Selecting Good Speech Features for Recognition

Youngjik Lee and Kyu-Woong Hwang

CONTENTS

ABSTRACT

This paper describes a method to select a suitable feature for speech recognition using information theoretic measure. Conventional speech recognition systems heuristically choose a portion of frequency components, cepstrum, mel-cepstrum, energy, and their time differences of speech waveforms as their speech features. However, these systems never have good performance if the selected features are not suitable for speech recognition. Since the recognition rate is the only performance measure of speech recognition system, it is hard to judge how suitable the selected feature is. To solve this problem, it is essential to analyze the feature itself, and measure how good the feature itself is. Good speech features should contain all of the class-related information and as small amount of the class-irrelevant variation as possible. In this paper, we suggest a method to measure the class-related information and the amount of the class-irrelevant variation based on the Shannon's information theory. Using this method, we compare the mel-scaled FFT, cepstrum, mel-cepstrum, and wavelet features of the TIMIT speech data. The result shows that, among these features, the mel-scaled FFT is the best feature for speech recognition based on the proposed measure.

I. INTRODUCTION

For three decades since 1960's, various speech recognition methods have been developed to recognize isolated words and continuous speech. They are dynamic time warping method [1], [2], [3], and hidden Markov model [4], [5], [6], [7], respectively. These methods have found their own application such as voice-controlled computers and automatic response systems.

Recently, spontaneous speech recognition has become one of the major research areas since it handles really natural human voice. There are uncertain pronunciations, varying speed, and incomplete sentences in spontaneous speech, which make the recognition problem difficult to solve. Currently, the best word recognition rate is 70 % for limited domain [8], 50 % for multiple domains [9], which is far from user requirements.

Good feature extraction has been an important research topic in speech recognition area. There has been many kinds of features for speech recognition. Formants and linear prediction parameters are such features in earlier days. Cepstrum [10] has been widely used since it is free from pitch variation. To incorporate the frequency response of human auditory system, mel-scaling was introduced [11]. Recently, the linear discriminant analysis [12] of features is commonly used in most speech recognition systems since it efficiently reduce the dimension of feature parameters.

Feature extraction is in the beginning stage of speech recognition. If this procedure loses any small portion of the class-related information, there is no way to recover it in the later stages. One usually tests various speech features on the same recognition system and compares the corresponding recognition ratio to select a good feature. Even in this case, different architectures of classifiers and/or the post-processors may give different results. For absolute comparison, we need to compare speech features themselves.

One can use the linear separability or the Fisher criterion [13], [14] for this purpose. The linear separability is the maximum probability of classification when discriminating the patterns with hyperplanes. This measure performs well in two-class problems when the patterns are clustered in convex regions. However, speech patterns are multi-class and are not well clustered. The Fisher criterion is the ratio of the average distance among classes to the average of the variations in each class. It has a critical drawback that it cannot measure how much the classes overlap.

In this paper, we propose a method to select a good speech feature for recognition. We apply Shannon's information theory [15] to build the method. The proposed method is applicable to multi-class, non-convex, and overlapped patterns such as speech signals and speech features. Bichel and Seitz [16] have proposed a method to measure the performance of hidden and output units of feedforward neural networks using the conditional class entropy when the patterns are binary vectors. Our method is an extension of this method to continuous vectors.

This paper is organized as follows: In section II we explain the linear separability, the Fisher criterion, and the information theoretic measure, and compare their capabilities. In section III, we apply this method to TIMIT speech data to select a good speech feature, and section IV concludes this paper.

II. COMPARISON OF MEASURES

In this section, we briefly introduce the multi-class versions of the linear separability, the Fisher criterion, and the proposed measure, and compare their complexity measuring capabilities.

