IEICE TRANS. INF. & SYST., VOL.E95–D, NO.3 MARCH 2012

LETTER

Improvement of SVM-Based Speech/Music Classification Using Adaptive Kernel Technique

Chungsoo LIM[†], ***Nonmember and*** **Joon-Hyuk CHANG**[††a)], ***Member***

SUMMARY In this paper, we propose a way to improve the classification performance of support vector machines (SVMs), especially for speech and music frames within a selectable mode vocoder (SMV) framework. A myriad of techniques have been proposed for SVMs, and most of them are employed during the training phase of SVMs. Instead, the proposed algorithm is applied during the test phase and works with existing schemes. The proposed algorithm modifies a kernel parameter in the decision function of SVMs to alter SVM decisions for better classification accuracy based on the previous outputs of SVMs. Since speech and music frames exhibit strong inter-frame correlation, the outputs of SVMs can guide the kernel parameter modification. Our experimental results show that the proposed algorithm has the potential for adaptively tuning classifications of support vector machines for better performance.

key words: *SVM, SMV, adaptive kernel, sigmoid*

1. Introduction

Recent progress in mobile communication and semiconductor technology enables us to receive diverse multimedia services with our personal wireless devices such as cell phones. Since these services and technologies are still emerging, there are still problems to be resolved. One issue is how to effectively utilize limited bandwidth. To fully take advantage of limited bandwidth, variable bit-rate speech coding has been researched. An example of variable bit-rate coding is the selectable mode vocoder (SMV) speech codec adopted by the third-generation partnership project 2 (3GPP2). Variable bit-rate coding requires speech/music classification, and the SMV codec incorporates a simple speech/music classification algorithm for different bit-rate allocations [1]. Among speech/music classification techniques, a technique that combines features from the SMV codec and support vector machines (SVMs) recently showed great potential [2]. To take advantage of the potential, we propose an algorithm that adaptively tunes classifications of SVMs specifically tailored for speech and music frames.

While existing techniques targeting SVMS [3]–[5] are employed during the training phase of SVMs, a technique that applies during the test phase of SVMs was proposed [6]. This technique assigns a different weight to each input element based on its contribution to generalization error in order to improve the performances of SVMs. Since this algorithm is designed for the test phase, it can be utilized with other techniques developed for the training phase, yielding an increase in synergistic performance.

In this paper, we propose a simple but effective algorithm employed during the test phase, in which the decision function of SVM is evaluated. For the kernel function, we use a radial basis function (RBF) [5]. With this classical kernel function, we can manipulate the RBF width parameter. Since the influence of the kernel parameter of RBF on SVM outputs is not clearly understood, we analyze the impact of the kernel parameter on SVM classifications and show how to control SVM classifications with the kernel parameter. Even though we know how to control outputs of the decision function, adaptively tuning SVMs requires guidance that minimizes classification error. For this purpose, we propose a novel scheme for adjusting the kernel parameter by using strong correlations of speech/music activity among neighboring input frames [7].

2. Brief Review of SMV Codec

At first, we briefly review the SMV codec since it is utilized for the system. SMV, an adaptive multi-rate speech codec adopted as a standard in 3GPP2, is capable of efficiently utilizing limited bandwidth [1]. It features four average data rate and four operational modes dynamically chosen based on the types of input frames and statuses of communication channels respectively. The parameters used for speech/music classification are running average of energy, running mean of the reflection coefficients, running mean of the partial residual energy, running mean of the normalized pitch correlation, running average of the periodicity counter, and music continuity counter as listed in [2].

3. RBF Kernel Parameter Modification to Enhance SVM

Radial basis function (RBF) is one of the classical kernel functions adopted in SVM [5] and is used for classification problems that are not linearly separable. In this section, we vary the parameter of RBF to examine the influence of it on the outputs of SVMs. When input vector $\mathbf{x}$ is linearly separable, the decision function results in the following:

$$f(\mathbf{x}(t)) = \sum_{i=1}^{M} \alpha_i^* y_i \langle \mathbf{x}_i^*, \mathbf{x}(t) \rangle + b^* \quad (1)$$

Manuscript received July 19, 2011.
Manuscript revised September 22, 2011.

