Effect of the Number of Codebook and MFCC on Chinese Isolated Words Recognition

Liu De, Wang Mingjiang, Wu Zejun[1]

Abstract. This paper aims to design and optimize the algorithm of Chinese isolated words speech recognition, to improve the recognition accuracy and robustness. Mel-Frequency Cepstral Coefficient (MFCC) has been widely used in the art-of-state speech recognition since it considers the characteristics of human voice and sounds receiving, and has better robustness for the recognition accuracy. The MFCCs of different speech frame are relevant, and each Chinese isolated word has its own optimal number of codebook. In this paper, we use the first-order differential and second-order differential of MFCC to improve the recognition accuracy and robustness, and the results show that the difference between the first recognition probability and the second recognition probability increases. By establishing an optimal codebook for each isolated word, and using codebook adaptive algorithm, the robustness of recognition is also improved greatly.

Keywords: Isolated Words Recognition, Feature Extraction, MFCC Differential, Codebook Number Adaptive

1 Introduction

The research object of speech recognition technology is speech signal processing and making machines understand human natural language, which lets the technolo-

[1] Liu De (✉)
Harbin Institute of Technology Shenzhen Graduate School, 518055, Shenzhen, China
e-mail: liude19832006@126.com

Wang Mingjiang
Key Laboratory of The Technology of The Internet of Things, 518055, Shenzhen, China

Wu Zejun
Harbin Institute of Technology Shenzhen Graduate School, 518055, Shenzhen, China

*Funded by: The Key Laboratory Projects of Shenzhen (CXB201104220018A)

gy become the interface of human-computer interaction. Specifically, speech recognition is defined more technically as the building of system for mapping acoustic signals to a string of words or a sentence. Speech Recognition has many aspects [1][2], such as isolated words recognition, continuous speech recognition, speaker recognition, etc. In general, all aspects of speech recognition aim to translate the input speech signals to an instruction which can be executed by computers or machines. The process of speech recognition is illustrated in **Fig.1**.

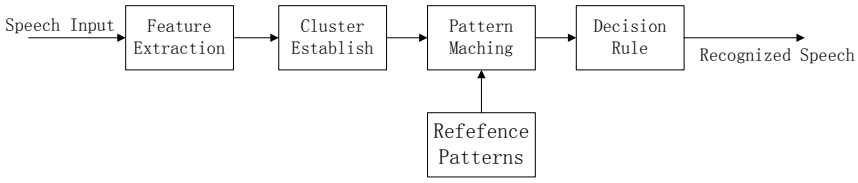

Figure 1. Speech recognition process

Of all these technologies, isolated words recognition is more mature than others, for its computational complexity is relatively simple and it needs less storage space, which is easy to implement in hardware.

The human ear can hear the voice signal from the environment with a noisy background, this is because the inner ear can regulate the external signals. At different frequencies and within the critical bandwidth, the external signals can cause different parts of inner ear membrane to vibrate. Therefore band-pass filter banks can be used to imitate the human auditory system, such that the interference of noise on voice can be reduced. The basic problem of speech recognition is choosing suitable feature parameters. At present, the mostly used feature parameters are Linear Predicted Cepstrum Coefficient (LPCC) [1][3] and Mel-Frequency Cepstrum Coefficient (MFCC) [1] which are both based on the human hearing model and human vocal tract model. But at low frequencies, MFCC parameter has high spectral resolution, which makes noise robustness better than LPCC. This is why MFCC parameter is more suitable for speech recognition and widely used.

The amount of feature parameters is very large. If the data of feature parameters are directly stored, the required storage space is increased in magnitude of multiplication, with the increasing of the number of words to be recognized. Assume there are 50 useful frames of each word, 14 parameters are used and each parameter need 2 bytes to store, then for a system of 32 isolated words, totally 32*50*14=22400 bytes are needed. This is a disaster for ASIC chip which has smaller memories. At the same time, the computation of such large amount of data is also tedious. In speech recognition, a key step is to establish a codebook to reduce the storage space. When subsequently compute the matching patterns, which is well known as Hidden Markov Model (HMM), the codebook is also needed.