1. The Complexity Measures of Multi-class Distributions

A. The Linear Separability

The linear separability [14] is the maximum probability of correct classification when discriminating the pattern distribution with hyperplanes. In two-class problems, it represents the probability of overlapping if each class is distributed in convex region. In order to measure the linear separability of the multi-class patterns that have non-convex distribution, one can use a single layer perceptron [17]. A single layer perceptron divides the pattern space into a finite number of classes. Training a single layer perceptron, we can find the hyperplanes that minimize the total classification error. Figure 1 shows such examples. The linear separability clearly shows whether the two convex-distributed classes overlap as shown in Fig. 1(a). However, it fails to show the overlapping factor if the number of classes is more than two as shown in Fig. 1(b). It also fails on two non-convex distributions as shown in Fig. 1(c). In such cases, training a single-layer perceptron gives the hyperplanes that minimize the total error with respect to all the classes.

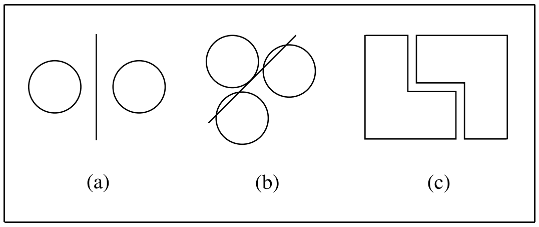

Fig. 1. Property of the linear separability: (a) Two convex-distributed patterns, (b) three convex-distributed patterns, and (c) two non-convex-distributed patterns.

In this paper, we define the linear separability as the recognition ratio of a single-layer perceptron after training. Therefore, the linear separability is

$$ls \triangleq \frac{N_{correct}}{P}, \tag{1}$$

where P is the number of the whole patterns and $N_{correct}$ is the number of correctly classified ones.

B. The Fisher Criterion

The Fisher criterion [13] is a measure that indicates how much the class distributions are separated. A multi-class version of the Fisher criterion is

$$fc = \frac{\sum_{m=1}^{M} P_m \| \mu_m - \mu_0 \|^2}{\sum_{m=1}^{M} P_m {\sigma_m}^2}, \qquad (2)$$

where M is the number of classes and P_m is the probability of the mth class. Here, μ_m is the mean and ${\sigma_m}^2$ is the variance of the patterns assigned to the mth class, and $\mu_0 = \frac{1}{M}\sum_{m=1}^{M}\mu_m$. When the patterns are multidimensional vectors, we can use the trace of the sample covariance matrix instead of the variance. It is easy to see that this value becomes large when the distances among the centers of classes are large and/or the variances of classes are small.

C. The Information Theoretic Measure

The main idea of the information theoretic measure is to calculate the class-related information content and the class-irrelevant variation from the pattern distribution using Shannon's information theory [15].

To understand this concept, we need to define two partitions, the spatial partition and the class partition. The spatial partition consists of a finite number of hypercubes that cover the whole patterns. To define the spatial partition, we first find the smallest rectangular hyper-hexahedron in the subspace that includes the whole patterns. Then, we divide it into given number of hypercubes. These hypercubes form the spatial partition **S** as shown in Fig. 2. The class partition **C** divides all patterns into each class. For phoneme recognition, for example, all Korean speech segments are naturally partitioned into 46 phoneme classes.

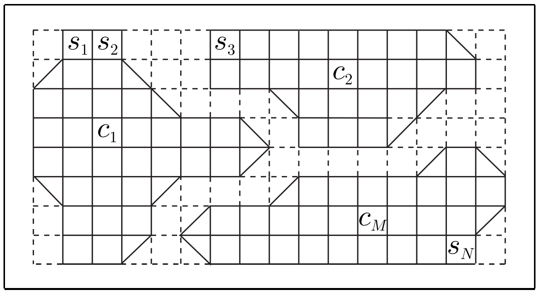

Fig. 2. An example of the two-dimensional spatial partition **S**. Here, it is drawn in the solid line. The largest dotted rectangle is the smallest rectangular hyper-hexahedron that includes the whole pattern distribution. $\mathbf{C} = \{C_1, C_2, \ldots, C_M\}$ is the class partition.