[†]The author is with School of Electronic Engineering, Mokpo National University, Mokpo, Korea.
[††]The author is with School of Electronic Engineering, Hanyang University, Seoul, Korea.
a) E-mail: jchang@hanyang.ac.kr (Corresponding author)

DOI: 10.1587/transinf.E95.D.888

where $\mathbf{x}_i^*$ is the i^{th} vectors of M support vectors and $\mathbf{x}(t)$ is the t^{th} input frame vector. Optimization bias b^* and Lagrange multiplier α^* are acquired by solving a quadratic programming problem. If input vectors are not linearly separable, its decision function should incorporate a kernel function, given by

$$f(\mathbf{x}(t)) = \sum_{i=1}^{M} \alpha_i^* y_i K(\mathbf{x}_i^*, \mathbf{x}(t)) + b^*. \quad (2)$$

As mentioned above, RBF is used as the kernel function and is defined as the following:

$$K(\mathbf{x}_i^*, \mathbf{x}(t)) = \exp(-\gamma \parallel \mathbf{x}_i^* - \mathbf{x}(t) \parallel^2) \quad (3)$$

where γ is the kernel parameter of RBF and is associated with the width of RBF. If we add a small positive value (δ) to γ to increase it, the new modified RBF kernel becomes

$$\widetilde{K}(\mathbf{x}_i^*, \mathbf{x}(t)) = \exp(-(\gamma + \delta) \cdot \|\mathbf{x}_i^* - \mathbf{x}(t) \|^2). \quad (4)$$

If we rewrite Eq. (4), it can be expressed as

$$\widetilde{K}(\mathbf{x}_i^*, \mathbf{x}(t)) = \exp(-\gamma\|\mathbf{x}_i^* - \mathbf{x}(t) \|^2) \cdot \exp(-\delta\|\mathbf{x}_i^* - \mathbf{x}(t) \|^2) \quad (5)$$

As seen above, the modified kernel function $\widetilde{K}(\mathbf{x}_i^*, \mathbf{x}(t))$ is the product of the original kernel function $K(\mathbf{x}_i^*, \mathbf{x}(t))$ and $\exp(-\delta\|\mathbf{x}_i^* - \mathbf{x}(t) \|^2)$. Here, if a positive δ is added to γ, the added term $\exp(-\delta\|\mathbf{x}_i^* - \mathbf{x}(t) \|^2)$ is a value between 0 and 1, making $\widetilde{K}(\mathbf{x}_i^*, \mathbf{x}(t))$ smaller than $K(\mathbf{x}_i^*, \mathbf{x}(t))$. On the contrary, if a negative δ is added to γ, $\exp(-\delta\|\mathbf{x}_i^* - \mathbf{x}(t) \|^2)$ is a value larger than 1, making $\widetilde{K}(\mathbf{x}_i^*, \mathbf{x}(t))$ larger than $K(\mathbf{x}_i^*, \mathbf{x}(t))$.

For a clear analysis, we vary δ and observe how the outputs of the decision function changes as Table 1 contains the result. The first row shows six δ values added to the kernel parameter γ. The second row holds the ratio between the number of transitions from positive to negative outputs and the number of positive outputs before γ is adjusted. The last row represents the ratio between the number of transitions from negative to positive outputs and the number of negative outputs before γ is modified. This table is populated with 50 data files that will be explained in Sect. 5.

If a positive δ is added to γ, outputs of SVMs are likely to change from positive to negative values, and the reverse transitions rarely occur. On the other hand, a negative δ produces the opposite behavior. To achieve more insight about this behavior, it is desirable to consider the decision function that follows.

$$f(\mathbf{x}(t)) = f^+(\mathbf{x}(t)) - f^-(\mathbf{x}(t)) + b^* \quad (6)$$

where $f^+(\mathbf{x}(t))$ is one part of $f(\mathbf{x}(t))$ that corresponds to the case where y_i is 1, and $f^-(\mathbf{x}(t))$ is the other part of $f(\mathbf{x}(t))$ that corresponds to the case where y_i is -1. If a negative δ is added to γ, both $f^+(\mathbf{x}(t))$ and $f^-(\mathbf{x}(t))$ become larger, and, at the same time, the difference between them tends to be larger too. Thus, for the case where $f^+(\mathbf{x}(t))$ is larger than $f^-(\mathbf{x}(t))$ but $f(\mathbf{x}(t))$ is negative due to a negative bias b^*, bigger difference between $f^+(\mathbf{x}(t))$ and $f^-(\mathbf{x}(t))$ resulted from a negative δ can switch the polarity of $f(\mathbf{x}(t))$. This is the major reason for the polarity switch from negative to a positive value when a negative δ is added to γ. The polarity switch from a positive to negative value when a positive δ is used can be similarly explained.

