Adaptive Kernel Function of SVM for Improving Speech/Music Classification of 3GPP2 SMV

Chungsoo Lim and Joon-Hyuk Chang

Because a wide variety of multimedia services are provided through personal wireless communication devices, the demand for efficient bandwidth utilization becomes stronger. This demand naturally results in the introduction of the variable bitrate speech coding concept. One exemplary work is the selectable mode vocoder (SMV) that supports speech/music classification. However, because it has severe limitations in its classification performance, a couple of works to improve speech/music classification by introducing support vector machines (SVMs) have been proposed. While these approaches significantly improved classification accuracy, they did not consider correlations commonly found in speech and music frames. In this paper, we propose a novel and orthogonal approach to improve the speech/music classification of SMV codec by adaptively tuning SVMs based on interframe correlations. According to the experimental results, the proposed algorithm yields improved results in classifying speech and music within the SMV framework.

Keywords: SVM, SMV, speech/music classification algorithm.

Manuscript received Dec. 24, 2010; revised Apr. 19, 2011; accepted May 6, 2011.

This work was supported by the Ministry of Knowledge Economy (MKE), Rep. of Korea, under the Information Technology Research Center (ITRC) support program supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2011-C1090-1121-0007) and by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0028295).

Chungsoo Lim (phone: +82 10 8615 6396, email: chungsoo.lim@gmail.com) was with the Department of Electronic Engineering, Inha University, Seoul, Rep of Korea, and is now with the Institute of Information Science and Engineering Research, Mokpo National University, Mokpo, Rep. of Korea.

Joon-Hyuk Chang (corresponding author, email: jchang@hanyang.ac.kr) was with the Department of Electronic Engineering, Inha University, Seoul, Rep of Korea, and is now with the Department of Electronic Engineering, Hanyang University, Seoul, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.11.0110.0780

I. Introduction

Since multimedia services provided through personal wireless communication devices such as cell phones are now commonplace due to recent progress in mobile communication technology, there has been growing technological demand to efficiently utilize limited bandwidth. To make the most of limited bandwidth, the variable bitrate speech coding concept has been introduced and extensively researched. As an example of this trend, the selectable mode vocoder (SMV) speech codec adopted by the third-generation partnership project 2 (3GPP2) incorporates a speech/music classification technique for different bitrate allocations [1], [2]. However, given the discovery that a simple heuristic logic inherently implemented for speech/music classifications in the SMV has room for improvement, a novel classification algorithm based on support vector machines (SVMs) was proposed by Kim and Chang [3] and achieved a substantial improvement for speech/music classification. Inspired by the potential they presented, our goal is to further improve the speech/music classification of SMV by adding a novel and orthogonal enhancement to SVMs. Actually, SVM is one of widely acknowledged and employed machine learning techniques that is particularly good at pattern recognition, such as face recognition, character recognition, and speech recognition as well as data mining [4], [5]. In SVM, kernel function plays a crucial role. First, it maps target input space to higher dimensional space in case the input is not linearly separable. After the mapping, the input becomes linearly separable. Second, classification performance tends to be sensitive to the choice of a kernel function. This implies that it is desirable to choose the right kernel function for optimal performance [6].

Once a kernel function is chosen, it is also important to

optimize parameters of the kernel function [7]. This is because not only the performance but also the training time of SVMs depends heavily on kernel parameters. A great deal of research has been focused on optimizing kernel parameters [6], [8], [9], and most of these kernel parameter optimization techniques are employed when SVMs are being trained. Meanwhile, a technique that applies when SVMs are in a classification phase was proposed [10]. This technique improves the performances of SVMs by assigning different weight to each input element according to its contribution to generalized error. Since it applies during classification, it becomes orthogonal to techniques that apply during training. It means that our proposed approach can be developed and employed in parallel with conventional optimizations, resulting in synergistic boost in classification performance.

Therefore, we propose a novel approach with this orthogonal nature. A simple way to develop an orthogonal scheme is based on a decision function obtained through training and used for classification. In a decision function, its kernel function is the most relevant parameter because it plays a crucial role in training, and no one has ever studied modification of a kernel function in classification phase. In this regard, we investigate kernel functions with respect to their kernel parameter in terms of classification performance. At first, we choose to analyze radial basis function (RBF) because it is widely used as the kernel function. Afterwards, our investigation shows that we can control the outputs of decision function with the kernel parameter. This means that it is possible to adaptively tune the performance of SVMs. However, adaptive-tuning SVMs require guidance that leads to reduction in classification error. For this reason, we first identify strong correlations in neighboring input frames and then propose a novel approach that utilizes the correlations as guidance for tuning SVMs since the strong correlation is a relevant property in speech and music signals [11]. For example, a music segment lasts for a while before it is interrupted by silence or speech segments. Thus, it is highly probable that the current frame belongs to music if previous frames belong to music.

