Regular paper http://dx.doi.org/10.13067/JKIECS.2015.10.9.993

음성 인식에서 위상 정보의 활용

이창영[*]

Utilization of Phase Information for Speech Recognition

Chang-Young Lee[*]

요 약

MFCC는 음성 신호 처리에서 귀중한 특징 벡터들 중 하나이다. MFCC에서 명백한 결점은 푸리에 변환의 크기를 취함에 의해 위상 정보가 손실된다는 것이다. 이 논문에서 우리는 푸리에 변환의 실수부와 허수부 크기를 따로 취급함으로써 위상 정보를 활용하는 방법을 생각한다. 퍼지 벡터 양자화와 은닉 마코브 모델을 이용한 음성 인식에 이 방법을 적용함으로써, 종전 방법에 비해 음성 인식 오류율을 줄일 수 있음을 보인다. 우리는 또한 수치 해석을 통하여, FFT의 실수부와 허수부 각각에서 6개의 성분을 취하여 모두 12개의 MFCC 성분을 사용하는 것이 음성인식에 최적임을 보인다.

ABSTRACT

Mel-Frequency Cepstral Coefficients(: MFCC) is one of the noble feature vectors for speech signal processing. An evident drawback in MFCC is that the phase information is lost by taking the magnitude of the Fourier transform. In this paper, we consider a method of utilizing the phase information by treating the magnitudes of real and imaginary components of FFT separately. By applying this method to speech recognition with FVQ/HMM, the speech recognition error rate is found to decrease compared to the conventional MFCC. By numerical analysis, we show also that the optimal value of MFCC components is 12 which come from 6 real and imaginary components of FFT each.

키워드

Complex Cpstrum, Phase Information, MFCC, Speech Recognition
복소 켑스트럼, 위상 정보, MFCC, 음성 인식

I. Introduction

As a method of communication between man and machine, speech recognition provides a very effective interface. Speech input to a machine is about twice as fast as information entry by a skilled typist[1].

The state of the art in the field of speech recognition has now reached such a level of performance and robustness, even in noisy environments, that permits lots of daily applications. As a result, we are now living in a world of various devices which deploy the relevant technology[2-5].

However, in practical applications currently, the machine provides several candidates for the recognized word(or phrase or sentence) and the user(human being) selects one of them. In other

* 교신저자 (corresponding author) : Div. of Mechatronics Engineering, Dongseo University (seewhy@dongseo.ac.kr)
접수일자 : 2015. 08. 01 심사(수정)일자 : 2015. 09. 13 게재확정일자 : 2015. 09. 23

words, the man-machine communication proceeds with confirmation for each response. This might be one of the reasons that hinder the wide spread of the speech recognition technology in our daily life. For example, there are seldom found in Internet banking who send money by invoking the speech recognition method. In this sense, the speech recognition technology needs to be much more refined to reach the level of human-human conversation with little confirmation.

In speech processing, the importance of the feature extraction cannot be overemphasized. There are several kinds of parametric representations for acoustic speech signals[6]. Among them, MFCC is currently one of the most popular methods of front-end processing for subsequent speech works such as vocoding, speaker identification, and speech recognition.

In obtaining MFCC, we take the Fourier transform of the speech signal, its magnitude, and cosine transform successively. An apparent drawback of this approach is that the phase information, i.e. the relative nature(including polarity) of the real and imaginary components of the Fourier transform is lost. Phase information plays great roles in many applications[7-8]. In this paper, we study on the method of enhancing the speech recognition performance by utilizing the phase information of the speech signal.

The organization of this paper is as follows. Section II reviews the conventional procedure of MFCC feature extraction. Section III provides several methods of remedying the drawback of this method. Experimental details on the application of the proposed method to speech recognition is given in Section IV. After providing the results and discussion in section V, concluding remarks will be given in section VI.

II. Review of the MFCC Extraction

Speech signal is first pre-emphasized with FIR filter for spectral flattening. This is usually intended to boost the signal spectrum approximately 20 dB per decade. For short-term analysis, the signal is blocked into frames of duration $\sim$10 ms. To reduce edge effects incurred by abrupt frame blocking, Hamming or Hanning window is applied to each frame. After performing FFT on this signal, log-energies in the filter banks are estimated and fed through discrete cosine transform to obtain MFCC. The concrete implementation procedures are as follows.

