Regular paper http://dx.doi.org/10.13067/JKIECS.2013.8.9.1299

음성인식에서 주 성분 분석에 의한 차원 저감

이창영*

Dimensionality Reduction in Speech Recognition by Principal Component Analysis

Chang-Young Lee*

요 약

이 논문에서 우리는 MFCC 특징벡터의 차원 저감을 통해 음성 인식에서의 계산량을 줄이는 방법을 조사한다. 특징벡터의 특성분해는 벡터의 성분을 분산의 크기에 따라 배치되도록 선형 변환 시켜준다. 첫 번째 성분은 가장 큰 분산을 가져서 패턴 분류에서 가장 중요한 역할을 한다. 따라서, 분산이 작은 성분들을 제외시키는 차원 저감을 통하여, 계산량을 줄이면서 동시에 음성 인식 성능을 저하시키지 않는 방법을 생각할 수 있다. 실험 결과, MFCC 특징벡터의 성분을 절반 정도로 줄여도 음성인식 오류율에 큰 악영향이 없음이 확인되었다.

ABSTRACT

In this paper, we investigate a method of reducing the computational cost in speech recognition by dimensionality reduction of MFCC feature vectors. Eigendecomposition of the feature vectors renders linear transformation of the vectors in such a way that puts the vector components in order of variances. The first component has the largest variance and hence serves as the most important one in relevant pattern classification. Therefore, we might consider a method of reducing the computational cost and achieving no degradation of the recognition performance at the same time by dimensionality reduction through exclusion of the least-variance components. Experimental results show that the MFCC components might be reduced by about half without significant adverse effect on the recognition error rate.

키워드

Dimensionality Reduction, Speech Recognition, Computational Cost Reduction, Principal Component Analysis
차원 저감, 음성 인식, 계산 비용 저감, 주 성분 분석

I. Introduction

The state-of-the-art technology in the field of speech recognition has now reached such a level of performance and robustness that permits lots of daily applications. As a result, we are now living in a world of various appliances which deploy the relevant technologies [1-7].

Redundancy is unavoidably inherent in any kind of information. There are four major causes for it: hardware, information, time, and software. In some applications, redundancy is utilized to improve the system performance [8-11]. For example, difference in perception between natural speech and

* 교신저자(corresponding author) : 동서대학교 산업경영공학과(seewhy@dongseo.ac.kr)
접수일자 : 2013. 07. 03 심사(수정)일자 : 2013. 08. 23 게재확정일자 : 2013. 09. 23

high-quality synthetic speech is inferred to be due to the redundancy of the acoustic-phonetic information encoded in the speech signal. Therefore, in this case, it is better not to remove redundancy for natural speech [12]. There are many other instances, on the other hand, that redundancy reduction enhances the relevant performance [13]. The ultimate goal in regards to the redundancy is to pursue both maximum relevance and minimum redundancy (MRMR) at the same time [14].

The speech signal is in itself somewhat redundant and it is usual to extract feature vectors that contain useful information in a reduced size. One of the popular feature vectors is mel-frequency cepstral coefficients (MFCC), which is commonly adopted in speech recognition.

The number of components in MFCC is usually taken to be 13 on empirical grounds [15]. In this paper, as a means for decreasing the calculational cost without deteriorating the recognition performance, we investigate a method of reducing the MFCC components by principal component analysis (PCA) [16].

The organization of this paper is as follows. Section II provides a brief description on PCA. Section III describes details of the experiments performed in our study. In section IV, experimental results are presented to demonstrate the efficacy of the method. Concluding remarks are given in section V finally.