Even if the proposed algorithm is similar in nature with the discriminative weight training algorithm, there is one significant distinction. The proposed scheme does not require a complex training process, which typically necessitate careful selection of training data. The proposed technique is capable of enhancing SVM-based speech/music classification without any training, provided that SVMs classify speech/music frames accurately enough to ensure reasonable class prediction performance before the enhancement is applied.

The rest of the paper is organized as follows. Section II briefly explains SMV and then lists parameters used for speech/music classifications. Section III contains analysis of an RBF in the beginning and then shows feasibility for controlling SVM outputs with a kernel parameter. A way to control SVM outputs with the kernel parameter is proposed in section IV. Section V describes experimental setup and results in detail, and conclusions and some future directions are presented in section VI.

II. Brief Review of SMV Codec

SMV is an adaptive multirate speech codec adopted as a standard in 3GPP2 and known for its high efficiency in utilizing limited bandwidth [12], [13]. It supports four different average data rates dynamically adjusted according to the types of input frames and four different operational modes dynamically selected based on the status of communication channels between communication stations. Because of this flexibility, an appropriate tradeoff between quality of service and system capacity can be selected for a given situation.

The speech/music classification step in SMV is conducted only for signals that have been identified as speech by voice activity detection (VAD), which is able to distinguish speech from silence and noise. The following are the employed feature parameters used for speech/music classification.

Running average of energy. The running mean energy is given by

$$\overline{E} = 0.75 \cdot \overline{E} + 0.25 \cdot E, \tag{1}$$

where E is the frame energy given by the ratio of the signal power and the window length.

Running mean of the reflection coefficients. Reflection coefficients $k_1(i)$ are calculated by the standard Levinson-Durbin algorithm [14] and the running average of the reflection coefficients is obtained by

$$\overline{k}_N(i) = 0.75 \cdot \overline{k}_N(i) + 0.25 \cdot \overline{k}_1(i), \quad i = 1, ..., 10. \tag{2}$$

Running mean of the partial residual energy. The running mean of the partial residual energy is calculated as

$$\overline{E}_N^{\text{res}} = 0.9 \cdot \overline{E}_N^{\text{res}} + 0.1 \cdot E_N^{\text{res}}, \tag{3}$$

where E_N^{res} is calculated using the signal power and the reflection coefficients [1].

Running mean of the normalized pitch correlation. The running mean of the normalized pitch correlation is given by

$$\overline{corr}_{\text{p}} = 0.8 \cdot \overline{corr}_{\text{p}} + 0.2 \cdot \left(\frac{1}{5} \sum_{i=1}^{5} corr_{\text{p}}(i) \right), \tag{4}$$

where $corr_{\text{p}}(i)$ is the pitch correlation obtained from the open loop pitch estimation.

Running average of the periodicity counter. The periodicity

counter c_{pr} is given by comparing some extracted parameters with a fixed threshold value [1]. The classification algorithm of the SMV determines a signal to music when c_{pr} is higher than 18. The running mean of the periodicity counter is updated as

$$\bar{c}_{pr} = \alpha \cdot \bar{c}_{pr} + (1-\alpha) \cdot c_{pr}, \quad (5)$$

where α is the specified weight.

Music continuity counter. The music continuity counter c_M is adaptively incremented and decremented by comparing the speech/music classification parameters to a set of fixed thresholds [1]. The original algorithm of the SMV classifies a signal into music when $\bar{c}_M$ is higher than 200 in which the running of the music continuity counter $\bar{c}_M$ is given by

$$\bar{c}_M = 0.9 \cdot \bar{c}_M + 0.1 \cdot c_M. \quad (6)$$

III. Impact of RBF Kernel Parameter on SVM Classification

Our choice of the kernel function, RBF, is one of the classical kernel functions in SVMs [9]. In this section, we vary the width parameter of RBF to see how it affects the outputs of SVMs. When inputs are linearly separable, the decision function is given by

$$f(X(t)) = \sum_{i=1}^{M} \alpha_i^* z_i \cdot \langle X_i^*, X(t) \rangle + b^* \underset{H_1}{\overset{H_0}{\gtrless}} 0, \quad (7)$$

where X_i^* is the i-th vector of M support vectors, z_i is the label for support vector X_i^*, and $<X_i^*, X(t)>$ is the inner product between the support vector and the t-th input frame $X(t)$. If the output of the decision function is greater than zero, $X(t)$ is classified as speech (H_0), but otherwise $X(t)$ is classified as music (H_1). Optimization bias b^* and Lagrange multiplier α^* are obtained by solving a quadratic programming problem. If input vectors are not linearly separable, its decision function is slightly modified incorporating a kernel function, as shown in the following:

$$f(X(t)) = \sum_{i=1}^{M} \alpha_i^* z_i \cdot K(X_i^*, X(t)) + b^* \underset{H_1}{\overset{H_0}{\gtrless}} 0. \quad (8)$$

