Events Detection for Audio Based Surveillance by Variable-Sized Decision Windows Using Fuzzy Logic Control

Ing-Jr Ding

Department of Electrical Engineering, National Formosa University,
Yunlin County, Taiwan 632, R.O.C

Abstract

In contrast to the use of fixed-length decision window for analyzing the stream of audio frames seen in many audio event detection applications, a variable-sized decision window approach is proposed in this paper. The control of the window size is governed by a fuzzy logic controller (FLC) which estimates the difference between the likelihood of a targeted audio event and that of the normal acoustic background in order to adjust the window size. The FLC is designed to stretch the window while the monitored environment remains "aurally hot" for collecting more audio frames to ensure the reliability and correctness of the detection and to do the opposite if the context gets "aurally calm". Such a situation-dependent behavior is essential to application where reliable and real-time response is the major concern, for which the fixed-length decision window may not suffice.

Key Words: Audio Event Detection, Decision Window, Fuzzy Logic Controller, Gaussian Mixture Model, Feature Extraction

1. Introduction

Audio event detection [1–4] is getting a lot more attentions in surveillance and security applications where video from cameras used to be the sole source of information input [5–8], as audio events may convey more useful, sometimes even dominant, clues indicating the occurrence of certain singular situation when video information is unreliable or even unavailable in the darkness.

A typical process for audio event detection would feed the stream of audio frames (vectors of extracted acoustic features, that is) into the event classifier by which successive analysis on a pre-determined number of audio frames is conducted and then the decision as to whether an audio event being detected over the associated time span, so called the decision window (DW), is made, as depicted in Figure 1.

As is clearly seen, for a fixed-length DW covering n audio frames of Δt ms time interval, the process makes a decision of event detection every $n \cdot \Delta t$ ms, regardless of the aural situation in the context, which may be calm or noisy. A too-long DW might face the concern of real-time response, which is essential to all surveillance and security applications, whereas a too-short one would instead encounter the problem of false alarms against sudden/intermittent acoustic changes in the background, which is equally undesired either.

The idea of variable-sized DW thus arises and is the core of the proposed audio event detection system described in the following sections.

2. Audio Event Detection System

Figure 2 shows the structure of the proposed audio event detection system, where each frame consists of 160 audio samples with 50% overlaps with leading and trailing neighbor frames, respectively. Note that at 8K Hz sampling rate, the time span for each frame is thus 20

*Corresponding author. E-mail: ingjr@nfu.edu.tw

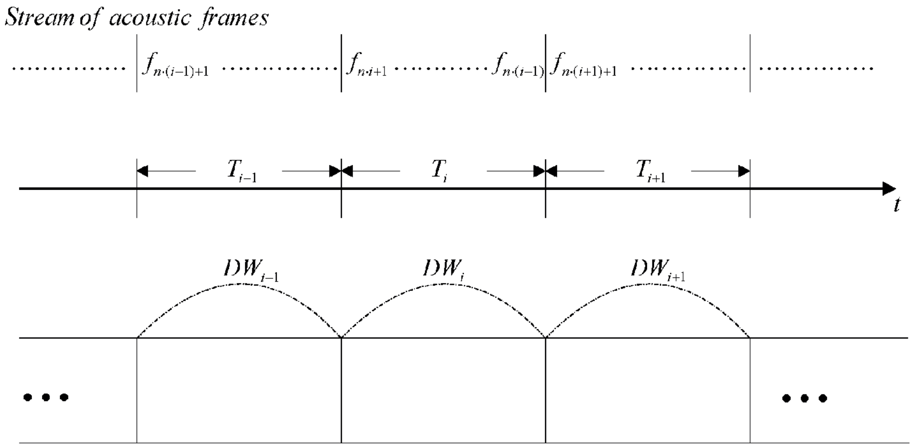

Figure 1. DW with fixed-length, each covering the same number of audio frames, *n*.

