NORTHERN ILLINOIS UNIVERSITY

DEKALB, ILLINOIS

APRIL 2021

FEATURE EXRACTION METHODS IN SPEECH PROCESSING

BY

MUSTAFA OZTOPRAK

27 APRIL 2021

ELE 653 DIGITAL SPEECH PROCESSING

FINAL PROJECT

ELECTRICAL ENGINEERING

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL ENGINEERING

Director:

LICHUAN LIU

ACKNOWLEDGEMENTS

We would like to express our special thanks to Professor Lichuan Liu who gave us critical and necessary insight for doing this project on feature extraction methods which also lead us to learn and emphasize so many new things through the semester.

ABSTRACT

Speech processing includes the various techniques such as speech coding, speech synthesis, speech recognition and speaker recognition. In the area of digital signal processing, speech processing has versatile applications, so it is still an intensive field of research. Speech processing mostly performs two fundamental operations such as Feature Extraction and Classification. The main criterion for the good speech processing system is the selection of feature extraction technique which plays an important role in the system accuracy. In this project, Linear Prediction coding (LPC), Linear Prediction Cepstral Coefficient (LPCC), Mel Frequency Cepstral Coefficient (MFCC), and Bark frequency Cepstral coefficient (BFCC) feature extraction techniques for recognition of male and female voice have been studied, and the corresponding recognition rates are compared. The experimental results show that a better recognition rate is obtained for MFCC as compared to LPC, LPCC, and BFCC for all the four types.

TABLE OF CONTENTS

Page
LIST OF FIGURES..... iii
CHAPTER
1. INTRODUCTION……1
2. FEATURE EXCTRACTION METHODS……..2
2.1. Linear Predictive Coding 3
2.2. Linear Predictive Cepstral Coefficients 6
2.3. Mel Frequency Cepstral Coefficients……...9
2.3. Bark Frequency Cepstral Coefficients…….13
3. EXPERIMENTAL RESULTS……15
4.SIMULATION AND RESULTS…….15
5.CONCLUSION…….18
6.REFERENCES……..19

LIST OF FIGURES

Page

1. Steps for LPCC 7
2. Mel frequency curve..... 10
3. MFCC feature extraction procedures. 11
4. Filter banks of MFCC … 12
5. MFCC spectrum. 13
6. Block Diagram of BFCC Process..... 14
7. LPC comparation on female and male voice 16
8. LPCC comparation on female and male voice.....16
9. MFCC comparation on female and male voice.....17
10. BFCC comparation on female and male voice.....17
11. Comparation of different features on same voice 18

LIST OF ACRONYM

LPC: Linear Predictive Coding

LPCC: Linear Predictive Cepstral Coefficients

MFCC: Mel Frequency Cepstral Coefficients

BFCC: Bark Frequency Cepstral Coefficients

1.INTRODUCTION

Feature extraction is an essential part of audio analysis. Analysis voice signal in the time or frequency domain highlights its characteristics, which is convenient for us to compare and classify. An excellent acoustic feature needs to consider the following aspects: First, it has excellent distinguishing characteristics so that various sounds can extract features. Second, feature extraction is also a compression coding process. It is necessary to retain or eliminate some related factors (such as the characteristics of the channel. When identifying the speaker, the information of the human throat channel is maintained, and the characteristics of the environmental channel are eliminated) and it is necessary to use it as low as possible without losing too much useful information. Parameter dimension is convenient for efficient and accurate model training. Finally, we need to consider robustness, that is, the ability to resist interference from environmental noise[1].

2.FEATURE EXCTRACTION

In order to achieve higher accuracy in speech recognition, selecting appropriate features from a speech signal is the central concern. Feature extraction process is grounded on the basis of discarding the irrelevant information from the speech signal and only keeping the useful content. As the raw speech signal is always complex thus feeding the said as an input to the classifier may not be suitable, hence the requirement for a high-quality front-end arises. The primary aim of feature extraction

is to find a set of properties of an utterance that have acoustic correlations to the speech-signal, that is parameters that in some way can be computed or estimated through processing of the signal waveform. Such parameters are termed as features. There are many properties of features which includes high discrimination between sub-word classes, low speaker variability, invariableness to degradations in the speech signal due to noise and channel [2]. The feature extraction methods considered in this project are Linear predictive coding (LPC), Linear Prediction cepstral coefficients (LPCC), Mel frequency cepstral coefficients (MFCC) and Bark Frequency cepstral coefficients (BFCC). A concise explanation of each of the feature extraction method is given below.