If RBF is used as the kernel function, the kernel function is defined as

$$K(X_i^*, X(t)) = \exp(-\gamma \left\| X_i^* - X(t) \right\|^2), \quad (9)$$

where γ is the kernel parameter of RBF and is related to the width of RBF. If we add a small positive value δ to γ

$$K'(X_i^*, X(t)) = \exp\left(-(\gamma + \delta) \cdot \left\| X_i^* - X(t) \right\|^2 \right). \quad (10)$$

If we rewrite (10), it can be expressed as

$$K'(X_i^*, X(t)) = \exp(-\gamma \left\| X_i^* - X(t) \right\|^2) \cdot \exp(-\delta \left\| X_i^* - X(t) \right\|^2). \quad (11)$$

This is the form that the original kernel function $K(X_i^*, X(t))$ is multiplied by $\exp(-\delta||X_i^*-X(t)||^2)$. Since $||X_i^*-X(t)||^2$ is positive and a positive number is usually chosen for γ, $\exp(-\gamma\ ||X_i^*-X(t)||^2)$ is a positive value between 0 and 1. Here, if a positive value δ is added to γ, $\exp(-\delta||X_i^*-X(t)||^2)$ is a positive number between 0 and 1, making the modified kernel function produce smaller values than the original kernel function does for a given $||X_i^*-X(t)||^2$. On the contrary, if a negative value δ is added to γ, $\exp(-\delta||X_i^*-X(t)||^2)$ is a positive value bigger than 1, making the modified kernel function produce bigger values than the original function does for a given $||X_i^*-X(t)||^2$. With this simple analysis, the relation between the additive modification δ and the kernel function can be clarified, but we do not know how much the decision function $f(X(t))$ is affected by δ due to $\alpha_i z_i$ in (8). In this regard, we vary δ and summarize the changes of the decision function in Table 1. The first row of Table 1 indicates δ added to the kernel parameter γ. The second row shows the ratio between the number of transitions from positive to negative outputs and the positive outputs before γ is modified. The last row represents the ratio between the number of transitions from negative to positive outputs and the number of negative outputs before the modification. This table is populated with 50 database files that will be described in section V.

If a positive value is added to γ, outputs of SVMs tend to change from positive to negative values, and the reverse transitions are rare. On the other hand, the opposite behavior is observed with negative δ. These behaviors tell us that if we can add a value to γ, more classifications are made for one class and subsequently fewer classifications are made for the other class. Because we label music as –1 and speech as 1, if a positive value is added to γ, more classifications are made for music, whereas the number of classifications as speech is decreased. One more thing to note from the table is that the number of transitions is proportional to δ. Judging from these

Table 1. Impact of kernel parameter γ on polarity of $f(X(t))$.

δ	–0.9	–0.6	–0.3	0.3	0.6	0.9
+ → –	0.23	0.29	0.17	23.4	46.7	55.8
– → +	62.1	35.9	7.44	0.18	0.37	0.56

two observations, we can conclude that we are able to control output of the decision function by varying δ. Even though we can control outputs of SVMs, a rigorous rule for adjusting γ is still required. The next section introduces a way to achieve such a rule.

IV. Guidance Based on Correlations among Neighboring Frames

In this section, we introduce a method to guide adaptive tuning of the kernel parameter based on correlations in adjacent frames. The speech/music signals used in our experiments are made up of three distinct segments: speech segment, music segment, and silence segment. Each segment lasts for seconds, so each segment consists of a group of frames. Therefore, it is highly probable that the current frame belongs to the same class as previous ones. Actually, we measured the probability that the current frame is in the same class as previous ones and found that the probability was almost 100%. Practically, however, we do not have a priori information about the class of each frame. Consequently, we have no choice but to use the classification results of previous frames made by the SVM. If a certain number of consecutive prior frames belong to a class, we predict the current frame to be in the same class as them and adjust the kernel parameter accordingly. Figure 1 shows the accuracy and the usefulness of frame class predictions as we vary the length of consecutive frames that we consider for class predictions. Note that the classification of our SVM is either 1 (speech) or –1 (music). The dashed lines and the solid lines represent the speech class and the music class, respectively. The bolder lines and the narrower lines denote prediction accuracy and prediction usefulness, respectively. The prediction accuracy is the ratio between the number of correct predictions for a class and the number of total predictions made for the class, and the prediction usefulness is defined as the ratio between the number of correct prediction for a class and the total number of frames of the class. The prediction accuracy does not directly represent the benefit from the predictions because it can be misleading especially when the number of predictions is either mush larger or smaller than the number of corresponding frames. Contrarily, the prediction usefulness directly conveys information on the potential benefit from the predictions. The x-axis represents the number of consecutive frames previously classified to be in the same class, and the y-axis denotes the probability that predictions based on previous classifications are correct. The figure is obtained from 50 database files that will be described in section V.

