Cepstral Analysis of Audio Signals

Yinan Zhou
yinan.zhou@mail.mcgill.ca

Abstract. The output signal of the system is due to the input excitation and the response of the system. From the perspective of signal processing, the output of a system can be regarded as the convolution of the input signal and the system impulse response. The objective of the cepstral analysis is to separate the input excitation and the system components. This project introduces the basic concepts of cepstrum and cepstrogram. Their implementations in pitch estimation, spectral envelope estimation, and audio feature extraction are also covered.

1 Introduction

In 1959, Bogert found that the ripples in the spectra were the characteristics of signals with echoes while studying seismic signals [1]. Tukey suggested that the spectral analysis of the logarithm spectrum can determine the frequency of these ripples. The term "cepstrum" was then created to describe the spectrum of the logarithm spectrum by Bogert et al [2]. Accordingly, the term "quefrency" refers to the frequency of the spectral ripples. Besides, "filtering" and "harmonic" respectively correspond to "liftering" and "rahmonic" in the quefrency domain. The essential idea of cepstral analysis is to convert multiplication into addition so that the original signal and its echo can be easily separated.

In 1962, Schroeder suggested to Noll that cepstral analysis might be suitable for speech signals because speech spectra also have ripples. Noll then applied short-time cepstrum to determine the pitch in speech signals [3, 4].

Unrelated to the work above, Oppenheim developed the theory of homomorphic systems in his dissertation in 1965 [5]. The main idea of the homomorphic system is to map nonadditive combination into addition, then addition into addition, and inversely addition into nonadditive combination.

When the nonadditive combination is convolution, homomorphic deconvolution resembles the main idea of cepstrum.

However, homomorphic deconvolution needs to perform complex Fourier transform and complex logarithm because the phase information is required. Therefore, the nonlinear mapping in homomorphic deconvolution is also known as complex cepstrum. By contrast, the cepstral analysis proposed by Bogert et al. loses phase information. Thus, the term ”real cepstrum” generally refers to the cepstrum defined by Bogert et al.

The cepstral analysis plays an important role in audio processing. For instance, power cepstrum is widely utilized in the pitch detector. Homomorphic deconvolution was also applied to separate vocal source and vocal tract in speech modelling. In addition, the Mel-Frequency Cepstral Coefficients (MFCC) are regarded as a robust and reliable feature in speech recognition, speaker identification, Music Information Retrieval and other audio processing related problems.

This project first introduces the definition of cepstrum and its use in pitch estimation. Then, the concept of short-time cepstrum is presented. The spectral envelope estimation using cepstral techniques is also discussed. Finally, the audio feature MFCC is extracted and visualized. This project presents several examples of speech and music instruments.

2 Cepstrum and Pitch Estimation

2.1 Cepstrum

The basic system of human speech generation includes two parts, vocal source and vocal tract (Figure 1). The vocal source signal $s(t)$ is the glottal pulses produced by vocal cords. The vocal tract is a complex system that consists of the tongue, nasal cavity, and other elements. It provides resonances and acts as a filter on the glottal pulses in terms of signal processing, which is specified by its impulse response $h(t)$.

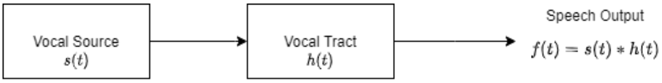


Figure 1: Basic Speech Generation System.

This excitation-resonance model is also called a source-filter model. Generally, the glottal pulses carry information about pitch, whereas the impulse response of the vocal tract filter carries information about the timbre and the content of the speech.

The output speech signal $f(t)$ is the convolution of $s(t)$ and $h(t)$. Thus, the spectrum of the speech signal $F(w)$ is the product of the spectrum of the vocal source $S(w)$ and that of the vocal tract impulse response $H(w)$. These relationships can be expressed as:

$$f(t) = s(t) * h(t) \tag{1}$$

$$F(\omega) = S(\omega) \cdot H(\omega) \tag{2}$$

The speech signal is quasiperiodic. If the period is T seconds, the power spectrum of the speech signal $|F(\omega)|^2$ will also be periodic, consisting of harmonics spaced T^{-1}Hz. Therefore, the direct way to measure this period is to take another Fourier transform. This method is known as autocorrelation. This approach, however, cannot separate the effects of the vocal source and vocal tract. Thus, the results consist of broad peaks, which leads to unsatisfactory performance in pitch determination.

