Feature selection algorithms for the creation of multistream speech recognizers

Yotaro Kubo[1]*, Shigeki Okawa*[2]*, Akira Kurematsu*[1]*, Katsuhiko Shirai*[1]

[1]Department of Computer Science and Engineering, Waseda University, Tokyo, Japan
[2]Chiba Institute of Technology, Narashino, Japan

yotaro@shirai.cs.waseda.ac.jp

Abstract

In this paper, we present a method to split a feature stream into multiple feature streams.

The efficiency of ensemble classifiers for speech recognition is confirmed by several experiments. The conventional methods for constructing multiple classifiers are done by splitting the feature stream by type of features or subbands where the features are associated. The splitting approach is well suited for obtaining high-dimensional features because it naturally leads to dimension reduction of features.

In order to take advantage of ensemble classifiers, each classifier should compensate for the errors due to the other classifiers. Because every streams depends on phonetic information in clean environments, the difference on noise robustness can be measured by using independency. Therefore, each classifier should be independent from others in noisy environments. We proposed a method to split a feature stream using stream independency criteria in order to constructing independent classifiers.

We evaluated several stream splitting methods and compare word error rate by conducting continuous digit recognition experiments on noisy speech. Our method can reduce 30.9% of the word error when compared with the single classifier method, while it reduces 3.2% of the word error when compared with conventional multistream approach.

Index Terms: Multistream speech recognition, ensemble classifiers, mutual information

1. Introduction

The efficiency of multistream speech recognizers based on ensemble classification has been confirmed by several experiments [1, 2]. The conventional methods for constructing ensemble classifiers employ classifiers that only process a particular subset of features.

The multiband approach is one of the ensemble creation methods in which the features associated with a particular subbands are selected [1]. Because environmental noise is often present in limited subbands, multiband approach can reduce the effect of noise.

The multistream approach is an alternative approach for ensemble creation; in this method, the features are selected according to their types [2]. This approach can be applied when the original features include heterogeneous features. Because the features derived by other analysis methods should be robust against different type of noise, multistream feature selection can be suitably used. All conventional methods of ensemble creation for speech recognition are based on prior knowledge of features.

In conventional pattern recognition problems, more complex feature selection methods are employed for the creation of ensemble classifiers. For example, Opitz *et al.* and César *et al.* proposed a genetic-algorithm-based searching of feature subsets, which minimizes the classification error of ensemble classifiers [3, 4]. We call this approach 'risk minimization approach'. However, it is difficult to apply this approach to pattern recognition problems that requires a large data set and complicated models such as speech recognition. Because the number of possible subsets is very large, measuring the classification error might take a very long time.

In this paper, we propose an intermediate approach between the prior knowledge approach and the risk minimization approaches for the creation of ensembles. In the intermediate method, the selection of features is determined optimally by using a statistical distribution of the behavior of features. The evaluation of classification errors is not required in this approach. Therefore, the optimization algorithm makes only one path through the data set.

One of the advantages of using ensemble classifiers is that errors from one classifier can be compensated by using the results of other classifiers. Therefore, it is important that at least a particular subset of features is error-free. We evaluate this property by using the mutual entropy between subsets of features.

In this paper, we introduce our feature extraction system and the technique to merge the results of classifiers in section 2. In section 3, we propose a method to determine feature subset. The experimental setup is presented in section 4, and the results are discussed in section 5.

2. Features and decoders

In this section, we describe the speech recognition system used in this study. Figure 1 shows the block diagram of the ASR system used.

2.1. Feature extraction

We use emphasised amplitude modulation (AM) components and frequency modulation (FM) components as features of classification.

First, the Bark filterbank [5] is applied in order to separate the signal into narrowband signals. Then, the amplitude of the n^{th} frame $a(n)$ is defined by the energy of short-time segment, and spectral centroid $f(n)$ is defined by the number of zero-crossing points in the short time segment.

We emphasise $a(n)$ and $f(n)$ by using MLP-OL [6]. MLP-OLs are used to extract the significant modulation components from the input signals. MLP-OL is the general MLP classifier during the training phase (Figure 2).

