these degradations are. (i.e. which sections of the MFCC extraction process are not invertible?)

The reconstructed signal is a little obscure and lose some detail. Although we still can understand what the people say in the audio, it is unclear. First, as our signal pass the filter banks, it can not preserve information perfectly. Besides, during extraction, we have ever taken absolute value of our signal and it will also cause the loss of information.

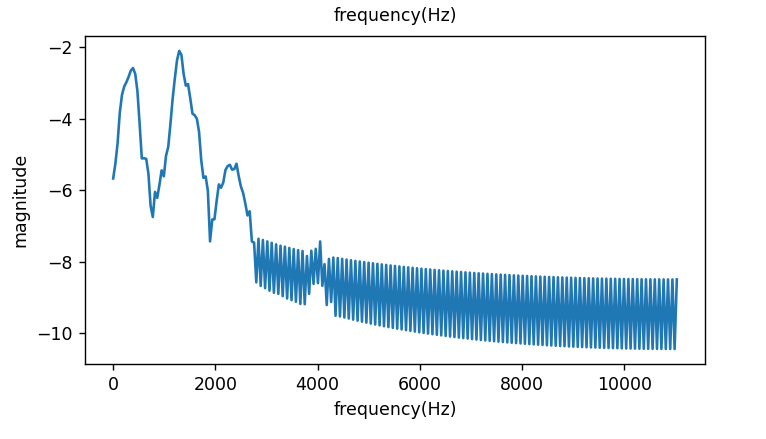
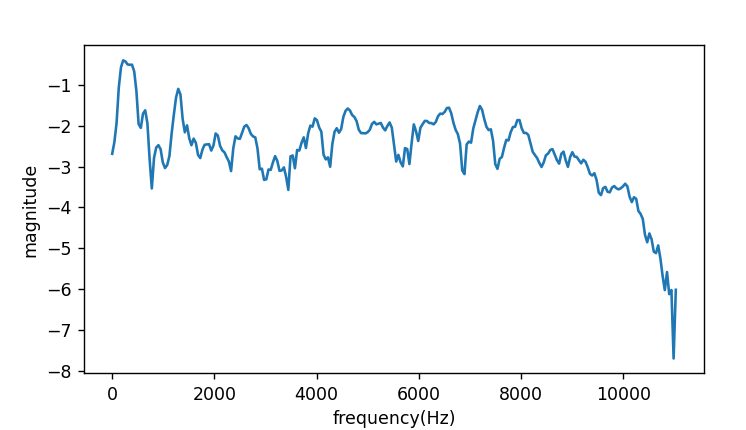
Question 2: Experiment with different frame length, step length, and number of fbanks; discuss what effects each of them has in the reconstruction process.

While the frame length is too large, the sound may be unstable. However, if it is very small, the frame number will be very large and the during of each frames will be very short. It make the frames not easy to process.

The step length will affect the overlapping between two adjacent frames , so it can not be larger than frame length. If step length is small, the number of frames will get very large. It will cost more time to extract the features. Besides, if step length is too large, the information in boundaries may be degraded. Because our frames will pass Hamming windows during processing and this kind of window cannot preserve the information in boundaries, some information will lose

As the number of fbanks increase, we can get more information from the signal. However, to construct a filter banks with larger number of fbanks, we need higher resolution. That is, the num\_FFT should be also larger. It will cost us more time to process the signal. Besides, once the number of fbanks is large enough, the larger fbanks cannot change performance obviously.

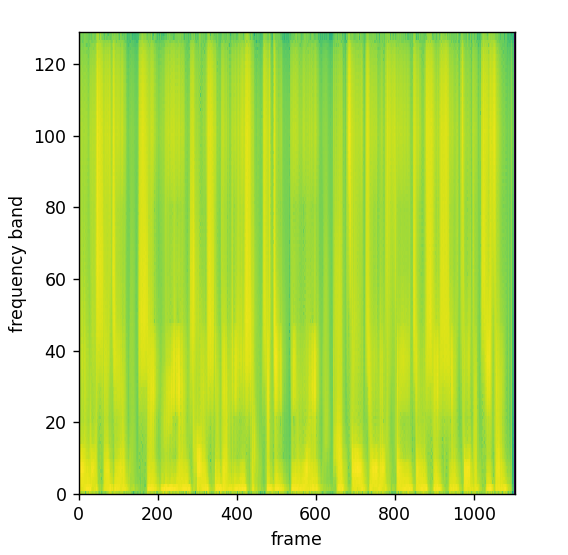
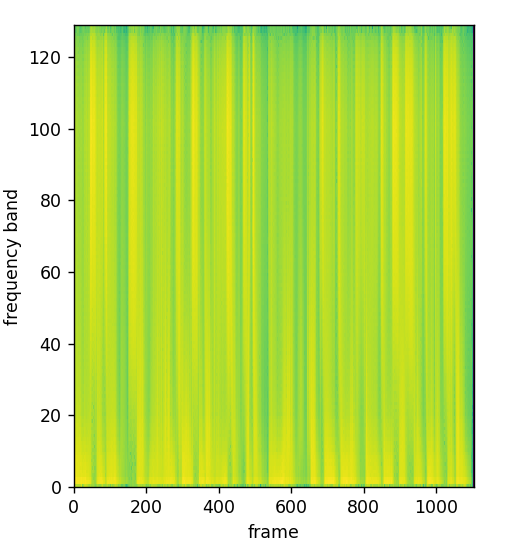
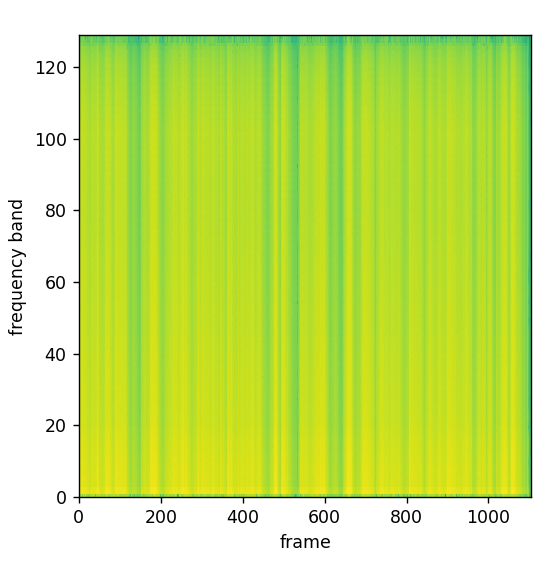
Bonus 1: Aside from setting optimal parameters, what can be added in the reconstruction algorithm to improve the end quality? Implement your proposal and present some experiments of it.