From the figure, it is observed that the accuracy increases and the usefulness decreases as a longer sequence of identically classified frames are required for a prediction. The reason why

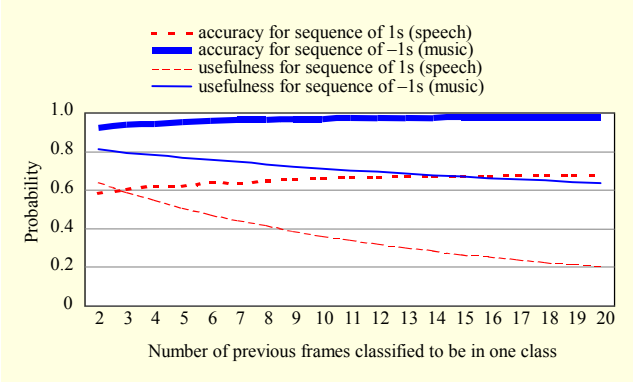


Fig. 1. Class prediction accuracy and usefulness with respect to number of previous frames classified as same class using TIMIT database [15] and commercial music CDs.

the usefulness drops, judging from the definition of the usefulness above, is that the number of occurrences of a longer identically-classified sequence is generally fewer than that of a shorter sequence.

The prediction accuracy can reflect the influence of mispredictions if the number of predictions is large enough. A misprediction becomes a significant problem because the kernel modification is dependent on the class predictions and this, in turn, affects the final SVM classification. Moreover, a misclassification caused by a misprediction may result in another misprediction because class predictions are based on the classifications from the SVM. Therefore, this detrimental chain of misprediction and misclassification must be avoided, and it may be possible to estimate such a devastating performance drop with the prediction accuracy. On the other hand, the prediction usefulness can reflect the influence of correct predictions. A correct class prediction enables the kernel modification to convert otherwise misclassified frames to correctly-classified frames. Thus, we can estimate the impact of correct predictions by the prediction usefulness.

Since it is evident that there is room for improvement for the class prediction from the analysis in terms of the two metrics shown in Fig. 1, we adopt a different approach to speculate frames. Instead of predicting the current frame with previous classifications, we predict the beginning of each segment. For instance, if a certain number of consecutive frames are classified as speech, it may indicate the beginning of a speech segment. After detecting the start of speech segment, every following frame is predicted as a speech frame until a sequence of frames is classified to be music. This method improves the accuracy of frame class prediction without increasing algorithmic complexity. Figure 2 shows a pseudocode for this algorithm.

Specifically, P_c is the variable that indicates the class prediction for the current frame, D_H is a fixed constant that

D_H: history depth
P_c: class prediction
$H_c[P_c]$: classification history
C_{H0}: counter representing the number of classification as H_0
C_{H1}: counter representing the number of classification as H_1
TH_{H0}: threshold for predicting as H_0
TH_{H1}: threshold for predicting as H_1
γ: RBF kernel parameter
γ': modified RBF kernel parameter
δ: additive modification to RBF kernel parameter

```
Initialize variables;
// count the number of previous classifications as H0 or H1
for i = 0 to D_H {
    if H_c[i] > 0
        C_H0 = C_H0 + 1;
    else
        C_H0 = C_H0;
    if H_c[i] < 0
        C_H1 = C_H1 + 1;
    else
        C_H1 = C_H1;
}
// make a prediction based on classification history
if C_H0 > TH_H0
    P_c=H0;
else
    P_c=P_c;
if C_H1 > TH_H1
    P_c=H1;
else
    P_c=P_c;

// adjust the kernel parameter according to the prediction
if P_c=H0
    γ' = γ − δ;
if P_c=H1
    γ' = γ + δ;
```

Fig. 2. Pseudocode for adaptive kernel parameter modification.

determines how many previous classifications are considered for the prediction, and $H_c[\bullet]$ is an array holding previous SVM classifications, which are represented as 1 for speech and 0 for music. In the proposed algorithm, the number of previous classifications for speech (H_0) and the one for music (H_1) within previous D_H frames are counted from the classification history array ($H_c[\bullet]$) and recorded in counter variables, C_{H0} and C_{H1}, respectively. If the number of classifications as speech reflected by C_{H0} is greater than a predefined threshold TH_{H0}, it is assumed that the beginning of a speech segment is detected, and P_c is set to H_0. It should be noted that the current prediction P_c is not changed until the opposite behavior ($C_{H1} > TH_{H1}$) is observed. This ensures that all subsequent frames as well as the current frame are predicted as speech until a music segment is encountered. Likewise, if C_{H1} is greater than TH_{H1}, P_c is set to H_1 and all subsequent frames are predicted to be music until a beginning of a speech segment is detected.

Table 2. Analysis about influence of kernel parameter modification on how fast music onsets can be detected in terms of number of frames between onsets and their corresponding detections.

