Multimed Tools Appl (2015) 74:5375–5400
DOI 10.1007/s11042-014-1859-8

Efficient implementation techniques of an SVM-based speech/music classifier in SMV

Chungsoo Lim · Joon-Hyuk Chang

Published online: 1 February 2014

Abstract For real-time speech and audio encoders used in various multimedia applications, low-complexity encoding algorithms are required. Indeed, accurate classification of input signals is the key prerequisite for variable bit rate encoding, which has been introduced in order to effectively utilize limited communication bandwidth. This paper investigates implementation issues with a support vector machine (SVM)-based speech/music classifier in the selectable mode vocoder (SMV) framework, which is a standard codec adopted by the Third-Generation Partnership Project 2 (3GPP2). While a support vector machine is well known for its superior classification capability, it is accompanied by a high computational cost. In order to achieve a more realizable system, we propose two techniques for the SVM-based speech/music classifier, aimed at reducing the number of classification requests to the classifier. The first technique introduces a simpler classifier that processes some of the input frames instead of the SVM-based classifier, and the second technique skips a portion of input frames based on strong inter-frame correlation in speech and music frames. Our experimental results show that the proposed techniques can reduce the computational cost of the SVM-based classifier by 95.4 % with negligible performance degradation, making it plausible for integration into the SMV codec.

Keywords Speech/music classification · Support vector machine · Selectable mode vocoder · Embedded system

This work was supported by NRF of Korea grant funded by the MEST (2012R1A2A2A01004895) and this research was supported by the MSIP, Korea, under the ITRC support program supervised by the NIPA (NIPA-2013-H0301-13-4005)

C. Lim
Korea National University of Transportation, 50 Daehak-ro, Choungju-si, Chungbuk, Republic of Korea

J.-H. Chang (✉)
Hanyang University, 222 Wangsimni-ro, Seongdong, Seoul, Republic of Korea
e-mail: jchang@hanyang.ac.kr

1 Introduction

Audio compression technology is essential to multimedia communication systems, including internet protocol (IP) telephony, video conferencing, and digital multimedia broadcasting (DMB). Thus, efficient utilization of limited bandwidth is gaining more attention with the increasing availability of multimedia services through wireless communication devices, resulting in a number of variable bit rate speech codecs. Variable bit rate coding technique assigns higher bit rates to signals with complex time-varying characteristics such as music signals in order to faithfully represent the signals with higher coding resolution, but, on the other hand, it assigns lower bit rates to signals with simple time-varying characteristics in order to efficiently use the bandwidth. If a low fixed bit rate coding is used regardless of signal types, quality of service cannot be guaranteed. In contrast, if a high fixed bit rate is used, bandwidth is wasted because the high bit rate does not improve the quality of service noticeably for some signals with simple characteristics. Variable bit rate coding adapts the coding mechanism to the input signals and delivers high quality at low average bit rates. To cope with variable bit assignments according to input signal type, accurate classification of signals should be performed as a prerequisite. If signal classifications are inaccurate, either simple signals are coded with high bit rates (when simple signals are classified as complex signals), or complex signals are coded with low bit rates (when complex signals are classified as simple signals). When the former occurs, valuable bandwidth is wasted, and when the latter happens, the quality of complex segments is severely degraded. Therefore, correct signal classification should be performed for effective variable bit rate coding. Among several classifications, discriminating speech from music is of prime importance because music is the most demanding class of signals that requires a high bit rate to preserve its quality. Hence, accurate classification of speech and music signals ensures both high quality of music signals and efficient utilization of limited bandwidth, enabling better experience with wireless network technologies such as bluetooth and long term evolution (LTE).

One of the codecs supporting variable bit rate encoding is the selectable mode vocoder (SMV) adopted by the Third-Generation Partnership Project 2 (3GPP2) [1, 8]. The SMV supports four different bit-rates and incorporates a simple speech/music classification technique and a refined silence/noise/voice classification scheme. Recently, speech/music classifiers based on a support vector machine (SVM) in the SMV framework were proposed by Kim and Chang [11], showing acceptable performance. Specifically, the SVM-based classifiers utilize parameters generated inside the SMV without any further processing and considerably reduce the classification error rate of the built-in classifier in the codec from 39.7 % to 9.8 %. Indeed, SVMs have been adopted for a large variety of multimedia applications such as human object detection for protecting privacy in video surveillance, gesture recognition for exercise games, defect classification in steel strip surfaces, and detection of transmission errors within H.264 decoder [6, 15]. However, SVMs typically require a tremendous amount of computation for a classification due to the inner vector products with a large number of support vectors. Therefore, naive introductions of SVMs to embedded systems that generally have limited computational resources are prohibitive. The SVM parameters directly related to computational intensity are the number and dimensionality of support vectors. A great deal of research has been conducted to optimize these parameters, including selection of relevant features and simplification of support vectors [2, 9]. While these approaches can successfully reduce the number and dimensionality of support vectors, another major contributor to the high computational intensity of the SVM-based classifier is the number of classifications executed by the classifier.

Therefore, we propose two simple but effective techniques to reduce the number of classification requests to the SVM-based classifier with little impact on the classification performance. The first technique uses a modified version of the built-in classifier in the SMV codec for filtering out classification requests that can be inherently handled by the built-in classifier, thereby reducing the number of classification requests to the SVM-based classifier. Since the computational load for the built-in classifier is incontestably lighter than that of the SVM-based classifier, this technique can significantly reduce the computational load. The second technique also lessens the classification burden on the SVM-based classifier by skipping some of the input frames. This technique is eventually based on the strong inter-frame correlations in speech and music. Extensive experiments show that these proposed techniques improve the practicality of the SVM-based speech/music classifier under various conditions. Further, in order to evaluate the general applicability of the techniques, we conduct relevant experiments with other speech/music classifiers and a different data set.

The rest of the paper is organized as follows: Section 2 introduces the classification request filtering mechanism, Section 3 presents the classification request omission technique, and Section 4 describes the combination of the filtering and skipping mechanism. Section 5 describes the experimental setup and results in detail, while Section 6 presents conclusions and some future research directions.

2 Filtering classification requests by hierarchical classification

In this section, a filtering mechanism that reduces the usage of the SVM-based classifier is presented. The basic idea behind this technique is to filter out classification requests that can be successfully handled by a much simpler speech/music classifier, lessening the burden on the SVM-based classifier. In other words, a hierarchical classifier is formed by combining the SVM-based classifier as the second level classifier and a simpler classifier as the first level classifier. Only the requests that cannot be handled by the first level classifier are passed to the second level classifier. There are three requirements for the first level classifier. First, it has to fit in the SMV codec without exhaustive effort. Second, it should be much simpler than the SVM-based classifier so that using the first level classifier instead of the SVM-based classifier results in shorter execution time and less energy consumption. Third, it needs to be versed in classifying at least one class. Before we evaluate the built-in speech/music classifier in the SMV with respect to the three requirements, let us first briefly review the classifier. The classifier chosen for the first level classifier is introduced and its eligibility as the first level classifier is verified against the three requirements in Section 2.1, and the accuracy of the hierarchical classifier is analyzed in Section 2.2.

2.1 Built-in classifier in SMV as the first level classifier

The built-in classifier in SMV uses two parameters for deciding the class of an input frame: running average of periodicity counter and music continuity counter. The periodicity counter c_{pr} is reset every 32 frames and incremented if the periodicity flag is high, which is set if the normalized pitch correlation is larger than a fixed threshold [1]. The classification algorithm determines a signal to be music when $\overline{c_{pr}}$ is greater than 18. The running mean of the periodicity counter is updated as:

$$\overline{c}_{pr} = \delta \cdot \overline{c}_{pr} + (1 - \delta) \cdot c_{pr} \qquad (1)$$

where δ is the specified weight. The music continuity counter c_M is adaptively incremented and decremented by comparing the speech/music classification parameters such as the running mean of input frame energy and running means of partial residual frame energy to a set of fixed thresholds [1]. A signal is classified as music when $\overline{c}_M$ is greater than 200 by according to the classification algorithm, in which the running mean of the music continuity counter $\overline{c}_M$ is given by

$$\overline{c}_M = 0.9 \cdot \overline{c}_M + 0.1 \cdot c_M \tag{2}$$

In the original classifier, if either one of the conditions given by (1) and (2) is satisfied, the corresponding frame is classified as music.

From the above description, it is obvious that the original speech/music classifier in the SMV meets the first two requirements. First, it is already a built-in classifier in the SMV, so there is no need for any extra processing. Second, it is much simpler than the SVM-based classifier because it only requires the comparison of two parameters with two given threshold values. To evaluate the classification capability of the original classification algorithm, we vary the two thresholds used in the algorithm and plot the accuracy and coverage for speech and music classes in Fig. 1. Note that the plots represent the average behaviors of the test data used in our experiments. The x-axis represents the threshold pair (music continuity counter and running average of the periodicity counter), and the y-axis shows both the accuracy of the classifier, which is defined as the ratio between the number of correctly classified frames of one class and the total number of frames in the corresponding class, and the coverage of the classifier, which is defined as the ratio between the number of correctly classified frames of one class and the total number of classifications made for the corresponding class. When the thresholds are large, most of the speech frames are correctly classified but more than half of the speech classifications are erroneous. On the other hand, only a small fraction of music frames is classified but with high accuracy. This behavior occurs also when the original thresholds (200 for the running mean of the periodicity counter and 18 for the music continuity counter) are used. This indicates that the original classification algorithm can satisfy the third requirement of the first level classifier for music frames. However, the coverage for the music class is not large enough to allow the original classifier to function as an effective filtering mechanism. This shortcoming resulting from the large threshold values of the original classifier can be resolved at the expense of classification accuracy. As the thresholds decrease, the accuracy for the music class increases slowly at the outset but decreases rapidly at small thresholds. The coverage reacts in the opposite manner to the

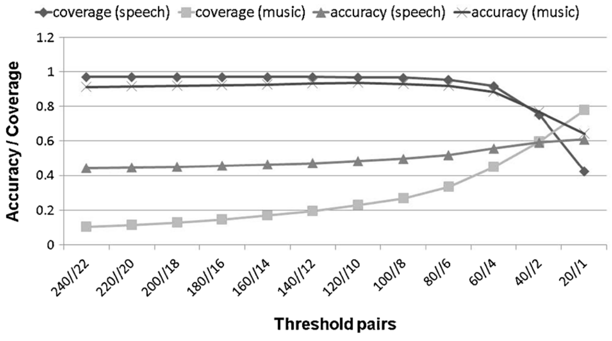

Fig. 1 The accuracy and the coverage of the original speech/music classifier in the SMV for various threshold pairs

accuracy. Therefore, we need to determine threshold values corresponding to a good tradeoff between the accuracy and the coverage. From the figure, threshold sets {80, 6}, {60, 4}, and {40, 2} are potential candidates for a reasonable tradeoff. Note that the accuracy and the coverage that have been described so far are for the first level classifier, not for the overall hierarchical classifier.

