MINIMUM DETECTION ERROR TRAINING FOR ACOUSTIC SIGNAL MONITORING

Hideyuki Watanabe, Yuji Matsumoto and Shigeru Katagiri

ATR Human Information Processing Research Laboratories
2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-02, JAPAN
e-mail: {hwatana, yuji, katagiri}@hip.atr.co.jp

ABSTRACT

In this paper we propose a novel approach to the detection of acoustic irregular signals using Minimum Detection Error (MDE) training. The MDE training is based on the Generalized Probabilistic Descent method, which was originally developed as a general concept for discriminative pattern recognizer design. We demonstrate its fundamental utility by experiments in which several acoustic events are detected in a noisy environment.

1. INTRODUCTION

We are usually surrounded by various kinds of acoustic events such as speech, traffic noise, bird songs, and music. Detection of (unexpected) irregular sounds in such a daily life environment is an important application of intelligent signal processing.

Acoustic signal monitoring has long been studied in the framework of machine failure monitoring [1, 2, 3], and its importance is growing in today's information society where the application range of acoustic signal monitoring now extends to medical care and security control, as well as the traditional manufacturing application. The main procedures of acoustic signal monitoring are to 1) segment an observed signal to acoustic events and 2) classify the events as either regular classes or irregular classes. The larger the size of classes that a system can handle, the more intelligent and useful the system. A desirable monitoring system should achieve the minimum detection (both detection-failure and false-alarm) error status for the irregular classes.

Conventionally, monitoring systems have been developed by simply combining the existing signal processing techniques of spectrum analysis and pattern recognition, and no design efforts have been applied that enable one to achieve the minimum error condition [1, 2]. Consequently, the systems work well to some extent but there is still much room for improvement in their design procedures. Motivated by this concern, we introduce in this paper a novel solution to a design problem of acoustic signal monitoring systems, which we refer to as Minimum Detection Error (MDE) training. The proposed method uses the design concept of Minimum SPotting Error training (MSPE) [4], which was developed based on the Generalized Probabilistic Descent method [5] for spoken keyword spotting, and newly incorporates several techniques suited for the online operation of acoustic signal monitoring.

There are many possibilities, in terms of system structure selection, for MDE implementation. Among them, we specially study in this paper a monitoring system based on a kernel-based neural network in the task of detecting irregular sounds observed in a computer room.

2. FORMALIZATION OF ACOUSTIC SIGNAL MONITORING

Let us assume $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ ...\ \mathbf{x}_t\ ...]$ to be an input (to a detection system) that contains acoustic events to be detected and is a sequence of D-dimensional acoustic feature vectors, where $\mathbf{x}_t$ is the acoustic feature vector at time (frame) index t. One should also notice that $\mathbf{X}$ is a right-endless sequence. Our task is to detect events of S classes of *regular* acoustic signal ($\mathbf{C}_s, s = 1, 2, ..., S$), assuming that the portion of input that is not detected is considered to belong to *irregular* class ($\hat{\mathbf{C}}$). We consider performing the task by using the procedure illustrated in Fig. 1; That is, we set a *signal detector* for each regular class, make a classification decision at every time t in the class-by-class mode, and finally emit irregular events.

The proposed procedure looks indirect with respect to the goal of irregular event detection. However, it is not realistic to model all the unexpected irregular phenomena, and thus the strategy of Fig. 1 is a good practical solution to the problem.

The quality of the monitoring procedure mainly relies on the detection accuracy of regular class events. The detection is basically equivalent to the process of

