# Speaker recognition based on MFCC

Chuyu Zhang ,Mingtao Wang
Xiaobo Li ,Yueyang Hu

***Abstract*— Speaker recognition is performed by using Meier cepstrum coefficients. Record multiple speakers who speak the same voice several times, then use this as a sample, and use dynamic time warping technology to generate speaker comparison template. After a newly recorded voice is obtained, it can be compared with the template to determine which speaker in the sample speaks the voice. Compared with the text-based traversal matching method, the template matching method based on DTW has a good improvement in time and accuracy.**

## I. INTRODUCTION

Speaker recognition, also known as voiceprint recognition, is a technique used by a computer to automatically recognize a speaker's identity by using voice feature parameters contained in the voice that reflect the physiological and behavioral characteristics of a particular speaker.

By using Mel Frequency Cepstrum Coefficient, analysis focuses on the auditory characteristics of the human ear, according to the results of auditory experiments to analyze the speech to get the cepstrum, to obtain a higher recognition accuracy.

Considering that the speech signal length is not necessarily the same, it has greater randomness. The simple method is to make the length of the speech signal same as the length of the reference template, but the accuracy is not up to the requirement. Therefore, DTW is proposed, which is a nonlinear normalization technique that combines the time and distance measurement calculation.

The typical characteristic of speech signal is the length of pronunciation is different, which has big randomness. Even if the same person sends the same tone at different times, it is impossible to have the same length of time. Moreover, the pronunciation speeds of different phonemes in the same word are different. For example, some people will drag the ”A” sound very long, or the ”I” will be very short. In these complex cases, the distance between two times series that cant be effectively solved using traditional Euclidean distances.

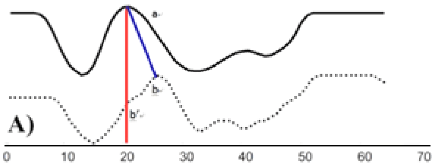

Fig. 1. traditional match

As Figure 1 shows, the solid line and the dotted line are the two speech waveforms of the same word ”open”. Their overall waveform shapes are similar, but they are not aligned on the timeline. For example, at the 20th time point, the point a of the solid line waveform corresponds to the ”$b'$” point of the dotted waveform, so that the conventional comparison of the distance to calculate the similarity is obviously unreliable.

Therefore, DTW is proposed, which is a nonlinear normalization technique that combines the time and distance measurement calculation. In Figure2, DTW can find the points where the two waveforms are aligned, so that calculating their distance is correct.

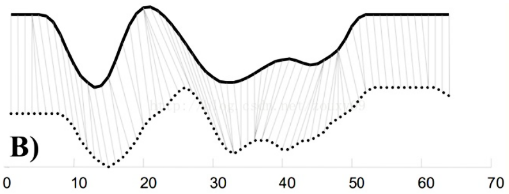

Fig. 2. DTW match

## II. METHODS

### A. MFCC

The human ear has different perceptual abilities to speech at different frequencies. It is found that the perception ability is linear with frequency below $1000Hz$, and the perception ability is logarithmically related to frequency above $1000Hz$. In order to simulate the perception characteristics of the human ear to different frequencies, the concept of Mel frequency is proposed. The meaning is: 1Mel is $1/1000$ of the degree of pitch perception of $1000Hz$, and the conversion formula between frequency $f$ and Mel frequency B is:

$$Mel(f) = 2595\lg(1 + f_{Hz}/700) \tag{1}$$

Where $f$ is the frequency and the unit is $Hz$.

The Mel Frequency Cepstral Coefficient (MFCC) is proposed based on the above-mentioned Mel frequency concept. The proposed process and calculation process are shown in Fig.3.

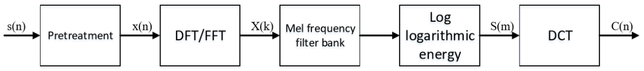

Fig. 3. the proposed process of MFCC

*1) :* The original signal s(n) is pre-emphasized, framed, windowed to the time domain signal x(n) of each speech frame.

*2) :* After the time domain signal x(n) is complemented by a number of 0, a sequence of length N (N should be an integer power of 2, where N is 256) is formed, and then a linear spectrum X (k) is obtained by Discrete Fourier Transform, the conversion formula is:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N}(0 \leq n, k \leq N-1) \qquad (2)$$

In practical applications, it is often calculated by a Fast Fourier Transfor, where N is generally referred to as the window length of the FFT.