In addition to adjusting threshold values, the operation between the two conditions described above can be modified for more accurate filtering. We attempt to replace the OR operation with the AND operation and compare them in terms of accuracy and coverage. Compared with the OR operation, the AND operation produces more accurate filtering but misses some filtering opportunities.

2.2 Analysis of the accuracy of the hierarchical classifier

The speech and music classification accuracies of the hierarchical classifier has a unique relationship with those of the SVM-based classifier at the second level. The speech classification accuracy of the hierarchical classifier is bound to be lower than that of the SVM-based classifier, and the music classification accuracy of the hierarchical classifier tends to be higher than that of the SVM-based classifier. While this behavior is shown in Section 5.2, this subsection is devoted to explain the reason behind the behavior.

The aforementioned inclination is attributable to the fact that the first level classifier can only make a classification as music. The speech classification accuracies of the hierarchical classifier for a test case can be written as:

$$P_S^H = \frac{N_{S_c}^H}{N_S}, \tag{3}$$

where $N_{S_c}^H$ denotes the number of correct speech classifications made by the hierarchical classifier, and N_S represents the number of speech frames in the test case. Similarly, the speech classification accuracies of the SVM-based classifier at the second level can be defined as:

$$P_S^2 = \frac{N_{S_c}^2}{N_S}, \tag{4}$$

where $N_{S_c}^2$ stands for the number of correct speech classifications made by the SVM-based classifier for the same test case. Since the classifications the first level classifier makes are only for the music class, and the incorrectly classified speech frames by the first level classifier cannot be corrected by the second level classifier, the number of correct speech classifications by the hierarchical classifier, $N_{S_c}^H$, can be represented as as $N_{S_c}^2 - \alpha \cdot N_{S_{ic}}^1$, where $N_{S_{ic}}^1$ and α denote the number of incorrectly classified speech frames by the first level classifier, and the proportion between the number of the speech frames that are incorrectly classified by the first level classifier but correctly classified by the second level classifier and the number of incorrectly classified speech frames by the first level classifier, respectively. Therefore, (3) can be rewritten as:

$$P_S^H = P_S^2 - \alpha \cdot \frac{N_{S_{ic}}^1}{N_S}. \tag{5}$$

This equation explains why the speech classification accuracy of the hierarchical classifier has a strong likelihood of being lower than that of the SVM-based classifier at the second level.

In a similar fashion, the music classification accuracy of the hierarchical classifier for the test case can be expressed as:

$$P_M^H = \frac{N_{M_c}^H}{N_M}, \tag{6}$$

where $N_{M_c}^H$ denotes the number of correct music classifications made by the hierarchical classifier, and N_M represents the number of music frames in the test case. Likewise, the music classification accuracies of the SVM-based classifier at the second level can be defined as:

$$P_M^2 = \frac{N_{M_c}^2}{N_M}, \tag{7}$$

where $N_{S_c}^2$ represents the number of correct music classifications made by the SVM-based classifier. Since the number of correct music classifications by the hierarchical classifier is the sum of the number of correctly classified music frames by the second level classifier and the number of frames that is not correctly classified by the second level classifier but correctly classified by the first level classifier, the number of correct music classifications by the hierarchical classifier, $N_{M_c}^H$, can be expressed as $N_{M_c}^2 + \beta \cdot N_{M_c}^1$, where $N_{M_c}^1$ denotes the number of correctly classified music frames by the first level classifier, and β represents the ratio between the number of the music frames that are correctly classified only by the first level classifier and the number of music frames that are correctly classified by the first level classifier. By incorporating this expression, (6) becomes

$$P_M^H = P_M^2 + \beta \cdot \frac{N_{M_c}^1}{N_M}, \tag{8}$$

from which it is evident that the music classification accuracy of the hierarchical classifier is at least equal to that of the SVM-based classifier at the second level.

3 Skipping classification requests by virtue of inter-frame correlations

In this section, we introduce an efficient technique to reduce the classification load on the SVM-based classifier by skipping classification requests based on inter-frame correlations generally found in speech/music frames. Before the skipping mechanism is explained in detail, motivations for the mechanism is presented in the following.

3.1 Motivations for the skipping mechanism

The skipping mechanism is motivated by the two observations: significant reduction in computational load of SVM-based classifications by reducing the number of classification requires and strong correlation between adjacent frames. In this subsection, these observations are described.

The decision function of SVMs is defined as

$$f(\mathbf{x}(t)) = \sum_{i=1}^{M} \alpha_i^* y_i K\left(\mathbf{x}_i^*, \mathbf{x}(t)\right) + b^* \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

where $\mathbf{x}_i^*$ is the i^{th} vector of M support vectors, $\mathbf{x}(t)$ is the t^{th} input frame vector, and $K(\cdot)$ represents the kernel function used when input vectors are not linearly separable. Optimization bias b^* and Lagrange multiplier α^* are obtained by solving a quadratic programming problem [17]. If the output of the decision function is larger than η, it is classified as H_1

(speech hypothesis), but if otherwise, it is classified as H_0 (music hypothesis). If a radial basis function (RBF) is used as the kernel function, $K\left(\mathbf{x}_i^*, \mathbf{x}(t)\right)$ is defined as follows:

$$K\left(\mathbf{x}_i^*, \mathbf{x}(t)\right) = \exp\left(-\gamma \parallel \mathbf{x}_i^* - \mathbf{x}(t) \parallel^2\right) \tag{10}$$

where γ is the kernel parameter of RBF associated with the width of RBF. Since RBF is frequently used for linearly inseparable cases and a related work [3] chose RBF as the kernel function, RBF is chosen as the kernel function in the present study.

From (9) and (10), it is shown that the entire set of support vectors need to be acquired from memory and used in computations for every input frame vector. Clearly, skipping input frames can significantly shorten execution time and conserve energy, and the savings in time and energy are linearly proportional to the number of input frames skipped. This is the first observation that motivate the skipping mechanism.

The skipping mechanism is also motivated by the strong correlations in speech/music frames, which are attributable to their inherent structures. The speech/music signals used in our experiments are composed of three distinct segments: speech, music, and silence. Each segment tends to last for at least several seconds, existing as a group of frames [13]. Accordingly, it is highly probable that the current frame belongs to the same class as the previous frames when considering the inter-frame correlation of consecutive frames. Using this property, we can predict the class of a sequence of input frames based on the classification information of the previous frames without actually performing classifications. Due to the strong correlation, we can assume that

$$P(H(n) = H_i \mid H(n-1) = H_i) > P(H(n) = H_i) \quad i = 0, 1 \tag{11}$$

where $H(n)$ denotes the correct hypothesis for the nth frame. Note that this assumption also holds when more previous frames are considered. Based on this assumption, the probability for the dth frame from the nth frame to belong to the same class as the nth frame is expressed as

$$P(H(n+d) = H_i \mid H(n) = H_i) = \varepsilon \cdot P(H(n+d) = H_i) \quad i = 0, 1 \tag{12}$$

where ε is described as

$$\varepsilon \propto P(H(n) = H_i) \cdot \frac{L_{seg} - d}{L_{seg}} \tag{13}$$

where L_{seg} represents the length of a speech or music segment. This equation reflects the following two aspects. First, we do not have the actual class information of the nth frame so that we have to rely on the probability that $H(n)$=H_i is true, which can be obtained from the SVM output. Second, the inter-frame correlation weakens as the distance between frames increases. ε is supposed to be larger than one when d is small due to the strong correlation depicted in (11) but decreases as d increases as expressed in (13). Due to these two negative aspects, the classification accuracy might be degraded if the skipping mechanism is not applied judiciously.

3.2 The proposed skipping mechanism

The core of the skipping mechanism is the condition that determines when to start skipping classification requests. The condition should be chosen so that classification requests are skipped as many time as possible without lowering classification accuracy. To avoid accuracy degradation with the skipping mechanism, we adopt two efficient mechanisms that incorporate inter-frame correlation. The first one considers previous classifications for initiating a series of skips. At first, N previous classifications are grouped into two sets S_{speech} and S_{music} according to the following conditions.

$$C(t-i) \in S_{speech} \quad \text{if} \quad C(t-i) = 1, \quad i = 1, \cdot\cdot N \tag{14}$$

$$C(t-i) \in S_{music} \quad \text{if} \quad C(t-i) = -1, \quad i = 1, \cdot\cdot N \tag{15}$$

where $C(t-i)$ denotes the i^{th} previous classification for the current classification $C(t)$, and its value is either one for speech class or negative one for music class. After N previous classifications are assigned to either the speech or the music set, the cardinality of each set is compared with its corresponding threshold to determine whether a sequence of skips can be triggered. The class is set to speech if the following condition is met:

$$|S_{speech}| \geq \eta_c^{speech} \tag{16}$$

On the other hand, the class is set to music if the following condition holds true.

$$|S_{music}| \geq \eta_c^{music} \tag{17}$$

Note that η_c^{speech} and η_c^{music} indicate the threshold for the speech class and the threshold for the music class, respectively.