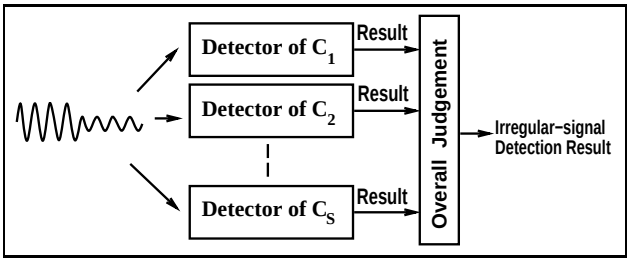

Figure 1: Irregular-signal Detection System

keyword spotting, and it is formalized as follows. We first define a block (a partial sequence of the acoustic feature vectors) $\mathbf{B}_t = [\mathbf{x}_{t-L_1} \ldots \mathbf{x}_t \ldots \mathbf{x}_{t+L_2}]$ around $\mathbf{x}_t$, where L_1 and L_2 are positive natural numbers. To measure the degree to which $\mathbf{B}_t$ belongs to class s, we next introduce a discriminant function. Since we consider a kernel-based network system, the discriminant function naturally outputs the distance measure between $\mathbf{B}_t$ and a kernel-based model of $\mathbf{C}_s$, λ_s, and it is denoted as $g_s(\mathbf{B}_t, \lambda_s)$. Due to the temporal structure of blocks and models, the nonlinear time warping mechanism, which is often used in speech pattern recognition, is incorporated in the distance calculation. Then, the rule of detection is expressed as

$$\begin{aligned} h_s \geq g_s(\mathbf{B}_t, \lambda_s) \quad &\Rightarrow \quad \mathbf{x}_t \text{ is classified as } \mathbf{C}_s, \\ h_s < g_s(\mathbf{B}_t, \lambda_s) \quad &\Rightarrow \quad \mathbf{x}_t \text{ is not classified as } \mathbf{C}_s, \end{aligned} \tag{1}$$

where h_s is a preset threshold for class s. It is clear here that the parameter set, $\mathbf{\Lambda}_s = \{\lambda_s, h_s\}$, determines the detection quality for class s, and our design target accordingly comes to find a state of $\mathbf{\Lambda}_s$ that results in the minimum detection error condition. Figure 2 illustrates the detection procedure.

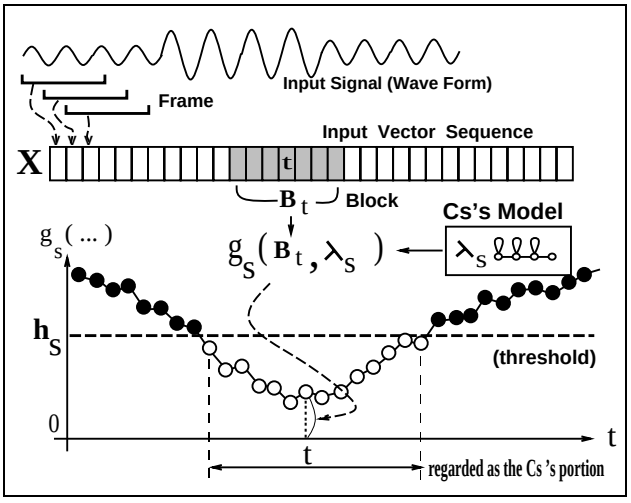

Figure 2: Procedure of Signal Detection

3. MINIMUM DETECTION ERROR TRAINING

3.1. Basic Formalization

The MDE training is applied individually to each regular-class signal detector. A key idea of the training is to formalize the entire detection operation and evaluation process as a *smooth* function of the trainable parameter set of detector $\mathbf{\Lambda}_s$ (first differentiable function with regard to $\mathbf{\Lambda}_s$) [4, 5].

The first step of the formalization is to express the detection decision of (1) by a smooth function of $\mathbf{\Lambda}_s$. To do this, we define a *detection measure* as

$$d_s(\mathbf{B}_t, \mathbf{\Lambda}_s) \quad = \quad h_s - g_s(\mathbf{B}_t, \lambda_s). \tag{2}$$

From (1), it can be noticed that $d_s(\cdot) \geq 0$ means that $\mathbf{x}_t$ is classified as $\mathbf{C}_s$, and $d_s(\cdot) < 0$ means that $\mathbf{x}_t$ is not classified as $\mathbf{C}_s$. Obviously, $d_s(\cdot)$ is a smooth function of $\mathbf{\Lambda}_s$.