One way to address this problem is to convert the multiplicative combination of the spectra into addition. The logarithm of a product equals the sum of the logarithms of the multiplicands. Therefore, the logarithm of the power spectrum can be calculated before taking another Fourier transform:

$$\begin{aligned} \log|F(\omega)|^2 &= \log\left[|S(\omega)|^2 \cdot |H(\omega)|^2\right] \\ &= \log|S(\omega)|^2 + \log|H(\omega)|^2 \end{aligned} \tag{3}$$

Then, the basic properties of Fourier transform preserve the addition:

$$\mathcal{F}[\log|F(\omega)|^2] = \mathcal{F}[\log|S(\omega)|^2] + \mathcal{F}[\log|H(\omega)|^2] \tag{4}$$

The real part of the result of Eq. 4 is the real cepstrum. Now, the effect of the vocal source and vocal tract is additive instead of multiplicative.

Performing Fourier transform on the spectrum is equivalent to performing the inverse Fourier transform. Thus, the common procedure of calculating the cepstrum is first performing Fourier transform, then taking the logarithm, and finally calculating inverse Fourier transform. In addition, the output is squared in this project to make the peaks in the cepstrum more pronounced, as did in Noll's paper [4].

In general, the term "cepstrum" is defined as the spectrum of the logarithm power or amplitude spectrum. Cepstra using power spectrum and amplitude spectrum are actually the same. After taking the logarithm, it is only a matter of the scale. Moreover, the result is neither in the time domain, nor in the frequency domain. To prevent confusion, the new representation domain is called the “quefrency domain”. The sharp peaks in the quefrency domain are defined as “rahmonics”. Besides, the filtering operation in the quefrency domain is called “liftering”.

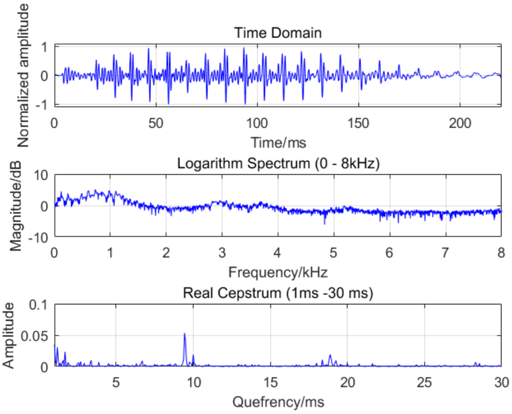


Figure 2: Word ”ALL” Pronounced by a Male.

An example of the cepstrum of a speech signal is shown in Figure 2. The speech content is the word “all” pronounced by a male. The top subfigure shows the signal in the time domain. The middle subfigure presents the logarithm spectrum. In the logarithm spectrum, the broad peaks are the formants. The bottom subfigure is the cepstrum calculated. The peak located near 9 ms is the first rahmonic.

In this project, the spectrum is calculated by the Fast Fourier Transform (FFT). To force the system to perform FFT, the signal is zero-padded to the length of the smallest power of two greater than the length of the signal.

2.2 Pitch Estimation

In the logarithm spectrum, the high-frequency ripples are produced by the vocal source, which corresponds to the higher part of the cepstrum. Thus, the higher part of the cepstrum carries information about the pitch.

The rahmonics in the higher part of the cepstrum correspond to the harmonics in the spectrum. Thus, to extract pitch information, the first rahmonic needs to be picked automatically. The idea is to find the maximum value in the part where the first rahmonic is most likely to appear. Thus, when picking the first rahmonic for speech signal, I found the maximum value in the range of 2 ms and 20 ms, which corresponds to 50-500 Hz in the frequency domain. Once having the first rahmonic, the reciprocal of its quefrency value is the corresponding frequency of the speech.

A set of examples used to estimate pitch is shown in Figure 3. The audio samples are the word "all" respectively pronounced by a male and a female. As shown in the figure, the rahmonic in the cepstrum of the female speech signal is in the lower part of the cepstrum than that of the male speech signal. This matches with the fact that the female voice is usually higher than the male voice. The pitches estimated by the model are 105.2632 Hz for the male, and 275.8621 Hz for the female. The results match with the fact that the frequency range for men is 100-120 Hz, and one octave higher for women.

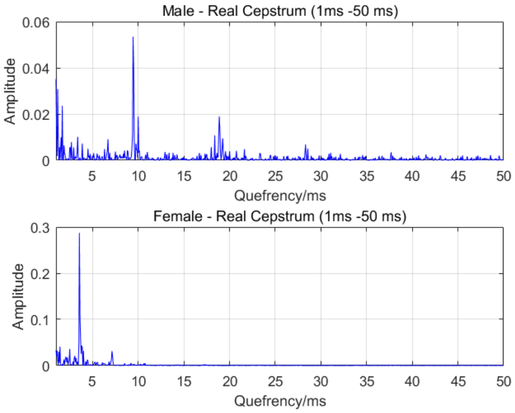


Figure 3: Ceptra of Male and Female Speech.