Based on the above point, this work optimized the number of codebook, used MFCC and high order differential of MFCC as feature parameters to increase the recognition rate and improve the robustness.

2 MFCC Feature Extraction and Codebook Establishment

2.1 MFCC Feature Extraction

The process of MFCC feature extraction is illustrated in **Fig.2**. The detailed steps are described as following.

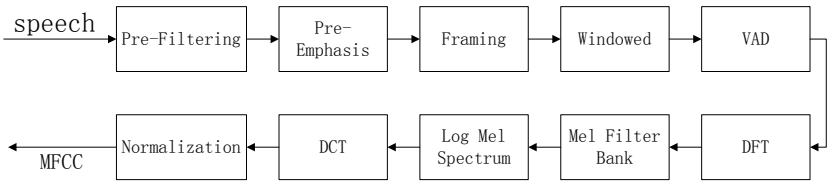

Figure 2. The process of MFCC feature extraction

(a) Firstly, use the finite impulse response filter (FIR) to pre-filter a Chinese isolated word speech to remove the high frequency noise. The FIR is represented as (1).

$$H_h(z) = \frac{0.989502 - 1.979004z^{-1} + 0.989502z^{-2}}{1 - 1.978882z^{-1} + 0.9799126z^{-2}} \quad (1)$$

The length of each Chinese isolated word is not exceeded 4 seconds, and the speech signal is sampled at the rate of 16KHz.

(b) Pre-emphasis of the high-frequency portion of voice can increase the high frequency resolution of the voice by a transfer function which is represented as (2). In fact, the transfer function is also a filter. In (2), α is the pre-emphasis coefficient, $0.9 < \alpha < 1.0$, here we use 0.95. Suppose at time n the sampled value of speech signal is $s(n)$, after pre-emphasis, the result is $y(n)$:

$$y(n) = s(n) - \alpha s(n-1) \quad (2$$

(c) Based on the short-term steady feature of voice, an isolated word speech can be framed to extract the short-term features so as to facilitate the establishment of the model. Here we take 25ms as the frame length, and the frame shift is 10 ms.

(d) The basic means of short-term analysis is to window each frame of the speech signal. That means using a sequence $w(n)$ with finite length to intercept the frame of speech signal to be analyzed. Since Hamming window has smooth low-pass characteristics and the minimum sidelobe height, by multiplying each frame with Hamming window, the discontinuity of signal at the beginning and end of the frame can be reduced. Hamming Window is showed in (3).

$$w(n) = 0.54 - 0.46\cos\left(\left(\frac{2\pi n}{N-1}\right)\right) \quad (0 \le n \le N-1) \tag{3}$$

In (3), N is the sampling number of one frame, here is 16000*0.025=400.

(e) Some frames of speech signals are useless, they are just non-speech frames, so effective voice detection is needed to determine the endpoint of speech signals, which is also called Voice Activity Detection (VAD) [6]. This paper uses the voice energy threshold method to detect the start and end point of useful frames. When the energy (square of the amplitude) of speech signals in one frame is greater than a threshold, it is considered that the frame is the start point of this word speech. When the energy of speech signals in one frame is less than a threshold value, it is considered that the frame is the end point.

(f) Because Mel cepstral coefficient is based on frequency domain, after VAD, the useful frames should be transformed to frequency domain. By the processing of Discrete Fourier Transform (DFT), we obtain the linear spectrum $X(k)$ of each useful frame. This paper uses N=512 points FFT processing.