II. Principal Component Analysis

Given a set of vectors

$$\mathbf{x}^{(i)} \in R^M,\ i = 1 \sim N, \tag{1}$$

the mean vector $\boldsymbol{\mu} \in R^M$ is first estimated:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}. \tag{2}$$

The new vectors with zero mean are obtained by

$$\mathbf{y}^{(i)} = \mathbf{x}^{(i)} - \boldsymbol{\mu},\ i = 1 \sim N. \tag{3}$$

Correlation matrix for new vectors are then estimated to be

$$R_{jk} = \frac{1}{N}\sum_{i=1}^{N} y_j^{(i)} y_k^{(i)},\ j = 1 \sim M,\ k = 1 \sim M. \tag{4}$$

The eigenvalue equation is given by

$$\mathbf{R}\,\mathbf{q}^{(j)} = \lambda_j \mathbf{q}^{(j)},\ j = 1 \sim M, \tag{5}$$

the eigenvalues of which, $\lambda^{(j)}$, are assumed to be arranged in descending order as follows:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M. \tag{6}$$

The normalized eigenvectors

$$\boldsymbol{q}^{(j)},\ j = 1 \sim M$$

form a set of basis vectors in the transformed vector space. The vectors transformed by

$$z_j^{(i)} = \sum_{k=1}^{M} y_k^{(i)} q_k^{(j)},\ i = 1 \sim N,\ j = 1 \sim M \tag{7}$$

are said to have "the principal components".

This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components.

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way that best explains the variance

in the data. If a multivariate data set is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the lower-dimensional picture, a "shadow" of this object when viewed from its most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced.

III. Reducing MFCC Orders with PCA

By applying PCA on MFCC and adopting the components of large variances, we reduced the calculational cost. For evaluation of the proposed methods, we consider two independent approaches. The first one is the speech recognition performance. The second approach, a supplement to the first one, is concerned with the separability of MFCC feature vectors. For this purpose, we employ the Fisher discriminant objective function.

Pattern classification is a very important task in many fields such as data mining, image and speech coding, pattern recognition, and other statistical analyses. An efficient procedure for this job should have the objective of separating the classes in multi-dimensional data space as discriminatively as possible. In pattern classification, separability of patterns is usually estimated by the Fisher discriminant objective function given by S_B/S_W. S_B and S_W represent the between-class and within-class scatters respectively, which are expressed by

$$S_B = \sum_{i=1}^{C} N_i(\boldsymbol{\mu_i} - \boldsymbol{\mu})(\boldsymbol{\mu_i} - \boldsymbol{\mu})^T$$

$$S_W = \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{x} - \boldsymbol{\mu_i})(\boldsymbol{x} - \boldsymbol{\mu_i})^T$$

$\boldsymbol{x}$'s are feature vectors and C_i denotes the i-th class. $\boldsymbol{\mu_i}$ and $\boldsymbol{\mu}$ represent mean values for the class C_i and for the whole feature vectors respectively. Given a set of feature vectors, principal component analysis and/or discriminant analysis might be utilized to find transformations of the extracted MFCC vectors aiming at efficient separability [19-20].

Figure 1 shows block diagram of the speech recognition with dimensionality reduction by PCA. The procedure without the broken box corresponds to the conventional method.

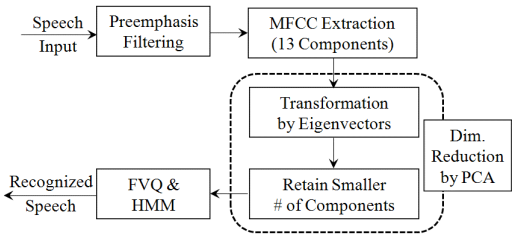

Fig. 1 Block diagram of the speech recognition with dimensionality reduction by PCA

IV. Experimental Results and Discussion

Our experiments were performed on a set of phone-balanced 300 Korean words. To see the effect of vocabulary size also, we divided the words into three sets as in Table 1. The sets A and B are disjoint each other and C is the union of them.

Table 1. Three sets of speech data divided for studying the effect of the vocabulary size

Word Set	Number of Words
A	100
B	200
C	300

Forty people including 20 males and 20 females produced the speech and the utterance tokens were divided into three disjoint groups as in Table 2.

Table 2. Division of the 40 people's speech production into three groups

Speaker Group	Number of People
I	32
II	4
III	4

Thirty-two people's speech tokens of the group I were used in generating codebooks 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 epoch. 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 performance of the speaker-independent speech recognition system. This prescription prevents the system from falling (overfitting) too deep into the local minimum driven by the training samples of the group I and hence becoming less robust against the speaker-independence when applied to the group III. It is a good strategy for balance between memorization and generalization [17].