3 Short-Time Cepstrum

However, speech signals change with time. The pitch estimation discussed above is just a rough estimation. The cepstrum of a long speech signal cannot show how the quefrencies evolve over time. Thus, a short-time cepstrum, also known as cepstrogram, is required. As the idea of the spectrogram, a sliding window $w(t)$ is applied to the speech signal:

$$g(t) = f(t) \cdot w(t) \tag{5}$$

Through the window, only the signal of a limited time can be viewed. This portion of the signal is analyzed to produce the cepstrum. Then, the window slides to the next portion and the next cepstrum is produced. This process repeats until the end of the signal.

To get the cepstrum of the kth windowed portion, the kth spectrogram $F_k(m)$ is first calculated:

$$F_k(m) = \sum_{l=-L}^{L} w(l) f[(k-1)K + l] e^{-jlm\Delta t \Delta \omega}, \tag{6}$$

where $L = T_w/\Delta t$, $K = T_H/\Delta t$, $m = 0, 1, ..., \omega_c/\Delta\omega$, T_w is the window length, T_H is the hop size and ω_c is the bandwidth of $f(t)$.

Then, after taking the logarithm and the real part of its Fourier transform, the result is the kth short-time cepstrum:

$$C_k(n) = \sum_{m=0}^{M} \log |F(\omega)|^2 \cos mn\Delta\gamma\Delta\omega, \tag{7}$$

where $M = \omega_c/\Delta\omega$, $n = 0, 1, ..., N$, N is the upper limit on the desired quefrencies in the cepstrum.

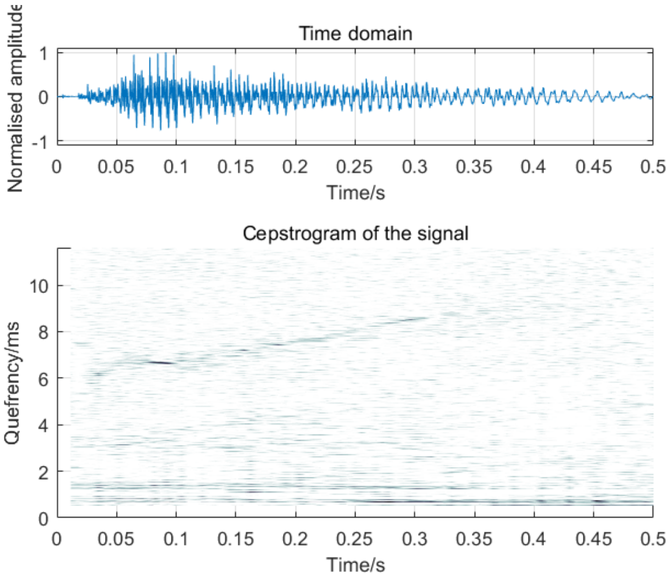


Figure 4: Ceptrogram of Diphthong Vowel /əʊ/ Pronounced by a Male.

Figure 4 shows an example of cepstrogram. The audio analyzed here is the diphthong vowel /əʊ/ pronounced by a male. With the cepstrogram, we can get how the pitch changes with time. In this example, the quefrency value that corresponds to the first rahmonic becomes higher and higher. This means that the pitch of the speech becomes lower and lower as time goes on. It is worth noting that by observing the examples of cepstrograms shown in Appendix A, it can be concluded that the pitches decrease for all the diphthong vowels.

4 Spectral Envelope

What is discussed above is the information carried in the vocal source signal. The impulse response of the vocal tract filter also contains important information.

A spectral envelope is a smoothing of a spectrum that tries to leave the detailed structure of the spectral line aside, while retaining the general shape of the spectrum. Audio signals generate a harmonic spectrum that is superimposed on a spectral envelope. The way that human beings identify sounds, whether through the ear or the brain, is highly influenced by the fact that the spectral envelope works as a guide to the detection and classification of sounds.

One way to estimate the spectral envelope is to use cepstral techniques. The spectral envelope can be estimated by the FFT values of the lower part of the cepstrum.

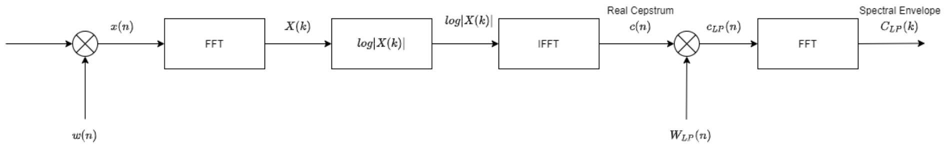


Figure 5: Spectral Envelope Computation Using Cepstrum.

