Music Genre Classification Using Timbral Feature Fusion on i-vector Framework

RAJEEV RAJAN [1]
HARISHANKER G. [2]
ATHIRASREE C.A [3]
HARITHA S.M. [4]

Department of Electronics and Communication Engineering
College of Engineering,Trivandrum
APJ Abdul Kalam Technological Univesrity, Thiruvananthapuram
Kerala,India
[1]rajeev@cet.ac.in
[2]harishankergopakumar@gmail.com
[3]athirasree96@gmail.com
[4]harithegreen123@gmail.com

Abstract. method for automatic music genre classification based on the fusion of high-level and low-level timbral descriptors is proposed. High-level features namely, i-vectors are computed from melfrequency cepstral coefficient (MFCC)-GMM framework. Low-level timbral descriptors namely MFCC, modified group delay features (MODGDF) and timbral feature set are also computed from the audio files. Initially, the experiment is performed using i-vectors alone. Later, low-level timbral features are appended with high-level i-vector features to form a high dimensional feature vector (55 dim). Support vector machine (SVM) and deep neural network (DNN) based classifiers are employed for the experiment. The performance is evaluated using GTZAN dataset on 5 genres. With high-level i-vector features, the baseline-SVM and DNN-based classifiers report average classification accuracies (in %) of 79.30 and 80.67, respectively. A further improvement (9 %) in performance was observed when low-level timbral descriptors are fused with the i-vectors in both SVM and DNN frameworks. The results demonstrate the potential of the timbral feature fusion in music genre classification task.

Keywords: genre, classification, i-vector, fusion

(Received July 19th, 2011 / Accepted September 1st, 2011)

1 Introduction

Music information retrieval (MIR) mainly focuses on the understanding and usefulness of music data [31], through research, development and application of computational approaches and tools [1, 14, 15, 30]. Recent worldwide popularization of online music distribution services and portable digital music players make the more important task. Automatic music genre classification, a task of fundamental importance in the context of music information retrieval is addressed using i-vector approach in this paper.

Automatic music genre classification extracts distinctive features from audio excerpts and then automatically categorizes them into respective musical genres. A genre or sub-genre can be distinguished from other by musical techniques, the style, the cultural context or the content and spirit of the themes. Although the division of music into genres is partially subjective and arbitrary, there exist perceptual criteria based on the texture, instrumentation and rhythmic structure of mu-

sic that can be used to characterize a particular genre. The music content distribution vendors can use an effective automatic music analysis system to attract customers [4, 37] in many day-to-day applications.

A comprehensive survey of both features and classification techniques used in genre classification can be found in [24]. In the front end, melodic features [32], local features [40] psychoacoustic features [33] and musical texture features [36] are extracted from the audio files. In the classification phase, researchers are employed both generative and discriminative models [19, 34]. An unsupervised approach using hidden Markov models (HMMs) was proposed in the work of Shao et. al [34]. In [2], a method is introduced with l-SVM classifier which combines the ideas of the classical SVM with the sparse approximation techniques in genre classification. A similar approach is also proposed in [41]. Two novel classifiers using inter-genre similarity (IGS) modeling are proposed in [38]. The SVM [6] based classifiers are widely used for multiclass pattern classification [8]. In [19], the best results were achieved using the SVM classifier. The performance of the SVM-based classifier can be further improved by fine-tuning the parameters of the selected kernel. In [42], the support vector metric learning (SVML) method combines the learning of Mahalanobis distance metric with the learning of the Gaussian kernel parameters.

A combination of max-average pooling and residual learning-based method, are introduced to improve the efficiency in [44]. Zhang et al. employed convolutional neural network (CNN) with k max-pooling layers for semantic modelling of music [43]. Visual representations of audio are used as input CNNs in [7, 27]. Results of multimodal experiments show how the aggregation of learnt representations from different modalities improves the accuracy, suggesting that different modalities embed complementary information [26]. Intermediate representations of deep neural networks (DNN) are learnt from audio tracks, text reviews, and cover art images, and further combinations. A masked conditional neural network (MCLNN) is designed to exploit the properties of multi-dimensional temporal signals by considering the sequential relationship across temporal frames for genre classification in [21]. CNN, which takes full advantage of low-level information of mel-spectrogram for genre classification can be seen in [20]. Deep learning architecture shows inferior performance with small datasets and often needs a large training set for improved performance. It is already established that DML techniques are effective for small datasets [11, 16].

