VOCAL SEGMENT ESTIMATION IN MUSIC PIECES BASED ON COLLABORATIVE USE OF EEG AND AUDIO FEATURES

Takuya Kawakami, Takahiro Ogawa and Miki Haseyama

Graduate School of Information Science and Technology, Hokkaido University
N-14, W-9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan
E-mail: {kawakami, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

This paper presents a novel estimation method of segments including vocals in music pieces based on collaborative use of features extracted from electroencephalogram (EEG) signals recorded while users are listening to music pieces and features extracted from these audio signals. From extracted EEG features and audio features, we estimate segments including vocals based on Support Vector Machine (SVM) by separately utilizing these two features. Furthermore, the final classification results are obtained by integrating these estimation results based on supervised learning from multiple experts. Therefore, our method realizes multimodal estimation of segments including vocals in music pieces. Experimental results show the improvement of our method over the methods utilizing only EEG or audio features.

Index Terms— electroencephalogram (EEG), multimodal scheme, classification, vocal segment.

1. INTRODUCTION

Recently, various services are provided to search music pieces which users desire from huge amount of music pieces. Generally, these services use metadata of music pieces such as singers' names (often referred to as artists' names), the genre of music pieces, and the meaning of lyrics. However, if such metadata are not attached to music pieces, it becomes difficult for conventional services to find out desired music pieces. In order to solve these problems, it is necessary to provide metadata to music pieces automatically, and we consider these data can be extracted from segments including vocals. Therefore, in this paper, we focus on estimation of these segments, and try to identify the segments where singing voices exist with instrumental accompaniment (vocal segments).

There have been proposed several vocal segment estimation methods [1, 2]. Generally, these methods utilize only audio features extracted from audio signals and input these audio feature vectors into classifiers. However, since their estimation accuracy is not satisfactory, it is necessary to improve the performance of classifiers by using new efficient audio features and introducing a new idea which uses the features extracted from signals other than audio signals.

In the fields of engineering, various trials, which analyze the relationships between music pieces and electroencephalogram (EEG) recorded while users are listening to music pieces, have intensively performed [3–5]. Especially, in [3], they propose an EEG-based emotion recognition method. In this method, it is reported that music stimuli affect the human brain and the effects are observable as EEG signals. Therefore, we can assume that there is validity in a method of utilizing EEG signals for the estimation of vocal segments. However, the conventional EEG-based method [3] is unimodal, and the performance is also limited. So far, there have been no studies where EEG signals are collaboratively utilized with other signals.

In this paper, we propose a novel vocal segment estimation method based on collaborative use of EEG features extracted from EEG signals recorded while users are listening to music pieces and audio features extracted from audio signals of the music pieces. Since our method implements multimodal vocal segment estimation, accurate classification can be expected even when characteristics of audio signals in vocal segments are similar to those in non-vocal segments. In order to utilize EEG and audio features collaboratively, we adopt the method proposed in [6]. This method focuses on obtaining the final classification results from multiple classification results estimated by multiple information sources (annotators). It is reported that the classification, which is based on the integration of multiple classification results by using the method in [6], is more accurate than the classification based on the majority voting. Furthermore, this method assigns higher weights to classification results of the best annotators. In this way, we realize a novel method based on the collaborative use of EEG and audio features for the vocal segment estimation. Consequently, the proposed method can achieve successful estimation of vocal segments in music pieces.

2. VOCAL SEGMENT ESTIMATION BASED ON COLLABORATIVE USE OF EEG AND AUDIO FEATURES

In this section, we explain the proposed method. First, our method extracts EEG feature vectors from EEG signals recorded while users are listening to music pieces and audio feature vectors from these audio signals. Secondly, we input these two feature vectors into classifiers separately and obtain classification results. We call this procedure the 1st step, hereafter. Finally, we estimate vocal segments by integrating the above results using the method in [6]. We call this procedure the 2nd step, hereafter.

This section is organized as follows. In **2.1**, we explain EEG features and audio features used in our method. In **2.2**, the proposed method estimates vocal segments in terms of EEG or audio features in the 1st step. Furthermore, we obtain the final classification results by collaborative use of the classification results based on EEG features and audio features in the 2nd step in **2.3**.