Figure 5 illustrates the steps to estimate the spectral envelope from the real cepstrum. The FFT of the windowed input signal $x(n)$ is first calculated:

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn} = |X(k)| e^{j\varphi_x(k)}, k = 0, 1, ..., N-1 \tag{8}$$

Then, by taking the logarithm of the amplitude and IFFT, the real cepstrum is obtained:

$$c(n) = 1/N \sum_{k=0}^{N-1} \log |X(k)| W_N^{-kn}. \tag{9}$$

To get the lower part that corresponds to the vocal tract, a low-pass window W_{LP} is applied to the real cepstrum.

$$W_{LP} = \begin{cases} 1, & n = 0, N_1 \\ 2, & 1 \leq n < N_1 \\ 0, & N_1 < n \leq N - 1 \end{cases} \tag{10}$$

where N_1 is the cutoff quefrency of the low-pass lifter.

Finally, the FFT of the windowed real cepstrum generates the spectral envelope:

$$C_{LP}(k) = FFT[c_{LP}(n)] \tag{11}$$

Figure 6 shows an example of the spectral envelope of speech. By subtracting the logarithm of the spectrum with the spectral envelope, only the details that correspond to the glottal pulses remain. In this way, the effects of the vocal source and vocal tract are separated. With regard to audio processing, however, what is of great importance is the identity of the sound, i.e., the formants carried by the spectral envelope.

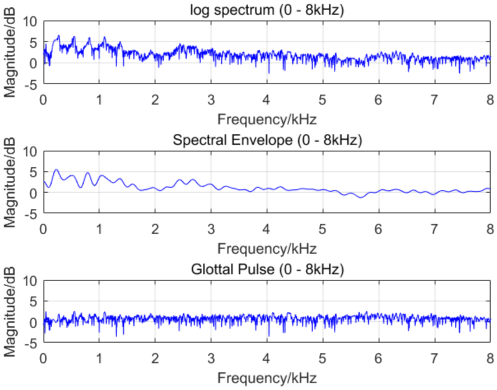
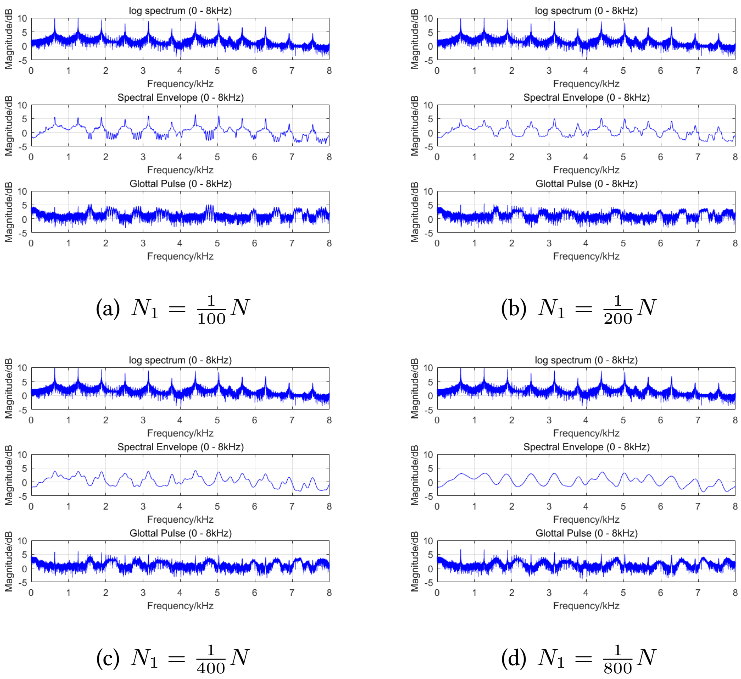


Figure 6: Spectral Envelope of Front Vowel /ə/ Pronounced by a Male.

More examples of spectral envelopes of different vowels and musical instruments can be found in Appendix B. It can be concluded that the spectral envelopes of different signals are unique. Moreover, during the experiments, I noticed that different kinds of signals need different cutoff quefrencies. To get a proper spectral envelope estimation of speech, I set the cutoff quefrency N_1 in the range of 5 ms to 10 ms. It is worth noting that the proper cutoff quefrencies for back vowels tend to be smaller than those for front vowels and center vowels. Besides, I set $N_1 = 10$ ms for the violin, and $N_1 = 8$ ms for the trumpet.



(a) $N_1 = \frac{1}{100}N$

(b) $N_1 = \frac{1}{200}N$

(c) $N_1 = \frac{1}{400}N$

(d) $N_1 = \frac{1}{800}N$

Figure 7: Spectral Envelopes of Violin at A4

In addition, the cutoff quefrency value is inversely proportional to the amount of details in the spectral envelope. As shown in Figure 7, the larger the cutoff quefrency is, the more details will appear in the spectral envelope.