We form these two partitions on a feature vector set to calculate the class-related information content and the class-irrelevant variation. We can formulate these concepts as follows:

A spatial partition $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$ of a speech feature space **E** is a collection of hypercubes such that $\bigcup_{n=1}^{N} S_n = \mathbf{E}$, $S_i \cap S_j = \phi$, $i \neq j$. In this case, the entropy $H(\mathbf{S})$ of the spatial partition **S** is [15]

$$H(\mathbf{S}) = -\sum_{n=1}^{N} P(S_n) \log P(S_n), \qquad (3)$$

where $P(S_n)$ is the probability that speech features fall into the hypercube S_n. When the Euclidean distance between two features is small, the features are more likely to appear in the same hypercube of **S**, resulting in small $H(\mathbf{S})$. While the distance is large, they are more likely to occur in different hypercubes of **S**, resulting in large $H(\mathbf{S})$. Thus, $H(\mathbf{S})$ indicates relative scatter of the whole features.

Consider the class partition $\mathbf{C} = \{C_1, C_2, \ldots, C_M\}$ in the same speech feature space **E**. The conditional entropy $H(\mathbf{S}|\mathbf{C})$ of **C** assuming the spatial partition **S** is

$$H(\mathbf{S}|\mathbf{C}) = -\sum_{m=1}^{M} P(C_m) \sum_{n=1}^{N} P(S_n|C_m) \times \log P(S_n|C_m). \tag{4}$$

Since $\sum_{n=1}^{N} P(S_n|C_m) \log P(S_n|C_m)$ is the relative variation of the class C_m, $H(\mathbf{S}|\mathbf{C})$ is the average relative variation of all classes.

The mutual information of the spatial partition **S** and the class partition **C** is

$$I(\mathbf{S};\mathbf{C}) = H(\mathbf{S}) - H(\mathbf{S}|\mathbf{C}). \tag{5}$$

When the features of different classes occur in the same element of **S**, $H(\mathbf{S})$ decreases, resulting in reduction of $I(\mathbf{S};\mathbf{C})$. When there is no element of **S** that contains the features of different classes, $I(\mathbf{S};\mathbf{C})$ has the maximum value of $H(\mathbf{C})$. Thus, we can use $I(\mathbf{S};\mathbf{C})$ to determine whether the classes overlap or not.

We have shown in Appendix that, whenever the classes overlap,

$$I(\mathbf{S};\mathbf{C}) < H(\mathbf{C}) \tag{6}$$

and $I(\mathbf{S};\mathbf{C})$ decreases more if the shape of overlapping becomes more complicated. In particular, if $P(C_m, S_n) = P(C_m)P(S_n)$, $m = 1, 2, \ldots, M$, $n = 1, 2, \ldots, N$ only in the overlapping area,

$$I(\mathbf{S};\mathbf{C}) = (1-\delta)H(\mathbf{C}), \tag{7}$$

where δ is the probability of overlapping.

When the whole features are distributed uniformly in all the hypercubes of **S** and all the classes do not overlap, $H(\mathbf{S}|\mathbf{C})$ has a maximum value of

$$H_{max} = \log N - H(\mathbf{C}), \tag{8}$$

where N is the number of hypercubes.

Using these properties, we define the inter-class separability s and the intra-class variation v as

$$s \triangleq \frac{I(\mathbf{S};\mathbf{C})}{H(\mathbf{C})}, \quad v \triangleq \frac{H(\mathbf{S}|\mathbf{C})}{H_{max}}. \tag{9}$$

The inter-class separability s represents whether the distributions of the classes overlap. If s is 1, that is, $I(\mathbf{S};\mathbf{C})$ is equal to $H(\mathbf{C})$, the classes do not overlap, even for non-convex multi-class distributions. When s is less than 1, the classes are mingled with each other. The intra-class variation v indicates the relative size of the class-irrelevant variation within each class. If v is 1, each feature occupies different hypercubes of **S**, which is the most scattered case. If v is small, most features fall into small number of hypercubes. Thus, v represents the degree of dispersion of features within the same class. Clearly s should be 1 and v is small to be a good feature.

2. Comparison on the Complexity-measuring Capability

In this section, we discuss the relations among the above criteria and compare their properties.

As mentioned in the previous section, both the linear separability *ls* and the inter-class separability *s* measure the degree of separation of the classes. The parameter *ls* represents the complexity of given features in the extent that a class is separable from the others using a hyperplane. In most cases, however, *ls* cannot represent whether the classes overlap or not. For example, when two classes do not overlap and are not linearly separable as shown in Fig. 1(c), *ls* is less than 1, while the *s* is 1.

