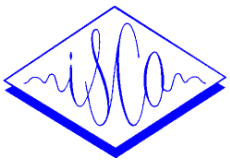

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# On the Importance of Components of the MFCC in speech and speaker recognition*

*ZHEN Bin, WU Xihong, LIU Zhimin, CHI Huisheng*

*(Center for Information Science, Peking University, Beijing, 100871, China)*

**Abstract**: In this paper, we analyzed the relative importance of components of MFCC for both speech recognition and speaker recognition using DTW recognizer in various noise environments. For English digit and under the Euclidean distance definition, the experiment results show cepstral components from $C_2$ to $C_{16}$ contain the most useful speaker information, while $C_0$ and $C_1$ are usually harm to speaker recognition. Cepstral terms from $C_1$ to $C_{12}$ are found to contain the most useful speech information. In both tasks, the additive noise decreases the relative importance of low MFCC terms faster than that of the middle and high MFCC terms, and the decrement depends on the speech SNR. The channel distortion will deteriorate low terms more than the middle and high MFCC terms in both tasks, also.

**Keywords:** MFCC, speech recognition, speaker recognition

## 1 Introduction

Automatic speech recognition and speaker recognition typically use features based on a short-term spectrum of speech which describe time-varying speech signal as a sequence of short-term feature vector.[1-3] Usually, the logarithmic or other kind of compressed spectral vectors have to be projected on the cosine basis to make the components independent each other in order to reduce the feature dimension before enter into the classifier, because more features will impose severe requirements on computation and storage in both training and testing.[1-3,5] MFCC, for example, is a dominant speech analysis feature in both speech and speaker recognition.[1-4] In almost all the recognition system, the zeroth and higher order cepstral coefficients are discarded in order to have better performance and less computation[1,2]. Juang etc. proposed a bandpass liftering method and gained better results in speech recognition.[4,8] He de-emphasized the low and high terms of the cepstral components and emphasized the middle term, because different transmission channels of speech generally affect more on the low cepstral terms and the variances of the high cepstral terms are relatively large. Recently, Zhen etc. proposed a new bandpass liftering for speaker recognition, where the low cepstral terms were de-emphasized, and both the middle and high terms were emphasized because the fine structures of spectrum are important to speaker recognition.[5] Not only in noise environment, but also in clean speech, the bandpass liftering technique increases the recognition performance.[4,5]

Why are the zeroth and higher order cepstral coefficients discarded in recognition? Should any other components be discarded? Why are there different liftering methods in speaker and speech recognition? It is therefore of necessary to determinate the relative importance of the cepstral components in both speech and speaker recognition.

Generally speaking, there are two methods to evaluate the relative importance of the cepstral components, one is using F-ratio of each component and the other is examining the influence of the cepstral components on recognition rate.[1,2] In this paper we provide an analysis on the relative importance of components of MFCC for both speech recognition and speaker recognition using DTW recognizer. The paper is organized as follows. Section 2 describes the analysis method we used. Section 3 introduces the experiment setup. Section 4 gives the experimental results and discussion. At last conclusions are given.

## 2 Method

The relative importance of cepstral terms can be simply determined through calculating the difference of DTW measurements with and without the term. For example, the average improvement $R(i)$ resulting from inclusion of the components $C_i$ is estimated as the average of the partial differential of recognition rate.[8,9] That is

$$R(i) = \frac{1}{n}\left(\sum_{j>i} p(i,j) - p(i,j-1) + \sum_{j<i} p(j,i) - p(j-1,i)\right) \quad (1)$$

where $n$ is the order of cepstrum and $p(i, j)$ is recognition rate using cepstrum components from $i$ to $j$. Fig. 1 depicts the grid generated by components $C_0$, $C_1$, $C_2$ and $C_3$.

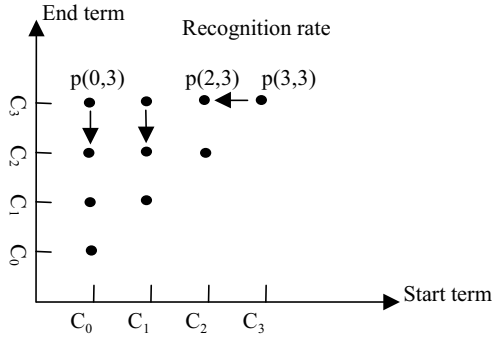


Fig. 1 Grid for evaluating the average relative importance of the MFCC components

* This work was supported by the key projects of National Nature Science Funds of China grants 69635052 and postdoctoral foundation.