	$\delta_{0,0}$	$\delta_{0.02,0.02}$	$\delta_{0.04,0.04}$	$\delta_{0.06,0.06}$	$\delta_{0.08,0.08}$
Number of frames	32.12	37.97	39.35	43.93	63.02

Once the prediction for the current frame is made, the kernel parameter γ is adjusted according to the prediction. For frames predicted to be speech, a predefined additive modification δ is subtracted from γ producing a smaller kernel parameter γ'. The reason for reducing the kernel parameter for speech frames is based on the observation that if a positive value is subtracted from γ, outputs of SVMs tend to change from negative values to positive values as shown in Table 1. The opposite modification is performed for frames predicted to be music.

The modified decision function incorporating the dynamically adjusted kernel parameter γ' is finally given by

$$\hat{f}(X(t)) = \sum_{i=1}^{M} \alpha_i z_i \exp(-\gamma' \left\| X_i^* - X(t) \right\|^2) + b^* \underset{H_1}{\overset{H_0}{\gtrless}} \eta. \quad (12)$$

In this way, a correct prediction elevates the probability that the modified decision function produces the correct classification for the frame for which the prediction is made. However, detecting the beginning of each segment has both advantages and disadvantages. One advantage is that it hinders sequences of misclassifications whose sizes are smaller than the history depth parameter in affecting class predictions once the beginning of a segment is detected. However, it is also hard to switch to correct predictions because one such switch requires D_H identical classifications. If a misclassification occurs before D_H identical classifications are encountered, it prevents the class prediction from being switched. Therefore, it is crucial to select a reasonable history depth. In addition, the modification δ affects how fast the beginning of a segment is discovered. Table 2 shows the influence of the modification δ on detecting the onsets of music segments. This table is tabulated from test files that contain metal genre.

The numbers presented in the table denote the number of frames between the first frame of a music segment and the frame from which algorithm detects the segment as music. The first row represents how the modification δ is set. For example, $\delta_{\alpha,\beta}$ means that the kernel parameter is incremented by α for music frames and is decremented by β for speech frames. From the figure, it is observed that bigger the modification is, slower the onset is detected. If each segment is very short (less than

100 frames), this may be a significant overhead. However, even commercials from radio broadcasts, known to be composed of short segments, have average segment length of 8.77 seconds, which is translated to 438.3 frames. If we use $\delta_{0.06,0.06}$, then the overhead of this onset detecting algorithm is 11.81 frames, which is 2.7% of a segment that contains 438.3 frames. If the benefit from the kernel parameter modification outweighs the overhead and the impact of incorrect kernel parameter modifications, the overall performance would be improved. Experiments with various test data will be presented in the next section.

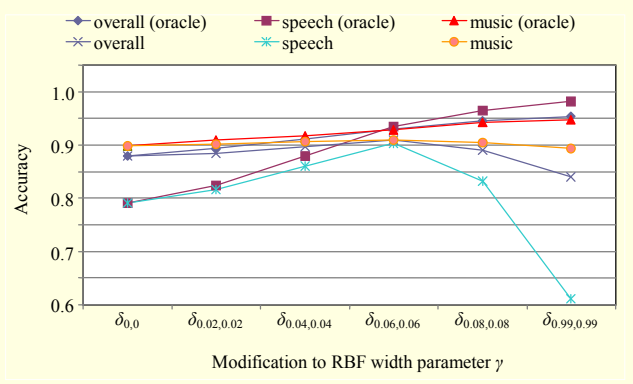


Fig. 3. Influence of RBF kernel width parameter γ on classification accuracy of proposed SVM.

V. Experiments and Results

To evaluate the proposed enhancement, comparisons with the original algorithm in SMV [1], the previously proposed SVM-based speech/music classification algorithm [3], and the discriminative weight training algorithm [10] were performed on the TIMIT speech database [15], commercial music CDs, and actual radio broadcasts. From TIMIT database and music CDs, 50 database files were constructed and used for 10-fold cross-validation. The speech portion of the database was created on utterances from 326 male and 138 female speakers from the TIMIT database and the music portion was created on music CDs of five different genres: metal, jazz, blues, hip-hop, and classical music. All data was sampled at 8 kHz with a frame size of 20 ms. Each database file consisted of five speech segments (6 s to 12 s each), five music segments (28 s to 32 s each), and ten periods of silence (randomly selected between 3 s and 15 s), and these segments alternated. Each of these files contained music segments from one genre only, and there were 10 database files for each genre.

In addition to the cross-validation, we recorded one-hour-long radio broadcasts from two different internet radio stations and used them for verifying the proposed algorithm. Because of the nature of radio broadcasting, the data contained segments where speech and music coexisted in addition to speech-only and music-only segments. We labeled these segments as music to provide enough bitrate for them.

To determine the correctness of classifications, we manually classified each frame and compared it with classification results from the SVM. As a feature vector, those six parameters introduced in section II were concatenated to form a feature vector for each frame. The kernel parameter for the baseline RBF was set to 0.1, and the history depth D_H is experimentally chosen to be 10.