In the signal detection, there are two types of errors: (1) *detection-failure (DF) error*, which corresponds to a failure to detect an existing event, and (2) *false-alarm (FA) error*, which corresponds to the detection of a false event. In the second step of the formalization, we incorporate these errors to the definition of loss that is used for evaluating detection decision results in the training stage. Since there are two types of errors, a loss is defined as

$$\ell_s(\mathbf{B}_t, \mathbf{\Lambda}_s, \alpha_s) = \frac{1}{1 + \exp\left(\alpha_s \cdot d_s(\mathbf{B}_t, \mathbf{\Lambda}_s)\right)}, \tag{3}$$

$$\alpha_s = \begin{cases} \alpha_{s,1} > 0 & \text{for } \mathbf{x}_t \in \mathbf{C}_s, \\ \alpha_{s,2} < 0 & \text{for } \mathbf{x}_t \notin \mathbf{C}_s. \end{cases}$$

Notice here that the smooth DF-error loss $\ell_s(\cdot, \alpha_{s,1})$ and the smooth FA-error loss $\ell_s(\cdot, \alpha_{s,2})$ approximate the DF error counts and the FA error counts, respectively.

The final stage of the formalization is to provide a training procedure for $\mathbf{\Lambda}_s$. Our training target is clearly to find the optimal status that leads to the minimum detection error condition over training input. Based on the probabilistic descent theorem [5, 6], we use the following adjustment rule for every feature vector $\mathbf{x}_t$:

$$\mathbf{\Lambda}_s^{(t+1)} \quad = \quad \mathbf{\Lambda}_s^{(t)} - \varepsilon_t \mathbf{U}_s \nabla_{\Lambda_s} \ell_s(\mathbf{B}_t, \mathbf{\Lambda}_s^{(t)}, \alpha_s), \tag{4}$$

where ε_t, called learning coefficient, is a positive, monotonically-decreasing (with regard to t) constant, $\mathbf{U}_s$ is a positive-definite matrix, ∇_{Λ_s} denotes the derivative (gradient) with regard to $\mathbf{\Lambda}_s$, and $\mathbf{\Lambda}_s^{(t)}$ represents the parameter status before updating at time t.

An infinite run of the above adjustment over an endless input is proved to lead almost surely to the (at

least local) minimum of the detection error rate in the sense of the following definition [5, 6]:

$$\text{Detection Error Rate} = \frac{\text{No. of DF errors} + \text{No. of FA errors}}{\text{Total No. of Frames } (t)}. \quad (5)$$

In a realistic training condition where only a finite run is possible, the adjustment approximates the minimum detection error status over the available length of the training input signal, or set of signals.

For clarity, let us re-explain in the final paragraph of this section that the signal monitoring system first uses the above-trained signal detectors for detecting regular class events and then emits, based on the result of the regular class event detection, the target events of irregular class.

3.2. Extension Using DMD

In the basic formalization, we considered the optimization (training) only of the model parameters and the threshold, $\{\lambda_s, h_s\}$. However, since the feature vectors are separately computed, the input sequence $\mathbf{X}$ and the blocks $\mathbf{B}_t$ $(t = 1, 2, ...)$ are not necessarily optimal for accurate detection decision. Improvement of the feature representation of the input would be useful for increasing the system performance. In this light, based on the Discriminative Metric Design (DMD) [7], we apply a feature transformation operator $\mathcal{T}_s(\cdot; \varphi_s)$ to $\mathbf{B}_t$, where φ_s is a trainable parameter used for transformation, and aim at jointly optimizing φ_s and $\{\lambda_s, h_s\}$ with the single objective of detection error minimization. This extended training enables one to find a useful, detection-oriented feature representation for every regular class, resulting in a more accurate detection of regular and irregular events.

