# Predominant Instrument Recognition in Polyphonic Music Using GMM-DNN Framework

Roshni Ajayakumar
Rajiv Gandhi Institute of Technology,
Kottayam, India
Email: roshni.rose.1993@gmail.com

Rajeev Rajan
College of Engineering, Trivandrum
Thiruvananthapuram, India
rajeev@cet.ac.in

***Abstract*—In this paper, the predominant instrument recognition in polyphonic music is addressed using timbral descriptors in three frameworks- Gaussian mixture model (GMM), deep neural network (DNN), and hybrid GMM-DNN. Three sets of features, namely, mel-frequency cepstral coefficient (MFCC) features, modified group delay features (MODGDF), and low-level timbral features are computed, and the experiments are conducted with individual set and its early integration. Performance is systematically evaluated using IRMAS dataset. The results obtained for GMM, DNN, and GMM-DNN are 65.60%, 85.60%, and 93.20%, respectively on timbral feature fusion. Architectural choice of DNN using GMM derived features on the feature fusion paradigm showed improvement in the system performance. Thus, the proposed experiments demonstrate the potential of timbral descriptors and DNN based systems in recognizing predominant instrument in polyphonic music.**

## I. Introduction

Predominant instrument recognition refers to the problem where the prominent instrument is identified from a polyphonic music audio file. In music with orchestration, there are several musical instruments playing together yielding strong harmonic interfering partials, which makes the identification of prominent instrument harder. Automatic instrument recognition in polyphonic music has a wide range of applications such as tailored instrument-specific audio equalization, melody extraction and music recommendation service in this era of enormous online music repertoire [1]. The efficiency of the source separation task can be improved significantly by knowing the number and the types of the instruments [1].

### A. Related Work

An extensive review of approaches for the isolated musical instrument classification can be found in [2]. A novel feature representation called sparse cepstral codes for instrument identification is proposed in [3]. It is shown that the use of sparse coding and power normalization can be used to derive a better representation of the spectrum. Features derived from the root-mean-square (RMS) energy envelope have also been exploited well for the task [4]. In the polyphonic environment, instrument recognition using non-negative matrix factorization (NMF)-based source/filter model with MFCC is attempted in [5]. A hierarchical classification scheme by exploiting the realistic genre-wise musical hypotheses is successfully utilized to recognize instruments in polyphonic music [6]. In a model proposed in [7], feature extraction and learning algorithms are trained together in an end-to-end fashion to achieve better results than traditional methods using hand-crafted features. Fuhrmann [8] proposed a scheme using a support vector machine (SVM) classifiers trained with features extracted from real musical audio signals. The authors investigated the importance and modeling accuracy of temporal characteristics in combination with statistical models. Convolutional neural networks (CNN) have been utilized to learn the spectral characteristics of the music recordings in [1].

### B. Motivation

Conventionally, the spectrum-related features used in instrument recognition take into account merely the magnitude information. However, there is often additional information concealed in the phase, which could be beneficial for recognition [9]. While the commonly applied MFCCs are capable of modelling the resonances introduced by the filter of the instrument body, it neglects the spectral characteristics of the vibrating source, which also, play their role in human perception of musical sounds [10]. It has already been established in the literature that the modified group delay function emphasizes peaks in spectra well [11]. Also, we want to experiment with the hybrid GMM-HMM scheme for the instrument classification task in the polyphonic environment.

Section 2 describes feature extraction and classification phase. The performance evaluation is described in section 3 followed by analysis of results in section 4. Finally, the paper is concluded in section 5.

## II. Proposed System

MFCC, MODGDF, and low-level timbral feature sets are extracted from the audio file. These features are considered because of their ability to extract the specific and distinguishing traits of variety of music styles. In the classification phase, three frameworks namely GMM, DNN, and GMM-DNN are employed. A detailed description is given in the following sections.

