A Study on the Optimal Mahalanobis Distance for Speech Recognition*

Chang-Young Lee**

ABSTRACT

In an effort to enhance the quality of feature vector classification and thereby reduce the recognition error rate of the speaker-independent speech recognition, we employ the Mahalanobis distance in the calculation of the similarity measure between feature vectors. It is assumed that the metric matrix of the Mahalanobis distance be diagonal for the sake of cost reduction in memory and time of calculation. We propose that the diagonal elements be given in terms of the variations of the feature vector components. Geometrically, this prescription tends to redistribute the set of data in the shape of a hypersphere in the feature vector space. The idea is applied to the speech recognition by hidden Markov model with fuzzy vector quantization. The result shows that the recognition is improved by an appropriate choice of the relevant adjustable parameter. The Viterbi score difference of the two winners in the recognition test shows that the general behavior is in accord with that of the recognition error rate.

Keywords: Mahalanobis Distance, MFCC, HMM, Fuzzy Vector Quantization, Speech Recognition

I. Introduction

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 the discipline of neural networks, unsupervised or self-organized learning algorithms such as ART [1] and SOM [2] have played central roles for that goal.

As a non-neural approach, the Linde-Buzo-Gray (LBG) clustering algorithm [3] is widely used in seeking optimal partitions iteratively. The basic idea is to give some reasonable initial partition to the data space and to move a given pattern from one group to another of closer distance. The representative vectors of the partitioned groups then constitute a codebook and an arbitrary input pattern is assigned (quantized) to one (or several in fuzzy case) class whose distance with the given pattern is the smallest. This is the rough content of the vector

* This work was supported in part by the research grant funded by Dongseo University.
** Division of Information System Engineering, Dongseo University

quantization (VQ) [4].

In the codebook design and subsequent VQ, a similarity measure should be chosen in order to compare the distances between vectors. The most simple, intuitive, and straightforward distance measure is to use the Euclidean norm. Though such a scheme allows easy calculation, it has drawback that the components of the feature vectors with seemingly non-comparable quantities are treated on equal footing. Geometrically, this method implies an isotropic feature space weighting, which might lead to trouble when the vector components correspond to different physical properties. In such cases, a clustering procedure based on Euclidean distance measure is inappropriate and a different procedure becomes inevitable. To remedy the problems caused by the Euclidean distance measure, a more general metric is adopted, which leads to the so-called Mahalanobis (or generalized Euclidean) distance.

Different scalings along different axes are also necessary in view of the effectiveness of the partitioning. The set of data usually falls in a hyperellipsoid rather than a hypersphere in the feature vector space. If we redistribute the data in the form of a hypersphere, then the separation would be easier and more effective. It is like that a 3-dimensional sphere is better for dividing into pieces compared to a thin line.

Historically, the Mahalanobis distance has long been used in many applications that need learning and designing pattern classifier. Competitive neural network approach using Mahalanobis metrics was studied by Martins et al. [5] and a different approach using ARTMAP network can be found in the work by Xu and Vuskovic [6]. Applications to the pattern recognition employing the concept of the Mahalanobis distance are innumerable, exemplary ones being the face recognition [7-8] and the hand-written character recognition [9].

As for the application to speech realm, Schwarz et al. [10] compared probability density function approach with the classification by Mahalanobis distance calculation in the text-independent speaker identification. More efforts on the speaker verification can be found in [11-13].

The application of Mahalanobis distance pervades in many fields utilizing VQ. This in turn signifies that the idea of Mahalanobis distance is invaluable in the pattern classification. However, not much effort utilizing it has been done in the field of speech recognition. In this paper, we try to use Mahalanobis distance in designing codebook and implementing vector quantization and see the effectiveness of the result by applying it to the speaker-independent speech recognition by hidden Markov model (HMM) combined with fuzzy vector quantizer (FVQ) [14].

II. Theory

The important question for classification of a set of feature vectors is how to determine the similarity measure between two vectors. The most intuitive and straightforward choice is to use the Euclidean distance

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{F} (x_i - y_i)^2} \quad (1)$$

where F is the dimension of the feature vector space. In this scheme, all the components are treated on equal footing and thus given the same weighting factors.