2.1. Feature Extraction

In this subsection, we explain EEG features and audio features used in the proposed method.

EEG Feature Extraction

EEG signals are electrical signals recorded as multiple channel signals from multiple electrodes placed on the scalp. We

Table 1. Features used for audio signals in the proposed method.

Spectral	Centroid
	Width
	Flux
	Rolloff
	Envelope
Volume	
Sideband Energy Ratio	
Zero Crossing Rate	
Pitch	
MFCC	

calculate EEG features from observed EEG signals and the power spectrum computed by applying short-time Fourier transform (STFT) to each channel's EEG signal. The details are shown below.

First, segmentation of each channel's EEG signal is performed at fixed intervals as the preprocessing. Next, we compute Zero Crossing Rate from each EEG segment and calculate content percentages of θ wave (4-7Hz), slow-α wave (7-9Hz), mid-α wave (9-11Hz), fast-α wave (11-13Hz), and β wave (13Hz-) of the power spectrum in every channel. We also calculate the weighted moving average of each percentage. Furthermore, features proposed in [3], which focus on the power spectrum of the hemispheric asymmetry, are adopted. By calculating these values, an EEG feature vector is generated.

Audio Feature Extraction

First, audio signal segmentation is performed at fixed intervals as the preprocessing. Then we extract audio features used in [1, 7, 8] from each audio segment and obtain audio feature vectors. Table 1 shows audio features used in our method. Spectral Envelope is Linear Prediction Coefficients [9] calculated from the amplitude spectrum obtained by applying STFT to each audio segment. Generally, an EEG feature vector is derived from a short period EEG segment (T seconds, *e.g.*, one or two seconds) and an audio feature vector is extracted from an audio segment in milliseconds, which is much shorter than an EEG segment. Therefore, we compute the mean, variance and standard deviation from audio features included in each EEG segment of T seconds, and a newly-defined audio feature vector is obtained.

2.2. Vocal Segment Estimation from Each Feature (1st step)

In this subsection, we explain the method to estimate vocal segments in the 1st step. From the previous subsection, our method can obtain EEG feature vectors and audio feature vectors. Then we apply the feature selection method based on Max-Relevance and Min-Redundancy (mRMR) criteria proposed in [10] to EEG features in order to obtain efficient feature set for classification. It is efficient to apply the feature selection method to EEG features since relationships between stimuli to human beings from the outside and which parts of the human brain are affected by these stimuli is not well-known.

In the proposed method, we employ SVM [11] as the classifier to estimate vocal segments in music pieces. We assign positive labels to samples of vocal segments and negative labels to those of non-vocal segments. We train the classifier by separately using EEG feature vectors and audio feature vectors. This means the two classifiers are respectively obtained based on EEG and audio features. Therefore, we can estimate vocal segments in terms of EEG and audio features by inputting feature vectors extracted from test data into each trained classifier.

2.3. Integration of Classification Results (2nd step)

In this subsection, we explain the method to obtain the final classification results in the 2nd step. We integrate the classification results obtained in the 1st step using the method in [6]. This method trains the classifier using labels estimated by multiple annotators and the target data. The details of the 2nd step are shown as follows.

2.3.1. Preparation: Performance of each annotator and classification model

We explain the performance of each annotator and the classification model in our method. Let $y^j \in \{0, 1\}$ be the label assigned to the feature vector $\boldsymbol{x}$ by j^{th} annotator, where the two classifiers based on EEG and audio features correspond to the annotators. Note that 1 and 0 respectively represent vocal and non-vocal segments. When $y \in \{0, 1\}$ is the actual label for the feature vector, the performance of each annotator, P^j_{se} (sensitivity) and P^j_{sp} (specificity) are respectively defined as follows:

$$P^j_{se} := \Pr[y^j = 1 | y = 1], \tag{1}$$

$$P^j_{sp} := \Pr[y^j = 0 | y = 0], \tag{2}$$

where $j \in \{\mathrm{E}, \mathrm{A}\}$ in our method, and E and A respectively correspond to EEG and Audio.