2.1 Linear Predictive Coding (LPC)

The basic idea of linear prediction coefficient (LPC) analysis of speech signals is a sample of speech can be approximated by a linear combination of several previous speech samples, and the model predicted by linear approximation to the actual speech sample in the sense of minimum mean square error can be obtained a unique set of prediction coefficients. The prediction coefficient is the weighting coefficient used in the linear combination. This linear prediction analysis was first used in speech coding.

LPC refers to an input signal with a model, that is, the speech signal as the output of a particular model or system so that the parameters of the model can be used to describe the target signal [3].

Assuming that the input of the model is u (n), the output is x (n), and the channel model is H (z), then the parameters of the model can be solved by using the transfer function method. The channel model is :

$$H(z)=G\frac{1+\sum_{l=1}^{q}b_l z^{-l}}{1-\sum_{i=1}^{p}a_i z^{-i}}=G\frac{B(z)}{A(z)} \tag{1}$$

Convert to time domain:

$$x(n)=\sum_{i=1}^{p}a_i x(n-i)+G\sum_{l=0}^{q}b_1 u(n-l), b_0=1 \tag{2}$$

Equation 2 is a linear constant-coefficient difference equation: u (n-l) is the past input of the model, u (n) is the current input to the model, and u (n-i) is the past input of the model. To simplify the solution, we simplified the model and adopted an autoregressive model; that is, the current output is a linear combination of the current input and the past output:

$$\begin{aligned} x(n)&=\sum_{i=1}^{p}a_i(n-i)+Gu(n)\\ H(z)&=\frac{G}{A(z)}=\frac{G}{1-\sum_{i=1}^{p}a_i z^{-i}} \end{aligned} \tag{3}$$

Since we got Equation 3, then we just need to solve it. First, we want to estimate x(n):

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i) \tag{4}$$

where $\hat{x}$ is the estimated value of the original signal x and a_i a linear combination of the past p outputs, where a_i is the linear prediction coefficient. Therefore, it is only necessary to ensure that the error between the predicted signal and the original signal is minimal to obtain the coefficient [4].

Equation 4 can be obtained by transformation to:

$$\sum_{i=1}^{p} a_i \phi(j,i) = \phi(j,0) \tag{5}$$

where $\phi(j,i) = \sum_{n}^{p} s(n-j)s(n-i)$

So we only need $\phi(j,i)$ to get the coefficient a_i. Based on the definition, we get

$$r(j) = \sum_{n=0}^{N-1} x(n)x(n-j), 1 \le j \le p \tag{6}$$

Therefore

$$r(|j-i|) = \phi(j,i) \tag{7}$$

So we can turn Equation 5 into

$$\sum_{i=1}^{p} a_i r(|j-i|) = r(j) \tag{8}$$

When i=0, $E_0 = r(0), a_0 = 1$.

When i= 1, 2, 3, ...p:

$$k_i = \frac{1}{E_{i-1}}\left[r(i) - \sum_{j=1}^{i-1} a_j^{i-1} r(j-i) \right] \tag{9}$$

$$a_i^{(i)} = k_i \tag{10}$$

For j=1 to i-1:

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \tag{11}$$

$$E_i = \left(1 - k_i^2\right) E_{i-1} \tag{12}$$

Repeating from 9 to12, we can get:

$$a_i = a_j^{(p)} \tag{13}$$

So far, we have obtained the p-order LPC eigenvalues of the speech signal.

2.2 Linear prediction cepstral coefficients (LPCC)

For estimating the basic parameters of a speech signal, LPCC has become one of the predominant techniques. The basic theme behind this method is that one speech sample at the current time can be predicted as a linear combination of past speech samples. Algorithm for LPCC is shown in Figure 1.

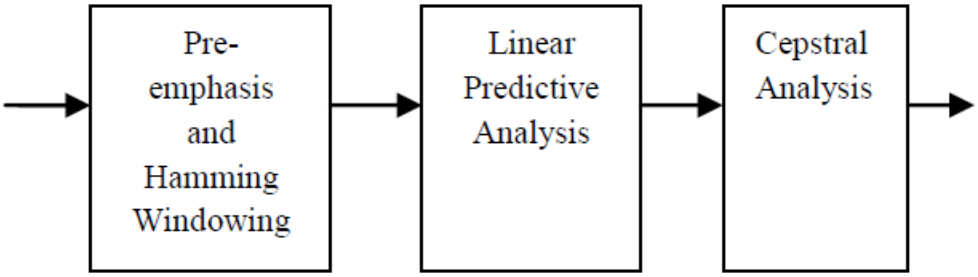

Figure 1. Steps for LPCC

To understand the difference between LPCC and LPC, we must first understand what cepstrum is. The cepstrum is somewhat similar to the autocorrelation series. We use the following formula to understand the cepstrum gradually [5].