(g) After processing the speech signals by Fourier transform, use Mel filter banks to process $X(k)$. The Mel filter banks are showed in (4):

$$H_m(k) = \begin{cases} 0 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (k < f(m-1)) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} \cdots\cdots\cdots\cdots (f(m-1) \le k \le f(m)) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} \cdots\cdots\cdots\cdots (f(m) < k \le f(m+1)) \\ 0 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (k > f(m+1)) \end{cases} \quad (0 \le k < M) \tag{4}$$

$f(m)$ is defined by (5)

$$f(m) = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \tag{5}$$

In (5), f_l-------the lowest frequency of the filter frequency range

f_h-------the highest frequency of the filter frequency range

N-------the width of DFT window

F_s------sampling rate

B^{-1}-----the inverse function of B

$$B^{-1}(b) = 700(e^{b/1125} - 1) \tag{6}$$

(h) By (7), we obtain the logarithmic energy $S(m)$ of spectrum $X(k)$, considering the Mel filter bank.

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right) \quad (0 \le m < M) \tag{7}$$

(i) The logarithmic spectrum $S(m)$ can be transformed to cepstrum domain by using discrete cosine transform (DCT). After the transformation, we obtain MFCC. The DCT process is showed in (8).

$$c(n) = \sum_{m=1}^{M-1} S(m) \cos\left(\frac{\pi n(m+1/2)}{M}\right) \qquad (0 \le m < M) \tag{8}$$

(j) Finally, in order to eliminate the impact of signal channel, the MFCC must be normalized. By suing (9), we obtain the final normalized MFCC parameters.

$$\hat{c}(n) = \frac{c(n) - \bar{c}(n)}{\sigma(n)} \tag{9}$$

In (9), $\hat{c}(n)$ is the normalized MFCC parameter, $\bar{c}(n)$ is the average value of MFCC, $\sigma(n)$ is the variance of MFCC.

The feature extraction process usually contains such a not very accurate assumption that the MFCCs of different frames are not relevant. But in fact, due to the limitation of physical pronunciation, different inter-frame voice must be relevant, the change is continuous. So we can use first-order differential and second-order differential of MFCCs which belong to different frames to approximately describe the inter-frame correlation of speech signals. The cepstrum feature of speech signals is normally referred to as the static characteristics of voice, and the differential of cepstrums is called dynamic characteristic. The static characteristics and dynamic characteristics are complementary, and can improve the recognition performance to a large extent.

High order MFCC parameters used in this paper is the differential of MFCCs in the front and rear frames. See (10) and (11)

$$\nabla c_n(m) = \frac{1}{T}\left(\sum_{t=-S}^{S} t \times c_{n-t}(m)\right) \tag{10}$$

$$\nabla^2 c_n(m) = \frac{1}{T}\left(\sum_{t=-S}^{S} t \times \nabla c_{n-t}(m)\right) \tag{11}$$

In (10) and (11), m represents the mth frame, n represents the nth MFCC of each frame. In this paper, S=2, n can be 13 or 14.

2.2 Codebook Establishment

In speech recognition systems, we need to process the MFCCs in some ways. Vector quantization [4][5] method is extremely important to establish a codebook, and also this method is the most basic way. For example, an N-dimensional vector corresponds to a point in the coordinate space, and the set of all N-dimensional vectors forms the coordinate space. In according to certain rules, the vector space can be divided into several sub spaces, e.g. M spaces, and the M spaces meet the following conditions, see (12) and (13)

$$\bigcup_{k=1}^{M} C_k = C \tag{12}$$

$$C_i \cap C_j = \varphi \qquad \forall i \neq j \tag{13}$$

In (12), C represents the whole vector space, each subspace is referred to as Voronoi (cell), and each voronoi has its own typical vector. The typical vector of subspace C_i is marked as Z_i, and Z_i is a Code. The set of all code is called a Codebook, M is the number of codebook.