*3) :* The linear spectrum X(k) is obtained by a Mel frequency filter bank to obtain a Mel frequency, and a logarithmic spectrum S(m) is obtained by processing a logarithmic energy, where in the Mel frequency filter bank is set in a frequency spectrum range of the speech. A number of bandpass filters Hm(k), 0m¡M, M is the number of filters, each filter has a triangular filter characteristic, and its center frequency is f(m), when the m value is small The spacing between f(m) is also small. As m increases, the interval between adjacent f(m) also increases.

*4) :* Calculate the logarithmic energy of each filter bank output:

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right), (0 \leq m < M) \qquad (3)$$

*5) :* The above-mentioned logarithmic spectrum S(m) is subjected to discrete cosine transform DCT to the cepstrum domain to obtain the Mel frequency cepstral coefficient C(n):

$$c(n) = \sum_{m=0}^{M-1} S(m)\cos\left(\frac{\pi n(m+1/2)}{M}\right), (0 \leq m < M) \qquad (4)$$

In the actual application process, the number of Mel filter banks is taken as 24, and the first 13 coefficients are selected.

### *B. Dynamic Time Warping*

Suppose the template for speech is Q and reference template is C. The length of Q is N, while the length of C is M. The task is to find a time warping function m=w(n) to match Q to C.

In order to align the two sequences, we need to construct a matrix grid of $m * n$,the matrix elements (i, j) represent the distance $d(q_i, c_j)$ between the two points$q_i$and$c_j$. For the similarity between each point, the smaller the distance, the higher the similarity. Each matrix element (i, j) represents the alignment of points $d(q_i, c_j)$ The DP algorithm can be attributed to finding a path through several grid points in the grid. The grid point through which the path passes is the aligned point for the two sequences to be calculated.

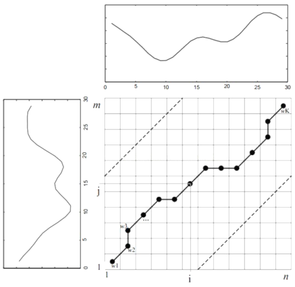

Fig. 4. an example wraping path

DP is used to gradually divide the problem of a total decision into several sub-problems. DTW is the representative of this technology. Its goal is to find the minimum cumulative distance, so we firstly find the correspondence between the speech vector and the reference vector to calculate the minimum value of the cumulative distance when matching two templates. We can use this minimum to get the regular function $m = w(n).m = w(n)$ needs to meet the following conditions:

$$w(1) = 1, w(N) = M \qquad (5)$$

Continuous condition:

$$w(n+1) - w(n) = \begin{cases} 0, 1, 2, w(n) \neq w(n-1) \\ 1, 2, w(n) = w(n-1) \end{cases} \qquad (6)$$

From another point to understand it: starting point(1,1) reaches destination(N,M) with the shortest path.

The minimum cumulative distance:

$$D(N, M) = \min_{w(n)} \sum_{n=1}^{N} d[n, w(n)] \qquad (7)$$

In this formula, $d[n, w(n)]$ represents the distance between the speech signal of the kth frame and the mth frame reference template. D is the distance of the two templates in the case where the optimization time is corrected.

This solution process is a reverse process, details as follows:

*1) :* Get $D(n, m)$ at the beginning, because it is smallest among $d(N, M)$,$D(N-1, M)$, $D(N-1, M-1)$ and $D(N-1, M-2)$.

*2) :* According to (5), $D(N-1, M)$ is not from $D(N-2, M)$; Then check $D(N-1, M)$,$D(N-1, M-1)$,$D(N-1, M-2)$. Work out $d(N-1, m)$, m is the value of $n = N-1$.

*3) :* From backwards to forwards, we can get the following recursion formula:

$$\begin{aligned} &D(n+1, m) = d[n+1, m] \\ &+ \min[D(n, m)g(n, m), D(n, m-1), D(n, m-2)] \end{aligned} \qquad (8)$$

$$g(n, m) = \begin{cases} 1, w(n) \neq w(n-1) \\ \infty, w(n) = w(n-1) \end{cases} \qquad (9)$$

In summary, Working out $D(n+1,m)$ needs to be calculated three times. If we get the best match length between speech template and reference template, speech analysis can be greatly simplified.

## III. Comparison and Analysis

### A. Dataset description

We recorded four people for a total of 75 speeches at two different place. And then, we split the total speeches to train set and test set.Train set include 32 speeches and Test set include 43 speeches.The detail is below:

| | hyy | lxb | wmt | zcy |
|---|---|---|---|---|
| Train set | 8 | 8 | 8 | 8 |
| Test set | 8 | 12 | 12 | 11 |

Fig. 5. hyy,lxb,wmt and zcy is four people

### B. Result analysis

We generate four templates from train set. And we calculate the distance between speech in test set and templates.Then, we classify speech to the nearest template.