In our method, a linear discriminating function is adopted based on the method in [6], and its classification model is specifically written as follows:

$$f_w(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}, \quad (\boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^d). \tag{3}$$

The final classification results $\hat{y}$ are obtained as follows:

$$\hat{y} = \begin{cases} 1 & \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} \geq Th \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\boldsymbol{w}$ is a weight and Th is a predetermined threshold.

2.3.2. Maximum likelihood estimator

Given the test data $\mathcal{D}$ consisting of N feature vectors with classification results by M ($=2$ in our method) annotators, $\mathcal{D} = \{\boldsymbol{x}_i, y^1_i, \cdots, y^{\mathrm{M}}_i\}^N_{i=1} = \{\boldsymbol{x}_i, y^{\mathrm{E}}_i, y^{\mathrm{A}}_i\}^N_{i=1}$, the estimation target is the weight $\boldsymbol{w}$. Given the test data $\mathcal{D}$, the likelihood of the weight $\boldsymbol{w}$ is defined as:

$$\Pr[\mathcal{D}|\boldsymbol{w}] \quad = \quad \prod_{i=1}^{N} \Pr[y^{\mathrm{E}}_i, y^{\mathrm{A}}_i | \boldsymbol{x}_i, \boldsymbol{w}]. \tag{5}$$

Using $\boldsymbol{P_{se}} = [P^{\mathrm{E}}_{se}, P^{\mathrm{A}}_{se}]$ and $\boldsymbol{P_{sp}} = [P^{\mathrm{E}}_{sp}, P^{\mathrm{A}}_{sp}]$, it is rewritten as

$$\Pr[\mathcal{D}|\boldsymbol{w}] = \prod_{i=1}^{N} \Big\{ \Pr[y^{\mathrm{E}}_i, y^{\mathrm{A}}_i | y_i = 1, \boldsymbol{P_{se}}] \cdot \Pr[y_i = 1 | \boldsymbol{x}_i, \boldsymbol{w}] + \Pr[y^{\mathrm{E}}_i, y^{\mathrm{A}}_i | y_i = 0, \boldsymbol{P_{sp}}] \cdot \Pr[y_i = 0 | \boldsymbol{x}_i, \boldsymbol{w}] \Big\}. \tag{6}$$

If we denote the actual label as y_i, and it is assumed that y_i^{E} and y_i^{A} are independent, $\Pr[y_i^{\mathrm{E}}, y_i^{\mathrm{A}}|y_i = 1, \boldsymbol{P_{se}}]$ can be written as

$$\begin{aligned}&\Pr[y_i^{\mathrm{E}}, y_i^{\mathrm{A}}|y_i = 1, \boldsymbol{P_{se}}]\\ &= \Pr[y_i^{\mathrm{E}}|y_i = 1, P_{se}^{\mathrm{E}}] \cdot \Pr[y_i^{\mathrm{A}}|y_i = 1, P_{se}^{\mathrm{A}}]\\ &= [P_{se}^{\mathrm{E}}]^{y_i^{\mathrm{E}}}[1 - P_{se}^{\mathrm{E}}]^{1-y_i^{\mathrm{E}}} \cdot [P_{se}^{\mathrm{A}}]^{y_i^{\mathrm{A}}}[1 - P_{se}^{\mathrm{A}}]^{1-y_i^{\mathrm{A}}}. \end{aligned} \quad (7)$$

Similarly,

$$\begin{aligned}&\Pr[y_i^{\mathrm{E}}, y_i^{\mathrm{A}}|y_i = 0, \boldsymbol{P_{sp}}]\\ &= \Pr[y_i^{\mathrm{E}}|y_i = 0, P_{sp}^{\mathrm{E}}] \cdot \Pr[y_i^{\mathrm{A}}|y_i = 0, P_{sp}^{\mathrm{A}}]\\ &= [P_{sp}^{\mathrm{E}}]^{1-y_i^{\mathrm{E}}}[1 - P_{sp}^{\mathrm{E}}]^{y_i^{\mathrm{E}}} \cdot [P_{sp}^{\mathrm{A}}]^{1-y_i^{\mathrm{A}}}[1 - P_{sp}^{\mathrm{A}}]^{y_i^{\mathrm{A}}}. \end{aligned} \quad (8)$$