As explained in section IV, the kernel parameter was modified based on the predicted class of current frame. As shown in Table 1, for the output of SVM to change from a positive value to a negative value, the kernel parameter should be incremented. Therefore, the parameter has to be incremented for music and decremented for speech. Figure 3 depicts how the classification accuracy changes as δ, the modification to the kernel parameter, varies. This figure is obtained from test files that contain metal genre. All other test files produce similar results, so their results are not shown here. The x-axis represents how the kernel parameter is modified, and the y-axis indicates the classification accuracy of the SVM enhanced by the proposed mechanism.

There are two sets of lines in the figure, and each set has three lines: speech, music, and overall. The set labeled as oracle is based on actual classes of frames, and the other set is based on predicted classes of frames.

For the case where actual class information is used for manipulating the kernel parameter, more misclassifications are fixed when the kernel parameter is modified by bigger values. On the other hand, for the case where predicted class information is used for predicting the class of current frame, it is observed that the accuracy begins to drop from $\delta_{0.06,0.06}$. To unveil the reason, we analyzed the transitions caused by the kernel parameter modification and show the observations in Tables 3 and 4. The tables are based on the same test files as the ones used for Fig. 3. Table 3 corresponds to the oracle case and Table 4 shows the other case. The leftmost column indicates different transitions entailed by changing the kernel parameter. 'S' means speech and 'M' represents music. Subscript 'c' means correct classifications and 'ic' indicates incorrect classifications. For instance, $S_{ic\text{-}c}$ represents the transitions of speech frames from being incorrectly classified as music to being correctly classified as speech. $M_{ic\text{-}ic}$ is the transitions of music frames that a kernel parameter modification fails to convert the original incorrect classification. The numbers in the tables represent how many times each transition takes place.

From Table 3, we can observe that the number of transitions from incorrect classification to correct classification ($S_{ic\text{-}c}$ and $M_{ic\text{-}c}$) increases, and this results in decrease in the number of

Table 3. Detailed analysis on changes in pattern classification made by adjusting RBF kernel parameter γ (actual frame class information is used).

	$\delta_{0,0}$	$\delta_{0.02,0.02}$	$\delta_{0.04,0.04}$
$S_{c\text{-}c}$	1,390	1,387	1,383
$M_{c\text{-}c}$	5,063	5,063	5,051
$S_{ic\text{-}c}$	0	54	109
$M_{ic\text{-}c}$	0	472	1,366
$S_{ic\text{-}ic}$	601	547	492
$M_{ic\text{-}ic}$	2,437	1,965	1,071
$S_{c\text{-}ic}$	0	3	7
$M_{c\text{-}ic}$	0	0	12
	$\delta_{0.06,0.06}$	$\delta_{0.08,0.08}$	$\delta_{0.99,0.99}$
$S_{c\text{-}c}$	1,381	1,379	1,379
$M_{c\text{-}c}$	5,032	5,012	5,000
$S_{ic\text{-}c}$	156	197	236
$M_{ic\text{-}c}$	1,749	1,891	1,972
$S_{ic\text{-}ic}$	445	404	365
$M_{ic\text{-}ic}$	688	546	465
$S_{c\text{-}ic}$	9	11	11
$M_{c\text{-}ic}$	31	51	63

Table 4. Detailed analysis on changes in pattern classification made by adjusting RBF kernel parameter γ (predicted frame classes are used).

	$\delta_{0,0}$	$\delta_{0.02,0.02}$	$\delta_{0.04,0.04}$
$S_{c\text{-}c}$	1,390	1,383	1,383
$M_{c\text{-}c}$	5,063	4,916	4,769
$S_{ic\text{-}c}$	0	46	94
$M_{ic\text{-}c}$	0	150	637
$S_{ic\text{-}ic}$	601	555	507
$M_{ic\text{-}ic}$	2,437	2,287	1,800
$S_{c\text{-}ic}$	0	7	7
$M_{c\text{-}ic}$	0	147	294
	$\delta_{0.06,0.06}$	$\delta_{0.08,0.08}$	$\delta_{0.99,0.99}$
$S_{c\text{-}c}$	1,386	1,222	1,074
$M_{c\text{-}c}$	4,564	3,297	3,082
$S_{ic\text{-}c}$	144	149	133
$M_{ic\text{-}c}$	1,217	769	760
$S_{ic\text{-}ic}$	457	452	468
$M_{ic\text{-}ic}$	1,220	1,668	1,677
$S_{c\text{-}ic}$	4	168	316
$M_{c\text{-}ic}$	499	1,766	1,981

Table 5. Comparison with original algorithm in SMV [1], discriminative weight training algorithm [10], and previous SVM-based algorithm [3] in terms of speech/music detection probability P_D and total error probability P_E on TIMIT database and music CDs.