A positive value for the average relative importance *R(i)* reflects a relative reduction in recognition error due to the inclusion into the acoustic features of the components in most case, and *vice versa*. It should be note that the message *R(i)* used here only provides an indication of average importance of a cepstral component and as such does not provide information on the inter-dependence of different cepstral component.

## 3 Experiment setup

The speech database used in the experiments was the ten isolated digits of standard speech database TI46. Each digit was spoken by sixteen speakers (eight females and eight males). The data of these digits were divided into two sets (training and testing). In training set, each digit was repeated 10 times by each speaker in one session. In the testing set, each digit was repeated 16 times by each speaker in eight different sessions. In each session, two utterances were recorded.

The ten utterances from the training set were used for training and the sixteen utterances from the testing set were used for testing. The two kinds of degraded speech are shown in Fig. 2, which were referred as additive degraded speech and additive-filtered degraded speech in the paper, respectively. The 3dB passband of the channel filter is 300Hz-3300Hz as shown in Fig. 3, and the additive noise is zero-mean white Gaussian noise. The recognition experiments were performed under the conditions of different SNR for two kinds of degraded speech. The SNR of a test utterance is defined as the power of the speech signal to the noise power.

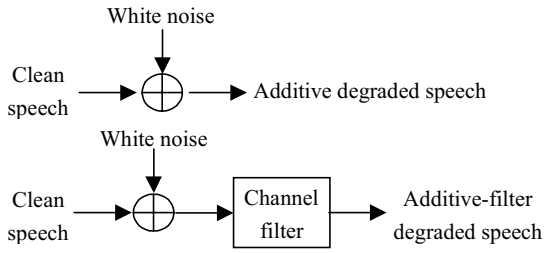


Fig. 2 The speech degraded process

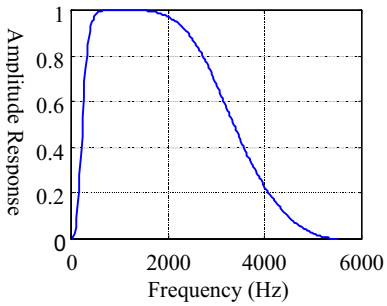


Fig. 3 Frequency response of channel distortion filter

Acoustic feature for recognition experiment was derived from a short-time analysis of the speech. MFCC was used as feature vector. The MFCC are derived directly from the FFT power spectrum after pre-emphasis of speech. The power spectrum are weighted by a triangle filter shape and then summed. The filters have a half-bandwidth of 100Hz up to center frequency 1KHz and a bandwidth of 1.149 times the center frequency above 1KHz. A Discrete Cosine Transform (DCT) converts the spectral estimation obtained from the logarithmic energy across filters into a final cepstral vectors.

As we mainly interested in the relative importance in performance change with the components of the cepstrum, the recognition was done using a simple-but-efficient DTW-based recognizer. During training phase, we obtained one template for each digit from training sets. Thus, we performed speaker-dependent speech recognition and text-dependent speaker recognition.

## 4 Recognition results and discussion

### 4.1 Results on clean speech

Table 1 shows the recognition results for different MFCC components in speech recognition with clean speech. The recognition rate is the average recognition rate of 16 people for digit from 0 to 9. The row is the start MFCC term and the column is the end MFCC term. The $C_0$ term specifies the log-spectrum average, the $C_1$ term approximates the spectrum tilt, etc., and the high cepstral terms represent quickly varying ripples across the log-spectrum. Fig. 4 is the average contribution of different components of MFCC for speech recognition calculated from Eq. 1. Here each bar indicates average improvement in speech recognition accuracy resulting from inclusion into the MFCC component. The most useful speech information is contained between $C_1$ and $C_{12}$, and the other components do not contain much useful information as far as the DTW classifier is concerned. Among the useful components, the contribution of components from $C_3$ to $C_9$ is larger than that of the other's.