Then the likelihood is rewritten as

$$\Pr[\mathcal{D}|\boldsymbol{w}] = \prod_{i=1}^{N}[\alpha_i p_i + \beta_i(1 - p_i)], \quad (9)$$

where

$$\begin{aligned} p_i = \Pr[y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}] &= \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)\\ &= \frac{1}{1 + \exp(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)}, \end{aligned} \quad (10)$$

$$\alpha_i = [P_{se}^{\mathrm{E}}]^{y_i^{\mathrm{E}}}[1 - P_{se}^{\mathrm{E}}]^{1-y_i^{\mathrm{E}}} \cdot [P_{se}^{\mathrm{A}}]^{y_i^{\mathrm{A}}}[1 - P_{se}^{\mathrm{A}}]^{1-y_i^{\mathrm{A}}}, \quad (11)$$

$$\beta_i = [P_{sp}^{\mathrm{E}}]^{1-y_i^{\mathrm{E}}}[1 - P_{sp}^{\mathrm{E}}]^{y_i^{\mathrm{E}}} \cdot [P_{sp}^{\mathrm{A}}]^{1-y_i^{\mathrm{A}}}[1 - P_{sp}^{\mathrm{A}}]^{y_i^{\mathrm{A}}}. \quad (12)$$

The maximum-likelihood estimator is found by maximizing the log-likelihood as follows:

$$\hat{\boldsymbol{w}}_{ML} = \arg\max_{\boldsymbol{w}}\{\ln\Pr[\mathcal{D}|\boldsymbol{w}]\}. \quad (13)$$

Let $\boldsymbol{y} = [y_1, \cdots, y_N]$ be the actual labels, and the complete data log-likelihood can be written as

$$\ln\Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{w}] = \sum_{i=1}^{N}\{y_i \ln p_i\alpha_i + (1 - y_i)\ln(1 - p_i)\beta_i\}. \quad (14)$$

In [6], the following Expectation-Maximization (EM) algorithm is adopted to maximize this likelihood.

(i)E-step

In the E-step, when the test data $\mathcal{D}$ and the current estimate of the weight $\boldsymbol{w}$ are given, the conditional expected value of log-likelihood is computed as follows:

$$\begin{aligned}&\mathrm{E}\{\ln\Pr[\mathcal{D}, \boldsymbol{y}|\boldsymbol{w}]\}\\ &= \sum_{i=1}^{N}\{\mu_i \ln p_i\alpha_i + (1 - \mu_i)\ln(1 - p_i)\beta_i\}, \end{aligned} \quad (15)$$

where μ_i is computed as follows:

$$\begin{aligned}\mu_i &\propto \Pr[y_i^{\mathrm{E}}, y_i^{\mathrm{A}}|y_i = 1, \boldsymbol{w}] \cdot \Pr[y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}]\\ &= \frac{\alpha_i p_i}{\alpha_i p_i + \beta_i(1 - p_i)}. \end{aligned} \quad (16)$$

(ii)M-step

In the M-step, based on the current estimate μ_i and the test data $\mathcal{D}$, the weight $\boldsymbol{w}$ is estimated by maximizing the conditional expected value in Eq.(15). Specifically, we obtain the following estimated weight $\boldsymbol{w}$ by equating the gradient of Eq.(15) to zero:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta\boldsymbol{H}^{-1}\boldsymbol{g}. \quad (17)$$

In Eq.(17), $\boldsymbol{g}$ is a gradient vector, $\boldsymbol{H}$ is a Hessian matrix and η is a step length. The gradient vector $\boldsymbol{g}$ and the Hessian matrix $\boldsymbol{H}$ are computed as follows:

$$\boldsymbol{g} = \sum_{i=1}^{N}[\mu_i - \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)]\boldsymbol{x}_i, \quad (18)$$

$$\boldsymbol{H} = -\sum_{i=1}^{N}[\sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)][1 - \sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)]\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}. \quad (19)$$

According to [6], the final classification results can be obtained by applying a threshold γ ($= 0.5$) to μ_i instead of directly using $\boldsymbol{w}$.

$$y_i = \begin{cases} 1 & \mu_i \geq \gamma \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

In our method, y_i^{E} and y_i^{A} are the classification results obtained by using EEG feature vectors and audio feature vectors in the 1st step, respectively. We select features based on mRMR [10] from all features in order to derive efficient features for classification and obtain $\boldsymbol{x}_i$. Then, from Eq.(20), we can obtain the final classification results.