Table 1 Impact of kernel parameter δ on the polarity of $f(\mathbf{x}(t))$.

δ	0.03	0.06	0.09	−0.03	−0.06	−0.09
$+ \rightarrow -$	11.41	23.69	53.15	0.22	0.27	0.26
$- \rightarrow +$	0.08	0.15	0.22	6.05	15.32	32.76

For our speech/music classification, because we label music as -1 and speech as 1, if a positive value is added to γ, more classifications are made for music, while the number of classifications as speech decreases. One more thing we can learn from the table is that the number of transitions is proportionate to δ. From these two observations, we can concluded that we are able to control output of the decision function with δ. Nonetheless, there is one thing missing: there is no rigorous rule for adjusting γ. Section 4 introduces a rule for when to adjust the kernel parameter.

4. Guidance Based on Correlations among Adjacent Frames

Although we know how to control outputs of SVMs with the kernel parameter of RBF, we still need a rule for adaptively tuning the outputs of SVMs. To adjust the kernel parameter in such a way as to reduce error, we propose a way to use strong correlations among adjacent frames in the actual speech and music signals. The speech and music signals used in our experiments consist of three distinct segments: speech segments, music segments, and silence segments. Since each segment is at least a few seconds long, there are a group of frames in each segment. Hence, a frame is likely to be in the same class as its previous frames with high probability. Actual probability that the current frame is in the same class as the previous ones is nearly 100%.

However, we can not use this correlation because we do not have *a priori* information about the class of each frame. Therefore, we should depend on previous classifications made by SVMs. Figure 1 shows the block diagram of the proposed algorithm. The output of the SVM ($f(\mathbf{x}(t))$) is first smoothed by using

$$f_s(\mathbf{x}(t)) = k_f f_s(\mathbf{x}(t-1)) + (1 - k_f) f(\mathbf{x}(t)) \quad (7)$$

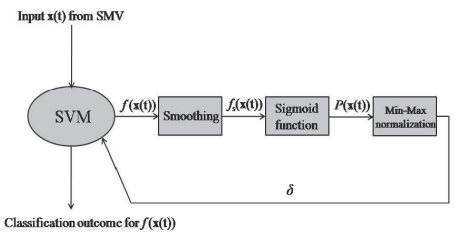

Fig. 1 Block diagram of the proposed algorithm.

where $f_s(\mathbf{x}(t))$ is a smoothed SVM output and k_f is the smoothing factor. Since $f(\mathbf{x}(t))$ depending on the input signal may change abruptly, particularly in onset and offset regions, $f_s(\mathbf{x}(t))$ could increase or decrease rapidly, respectively. Thus, the smoothing operation can reduce intermediate mis-classification while the delay due to the smoothing does not causes a serious problem at the onset region [8]. The smoothed outputs are then converted to probabilities according to

$$P(f_s(\mathbf{x}(t))) = P(H(t) = H_0 \mid f_s(\mathbf{x}(t)) = \frac{1}{1 + \exp(Af_s(\mathbf{x}(t)) + B)} \quad (8)$$

where $H(t)$ denotes the correct hypothesis for the tth frame, and H_0 and H_1 designate hypothetical music and speech, respectively. Also, A and B are the parameters obtained via maximum likelihood estimation for modeling the distribution of SVM outputs. To obtain these parameters, the model trust algorithm in [9] based on Levenberg Marquardt algorithm is adopted. While a probability is a value between zero and one, δ, an additive modification to the kernel parameter γ, can be either a positive or a negative value. Therefore, the output from the sigmoid function needs to be mapped to an appropriate range by the following Min-Max normalization.

$$\delta = \frac{P(f_s(\mathbf{x}(t))) - P_{min}}{P_{max} - P_{min}} \cdot (\delta_{max} - \delta_{min}) + \delta_{min} \quad (9)$$

where P_{max} and P_{min} are the maximum and minimum value of $P(\mathbf{x}(t))$, respectively, which are observed over the whole training set, and δ_{max} and δ_{min} are the maximum and minimum values of δ, respectively, which should be set to limit δ to a proper range. This normalization step ensures that δ has an appropriate value derived from the outputs of the SVM.