Step 1: Preprocessing of the speech signal such as Hamming windowing and spectral flattening are performed on the input speech signal. The resultant signal is given by $x(n),\ n=0,1,2,\cdots,N-1$.

N is the frame size which is usually of $\sim$ 10 ms time duration and chosen as a power of two.

Step 2: Spectrum in the frequency domain is obtained by FFT on $x(n)$, the result being

$$X(m)=\sum_{n=0}^{N-1} x(n)\exp\left(-i\frac{m}{N}2n\pi\right), \quad m=0,1,2,\cdots,N/2. \tag{1}$$

It should be noted that, only the components of $m=0\sim N/2$ are meaningful among the array $X(m)$ returned by FFT.

Step 3: The energy content in each Mel window is evaluated:

$$S(k)=\sum_{m=0}^{N/2} W_k(m)\,|X(m)|^2, \quad k=1,2,\cdots,K. \tag{2}$$

K is the number of windows ranging usually from 20 to 24. The windows are arranged according to the Mel-scale[9].

Step 4: MFCC is obtained by cosine transform on the log of the Mel-window energies:

$$C(l) = \sum_{k=1}^{K}\left[\log(|S(k)|)\cos\left\{l(k-0.5)\frac{\pi}{K}\right\}\right], \qquad (3)$$
$$l = 1, 2, \cdots, L.$$

L is the order of MFCC, which is usually taken to be 13 on empirical grounds.

III. Utilization of the Phase Information

Fig. 1 is a portion of a speech phoneme /ah/ pronounced by a young female speaker and Fig. 2 is its FFT.

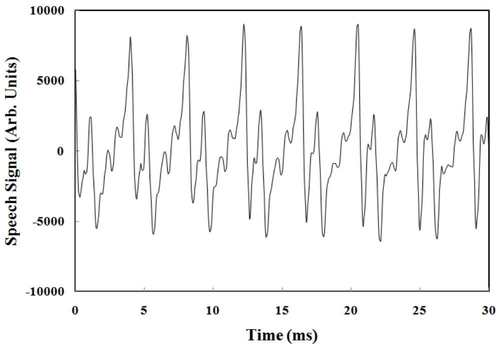

Fig. 1 A portion of a speech phoneme /ah/ pronounced by a young female speaker

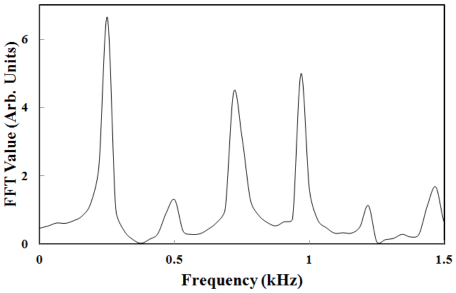

Fig. 2 FFT of Fig. 1

We note that the phase information of FFT is lost by taking the absolute square, in Eq. (2), of the real and imaginary components. This is an evident drawback of FFT. It is worthy of utilizing the phase information embodied in Eq. (1).

An approach for this purpose is to include the phase angle

$$\log[X(\omega)] = \log(|X(\omega)|) + i\angle X(\omega)$$
$$\angle X(\omega) = \arctan\left(\frac{X_I(\omega)}{X_R(\omega)}\right)$$

in between the steps 2 and 3 of section II. However, this method causes problems associated with wrapping and aliasing[10]. Another trouble is that it is not easy to devise a method of treating the phase angle and other parameters on equal footing.

Another method is to use complex MFCC[11]. Fig. 3 shows separately the real and imaginary components of FFT given in Fig. 2. The vertical bars within the graph denote not the graph grids but the frequency positions of the six peaks in Fig. 2.

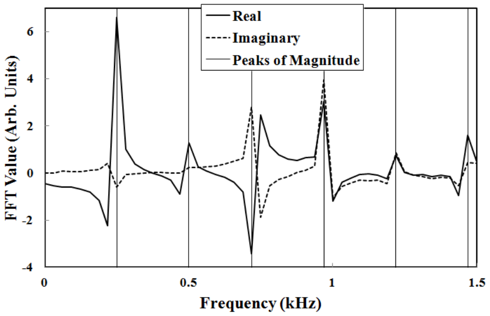

Fig. 3 The real and imaginary components of FFT for Fig. 1

Since we need log energy content in each window in estimation of MFCC, what matters is the absolute value of FFT. If we treat the real and imaginary components separately, we could acquire two kinds of advantages. Firstly, a single peak in magnitude spectrum(as shown in Fig. 2) is generally resolved into double peaks. The left three peaks of Fig. 3 correspond to this case. Secondly, even if the peak positions of real and imaginary components coincide(at the same frequency), the relative magnitudes of them are different in general. This is another sort of information. The 4th and 6th peaks of Fig. 3 correspond to this case.