4. IMPLEMENTATION EXAMPLE USING KERNEL-BASED NETWORK

4.1. Implementation

As an implementation example, we designed a monitoring system that computed filter-bank power spectrum feature vectors and consisted of kernel-based signal detectors. An input here is a sequence of D-dimensional feature vectors, where each vector component is a logarithmic power spectrum coefficient that is calculated with its corresponding narrow-band filter. Similar to [8], the detector has a state-transition structure and assigns multiple kernels (prototypes or reference vectors) to each state. Then, for $\mathbf{C}_s$, the discriminant function is defined as

$$g_s(\mathbf{B}_t, \lambda_s) = \min_{\mathbf{X}_{T_1}^{T_2} \subset \mathbf{B}_t} \min_{\Theta_{T_1}^{T_2}} \mathcal{D}_s(\mathbf{X}_{T_1}^{T_2}, \lambda_s, \Theta_{T_1}^{T_2}), \quad (6)$$

$$\mathbf{X}_{T_1}^{T_2} = [\ \mathbf{x}_{T_1}\ ...\ \mathbf{x}_t\ ...\ \mathbf{x}_{T_2}\ ] \subset \mathbf{B}_t,$$

$$\Theta_{T_1}^{T_2} = [\ \theta_{T_1}\ ...\ \theta_t\ ...\ \theta_{T_2}\ ] \quad (\theta_t \in \{1, 2, ..., N\}),$$

$$\mathcal{D}_s(\mathbf{X}_{T_1}^{T_2}, \lambda_s, \Theta_{T_1}^{T_2}) = \frac{1}{T_2 - T_1 + 1} \sum_{\tau=T_1}^{T_2} \delta_s(\mathbf{x}_\tau, \lambda_{s,\theta_\tau}),$$

where $\mathbf{X}_{T_1}^{T_2}$ is a subsequence included in $\mathbf{B}_t$ (it is assumed that $\mathbf{X}_{T_1}^{T_2}$ involves $\mathbf{x}_t$), $\Theta_{T_1}^{T_2}$ is a sequence of the state indices, $\delta_s(\mathbf{x}_\tau, \lambda_{s,j})$ is a local distance measure of $\mathbf{x}_\tau$ at the j-th state of the $\mathbf{C}_s$'s model, and $\mathcal{D}_s(\mathbf{X}_{T_1}^{T_2}, \lambda_s, \Theta_{T_1}^{T_2})$ is an accumulated distance that is defined as the arithmetic mean of the local distance measures $\delta_s(\cdot)$ along the path $\Theta_{T_1}^{T_2}$. Theoretically, discontinuous operations such as min should not be included in GPD-based training. However, for computational simplification, we use such theoretically-inadequate but convenient operations in our implementation.

To perform the DMD-based transformation optimization, we specially define $\delta_s(\cdot)$ as

$$\delta_s(\mathbf{x}_\tau, \lambda_{s,j}) = \min_m ||\mathbf{W}_{s,j}(\mathbf{x}_\tau - \mathbf{r}_{s,j,m})||^2, \quad (7)$$

where $||\cdot||$ is the Euclidean norm, $\mathbf{W}_{s,j}$ denotes the $D \times D$ feature transformation matrix at the j-th state, and $\{\mathbf{r}_{s,j,m}\}_{m=1}^{M}$ denotes the set of kernels at the j-th state. A trained state of $\{\mathbf{W}_{s,j}\}_{j=1}^{N}$ leads to the most useful features in the sense of MDE training optimization.

Incorporating the signal detectors into the scheme of Fig. 1 results in a kernel-based monitoring system that enables one to detect irregular acoustic events, represented in the filter-bank output features.

4.2. Experimental Evaluation

The task was to detect irregular sound events in a noisy machine room environment. Five input signal sets were recorded at a sampling frequency of 8 kHz. Each set is about 30 sec in length. The first four sets were used for training, and they contain four kinds of acoustic events: background noise (BN), male voice (MV), portable phone ringing (PP) and hand clap (HC). The remaining set was used for testing, and they contain the sound of hitting two screwdrivers against each other (SD) as well as the aforementioned four events. We consequently regarded the SD class as an unknown, irregular signal class in the experiment.

To simulate a 16-channel filter-bank, we calculated FFT-based 16-dimensional, short-time log-power spectral vectors, by moving a 32-ms Hamming time window with an 8-ms shift length. $\mathbf{X}$ was consequently

a sequence of the 16-dimensional, log-power spectral vectors, calculated at every 8 ms.