Suppose there is a signal x(n); taking the discrete Fourier transform we get complex spectrum:

$$DFT(x(n)) \rightarrow X(k) \tag{14}$$

when we take the inverse discrete Fourier transform, we can get the signal back:

$$IDFT(X(k)) \rightarrow x(n) \tag{15}$$

So, if we take the square of the absolute of the DFT of x(n), we get the power spectrum:

$$\left|DFT(x(n))\right|^2 \rightarrow P(k) \tag{16}$$

Now, if we take the inverse discrete Fourier transform, instead of getting the original signal x(n) we get autocorrelation sequence:

$$IDFT(\log(P(k))) \rightarrow A(n) \tag{17}$$

And if we take the log of the power spectrum first, then we take the inverse discrete Fourier transform. We will get the standard cepstrum:

$$IDFT(\log(P(k))) \rightarrow C(n) \tag{18}$$

Linear prediction cepstral coefficients are computed in the same way as standard cepstrum, except they are calculated from the smoothed autoregressive power spectrum instead of the periodogram estimate of the power spectrum. A simple recursive formula for computing linear prediction cepstral coefficients directly from LPCs without any DFTs shown in equation 19.

$$c(n) = \begin{cases} 0 & n < 0 \\ \ln(G) & n = 0 \\ a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & 0 < n \le p \\ \sum_{k=n-p}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & n > p \end{cases} \tag{19}$$

where a_k is the linear prediction coefficient

The main advantage of LPCC is that it completely removes the incentive information in the voice generation process. It mainly reflects the channel response. From a finite number of LPC coefficients, an infinite number of cepstral coefficients can be calculated. Often only a dozen cepstrum coefficients are needed to better describe the formant characteristics of speech [6].

2.3 Mel-frequency cepstral coefficient (MFCC)

The most used speech feature is Mel-frequency Cepstral Coefficient (MFCC) in terms of speech processing. According to the human ear hearing mechanism study, human hearing sensitivity to sound waves of different frequencies is not the same. Speech signals from 200Hz to 5000Hz have a more significant impact on speech intelligibility. When two sounds with different loudness are acting on the human ear simultaneously, the higher loudness frequency component will affect the frequency perception of the lower loudness component to the human ear, making the low loudness part challenging to detect. This phenomenon is called the "masking effect." Because of the characteristics of the human ear, low-frequency sounds are more likely to cover up high-frequency sounds.

Conversely, high-frequency sounds are relatively difficult to cover up low-frequency sounds. Therefore, a group of band-pass filters is arranged from dense to sparse according to the critical bandwidth from low to high frequencies. The output results obtained when the input signal passes through this set of filters can be used as the essential characteristics of the original signal. After processing this primary feature, it can be used as the input feature of speech [7].

MFCC is a cepstrum coefficient extracted in the frequency domain of the mel scale. The mel scale describes the non-linear characteristics of the human ear, and its relationship with frequency can be approximated as:

$$Mel(f) = 2595ln\left(1 + \frac{f}{600}\right) \tag{20}$$

where f is the frequency, and the unit is Hz. Figure 2 shows the relationship between mel scale and linear frequency.

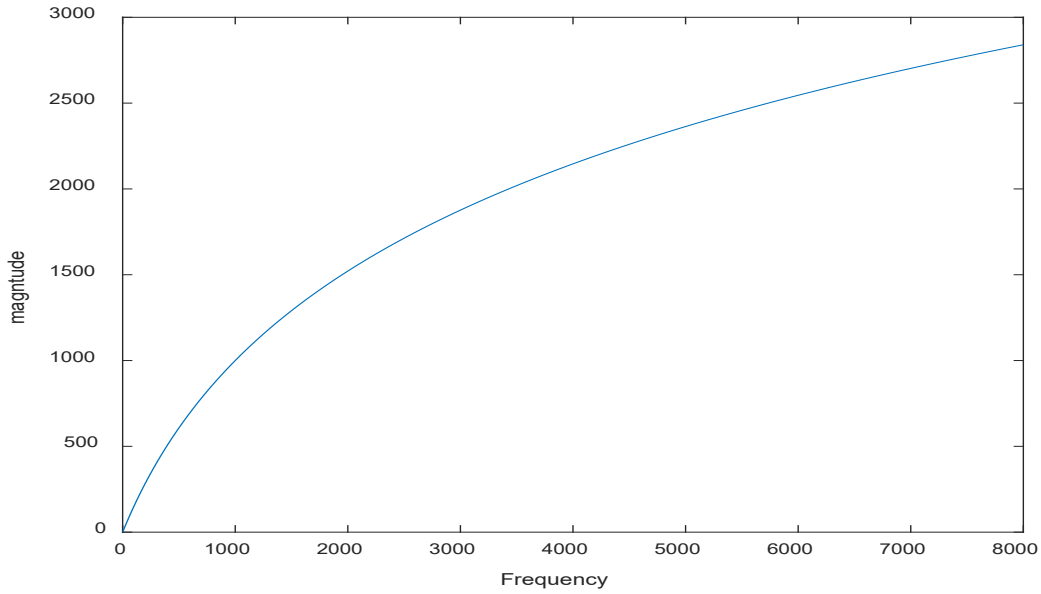

Figure 2. Mel frequency curve.