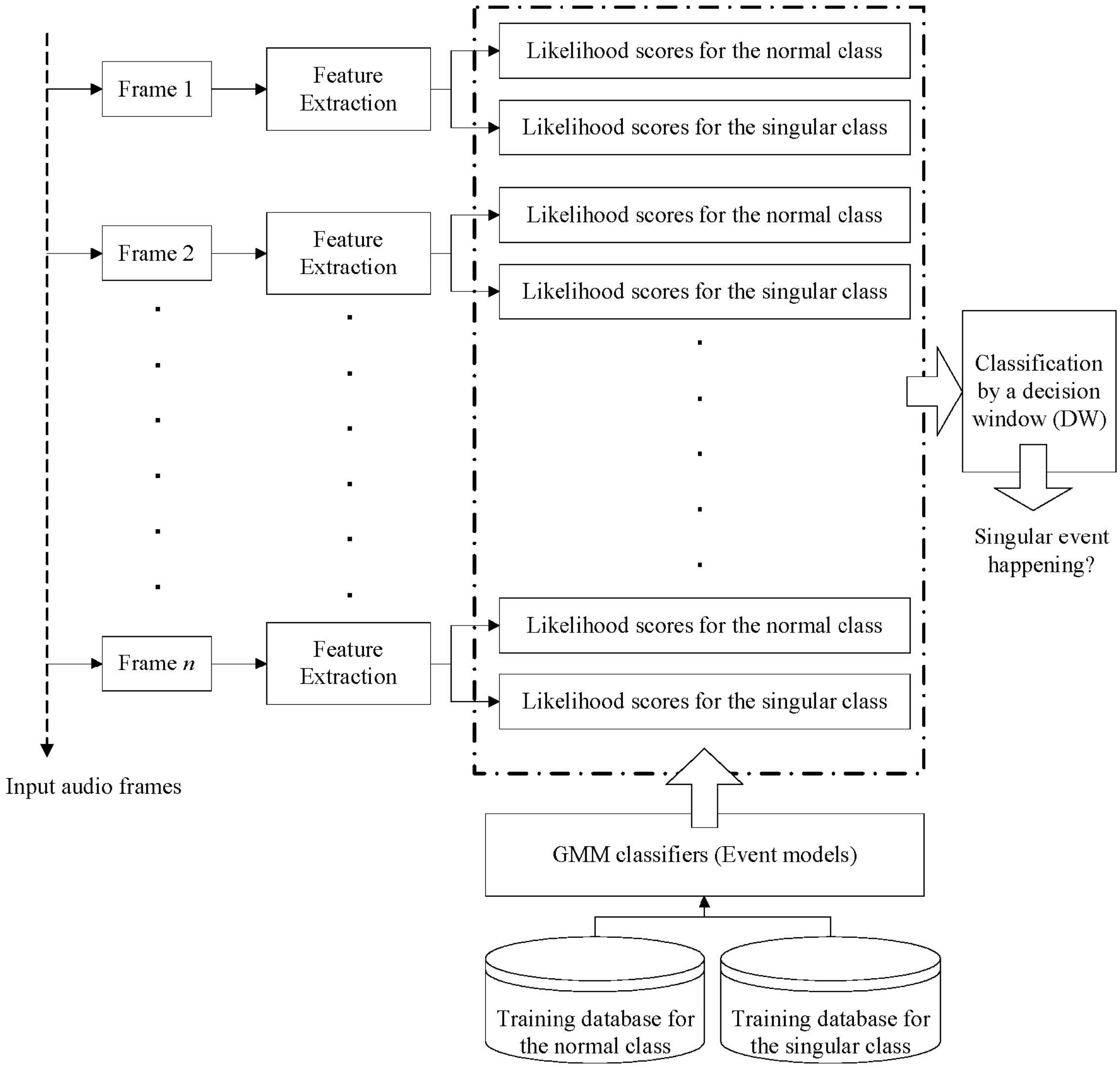
Frame 1
Frame 2
Frame *n*
Feature Extraction
Feature Extraction
Feature Extraction
Likelihood scores for the normal class
Likelihood scores for the singular class
Likelihood scores for the normal class
Likelihood scores for the singular class
Likelihood scores for the normal class
Likelihood scores for the singular class
Classification by a decision window (DW)
Singular event happening?
Input audio frames
GMM classifiers (Event models)
Training database for the normal class
Training database for the singular class

Figure 2. Audio event detection system.

ms and, as such, for a fixed-length DW covering *n* frames, the event decision will be made every $(n + 1) \cdot 10$ ms.

2.1 Acoustic Features

The acoustic features extracted from the audio samples include LPC (Linear Prediction Coefficients), LPCC (Linear Prediction Cepstral Coefficients) and MFCC (Mel Frequency Cepstral Coefficients), and the use of which is common practices in audio segmentation and classification [9–11]. As LPC is sensitive to vocal sounds, it is widely exploited in speech recognition and speaker identification, and is thus chosen in the proposed system for monitoring singular audio event: female screaming in this case. The details for LPC computation can be found in [12,13].

LPCC is derived from the impulse response of the acoustic model [13] and is thus good for catching signal changes, background noises (normal) or monitored sounds (singular) alike. It is chosen for comparing with the effect of LPC so that a sudden or abrupt variation in background acoustic condition will not be mistaken for a singular event.

MFCC is a more delicate human auditory model in the form of multivariate Gaussian distributions [13] and has been known to be not only robust against noises or sudden signal changes, but also effective in discriminating vocal and non-vocal audio event, and is thus chosen.