This paper uses simulated annealing algorithm to establish the codebook [7]. With different word, the corresponding number of codebook should not be the same. It is very apparent, for words with different length, their implicit numbers of state are different, which means that the number of voronois included in each words is different. Secondly, for some words, even if their numbers of codebook are equal, the codebooks should differ with each other. Based on the above two points, this paper established an independent codebook for each word and then optimized the number of codebook for each word. For each word, we increased the number of codebook from 16 to 32 with step size of 2 to find the optimal number of codebook whenever the recognition accuracy is the best.

3 Experimental Results and Discussion

This paper uses 32 words for training to establish the matching patterns (HMM model). But each word has 40 speech signal files—20 male and 20 female reading the same words. So there are totally 1280 speech files are used to extract the MFCCs, establish codebook and model the matching patterns (HMM). To verify the recognition rate and check the robustness, 12 words are used. Each test word is read by 6 male and 6 female, so the test set has totally 384 speech files.

The experiment results show that the recognition rate is 100% when only the MFCC is used. But for some monosyllabic words, the robustness of recognition is not high enough. The key indicator of evaluating the recognition robustness is the difference (denoted as $\triangle P=P_1-P_2$) between the first recognition probability (P_1) and the second recognition probability (P_2) of each word. The greater the difference, the stronger the recognition anti-interference when there are other factors, such as noises or the pronunciation may be not clear. Conversely, if the difference ($\triangle P$) is relatively small and there are interference, the difference may become negative. That is to say, the second recognition probability becomes the first recognition probability, the recognition result is wrong.

When high order MFCC differentials are also used as feature parameters, the recognition robustness of most words are improved, as illustrated in **Table.1** and **Fig.3**.

Table 1. The differential between 1st recognition probability and 2nd recognition probability

Word No. Difference between 1st P and 2nd P	**1,**	**2,**	**3,**	**4,**	**5,**	**6,**	**7,**	**8,**	**9,**	**10,**	**11,**	**12,**	**13,**	**14,**	**15,**	**16**
MFCC	21	22	27	34	34	53	38	46	42	49	45	43	52	42	38	66
MFCC+▽	24	23	35	43	35	57	45	62	59	54	55	50	55	54	52	69
MFCC+▽+▽^{2}	25	32	41	43	37	61	46	57	63	50	48	51	53	58	44	66
Word No. Difference between 1st P and 2nd P	**17,**	**18,**	**19,**	**20,**	**21,**	**22,**	**23,**	**24,**	**25,**	**26,**	**27,**	**28,**	**29,**	**30,**	**31,**	**32**
MFCC	60	89	51	71	48	54	55	54	89	55	40	79	72	44	65	91
MFCC+▽	64	89	54	78	51	59	60	64	84	76	44	84	64	40	81	90
MFCC+▽+▽^{2}	62	78	52	79	56	55	63	52	92	74	49	73	58	45	73	92

In Table.1, MFCC means only use MFCCs as feature parameters, MFCC+▽ means using MFCCs and first order MFCC differential, MFCC+▽+▽^{2} means using MFCC, first order MFCC differentials and second order MFCC differentials. In Table.1, by comparing the number in the same column, it can be seen that the difference between the first recognition probability and the second recognition probability increases when ▽MFCC is used. As can be seen in Fig.1, the recognition robustness is improved when high order MFCC differentials are used.