We can illustrate the difference between *ls* and *s* in Fig. 3 more precisely. In Fig. 3(a) and (b), the areas of the class A and the class B are unchanged. We can classify the two classes in (b) using some nonlinear classifier, but we cannot have 100 % classification in (a). The *s* of (a) is less than 1 since two classes overlap, while it is 1 in (b). However, the *ls* is the same value in both (a) and (b). Thus, the *s* can represent the overlapping factor of multi-class patterns distributed in non-convex region.

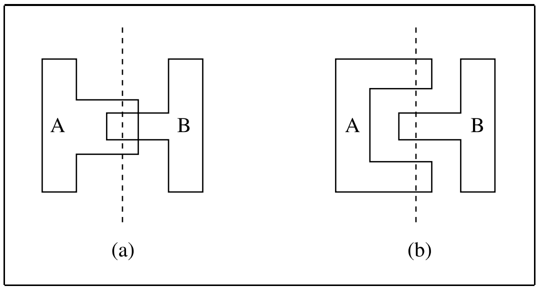

Fig. 3. The comparison of the linear separability and the inter-class separability. In (a) and (b), the areas of the class A and the class B are unchanged. The *ls* is less than 1 in both (a) and (b). The *s* is less than 1 in (a) while it is 1 in (b).

The Fisher criterion fc and the intra-class variation v are the measures that reflect the degree of dispersion of the pattern distribution. The Fisher criterion has the same value under the transformations such as translation, rotation and magnification since the numerator and the denominator in (2) change with the same factor. The intra-class variation represents relative scatter of patterns. It is also invariant under similar transformation as shown in Fig. 4.

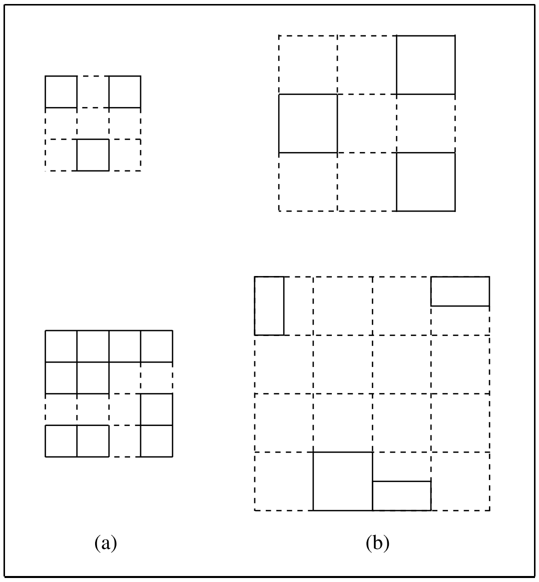

Fig. 4. Invariance of the intra-class variation and the Fisher criterion. The pattern distribution (a) is rotated, magnified, and flipped over, resulting in (b). They have the same values of the intra-class variation as well as the Fisher criterion.

However, the Fisher criterion has a critical drawback that it cannot indicate the overlapping factor. Fig. 5 shows such an example. In Fig. 5(a) and (b), the distance between the centers of two classes and the variance of each class remain equal. When the two classes are distributed uniformly in the rectangles, the fc

has the same value as 1.8808 in both (a) and (b). However, the two classes overlap in (a) while they do not in (b). In other words, the fc cannot represent the difference of the complexity in this case. In contrast, the s is 1.0 in (a) but it is 0.6666 in (b).

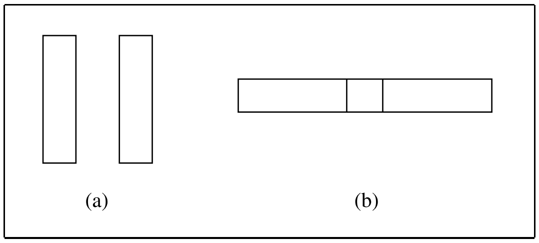

Fig. 5. The comparison of the Fisher criterion and the inter-class separability. When two distributions are uniform, the distance between the two class centers and the variance of each class remain equal. Thus, the fc is 1.8808 in both (a) and (b), However, the s is 1.0 in (a) and 0.6666 in (b).