MFCC feature extraction procedure is illustrated in Figure 3.

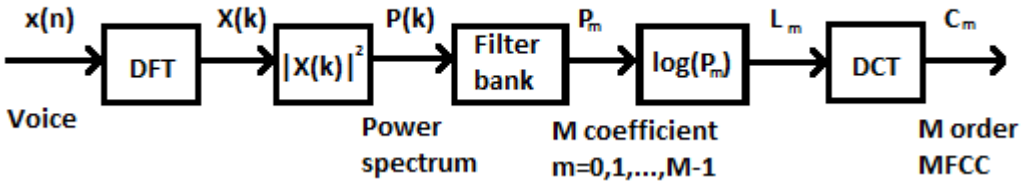

Figure 3. MFCC feature extraction procedures.

When we get the word pieces, we are using MFCC to get the feature of each piece by the following steps:

First, we take discrete Fourier transform (DFT) of signal; mathematic expression of N points DFT can be described as follows:

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-\frac{j2\pi nk}{N}} \tag{21}$$

Then square each spectrum amplitude value to get power spectrum:

$$P(k)=|X(k)|^2 \tag{22}$$

Convolute the power spectrum P(k) with a mel-scaled triangular filter bank shown in Figure 4.

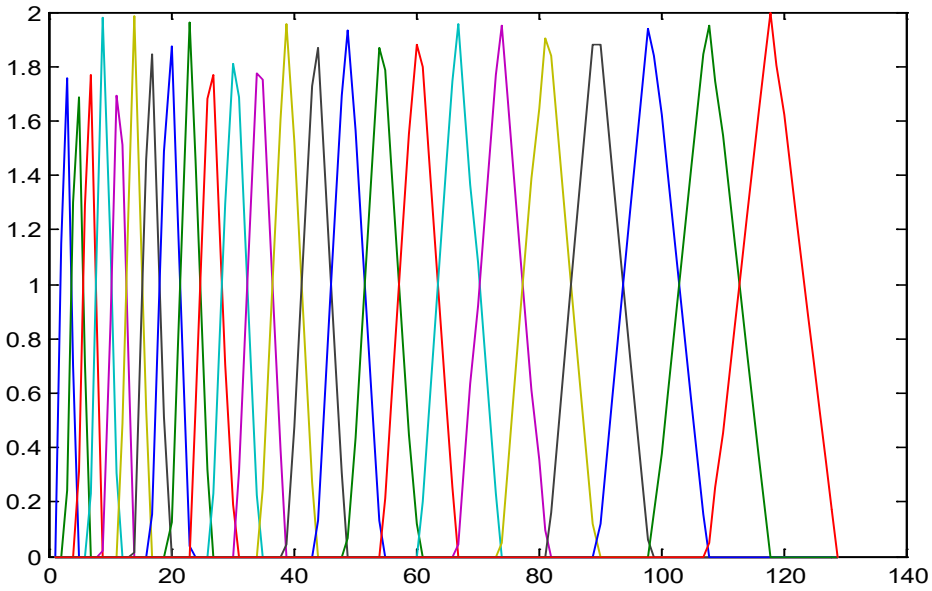

Figure 4. Filter banks of MFCC

Take logarithm

$$L_m = \log\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right), 0 \le m \le M \tag{23}$$

Finally, take discrete cosine transform (DCT) to get MFCC:

$$C_m = \sum_{n=0}^{M-1} L_m \cos\left(\frac{\pi m(n+0.5)}{M}\right), 0 \le m \le M \tag{24}$$

Then we can get the MFCC of each frame of the input audio and then integrate it into the MFCC of the entire audio segment (Figure 5).