978-1-7281-8895-9/20/$31.00 © 

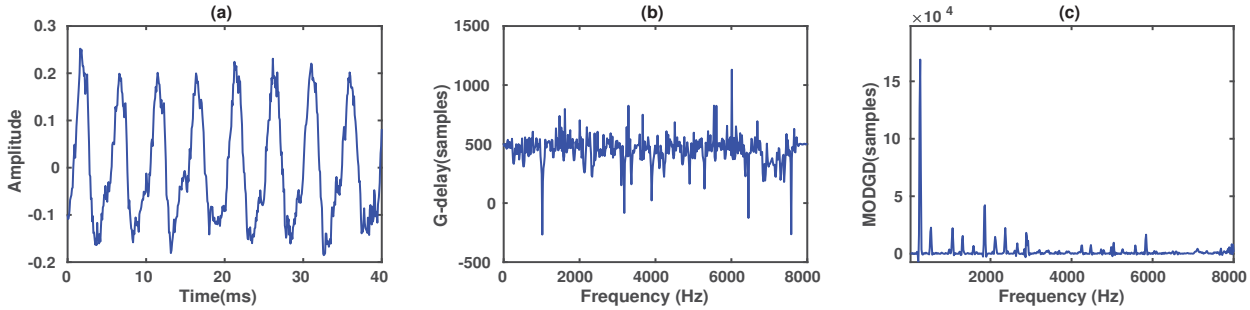

Fig. 1: (a) Frame of music, (b) Group delay function for the frame in (a), (c) Modified group delay function for the frame in (a).

### A. Feature Extraction

*1) MFCC:* Due to the capacity to capture "global" spectral envelope properties, MFCCs are employed in numerous perceptually motivated audio classification tasks, despite their widespread use as predictors of perceived timbre similarity [12].

*2) Modified Group Delay Feature (MODGD):* Group delay features have already been employed in numerous speech and music processing applications[11, 13, 14, 15, 16]. The group delay function, $\tau(e^{j\omega})$ is defined as the negative derivative of unwrapped Fourier transform phase. The group delay function of minimum phase signals can be computed directly from the signal by [17].

$$\tau(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|X(e^{j\omega})|^2} \quad (1)$$

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 denominator is replaced by its spectral envelope, $S(e^{j\omega})$ to mask the spiky nature. The modified group delay function (MODGD) $\tau_m(e^{j\omega})$ is obtained as

$$\tau_m(e^{j\omega}) = (\frac{\tau_c(e^{j\omega})}{|\tau_c(e^{j\omega})|})(|\tau_c(e^{j\omega})|)^{\alpha}, \quad (2)$$

where,

$$\tau_c(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2\gamma}}. \quad (3)$$

Two new parameters, $\alpha$ and $\gamma$ ($0 < \alpha \leq 1$ and $0 < \gamma \leq 1$) are introduced to control the dynamic range of MODGD [11]. Modified group delay functions are converted to spectra using DCT as shown in [18]. The group delay functions and modified group dealy functions computed for a frame of music is shown in Figure 1. In the proposed experiment, 20 dimensional modified group delay features (MODGD) are computed using frame size of 40 ms and hopsize of 10 ms.

*3) Low-level timbral features:* Timbral features that serve as a physical correlate to perceptual attributes differentiate a mixture of sounds that are with the same or similar rhythmic and pitch contents [19, 2]. 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. In our experiment, five features, namely, spectral centroid, spectral roll-off, spectral entropy, zero crossings, and low-energy are computed in track-level [19]. The mapping of 5 instrument classes in two-dimensional timbral space in Figure 2 clearly shows the importance of timbral descriptors for the instrument classification task.

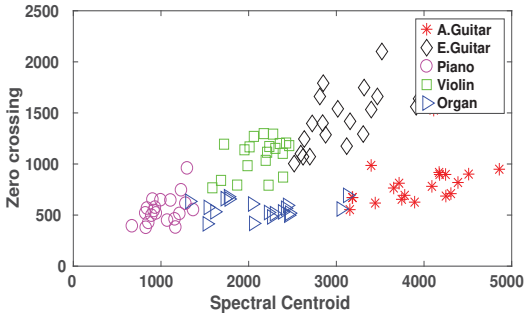

Fig. 2: Two dimensional mapping of five classes using spectral centroid and zero crossing for training files