As shown in Table 1, in order for the output of SVM to change from a positive value to a negative value, the kernel parameter should be incremented, and vice versa. Therefore, the parameter should be incremented for music and decremented for speech. To achieve this, $P(f(\mathbf{x}(t)))$ in Eq. (8) is defined to be the probability that a frame is a music frame. For example, $P(f(\mathbf{x}(t)))$ tends to be low for a speech frame if the previous SVM outputs have been correct. Consequently, low $P(f(\mathbf{x}(t)))$ results in a negative δ according to the Min-Max normalization, increasing the probability for the frame to be a speech frame.

5. Experiments and Results

In this section, the proposed technique is evaluated. For experiments, we use the TIMIT speech database [10] and commercial music CDs. 50 database files were formed from TIMIT database and music CDs and used for 10-fold cross-validation. The speech portion of the database was extracted from TIMIT database and the music portion was created from music CDs of five different genres: metal, jazz, blues, hip-hop, and classical music. Note that the speech-overlapped music frames in music segments were classified as music for proper training because it is evident that we need to assign the higher bit rate to this signal in enhancing music quality at the SMV encoding. All data were sampled at 8 kHz with a frame size of 20 ms. Each database file is composed of five speech segments (6 - 12 s each), five music segments (28 - 32 s each), and ten periods of silence (randomly selected between 3 and 15 s), and these segments alternated. Each of these files contained music segments from one genre only. The six parameters introduced in Sect. 2 were concatenated to form a feature vector for each frame.

For training sigmoid parameters described in Sect. 4, three-fold cross-validation was adopted. As a result of the training, parameters A and B were set to -1.91531 and 1.29230, respectively. The two parameters that define the range for δ are 0.08 for the upper bound and -0.08 for the lower bound, and the smoothing factor k_f was set to 0.9.

Figure 2 shows how δ improves classification performance for a sequence of speech/music segments. Figure 2 (a) is an actual waveform of a test file, and Fig. 2 (b) is the classification result of the previous SVM-based scheme [2]. Figure 2 (d) shows δ to be added to the kernel parameter, and its impact can be seen by comparing Fig. 2 (b) and Fig. 2 (c), which is the result of the proposed algorithm. During speech segments, δ becomes negative making the mis-classified frames switch to speech frames, and vice versa. Note that classification outcome 1 and 2 denote the speech and music classes, respectively.

Table 2 shows the performance improvement due to the proposed enhancement. We compared the proposed algorithm with the original algorithm in [1], the previous SVM-based algorithm in [2], and the weight training algorithm in [6], which are denoted by *SMV*, *SVM*, and *WT*, respectively. The first column shows five music genres used for the experiments, and the second column has four speech/music classification algorithms. *SVM*, whose kernel parameter was set to 0.1 by the kernel parameter optimization algorithm in [3], is used as a baseline for the proposed and the weight training algorithm. The results summarized in the table are average values from 10 validation runs. P_d for speech and music represents the probability that non-overlapping music and speech frames are correctly classified, and P_d for speech-overlapped music (denoted by *overlap*) is the probability that overlapped frames are classified as music. Note that P_e denotes the error probability that encompasses speech, music, and overlapped frames.

From the table, we can observe that the proposed technique successfully improves the accuracy of SVM-based classification for both non-overlapped and overlapped frames by adaptively adjusting the kernel parameter based on previous outputs of the decision function. It is also discovered that the proposed algorithm outperforms or at least produces comparable performance to the discriminative weight training, the previous SVM-based algorithm, and the original algorithm in SMV.