5 MFCC

In any automatic speech recognition scheme, the first step is always to extract features, i.e., to recognize the attributes that are useful in characterizing the phonetic content and discarding other elements that carry information such as background noise and emotion. In the speech generation system, the shape of the vocal tract determines what sounds come out, which manifests itself in the spectral envelope. Proposed by Davis and Mermelstein in 1980 [6], MFCC is a widely used audio feature that can accurately represent the spectral envelope.

Figure 8 shows the flow chart of the algorithm used to get MFCC. The first step is to filter the input signal with a high-pass filter. The aim is to emphasize the high-frequency components and make the spectrum flat. This operation can compensate for the high-frequency part suppressed by the vocal tract and highlight the formants of high frequencies.

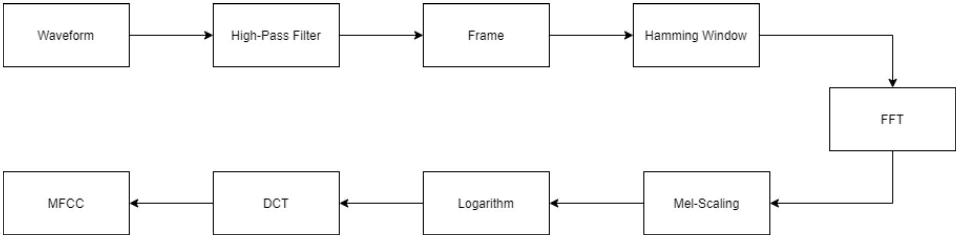


Figure 8: Flow Chart of MFCC Computation.

The output of the filter is then framed into short frames of 20-40 ms. As mentioned above, the speech signal is constantly changing. For simplicity, we assume that the signal does not change statistically on short time scales. The length of the frame is a compromise. If it is too short, there are not enough samples to get a reliable estimate of the spectrum. By contrast, if the frame length is too long, the signal changes too much throughout the frame.

The next step is to pass each frame through a window function. The purpose is to eliminate the discontinuity caused at both ends of each frame. The window function used in this project is a hamming window.

Then, the power spectrum of each frame is calculated in order to observe the power distribution in the spectral domain. The way used here is to implement the FFT and take the square of the amplitude.

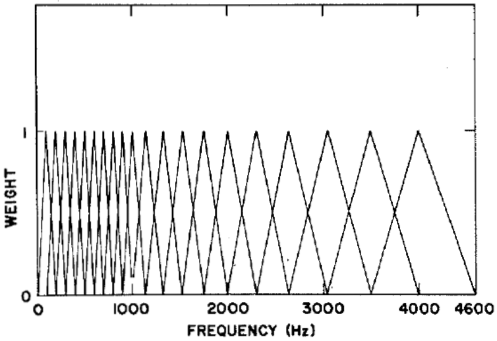


Figure 9: Mel Filterbank [6].

The human auditory system is a special nonlinear system. Its sensitivity to signals of different frequencies is different. Thus, the next step is to map the linear spectrum to the Mel nonlinear spectrum based on auditory perception. The Mel scale relates the perceived frequency of a signal to its actual frequency. This is fulfilled by passing the spectrum through the Mel filterbank (Figure 9). In this project, 24 triangular bandpass filters are performed, whose center frequencies are $f(m), m = 1, 2, ..., 24$. The interval between $f(m)$ decreases as the value of m decreases, and widens as the value of m increases.

The following step is to take the logarithm of the Mel spectrum and to compute the inverse transform. Discrete Cosine Transform (DCT) is widely used to perform the inverse transform here. The reason is that the DCT can decorrelate the overlapping filterbank energies in the classifier. Besides, higher coefficients provide information about rapidly changing spectral details, which does not matter much in terms of speech recognition. Traditionally, only the first 12-13 DCT coefficients are kept because they keep the most relevant information, which is the information about the formants, spectral envelope, etc.

However, the standard MFCC only reflects the static characteristics of speech. The feature extracted from each frame only reflects the characteristics in the specific frame. To make the feature reflect continuity in the time domain better, information about dynamic characteristics can be added to the feature matrix. Difference spectra contain information about the previous and subsequent frames. Thus, the following step is to take the first and second derivatives of MFCC. Finally, by concatenating the MFCC and its first and second derivatives, the MFCC feature is extracted from the signal.

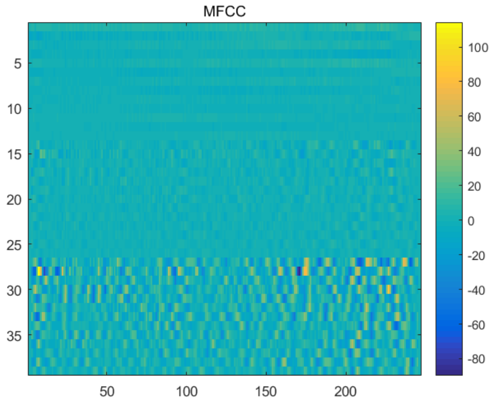


Figure 10: MFCC Feature of Diphthong Vowel /əʊ/. The x-axis represents time in milliseconds; the y-axis represents MFCC index.