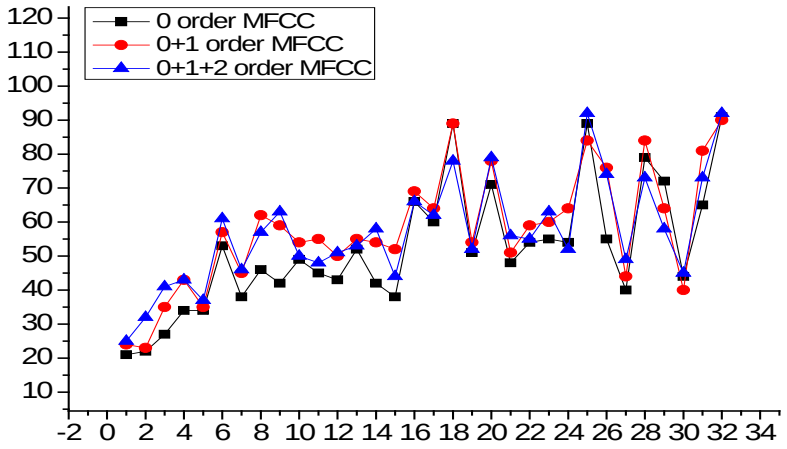

Figure 3. △*P* of ▽MFCC, ▽^{2}MFCC

The experiment results show that the recognition rate is 96.875% when all the 32 words share a common codebook. The recognition rate is 100% when each word has its own codebook. **Fig.4** shows the difference (△*P*) between the first recognition probability (P_1) and the second recognition probability (P_2) of each word. The horizontal axis represents the serial number of each word, the vertical axis represents the value of △*P*, and the curve with different color represents the △*P* when recognized with different number of codebook. For example, the black curve

with rectangle point on it means $\triangle P$ is computed on the condition that the number of codebook is 16.

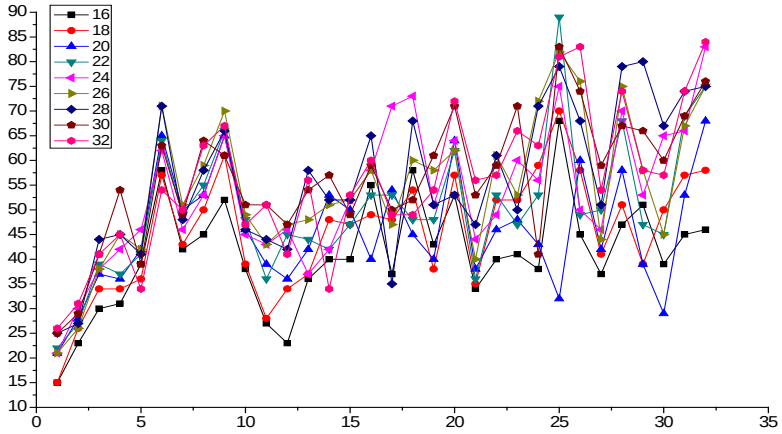

Figure 4. The difference between 1st Prob and 2nd Prob

From Fig.4, we can see that the recognition robustness of different words varies with the increasing of the number of codebook. The $\triangle P$ of each word does not simply become greater as the number of codebook increases, which means the recognition robustness is not always improved when the number of codebook increases. For example, the recognition robustness of the sixth word is the best when its codebook number equals to 28; for the seventeen and eighteen word, the best codebook number is 24. Accordingly, for various words, finding the most suitable number of codebook can greatly improve the recognition robustness.

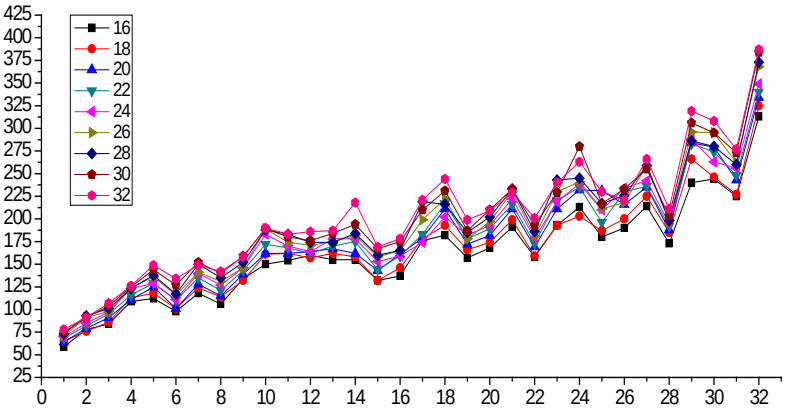

Figure 5. The means of first probability

With different number of codebook, the average value ($\overline{P_1}$) of the first recognition probability is illustrated in **Fig.5**, the variance of the first recognition probability is illustrated in **Fig.6**.