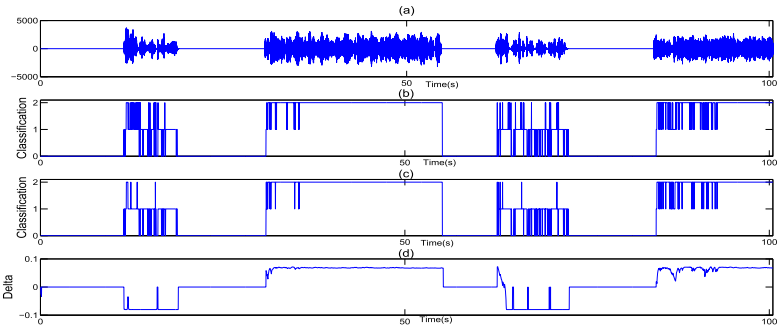

Fig. 2 (a) Waveform of a test file (repeating sequence: silence-speech-silence-music-silence) (b) Results of the previous SVM-based scheme (c) Results of the proposed scheme (d) Delta.

Table 2 Comparison with the original algorithm in SMV, the previous SVM-based algorithm, and weight training algorithm in terms of speech/music detection probability P_d and total error probability P_e.

Class	Method	Speech P_d	Overlap P_d	Music P_d	Total P_e
Blues	SMV [1]	0.882	0.309	0.453	0.488
	SVM [2]	0.839	0.911	0.924	0.093
	WT [6]	0.872	0.944	0.930	0.077
	Proposed	0.926	0.934	0.935	0.066
Classic	SMV	0.860	N/A	0.394	0.511
	SVM	0.739	N/A	0.681	0.307
	WT	0.816	N/A	0.721	0.261
	Proposed	0.753	N/A	0.734	0.256
Hiphop	SMV	1.000	0.120	0.035	0.707
	SVM	0.821	0.901	0.824	0.116
	WT	0.844	0.914	0.648	0.105
	Proposed	0.927	0.931	0.868	0.070
Jazz	SMV	0.975	N/A	0.558	0.358
	SVM	0.719	N/A	0.909	0.130
	WT	0.750	N/A	0.918	0.124
	Proposed	0.845	N/A	0.928	0.088
Metal	SMV	0.989	0.024	0.301	0.727
	SVM	0.758	0.886	0.789	0.157
	WT	0.776	0.915	0.757	0.147
	Proposed	0.839	0.917	0.830	0.116
Avg.	SMV	0.897	0.151	0.348	0.558
	SVM	0.773	0.899	0.825	0.161
	WT	0.812	0.924	0.795	0.143
	Proposed	0.858	0.927	0.859	0.119

6. Conclusions

We have proposed a novel and orthogonal technique that adaptively tunes classifications of SVMs by modifying the RBF kernel parameter based on correlations in speech and music frames. Our experiments show that the proposed enhancement is capable of improving the classification accuracies of SVMs.

Acknowledgement

This work was supported by Priority Research Centers Program through the NRF of Korea funded by the MEST (2011-0022980), and this work was partly supported by the IT R&D program of MKE/KEIT [KI001824] and this work was supported by the research fund of Hanyang University (HY-2011-201100000000210).

References

[1] Y. Gao, E. Shlomot, A. Benyassine, J. Hyssen, H. Su, and C. Murgia, "The SMV algorithm selected by TIA and 3GPP2 for CDMA applications," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing., vol.2, pp.709–712, Orlando, USA, May 2002.

[2] S.-K. Kim and J.-H. Chang, "Speech/music classification enhancement for 3GPP2 SMV codec based on support vector machine," IEICE Trans. Fundamentals, vol.E92-A, no.2, pp.630–632, Feb. 2009.

[3] L.-P. Bi, H. Huang, Z.-Y. Zheng, and H.-T. Song, "New heuristic for determination Gaussian kernel's parameter," Proc. International Conference on Machine Learning and Cybernetics., vol.7, pp.4299–4304, Guangzhou, China, Aug. 2005.

[4] S.S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," Neural Comput., vol.15, no.6, pp.1667–1689, July 2003.

[5] N.E. Ayat, M. Cheriet, and C.Y. Suen, "Automatic model selection for the optimization of SVM kernel," Pattern Recognit., vol.38, pp.1733–1745, Oct. 2005.

[6] 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, vol.E93-A, no.1, pp.316–319, Jan. 2010.

[7] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing., vol.2, pp.1331–1334, Munich, Germany, April 1997.

[8] Y.-D. Cho and A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," IEEE Signal Process. Lett., vol.8, no.10, pp.276–278, Oct. 2001.

[9] J.C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, pp.61–74, MIT Press, 1999.

[10] 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, pp.93–99, Feb. 1986.