3. Practical Considerations

Up to this point, we have assumed that the feature distribution is available. However, in real situations, only a finite number of samples are available. In this case, H_{max} in (8) can be replaced by $\log P - H(\mathbf{C})$, where P is the number of samples, provided that $N > P$. The number of samples should be large enough to represent the feature distribution fairly well.

The number of hypercubes N in the spatial partition $\mathbf{S}$ plays an important role in the proposed method. If N is too small, features from different classes may fall into the same hypercube of $\mathbf{S}$ since the volume of the hypercube is large. In this case, s becomes less than 1 even though the classes do not overlap. Thus, N should be large so that the size of the hypercube is comparable to the smallest gap between the classes. If N is too large, all features may fall into distinct hypercubes since the volume of the hypercube is very small. In this case, s is 1, but v is also 1. Thus, N should not be too large so that the size of the hypercube is larger than the smallest gap between the feature samples. To select a good speech feature, we use N from 10^3 to 10^{30} when the number of sample is 700,000 - 800,000 and the number of class is 61.

To calculate (4), it is necessary to estimate $P(S_n|C_m)$ for all n and m. If N is large, it is impossible to store all of these values. If the elements of the pattern set are statistically independent, the total entropy is just a sum of the entropies of all elements. When all the classes are Gaussian random vectors, the isometric transformation[1] under the eigenvector matrix makes the resulting elements to be statistically independent. Thus, the total entropy can be calculated by adding entropy of each element, which needs very small amount of memory and processing time. This process is mathematically equivalent to the linear discriminant analysis (LDA) [18] that is frequently used in speech recognition.

To decide the number N of the hypercubes in $\mathbf{S}$, we follow the next steps.

[1] A transformation $\mathbf{T}$ is called *isometric* if $d(x, y) = d(\mathbf{T}x, \mathbf{T}y)$ for all elements x and y where $d(x, y)$ denotes the distance between x and y.

1. Extract feature vectors from all speech samples.

2. Transform all the feature vectors using the eigenvector matrix of the sample covariance matrix into new vectors, namely, aligned vectors.

3. Set the size of the hypercube to α.

4. Since the eigenvalue represents the variance of the corresponding element, we calculate the corresponding standard deviation σ.

5. Divide 6σ by α. If it is greater than 1, divide that element into $\lceil \frac{6\sigma}{\alpha} \rceil$. If it is less than 1, do not divide that element.

6. Multiply all $\lceil \frac{6\sigma}{\alpha} \rceil$ for each element to get the total number of hypercubes N.

After finding N, we estimate the probability that the corresponding element of the aligned vector falls into each sector, and calculate the entropy.

We apply these criteria to various speech features to select good features for speech recognition in the next section.

III. ANALYSIS OF SPEECH FEATURES FOR RECOGNITION

We apply the proposed method to various speech features extracted from the phonetically labeled TIMIT speech data. The total number of frame is 782,253 and the number of classes is 61. In 92,327 frames out of the total frames, there are more than a class in a frame. In this case, we assign a class with the largest time portion to incorporate coarticulation effect. We also measure the same values from 689,926 unoverlapped frames for clear performance comparison. Fig. 6 shows the histogram of each phoneme for all frames and the unoverlapped frames. The two distributions are very similar in shape, meaning that the measurement on the unoverlapped frame is not far from the measurement on all frames.

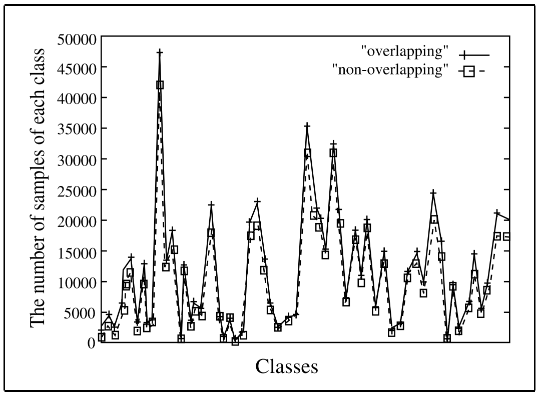

Fig. 6. Histogram of phoneme classes for overlapped and unoverlapped frames.