The problem with this approach, however, is that the components might have in general different physical properties. As an example, let's suppose we are to divide a number of persons into some groups. As the feature vector for classification, we use two features of "gender" and "age", which together form a two-dimensional vector. If binary representation for the first component "gender" is adopted, then its value will be 1 or 0 according whether the person is male or female. If we calculate the "distance" of the two persons based on (1), then the "gender" would be smeared out by "age" and do not play a significant role in the classification, the reason simply being that the difference in the age is mostly much bigger than that of the gender. It is not difficult, at least in this example, to see that a naive scheme of (1) would lead to trouble and inconsistency.

In order to solve this problem, it is desirable to give larger weighting factor to the gender. As a concrete example, it might be a plausible solution to give, say, 100 times bigger weighting factor to the gender compared to the age in the illustrative example given above. This comprises the motivation of our study in this paper. The relative magnitude of the weighting factors for feature vector components should be determined eventually by examining some objective function such as the recognition accuracy.

In the case of mel-scale frequency cepstral coefficients (MFCC) [15] which are widely used for speech coding and recognition, the components are divided by frequency range according to the human auditory response [16]. Though the components have thus the same physical dimension, there's no a priori reason that the components are calculated by the same weighting factors as in (1).

As a remedy to the problem caused by (1), the following Mahalanobis distance is generally used:

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T A^{-1} (\boldsymbol{x} - \boldsymbol{y})} \quad (2)$$

where the superscript T denotes matrix transpose and A is any positive definite $F \times F$ matrix. Some research works have used for A the covariance matrix formed by the clusters corresponding to the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ [17-18]. When the matrix A is set to the identity matrix I, (2) is reduced to the special case (1).

For the sake of cost reduction in memory and time of calculation, we assume the matrix A be diagonal. Then (2) can be written as

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{F} a_i (x_i - y_i)^2} \tag{3}$$

The remaining question is how to choose the diagonal elements a_i which can be viewed as a set of eigenvalues for the matrix A in (2).

In order to see our motivation for the choice of the parameters a_i, we revisit the example classification of people by two-dimensional vectors formed by gender and age. If the two axes are scaled the same way, then the variation of the first component is negligible compared to the second one and the resulting data set would look almost like a straight line. This means that the first component does not play significant role in classifying the patterns. In the extreme case where the variation in the first component becomes zero, that feature is actually missed in the calculation. A simple solution to this problem is to elongate the data along the first axis. Geometrically, this corresponds to converting the distribution of the data set from sharp ellipsoidal to circular geometry in the feature vector space.

A plausible candidate for the parameters a_i is therefore to choose in such a way that will give larger weights to the components with small variations. An intuitive choice is to use the following prescription:

$$a_i = \sigma_i^{-\alpha} \tag{4}$$

where σ_i is the variation of the i-th component of the feature vectors and α is an adjustable parameter. Larger values of α will give relatively larger weights to the components of smaller variations.

The procedure above may be addressed in different context. If we combine (3) and (4), the result can be written as

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{F} (x_i' - y_i')^2} \tag{5}$$

with the new feature vectors given by

$$x_i' = \sigma_i^{-\alpha/2} x_i \quad , \quad y_i' = \sigma_i^{-\alpha/2} y_i \tag{6}$$

From this, we see that the proposed method is equivalent to using the Euclidean distance with the feature vector replaced by a different one. In this sense, our procedure acts in adverse direction to the spirit of the feature vector extraction. That means that the exponent α in the above equations should be not too far from 0. Otherwise, the features are deformed severly from that of the extracted feature for the sake of rendering the set of data into a spherical shape. Eventually, the adjustable parameter α should be determined by the performance of the speech recognizer.

III. Experiments

Our study was performed on a set of phone-balanced 100 isolated Korean words. 40 people including 20 males and 20 females participated in speech production. Each utterance was sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32 ms of time duration were taken to be a frame. The next frame was obtained by shifting 170 data points, thereby overlapping the adjacent frames by 2/3. For each frame, Hanning window was applied after pre-emphasis for spectral flattening. As the feature vector, MFCC of order 13 was obtained with the mel-scale adopted from [19].

A codebook of size 2,048 was generated by LBG clustering algorithm [3]. The distances between the input vector and the codebook cluster centroids were calculated according to (3) and (4) and quick-sorted in order of distances. Two nearest centroids were selected and assigned membership values according to [20] and then normalized.