In our approach, we apply i-vector framework for genre classification task.

The outline of the rest of the paper is as follows: Proposed system in Section 2 gives the description about the front-end and the classification scheme. The evaluation of the proposed system on GTZAN dataset and its analysis are given in the Section 3. Subsequently, conclusion is drawn in the final section.

2 Proposed System

In the front-end, i-vectors and low-level timbral features namely, MFCC, MODGDF, and timbral features are computed. Timbral descriptors are considered because of their ability to extract the specific and distinguishing traits of a variety of music styles. Initially, track-level i-vectors are computed from MFCC (including delta, delta-delta) and experimented with SVM and DNN. Later i-vectors are fused with low-level timbral feature (LTF) and the experiment is repeated. A detailed description is given in the following sections.

2.1 Feature Extraction

Two features namely, high-level features,low-level feature sets are employed for the proposed task.

2.1.1 High level features:

I-vector subspace modeling is one of the prominent methods that has become the state of the art technique in speech and music processing applications [12, 45]. It models both the intra-domain and inter-domain variabilities into the same low-dimensional space [39]. Factor analysis (FA) on the frame-level features results in high-level descriptors called 'identity vectors' or 'i-vectors'. The i-vector method applies FA to extract low-dimensional features from Gaussian mixture model(GMM)-supervectors. This approach estimates hidden variables in GMM supervector space, which provides better discrimination ability than GMM supervectors [12]. The method of modelling GMM-super vectors has achieved superior recognition performance in recent works [12, 45] and it motivated us to use the scheme as a baseline in the proposed task. The i-vector-based statistical feature has been shown promise in music genre classification [9]. We computed i-vectors from frame-level extracted MFCC features using Alize tool kit [5].

The method of modeling Gaussian mixture model (GMM) super vectors has provided for superior speaker recognition performance in recent works. I-vector system [10] is a technique to map the high dimensional

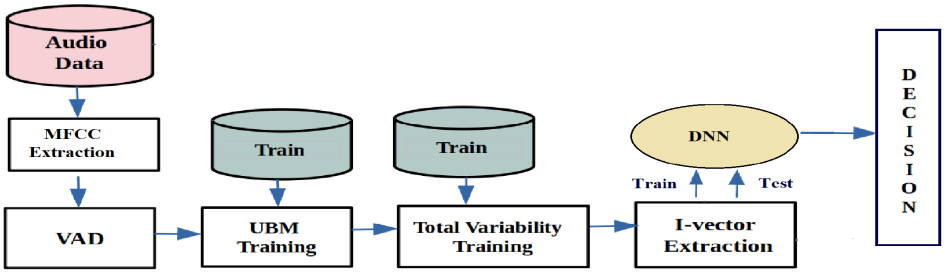

Figure 1: Basic steps in i-vector extraction and classification

GMM super vector space (generated from concatenating all the mean values of GMM) to low dimensional space called total variability space. The main idea is to adapt the target utterance GMM from a universal background model(UBM) using the eigenvoice adaption method introduced in [17, 28]. The target GMM super vector can be viewed as shifted from the UBM. Formally, a target GMM super vector M can be written as:

$$M = m + Tw \tag{1}$$

where m represents the UBM super vector, T is a low dimensional rectangular total variability matrix, and w is termed as i-vector. Using training data, the UBM and TV matrix will be modeled by expectation maximization (EM) method. In the E-step, w is considered as a latent variable with normal prior distribution $N(0, I)$. Eventually, the i-vectors will be estimated as the mean of posterior distribution of w, that is [10],

$$w(u) = (I + T^T\Sigma^1.N(u).T)^1T^T\Sigma^1S(u) \tag{2}$$

where for utterance u, the terms $N(u)$ and $S(u)$ represent zeroth and centralized first order Baum-Welch statistics respectively, and Σ is the covariance matrix of UBM. The basic steps in i-vector based classification is shown in Fig 1.