The input vector x of MLP-OLs are

$$x_i = a\left(n + i - \frac{L+1}{2}\right) \quad (1)$$

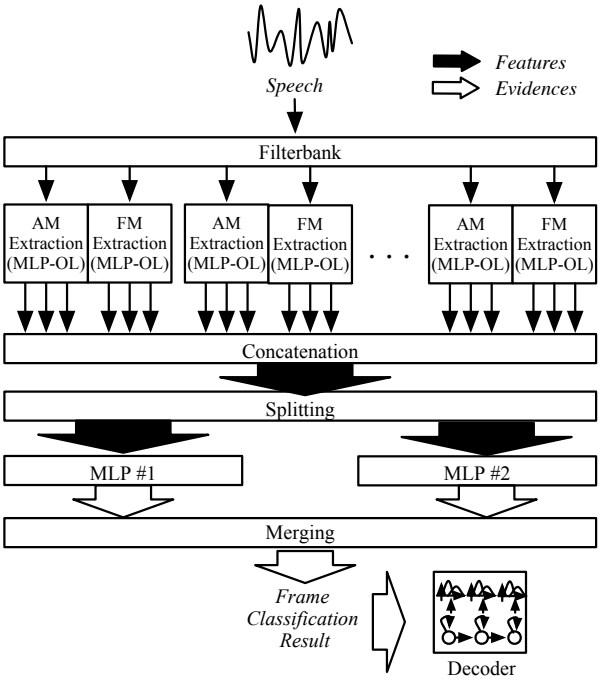

Figure 1: *System diagram*

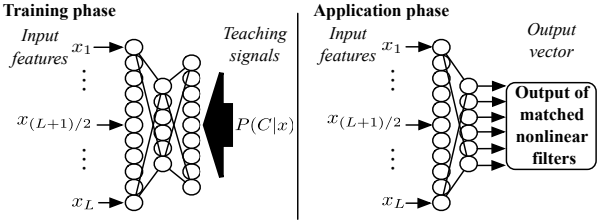

Figure 2: *Diagram of MLP-OL.*

and

$$x_i = f\left(n + i - \frac{L+1}{2}\right) \quad (2)$$

for AM and FM emphasis, respectively.

As is typical for the MLPs trained to estimate the posterior probabilities over monophones, all the MLPs are trained using the teaching signal, which is '1.0' for the monophone associated with the central frame (n) and '0' for all the others. We employ the standard error back-propagation algorithm to optimize the weights of the connections between layers so that the mean squared error is minimized.

During the application phase, the output layer of the MLP is removed. Because the input vector x_i can be interpreted as the time-series signal, the output of hidden neurons can be interpreted as the convolution of x and the weights between the input neurons and the hidden neuron with a nonlinear sigmoid function. Therefore, the output of the hidden neurons has a fixed frequency response that can improve the distinguishability of x. The filter constructed using the above procedure is called a 'matched filter'.

We use the concatenation of the output vectors of all the matched filters as an original feature vector.

2.2. Classifiers

We use MLPs as a base classifier of ensemble classifiers. Each MLP considers only a subset of the features. MLPs are trained using the standard error back-propagation algorithm. We use monophones as the target classes of classification. Every MLP in an ensemble has the same number of parameters.

2.3. Merging and decoding

We observed multiple streams of the classification results. According to the tandem-approach of acoustic modeling [7], we decode these using Gaussian mixture hidden Markov models (GM-HMMs).

In order to reduce the errors due to unreliable classifiers, we use the entropy-based combination of tandem acoustic models, which was introduced by Ikbal *et al.* [2]. This method can suppress the evidence from unreliable classifiers by weighting them with conditional entropies of the classification results.

First, we calculate the approximate posterior probability for the m^{th} MLP $p(c|x^m)$ as follow:

$$p(c|x^m) = \frac{(\exp(-y^m_{i(c)}) - 1)^{-1}}{\sum_{d \in C}(\exp(-y^m_{i(d)}) - 1)^{-1}}. \quad (3)$$

Here, x^m is the input vector of the m^{th} MLP; y^m, the output vector of the m^{th} MLP; C, the set of target classes (in this study, C is the set of monophones); and $i(c)$, the mapping from the elements of C to the dimension index of y^m.

In concrete terms, x^1 is the AM feature vector; x^2, the FM feature vector; y^1, the output of the AM classifier; and y^2, the output of the FM classifier.

The conditional entropy of x^m is estimated as follow:

$$H^m(C|x^m) = \sum_{c \in C} -p(c|x^m) \log p(c|x^m). \quad (4)$$

The weights of y^m are determined by taking the inverse of the conditional entropy h^m:

$$w^m = \frac{\{H^m(C|x^m)\}^{-1}}{\sum_{j=1}^{M}\{H^j(C|x^j)\}^{-1}}, \quad (5)$$

where M is the number of MLPs (in this study, M is 2).