### B. Classification Framework

In the classification phase, GMM, DNN, and hybrid GMM-DNN frameworks are employed. Initially, the classification is performed using a 256 mixture-GMM based baseline system which employs the maximum likelihood (ML) criteria for the identification task. Later, the experiment is extended using the DNN framework. Our proposed DNN architecture [20] is based on three hidden layered feed-forward neural network (100 nodes per layer) and the AdaMax optimization algorithm. Rectified linear units (ReLUs) have been chosen as the activation function for hidden layers and softmax function for the output layer. The network is trained in 1000 epochs with a batch size of 10. In the final phase, a combined discriminative/generative formulation is derived that leverages the complementary strengths of both models. Thus, a hybrid GMM-DNN framework is adopted to investigate the promise of GMM log-likelihood features, in training a DNN framework. GMM is a generative model and it fits the training data so that the likelihood of the data given the model is maximized. In contrast, DNN is a discriminative model, and its parameters are trained to minimize the classification error.

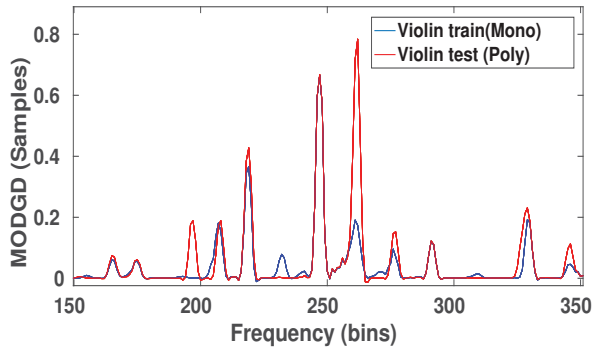


Fig. 3: The summary-MODGD gram of audio segments.(a) monophonic(violin),(b) polyphonic (violin,predominant).

Thus, the DNN is trained using generative features which is the log-likelihood of the features, given the instrument model obtained from GMM.

## III. Performance Evaluation

### A. Dataset

We considered five classes, namely, acoustic guitar (A.Gu), electric guitar (E.Gu), organ (Org), piano (Pia) and violin (Vio) for the proposed experiment from IRMAS dataset [21]. Since the dataset comprises of audio samples with multiple predominant instruments in the same audio file, we have considered audio test files with only one predominant instrument. The testing set comprises of 250 polyphonic excerpts (50 per class) annotated with one instrument as predominant.

TABLE I: Overall accuracy for the experiments (in %)

| Feature ↓ \ Method → | GMM | DNN | **GMM+DNN** |
|---|---|---|---|
| MFCC | 50.80 | 57.60 | 61.20 |
| MODGDF | 56.40 | 65.20 | 69.60 |
| MFCC+MODGDF | 58.80 | 79.60 | 79.60 |
| **MFCC+MODGDF+Timbral** | **65.60** | **85.60** | **93.20** |

TABLE II: Confusion matrix of MFCC-GMM-DNN experiment. Class-wise accuracy is entered in the last column

| | A.Gu | E.Gu | Org | Pia | Vio | **Accr.(%)** |
|---|---|---|---|---|---|---|
| A.Gu | **41** | 2 | 6 | 1 | 0 | 82 |
| E.Gu | 8 | **28** | 0 | 2 | 12 | 56 |
| Org | 20 | 7 | **16** | 7 | 0 | 32 |
| Pia | 9 | 1 | 5 | **34** | 1 | 68 |
| Vio | 14 | 1 | 1 | 0 | **34** | 68 |

### B. Experimental Set-up

MFCC (20 dim) and MODGDF (20 dim) are frame-wise computed for every 10 ms with a frame size of 40 ms. MFCC is fused with MODGDF in feature level to form 40 dim feature vector for fusion. Later the track level computed timbral feature set is appended to the previously fused feature vector yielding 45 dim. The low-level timbral features are computed using MIRToolbox [22]. The experiment is progressed in four stages namely MFCC, MODGDF, early fusion of MFCC and MODGDF and finally, the fusion of the entire feature set. Classifier is trained using monophonic files (600 files approx.) and tested on polyphonic files with one leading instrument. DNN is implemented using Keras-TensorFlow.