Figure 10 shows the MFCC feature extracted from a speech signal. The x-axis represents time in milliseconds. The top 13 coefficients are the standard MFCC of the signal. The middle 13 coefficients are the first derivatives of the MFCC. The bottom 13 coefficients are the second derivatives. More examples of MFCC and MFCC features of musical instruments can be found in Appendix C.

The most significant advantage of MFCC is that it provides information about formants. It describes the structure of the spectrum and ignores the fine details. With the advent of machine learning algorithms, MFCC becomes a main audio feature in audio processing problems.

One of the limitations of MFCC is that it is based on the biases of the human auditory system. In other words, it is not the machine that decides which element is the most relevant in an audio signal. Another limitation is that MFCC is only suitable for analysis, but not for synthesis. MFCC does not have an inverse so audio signals cannot be synthesized using MFCC.

6 Conclusion

This project introduced the cepstral analysis of audio signals. The general idea of cepstrum is to separate the excitation signal and the impulse response of the resonance system.

With the information carried by the excitation signal, the pitch can be estimated. Besides, the short-time cepstrum shows how the source signal evolves over time.

The spectral analysis of the system response produces the spectral envelope of the audio signal, which carries information about the audio identity. When calculating the spectral envelope, different signals need different cutoff quefrencies. In addition, the value of cutoff quefrency is inversely proportional to the amount of detail in the spectral envelope.

Furthermore, the audio feature MFCC that represents the formants can be obtained by mapping the spectrum to Mel spectrum based on human perception. Admittedly, MFCC has certain limitations. It is based on how human beings perceive sounds. Besides, it is not efficient in sound synthesis. Despite all the limitations, MFCC is a reliable audio feature of audio processing in general. In recent research, MFCC is combined with other algorithms and features to improve the result.

References

[1] Alan V Oppenheim and Ronald W Schafer. From frequency to quefrency: A history of the cepstrum. *IEEE signal processing Magazine*, 21(5):95–106, 2004.

[2] B Bogert, M Healy, and J Tukey. The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe craking. In *Proc. Symp. On Time Series Analysis*, pages 209–243, 1963.

[3] A Michael Noll. Short-time spectrum and "cepstrum" techniques for vocal-pitch detection. *The Journal of the Acoustical Society of America*, 36(2):296–302, 1964.

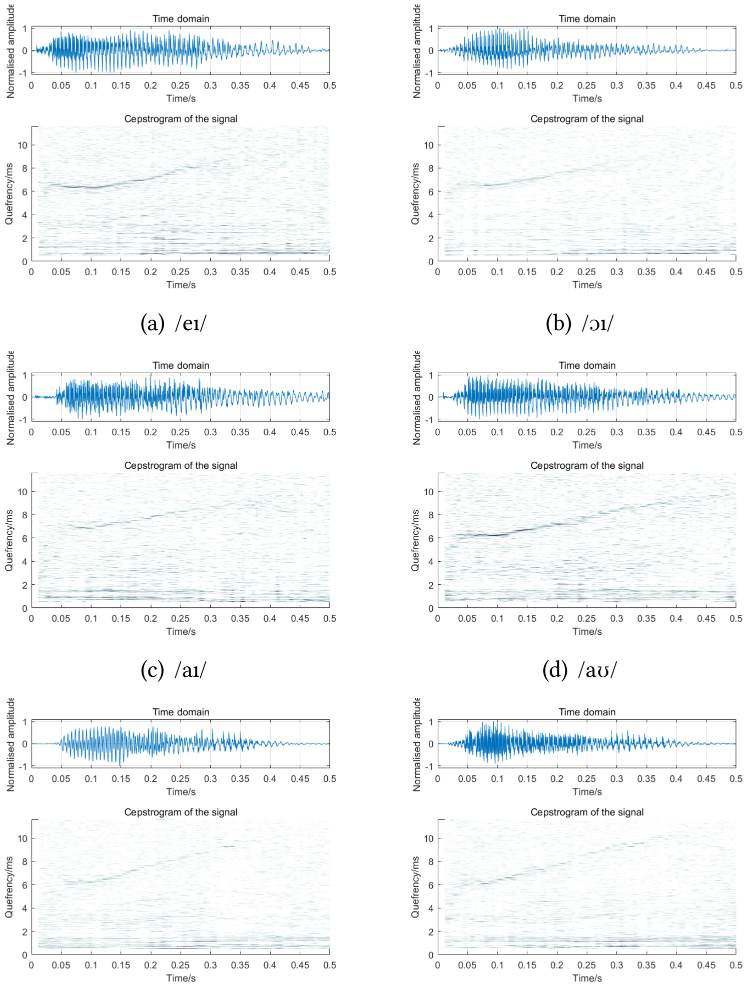
[4] A Michael Noll. Cepstrum pitch determination. *The journal of the acoustical society of America*, 41(2):293–309, 1967.