Finally, we obtain the merged output $\hat{y}$ as follows:

$$\hat{y_i} = \sum_{m=1}^{M} w^m \log(y^m_i). \quad (6)$$

We decode a sequence of frame-by-frame classification results using diagonal GM-HMMs. Because the dimensions of $\hat{y}$ correlate with each other, $\hat{y}$ is incompatible with the diagonal GM-HMMs. It is necessary to transform $\hat{y}$ for dimensionality reduction and decorrelation. For this purpose, we use the Karhunen-Loeve transformation (KLT). The number of features for GM-HMMs is equal to the number of monophones.

3. Independent stream determination

In this section, we propose a method to select the features. In this method, features are selected by splitting original features

into groups using independency criteria so that every feature belongs exactly to one group.

We denote the feature group by $X = \{x^1, \cdots, x^M\}$, where M is the number of classifiers, $x^m = \{x_1^m, \cdots, x_N^m\}$ is a set of features, and x_n^m is a single feature.

First, we measure the independency by using the performance function $Q(X)$:

$$Q(X) = H(X) - \sum_m H(x^m) \quad (7)$$

where $H(X)$ represents the entropy X.

Because $H(X)$ is constant by varying the grouping rule X, we omit $H(X)$ and reformulate the performance function $Q(X)$ as follows:

$$Q(X) = -\sum_m H(x^m) \quad (8)$$

$$= \sum_m - \left[\sum_n H(x_n^m) - \left\{ \sum_{g \in \mathcal{P}(x^m)} (N(g)-1) M(g) \right\} \right] \quad (9)$$

$$M(g) = I(g_1; ...; g_{N(g)}), \quad (10)$$

where $N(g)$ is the number of elements in the set g, $\mathcal{P}(x^m)$ is the power set of x^m, and $M(g)$ denotes the mutual information across the members of g.

Because $H(x_n^m)$ is a constant, we finally obtain the following performance function:

$$Q(X) = \sum_m \sum_{g \in \mathcal{P}(x^m)} (N(g)-1) M(g). \quad (11)$$

It is difficult to solve this optimization efficiently because it implies the estimation of the multivariate mutual entropy. However, by using the property of $M(g)$ that denotes the mutual information between the elements in the group, the maximization problem can be solved by the clustering algorithm, which uses the distance function based on independency criteria.

There are several methods to measure the independency of two variables. Because we would like to argue about the independency of behavior, we require a measure that indicates the independency between non-stationary signals. We employ the method proposed by Ando *et al.*, which can deal with non-stationary time-series variables [8].

The distance function $D(x_1, x_2)$ is defined as follows:

$$D(x_1, x_2) = \left\{ \sum_t \log E_\tau [\tilde{x_1}(t,\tau) \cdot \tilde{x_2}(t,\tau)^T] \right\}^{-1} \quad (12)$$

$$\tilde{x_n}(t,\tau) = x_n(t+\tau) - \overline{x_n}(t), \quad (13)$$

$$\overline{x_n}(t) = E_\tau [x_n(t+\tau)]. \quad (14)$$

Here, x denotes the original feature vectors and $E[.]$ denotes the operations that measure the empirical expectation by varying $\tau \in [-T, T]$. In [8], it is suggested that T must be determined so that observed signals can be assumed as quasi-stationary in segments $x_n(t+\tau)$. We determined $T = 25$.

We apply the aggregative clustering method (group average method) using the distance function D and obtaining feature group X.

Table 1: *Word error rates of compared methods.*

System Description	Word Error Rate (%)	Baseline Improv. (% Rel.)
Single	17.15	-
Multiband	13.01	24.14
Multistream	12.24	28.63
Independent-2	11.85	30.90
Independent-4	11.63	32.19

4. Experimental setup

In this section, we evaluate the performance of the proposed method. We conducted continuous digit recognition of noisy speech in this experiment.

The training set and test set are taken from CENSREC-1 [9], which is the Japanese translation of the AURORA-2 data set. The training set used for MLP, HMM, and the ensemble creator comprises 4,220 utterances of noisy speech. The test set comprises 4,004 utterances, which are disturbed by realistic environmental noise at 10 dB.

Because our system uses heterogeneous features and band-limited features, there are many rational methods for selecting the features and creating ensemble classifiers. In an experiments, we compared several ensemble creation methods that are described below:

- Single
  Constructing single classifier.
- Multiband
  Constructing low-band features classifier and high-band features classifier.
- Multistream
  Constructing AM features classifier and FM features classifier.
- Independent-2
  Constructing ensembles using the proposed method. The number of classifiers is set to 2.
- Independent-4
  Constructing ensembles using the proposed method. The number of classifiers is set to 4.