The candidates are cepstrum, mel-cepstrum, and mel-scaled FFT [19] features that are most frequently used in speech recognition. We also extract an energy-normalized filter-bank-type wavelet feature.

To extract cepstrum, mel-cepstrum, and mel-scaled FFT, we extract 16-order feature from 20 msec speech signal frame with Hamming window. The frame rate is 10 msec. To extract wavelet features, we use a Gaussian window.

Fig. 7 shows the shape of the mother wavelet of size 20 msec. To construct the filter-bank-type wavelet feature, we use different analysis window width for each element. In this case, We extract one feature using the mother wavelet in 20 msec window, one feature using the $\frac{1}{16^{1/15}}$ daughter wavelet, one feature using the $\frac{1}{16^{2/15}}$ daughter wavelet, ... one feature using the $\frac{1}{16^{15/15}}$ daughter wavelet in 5 msec window, resulting in 16 dimensional vector.

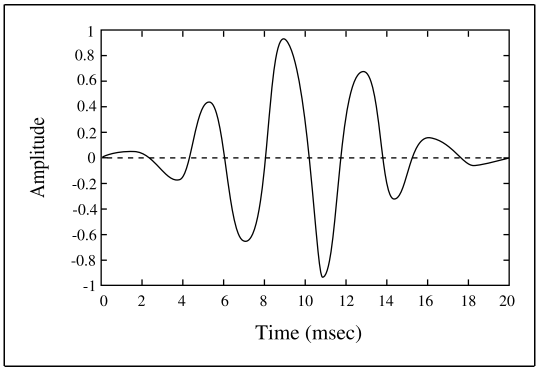

Fig. 7. The shape of mother wavelet.

We transform all the feature vectors into aligned vectors using the eigenvector matrix of the sample covariance matrix.

Fig. 8 shows the inter-class separability of all the features as a function of the number N of the hypercubes in $\mathbf{S}$. Clearly, the mel-scale FFT, and mel-scale cepstrum feature are better than the cepstrum and wavelet feature since they reach $s = 1$ at smaller N.

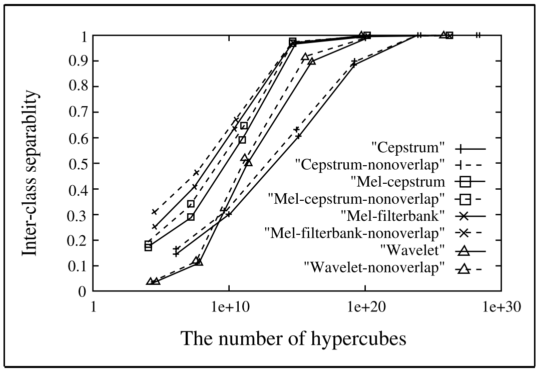

Fig. 8. The inter-class separability of various speech features *vs.* the number of hypercubes.

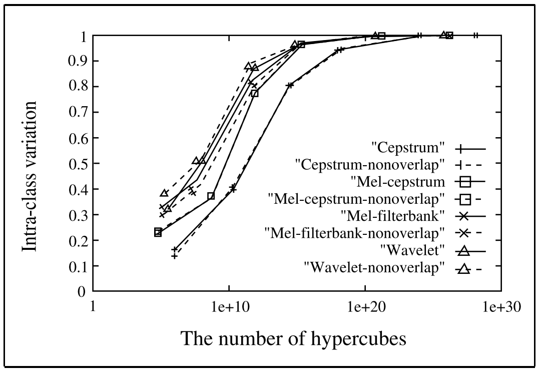

Fig. 9. The intra-class variation of various speech features *vs.* the number of hypercubes.

Fig. 9 shows the intra-class variation of all the features as a function of N. Interestingly, the wavelet feature shows an opposite trend in the comparison between the overlapped and unoverlapped data set. The intra-class variations of the other features on the overlapped data set is larger than those on the unoverlapped data set. This can be explained as follows: The conditional entropy $H(\mathbf{S}|\mathbf{C})$ in (9) does not decrease much even when the overlapped data samples are excluded, while H_{max} decreases as the number of samples decreases.