[5] Alan V Oppenheim. Superposition in a class of nonlinear systems. 1965.

[6] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

Appendices

Appendix A Examples of Cepstrograms



(a) /eɪ/

(b) /ɔɪ/

(c) /aɪ/

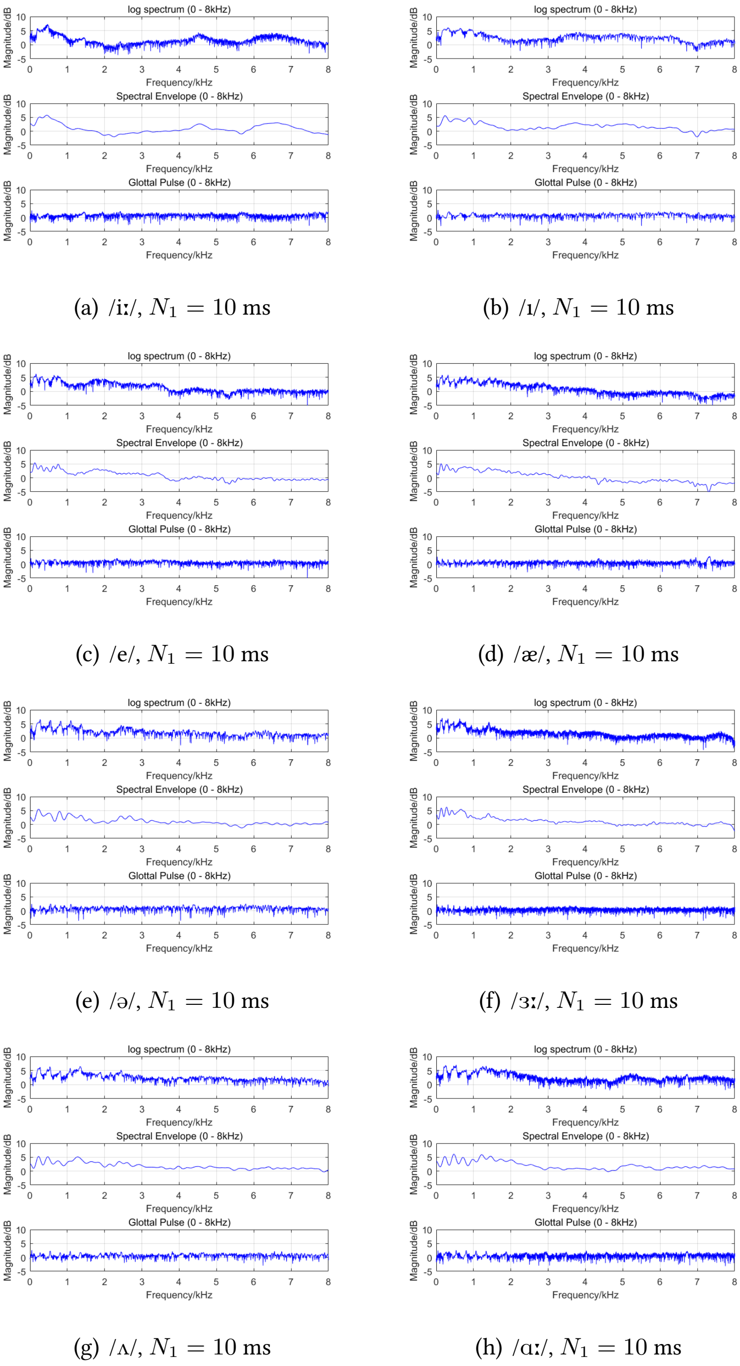
(d) /aʊ/

(e) /ɪə/

(f) /eə/

Figure 11: Cepstrograms of Diphthong Vowels

Appendix B Examples of Spectral Envelopes



(a) /iː/, $N_1 = 10$ ms

(b) /ɪ/, $N_1 = 10$ ms

(c) /e/, $N_1 = 10$ ms

(d) /æ/, $N_1 = 10$ ms

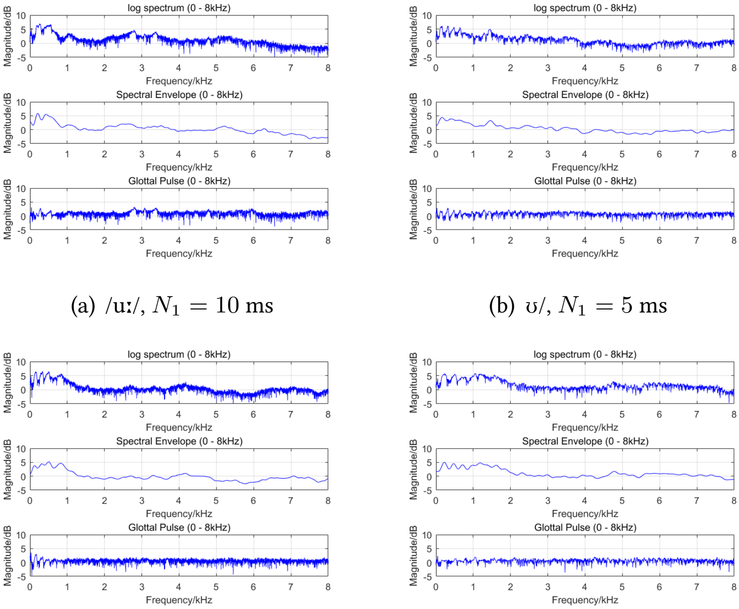
(e) /ə/, $N_1 = 10$ ms

(f) /ɜː/, $N_1 = 10$ ms

(g) /ʌ/, $N_1 = 10$ ms

(h) /ɑː/, $N_1 = 10$ ms

Figure 12: Spectral Envelopes of Front Vowels



(a) /uː/, $N_1 = 10$ ms