Class	Method	Speech P_D	Music P_D	Total P_E	Onset P_E
Blues	SMV [1]	0.882	0.424	0.488	0.201
	SVM [3]	0.839	0.925	0.101	0.243
	SVM+WT [10]	0.872	0.937	0.083	0.207
	Proposed	0.899	0.932	0.078	0.202
Classic	SMV [1]	0.86	0.394	0.511	0.290
	SVM [3]	0.739	0.681	0.302	0.219
	SVM+WT [10]	0.816	0.721	0.251	0.264
	Proposed	0.778	0.694	0.288	0.204
Hiphop	SMV [1]	1.000	0.111	0.707	0.202
	SVM [3]	0.821	0.901	0.123	0.229
	SVM+WT [10]	0.844	0.909	0.111	0.195
	Proposed	0.917	0.92	0.081	0.194
Jazz	SMV [1]	0.975	0.558	0.358	0.226
	SVM [3]	0.719	0.909	0.148	0.227
	SVM+WT [10]	0.75	0.918	0.132	0.267
	Proposed	0.82	0.932	0.102	0.196
Metal	SMV [1]	0.989	0.104	0.727	0.201
	SVM [3]	0.758	0.862	0.169	0.243
	SVM+WT [10]	0.776	0.869	0.159	0.203
	Proposed	0.843	0.874	0.135	0.208
Avg	SMV [1]	0.897	0.304	0.556	0.225
	SVM [3]	0.775	0.856	0.169	0.232
	SVM+WT [10]	0.812	0.871	0.147	0.227
	Proposed	0.851	0.871	0.135	0.203

cases where incorrect classification cannot be fixed by kernel parameter modifications ($S_{ic\text{-}ic}$ and $M_{ic\text{-}ic}$). This explicates the ideal performance improvement of oracle case shown in Fig. 3. Likewise, Table 4 holds information that explains why accuracy drops after a certain point in non-oracle case. Both significant modifications to the kernel parameter and error-prone class prediction increase the number of undesirable transitions from correct classifications to incorrect classifications ($S_{c\text{-}ic}$ and $M_{c\text{-}ic}$). Thus, the number of correct classifications regardless of kernel parameter adjustments also decreases. These phenomena then lower the accuracy of class prediction, leading to the reduction in the number of cases where kernel parameter modifications successfully convert incorrect classifications to correct classifications ($S_{ic\text{-}c}$ and $M_{ic\text{-}c}$).

Table 6. Comparison with original algorithm in SMV [1], discriminative weight training algorithm [10], and previous SVM-based algorithm [3] in terms of speech/music detection probability P_D and total error probability P_E on radio broadcast.

Class	Method	Speech P_D	Music P_D	Total P_E
Radio	SMV [1]	0.706	0.229	0.541
	SVM [3]	0.580	0.885	0.283
	SVM+WT [10]	0.602	0.891	0.270
	Proposed	0.634	0.888	0.254

Consequently, the number of misclassifications not corrected by kernel parameter adjustments ($S_{ic\text{-}ic}$ and $M_{ic\text{-}ic}$) is increased.

Table 5 shows the performance improvement of the proposed enhancement. We compared the proposed algorithm with the original algorithm in SMV in [1], the previous SVM-based algorithm (denoted by SVM) in [3] and the discriminative weight training algorithm (denoted by SVM+WT) in [10]. The first column shows the different music genres each test file represents, and the second column has the four classification algorithms under comparison. The results summarized in the table are average values obtained from all 50 database files. P_D is the probability that music and speech are correctly classified, and P_E is the error probability that encompasses both music and speech. While total P_E is for the entire frames, onset P_E is only for the onset frames. We vary δ, an additive modification to the kernel parameter, to see how classification performance is affected. The δ value for the Table 5 is chosen in such a way that it produces the lowest P_E while improving both speech and music classification rates.

From the table, it can be observed that the proposed enhancement successfully improves the performance of SVM-based classification by adaptively modifying the kernel parameter according to class predictions. It is also discovered that the proposed algorithm outperforms or at least produces comparable performance to the discriminative weight training algorithm and the algorithm in SMV in terms of both total P_E and onset P_E. Another advantage of the proposed scheme over the weight training scheme is that the proposed technique does not require a training process whereas the weight training technique necessitates it.

Additionally, we verified the proposed algorithm with an actual radio broadcast that included fast-paced switches between speech and music segments, and present the result in Table 6. Compared with the results in Table 5, it can be easily seen that the detection probability for speech is lower when an actual radio broadcast is used. This might be attributable to the fact that the length of speech is generally much longer in the radio broadcast than in the data used for the training affecting the running average of the periodicity counter. Through the experiment with the actual radio broadcast, we can observe that the proposed algorithm shows the best performance in terms of classification accuracy.

VI. Conclusion

We have proposed a novel and orthogonal algorithm that adaptively tunes classifications of SVMs with the RBF kernel parameter utilizing interframe correlations abundant in speech and music frames. Our experiments show that with this enhancement, classification accuracies of SVMs can be improved and that the enhancement still has potential for further improvement. In addition to performance improvement, this approach can be combined with existing techniques without any side effect. To take advantage of the full potential of this approach, future work may include an algorithmic way of determining the additive modification to the kernel parameter and a scheme for more accurate frame class predictions.