TABLE III: Confusion matrix of MODGDF-GMM-DNN experiment.

| | A.Gu | E.Gu | Org | Pia | Vio | **Accr.(%)** |
|---|---|---|---|---|---|---|
| A.Gu | **34** | 2 | 3 | 10 | 1 | 68 |
| E.Gu | 16 | **22** | 8 | 3 | 1 | 44 |
| Org | 10 | 3 | **30** | 6 | 1 | 60 |
| Pia | 4 | 2 | 4 | **39** | 1 | 78 |
| Vio | 0 | 0 | 0 | 1 | **49** | 98 |

TABLE IV: Confusion matrix of (MFCC+MODGDF)-GMM-DNN expt.

| | A.Gu | E.Gu | Org | Pia | Vio | **Accr.(%)** |
|---|---|---|---|---|---|---|
| A.Gu | **33** | 7 | 3 | 6 | 1 | 66 |
| E.Gu | 0 | **49** | 0 | 0 | 1 | 98 |
| Org | 9 | 6 | **26** | 8 | 1 | 52 |
| Pia | 3 | 0 | 0 | **47** | 0 | 94 |
| Vio | 2 | 2 | 1 | 1 | **44** | 88 |

## IV. Results and Analysis

The results of the experiments are tabulated in Table I. It can be observed that for the GMM framework, the results reported are 50.80%, 56.40%, 58.80% and 65.60% for MFCC, MODGDF, MFCC+MODGDF, and fusion of entire timbral feature set, respectively. It supports the hypothesis that timbral feature fusion shows improvement in the system performance due to the complementary information captured by the individual set. Also, From Table I, It is observed that the architectural choice of deep learning methodologies showed improvement in all stages with the best result for fusion scheme (85.60%, ref. third column Table I). In the hybrid GMM-DNN system, the results reported are 61.20%, 69.60%, 79.60% and 93.20% for the above feature sets. It can be seen that by inputting GMM log-likelihoods to DNN, we obtain about 4% absolute improvement over the naive DNN-MFCC framework. In the case of the entire feature combination, 8% improvement is observed for the hybrid GMM-DNN framework (93.20%) over DNN (85.60%). It appears that the likelihoods are more effective as features to the neural network due to the large dynamic range of the GMM likelihoods[23]. It is worth noting that, complementary strengths of the combined discriminative/generative formulation played a crucial role in improving the accuracy of the hybrid system.

As the final step in the analysis, the effectiveness of the feature fusion on timbral descriptors can be examined using the results of the hybrid GMM-DNN framework. The confusion matrices of MFCC and MODGDF experiments are shown

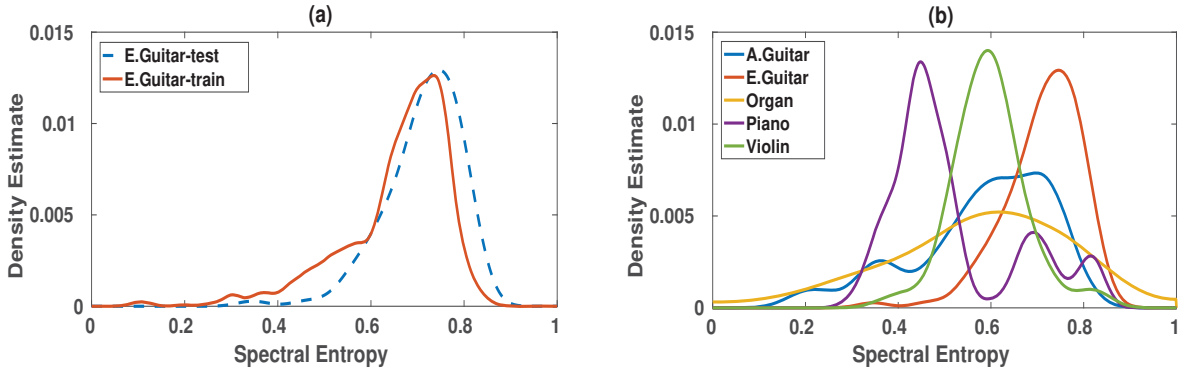


Fig. 4: (a) KDE of spectral entropy computed for train and test files of Electric guitar. Train file is monophonic and Test file is polyphonic with Electric guitar as predominant. (b) KDE of five classes of test data with each one is predominant