The second mechanism utilizes previous SVM outputs, defined by (9). First, the average of N previous SVM outputs is computed as follows:

$$\mu(t) = \frac{1}{N} \sum_{i=1}^{N} f(x(t-i)) \tag{18}$$

Then, the average value is compared with the class-specific thresholds to determine the initiation of a series of skips, as given by

$$\mu(t) \geq \eta_o^{speech} \tag{19}$$

$$\mu(t) \leq \eta_o^{music} \tag{20}$$

Note that (19) and (20) are the conditions for the speech and music classes, respectively. If one of the conditions is met, a predefined number of incoming input frames are skipped. This predefined number will be referred to as the skipping length, NUM_{skip} in this paper.

4 Combining filtering and skipping for more efficient implementation

The filtering and skipping algorithms described in previous sections can be combined for greater reductions in execution time and energy consumption. Roughly speaking, there are two ways of combining the two algorithms. In the first method, the filtering technique is performed before the skipping technique, so that the skipping technique is applied only to the input frames that have not been filtered. In the other method, the filtering mechanism is followed by the skipping mechanism. The first method preserves the filtering opportunities but misses some of the skipping opportunities because filtering interferes with skipping. On the other hand, the second combination preserves the skipping opportunities but loses some of the filtering opportunities. In the following subsection, the first method is described in detail. The second combination method is not explained here because the details for the second combination are similar to those for the first one.

4.1 Algorithm for combining filtering and skipping mechanisms

Figure 2 depicts the flow chart of the first combination method. Before the flow chart is explicated in detail, a few variables in the chart need to be described. Here, $N_{skip}(t)$ is a variable that represents the number of skips to be performed at a given input frame index t. If $N_{skip}(t)$ is positive, input frame $x(t)$ is skipped, and the variable is decremented by one. If $N_{skip}(t)$ is zero, no more skipping is allowed. When a series of skips is initiated, $N_{skip}(t)$ is set to NUM_{skip}, which indicates a predefined length of skipping. Here, C_{prev} represents the class (either speech or music) for the skipped frames, which is determined by the skipping mechanism according to the recorded previous classifications denoted as $C(t-i)$ (i=1, 2, · ·, N). Note that N denotes the number of considered previous classifications. To indicate the eligibility for the filtering operation, a flag F_{filter} is introduced, and similarly, F_{skip} indicates whether the conditions for the skipping operation is satisfied. The following equations describe how these flags are set.

$$F_{filter}(t) = \begin{cases} 1 & \text{if } \left(\overline{C}_{pr}(t) > \eta_{pr}\right) \; OP \; \left(\overline{C}_M(t) > \eta_M\right) \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where η_{pr}, η_M, and OP represent the threshold triplet described in Section 2.

$$F_{skip} = \begin{cases} 1 & \text{if } \left(|S_{speech}| \geq \eta_c^{speech}\right) \; OR \; \left(|S_{music}| \geq \eta_c^{music}\right) \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where the condition for F_{skip} to be set to one was presented in Section 3.

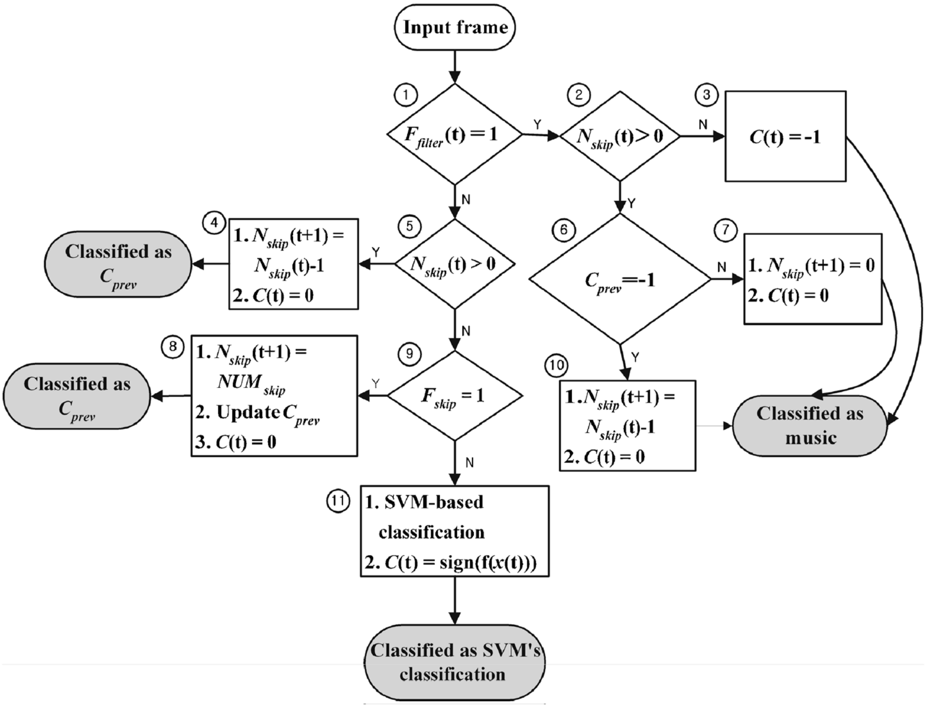

Fig. 2 Flow chart for the combined mechanism, in which filtering is applied before skipping

For a given input frame, F_{filter} is first checked (block 1). If filtering is possible, the current input is classified as music because the first level classifier only filters out frames classified as music. In cases where the filtering mechanism interrupts a streak of skips (e.g., the Y path from block 2), $N_{skip}(t+1)$ is updated differently depending on C_{prev}. If C_{prev} is not music, then $N_{skip}(t+1)$ is reset to zero (block 7) because the contradiction may lead to a series of misclassifications. Otherwise, $N_{skip}(t+1)$ is set to $N_{skip}(t)$ minus one (block 10) as in the original skipping mechanism to avoid skips after a long sequence of filtered frames. Note that frames are inclined to be filtered out in groups due to the way the running average of the periodicity counter and the music continuity counter are updated, as described in Section 2. These update details can be formulated as:

$$N_{skip}(t+1) = \begin{cases} NUM_{skip} & \text{if } (F_{filter}(t) = 0)\ \ AND\ \ (N_{skip}(t) = 0) \\ & \qquad AND\ \ (F_{skip} = 1)\ \text{ is true} \\ 0 & \text{if } (F_{filter}(t) = 1)\ \ AND\ \ (N_{skip}(t) > 0) \\ & \qquad AND\ \ (C_{prev} = 1)\ \text{ is true} \\ N_{skip}(t) - 1 & \text{if } (F_{filter}(t) = 0\ AND\ N_{skip}(t) > 0)\ \ OR \\ & \qquad (F_{filter}(t) = 1\ AND\ N_{skip}(t) > 0\ AND \\ & \qquad C_{prev} = -1)\ \text{ is true} \end{cases} \tag{23}$$

Note that this equation also contains the condition for setting $N_{skip}(t+1)$ to NUM_{skip}, which will be explained shortly. If the filtering does not interfere with a series of skips (the N path from block 2), then $C(t)$ is updated by the classification of the first level classifier. The complete rule for updating $C(t)$ is given by

$$C(t) = \begin{cases} sign(f(x(t))) & \text{if } (F_{filter}(t) = 0)\ \ AND\ \ (N_{skip}(t) = 0)\ \ AND \\ & \qquad (F_{skip} = 0)\ \text{ is true} \\ -1 & \text{if } (F_{filter}(t) = 1)\ \ AND\ \ (N_{skip}(t) = 0)\ \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

where $sign(f(x(t)))$ denotes the binary classification outcome (-1 or 1) made by the support vector machine.

On the other hand, if filtering is not possible, the hybrid mechanism is identical to the skipping mechanism. At first, $N_{skip}(t)$ is checked to determine whether the current frame can be included as a part of a series of skipped frames (block 5). If the skipping of the current frame is allowed (positive $N_{skip}(t)$ value), then $N_{skip}(t)$ is decremented (block 4) before the classification for the current frame is made according to C_{prev}, which was determined based on recorded previous classifications $C(t-i)$ ($i = 1, 2, \cdot\cdot, N$). The updating rule for C_{prev} is given by

$$C_{prev} = \begin{cases} 1 & \text{if } \left(|S_{speech}| \geq \eta_c^{speech}\right)\ \text{ is true} \\ -1 & \text{if } \left(|S_{speech}| \geq \eta_c^{speech}\right)\ \text{ is true} \end{cases} \tag{25}$$

If the current frame is not allowed to be skipped, F_{skip} is checked, which is updated based on (22) (block 9). Apparently, right after the end of a series of skips, it is not possible for the conditions to be satisfied because $C(t)$ is reset when a new sequence of skips begins (block 8). If a new series of skips is allowed, $N_{skip}(t+1)$ is set to NUM_{skip} as described in (23), and C_{prev} is set according to (25) (block 8). If the skipping conditions are not yet met, the SVM-based classifier at the second level is used to produce the classification for the current frame, and $C(t)$ is updated accordingly.