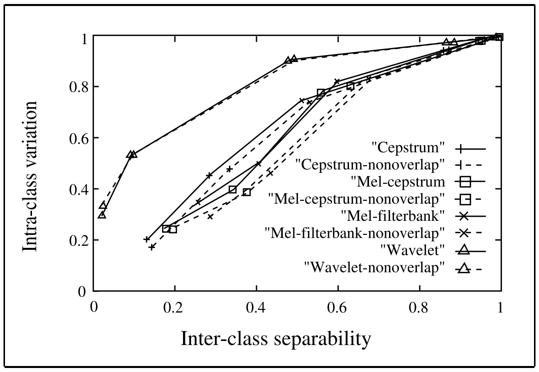

Fig. 10. The intra-class variation *vs.* the inter-class separability.

We show the relation between the inter-class separability s and the intra-class variation v of different number of hypercubes in Fig. 10. Clearly, all features hit the point (1, 1). This means that all classes can be separated only when every sample points fall into distinct hypercubes. Thus, we need more samples to get more precise result. The mel-scaled FFT and the mel-cepstrum are good speech features since their inter-class separabilities become 1 with smaller number of N and their intra-class variation is relatively smaller than the other features. However, we can easily see that the mel-scaled FFT is slightly better than the mel-cepstrum since the inter-class separability is slightly larger.

Summarizing all the above results, we can say that the mel-scaled FFT feature is good for speech recognition.

IV. CONCLUSION

In this paper, we proposed a method to select a good speech feature for recognition. We introduce the linear separability and the Fisher criterion, and propose a new method using Shannon's information theory. We showed that the proposed method is applicable to multi-class non-convex distributions including the speech features, while the other two methods are not. We compared cepstrum, mel-cepstrum, mel-scaled spectrum, and a wavelet feature of the TIMIT data. The result showed that the mel-scaled FFT feature is the best feature for speech recognition among them.

ACKNOWLEDGMENT

This research is funded by the Ministry of Information and Communication, Republic of Korea.

APPENDIX

In this Appendix, we prove (6) and (7). Toward this purpose, let

$$\Lambda_m = \{n \mid S_n \subset C_m\}, \; m = 1, 2, \ldots, M,$$
$$\Lambda_c = \{n \mid S_n \text{ contains elements of more than two classes}\},$$

and assume that Λ_c is not empty. Then, the mutual information is

$$I(\mathbf{S};\mathbf{C}) = H(\mathbf{C}) - H(\mathbf{C}|\mathbf{S})$$

$$= H(\mathbf{C}) - \sum_{n=1}^{N} P(S_n) H(\mathbf{C}|S_n).$$

But

$$H(\mathbf{C}|S_n) = \begin{cases} 0 & \text{if } n \in \Lambda_m \text{ for some } m \\ -\sum_{m=1}^{M} P(C_m|S_n) \log P(C_m|S_n) & \text{if } n \in \Lambda_c \end{cases}.$$

Thus,

$$I(\mathbf{S};\mathbf{C}) = H(\mathbf{C}) + \sum_{n\in\Lambda_c} P(S_n) \sum_{m=1}^{M} P(C_m|S_n) \times \log P(C_m|S_n). \qquad (10)$$

Since $H(\mathbf{C}|S_n) > 0$ for $n \in \Lambda_c$ and Λ_c is not empty,

$$I(\mathbf{S};\mathbf{C}) < H(\mathbf{C})$$

which is the case when the classes overlap.

If the probability of overlapping increases, one can say that the pattern set becomes more complicated. In this case, $\sum_{n\in\Lambda_c} P(S_n)$ increases, and $I(\mathbf{S};\mathbf{C})$ decreases. When $P(S_n)$ is fixed, one can say that the pattern set becomes more complicated if the number of classes in a hypercube increases. In this case, $I(\mathbf{S};\mathbf{C})$ decreases more since $H(\mathbf{C}|S_n)$ become larger for $n \in \Lambda_c$. Based on these facts, we argue that $I(\mathbf{S};\mathbf{C})$ decreases more if the shape of overlapping becomes more complicated.