2.1.2 Law-level timbral descriptors

Timbral descriptors which reflect mainly the instruments and arrangement of the music differentiate mixture of sounds that are possibly with the same or similar rhythmic and pitch contents [3, 19]. Timbral features which serve as a physical correlate to perceptual attributes differentiate mixture of sounds that are with the same or similar rhythmic and pitch contents [13]. In timbre space, the perceived (dis)similarity between the sounds is projected to a low-dimensional space where dimensions are assigned a semantic interpretation such as brightness and temporal variation. Three features namely MFCC, MODGDF, and timbral feature set are considered in the proposed experiment.

MFCC:

MFCCs are widely employed in numerous perceptually motivated audio classification tasks, despite their widespread use as predictors of perceived similarity of timbre [29]. 20 dim MFCCs are computed using frame-size of 40ms and frame-shift of 10ms. Perceptual filter banks-based cepstral features are based on the computation of cochleagram, which in some sense try to model the frequency selectivity of the cochlea.

Modified Group Delay Feature (MODGDF)

Earlier efforts have established the significance of modified group delay functions (MODGD) in various applications for speech and music. In this paper, we extend the use of modified group delay features for genre classification. Group delay feature has already been used in pitch estimation [22], formant extraction, speech recognition and speech synthesis [23]. The group delay function $\tau(e^{j\omega})$, of a discrete time signal $x[n]$, is defined by,

$$\tau(e^{j\omega}) = -\frac{d\{\arg(X(e^{j\omega}))\}}{d\omega}, \tag{3}$$

where ω, $X(e^{j\omega})$, $\arg X(e^{j\omega})$ represent angular frequency, Fourier Transform (FT) and phase function, respectively.

Consider a discrete time signal $x[n]$. Then

$$X(e^{j\omega}) = |X(e^{j\omega})|e^{j\arg(X(e^{j\omega}))} \tag{4}$$

From Equation 4

$$\log\ X(e^{j\omega}) = \log\left(|X(e^{j\omega})|\right) + j\{\arg(X(e^{j\omega}))\} \tag{5}$$

$$\arg(X(e^{j\omega})) = Im[\log\ X(e^{j\omega})]. \tag{6}$$

Equations 3 and 6, enable the computation of group delay function directly for minimum phase signals [25]:

$$\tau(e^{j\omega}) = -Im\frac{d(\log(X(e^{j\omega})))}{d\omega} \tag{7}$$

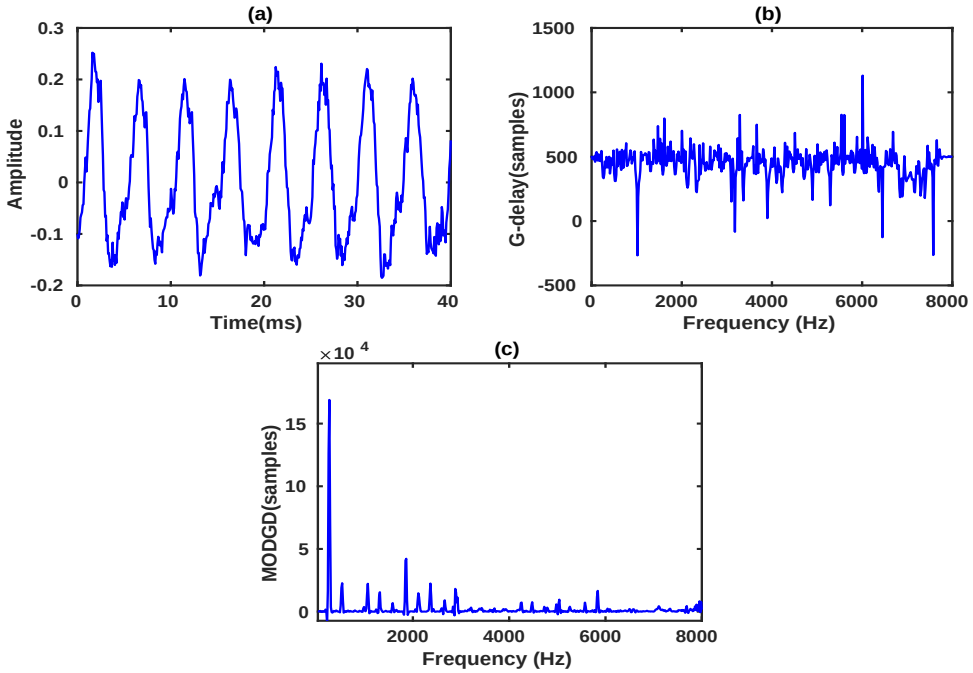

Figure 2: (a) Frame of music (b) Group delay function computed for the frame in (a), (c) Modified group delay function computed the frame in (a).

Group delay can be obtained by by [25],

$$\tau(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|X(e^{j\omega})|^2} \quad (8)$$

where the subscripts R and I denote the real and imaginary parts, respectively. $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the Fourier transforms of $x[n]$ and $n.x[n]$, respectively. The zeros close to the unit circle in the z-domain cause the group delay function to be ill behaved. The zeros of the transfer function can be pushed inside the unit circle to restore the spectrum by replacing denominator in Equation 8 by its spectral envelope, $S(e^{j\omega})$. The modified group delay function (MODGD) $\tau_m(e^{j\omega})$ for a signal x[n] is obtained by [23],

$$\tau_m(e^{j\omega}) = (\frac{\tau_c(e^{j\omega})}{|\tau_c(e^{j\omega})|})(|\tau_c(e^{j\omega})|)^{\alpha}, \quad (9)$$

where,

$$\tau_c(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2\gamma}}. \quad (10)$$

Two new parameters, α and γ ($0 < \alpha \leq 1$ and $0 < \gamma \leq 1$) are introduced to control the dynamic range of MODGDF. The group delay function and modified group delay function for a frame of music is shown in Figure 2.

The MODGDGRAM [1] of a pop song is plotted in Figure 3. MODGDGRAM emphasizes system-specific information as compared to spectrogram. Modified group delay features (MODGDF) are given by,

$$c(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos(\frac{n(2k+1)\pi}{N_f}) \quad (11)$$

where N_f is the discrete Fourier transform order and $\tau_m(k)$ is the modified group delay spectrum. 20 dimensional MODGDF is computed using frame-length of 40ms and frame-shift of 10ms.

Timbral feature set:

In our experiment, five features, namely, spectral centroid, spectral roll-off, spectral flux, zero crossings, and low energy are computed in track-level. A low-level spectral feature set is also computed which serve as a physical correlate to perceptual attributes, such as timbre and coloration [13]. Low-level timbral features are defined below;

1. Spectral centroid: Spectral centroid is defined as

[1] Visual representation of MODGD with time and frequency in horizontal and vertical axis, respectively. A third dimension, indicating the amplitude of group delay function at a particular time is represented by the intensity or color of each point in the image

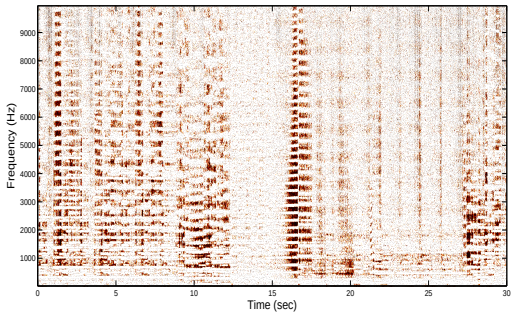

Figure 3: MODGDGRAM of a song from pop genre

the center of gravity of the magnitude spectrum of the STFT.

2. Spectral roll-off: Spectral roll-off is defined as the frequency below which 85% of the magnitude distribution is concentrated.