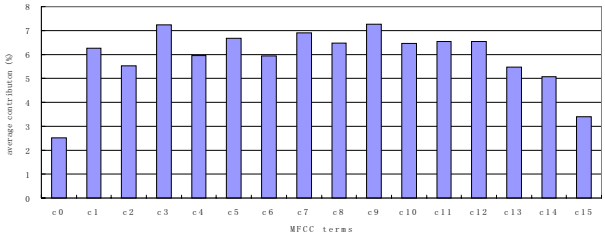


Fig. 4 Improvement of recognition accuracy by including each MFCC component in speech recognition with clean speech

Table 1. The average speech recognition ratio (%)

| | $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ | $C_{10}$ | $C_{11}$ | $C_{12}$ | $C_{13}$ | $C_{14}$ | $C_{15}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_0$ | 60.09 | 97.49 | 93.35 | 96.35 | 96.93 | 97.25 | 97.56 | 97.68 | 97.76 | 97.91 | 98.11 | 98.03 | 98.03 | 98.03 | 98.11 | 98.11 |
| $C_1$ | | 84.27 | 96.54 | 98.71 | 98.9 | 99.18 | 99.18 | 99.22 | 99.29 | 99.22 | 99.25 | 99.33 | 99.37 | 99.37 | 99.33 | 99.37 |
| $C_2$ | | | 86.75 | 98.82 | 99.45 | 99.65 | 99.61 | 99.65 | 99.76 | 99.84 | 99.84 | 99.84 | 99.80 | 99.80 | 99.84 | 99.88 |
| $C_3$ | | | | 88.54 | 97.6 | 98.82 | 99.02 | 99.37 | 99.41 | 99.65 | 99.57 | 99.76 | 99.27 | 99.76 | 99.65 | 99.65 |
| $C_4$ | | | | | 75.81 | 92.17 | 95.91 | 97.48 | 98.23 | 98.74 | 98.90 | 99.10 | 99.25 | 99.14 | 99.02 | 98.9 |
| $C_5$ | | | | | | 76.82 | 89.42 | 95.09 | 96.43 | 97.21 | 97.92 | 98.08 | 98.23 | 98.31 | 98.31 | 98.35 |
| $C_6$ | | | | | | | 69.76 | 87.7 | 92.89 | 95.45 | 96.23 | 97.21 | 97.49 | 97.45 | 97.45 | 97.53 |
| $C_7$ | | | | | | | | 69.43 | 85.84 | 93.20 | 95.29 | 96.19 | 96.66 | 96.98 | 97.14 | 96.70 |
| $C_8$ | | | | | | | | | 64.44 | 84.87 | 90.01 | 93.99 | 95.33 | 95.56 | 95.41 | 95.41 |
| $C_9$ | | | | | | | | | | 66.48 | 81.84 | 88.72 | 91.44 | 93.37 | 93.76 | 93.99 |
| $C_{10}$ | | | | | | | | | | | 59.74 | 78.24 | 86.71 | 89.07 | 90.96 | 91.51 |
| $C_{11}$ | | | | | | | | | | | | 55.22 | 75.92 | 84.27 | 87.73 | 88.51 |
| $C_{12}$ | | | | | | | | | | | | | 56.50 | 72.33 | 80.17 | 83.95 |
| $C_{13}$ | | | | | | | | | | | | | | 52.91 | 67.91 | 75.37 |
| $C_{14}$ | | | | | | | | | | | | | | | 52.14 | 64.13 |
| $C_{15}$ | | | | | | | | | | | | | | | | 43.58 |

Similarly, the average contribution of different MFCC terms for speaker recognition, shown in Fig. 5, can be obtained from the speaker recognition of different sequent MFCC components using Eq. 1. The recognition rate is the average recognition rate of digit from 0 to 9 for 16 people. The results indicate that most of useful speaker information is contained between $C_2$ and $C_{16}$ and the other terms are not very useful. Among the useful components, the relative importance of components from $C_6$ to $C_{13}$ is larger than that of the others. Surprisingly, the contribution of two components, $C_0$ and $C_1$ are negative, which indicates an average decrease in performance caused by including the two MFCC components. For example, the average recognition with components from $C_0$ to $C_{16}$ is 70.62%, from $C_1$ to $C_{16}$ is 85.6%, but from $C_2$ to $C_{16}$ is 91.38%. Although $C_0$ is discarded, $C_1$ is included in most speaker recognition.[4,6]

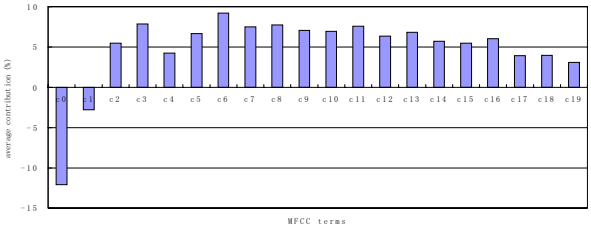


Fig. 5 Improvement of recognition accuracy by including each MFCC component in speaker identification with clean speech