2.2 Sound Classification Model

The Event-Model deployed here is basically a GMM classifier consisting of two separate GMM models, one for background sound, and the other for singular sound.

Mathematically, a GMM is a weighted sum of *M* Gaussians, denoted as

$$\lambda = \{w_i,\ \mu_i,\ \Sigma_i\},\ i = 1,\ 2,\ ...,\ M,\ \sum_{i=1}^{M} w_i = 1$$

where w_i is the weight, μ_i is the mean and Σ_i is the covariance.

To establish a GMM model given a set of acoustic feature vectors $X = \{x_n \mid n = 1, 2, \ldots, N\}$, the Expectation-Maximization (EM) algorithm [14] is adopted in the system and implemented as follows:

(1) λ initialization is performed by a binary splitting vector quantization algorithm [15]; Σ_i is in diagonal form for computational consideration; *M* is determined by the Bayesian Information Criterion as suggested in [16].

(2) The computation for GMM parameters is, as suggested by the name EM, basically an iterative process through which GMM parameters are progressively updated for maximizing the expectation value of the acoustic data.

REPEAT

{Expectation computation:

$$f(i \mid x_n,\ \lambda) = \frac{w_i \cdot b_i(x_n)}{\sum_{k=1}^{M} w_k b_k(x_n)} \tag{1}$$

where

$$b_i(x_n) = \frac{1}{(2\pi)^{D/2} \cdot |\Sigma_s|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x_n - \mu_s)^T (\Sigma_s)^{-1} (x_n - \mu_s)\right\} \tag{2}$$

λ-update for $f(\cdot)$ maximization:

$$w_i = \frac{1}{N}\sum_{n=1}^{N} f(i \mid x_n,\ \lambda) \tag{3}$$

$$\mu_i = \frac{\sum_{n=1}^{N} f(i \mid x_n,\ \lambda) \cdot x_n}{\sum_{n=1}^{N} f(i \mid x_n,\ \lambda)} \tag{4}$$

$$\Sigma_i = \frac{\sum_{n=1}^{N} f(i \mid x_n,\ \lambda) \cdot (x_n - \mu_i) \cdot (x_n - \mu_i)^T}{\sum_{n=1}^{N} f(i \mid x_n,\ \lambda)} \tag{5}$$

}*UNTIL* (λ convergence achieved)

The number of iterations typically goes as high as several thousands. In the training phase, three GMM models for the auditory contexts “office space”, “parking lot" and "home residence" are established, respectively; also build are three GMMs for "female screaming" in each of the three auditory contexts with recording collected from 15 females as the singular audio event to be detected.

2.3 Classification with Fixed-Length DW

Consider the classifier operating with a decision window covering *n* acoustic vectors of *D* dimensions, X = $\{x_i \mid i = 1, 2, \ldots, n\}$, together with two sound models, λ_1 for normal events and λ_2 for singular events.

The class of X is determined by maximizing the a *posteriori* probability $P(\lambda_s \mid X)$,

$$\hat{s} = \max_{s=\{1,2\}} P(\lambda_s \mid X) = \max_{s=\{1,2\}} \frac{f(X \mid \lambda_s)}{f(X)} \cdot P(\lambda_s) \tag{6}$$

Note that

$$f(x_i \mid \lambda_s) = \sum_{j=1}^{M} w_j \cdot b_j(x_i) \tag{7}$$

and

$$b_j(x_i) = \frac{1}{(2\pi)^{D/2} \cdot |\Sigma_s|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x_i - \mu_s)^T (\Sigma_s)^{-1} (x_i - \mu_s)\right\} \tag{8}$$

However, in real implementation, Eq. (6) is replaced by

$$\hat{s} = \max_{s=\{1,2\}} \sum_{i=1}^{n} \log f(x_i \mid \lambda_s) \tag{9}$$

for simplicity.

For audio event detection using fixed-length DW, as are in many currently existing cases, the number of audio frames in Eq. (9), *n*, is therefore also fixed. The setting of a relatively narrow DW may potentially increase the rate of false alarms in the case of sudden and abrupt fluctuations in the background acoustic condition, and that of a too wide DW may not suffice the need of real-time response due to making decisions at a longer periodicity. It is therefore desirable to have the width of DW be situation-dependent for enhancing overall performance of audio event detection.

3. Variable-Sized DW by FLC

The length of the decision window should be small when encountering a somewhat "aurally calm" situation so that decision of event detection could be undertaken at higher rate and be stretched at "aurally hot" moments. An FLC mechanism is conceived for this purpose.