We find two implementation of DTW, fastdtw and accelerated dtw.Below is the results of different dtw.

| DTW implementation | Time(s) | Acc(%) |
|---|---|---|
| accelerated dtw | 92 | 83.72 |
| fastdtw | 71 | 60.47 |

Fig. 6. Time and accuracy of fastdtw and accelerated dtw

Accelerated dtw spend more 21 second than fastdtw, and accuracy is 23.25% higher than fastdtw. We analysis accelerated dtw results.The confusion matrix is figure 7:

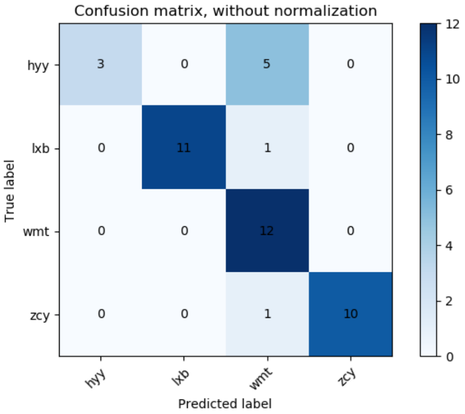

Fig. 7. the confusion matrix

According to confusion matrix, we find that most hyy speech are classified wrong.Whats more, all the wrong speeches are classified to wmt. To find the reason, we calculate the distance between template.

| | hyy | lxb | wmt | zcy |
|---|---|---|---|---|
| hyy | 0 | 0.238 | 0.242 | 0.239 |
| lxb | 0.238 | 0 | 0.190 | 0.203 |
| wmt | 0.242 | 0.190 | 0 | 0.213 |
| zcy | 0.239 | 0.203 | 0.213 | 0 |

Fig. 8. the distance between template

The template of wmt is close to lxb and zcy, so it is understandable that lxb and zcys speech in test set will be classified to wmt. But it is still confused why most hyy speeches are classified to wmt. We calculate the distance between hyy speech in test set and templates

| | hyy | lxb | wmt | zcy |
|---|---|---|---|---|
| hyy | 0.229 | 0.214 | 0.198 | 0.225 |
| hyy | 0.172 | 0.272 | 0.279 | 0.302 |
| hyy | 0.285 | 0.253 | 0.221 | 0.233 |
| hyy | 0.266 | 0.251 | 0.219 | 0.228 |
| hyy | 0.218 | 0.230 | 0.201 | 0.206 |
| hyy | 0.231 | 0.250 | 0.220 | 0.245 |
| hyy | 0.141 | 0.309 | 0.318 | 0.332 |
| hyy | 0.191 | 0.275 | 0.279 | 0.316 |

Fig. 9. the distance between hyy speech in test set and templates

After carefully analysis, we find the distance is below 0.2 when hyy speeches are classified to hyy. This applies to other samples.

So we conclude that there is something wrong in hyy template.When first recorded, hyy caught a cold. The second time, hyy recovered from a cold. If we don't think about hyy, the accuracy is 94.3%.

## IV. FURTURE WORK

### A.

The results of this experiment show that the accuracy of speaker recognition is inversely proportional to the running time of the system. In order to improve the recognition rate and reduce the running time, we will

- optimal algorithm, using DTW algorithm with higher accuracy and faster speed.
- using multi-threading technology to accelerate program execution.
- increase the sample base and build template vectors for each individual.

### B.

The content of this experiment is text-related, that is to say, the subjects must speak the same text as recorded in advance, and the next step is text-independent speaker recognition.

### C.

This experimental environment is quiet and the background noise is very small. In the future, speakers in noisy environment will be tested and identified to observe the impact of noise on the recognition accuracy and consider how to reduce the impact of noise.

# REFERENCE

[1] Dalmiya, C.P., Dharun, V.S., Rajesh, and K.P., ”An efficient method for Tamil speech recognition using MFCC and DTW for mobile applications,” 2013.

[2] Z. B. W. X. L. Z. C. Huisheng, Beijing, 100871, and China), ”On the Importance of Components of the MFCC in speech and speaker recognition,” , 2000, p. 4.

[3] Mohan, B.J., Babu, and N.R., ”Speech recognition using MFCC and DTW,” Advances in Electrical Engineering (ICAEE), 2014 International Conference on, 2014.