MFCC components from $C_1$ to $C_{12}$ and components from $C_2$ to $C_{16}$ contain the most useful speech and speaker information, respectively. Note that both the start MFCC term and the end MFCC term of the useful components for speaker recognition are higher than those for speech recognition. This indicates that the speech recognition requires only the spectral contour, but the speaker recognition requires the detail information of spectrum, as well as spectral contour.[5,6] Thus, discarding the zeroth cepstral term and the higher terms in both tasks is reasonable and necessary.[1,2,4] The negative value of $C_1$ and $C_0$ indicate that the spectrum tilt, as well as spectrum energy, should be normalized in speaker recognition.

### 4.2 Recognition results in various noisy environments

The relative importance of different MFCC components may change with speech condition. We hence investigate the relative importance of different MFCC components in various noisy environments. Fig. 6 and Fig. 7 show the normalized improvements of recognition accuracy in various noise environments for speech and speaker recognition. To compare the different case, the relative improvements are normalized by the maximum value.

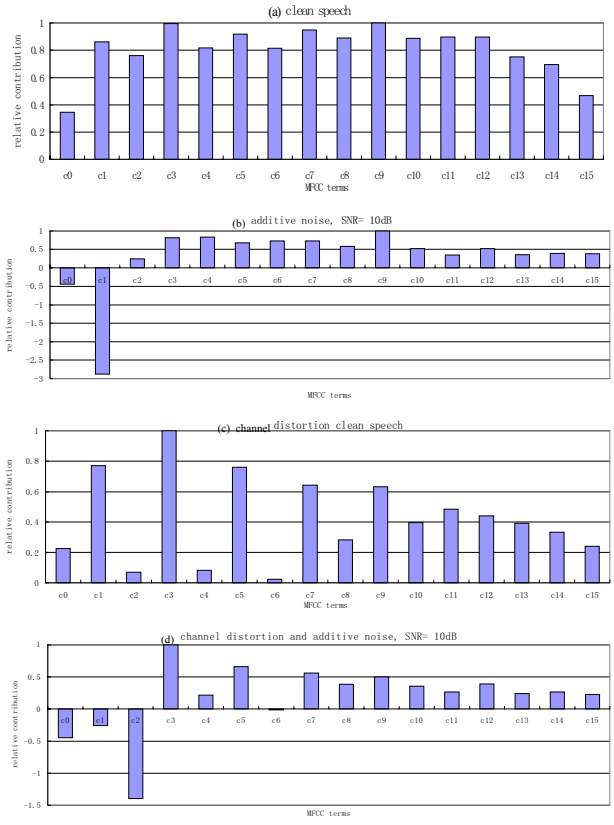


Fig. 6 Relative contribution of MFCC terms for speech recognition. (a) clean speech; (b) additive degraded speech, SNR=10dB; (c) filtered speech; (d) additive-filtered degraded speech, SNR=10dB;