3.1 Short Timeslot Likelihood Difference

An index *STLD* (Short Timeslot Likelihood Difference) for governing the length of the decision window in the case of two sound models is devised as follows:

$$STLD = \left|\sum_{i=1}^{m} \log f(x_i \mid \lambda_1) - \sum_{i=1}^{m} \log f(x_i \mid \lambda_2)\right| \tag{10}$$

where λ_1 and λ_2 are the sound models in consideration, $f(x_i \mid \lambda_1)$ and $f(x_i \mid \lambda_2)$ are given by Eq. (7), representing the likelihood of λ_1 and λ_2 model classification, respectively, for frame x_i.

The rationale behind Eq. (10) is that at the beginning stage covering *m* frames, say, of a decision window, if the class inclination of the frames has clearly exhibited, one term in Eq. (10) will be substantially greater than the other. As a consequence, a salient *STLD* value is acquired, indicating that a narrow decision window would suffice. If the class of the *m* frames can not be resolved, both terms in Eq. (10) would be trivial and lead to an insignificant *STLD* implying the need of a wide DW in order to collect more frames for classification. Figure 3 illustrates the "phenomenon" implicated by Eq. (10).

3.2 *STLD*-Driven FLC

Fuzzy Logic Controller has been deployed in many applications where empirical expertise is engineered in the form of rule set/base [17,18], including audio applications [19]. As already explained, the *STLD* index can be used as the key to DW size control and, as a result, an FLC driven by two IF-THEN fuzzy rules is designed accordingly:

Rule 1: If *STLD* is small,
Then *WL* is big.

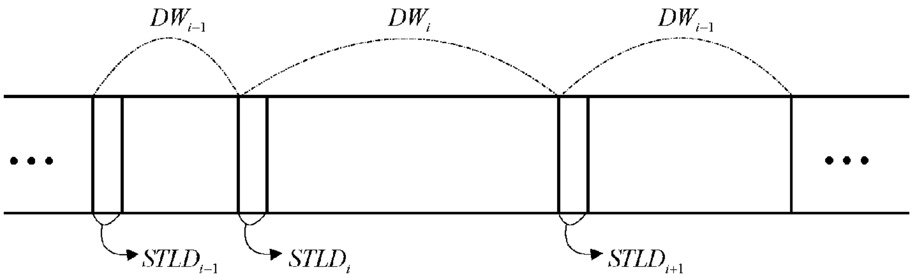

Figure 3. DWs with variable length governed by *STLD* (Short Timeslot Likelihood Difference) indices.

Rule 2: If *STLD* is big,
Then *WL* is small.

Where *STLD* is the input for the FLC and *WL*, the window length, is the output of the FLC.

Quantitatively, the FLC rule set is transformed into

Rule 1: If *STLD* is M_1 (*STLD*),
Then $WL_B = f_1$ (*STLD*),
Rule 2: If *STLD* is M_2 (*STLD*),
Then $WL_S = f_2$ (*STLD*), (11)

where

$$WL = \frac{\sum_{i=1}^{2} M_i(STLD)\cdot f_i(STLD)}{\sum_{i=1}^{2} M_i(STLD)} \tag{12}$$

$$M_1(STLD) = \begin{cases} 1 & STLD \le STLD_1, \\ \dfrac{STLD_2 - STLD}{STLD_2 - STLD_1} & STLD_1 < STLD < STLD_2, \\ 0 & STLD \ge STLD_2, \end{cases} \tag{13}$$

$$M_2(STLD) = \begin{cases} 0 & STLD \le STLD_1, \\ \dfrac{STLD - STLD_1}{STLD_2 - STLD_1} & STLD_1 < STLD < STLD_2, \\ 1 & STLD \ge STLD_2, \end{cases} \tag{14}$$

$$f_1(STLD) = a_1 \cdot STLD + b_1 \tag{15}$$

$$f_2(STLD) = a_2 \cdot STLD + b_2 \tag{16}$$

In the formulation, $M_1(\cdot)$ and $M_2(\cdot)$ are membership functions of *STLD*, as shown in Figure 4, and *WL*, the DW length to be determined by the *STLD*-controlled FLC, is a weighted sum of $f_1(\cdot)$ and $f_2(\cdot)$. It is observed from Eqs. (12–14) that for $STLD < STLD_1$, *WL* is solely determined by $f_1(\cdot)$, simply the case of Rule 1; whereas for $STLD > STLD_2$, *WL* is determined by $f_2(\cdot)$ alone, as is the case of Rule 2.

The FLC now has six hyper-parameters (a_1, a_2, b_1, b_2, $STLD_1$ and $STLD_2$) to be fixed, for which an iterative process is devised as follows

Step 1: Let $STLD_1 : STLD_2 = 1 : 3$ and give an initial value to $STLD_1$.