The sample rate of the speech signals in the experiments is fixed to 8,000 Hz. Therefore, the Bark filterbank separates the original signal into 14 filtered signals. All MLP-OLs are constructed with 20 hidden neurons; therefore the number of original features is fixed to 560 (product of the number of filters and the number of hidden neurons for the AM and FM.)

5. Discussions

Table 1 shows the word error rates of the compared methods.

The table shows that our proposed method achieves the highest accuracy when compared with conventional ensemble creation methods. Our method can reduce 30.9% of the word error when compared with the single classifier method, and reduce 3.2% of the word error when compared with the conventional multistream method. It is confirmed that the independence between two streams is important for accurate speech recognition.

The results of the multistream speech recognizer which defines streams using type of features are more accurate than the

Table 2: *Similarity indices of feature subsets between subset determined by proposed method and subset defined in conventional methods.*

Method	Similarity Index (%)
Multiband	52.86
Multistream	77.14

results of the multiband speech recognizer. We consider that this difference of accuracy also due to the independency of each stream. In order to compare the conventional selection methods with the proposed method, we define similarity index S of ensemble classifiers as follows:

$$S(X;Y) \quad = \quad \frac{N(X^1 \cap Y^1)}{N(X^1)}, \tag{15}$$

where $N(X)$ denotes the number of elements in the set X and X^1 is the subset of features associated with the first classifier.

Table 2 shows the similarity indices between the feature subset determined by the proposed method (Independent-2) and the subset defined in conventional methods.

The table shows that the multistream approach is more similar to the proposed method than the multiband approach. It is confirmed that the independency between two streams of the multistream speech recognizer is higher than that of the multiband speech recognizer.

From the relation between the similarity indices and the word error rates, it is confirmed that the independent behavior of feature subsets is important for ensemble creation.

6. Conclusions

In this paper, we proposed a method to optimize feature streams based on the independency criterion in order to obtain accurate ensemble classifiers. We understood that the independency of two streams is guaranteed by maximizing the dependency between the components in each stream. Then, we searched the dependent components by aggregative clustering using the distance measure, which was defined using the summation of segmental correlations.

We confirmed that our method can reduce 30.9% of the word error when compared with the single classifier method and 3.2% of the word error when compared with the conventional multistream method. We also confirmed that similar splitting method results in higher accuracy. Therefore, independent behavior of feature subsets is important for ensemble creation.

7. Acknowledgements

This research was supported by "Ambient SoC Global COE Program of Waseda University" of the Ministry of Education, Culture, Sports Science and Technology, Japan, and the Advanced Research Institute for Science and Engineering, Waseda University, under the project "Research on Multi-Modal Human Interface Aiming for Hybrid Smart Room System."

8. References

[1] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-Band Speech Recognition in Noisy Environments," Proc. ICASSP-98, pp. 641–644, Seattle, Washington, USA, May 1998.

[2] S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, and H. Bourlard, "Entropy Based Combination of Tandem Representations for Noise Robust ASR," Proc. INTERSPEECH-ICSLP-2004, pp. 2553–2556, Jeju Island, Korea, October 2004.

[3] D.W. Opitz, "Feature Selection for Ensembles," Proc. of 16^{th} National Conference on Artificial Intelligence, pp. 379–384, Palo Alto, California, USA, March 1999.

[4] C. Guerra-Salcedo, D. Whitley, "Genetic approach to feature selection for ensemble creation," Proc. of the Genetic and Evolutionary Computation Conference, pp. 236–243, Orlando, Florida, USA, July 1999.

[5] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," Journal of the Acoustical Society of America, Vol. 87, pp. 1738–1752, April 1990.

[6] B. Chen, S. Chang, S. Sivadas, "Learning Discriminative Temporal Patterns in Speech: Development of Novel TRAPS-like Classifiers," Proc. Eurospeech, pp. 429–432, Geneve, Switzerland, September 2003.

[7] N. Morgan, H. Bourlard, "An Introduction to the Hybrid HMM/Connectionist Approach," IEEE Signal Processing Magazine, pp. 25–42, May 1995.

[8] A. Ando, M. Iwaki, "Blind Separation of Nonstationary Sources by Block Decorrelation of Output Signal," Technical Report of IEICE (EA), Vol. 9, No. 57, September 2002.

[9] CENSREC-1: `http://sp.shinshu-u.ac.jp/CENSREC/ja/CENSREC/AURORA-2J/`