3. Spectral flux: Spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions.

4. zero-crossing: It is a weighted measure of the number of times the signal changes sign in a frame:

The significance of low-level timbral features can be well understood from the 2-dimensional mapping of five genres in Figure 4.

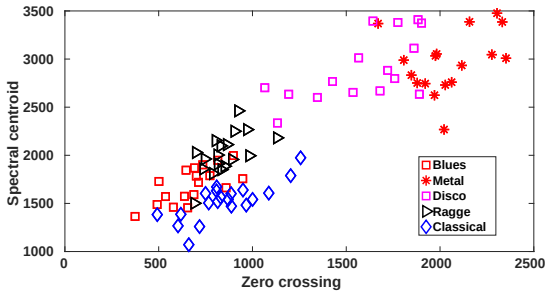

Figure 4: Significance of timbral features in genre classification

2.2 Classification Scheme:

In the classification phase, SVM and DNN are employed. Deep neural networks are a powerful category of machine learning algorithms implemented by stacking layers of neural networks along the depth and width of smaller architectures. Deep networks have recently demonstrated discriminative and representation learning capabilities over a wide range of applications. In the first phase, the baseline-SVM classifier with a linear kernel is used. Later, the experiment is carried out using a DNN framework on the same feature set. Our proposed DNN architecture uses three hidden layers (100 nodes per layer) with Adam optimization algorithm. Rectified linear units (ReLUs) have been chosen as the activation function for hidden layers and softmax function for the output layer. In the final phase, the experiment is extended with DNN using the fusion of timbral features with i-vectors computed using MFCCs.

3 PERFORMANCE EVALUATION

The performance evaluation is done using GTZAN dataset. The description of the dataset, experimental framework and results and analysis are given below.

3.1 Dataset :

The GTZAN dataset is the most-used public dataset for evaluation in machine listening research for music genre recognition (MGR). The files were collected from a variety of sources including personal CDs, radio, microphone recordings, in order to represent a variety of recording conditions. The performance of the proposed system is evaluated using GTZAN dataset [35]. GTZAN dataset is created by Tzanetakis and Cook and it includes 1000 music excerpts of 30 seconds of duration. The system is evaluated using subset which includes 5 genres namely, classical, country, hip-hop, metal, and pop, are considered for the experiment. The classical dataset has the following classes: choir, orchestra, piano, string quartet. The tracks are all 22050 Hz Mono 16-bit audio files in .wav format.

3.2 Experimental Framework :

Initially MFCCs, MODGDF features are frame-wise computed. I-vectors are also extracted using ALIZE toolkit [5]. MFCC and MODGDF feature set are averaged across dimensions to form a representative feature vector. In the first phase, experiments are carried out with i-vectors computed from the MFCCs using DNN and SVM.

In the final stage, the experiment is extended with the early fusion of i-vectors (10 dim) and low-level feature set (45 dim). The i-vectors are computed using 128 mixture GMM built using 60-dimensional MFCC (including delta and delta-delta features). I-vectors are

Table 1: Confusion matrix of i-vectors-SVM baseline system. Entries are in number of files

Class	Classical	Country	Hip-hop	Metal	Pop
Classical	**20**	8	1	0	1
Country	5	**20**	0	2	3
Hip-hop	2	1	**23**	2	2
Metal	0	0	0	**29**	1
Pop	0	1	0	2	**27**

Table 2: Confusion matrix of i-vectors + LTF fusion using DNN framework. Entries are in number of files

Class	Classical	Country	Hip-hop	Metal	Pop
Classical	**29**	1	0	0	0
Country	1	**27**	0	1	1
Hip-hop	0	0	**24**	1	5
Metal	1	1	2	**25**	1
Pop	0	2	1	0	**27**

computed using ALIZE open source speaker recognition tool kit [5]. The track level timbral feature set is computed using MIRToolbox [18]. In the i-vector framework, first, a UBM-GMM model is built from MFCCs computed from an auxiliary database and audio samples from the corpus under study. Total variability matrix, T is also trained using the audio files from the corpus, covering all genres. 150 files with 30 files/ genre is used for testing phase. Baseline-SVM and DNN classifiers are implemented using LibSVM and Keras-tensorflow respectively. In DNN experiment, the system was trained for 1000 epochs with learning rate of 0.002.