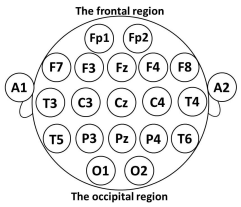

Fig. 1. Electrode layout of the international 10-20 system.

3. EXPERIMENTAL RESULTS

In this section, we show experimental results to verify the effectiveness of the proposed method. We explain EEG signal collection and the experimental procedures in **3.1**. Furthermore, the results of vocal segment estimation in music pieces are shown in **3.2**.

3.1. EEG Signal Collection and Experimental Procedures

In this subsection, we explain how to collect EEG signals in this experiment. EEG signals in this study are collected from six healthy subjects while they are listening to music pieces. The average age of subjects is about 23 years old. We record EEG signals from 12 channels (Fp1, Fp2, F7, F8, C3, C4, P3, P4, O1, O2, T3, T4) according to the international 10-20 system shown in Fig.1. Since EEG signals are weak, we amplify these signals by using an amplifier (MEG-6116M, NIHON KOHDEN). All leads are referenced to linked earlobes, and a ground electrode is located in the forehead. We also apply a band-pass filter to recorded EEG signals to avoid artifacts, and set the filter bandwidth to 0.04-30Hz.

Table 3. Results of our method and conventional methods. Note that results of only audio features are not depend on each subject. Thus, the result values become the same. F-measure is a harmonic mean between Recall and Precision.

	Only EEG features			Only Audio features			Proposed method		
	Avg.Recall	Avg.Precision	Avg.F-measure	Avg.Recall	Avg.Precision	Avg.F-measure	Avg.Recall	Avg.Precision	Avg.F-measure
Subject A	**0.9460**	0.6191	0.7436	0.8911	0.8838	0.8869	0.8878	**0.9116**	**0.8988**
Subject B	**0.9914**	0.5522	0.7081	0.8911	0.8838	0.8869	0.8955	**0.8856**	**0.8897**
Subject C	**0.9279**	0.6415	0.7556	0.8911	0.8838	0.8869	0.8921	**0.9141**	**0.9024**
Subject D	**0.9371**	0.6321	0.7502	0.8911	0.8838	0.8869	0.8911	**0.9131**	**0.9019**
Subject E	**0.9708**	0.5677	0.7124	0.8911	0.8838	0.8869	0.8933	**0.9007**	**0.8964**
Subject F	**0.9183**	0.6145	0.7374	0.8911	0.8838	0.8869	0.8888	**0.9077**	**0.8974**
Avg.	**0.9486**	0.6045	0.7346	0.8911	0.8838	0.8869	0.8914	**0.9055**	**0.8978**

Table 2. Experimental conditions.

	Audio signals	EEG signals
Sampling rate	44.1kHz	500Hz
Quantifying bit number	16bit	12bit
Time length of a segment	30ms	1s
Time length of an overlapping	20ms	0s

All music pieces which subjects are listening to are the two kinds of excerpts that include vocals or non-vocals, and the genre is Japanese POP music. One minute silence is inserted between every two music pieces. Subjects are instructed to keep their eyes closed, relax and remain seated when listening to music pieces. Furthermore, other experimental conditions are shown in Table 2.

3.2. Classification Results of Vocal Segments

In this subsection, we show experimental results to verify the effectiveness of our method. We use 60 pieces of music as a dataset including 30 positive pieces and 30 negative pieces, and the average of each time length is about 24 seconds. We also employ 5-fold cross-validation and each five subdataset contains the same number of music pieces labeled as positive and negative.