In order to utilize the phase information based on the idea mentioned above, we modify the procedures of MFCC extraction as follows.

Step 1 and Step 2 are the same as the conventional ones described above. By these procedures, $X(m)$ is obtained. Complex MFCC goes another way from Step 3 in order not to discard phase information contents. The procedures are as follows.

Step 3: Instead of adding the real and imaginary components, we estimate the energy contents separately as follows.

$$Y_R(m) = X_R^2(m), \quad Y_I(m) = X_I^2(m) \tag{4}$$

Step 4: The energy contents in each Mel window for real and imaginary components are evaluated separately:

$$\left.\begin{aligned} S_R(k) &= \sum_{j=0}^{F/2} W_k(j)\, Y_R(j) \\ S_I(k) &= \sum_{j=0}^{F/2} W_k(j)\, Y_I(j) \end{aligned}\right. , \quad k = 1 \sim K. \tag{5}$$

where K is 20~24.

The final procedure for new MFCC extraction is, as usual, to perform the cosine transform of the log energies:

$$\begin{aligned} C_R(n) &= \sum_{k=1}^{M}\left[\log(S_R(k))\cos\left\{n(k-0.5)\frac{\pi}{M}\right\}\right] \\ C_I(n) &= \sum_{k=1}^{M}\left[\log(S_I(k))\cos\left\{n(k-0.5)\frac{\pi}{M}\right\}\right] \\ &\qquad n = 1, 2, \cdots, L \end{aligned} \tag{6}$$

Another motivation for our approach might be inferred from Fig. 4. The abscissa and ordinate represent the order and variance of MFCC respectively. We see that the variance decreases as the MFCC order increases. We might conjecture that it is desirable to get more information from the lower portion of the MFCC order.

In conventional method, the MFCC order L is usually 13. However, in our new method, a search for the optimal value for it will be performed. Otherwise, doubling of L due to inclusion of real and imaginary components might incur somewhat heavy computational load. In this paper, therefore, we vary the value of L and investigate its effect on the performance of speech recognition.

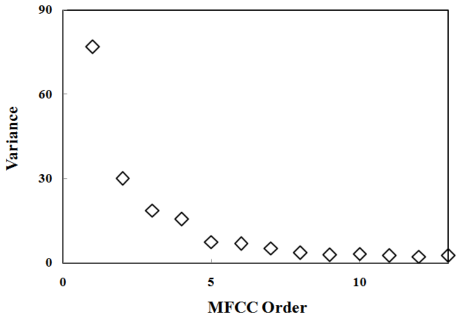

Fig. 4 Relative variances of the MFCC values

IV. Experiment

Our experiments were performed on a set of phone-balanced 300 Korean words. Forty people of 20 male and female speakers each produced speech utterances, which were divided into three disjoint groups as in Table 1.

Table 1 Division of the 40 people's speech

Speaker Group	Number of People
I	28
II	6
III	6

Speech tokens of the group I were used in generating codebook of size 512, whose centroids serve for Fuzzy Vector Quantization(FVQ) of all the speeches of 40 people.

HMM parameters were updated on each iteration of training. In order to choose which values of parameters to use in the final test of speech recognition, some test speeches are necessary. The parameters that yield the best performance on the group II were stored and used for the test on the group III to obtain the final result of the system. This prescription prevents the system from falling into the local minimum driven by the training samples of the group I. Otherwise, the system becomes less robust against the speaker-independence when applied to the group III. It is a good strategy for balance between memorization and generalization[12].

The speech utterances were sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32 ms of time duration were taken to be a speech frame for short-term analysis. The next frame was obtained by shifting 256 data points, thereby overlapping the adjacent frames by 50% in order not to lose any information contents of coarticulation[13].