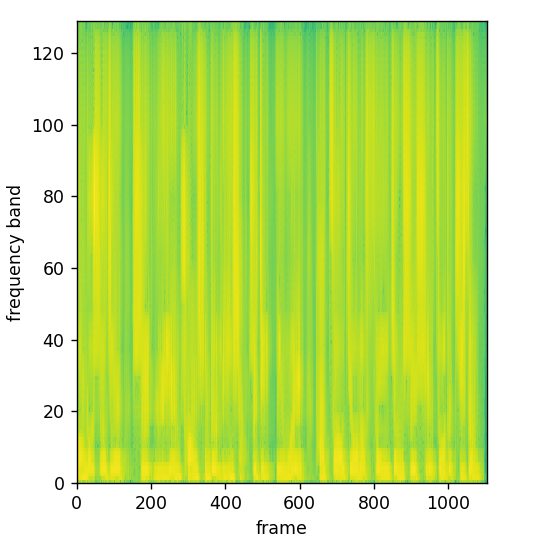
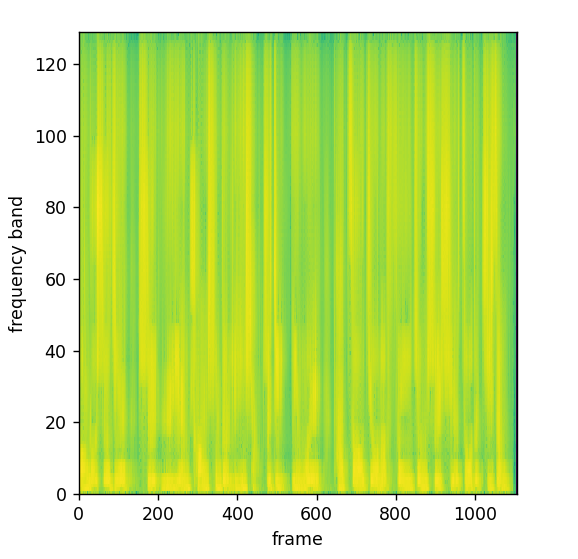
1. (b)
2. the reconstructed audio without noise reduction
3. the reconstruction audio with noise reduction

The reconstructed audio contain much noise in high frequency. If we want to make the speech more clear, performing noise reduction can be a choice. The picture(b) is the spectrum of random frame after noise reduction. It has smaller magnitude in high frequency. Most of time, frequency of human speech is below 1100 Hz ,so reduce noise in high frequency can make speech more clear.

Bonus 2: We did not perform dimension reduction/reconstruction in the DCT/inv-DCT sections. Modify those parts such that we have a complete algorithm that performs compression/decompression. Discuss how this influences the reconstruction quality



(a) (b) (c)



(d) (e)

(a)reduction from 12 to 1 (b)reduction from 12 to 3(c)reduction from 12 to 6 (d)reduction from 12 to 9 (e)no reduction

When we only preserve one dimension, we can hear something but it is extremely obscure. If we preserver more than three dimension, it is still unclear, but we can understand what the people say. If the dimension more than six, it is almost clear, but some detail is still lost. If the dimension is more than nine, it is almost the same as the no reduction one and only lose some unimportant information in high frequency. Preserving more dimension means that we preserve more information in high frequency. However, to recognize human speech, we do not need much high dimension information. Therefore, even if we lose some high frequency information, it may not affect the quality seriously

**Conclusion**

In lab2, I learn how to reconstruct the audio from MFCC. We can know what the people say in the reconstructed audio. Although the reconstructed audio may still lose some information, it seems that it successfully preserves some important features that can represent human speech. The quality of reconstructed audio is affected by parameters such as frame length, frame step and fbanks, so we should set these parameters properly.

**References**

<https://paperswithcode.com/method/griffin-lim-algorithm>

In this reference, I learn some detail about Griffin-Lim Algorithm. With the basic understanding about this algorithm, I can successfully implement it with my python code.