Step 2: Estimate the parameters a_1 and b_1 under the condition $STLD < STLD_1$, wherein $M_1(STLD) = 1$, $M_2(STLD) = 0$, and

$$WL = \frac{M_1(STLD)\cdot f_1(STLD)}{M_1(STLD)} = f_1(STLD) = a_1 \cdot STLD + b_1.$$

The procedure for fixing a_1 and b_1 is explained in the following pseudo-code sequence:

a_1 = initial value; $b_1 = 0$; $k = 0$;

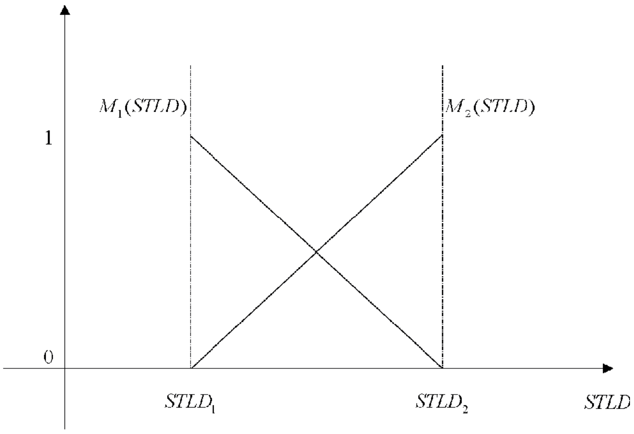

Figure 4. Membership functions.

```
$F^0$ = event_detection_ rate($WL = a_1 \cdot STLD + b_1$,
training_database);
$a_1$ += $\Delta a_1$; $k$ ++;
$F^k$ = event_detection_ rate($WL = a_1 \cdot STLD + b_1$,
training_database);
if ($F^k > F^{k-1}$)
    Repeat
    {$a_1$ += $\Delta a_1$; $k$ ++;
    $F^k$ = event_detection_ rate($WL = a_1 \cdot STLD$
    + $b_1$, training_database);
    } while ($F^k > F^{k-1}$);
else
    Repeat
    {$a_1$ −= $\Delta a_1$; $k$ ++;
    $F^k$ = event_detection_ rate($WL = a_1 \cdot STLD$
    + $b_1$, training_database);
    } while ($F^k > F^{k-1}$);
$b_1$ += $\Delta b_1$; $k$ ++;
$F^k$ = event_detection_ rate($WL = a_1 \cdot STLD + b_1$,
training_database);
if ($F^k > F^{k-1}$)
    Repeat
      {$b_1$ += $\Delta b_1$; $k$ ++;
        $F^k$ = event_detection_ rate($WL = a_1 \cdot$
        $STLD + b_1$, training_database);
    } while ($F^k > F^{k-1}$);
else
    Repeat
      {$b_1$ −= $\Delta b_1$; $k$ ++;
        $F^k$ = event_detection_ rate($WL = a_1 \cdot$
        $STLD + b_1$, training_database);
      } while ($F^k > F^{k-1}$);
return $F^k$;
```

In the pseudo-code sequence, the rate of correct detection returned by event_detection_ rate(WL, X) is defined below

$$\text{detection rate} = \frac{\text{numbers of decision windows with correct detection}}{\text{numbers of all decision windows}} \times 100\ (\%) \tag{17}$$

Step 3: Estimate the parameters a_2 and b_2 under the condition $STLD > STLD_2$, wherein $M_1(STLD) = 0$, $M_2(STLD) = 1$, and

$$WL = \frac{M_2(STLD) \cdot f_2(STLD)}{M_2(STLD)} = f_2(STLD) = a_2 \cdot STLD + b_2.$$

The determination of a_2 and b_2 is done by the same process for a_1 and b_1.

Step 4: Re-estimate the parameter $STLD_2$ under the condition $STLD_1 < STLD < STLD_2$, wherein $M_1(STLD) = \dfrac{STLD_2 - STLD}{STLD_2 - STLD_1}$, $M_2(STLD) = \dfrac{STLD - STLD_1}{STLD_2 - STLD_1}$, and

$$WL = \frac{M_1(STLD)f_1(STLD) + M_2(STLD)f_2(STLD)}{M_1(STLD) + M_2(STLD)}$$

$$= \frac{\dfrac{STLD_2 - STLD}{STLD_2 - STLD_1} \cdot (a_1 \cdot STLD + b_1) + \dfrac{STLD - STLD_1}{STLD_2 - STLD_1} \cdot (a_2 \cdot STLD + b_2)}{\dfrac{STLD_2 - STLD}{STLD_2 - STLD_1} + \dfrac{STLD - STLD_1}{STLD_2 - STLD_1}}$$

$$= \frac{(STLD_2 - STLD)(a_1 \cdot STLD + b_1) + (STLD - STLD_1)(a_2 \cdot STLD + b_2)}{STLD_2 - STLD_1}.$$

With a_1 and b_1 together with a_2 and b_2 already obtained at step 2 and step 3 respectively, a new value for $STLD_2$ can be found through the tuning for best recognition rate process too.