3.3 Results and Analysis

The results of the experiments are tabulated in Table 3. The baseline-SVM system reports an overall accuracy of 79.30% on i-vectors. The overall accuracy of 80.67% is reported with DNN framework. When we fuse i-vectors with the low-level timbral feature set, a significant improvement is observed with an overall accuracy of 88.40% and 88.00% for SVM and DNN frameworks respectively. It demonstrates the potential of fusion system in the automatic genre classification task.

The confusion matrices of the baseline system and the best performing systems are given in Table 1 and Table 2, respectively. From Table 2, it is observed that genres, classical and country show the least accuracy with 66.66%. By incorporating the fusion strategy, we could improve the accuracy of the country to 96.66% and classical to 90.00%. The classification accuracy of all other genres except for metal is improved. It is worth noting that the i-vectors alone gives an accuracy of 96.66%. The genre-wise accuracy is shown in Figure 5. It is observed that class wise accuracy of more than 80% is reported for all genres in feature-fusion paradigm for both SVM and DNN.

To end the discussion, the experiments show the promise of fusion of timbral descriptors in genre classification task.

4 CONCLUSION

We presented a genre classification method based DNN on the fusion of timbral descriptors. High-level timbral feature namely i-vectors, low-level timbral features namely MFCC, MODGDF, and timbral feature set is computed. The experiment is performed using a baseline-SVM and DNN classifier. The performance is evaluated using GTZAN dataset. An improvement of 9% is observed with timbral feature fusion as compared to baseline-i-vector/SVM system. The results demonstrate the potential of timbral feature fusion in automatic music genre classification task. While the baseline system reports an overall accuracy of 79.30%, the fusion system reports an overall accuracy of 88.04%. The results demonstrate the potential of timbral feature fusion in automatic music genre classification task.

References

[1] Affonso, E. T., Nunes, R. D., Rosa, R. L., Pivaro, G. F., and Rodriguez, D. Z. Speech quality assessment in wireless voip communication using deep belief network. *IEEE Access*, 6:77022–77032, 2018.

Table 3: Overall accuracy of four phases of experiments.

No	Method	Accr.(%)
1	i-vector + SVM	79.30
2	i-vector + DNN	80.67
3	i-Vector + LTF (Fusion) + SVM	**88.40**
4	i-Vector + LTF (Fusion) + DNN	**88.00**

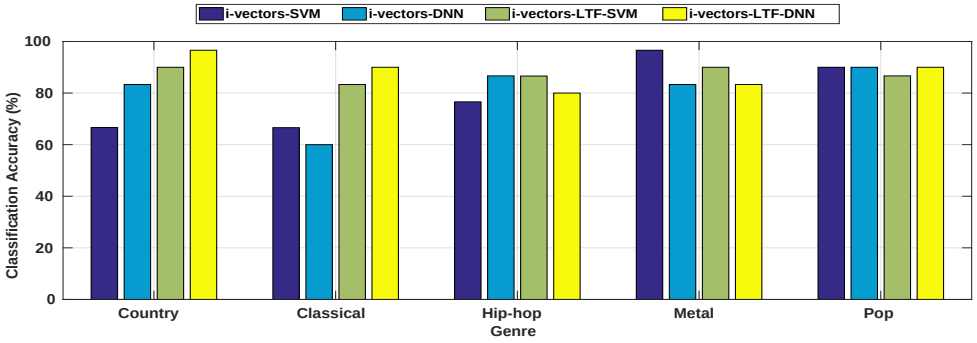

Figure 5: Genre-wise classification accuracy for all the experiments.

[2] Aryafar, K., Jafarpour, S., and Shokoufandeh, A. Music genre classification using sparsity-eager support vector machines. *Pattern Recognition (ICPR), 21st International Conference on*, pages 1526–1529, 2012.