5 Experiments and results

5.1 Experimental setup

To evaluate the proposed techniques, we measured execution time, energy consumption, and classification accuracy of the proposed techniques and compared them. The SVM-based classifier in the SMV [3] was used as the base case for which the proposed techniques were applied. The experiments were performed on the TIMIT speech database [7] and commercial music CDs. From the CDs, songs from five different genres (blues, classic, hip-hop, jazz, and metal) were collected, each of them about 5 min in length. For training the SVM-based classifier, 20 min of data from the speech and music database was selected. The length of each speech segment varied from 6 to 12 s, and the length of each music segments varied from 28 to 30 s. For testing, four sets of data with different segment lengths were prepared. One set had the same segment length as the training data, but the other three sets had segment lengths of 2, 5, and 10 s, respectively. The reason for having different segment lengths was to examine the sensitiveness of the proposed techniques to segment length. Each set contained ten files (two files per genre), and each file was composed of 15 alternating speech, music, and silence segments. Note that the length of silence segments was set to 0.5 s for all test sets except for the one that had the same segment length as the training data, whose silence segments were randomly selected to be between 3 and 15 s. Test data were selected to have no overlap with the training data. All data were sampled at 8 kHz with a frame size of 20 ms. In order to determine the correctness of a classification, we manually labeled each frame and compared it with the corresponding classification result from the classifier.

The six features adopted for the SVM-based classifier [11] were used with no modification. They consisted of the running average of energy, the running mean of the reflection coefficients, the running mean of the partial residual energy, the running mean of the normalized pitch correlation, the running average of the periodicity counter, and the music continuity counter, as listed in [8]. Note that these six features were the parameters originally generated within the SMV. As mentioned previously, RBF was used as the kernel function for the SVM, and its kernel width was set to 0.01, for which reasonable performance can be achieved with a reasonable number of support vectors.

To measure the execution time and energy consumption, we utilized a processor simulator based on simplescalar [3]. Since this processor simulator models both processor architecture and power consumption in detail, execution time and energy consumption can be measured simultaneously. We configured the simulation environment to simulate an embedded system with an ARM instruction set architecture (ISA). In order to compile the classifiers and control logic to ARM ISA, a cross-compiler based on gcc-2.95 was used. Table 1 captures the processor configuration used for the experiment.

Table 1 Parameters for the processor simulator

Pipeline	800 Mhz, 5 stages, in-order, dual issue
Cache	64KB separate Inst/Data Cache 4way, 32B line, 1 cycle latency
Memory	64bit bus, 100 cycle latency, 1 port
FPU	1 adder (2 cycle, pipelined) 1 multiplier (4 cycle, pipelined)

5.2 Results for the filtering mechanism

In this subsection, the experimental results for the hierarchical classifier described in Section 2 are presented: the accuracy and filtered ratio of the classifier and the influence of segment length on the classifier's performance. For experiments, we selected three sets of threshold triplets, which consisted of two thresholds and one operation (either AND or OR), based on filtering accuracy and coverage: a balanced set, an accuracy-oriented set, and a coverage-oriented set. In each set, two threshold triplets were prepared. The threshold triplets in the accuracy-oriented set were chosen such that their corresponding filtering accuracy and coverage were higher that 95 % and 10 %, respectively. Similarly, filtering accuracy and coverage of at least 85 % and 25 % were required for the balanced set, and 75 % and 35 % were required for the coverage-oriented set. The purpose of having three sets of threshold triplets was to attain more accurate insight into the tradeoff between classification accuracy and filtered ratio.

5.2.1 The classification accuracy and filtered ratio of the hierarchical classifier

In this subsection, we present the classification accuracy and filtered ratio obtained with the threshold triplets selected based on filtering accuracy and coverage. Figure 3 presents the classification accuracies (overall, speech, and music) and filtered ratios of the hierarchical classifier when the selected threshold triplets are used. The x-axis shows six threshold triplets from left to right: the accuracy-oriented set ("80 AND 4" and "70 AND 2"), the balanced set ("80 OR 6" and "60 OR 4"), and the coverage-oriented set ("30 AND 1" and "40 OR 2"). The classification accuracies presented in the figure were normalized to those of the original SVM-based classifier. The filtered ratio is defined as the ratio between

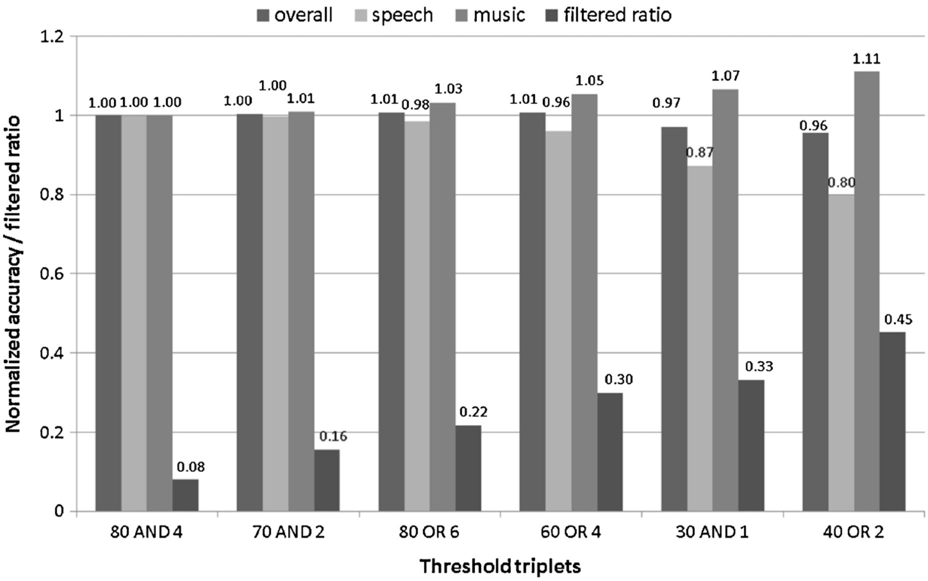

Fig. 3 Classification accuracies (overall, speech, and music) and filtered ratios of the hierarchical classifier for the selected threshold triplets. Presented accuracies were normalized to those of the original SVM-based classifier

the number of filtered frames and the total number of frames, representing the fraction of frames for which the SVM-based classifier at the second level is not used. Note that the filtered ratio is different from the filtering coverage introduced earlier in that the filtering ratio also considers the number of incorrect classifications made by the first level classifier, and that it is computed with respect to all frames instead of only the music frames.

For the accuracy-oriented threshold triplets, which achieve the highest filtering accuracy and the lowest filtering coverage, classification accuracies of the original SVM-based classifier are successfully maintained, but the filtered ratios are limited to less than 20 %. With the balanced threshold triplets, the classification accuracies for the speech class are degraded lightly (2 % and 4 %), but a filtered ratio of up to 30 % is obtained. The overall accuracies are maintained due to the improved classification accuracies for music frames. The coverage-oriented triplets attained the highest filtered ratios at the expense of significant degradation in classification accuracy. However, the overall accuracies are not reduced as much as the accuracies for the speech class because of the substantial classification improvements in the music class. Among the three sets, the balanced set shows the best tradeoff between classification accuracy and filtered ratio. Also, the increased music classification accuracy and decreased speech classification accuracy of the hierarchical classifier compared with the SVM-based classifier are observed, which is in accordance with (5) and (8).

Choosing a right threshold triplet depends on the design requirement imposed on the hierarchical speech/music classifier. If classification accuracy outweighs execution time and power consumption in design priority, from the threshold triplets given in the figure, “70 AND 2” should be chosen; however, if practicality is more important, a threshold triplet with a higher filtered ratio (than “70 AND 2) could be chosen depending on the allowable accuracy degradation. The applicability of a chosen threshold triplet to a different data set and the applicability of the filtering mechanism to other classifier in SMV context will be examined and reported in Sections 5.6 and 5.7.

5.2.2 The impact of the segment length on the classification accuracy and filtered ratio

Real-life speech and music segments inherently have various lengths. Thus, evaluating how the filtering mechanism works with various segment lengths is essential in estimating the performance of the proposed mechanism. Figures 4 and 5 show the impact of segment length on the performance of the hierarchical classifier. Note that the segment length denoted by ’varied’ represents a speech segment length between 6 and 12 s and a music segment length between 28 and 30. Figure 4 shows the threshold triplet “80 OR 6” of the balance set, and Fig. 5 shows for the threshold triplet “30 AND 1” of the coverage-oriented set. These two threshold triplets were chosen to contrast the case of high filtering accuracy and decent filtered ratio (“80 OR 6”) with the case of low accuracy and high filtered ratio (“30 AND 1”).

For “80 OR 6”, the classification accuracies do not seem to be affected by the segment length, but the filtered ratio increases as segment length increases. The lower filtered ratios for the shorter segments are attributable to the fact that music segments are not long enough for the running average of periodicity counter and music continuity counter to reach their respective thresholds. Similar behavior is observed with “30 AND 1” except for its higher filtered ratio and noticeably lower speech classification accuracy. Apparently, “30 AND 1” achieves higher filtered ratio than “80 OR 6” due to its lower threshold values. The larger degradation in speech classification accuracy is caused by the lower filtering accuracy of the threshold triplet “30 AND 1” (77.5 %) compared with that of “80 OR 6” (92.2 %).