Step 5: Update $STLD_1$ such that $STLD_1 : STLD_2 = 1 : 3$. Repeat from step 2 until the settings of a_1, a_2, b_1, b_2, $STLD_1$ and $STLD_2$ maximize the system performance over the training dataset (60 sec. screaming sounds, 20 sec. in each of three types of background environments, recorded from each of the female subjects).

3.3 Time Complexity Analysis in Online Applications

In an online audio event detection application, the decision window length is appropriately calculated by the well-trained $STLD$-driven FLC. The overhead of finding the decision window length in terms of the number of multiplications, as compared to the conventional fixed-length decision window approach, can be analyzed through its computation defined by Eq. (12). For $STLD < STLD_1$, $WL = a_1 \cdot STLD + b_1$ which requires 1 multiplication, as is for the case when $STLD > STLD_2$, $WL = a_2 \cdot STLD + b_2$.

For $STLD_1 < STLD < STLD_2$,

$$
\begin{aligned}
WL &= \frac{M_1(STLD)\cdot f_1(STLD) + M_2(STLD)\cdot f_2(STLD)}{M_1(STLD) + M_2(STLD)} \\
&= \frac{\begin{array}{l} STLD^2\cdot(a_2 - a_1) + STLD\cdot(a_1\cdot STLD_2 - a_2\cdot \\ STLD_1 + b_2 - b_1) + b_1\cdot STLD_2 - b_2\cdot STLD_1 \end{array}}{STLD_2 - STLD_1} \\
&= p\cdot(c_1\cdot STLD^2 + c_2\cdot STLD + c_3),
\end{aligned}
$$

the computation of which involves 4 multiplications.

Thus, compared to the fixed-length decision window method, the computation of the decision window length by the *STLD*-driven FLC does not increase any time complexity in calculation. The proposed approach in this paper will be efficient and effective for the audio event detection system to be implemented in an online application.

4. Experiments

The training and performance of the proposed audio event detection system are respectively reported in the following subsections.

4.1 Database and Experiment Design

In the training phase, three GMM models for "office space", "parking lot" and "living room" were built using 10-minute recording in each environment. The recording was undertaken at 8K Hz sampling rate, from which LPC, LPCC and MFCC were extracted for each 20 ms frame (consisting of 160 samples, i.e.). Note that a 12-th order LPC, a 12-th order LPC/mel cepstrum and a 12-th order delta cepstrum were utilized [13]. Three GMM models for "female screaming" in each of the three environments were also built using 2/3 of a 180-second (60 sec. for each environment) recording from each of a group of 15 female subjects for extracting the same set of 3 acoustic features; the subjects were requested to scream in every possible way they could during the recording.

The rest one-third of the screaming data (20 sec. for each environment and totally 900 sec. for all 15 females in all the three environments) was used for FLC parameter-tuning as previously described.

In the event detection testing phase, an entirely new group of 15 females was recruited for the screaming recording of 60 sec. each (20 sec. for each of the three environments).

4.2 Experiment Results

During the testing phase, the GMM classifier with the proposed FLC-regulated DW was put to detect audio events occurring in a background audio stream of 15 minutes in length. Three experiments were conducted in "office space", "parking lot" and "living room" respectively, and several observations on the effects of the proposed approach are presented in tabulation for comparison, as are briefed below.

(1) Table 1 shows that, using LPC alone, the approach exploiting variable-sized DW governed by FLC achieves an average of 92%, 93.5% and 95% accuracy for event detection in the three testing contexts respectively, where the window size varies between $W_{\min}$ and $W_{\max}$, with an average of W_{avg}. With LPC alone, Table 2 shows the performance of the fixed-length DW scheme with a variety of fixed DW settings, from 0.5 sec. to 5 sec. at an increment of 0.5 sec., and in all cases the accuracy is inferior to the scores in Table 1. It is further noted that, against the variable-sized DW, the fixed DW reaches competitive scores of 91% at 3-sec. *WL*, 93.33% at 2.5-sec. *WL* and 95% at 1.5-sec. *WL*, respectively in the three testing contexts: the settings of DW fall within the corresponding ranges of DW variation [$W_{\min}$, $W_{\max}$] associated with the FLC-regulated DW operation.

(2) Similar observations from the case of using LPCC alone are also made, as shown in Table 3 and 4.

(3) Table 5 and 6 present the experiment results in the case of using MFCC feature.