References

[1] 3GPP2 Spec., "Source-Controlled Variable-Rate Multimedia Wideband Speech Codec (VMR-WB), Service Option 62 and 63 for Spread Spectrum Systems," 3GPP2-C.S0052-A, vol. 1.0, Apr. 2005.

[2] Y. Gao et al., "The SMV Algorithm Selected by TIA and 3GPP2 for CDMA Applications," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 2, May 2002, pp. 709-712.

[3] S.-K. Kim and J.-H. Chang, "Speech/Music Classification Enhancement for 3GPP2 SMV Codec Based on Support Vector Machine," *IEICE Trans. Fundamentals Electron., Commun. Comput. Sci.*, vol. E92-A, no. 2, Feb. 2009.

[4] X. Wang et al., "Infrared Human Face Auto Locating Based on SVM and a Smart Thermal Biometrics System," *Proc. 6th Int. Conf. Intell. Syst. Design Appl.*, vol. 2, Oct. 2006, pp. 1066-1072.

[5] A. Ganapathiraju, J.E. Hamaker, and J. Picone, "Applications of Support Vector Machines to Speech Recognition," *IEEE Trans. Signal Process.*, vol. 52, no. 8, Aug. 2004, pp. 2348-2355.

[6] L.-P. Bi et al., "New Heuristic for Determination Gaussian Kernel's Parameter," *Proc. Int. Conf. Mach. Learning Cybern.*, vol. 7, Aug. 2005, pp. 4299-4304.

[7] S.S. Keerthi and C.-J. Lin, "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel," *Neural Comput.*, vol. 15, no. 7, July 2003, pp. 1667-1689.

[8] J. Tian and L. Zhao, "Weighted Gaussian Kernel with Multiple Widths and Support Vector Classifications," *Proc. Int Symp. Info.*

Eng. Electron. Commerce, May 2009, pp. 379-382.

[9] N.E. Ayat, M. Cheriet, and C.Y. Suen, "Automatic Model Selection for the Optimization of SVM Kernels," *Pattern Recognition*, vol. 38, no. 10, Oct. 2005, pp. 1733-1745.

[10] S.-K. Kim and J.-H. Chang, "Discriminative Weight Training for Support Vector Machine-Based Speech/Music Classification in 3GPP2 SMV Codec," *IEICE Trans. Fundamentals of Electron., Commun. Comput. Sci.*, vol. E93-A, no. 1, Jan. 2010, pp. 316-319.

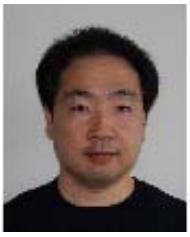
[11] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 2, Apr. 1997, pp. 1331-1334.

[12] S.C. Greer and A. Dejaco, "Standardization of the Selectable Mode Vocoder," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 2, May 2001, pp. 953-956.

[13] C.V. Goudar et al., "SMVLite: Reduced Complexity Selectable Mode vocoder," *Proc. IEEE Int. Conf. Speech Signal Process.*, vol. 1, May 2006, pp. 701-704.

[14] P. Vary and R. Martin, "Digital Speech Transmission: Enhancement, Coding and Error Concealment," *Proc. IEEE Int. Conf. Acoutics, Speech, Signal Process.,* vol. 1, May 2006, pp. 701-704.

[15] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proc. DARPA Workshop Speech Recognition*, Feb. 1986, pp. 93-99.



Chungsoo Lim received the BS and ME in electrical engineering from Inha University, Rep. of Korea, in 1996 and 1999, respectively, and the MS and PhD in computer engineering from the University of Maryland, College Park, 2004, and from North Carolina State University, 2009, respectively. From April 2010 to February 2011, he was with Inha University, Incheon, Rep. of Korea, in a postdoctoral position. Currently, he is a research professor at Mokpo National University, Mokpo, Rep. of Korea. His research interests include computer architecture, embedded systems, and digital signal processing.

Joon-Hyuk Chang received the BS in electronics engineering from Kyungpook National University, Daegu, Rep. of Korea, in 1998, and the MS and PhD in electrical engineering from Seoul National University, Rep. of Korea, in 2000 and 2004, respectively. From March 2000 to April 2005, he was with Netdus Corp., Seoul, as a chief engineer. From May 2004 to April 2005, he was with the University of California, Santa Barbara, in a postdoctoral position to work on adaptive signal processing and audio coding. In May 2005, he joined the Korea Institute of Science and Technology, Seoul, as a research scientist to work on speech recognition. From August 2005 to February 2011, he was an assistant professor in the School of Electronic Enginnering at Inha University, Incheon, Rep. of Korea. Currently, he is an associate professor in the School of Electronic Engineering at Hanyang University, Seoul, Rep. of Korea. His research interests are in speech coding, speech enhancement, speech recognition, audio coding, and adaptive signal processing.