In spite of insufficient training data, speech utterances of 40 people were divided into three disjoint groups. The first group consisting of 32 persons' speech was used for training of HMM parameters. After each training iteration, the recognition error rate was examined on the second group consisting of speeches from 4 people. The HMM model parameters $\lambda = (\pi, A, B)$ for each word that yields the best recognition rate for this second group were recorded and used for the final test of the speaker-independent speech recognition on the third group of the remaining 4 persons.

For the HMM, a non-ergodic left-right (or Bakis) model of 20 states was adopted. Initial estimation of the parameters was obtained by K-means segmental clustering after the first training. By this procedure, the convergence of the iteration was so fast that enough

convergence was reached only after several iterations. Backward state transitions were prohibited by suppressing a_{ij} with $i > j$ to a very small value (10^{-20}) but skipping of states was allowed.

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 parameters $b_i(j)$ were boosted above 10^{-6} in view of the results from [21]. Three features were monitored during the iterations: (1) the recognition error rate for the second group, (2) the total probability likelihood of events for all the words of the training set according to the trained model, and (3) the event observation probabilities for one word. Iteration was terminated when the convergences for these three features were thought to be sufficient.

IV. Results and Discussion

<Figure 1> shows the relative variations of the 13 MFCC components. The variations of the 13 components were divided by that of the first component. The first component has by far the largest variation. Except that, the 4th component has also relatively large value compared to the other ones. Therefore, in our experiment, relatively small weighting factors were assigned to those two components in the calculation of the Mahalanobis distance of the feature vectors.

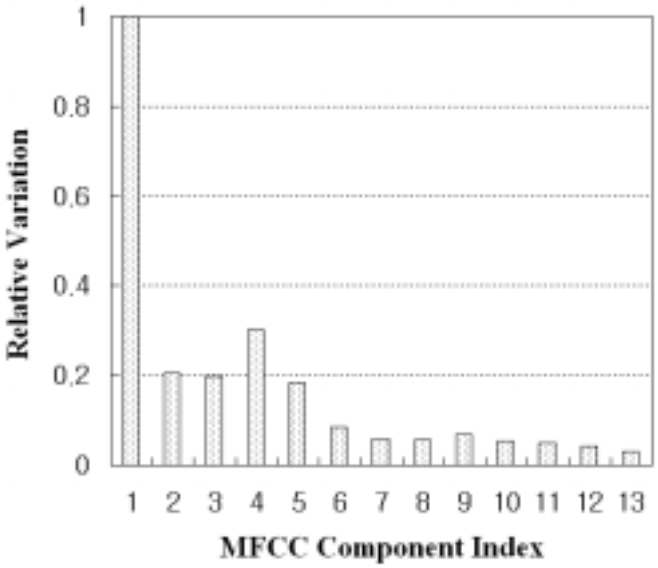

Figure 1. Relative variations of the MFCC components.

<Figure 2> shows the recognition error rate vs. the parameter α. Though it is not included in the figure, we also tested for negative values of α and the result showed that the error rate increased rapidly for $\alpha < -0.1$, which reflects the agreement with our expectation that those values would shrink the shape of the distribution of the set of feature vectors into sharper ellipsoids. For positive values of α, the recognition error rate reached the minimum at $\alpha = 0.3$ and increased beyond that value. This result suggests that the gradual redistribution of the feature vectors from ellipsoidal to spherical geometry works in the direction of improving the recognition accuracy. Along with this change, however, the deterioration of the MFCC extraction is inevitably accompanied and this effect results in the increase of the recognition error rate. It might be inferred from these observations that, as α increases, data redistribution occurs in the cooperative direction on the one hand, and feature deterioration happens in the other direction on the other hand. These two competing effects result in the minimum error rate around a certain value of α around 0.3.