Table 1. Event detection by an FLC-regulated DW, using only LPC feature

Variable-sized DW	Living room	Parking lot	Office space
$W_{min.}$	3.12 sec.	2.23 sec.	1.12 sec.
$W_{max.}$	3.96 sec.	2.88 sec.	1.58 sec.
$W_{avg.}$	3.55 sec.	2.56 sec.	1.33 sec.
Accuracy$_{\text{avg.}}$	92%	93.5%	95%

(4) In the experiment, acoustically the noisiest background is the living room, followed by the parking lot, and then the office space. Such a phenomenon seems to be reflected by the range of *WL* variation, *WR* = [$W_{\min}$, $W_{\max}$], when the -driven FLC operated in the three contexts. To be specific,

WR (office space) < *WR* (parking lot) <
WR (living room),

regardless of whichever of the three acoustic features used.

(5) For all the testing in the 3 backgrounds, MFCC leads to the best performance in audio event detection, LPCC the second and LPC the third, regardless of whichever control scheme on DW size being taken, as shown in Figures 5, 6 and 7.

5. Conclusion

An *STLD*-driven FLC mechanism is devised for regulating the DW size in the application of audio event detection, and for all testing cases exploiting 3 individual acoustic feature in three operating backgrounds where the performance of audio event detection is examined, the proposed scheme of variable-sized DW surpasses the fixed-length DW. Moreover, it is noted that

Table 2. Event detection by fixed-length DW, using only LPC feature

DW length	Living room	Parking lot	Office space
0.5 sec.	80.83%	83.33%	91.67%
1 sec.	81.67%	86%	93.33%
1.5 sec.	84%	87%	95%
2 sec.	86%	91.33%	94.67%
2.5 sec.	89.33%	93.33%	95%
3 sec.	91%	93%	95%
5 sec.	91.67%	93.33%	95%
Average	86.36%	89.62%	94.24%

Table 3. Event detection by an FLC-regulated DW, using only LPCC feature

Variable-sized DW	Living room	Parking lot	Office space
$W_{min.}$	3.18 sec.	2.31 sec.	1.15 sec.
$W_{max.}$	3.98 sec.	2.92 sec.	1.63 sec.
$W_{avg.}$	3.57 sec.	2.61 sec.	1.36 sec.
$\text{Accuracy}_{\text{avg.}}$	93.5%	95%	97%

Table 4. Event detection by fixed-length DW, using only LPCC feature

DW length	Living room	Parking lot	Office space
0.5 sec.	83.67%	87.5%	92.5%
1 sec.	87.67%	90%	94.67%
1.5 sec.	89%	90.5%	96.5%
2 sec.	90.67%	93.33%	96.67%
2.5 sec.	91.67%	95%	96.67%
3 sec.	93%	95%	96%
5 sec.	93.33%	95%	96.67%
Average	89.86%	92.33%	95.67%

Table 5. Event detection by an FLC-regulated DW using only MFCC feature

Variable-sized DW	Living room	Parking lot	Office space
$W_{min.}$	3.15 sec.	2.18 sec.	1.17 sec.
$W_{max.}$	3.92 sec.	2.91 sec.	1.68 sec.
$W_{avg.}$	3.52 sec.	2.55 sec.	1.41 sec.
$\text{Accuracy}_{\text{avg.}}$	95%	98.5%	98.5%

Table 6. Event detection by fixed-length DW, using only MFCC feature

DW length	Living room	Parking lot	Office space
0.5 sec.	84.5%	88.33%	93.33%
1 sec.	88.33%	90.67%	95.33%
1.5 sec.	90.5%	91.5%	98%
2 sec.	91.33%	94.67%	98%
2.5 sec.	92.5%	98.33%	98.33%
3 sec.	94%	98%	98%
5 sec.	95%	98.33%	98.33%
Average	90.88%	94.26%	97.05%

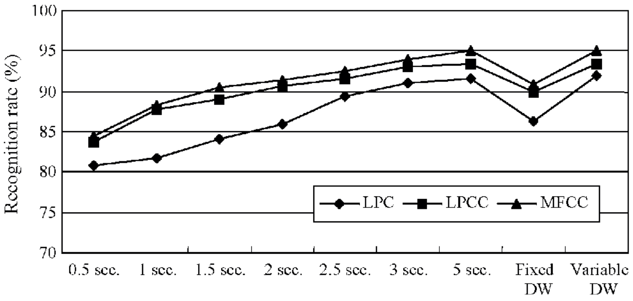

Figure 5. Living room audio event detection.

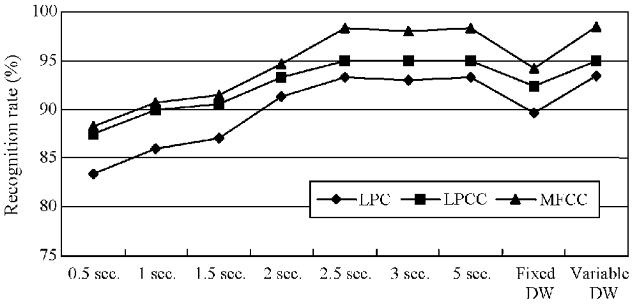

Figure 6. Parking lot audio event detection.