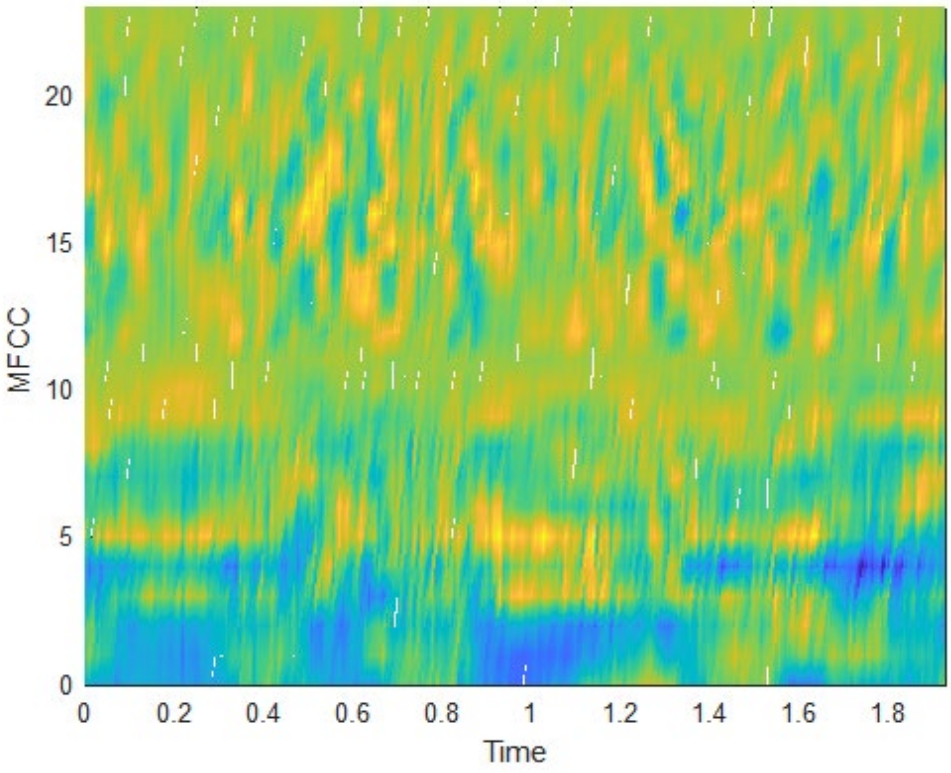

Figure 5. MFCC spectrum.

Because this feature does not depend on the nature of the signal, it makes no assumptions or restrictions on the original input signal. It makes full use of the research results of the auditory model. Therefore, this parameter has better robustness and is more in line with the hearing characteristics of the human ear. In addition, when the signal-to-noise ratio is reduced, this method still has good recognition performance [8].

2.4 Bark frequency cepstral coefficients (BFCC)

BFCC is another method for extracting the features from the speech signal. Figure 6 shows the block diagram of BFCC algorithm. This method is similar to MFCC. Implementation of bark scale filters in place of mel filters is quite obvious. Mathematically bark scale filter is represented by following formula:

$$f(bark) = 6\ln\left[\frac{f}{600} + \left(\frac{f}{600}\right)^2 + 1\right]^{\frac{1}{2}} \qquad (25)$$

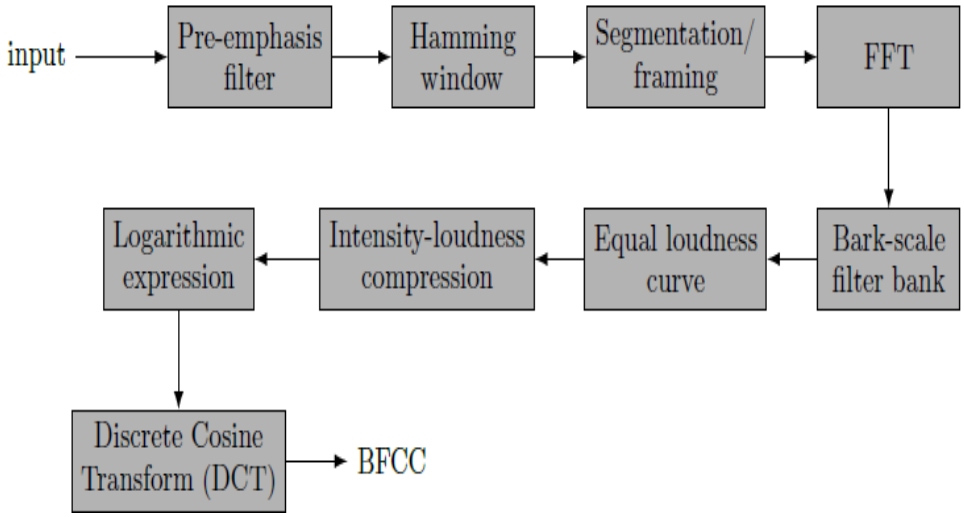

Figure 6. Block Diagram of BFCC Process

Where $fbark$, is the corresponding Bark frequency in Barks and f is the linear frequency in hertz. The filter outputs are weighted according to the equal-loudness curve, that approximates the sensitivity of human hearing. The signal is then compressed under the intensity-loudness power law where the amplitude of the signal is compressed by the cubic-root to match the non linear relationship between intensity

of sound and perceived loudness. Finally signal is first compressed through the logarithmic function and finally DCT is used to decorrelate the features as in case of MFCC[2].