TABLE V: Confusion matrix of (MFCC+MODGDF+Timbral)–GMM-DNN experiment.

| | A.Gu | E.Gu | Org | Pia | Vio | **Accr. (%)** |
|---|---|---|---|---|---|---|
| A.Gu | **48** | 2 | 0 | 0 | 0 | 96 |
| E.Gu | 2 | **48** | 0 | 0 | 0 | 96 |
| Org | 2 | 0 | **46** | 1 | 1 | 92 |
| Pia | 3 | 2 | 0 | **44** | 1 | 88 |
| Vio | 1 | 2 | 0 | 0 | **47** | 94 |

in Tables II and III respectively. In the MFCC experiment, it is observed that acoustic guitar, piano, and violin are the better-perceived instruments (82%,68%, and 68%), organ on the contrary, the worst (32%). The identification accuracy for organ in the experiment of Y.Han et. al [1] shows a similar trend, owing to the fact that onset is an important clue in judging the predominant instrument in the audio clips. Meanwhile, during the MODGDF experiment (Ref. Table III), the classification accuracy of the organ is considerably improved (60%) as compared to the rest and the overall accuracy is 69.60% with an improvement of 8.40% over MFCC-GMM-DNN system. The significance of MODGDF as stated in [9], is reflected in this result. In Figure 3, summary-MODGDgram [1]for 500 successive frames of violin (isolated and polyphonic environment). It appears that the peaks in the group delay function coincide, which emphasizes the promise of MODGDF in detecting predominant instruments. The confusion matrix of MFCC+MODGDF is shown in Table IV. Fusion of MFCC and MODGDF resulted in an improvement of 18.40% over MFCC for the hybrid framework.

In the final phase, all the timbral descriptors are fused, which results in an overall accuracy of 93.20% (for hybrid framework) with individual accuracy greater than 88% (Ref-Table V). From the table, we found that considerable confusion occurs for acoustic guitar and organ with other classes during the MFCC-MODGDF experiment. Combining the musical texture features with the existing feature set reduces this confusion, leading to an overall increase in accuracy. When we compare the results of feature-fusion across the classifiers, the GMM-DNN shows an improvement of 27.60% over the baseline GMM (ref. last row in Table V). The complementary information captured by the low level feature set improved the result by 13.60% over MFCC-MODGDF fusion.

The effectiveness of low-level timbral features can be explained using Figure 4. From Figure 4 (a) it can be seen that kernel density estimate (KDE) [24] of spectral entropy of acoustic guitar in isolated and in polyphonic environment matches even in the presence of accompaniments. Moreover, the nature of KDE of spectral entropy varies across instrument classes as shown in Figure 4 (b). It supports our hypothesis that low-level timbral features are more important acoustic cues in instrument recognition.

From [25] it can be observed that most of the techniques report classification accuracy, a maximum of 90% for 10 instruments. In [26], the recognition rates of 84.1% for duo, 77.6% for trio, and 72.3% for quartet are reported for RWC dataset using both features weighting and musical context. It is quite understandable that the results of the proposed system are of the same order if not better than the results achieved by using several DNN/CNN -based methods. To end the discussion, the performance evaluation supports our claim that the architectural choice of deep learning methodology on timbral feature fusion has merit in the predominant instrument recognition task.

## V. Conclusion

Predominant instrument recognition in polyphonic music is addressed in this paper. Three timbral descriptors namely, MFCC, MODGDF, and low-level timbral feature sets are computed and experimented with GMM, DNN and GMM-DNN classifiers. The performance is evaluated using a subset of IRMAS dataset. The fusion of features on the GMM-DNN framework resulted in an overall accuracy of 93.20% with a significant improvement over the GMM-MFCC system. The results show the potential of timbral descriptors and deep learning methodologies in predominant instrument recognition.

[1]summary-MODGDgram is obtained by bin-wise addition of MODGDSs of consecutive frames for a music segment

## REFERENCES


[1] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Trans. Acou., Speech, Lang. Proces.*, vol. 25, no. 1, pp. 208–221, 2017.