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.

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) were applied on utterance basis to endow robustness against various adverse effects such as system dependence and noisy environment.

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 two doublets (cluster index / fuzzy membership) were fed into HMM for speech recognition processing.

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 [18]. Initial estimation of HMM parameters $\lambda=(\pi,A,B)$ was obtained by K-means segmental clustering after the first training. This procedure facilitates the convergence of the parameters so fast.

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 that might be 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 2, (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 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.

Figure 2 shows the variances for MFCC feature

vector components with and without application of PCA.

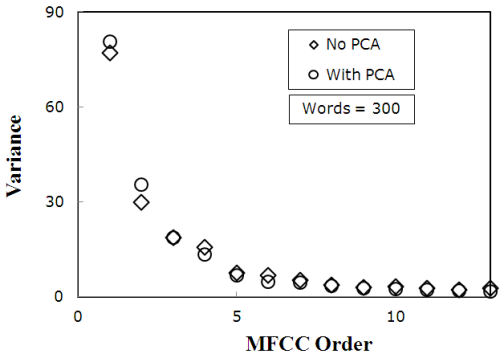

Fig. 2 The variances for MFCC feature vector components with and without application of PCA

We see that the first few components have relatively large variances. Another feature is that the effect of PCA is slightly meaningful only for the first two components. This means that the feature vectors are already distributed almost along the principal axes.

As for the computational cost, it is roughly proportional to the number of feature vector components. Since the eigenvectors $\boldsymbol{q}$ are already determined in the training phase, the additional cost accompanying PCA in the recognition phase comes from the vector transformation of Eq. (7), which is practically insignificant. Therefore, if we retain six MFCC components, for example, then the computational cost might be reduced by half or so.

Figure 3 shows Fisher discriminant score (squares) and recognition error rate (diamonds) for vocabulary size of $W=200$. The abscissa is the included number of MFCC components of large variances. The horizontal dotted line denotes the recognition accuracy for conventional method where 13 components with no PCA are all used.

We see that no significant degradation of recognition performance and Fisher discriminability occur if we include six or more MFCC components.

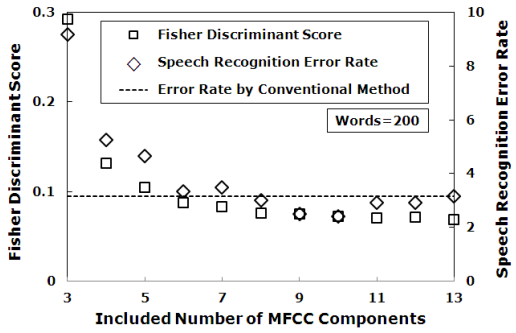

Fig. 3 Fisher discriminant score (squares) and recognition error rate (diamonds) for vocabulary size of $W=200$

In other words, we might lessen the computational cost for speech recognition by including only six components of feature vectors. In short, the dimensionality reduction by PCA eigendecomposition is expected to be effective in decreasing the computational cost of speech recognition. Similar results were reported by Hu et al. for phoneme recognition [21].

IV. Conclusion

In this paper, we studied the method of reducing the computational cost in speech recognition by dimensionality reduction through principal component analysis. PCA was found to be of a little help in enhancing the variances of the vector components.

The test of the proposed method was performed in two ways: speech recognition with FVQ/HMM and Fisher discriminant analysis. Experimental results have shown that about half of the MFCC components might be discarded without significant adverse effects on the speech recognition performance and the Fisher discriminant score.

References

[1] M. Pleva, "Speech and Mobile Technologies for Cognitive Communication and Information Systems", 2011 2nd International Conference on Cognitive Infocommunications, pp. 1-5, 2011.

[2] S. Primorac & M. Russo, "Android Application for Sending SMS Messages With Speech Recognition Interface", 2012 Proceedings of the 35th International Convention, pp. 1763-1767, 2012.