To each frame, Hanning window was applied after pre-emphasis for spectral flattening. MFCC feature vectors of order 13 were obtained and then Cepstral Mean Subtraction(CMS)[14] were applied on utterance basis to endow robustness against various adverse effects such as system dependence and noisy environment. CMS was performed on the real and imaginary components independently.

Codebooks of 512 clusters were generated by the k-means clustering algorithm on the MFCC feature vectors obtained from the speeches of the group I of Table 2. The distances between the vectors and the codebook centroids were calculated and sorted. Appropriately normalized fuzzy membership values were assigned to the nearest two clusters and a train of doublets(cluster index / fuzzy membership) fed into the machine of Hidden Markov Model (HMM) for speech recognition processing.

As for the HMM, a non-ergodic left-right (or Bakis) model was adopted. The number of states that is set separately for each class(word) was made proportional to the average number of frames of the training samples in that class[15]. Initial estimation of HMM parameters $\lambda=(\pi,A,B)$ was obtained by K-means segmental clustering after the first training. By this procedure, convergence of the parameters became so fast that enough convergence was reached mostly in several epochs of training iterations.

Backward state transitions were prohibited by suppressing the state transition probabilities a_{ij} with $i>j$ to a very small value, but skipping of states was allowed. The last frame was restricted to end up with the final state associated with the word being scored within a tolerance of 3. Parameter reestimation was performed by Baum-Welch reestimation formula with scaled multiple observation sequences to avoid machine-errors caused by repetitive multiplication of small numbers. After each iteration, the event observation probabilities $b_i(j)$ were boosted above a small value.

Three features were monitored while training the HMM parameters: (1) the recognition error rate for the group II of Table 1, (2) the total probability likelihood of events summed over all the words of the training set according to the trained model, and (3) the event observation probabilities for the first state of the first word in the vocabulary list. Training was terminated manually when the convergences for these three features were thought to be enough. The parameter values of $\lambda=(\pi,A,B)$ that give the best result for the group II were stored and used in speech recognition test on the group III of Table 2.

V. Results and Discussion

Fig. 5 shows the speech recognition error rate E vs. the number N of total MFCC components(for separate real and imaginary FFT components).

The experimental values are shown by diamond symbols. The horizontal solid line denotes the average value(2.13%) of E for the seven data of the right side. The horizontal dotted line denotes the value(3.50%) of E for the conventional MFCC with 13 components.

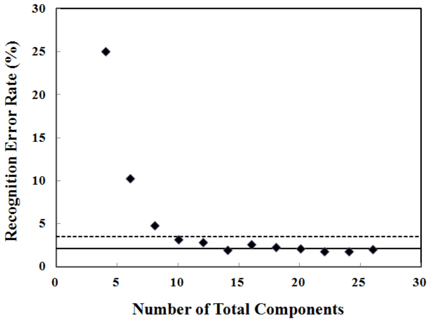

Fig. 5 Recognition error rate E vs. the number N of total FFT components

The data in Fig. 5 might be phrased in terms of two regimes. As the number N of total components is decreased from large to small values, the recognition error rate E does not show significant changes(regime I). However, below a certain threshold around $N = 10$, it begins to increase rapidly(regime II).

To examine the behavior in regime II and thereby determine the optimal value for N, we need numerical analysis. We might try employing the exponential model

$$E = E_0 \exp(-\alpha N)$$

with adjustable parameters E_0 and α, which is to be determined from curve-fitting. However, this does not seem to be a good choice, since it implies

$$E \rightarrow E_0 \quad \text{as} \quad N \rightarrow 0$$

which should not be the case.

For this reason, we employ a power law model

$$E = ax^b$$

with adjustable parameters a and b. By taking the logarithm of both sides and applying the routine of the least square method, a and b can be calculated. The best fit for the left four data of Fig. 5 was found to be

$$E = 588\,x^{-2.27} \tag{7}$$

and is shown in Fig. 6 by the solid curve.

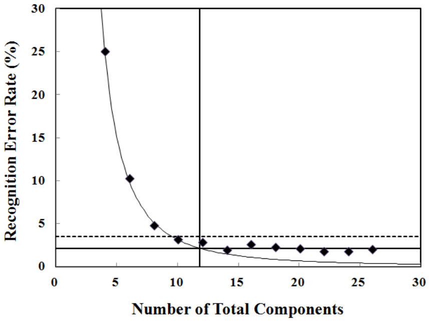

Fig. 6 Recognition error rate vs. the number of total FFT components

The vertical line within the graph denotes intersection of Eq. (7) and the horizontal solid line. Its value of N represents the optimal one in that it affords the minimum number of total components with no significant degradation of the recognition error rate. The value is around 12, which means that six components of real and imaginary components each are the best choice for the recognition task of our experiments.