In particular, if $P(C_m, S_n) = P(C_m)P(S_n)$, $m = 1, 2, \ldots, M$ for $n \in \Lambda_c$, it is easy to see that

$$H(\mathbf{C}|S_n) = H(\mathbf{C}), \qquad n \in \Lambda_c. \qquad (11)$$

With (11) in (10), we have

$$I(\mathbf{S};\mathbf{C}) = [1 - \sum_{n\in\Lambda_c} P(S_n)]\, H(\mathbf{C}),$$

where $\sum_{n\in\Lambda_c} P(S_n)$ is the probability of overlapping.

REFERENCES

[1] L. R. Rabiner and C. E. Schmidt, "Application of dynamic time warping to connected digit recognition," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-28, pp. 377-388, Aug. 1980.

[2] C. S. Myers and L. R. Rabiner, "Connected digit recognition using a level-building DTW algorithm," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-29, pp. 351-363, Jun. 1981.

[3] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. ASSP-32, pp. 263-271, 1984.

[4] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *The Bell System Technical Journal*, vol. 62, pp. 1075-1105, Apr. 1983.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition," *Proc. IEEE*, vol. 77, pp. 257-286, Feb. 1989.

[6] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker independent continuous speech recognition," *IEEE Trans. Acoust., Speech, Signal Proc.*, vo. 38, no. 4, pp. 599-609, 1990.

[7] J. Picone, "Continuous speech recognition using hidden Markov models," *IEEE ASSP Magazine*, pp. 26-41, Jul. 1990.

[8] M. Woszczyna *et al*., "Janus 93: Towards spontaneous speech translation," *Proc. ICASSP'94*, vol. I,

pp. 345-348, 1994.

[9] P. Jeanrenaud *et al.*, "Reducing word error rate on conversational speech from the switchboard corpus," *Proc. ICASSP'95*, vol. 1, pp. 53-56, May 1995.

[10] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 29, pp. 254-272, Mar. 1981.

[11] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 28, pp. 357-366, July 1980.

[12] O. Siiohan, "On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition," *Proc. ICASSP'95*, pp. 125-128, Apr. 1995.

[13] M. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic Press, 1972.

[14] J. A. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Redwood city, CA: Addison-Wesley, 1991.

[15] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 2nd edition, 1984.

[16] M. Bichel and P. Seitz, "Minimum class entropy: A maximum information approach to layered networks." *Neural Networks*, vol. 2, pp. 133-141, 1989.

[17] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*. Cambridge, MA: MIT Press, 1986.

[18] L. R. Bahl *et al.*, "Robust methods for using context-dependent features and models in a continuous speech recognizer," *Proc. ICASSP'94*, vol. I, pp.533-536, Apr. 1994.

[19] A. Waibel and B. Yegnanarayana, "Comparative study of nonlinear time warping techniques in isolated word speech recognition systems," *Tech. Rep., Carnegie-Mellon Univ.*, June 1981.

Youngjik Lee received the B.S. degree in Electronics Engineering form Seoul National University, Seoul, Korea in 1979, the M.S. degree in Electrical Engineering from Korea Advanced Institute of Science, Seoul, Korea in 1981, and the Ph.D. degree in Electrical Engineering from the Polytechnic University, Brooklyn, New York, U.S.A.

From 1981 to 1985 he was with Samsung Electronics Company, Suwon, Korea where he was involved in the development of video display terminal. From 1985 to 1988 his research topic was concentrated on the theories and applications of sensor array signal processing. Since 1989, he has been with Research Department and Spoken Language Processing Section of ETRI, Taejon, Korea pursuing interests in theories, implementations, and applications of spoken language translation, speech recognition and synthesis, and neural networks.

Kyu-Woong Hwang received B.S. degree in Electronics Engineering at Seoul National University, Seoul, Korea in 1991 and M.S. degree in Electronics Engineering at Korean Advanced Institute of Science and Technology in 1993. He joined ETRI in 1993 where he works for Spoken Language Processing Section. His interests include speech recognition, artificial neural networks, and pattern recognition.