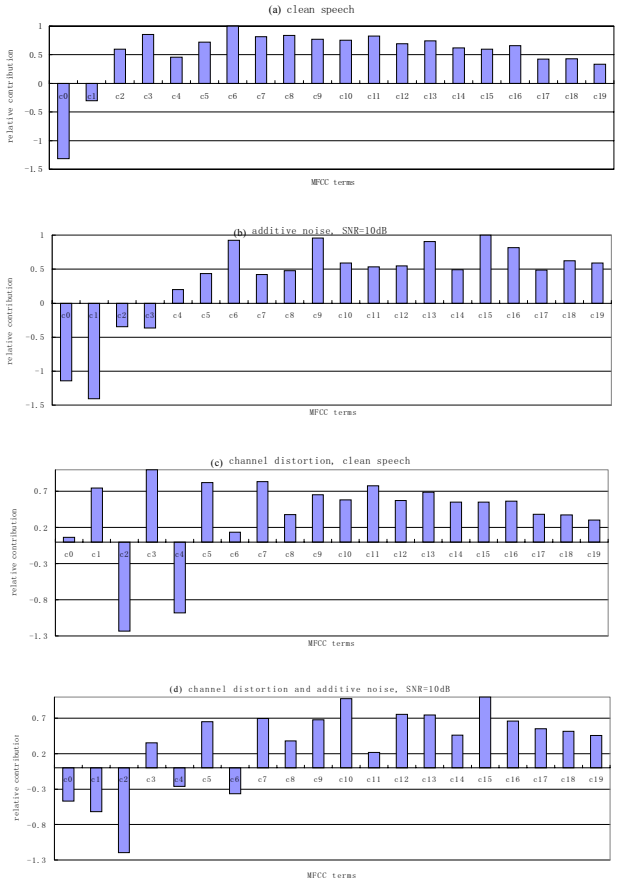


Fig. 7 Relative contribution of MFCC terms for speaker recognition. (a) clean speech; (b) additive degraded speech, SNR=10dB; (c) filtered speech; (d) additive-filtered degraded speech, SNR=10dB;

When the speech is contaminated only by the additive white noise, with the decrease of speech SNR, the relative importance of the low MFCC terms tend to decrease and that of the middle and high MFCC terms remain almost the same. This indicates that the middle and high terms are more robust to additive noise than the low MFCC terms. That is why the bandpass liftering methods increase the recognition performance in both tasks, they emphasize the useful and noise robust terms.[5,6] Although the relative importance varies with the speech SNR, the low cepstral terms until $C_2$ and high cepstral terms from $C_{13}$ are useless in speech recognition, and the low MFCC terms until $C_5$ and high MFCC terms from $C_{17}$ are useless in speaker recognition in whatever condition. That is why the high terms are discarded in both tasks.

When channel distortion is introduced, surprisingly, the relative importance of even terms decrease, but that of the odd terms increase in both the speech and speaker recognition. The reason why this happened is still under investigation. It may be related to the frequency response of the channel distortion filter we introduced. When there is both channel distortion and additive noise, the two effects seem to combine together. The relative importance of low MFCC terms decrease faster than that of the middle and high MFCC terms', and the relative importance of the even MFCC terms decreases while that of the odd MFCC terms' increase again.

## 5 Conclusions

In conclusion, in this paper, the average relative importance of MFCC components is analyzed in both speech and speaker recognition tasks in various noisy environments. For English digit and under Euclidean distance measurement, the MFCC terms from $C_2$ to $C_{16}$ are found to contain the most useful speaker information, while $C_0$ and $C_1$ detract the recognition performance. MFCC terms from $C_1$ to $C_{12}$ are found to contain the most useful speech information. This indicates that the speaker recognition require more detail information of speech spectrum compared with the speech recognition. In both tasks, the additive noise decreases the contribution of low MFCC terms faster than that of the middle and the high terms, and the decrement depends on speech SNR. The channel distortion and will deteriorate low terms more than the middle and high terms in both tasks too. Thus, better performance can be achieved in noisy environment by using the middle and high MFCC components because the low MFCC terms are sensitive to noise. And that is why the bandpass liftering techniques in cepstral domain will increase the performance in both tasks.

## References

1. L. Rabiner and B. H. Juang, "Fundamental of speech recognition", Prentice Hall, 1993
2. H. Chi and X. Yang, "Digital Processing of Speech Signal", Press of Electric Industry, 1995.
3. J. P. Campbell, "Speaker Recognition: a Tutorial", Proc. IEEE, vol.85, No.9, pp. 1437-1462, 1997.
4. D. A. Reynolds, "Experimental evaluation of features for robust speaker identification ", IEEE Trans. Speech and Audio Proc., vol. 2, no. 4, pp. 639-643, 1994.
5. B. H. Juang, L. R. Rabiner and J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition", IEEE Transactions on ASSP, Vol. 35, No.7, pp. 947-953, 1987.
6. B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the Use of Bandpass Liftering in Speaker Recognition", (see proceeding of this conference)
7. H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. Speech and Audio Proc., vol. 2, no. 4, pp. 578-589, 1994
8. B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition", IEEE Trans. Speech and Audio Proc., vol.5, no. 5, pp. 451-464, 1997
9. N. Kanedera, T. Arai. H. Hermansky and M. Pavel, "On the importance of various modulation frequencies for speech recognition", in Proc. of EUROSPEECH'97, Rodos, Greece, 1997
10. A. Van Vuren and H. Hermansky "On the importance of components of the modulation spectrum for speech verification", in ICSLP'98, Sydney, Australia, 1998