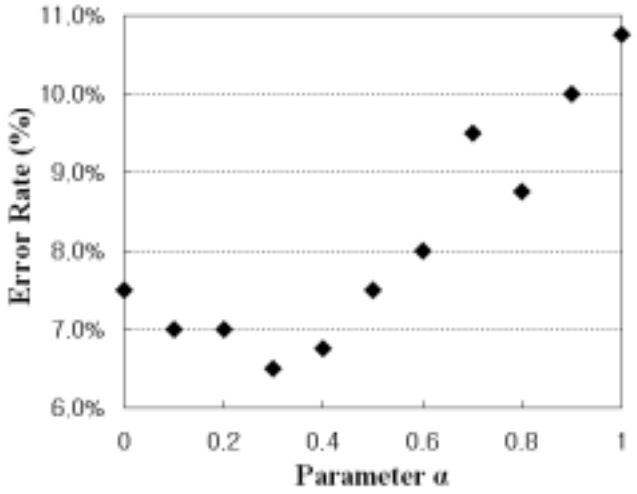

Figure 2. Recognition error rate vs. the parameter α. As α increases, recognition error rate decreases due to the gradual redistribution of the data set. For $\alpha > 0.3$, the effect of deterioration in the feature vectors results in the increase of the error rate.

The result above might be a coincidence due to displacement of the cluster centroids of the codebook and the resultant different vector quantization. In order to check whether the use of Mahalanobis distance with our prescription of the metric matrix is actually effective or not, we examined the discrimination in the Viterbi score of HMM of the two winners. <Figure 3> shows the average (log) Viterbi score difference of the two winners of the highest score for the correctly matched (recognized) word and the next highest one. The larger the value of this difference, the better the discrimination of the recognition. Though the value of α for minimum

value of the recognition error rate and that of the maximum Viterbi discrimination do not coincide exactly, the general behaviors are well within the resonable accord with each other and our expectation.

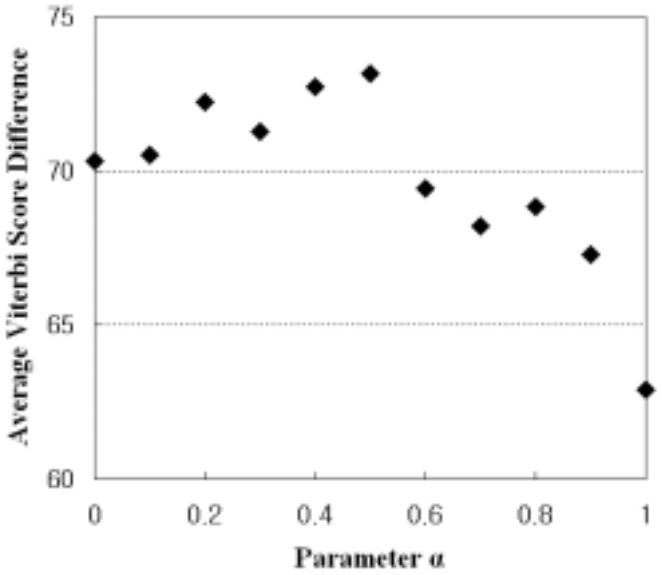

Figure 3. The average Viterbi score difference for the correctly matched (recognized) words vs. the parameter α.

V. Conclusion

In this paper, a search for the optimal Mahalanobis distance for speech recognition was performed. For the metric in the calculation of the similarity measure for clustering and subsequent vector quantization, we considered diagonal matrix whose eigenvalues are taken to be some inverse power of the variation in the MFCC feature vector components. This prescription has the effect of redistributing the feature vectors in such a way that change their distribution from ellipsoidal to spherical one in feature vector space. Mathematically, however, the proposed method leads inevitably to the deterioration in the extracted feature vector due to different scaling for each vector component. As a result, the speaker-independent speech recognition error rate showed its minimum by an appropriate choice of the adjustable parameter for the Mahalanobis distance. To confirm the effectiveness of our work, we also investigated the difference in the Viterbi scores of the recognized word and the next candidate. The result also showed similar behavior with the recognition rate, which might be considered as

suggesting that the Mahalanobis distance with our prescription of the elements do work in improving the pattern classification and recognition.

References

[1] Carpenter, G. A., & Grossberg, S. 1987. "A Massively Parallel Architecture For A Self-Organizing Neural Pattern Recognition Machine." *Computer Vision, Graphics, and Image Processing*, 37, 54-115.

[2] Kohonen, T. 1989. "*Self-Organization And Associate Memory.*" Springer-Verlag.

[3] Linde, Y., Buzo, A., & Gray, R. M. 1980. "An Algorithm for Vector Quantizer Design." *IEEE Trans. on Communications*, Vol. 28, 84-95.