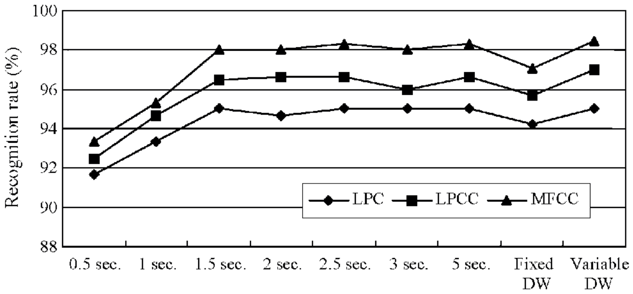

Figure 7. Office space audio event detection.

the performance of the fixed-length DW reaches the score competitive against FLC-regulated DW at a DW setting that falls within the range of size variation of the latter, which declares the effectiveness of the design proposed in this paper.

References

[1] Clavel, C., Ehrette, T. and Richard, G., "Event Detection for an Audio-Based Surveillance System," *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 1306–1309 (2005).

[2] Harma, A., McKinney, M. F. and Skowronek, J., "Automatic Surveillance of the Acoustic Activity in Our Living Environment," *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 634–637 (2005).

[3] Besacier, L., Dufaux, A., Ansorge, M. and Pellandini, F., "Automatic Sound Recognition Relying on Statistical Methods, with Application to Telesurveillance," *Proceedings of International Workshop on Intelligent Communication Technologies and Applications, with Emphasis on Mobile Communications*, pp. 116–120 (1999).

[4] Atrey, P. K., Maddage, N. C. and Kankanhalli, M. S., "Audio Based Event Detection for Multimedia Surveillance," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 813–816 (2006).

[5] Evans, R. J., Brassington, E. L. and Stennett, C., "Video Motion Processing for Event Detection and Other Applications," *Proceedings of International Conference on Visual Information Engineering*, pp. 93–96 (2003).

[6] Albiol, A., Sandoval, C., Naranjo, V. and Mossi, J. M., "Robust Motion Detector for Video Surveillance Applications," *Proceedings of International Conference on Image Processing*, Vol. 3, pp. II-379–382 (2003).

[7] Amano, T., Hiura, S., Yamaguti, A. and Inokuchi, S., "Eigen Space Approach for a Pose Detection with Range Images," *Proceedings of International Conference on Pattern Recognition*, pp. 622–626 (1996).

[8] DuPont, E. M., Yu, H. and .Roberts, R. G., "Object Pose Detection in the Presence of Background Clutter and Occlusion," *Proceedings of the Thirty-Sixth Southeastern Symposium on System Theory*, pp. 446–450 (2004).

[9] Lu, L., Jiang, H. and Zhang, H. J., "A Robust Audio Classification and Segmentation Method," *Proceedings of the 9th ACM International Conference on Multimedia*, pp. 203–211 (2001).

[10] Lu, L., Li, S. Z. and Zhang, H. J., "Content-Based Audio Segmentation Using Support Vector Machines," *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 956–959 (2001).

[11] Li, S. Z., "Content-Based Audio Classification and Retrieval Using the Nearest Feature Line Method," *IEEE Transactions on Speech and Audio Processing*, Vo1. 8, pp. 619–625 (2000).

[12] Markel, J. D. and Gray, A. H., *Linear Prediction of Speech*, Springer-Verlag, New York (1976).

[13] Rabiner, L. and Juang, B. H., *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey (1993).

[14] Dempster, A. P., Laird, N. M. and Rubin, D. B., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Vol. 39, pp. 1–38 (1977).

[15] Linde, Y., Buzo, A. and Gray, R. M., "An Algorithm

for Vector Quantizer Design," *IEEE Transactions on Communications*, Vol. 28, pp. 84–95 (1980).

[16] Fraley, C. and Raftery, A. E., "How Many Clusters? Which Clustering Method? Answers via Model Based Cluster Analysis," *The Computer Journal*, Vol. 41, pp. 578–588 (1998).

[17] Yager, R. and Filev, D., *Essentials of fuzzy modeling and control*, Wiley, New York (1994).

[18] Takagi, T. and Sugeno, M., "Fuzzy Identification of Systems and Its Applications to Modeling and Control," *IEEE Transactions on System, Man, and Cybernetics*, Vol. 15, pp. 116–132 (1985).

[19] Yen, J., Langari, R. and Zadeh, L. A. (eds.), *Industrial applications of fuzzy logic and intelligent systems*, IEEE Press, New York (1995).

Manuscript Received: Jul. 20, 2007
Accepted: Mar. 6, 2009