3.EXPERIMENTAL RESULTS

In this section, we present two experiments such as feature extraction on different female and male voice with the evaluation of four different feature extraction methods. Experimentation has been started with the same signal from different speakers with different gender, age groups, origin. Eight speakers, four male and four female speaking each of the word, each with different emotions have been analyzed. LPC, LPCC, MFCC and BFCC have been employed as feature extraction methods and the corresponding outputs are given as input to the classifier.

4.SIMULATION AND RESULTS

We selected two speakers, one female and one male. For each person, we took two five-second speech signals for processing. First, we use STE to extract the effective part of each file and make it into word pieces. Then we make each word piece into a frame through the window. Then LPC, LPCC, MFCC and BFCC are used for feature extraction. Finally, the features are combined to obtain a feature matrix.

We compare the feature matrix:

In figure 7, Observation shows that when LPC compares female and male voices, there is a significant difference in high-level coefficients.

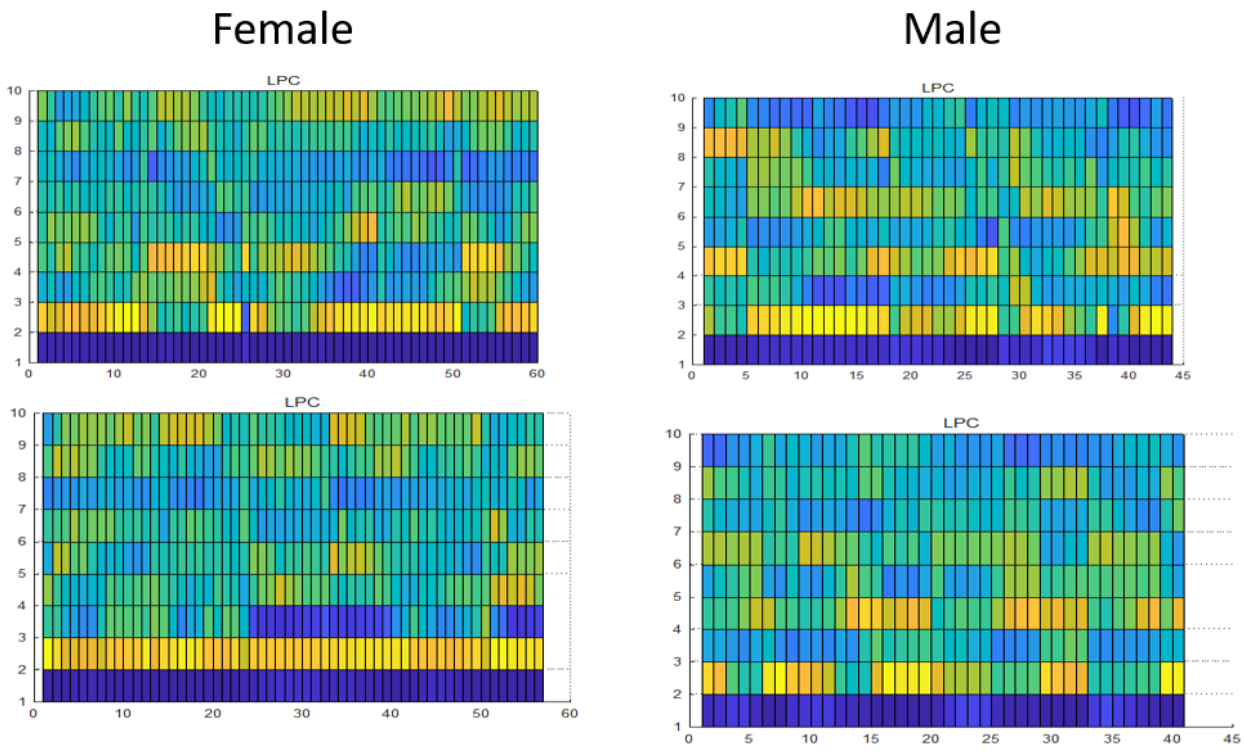

Figure 7. LPC comparation on female and male voice

In figure 8, Because LPCC is the cepstrum of LPC, we can see obvious differences in the low-level coefficients.