[4] Rabiner, L. & Juang, B. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall International, Inc., 122-132.

[5] Martins, A., Neto, A., & Melo, J. 2003. "Neural Networks Applied to Classification of Data Based on Mahalanobis Metrics." *Proceedings of the International Joint Conference on Neural Networks*, Vol. 4, 3071-3076.

[6] Xu, H., & Vuskovic, M. 2004. "Mahalanobis Distance-Based ARTMAP Network." *IEEE International Joint Conference on Neural Networks*, 3, 2353-2359.

[7] Fraser, A., Hengartner, N., Vixie, K., & Wohlberg, B. 2003. "Incorporating Invariants in Mahalanobis Distance Based Classifier: Application to Face Recognition." *Proceedings of the International Joint Conference on Neural Networks*, Vol. 4, 3118-3123.

[8] Beveridge, J., She, K., & Draper, B. 2001. "A Nonparametric Statistical Comparison of Principal Component and Linear Discriminant Subspaces for Face Recognition." *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, I535-I542.

[9] Kato, N., Omachi, S., Aso, H., & Nemoto, Y. 1999. "A Handwritten Character Recognition System Using Directional Element Feature and Asymmetric Mahalanobis Distance." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 3, 258-262.

[10] Schwartz, R., Roucos, S., & Berouti, M. 1982. "The Application of Probability Density Estimation to Text-Independent Speaker Identification." *IEEE International Conference on ICASSP '82*, Vol 7, 1649-1652.

[11] Shridhar, M., Mohankrishnan, N., & Sid-Ahmed, M.A. 1983. "A Comparison of Distance Measures For Text-Independent Speaker Identification." *IEEE International Conference on ICASSP '83*, Vol 8, 559-562.

[12] Ward, R., & Gowdy, J. 1989. "An Investigation of Speaker Verification Accuracy Using Fundamental Frequency and Duration as Distinguishing Features." *Twenty-First Southeastern Symposium on System Theory*, 390-394.

[13] Ramachandran, R., Zilovic, M., & Mammone, J. 1995. "A Comparative Study of Robust Linear Predictive Analysis Methods with Applications to Speaker Identification." *IEEE. Trans. on Speech and Audio Processing*, Vol. 3, No. 2, 117-125.

[14] Tsukoba, E., & Nakahashi, J. 1994. "On the Fuzzy Vector Quantization Based Hidden Markov Model." *Proc. ICASSP*, 1637-1640.

[15] Wang, Jia-Ching., Wang, Jhing-Fa., & Weng, Yu-Sheng. 2002. "Chip Design of MFCC Extraction for Speech Recognition." *The VLSI Journal*, Vol. 32, 111-131.
[16] Rabiner, L. & Juang, B. 1993. op. cit., 183-190.
[17] Younis, K., Rogers, S., & DeSimio, M. 1996. "Vector Quantization Based On Dynamic Adjustment of Mahalanobis Distance." *Proceedings of the IEEE 1996 National Aerospace And Electronics Conference*, Vol. 2, 858-862.
[18] Deer, P.J., Eklund, P.W., & Norman, B.D. 1996. "A Mahalanobis Distance Fuzzy Classifier." *Australian and New Zealand Conference on Intelligent Information Systems*, 220-223.
[19] Picone, J. W. 1993. "Signal Modeling Techniques in Speech Recognition." *Proc. IEEE*, Vol. 81, No. 9, 1215-1247.
[20] Lee, C.-Y., Nam, H., Jung, H., & Lee, C.-B. 2005. "The Effect of Membership Concentration in FVQ/HMM for Speaker-Independent Speech Recognition." *Speech Sciences*, Vol. 12, No. 4, 7-15.
[21] Levinson, S. E., Rabiner, L. R., & Sondhi, M. M. 1983. "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition." *Bell System Tech. J.*, Vol. 62, No. 4, 1035-1074.

received: October 3, 2006
accepted: November 27, 2006

▲ Chang-Young Lee
Div. of Information System Engineering
Dongseo University
Jurye San 69-1, Sasang, Pusan 617-716, Korea
Tel: +82-51-320-1719 Fax: +82-51-320-2389
E-mail: seewhy@dongseo.ac.kr