[2] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *J. Acoust. Soc. Amer.*, vol. 130, no. 5, pp. 2902–2916, 2011.

[3] Y. Li-Fan, S. Li, and Y. Yi-Hsuan, "Sparse cepstral codesand power scale for instrument identification," *in Proc. of Int. Conf. Acou., Speech, and Sig. Proces.*, pp. 7460–7464, 2014.

[4] I. Kaminsky and Materka., "Automatic source identification of monophonic musical instrument sound," *in Proc. of the IEEE Int. Conf. on Neu. Networks*, pp. 185–194, 2011.

[5] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," *in Proc. of Int. Soc. Music Inf. Retri. Conf*, pp. 327–332, 2009.

[6] S. Essid, G. Richard, and B. David., "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. on Audio, Speech, and Lang. Proces.*, vol. 14, no. 1, pp. 68–80, 2006.

[7] P. Li, J. Qian, and T. Wang, "Automatic instrument recognition in polyphonic music using convolutional neural networks," *arXiv preprint arXiv:1511.05520,*, 2015.

[8] F. Fuhrmann and P. Herrera, "Polyphonic instrument recognition for exploring semantic similarities in music," *in Proc. of 13th Int. Conf. on Digital Audio Effects DAFx10, Graz Austria*, vol. 14, no. 1, pp. 1–8, 2010.

[9] A. Diment, P. Rajan, T. Heittola, and T. Virtanen, "Modified group delay feature for musical instrument recognition," *in Proc. of 10th Int. Symp. Comput. Music Multidiscip. Res., Marseille, France*, pp. 431–438, 2013.

[10] F. Fuhrmann., "Automatic musical instrument recognition from polyphonic music audio signals," *PhD thesis, Universitat Pompeu Fabra,*, 2012.

[11] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its application to speech technology," *Sadhana*, vol. 36, no. 5, pp. 745–782, 2011.

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

[13] R. Rajan and H. A. Murthy, "Group delay based melody monopitch extraction from music," *in Proc. of the IEEE Int. Conf. on Audio, Speech and Sig. Proces.*, pp. 186–190, 2013.

[14] ——, "Melodic pitch extraction from music signals using modified group delay functions," *In Proc. of the Communications (NCC), 2013 National Conference on*, pp. 1–5, February 2013.

[15] ——, "Music genre classification by fusion of modified group delay and melodic features," *in Proc. of National Conference on Communications*, 2017.

[16] ——, "Two-pitch tracking in co-channel speech using modified group delay functions," *Speech Communication*, vol. 89, pp. 37–46, 2017.

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

[18] Hegde, Rajesh M., "Fourier transform based features for speech recognition," PhD Dissertation, Indian Institute of Technology Madras, July 2005.

[19] T. Li, M. Ogihara, and Q. Li, "A comparitive study on content-based music genre classification," *in Proc. of the 26th Annual Int. Conf. on Research and Development in Inf. Retri.*, pp. 282–289, 2003.

[20] G. Dahl, "Deep learning approaches to problems in speech recognition, computational chemistry, and natural language text processing," PhD Dissertation, University of Toronto, 2015.

[21] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." *in Proc. of 13th Int. Soc. for Music Inf. Retri. Conf.*, pp. 559–564, 2012.

[22] O. Lartillot, P. Toiviainen, and T. Eerola, "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*, pp. 261–268, 2008.

[23] J. Pinto and H. Hermansky, "Combining evidence from a generative and a discriminative model in phoneme recognition," *in Proc. of Int. Conf. Spoken Lang. Proces.*, pp. 2414–2417, 2008.

[24] Y.-C. Chen, "A tutorial on kernel density estimation and recent advances," *arXiv:1704.03924*, 2017.

[25] G. P. P. Herrera-Boyer and S. Dubnov, "Automatic classification of musical instrument sound," *Journal of New Music Research,*, vol. 32, no. 1, pp. 3–21, 2003.

[26] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H.G.Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," *EURASIP J. Appl. Sig. Proces.,*, pp. 155–175, 2007.