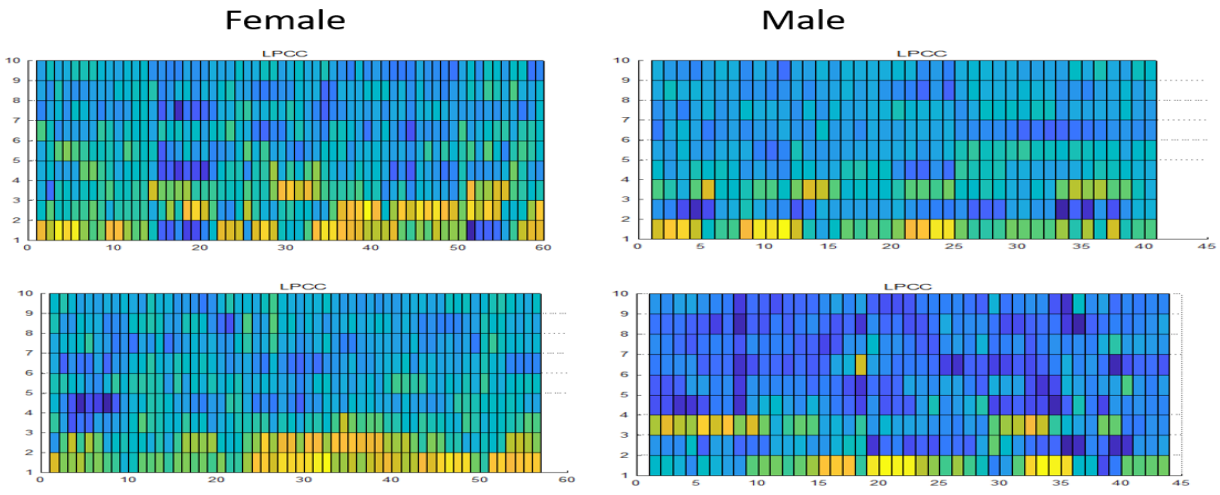

Figure 8. LPCC comparation on female and male voice

In figure 9, The coefficients of MFCC are relatively more uniform, and we can see obvious differences in the middle coefficients of MFCC.