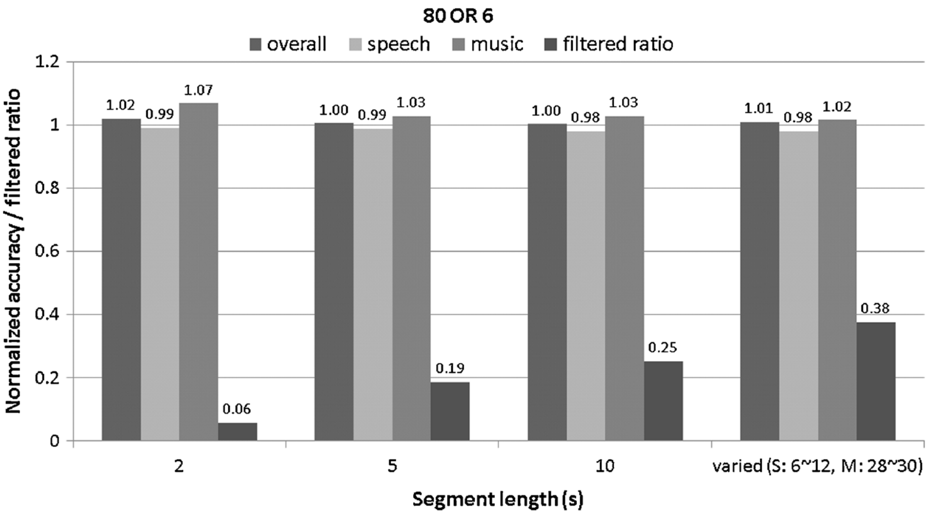

Fig. 4 Impact of segment length on the accuracy and the filtered ratio of the hierarchical classifier when {80 *OR* 6} is used as the threshold triplet. Presented accuracies were normalized to those of the original SVM-based classifier

From the figures, it is observed that the filtering mechanism embedded in the hierarchical classifier is not susceptible to segment length with respect to classification accuracy but generates higher filtered ratio for longer segments. This insensitivity to segment length and the inverse proportionality of the filtered ratio and the classification accuracy suggest that one set of threshold can accomplish the goal of maximizing the filtered ratio for a given accuracy requirement.

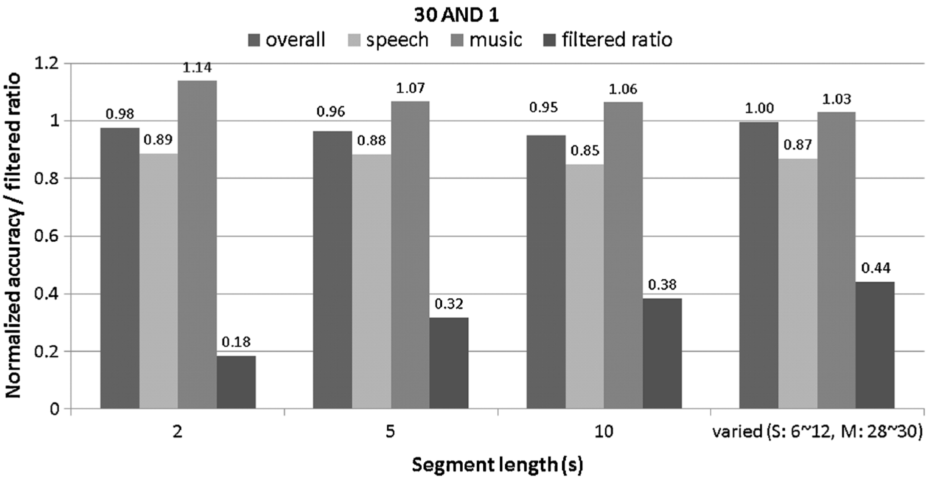

Fig. 5 Impact of segment length on the accuracy and the filtered ratio of the hierarchical classifier when {30 *AND* 1} is used as the threshold triplet. Presented accuracies were normalized to those of the original SVM-based classifier

5.3 Results for the skipping mechanisms

In this subsection, the experimental results for the skipping mechanisms described in Section 3 are presented. We analyze the impact of the number of previous classification results considered for determining the initiation of a series of skips, the impact of the thresholds, and the impact of skipping length in terms of classification accuracy and skipped ratio for various segment lengths. But, the impact of the thresholds is not presented here because no significant changes were observed in classification accuracy and skipped ratio. Since two skipping mechanisms show similar results, only the result of the mechanisms based on previous classifications is presented.

5.3.1 The impact of the number of previous classification results considered in the skipping mechanisms

While it is evident that the number of previous classification results considered for determining the suitability of skipping is directly related to the skipped ratio, it is not obvious how the number of considered previous classification results affects the classification accuracy. Therefore, in this subsection, we analyze the impact of previous classification results used in the skipping mechanisms on the classification accuracy and the skipped ratio.

Figure 6 shows the influence of the number of previous classifications considered in the skipping mechanism on the classification accuracy and the skipped ratio for various segment lengths. Figure 6a and b show the results for the mechanism that considers previous classifications when the skipping length is set to 8 and 32, respectively. Skipping lengths 8 and 32 represent the case with no classification accuracy degradation and the case with significant classification accuracy degradation, respectively. While the x-axis indicates the length of each segment in the test data, the figure shows how the overall classification accuracy and skipped ratio are affected by the number of considered previous classifications and the segment length. Note that the accuracy was normalized to that of the original SVM-based classifier. The threshold used for determining the initiation of a series of skips was set conservatively such that all of the considered previous classifications were required to be in the same class.

When the skipping length is set to 8 (Fig. 6a), the overall accuracy is well maintained close to that of the original SVM-based classifier for every segment length, implying that the classification accuracy is not severely affected by the number of considered previous classifications. On the other hand, the filtered ratio is a function of the number of previous classifications considered because skipping is not allowed until the predefined number of consecutive SVM classifications is collected. In contrast, when the skip length is set to 32 (Fig. 6b), the overall accuracy drops noticeably, and the impact of the number of considered previous classifications is slightly more evident, especially when the segment length is 2 or 5 frames. The reason why a short segment suffers from lower classification accuracy than longer segments for a long skip length such as 32 frames is simply that a shorter segment contains fewer frames. Specifically, since the classification accuracy of a segment is the ratio between the number of correctly classified frames and the number of frames in the segment, the similar number of mis-classifications caused by a long skip length at the onset of a segment results in lower classification accuracy for a shorter segment.

As we can see from Fig. 6, the number of previous classification outcomes does not significantly affect the classification accuracy unless the skipping length is too long. However, since the skipped ratio is strongly influenced by the number of previous classification

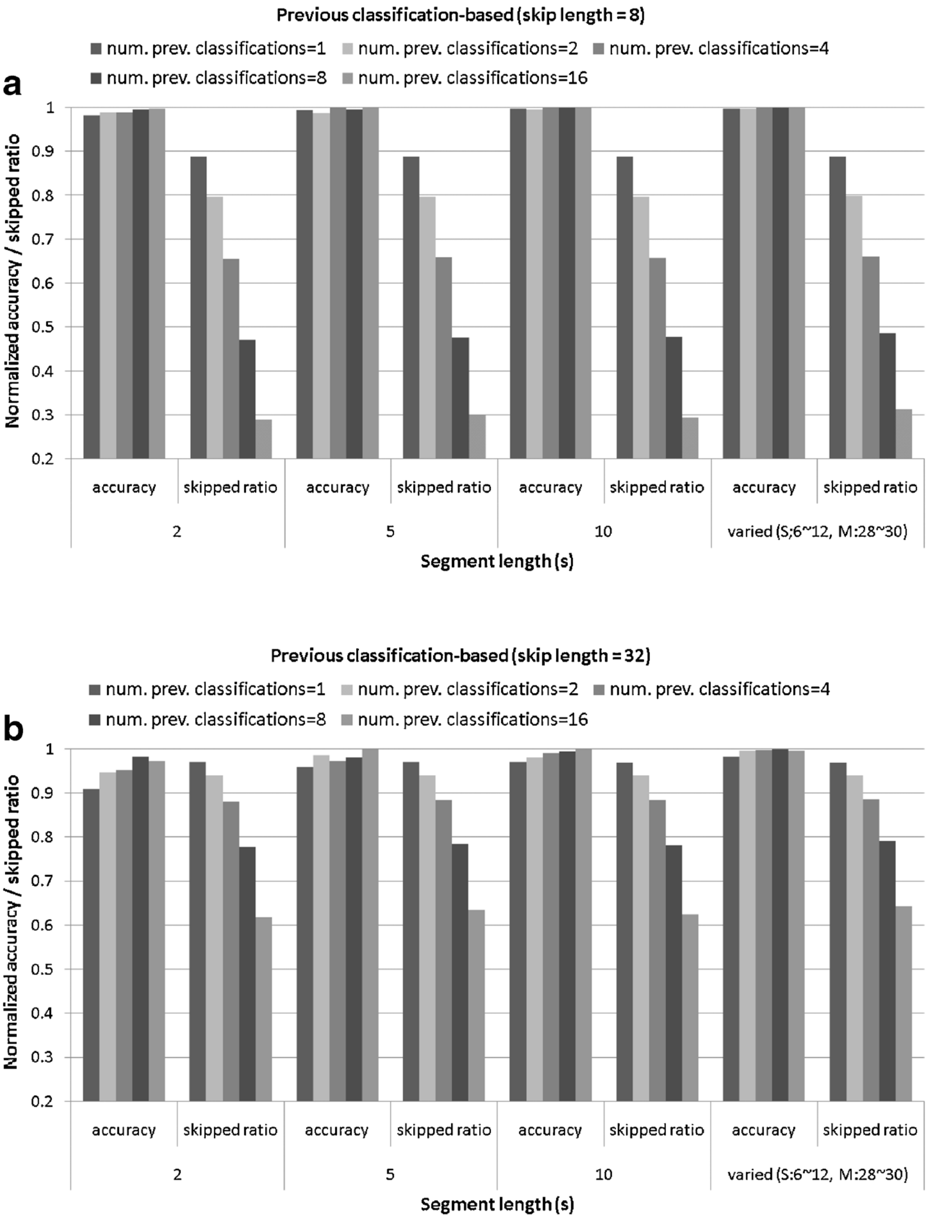

Fig. 6 Impact of the number of previous classifications considered in the skipping mechanism on the classification accuracy and the skipped ratio for various segment lengths at a skipping length of (**a**) 8 and (**b**) 32. Presented accuracies were normalized to those of the original SVM-based classifier

outcomes, the latter should be chosen judiciously, according to the given design requirement. For example, 16 should be selected for designs with stringent requirements on classification accuracy, but a shorter length can be used for designs with stringent resource constraints.