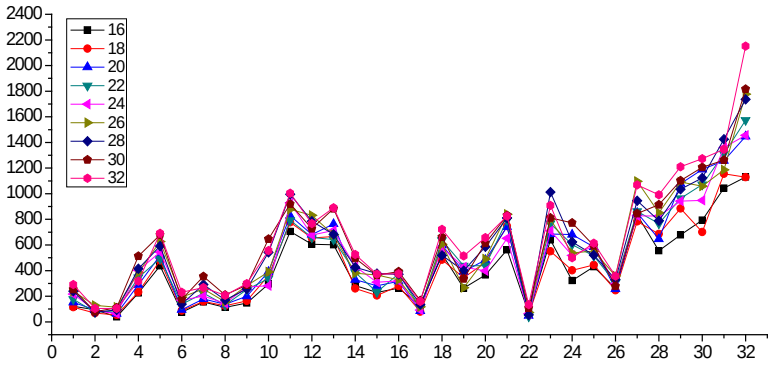

Figure 6. The variance of first Probability

Fig.5 shows that the first recognition probability of each word generally increases when the number of codebook increases. Fig.6 shows that the variance of the first recognition probability also increases as the number of codebook increases. When the number of codebook increases, each frame can be better classified into the appropriate category and the subsequent matching patterns can be created more precisely. The increasing of variance does not mean that the recognition stability getting worse, but means that the first recognition probability getting less stable. The key standard to evaluate whether the recognition robustness is good or not is the difference between the first recognition probability and the second recognition probability.

4 Conclusions

In order to increase the Chinese isolated words speech recognition rate and improve the recognition robustness, this paper proposes a method to create a codebook for each isolated word. At the same time, let each word have its own optimal number of codebook. At last, this paper made a contrast of recognition rate and robustness in the 3 following conditions: (1) simply MFCC, (2) MFCC and first order differential of MFCC, (3) MFCC, first order differential of MFCC and second order differential of MFCC.

The experiments show the following results:

(1) The recognition rate is 100% in all three conditions; by introducing high order MFCC differential, the difference between first recognition probability and the second recognition probability gets larger, so the recognition robustness improved.

(2) By establishing a separate codebook for each isolated word, the recognition rate has been improved from 96.875% to 100. With the foregoing method, by changing the number of the codebook, it is found that different word has its own optimal number of codebook. When the number of codebook continue to increase, especially exceeding a certain value, the robustness of some word even gets worse.

References

1. Antonio M. Peinado, Jose C. Segura, *Speech Recognition Over Digital Channels—Robustness and Standards*, John Wiley & Sons, Ltd, 2006.
2. John Holmes, Wendy Homes, S*peech Synthesis and Recognition*, 2nd ed., London and New York, 2001.
3. Assaleh K T, Mammone R J. "New LP-derived features for speaker identification". IEEE Trans. on Speech and Audio Proc., 1994, 2(4): 630-637.
4. Burton, D., Shore, J., Buck, J., "Isolated word speech recognition using multsecton vector quantization codebooks", IEEE Trans. on Acoustics, Speech and Signal Processing, 33(4): 837-849, 1985.
5. T. Pham, M. Brandl, D. Beck, "Fuzzy declustering-based vector quantization", Pattern Recognition 42(11): 2570-2577, 2009.
6. E. A. Escoto-Sotelo, E.Escamilla-Hernandez, E. Gracia-Rios, H. M. Perez-Meana, "End-pint Detector Algorithm for Speech Recognition Application", Instituto Politecnico Nacional SEPI ESIME Culhuacan.
7. Fatma zohra. Chelali, Amar. DJERADI, "MFCC and vector quantization for Arabic fricatives Speech/Speaker recogniton", Multimedia Computing and Systems, 2012: 284-289