[3] B. Jean-Marc and R. Christophe. Developement of a speech recognizer using a hybrid HMM/MLP system. *in Proc. of European Symp. on Arti. Neu. Netw., Belgium*, 4:441–446, 1999.

[4] Barbedo, J. G. and Lopes, A. Automatic genre classification of musical signals. *EURASIP J. Adv. in Sig. Proces., Article ID 64960*, 2007.

[5] Bonastre, J.-F., Wils, F., and Meignier, S. Alize, a free toolkit for speaker recognition. *in Proc. of the Annual Conference of the International Speech Communication Association,Interspeech*, 1, 01 2005.

[6] Boser, I. and Vapnik, V. A training algorithm for optimal margin classifiers. *in Proc. of the Fifth Ann. Work. on Comp. Learning Theory*, pages 144–152, 1992.

[7] Choi, K., Fazekas, G., and Sandler, M. Automatic tagging using deep convolutional neural networks. *in Proceedings of International Society. for Music Information Retrieval Conference.*, pages 805–811, 2016.

[8] Cristianini, N. and Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. *Cambridge University Press, Cambridge*, 2000.

[9] Dai, J., Xue, W., and Liu, W. Multilingual i-vector based statistical modeling for music genre classification. *Proc. of Interspeech*, pages 459–463, 2017.

[10] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. on Audio, Speech, and Language Processing,*, 19:788–798, 2011.

[11] Du, Y., Liu, C., and Zhang, B. Detection of GH pituitary tumors based on MNF. *in Proceedings of Chinese Control And Decision Conference*, pages 635–639, 2019.

[12] Eghbal-zadeh, H., Lehner, B., Schedl, M., and Widmer, G. I-vectors for timbre-based music similarity and music artist classification. *in Proc. of 16th Int. Society for Music Information Retrieval Conference*, pages 554–560, 2015.

[13] G.Peeters, Giordano, B. L., Susini, P., N.Misdariis, and McAdams., S. The timbre toolbox: Extracting audio descriptors from musical signals. *J. the Acou. Soc. of Amer.*, 130(5):2902–2916, 2011.

[14] Guimarães, R., Rodríguez, D. Z., Rosa, R. L., and Bressan, G. Recommendation system using sentiment analysis considering the polarity of the adverb. In *2016 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 71–72. IEEE, 2016.

[15] Guimaraes, R. G., Rosa, R. L., De Gaetano, D., Rodriguez, D. Z., and Bressan, G. Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816, 2017.

[16] Kaya, M. and Bilge, S. H. Deep metric learning: A survey. *Symmetry*, 11(9):1–26, 2019.

[17] Kenny, P., Boulianne, G., , and Dumouchel, P. Eigenvoice modeling with sparse training data. *IEEE Trans. on Speech and Audio Processing*, 13:345–354.

[18] Lartillot, O., Toiviainen, P., and Eerola, T. A matlab toolbox for music information retrieval. *in Preisach C., Burkhardt H., Schmidt-Thieme L., Decker R. (eds) Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg*, pages 261–268, 2008.

[19] Li, T., Ogihara, M., and Li, Q. A comparative study on content-based music genre classification. *in Proc. 26th Int. ACM Conf. Research and Development in Infor. Retri.*, pages 282–289, 2003.

[20] Liu, C., Feng, L., Liu, G., Wang, H., and Liu, S. Bottom-up broadcast neural network for music genre classification. *Pattern Recognition Letters*, pages 1–7, 2018.

[21] Medhat, F., Chesmore, D., and Robinson, J. Masked conditional neural networks for audio classification. *In: Lintas, A., Rovetta, S., Verschure, P.F.M.J., Villa, A.E.P. (eds.) ICANN. LNCS*, pages 349–358, 2017.

[22] Murthy, H. A. *Algorithms for Processing Fourier Transform Phase of Signals*. PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, December 1991.

[23] Murthy, H. A. and Yegnanarayana, B. Group delay functions and its application to speech processing. *Sadhana*, 36(5):745–782, 2011.