5.3.2 The impact of skipping length in the skipping mechanisms

Once the proposed skipping mechanisms decide to skip incoming input frames, a predefined number of frames are skipped consecutively. Here, it is easily conceivable that this predefined skipping length affects the classification accuracy and the skipped ratio, to which execution time and energy consumption are proportional. However, the impact of the skipping length should be known in order to fine-tune the skipping mechanisms for achieving the highest possible skipped ratio without influencing the classification accuracy. Therefore, in this subsection, we investigate the impact of skipping length on the classification accuracy and the skipped ratio.

Figure 7 depicts the impact of the skipping length on the classification accuracy and the skipped ratio for four different segment lengths with the number of considered previous classifications set to 1 (Fig. 7a) or 16 (Fig. 7b). Note that the threshold was set to the number of considered previous classifications. As shown in the figure, remarkable skipped ratios around 97 % are achieved when the skipping length and the number of considered previous classifications are 32 and 1, respectively, at the expense of accuracy deterioration for all segment lengths. Particularly, when the segment length is 2 s, the degradation is the most obvious (9 %). In contrast, when 16 previous classifications are considered, with the skipping length of 32, the degradation in the classification accuracy is alleviated, but the skipped ratio is significantly reduced. This drastic change in the skipped ratio is caused because at least 16 classifications should be made by the SVM-based classifier before a skip is allowed. Note that the skipped ratio is not as sensitive as the classification accuracy to the segment length even when one previous classification is considered.

For the selection of the skipping length, it is generally advised to conform to the design requirement. However, it should be noted that the number of considered previous classification outcomes and the skipping length need to be selected together because they are two dominant parameters that determine the skipped ratio. For example, in Fig. 7a, which shows the results when a previous SVM classification is considered, almost perfect classification accuracy of 98.8 % is obtained with high skipped ratio of 94.1 % when the skipping length is set to 16 frames. On the other hand, in Fig. 7b, which depicts the results when 16 previous SVM classifications are considered, a similar accuracy (99.9 %) is achieved with a lot lower skipped ratio of 46.2 % when the skipping length is also set to 16 frames. From this example, it is seen that the same skipping length results in totally different skipped ratios according to the number of previous classification results.

From the sensitiveness analysis of the parameters of the skipping mechanism to the classification accuracy and skipped ratio, it has been observed that unless the skipping length is too long (e.g., 32 in our analysis), the skipping length, the number of considered previous classifications, and the threshold are not affected by segment length in terms of classification accuracy. This indicates a good general applicability of the parameters of the skipping mechanism to various segment lengths. The applicability to other classifiers and a different data set will be presented in Sections 5.6 and 5.7.

5.4 Results for the combined mechanisms

Two different ways of combining the filtering and skipping mechanisms were described in Section 4. In this subsection, we examine how these two combinations differ in terms of classification accuracy and filtering/skipping ratios. In addition, the filtering threshold

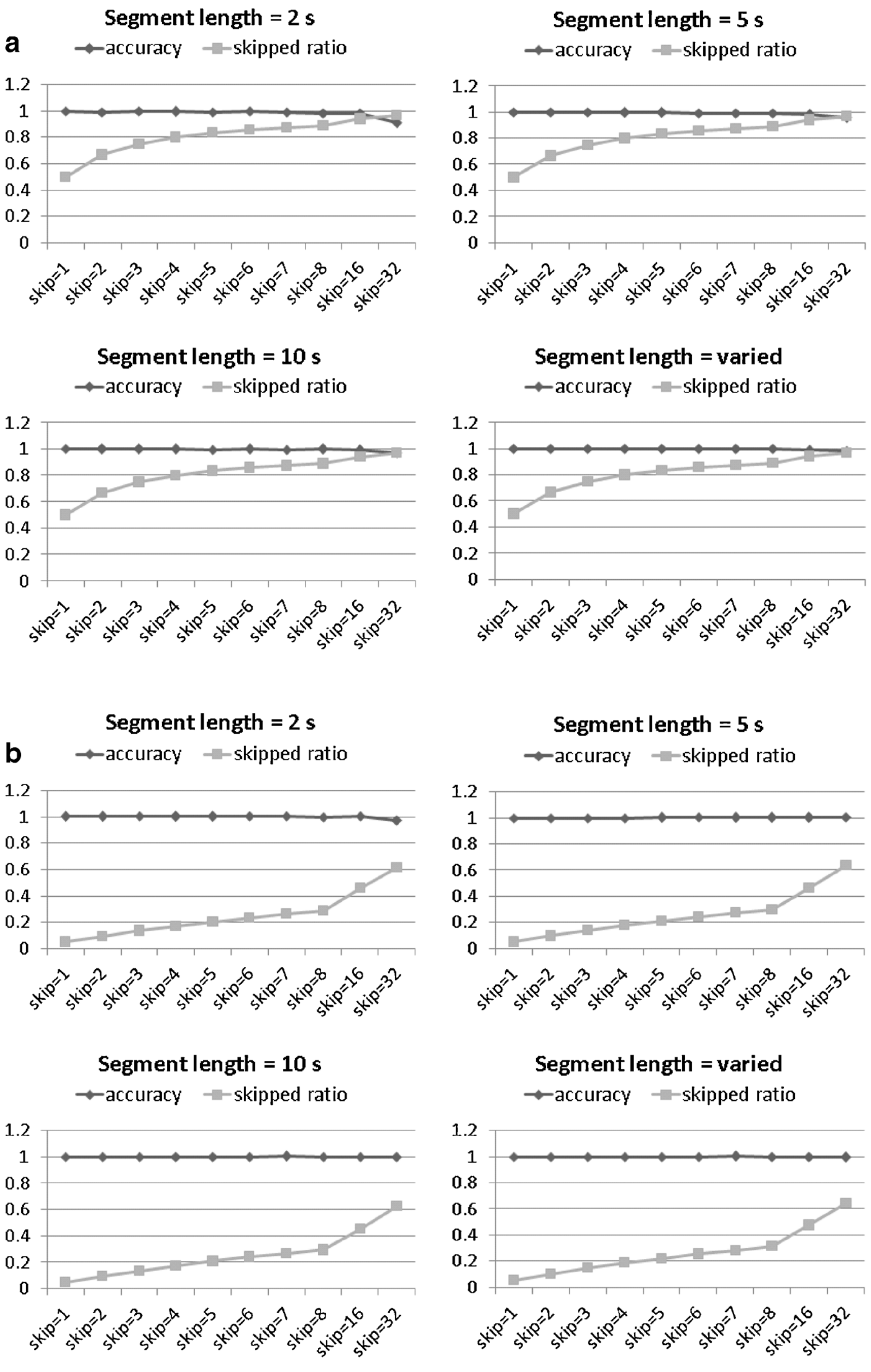

Fig. 7 Impact of the skipping length on the classification accuracy and the skipped ratio for various segment lengths (**a**) the number of considered previous classification is 1 (**b**) the number of considered previous classification is 16. Presented accuracies were normalized to those of the original SVM-based classifier

values and the skipping mechanism parameters such as the number of considered previous classification results and the skipping length were varied to analyze their impacts on classification accuracy and filtering/skipping ratios.

To facilitate understanding of the difference between the two combined mechanisms, Table 2 shows the classification accuracies and filtered/skipped ratios of the two mechanisms for four different segment lengths. In the table, $F1$ refers to the filtering mechanism based on a threshold triplet in the balanced set (“60 OR 4”) introduced in Section 5.2, and $S1$ represents the skipping mechanism whose skipping length, number of considered previous classifications, and the thresholds for skipping decisions are set to 8, 4, and 4, respectively. This choice of parameters causes the corresponding skipping mechanism produce a skipped ratio of 0.66 and 1 % degradation in music classification accuracy. Note that, in all the tables in this subsection, the accuracies were normalized to those of the original SVM-based classifier.