VI. Conclusion

In an effort to improve speech recognition performance, we considered a method of utilizing the phase information in the Fourier transform of the speech signal, which is discarded in conventional MFCC. Among several approaches for that purpose, we studied of treating the real and imaginary FFT components separately for MFCC extraction.

This idea was tested by speaker-independent speech recognition of 300 Korean isolated words by FVQ and HMM. The number of total(real and imaginary) components was varied from 4 to 26.

The experimental result showed largely two stages of changes as the value of the total components is decreased from large values. For relatively large values, the recognition error rate does not show significant change. Below a certain threshold value, however, it increases nonlinearly.

From numerical analysis, it was shown that the optimal value of the number of total components be 12 corresponding to six real and imaginary components each. Above this value, the recognition error rate was about 2.13% which is to be compared with 2.58% obtained by the conventional MFCC. Below 12 total components, the recognition error rate increases according to a power function.

References

[1} G. Kaplan, "Words into action: I," *IEEE Spectrum*, vol. 17, 1980, pp. 22-26.

[2] Y. Chang, S. Hung, N. Wang, and B. Lin, "CSR: A Cloud-assisted speech recognition service for personal mobile device," *Int. Conf. on Parallel Processing*, Taipei, Taiwan, Sep. 2011, pp. 305-314.

[3] M. Kang, "A Study on the Design of Multimedia Service Platform on Wireless Intelligent Technology," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 4, no. 1, 2009, pp. 24-30.

[4] J. Yoo, H. Park, H. Shin, and Y. Shin, "A Study of the Communication Infrastructure Construction for u-City in Korea," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 1, no. 2, 2006, pp. 127-135.

[5] B. Kim, "Service Quality Criteria for Voice Services over a WiBro Network," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 6, no. 6, 2011, pp. 823-829.

[6] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, no. 9, 1993, pp. 1215-1247.

[7] B. Bozkurt and L. Couvreur, "On the use of phase information for speech recognition," In *Proc. of Eusipco*, Antalya, Turkey, 2005, pp. 1-4.

[8] K. K. Paliwal, "Usefulness of phase in speech processing", *Proc. IPSJ Spoken Language Processing Workshop*, Gifu, Japan, Feb. 2003, pp. 1-6.

[9] J. C. Wang, J. F. Wang, and Y. Weng, "Chip design of MFCC extraction for speech recognition," *The VLSI Journal*, vol. 32, 2002, pp. 111-131.

[10] J. M. Bioucas-Dias and G. Valadao, "Phase Unwrapping via Graph Cuts," *IEEE Trans. on Image Processing*, vol. 16 no. 3, 2007, pp. 698-709.

[11] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex Cepstrum-Based Decomposition of Speech for Glottal Source Estimation," *Interspeech*, Brighton, Sep. 2009, pp. 116-119.

[12] L. Fausett, *Fundamentals of Neural Networks*, New Jersey: Prentice-Hall, 1994.

[13] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, New York: Macmillan, 1994.

[14] W. Xu, Zhengzhou, Y. Guo, B. Wang and X. Wang, "A Noise Robust Front-End Using Wiener Filter, Probability Model and CMS for ASR," *Int. Conf. on Natural Language Processing and Knowledge Engineering*, Zhengzhou, China, 2005, pp. 102-105.

[15] M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, "Unconstrained Farsi Handwritten Word Recognition Using Fuzzy Vector Quantization and Hidden Markov models," *Pattern Recognition Letters*, vol. 22, 2001, pp. 209-214.

저자 소개

이창영 (Chang-Young Lee)

1982년 2월 서울대학교 물리교육학과 졸업 (이학사)
1984년 2월 한국과학기술원 물리학과 졸업 (이학석사)
1992년 8월 뉴욕주립대학교 (버펄로) 물리학과 졸업 (이학박사)
1993년～현재 동서대학교 시스템경영공학과 교수
※ 주 관심분야 : 패턴인식, 신호처리