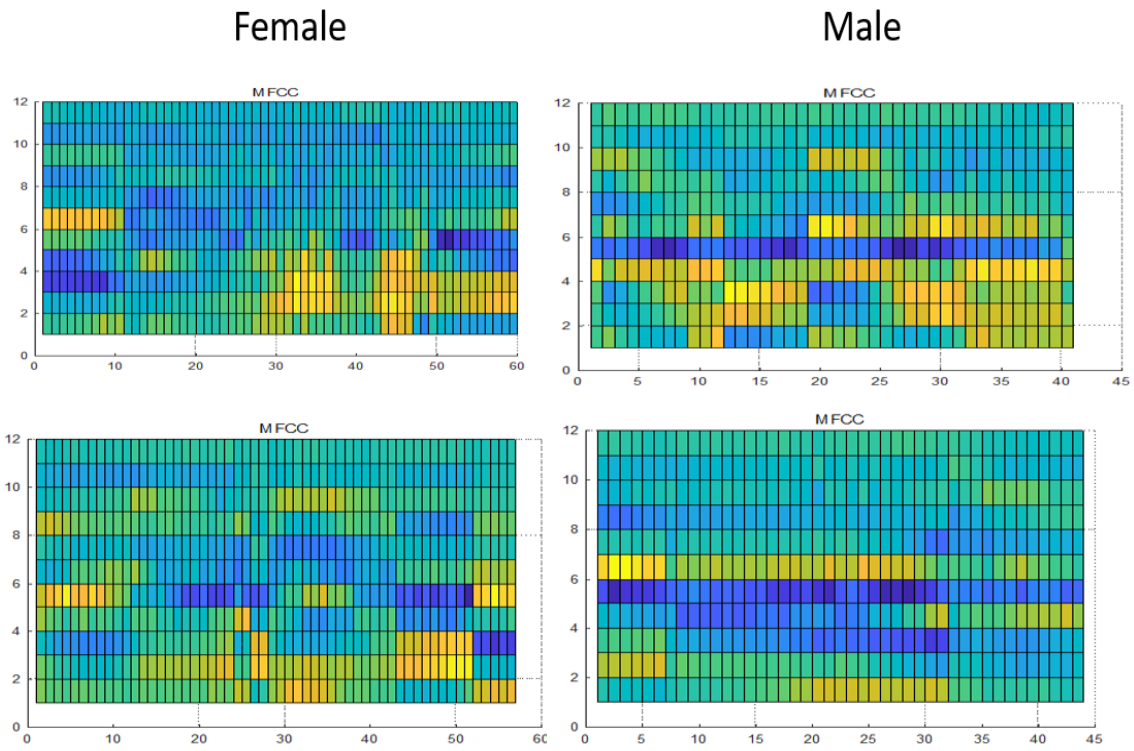

Figure 9. MFCC comparation on female and male voice

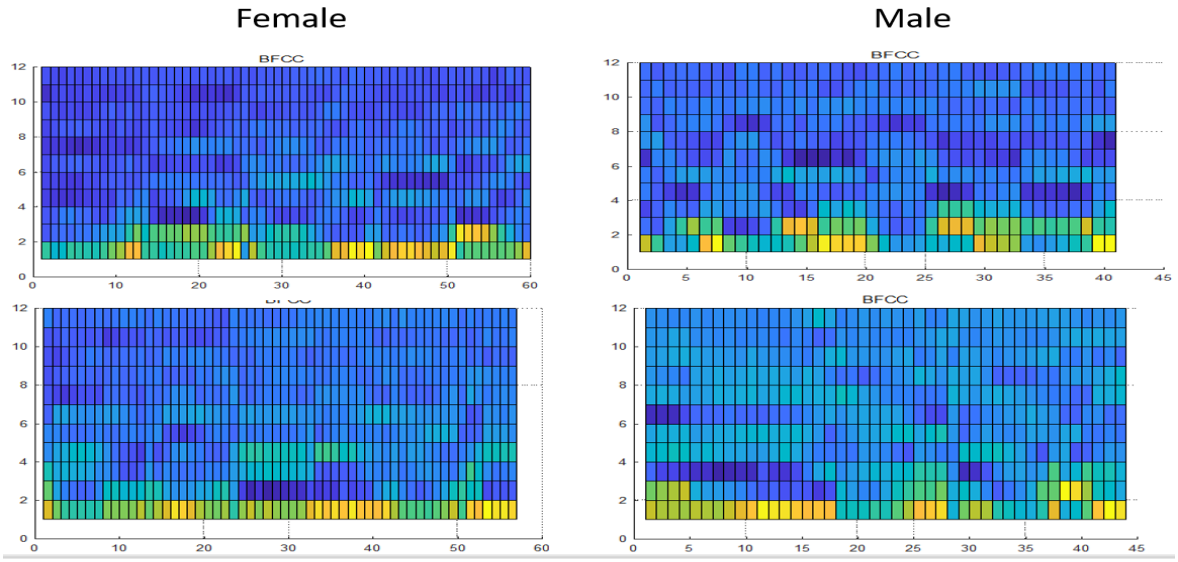

Figure 10. BFCC comparation on female and male voice

In figure 10, The effect of BFCC is not very satisfactory, and the value in the high coefficient is obviously low. But we can still find the difference between the two in this section.

5.CONCLUSION

First of all, we can know that all feature extraction methods can effectively realize the function of distinction.

In terms of the number of coefficients, MFCC and BFCC can obtain far more coefficients in each frame than LPC and LPCC. However, LPC and LPCC pay more attention to the change relationship between adjacent frames. The relationship between the initial MFCC and BFCC is relatively independent for each frame. But we can make up for this shortcoming by increasing the DCT step. The difference between MFCC and BFCC is the choice of different filter-banks. Because the Mel filter-bank is closer to the human ear, the effect of MFCC is better than that of BFCC. This is why MFCC is currently the most popular feature extraction method.

6. REFERENCES

[1] Taabish, G., Anand, S., Rajouriya, D.K. and Najma, F. 2014, A Systematic Analysis of Automatic Speech Recognition: An Overview, International Journal of Current Engineering and Technology, Vol.4, No.3

[2] Sandeep, S., Anand, S., 2014, A Comperative analysis of LPCC,MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks: An Overview, International Journal of Computer Applications · September 2014

[3] Tingxiao Yang, "The Algorithms of Speech Recognition, Programming and Simulating in MATLAB", *University of Gavale*, pp.1-49, January 2012.

[4] Dr.Yousra F., Al-Irhaim Enaam Ghanem Saeed, "Arabic word recognition using wavelet neural network", *Scientific Conference in Information Technology*, November 2010

[5] Rekha Hibare., Anup Vibhute, Feature Extraction Techniques in Speech Processing A Survey International Journal of Computer Applications (0975 – 8887) Volume 107 – No 5, December 2014

[6] S. B. Dhonde, Amol A. Chaudhari, M. P. Gajare, Performance Evaluation of Mel and Bark Scale based Features for Text-Independent Speaker Identification: International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-11, September 2019

[7] Cabral, F.S.; Pinto, M.; Mouzinho, F.A.; Fukai, H.; Tamura, S. An Automatic Survey System for Paved and Unpaved Road Classification and Road Anomaly Detection using Smartphone Sensor. In Proceedings of the 2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), Singapore, 31 July–2 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 65–70.

[8] Tomyslav Sledevic, Arturas Serackis, Gintautas Tamulevicˇ us, Dalius Navakauskas, International Journal of Electrical,Computer, Electronics and Communication on Evaluation of Features Extraction Algorithms for a Real-Time Isolated Word Recognition System Vol:7 No:12, 2013