[24] N.Scaringella, G.Zoia, and Miyen.K. Automatic genre classification of music content. *IEEE Sig. Proces. Magazine*, 23(2):133–141, 2006.

[25] Oppenheim, A. V. and Schafer, R. W. *Discrete Time Signal Processing*. Prentice Hall, Inc, New Jersey, 1990.

[26] Oramas, S., Barbieri, F., Nieto, O., and Serra, X. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval*, 1:4–21, 2018.

[27] Pons, J., Lidy, T., and Serra, X. Experimenting with musically motivated convolutional neural networks. *in Proceedings of 14th International Workshop on Content-Based Multimedia Indexing*, pages 1–6, 2016.

[28] Reynolds, D. and Rose, R. Robust text-independent speaker identification using gaussian mixture speaker models,. *IEEE Trans. on Speech and Audio Processing*, 3:72–83.

[29] Richard, G., Sundaram, S., and Narayanan, S. An overview on perceptually motivated audio indexing and classification. *in Proc. of the IEEE*, 101:1939–1954, 2013.

[30] Rosa, R. L., Rodriguez, D. Z., and Bressan, G. Sentimeter-br: A social web analysis tool to discover consumers' sentiment. In *2013 IEEE 14th international conference on mobile data management*, volume 2, pages 122–124. IEEE, 2013.

[31] Rosa, R. L., Rodriguez, D. Z., and Bressan, G. Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics*, 61(3):359–367, 2015.

[32] Salamon, J., Rocha, B., and Gomez, E. Musical genre classification using melody features extracted from polyphonic music signals. *in Proc. of IEEE Int. Conf. Aco., Speech, and Sig. Proces.*, pages 81–85, 2012.

[33] Scheirer, E. Music listening systems. *Ph.D Thesis., School of Architecture and Planning, Massachusetts Institute of Technology*, 2000.

[34] Shao, X., Xu, C., and Kankanhalli., M. S. Unsupervised classification of music genre using hidden markov model. *in Proc. of IEEE Int. Conf. on Multi. and Expo,*, 3:2023–2026, 2004.

[35] Tzanetakis, G. and Cook, P. Musical genre classifcation of audio signals. *IEEE Trans. Speech Audio Proces.*, 10:293–302, July 2002.

[36] Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Process.*, 10(5):293–302, 2002.

[37] Tzanetakis, G., Essl, G., and Cook, P. Automatic musical genre classification of audio signals. *in Proc. of Int. Soc. Music Infor. Retri. Conf.*, 2001.

[38] UlaBagci and Erzin, E. Musical genres using inter-genre similarity. *IEEE Sig. Proces. Letters*, 14(8):521–524, 2007.

[39] Verma, P. and Das, P. i-vectors in speech processing applications: A survey. *International Journal of Speech Technology*, 18:529–546, 12 2015.

[40] Wulfing, J. and Riedmille, M. Unsupervised learning of local features for music classifications. *in Proc. of Int. Soc. for Music Infor. Retri. Conf.*, pages 139–144, 2012.

[41] Xu, C., Maddage, N. C., Shao, X., Cao, F., and Tian, Q. Musical genre classification using support vector machines. *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 5:429–432, 2003.

[42] Xu, Z., K. Weinbergerger, Q., and Oliver, C. Distance metric learning for kernel machines. *arXiv:1208.3422v2 [stat.ML]*, 2013.

[43] Zhang, P., Zheng, X., Zhang, W., Li, S., Qian, S., He, W., and S. Zhang, a. Z. W. A deep neural network for modeling music. *in Proceedings of the ACM International Conference on Multimedia Retrieval.*, pages 379–386, 2015.

[44] Zhang, W., Lei, W., Xu, X., and Xing, X. Improved music genre classification with convolutional neural networks. *in Proceedings of International Society for Music Information Retrieval Conference*, 19:3304–3308, 2016.

[45] Zhong, J., Hu, W., Soong, F., and Meng, H. DNN i-vector speaker verification with short, text-constrained test utterances. *in Proc. of the Annual Conference of the International Speech Communication Association, Interspeech*, pages 1507–1511, 08 2017.