In this experiment, we perform the comparison with recent representative classifiers, *i.e.*, SVM, based only on EEG or audio features. It is reasonable to adopt these comparisons since our main contribution is the multimodal approach. Note that we employ the Gaussian kernel for SVM in all methods. To evaluate the accuracy of our vocal segment estimation, we use Recall, Precision and F-measure as defined below:

$$\text{Recall} = \frac{N_c}{N_m}, \tag{21}$$

$$\text{Precision} = \frac{N_c}{N_a}, \tag{22}$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}, \tag{23}$$

where N_c is the number of segments estimated correctly as vocal segments, N_m is the number of true segments estimated manually as vocal segments and N_a is the number of segments estimated by the method automatically as vocal segments.

The results are shown in Table 3. As shown in this table, the proposed method realize more accurate vocal segment estimation than the method based only on EEG or audio features. From these results, we can confirm that the multimodal approach is effective in the proposed method. We also note that our method can modify the classification results estimated incorrectly by utilizing only audio features based on the collaborative use of EEG and audio features. This effectiveness is also shown in Fig.2. In this figure, the whole part is non-vocal segment, but the classification results using only audio features suffer from some errors. On the other hand, our method can reduce such errors. Therefore, from the results in this subsection, the effectiveness of our main contribution, *i.e.*, the multimodal approach, is verified. This provides a solution to the performance limitation of the recent conventional methods.

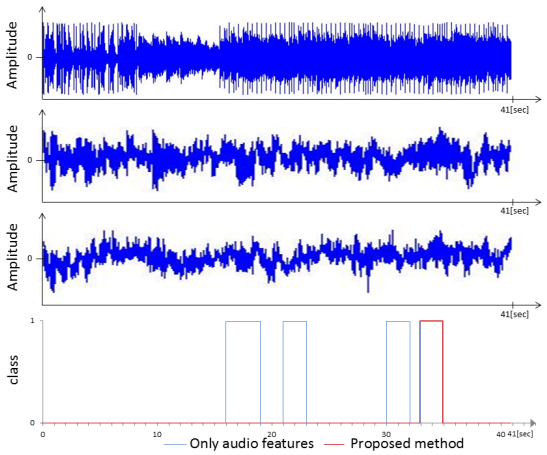

Fig. 2. A part of experimental results: the top is the waveform of audio signals, the two middle are waveforms of EEG signals (the upper is O1 and the lower is O2) and the bottom is the classification results. Class 1 and 0 correspond to vocal segment and non-vocal segment, respectively. Whole ground truth is 0, *i.e.*, there is no vocal segment in this example.

4. CONCLUSIONS

In this paper, we have proposed a novel vocal segment estimation method based on collaborative use of EEG and audio features. In the proposed method, we utilize EEG features extracted from EEG signals recorded while users are listening to music pieces and audio features extracted from audio signals of the music pieces. Therefore, our method implements multimodal vocal segment estimation. The experimental results show that our multimodal approach can realize more accurate vocal segment estimation than the unimodal approach.

5. ACKNOWLEDGEMENT

This work was partly supported by Grant-in-Aid for Scientific Research on Innovative Areas 24120002 from the MEXT.

6. REFERENCES

[1] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1885 –1888, 2008.

[2] T. L. Nwe, "Automatic detection of vocal segments in popular songs," in *Proceedings of the International Conference on Music Information Retrieval*, pp. 138 –145, 2004.

[3] Y. P. Lin, C. H. Wang, T. P. Jung, Wu T. L., S. K. Jeng, J. R. Duann, and J. H. Chen, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798 –1806, 2010.

[4] H. Lu, M. Wang, and H. Yu, "EEG model and location in brain when enjoying music," in *Proceedings of IEEE-EMBS 2005, 27th Annual International Conference of the Engineering in Medicine and Biology Society*, pp. 2695 –2698, 2005.

[5] S.-I. Ito, Y. Mitsukura, M. Fukumi, and N. Akamatsu, "A feature extraction of the EEG during listening to the music using the factor analysis and neural networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 3, pp. 2263 –2267, 2003.

[6] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 889 –896, 2009.

[7] N. Nitanda and M. Haseyama, "Audio-based shot classification for audiovisual indexing using pca, mgd and fuzzy algorithm," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 90, no. 8, pp. 1542 –1548, 2007.

[8] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293 –302, 2002.

[9] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 3, pp. 247 –254, 1979.

[10] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226 –1238, 2005.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273 –297, 1995.