[3] G. Nemeth, "Speech-Enhanced Interaction with TV", 2011 2nd International Conference on Cognitive Infocommunications, pp. 1-5, 2011.

[4] O. Viikki, I. Kiss, & J. Tian, "Speaker- and Language-Independent Speech Recognition in Mobile Communication Systems", ICASSP '01, Vol. 1, pp. 5-8, 2001.

[5] M. Kang, "A Study on the Design of Multimedia Service Platform on Wireless Intelligent Technology", The Journal of the Korea Institute of Electronic Communication Sciences, Vol. 4, No. 1, pp. 24-30, 2009.

[6] J. Yoo, H. Park, H. Shin, & Y. Shin, "A Study of the Communication Infrastructure Construction for u-City in Korea", The Journal of the Korea Institute of Electronic Communication Sciences, Vol. 1, No. 2, pp. 127-135, 2006.

[7] Y. Kim & H. Lee, "A Study on Improved Method of Voice Recognition Rate", The Journal of the Korea Institute of Electronic Communication Sciences, Vol. 8, No. 1, pp. 77-83, 2013.

[8] I. Spiro, G. Taylor, G. Williams, & C. Bregler, "Hands by Hand: Crowd-Sourced Motion Tracking for Gesture Annotation", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 17-24, 2010.

[9] W. Sun, Z. Wu, H. Hu, & Y. Zeng, "Multi-Band Maximum a Posteriori Multi-Transformation Algorithm Based on the Discriminative Combination", International Conference on Machine Learning and Cybernetics, Vol. 8, pp. 4876-4880, 2005.

[10] H. Tohidypour, S. Seyyedsalehi, H. Roshandel, & H. Behbood, "Speech Recognition Using Three Channel Redundant Wavelet Filterbank", 2nd International Conference on Industrial Mechatronics and Automation (ICIMA), Vol. 2, pp. 325-328, 2010.

[11] M. Paulik & A. Waibel, "Spoken Language Translation from Parallel Speech Audio: Simultaneous Interpretation as SLT Training Data", ICASSP, pp. 5210-5213, 2010.

[12] D. Pisoni, H. Nusbaum, & B. Greene, "Perception of Synthetic Speech Generated by Rule", Proceedings of the IEEE, Vol. 73, No. 11, pp. 1665-1676, 1985.

[13] S. Alizadeh, R. Boostani, & V. Asadpour, "Lip Feature Extraction and Reduction for HMM-Based Visual Speech Recognition Systems", 9th International Conference on Signal Processing (ICSP), pp. 561-564, 2008.

[14] V. Estellers, M. Gurban, & J. Thiran, "Selecting Relevant Visual Features for Speech Reading", IEEE International Conference on Image Processing (ICIP), pp. 1433-1436, 2009.

[15] J. Deller, J. Proakis, & J. Hansen, "Discrete-Time Processing of Speech Signals", Macmillan, New York, pp. 246-251, 1994.

[16] S. Haykin, "Neural Networks", Prentice Hall, New Jersey, pp. 392-440, 1999.

[17] L. Fausett, "Fundamentals of Neural Networks", Prentice Hall, New Jersey, p. 298, 1994.

[18] M. Dehghan, K. Faez, M. Ahmadi, & M. Shridhar, "Unconstrained Farsi Handwritten Word Recognition Using Fuzzy Vector Quantization and Hidden Markov Models", Pattern Recognition Letters, Vol. 22, pp. 209-214, 2001.

[19] J. Hung, "Optimization of Filter-Bank to Improve the Extraction of MFCC Features in Speech Recognition", Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 675-678, 2004.

[20] A. Martin, D. Charlet, & A. Mauuary, "Robust Speech/Non-Speech Detection Using LDA Applied to MFCC", 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing," Vol. 1, pp. 237-240, 2001.

[21] H. Hu & S. Zahorian, "Dimensionality Redu-

ction Methods for HMM Phonetic Recognition", 2010 ICASSP, pp. 4854-4857, 2010.

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 정보시스템공학부 교수
※ 관심분야 : 음성인식, 화자인식, 신호처리