From Table 2, it is easily seen that both combined mechanisms achieve very similar accuracies and overall ratios (sum of filtered and skipped ratios). The overall accuracies attained by the combined mechanisms are higher than those of the original SVM-based classifier for three out of four segment lengths and show, in the worst case, a small 1.2 % degradation. However, some accuracy loss and accuracy gain are observed for speech and music classes, respectively because the filtering mechanism, a part of the combined mechanisms, can only classify frames as music as explained in Section 5.2.1. In the combined mechanism “$S1 \rightarrow F1$”, since skipping is performed first, the skipped ratios are the same as those obtained when the skipping mechanism is used alone and does not change much for different segment lengths. Similarly, in the mechanism “$F1 \rightarrow S1$”, in which filtering is performed first, the filtered ratios are the same as those obtained when the filtering mechanism is used alone. However, unlike the skipped ratio, the filtered ratio increased as the segment length increased, implying that, for shorter segments, the threshold triplet is not as effective as for longer segments. This increase in filtered ratio results in a decrease in the

Table 2 Comparison of classification accuracies and filtered/skipped ratios between the combined mechanism in which the filtering mechanism is applied first and the combined mechanism in which the skipping mechanism is applied first

	$F1 \rightarrow S1$	$S1 \rightarrow F1$
Overall (speech / music) accuracy		
2 s	1.028 (0.982 / 1.110)	1.017 (0.984 / 1.075)
5 s	1.003 (0.978 / 1.036)	1.002 (0.978 / 1.032)
10 s	0.988 (0.934 / 1.050)	0.988 (0.938 / 1.047)
Varied	1.007 (0.962 / 1.019)	1.009 (0.965 / 1.020)
Overall (filtered / skipped) ratio		
2 s	0.700 (0.133 / 0.567)	0.703 (0.046 / 0.657)
5 s	0.741 (0.245 / 0.496)	0.741 (0.082 / 0.659)
10 s	0.778 (0.355 / 0.424)	0.778 (0.118 / 0.660)
Varied	0.818 (0.464 / 0.354)	0.817 (0.155 / 0.663)

$S1$ and $F1$ denote balanced skipping and filtering mechanism, respectively. Presented accuracies were normalized to those of the original SVM-based classifier

Table 3 Impact of filtering threshold values on classification accuracies and filtered/skipped ratios of the combined mechanism in which filtering is applied first

	$F1 \rightarrow S1$	$F2 \rightarrow S1$
Overall (speech / music) accuracy		
2 s	1.028 (0.982 / 1.110)	0.990 (1.004 / 0.965)
5 s	1.003 (0.978 / 1.036)	0.999 (1.006 / 0.990)
10 s	0.988 (0.934 / 1.050)	1.001 (1.002 / 1.001)
Varied	1.007 (0.962 / 1.019)	1.002 (1.006 / 1.001)
Overall (filtered / skipped) ratio		
2 s	0.700 (0.133 / 0.567)	0.655 (0.000 / 0.655)
5 s	0.741 (0.245 / 0.496)	0.662 (0.011 / 0.651)
10 s	0.778 (0.355 / 0.424)	0.689 (0.095 / 0.594)
Varied	0.818 (0.464 / 0.354)	0.732 (0.211 / 0.521)

$S1$ and $F1$ represent balanced skipping and filtering mechanisms, respectively, and $F2$ denotes accuracy-oriented filtering mechanism. Presented accuracies were normalized to those of the original SVM-based classifier

skipped ratio because the filtering mechanism reduces the skipping opportunities. Nevertheless, the overall ratio, which is the sum of the filtered and skipped ratios, increases as the segment length increases.

In Table 3, the impact of different filtering threshold triplets on the combined mechanism that favors the filtering mechanism is illustrated. In the table, $F2$ represents the filtering mechanism with a threshold triplet in the accuracy-oriented set (“80 AND 4”) introduced in Section 5.2. This threshold triplet is expected to be more accurate in classification and less effective in coverage compared with the “60 OR 4” used for $F1$. An evidence of this more accurate classification capability was found in speech classification accuracy. No degradation in speech classification accuracy is observed with “$F2 \rightarrow S1$”, but the accuracy is obtained at the expense of coverage; its overall and filtered ratio decrease significantly compared with those of “$F1 \rightarrow S1$”. As for the skipped ratio, “$F2 \rightarrow S1$” achieves a higher ratio than “$F1 \rightarrow S1$” because more skipping opportunities are preserved for the former due to its lower filtered ratio. Note that these higher skipping ratios of “$F2 \rightarrow S1$” partially compensate for its lower filtered ratios, resulting in smaller differences in the overall ratios compared with those in the filtered ratios.

Table 4 shows the effects when different skipping parameters are used in a combined mechanism for a given filtering mechanism. In the table, $S2$ denotes a conservative skipping mechanism whose skipping length, number of considered previous classifications, and the skipping threshold were set to 4, 16, and 16, respectively. Originally, as a skipping mechanism, $S2$ achieves a slightly better accuracy with a considerably lower skipped ratio compared with $S1$. When $S2$ is combined with $F2$, this behavior is transferred to “$F2 \rightarrow S2$”. While a slightly higher accuracy is obtained with “$F2 \rightarrow S2$”, the overall ratio significantly decreases due to its much lower skipped ratio than that of “$F2 \rightarrow S1$”, making it a futile design attempt.

Table 4 Impact of different skipping mechanism parameters on classification accuracies and filtered/skipped ratios of the combined mechanism in which filtering is applied first

	$F2 \rightarrow S1$	$F2 \rightarrow S2$
Overall (speech / music) accuracy		
2 s	0.990 (1.004 / 0.965)	1.001 (1.004 / 0.996)
5 s	0.999 (1.006 / 0.990)	1.001 (1.002 / 0.999)
10 s	1.001 (1.002 / 1.001)	1.003 (1.003 / 1.004)
Varied	1.002 (1.006 / 1.001)	1.004 (1.003 / 1.004)
Overall (filtered / skipped) ratio		
2 s	0.655 (0.000 / 0.655)	0.172 (0.000 / 0.172)
5 s	0.662 (0.011 / 0.651)	0.188 (0.011 / 0.177)
10 s	0.689 (0.095 / 0.594)	0.252 (0.095 / 0.157)
Varied	0.732 (0.211 / 0.521)	0.356 (0.211 / 0.145)

$F2$ denotes accuracy-oriented filtering mechanism, and $S1$ and $S2$ represent balanced and accuracy-oriented skipping mechanisms, respectively. Presented accuracies were normalized to those of the original SVM-based classifier

5.5 Comparison of the effectiveness of the proposed mechanisms

In this subsection, we compare the proposed techniques in terms of execution time and energy consumption as well as classification accuracy and filtered and/or skipped ratio. For this purpose, we first set a design constraint, then selected parameters that conformed to the constraint, and finally evaluated the techniques with the test data. The assumed design constraint stated that no more than 2 % degradation was allowed in speech and music classification accuracies as well as overall accuracy. The goal was to achieve as large as possible reductions in execution time and energy consumption abiding by the constraint. Once the design constraint is given, parameters for the skipping and filtering mechanisms should be selected such that the design constraint is not violated.

Table 5 summarizes the parameters chosen experimentally. Note that these parameters are not the only ones but exemplary ones that achieve the goal. The threshold triplet is the one presented in Section 5.2. Among the triplets that conform to the constraint, it attains the highest filtered ratio. For the skipping mechanism, only one previous classification was used for determining whether to start skipping. This implies that there is no need to examine more previous classifications or for more sophisticated algorithms. Note that the combined mechanism based on previous classifications was used.

The effectiveness of each mechanism was evaluated in terms of classification accuracy, execution time, and energy consumption, as shown in Table 6. Note that the accuracies

Table 5 Parameters selected for the filtering, skipping, and combined mechanisms

Filtering	Filtering threshold triplet	{80, OR, 6}
Skipping	Skip length	15
	The number of considered classifications / threshold	1 / 1
Combined	Filtering threshold triplet	{80, OR, 6}
	Skip length	16
	The number of considered classifications / threshold	1 / 1

Table 6 Comparison of the effectiveness of the proposed mechanisms in terms of classification accuracy, execution time per frame, and energy consumption per frame

	Overall	Speech	Music	Filtered and/or skipped ratio	Exec. Time (ms)	Energy (mJ)
Baseline	1.000	1.000	1.000	0.000	16.501	58.808
Filtering	1.011	0.983	1.032	0.221	12.914	45.843
Skipping	0.993	0.998	0.985	0.937	1.021	3.541
Combined	1.005	0.986	1.010	0.954	0.804	2.927

Presented accuracies were normalized to those of the original SVM-based classifier

were normalized to those of the original SVM-based classifier (denoted as *baseline* in the table) [11] for easier comparison. The execution time was measured in milli-seconds, and the energy consumption was measured in milli-joules. As shown in the table, no classification accuracy degradation is larger than 2 %. While the accuracies in the table are average accuracies over four different segment lengths, no accuracy degradation larger than 2 % is caused for any segment length. For the filtered and /or skipped ratio, as expected, the combined mechanism is the best, resulting in the shortest execution time and the lowest energy consumption. Compared with the original SVM-based classifier, the combined mechanism, on average, can process one input frame about 20 times faster while consuming 20 times less energy. The reductions in execution time and energy consumption achieved by the skipping mechanism are comparable to those achieved by the combined mechanism. Again, it should be noted that this significant improvement is obtained by using only one previous classification outcome. The high skipped ratio and the high classification accuracy of the skipping and combined mechanisms indicate that using more previous classification outcomes or more sophisticated algorithms is not necessary.

5.6 Applicability of the proposed mechanisms to other speech/music classifiers

Since the skipping mechanism utilizes a general property of speech and music signals, it can be applied to speech/music classifiers that process signals on a short frame basis. While the filtering mechanism, as it is, is limited to the classifiers that target the SMV codec, the idea of having a simpler classifier that filters out classification requests to a more sophisticated

Table 7 Parameters selected for the GMM-based speech/music classifier and the WSVM-based speech/music classifier

Mechanism	Parameter	GMM-based	WSVM-based
Filtering	Filtering threshold triplet	{50, AND, 2}	{80, OR, 6}
Skipping	Skip length	10	16
	The number of considered classifications / threshold	1 / 1	1 / 1
Combined	Filtering threshold triplet	{100, OR, 8}	{80, OR, 6}
	Skip length	14	16
	The number of considered classifications / threshold	1 / 1	1 / 1

Table 8 The effectiveness of the proposed mechanisms for the GMM-based speech/music classifier and the WSVM-based speech/music classifier (F: filtering, S: skipping, C: combined)

	Overall	Speech	Music	Filtered and/or skipped ratio	Exec. Time (ms)	Energy (mJ)
GMM	1.000	1.000	1.000	0.000	0.107	0.429
GMM+F	0.995	0.982	1.012	0.173	0.089	0.355
GMM+S	0.992	1.000	0.981	0.909	0.011	0.042
GMM+C	0.984	0.980	0.985	0.945	0.007	0.027
WSVM	1.000	1.000	1.000	0.000	17.200	61.148
WSVM+F	1.006	0.982	1.028	0.232	13.213	46.982
WSVM+S	0.990	1.006	0.980	0.944	0.961	3.423
WSVM+C	1.007	1.000	1.013	0.959	0.699	2.493

Presented accuracies were normalized to those of the baseline classifiers without filtering and/or skipping mechanisms

main classifier is generally applicable. In this subsection, we apply the proposed techniques to two speech/music classifiers that target the SMV codec: the gaussian mixture model-based classifier [16] and the weighted support vector machine-based classifier [12]. The speech/music classifier based on the gaussian mixture model (GMM) was motivated by the observation that the SMV features could be well represented by GMM. Through training, the speech and music classes are each modeled by separate GMMs, and input frames are classified based on the likelihood ratio (LR) test. The weighted support vector (WSVM), an enhanced version of the original SVM, was motivated by the observation that each feature had a different contribution to the final SVM output. In the WSVM, each feature in an input vector is multiplied by a different weight, which is obtained based on a minimum classification error (MCE) method.