For the four regular classes, the acoustic models of the signal detector had a 3-state, 1-kernel structure. For the MDE training, the kernels were initialized by using segmental K-means clustering. The clustering was done in a class-by-class mode. All the feature transformation matrices $\{\mathbf{W}_{s,j}\}$ were initialized at the identity matrix. Moreover, for every class, the initial value of the decision threshold h_s was selected through some preliminary analysis of the DF/FA errors produced with the initial configuration of the signal detector model so that the total number of these errors can be close to the minimum.

We summarize the detection error rates for the training data and test data in Table 1. In the table, K-means stands for the segmental K-means clustering; MDE stands for the 200-epoch ("epoch" means one full presentation of training sequence) MDE training without the adjustment of $\{\mathbf{W}_{s,j}\}$; MDE-DMD stands for the 200-epoch MDE training with the adjustment of $\{\mathbf{W}_{s,j}\}$; and "Av" denotes the average error rate over the four regular classes (BN, MV, PP and HC). Here, the error rate was calculated according to (5).

Table 1: Detection error rates (%)

(a) Training Data

	K-means	MDE	MDE-DMD
BN	2.95	2.53	2.31
MV	2.16	0.393	0.237
PP	0.868	0.192	0.128
HC	1.28	0.347	0.374
Av	1.81	0.866	0.762

(b) Test Data

	K-means	MDE	MDE-DMD
BN	3.72	3.20	3.46
MV	2.54	1.56	0.635
PP	0.404	0.289	0.144
HC	0.981	0.577	0.548
Av	1.91	1.41	1.20
SD	2.97	1.59	1.41

For all of the four regular classes, the MDE-trained detectors achieved lower detection error rates than the segmental-K-means-based detectors. Also, in most cases, the detector, for which the feature transformation matrix was further optimized, showed improvement. The superiority of the MDE training is observed over training and more importantly over testing data. The improvement in the regular class event detection causes one to expect a more accurate detection result for the irregular class events. Actually, in (b) of Table 1, we can clearly see improvements in the detection error rates of SD. The MDE training achieved about a 50 % error rate reduction, compared with the K-means-trained system. As a whole, the results clearly demonstrate the utility of the proposed strategy of sound monitoring and the MDE training.

5. CONCLUSION

This paper has introduced a novel approach to acoustic signal monitoring, proposing a new design method called Minimum Detection Error training. It has also demonstrated the utility of the MDE-based approach in an experimental task of detecting irregular sound events in a noisy room environment.

The proposed monitoring method is quite general. One can easily apply the method to various kinds of signal detection and monitoring tasks, and also use any reasonable system structure, such as multi-layer perceptron, in its implementation. The well-formalized mathematical basis of the method can also contribute to making the theoretical framework of intelligent signal processing more sound.

6. REFERENCES

[1] E. Emel and E. Kannatey-Asibu, Jr., "Tool failure monitoring in turning by pattern recognition analysis of AE signals, " *ASME Journ. of Engineering for Industry,* vol. 110, pp. 137-145, May 1988.

[2] L.P. Heck, "Signal processing research in automatic tool wear monitoring," *Proc. ICASSP 93,* vol. 1, pp. 55-58, 1993.

[3] M.J. Dowling, "Application of non-stationary analysis to machinery monitoring," *Proc. ICASSP 93,* vol. 1, pp. 59-62, 1993.

[4] T. Komori and S. Katagiri, "A minimum error approach to spotting-based pattern recognition," *IEICE Trans. Inf. and Syst.,* vol. E78-D, No. 8, pp. 1032-1043, Aug. 1995.

[5] B.H. Juang and S. Katagiri, "Discriminative learning for minimum error classification", *IEEE Trans. Signal Processing,* vol. 40, No. 12, pp. 3043-3054, Dec. 1992.

[6] S. Amari, "A theory of adaptive pattern classifiers," *IEEE Trans. EC,* vol. EC-16, No. 3, pp. 299-307, Jun. 1967.

[7] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative metric design for robust pattern recognition," *IEEE Trans. Signal Processing,* vol. 45, No. 11, Nov. 1997 (in press).

[8] E. McDermott and S. Katagiri, "Prototype-based minimum classification error/generalized probabilistic descent training for various speech units," *Computer Speech and Language*, vol. 8, pp. 351-368, 1994.