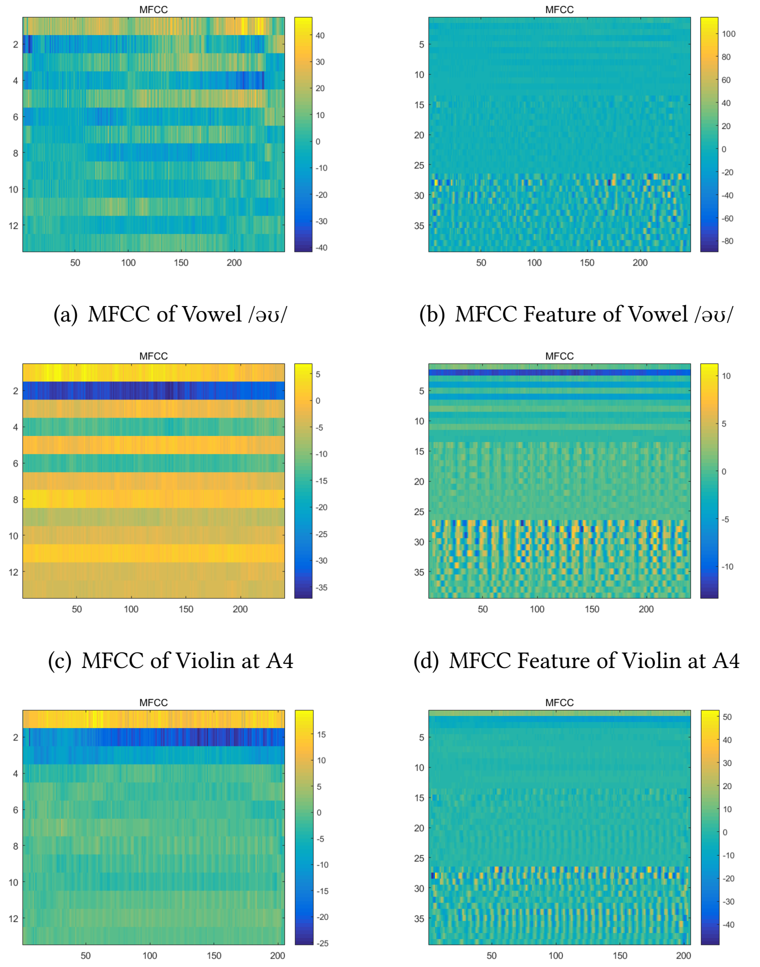
(b) ʊ/, $N_1 = 5$ ms

(c) /ɔː/, $N_1 = 7$ ms

(d) /ɐ/, $N_1 = 8$ ms

Figure 13: Spectral Envelopes of Front Vowels

Appendix C Examples of MFCC Features



(a) MFCC of Vowel /əʊ/

(b) MFCC Feature of Vowel /əʊ/

(c) MFCC of Violin at A4

(d) MFCC Feature of Violin at A4

(e) MFCC of Trumpet D♯4

(f) MFCC Feature of Trumpet at D♯4

Figure 14: MFCC Examples. The x-axis represents time in milliseconds; the y-axis represents MFCC index.

Appendix D Links

- The code of this project can be found via this link: `https://github.com/yinanazhou/cepstral-analysis`

- The audio samples are downloaded from this website: `https://pronunciationstudio.com/45-sounds-ebook/`