Table 7 shows the parameters selected for the two classifiers with the maximum allowable accuracy degradation set to 2 %. Note that the parameters are not the same as those for the SVM-based classifier. Although the differences in the parameters are not significant and do not produce substantially different results, they need to be chosen such that any accuracy deteriorations stay within the 2 % margin. In other words, if there is no stringent requirement on classification accuracy, one set of parameters or at most a little tuning is enough. The effectiveness of the proposed techniques with the two classifiers is presented in Table 8 in terms of classification accuracies, filtered and/or skipped ratio, execution time in milliseconds, and energy consumption in milli-joules. Note that the accuracies are normalized

Table 9 Parameters selected for the new data set

Filtering	Filtering threshold triplet	{90, OR, 7}
Skipping	Skip length	16
	The number of considered classifications / threshold	1 / 1
Combined	Filtering threshold triplet	{90, OR, 7}
	Skip length	17
	The number of considered classifications / threshold	1 / 1

Table 10 The effectiveness of the proposed mechanisms for the new data set

	Overall	Speech	Music	Filtered and/or skipped ratio	Exec. Time (ms)	Energy (mJ)
Base	1.000	1.000	1.000	0.000	16.502	58.809
Filtering	1.008	0.980	1.051	0.288	11.757	42.132
Skipping	0.979	0.989	0.987	0.940	1.012	3.544
Combined	0.993	0.987	1.011	0.959	0.681	2.486

Presented accuracies were normalized to those of the original SVM-based classifier

to those of their corresponding baseline classifier (either the GMM-based classifier or the WSVM-based classifier).

According to the execution times of the two classifier, the GMM-based classifier is far less complex than the SVM-based classifier, and the WSVM-based classifier is a little more complex due to the additional weight assigning step. However, regardless of the complexity, the proposed techniques can achieve the effectiveness similar to that attained for the SVM-based classifier. From this, it is proved that the proposed mechanisms can be applied to other classifiers without compromising their effectiveness. It is also worth noting that the filtered ratio of the filtering mechanism for the GMM-based classifier is lower than those for the other classifiers. This is because the number of music frames incorrectly classified as speech by the simpler first level classifier in the filtering mechanism is larger for the GMM-based classifier due to its relatively higher classification accuracy for speech class. Therefore, in order to conform to the requirement on the classification accuracy, a more conservative threshold triplet ({50, AND, 2}) was selected, thus reducing the filtered ratio.

5.7 Effectiveness of the proposed mechanisms with a different data set

In this subsection, we evaluate the effectiveness of the proposed techniques against a different test data set and verify the applicability to the other fields. To obtain the new data set, we transmitted the original test data set by using the Bluetooth protocol and the G.722 codec [14] running on the transmitter and the receiver powered by CSR's $Kalimba\ BC05$ chip [4]. The data collected at the receiver are used as the new test data. This new test data represent a more realistic situation where received data via a different codec is being retransmitted by using the SMV codec. The new test data contain audible degradation in sound quality, which may be caused by the lossy codec and channel noise. Note that the support vectors trained with the original training data are used for testing.

The classification accuracy achieved with the new test data is comparable to the original classification accuracy; 2.57 % degradation for speech, 0.3 % improvement for music, and 0.3 % decrease in overall accuracy. This shows the robustness of the SVM-based classifier against different test data. As with the previous section, the parameters for the skipping, filtering, and combined mechanisms are selected such that any accuracy degradation stay within the 2 % margin. The selected parameters, summarized in Table 9, are similar to the original parameters shown in Table 5. This indicates that no significant tuning of the parameters is necessary, thereby proving the general applicability of the chosen parameters. The effectiveness of the proposed mechanisms with the new test data is also very similar to that with the original test data as shown in Table 10. From this table, it is found that the proposed techniques are also very effective even with different test data, confirming the general applicability of the techniques under various multimedia services.

6 Conclusion

This paper proposed two techniques that reduce the computational complexity and energy consumption of an SVM-based classifier in the SMV framework in order to facilitate effective implementation of the classifier. First, we proposed a hierarchical classifier in which the first level classifier can filter out a significant portion of input frames, reducing the burden on the SVM-based classifier at the second level. Second, a skipping mechanism based on inter-frame correlation considerably reduced the number of classification requests to the SVM-based classifier, leaving only a fraction of the original requests for the SVM-based classifier. Furthermore, these techniques can be easily combined for more reduction in execution time and energy consumption. Through experimenting with other speech/music classifiers and a different data set, it was found that the mechanisms are generally applicable and their parameters do not require an extensive tuning. Future research directions may include the development of techniques that reduce the execution time and energy consumption of the SVM-based classifier at lower levels, such as the architecture or instruction level.

References

1. 3GPP2 Specification (2004) Selectable Mode Vocoder (SMV) service option for wideband spread spectrum communication systems. 3GPP2-C.S0030-0, v3.0
2. Burges C (1996) Simplified support vector decision rules. In: Proceedings of IEEE international conference on machine learning. Bari, Italy, pp 71–77
3. Burger D, Austin TM (1997) The simplescalar tool set, version 2.0, Tech Rep 1342. University of Wisconsin-Madison, Computer Sciences Department
4. CSR (2006) BlueCore5 Multimedia. http://www.csr.com/products/16/bluecore5-multimedia. Accessed 28 June 2013
5. Dardas NH, Silva JM, Saddik AE (2012) Target-shooting exergame with a hand gesture control. Multimed Tools Appl doi:10.1007/s11042-012-1236-4
6. Farrugia RA, Debono CJ (2012) A support vector machine approach for detection and localization of transmission errors within standard H.263++ decoders. IEEE Trans Multimed 11(7):1323–1330
7. Fisher WM, Doddington GR, Goudie-Marshall KM (1986) The DARPA speech recognition research database: specifications and status. In: Proceedings of DARPA workshop speech recognition, pp 93–99
8. Gao Y, Shlomot E, Benyassine A, Hyssen J, Su H, Murgia C (2001) The SMV algorithm selected by TIA and 3GPP2 for CDMA applications. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing. Salt Lake City, pp 709–712
9. Ho T (2005) An efficient method for simplifying support vector machines. In: Proceedings of international conference on machine learning. Bonn, pp 617–624
10. Hu H, Li Y, Liu M, Liang W (2012) Classification of defects in steel strip surface based on multiclass support vector machine. Multimed Tools Appl. doi:10.1007/s11042-012-1248-0
11. Kim SK, Chang JH (2009) Speech/music classification enhancement for 3GPP2 SMV codec based on support vector machine. IEICE Trans Fundam Electron Commun Comput Sci E92-A(2):630–632
12. Kim SK, Chang JH (2010) Discriminative weight training for support vector machine-based speech/music classification in 3GPP2 SMV codec. IEICE Trans Fundam Electron Commun Comput Sci E93-A(1):316–319
13. Lavner Y, Ruinskiy D (2009) A decision-tree-based algorithm for speech/music classification and segmentation. EURASIP J Audio Speech Music Process 2009:1–14

14. Maitre X (1988) 7 KHz audio coding within 64 kbit/s. IEEE J Sel Areas Commun 6(2):283–298
15. Nakashima Y, Babaguchi N, Fan J (2012) Intended human object detection for automatically protecting privacy in mobile video surveillance. Multimedia Systems 18(2):157–173
16. Song J, An H, Song Y, Choi S, Jeong D, Lee S (2011) Enhancement of speech/music decision employing GMM for SMV codec. In: Proceedings of international congressional image and signal processing, pp 2182–2185
17. Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10(5):988–999
18. Zhan Y (2005) Design efficient support vector machine for fast classification. Pattern Recog 38(1):157–161

Chungsoo Lim received the BS and ME degree in electrical engineering from Inha University in 1996 and 1999, respectively and the MS and Ph.D. degree in computer engineering from University of Maryland at College Park in 2004 and from North Carolina State University in 2009, respectively. He was a research professor at Hanyang University, Seoul, Korea. Currently, he is an assistant professor at Korea National University of Transportation, Chungju, Korea. His research interests include computer architecture, digital signal processing, and software development and optimization.

Joon-Hyuk Chang received the BS degree in electronics engineering from Kyungpook National University, Daegu, Korea in 1998 and the MS and PhD degrees in electrical engineering from Seoul National University, 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 Korea Institute of Science and Technology (KIST), 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 Engineering at Inha University, Incheon, Korea. Currently, he is an associate professor in the School of Electronic Engineering at Hanyang University, Seoul, Korea. His research interests are in speech coding, speech enhancement, speech recognition, audio coding, and adaptive signal processing. He is a senior member of IEEE. He is a winner of IEEE/IEEK IT young engineer of the year 2011. He is serving as Editor-in-chief of the Signal Processing Society Journal of the IEEK.