World Academy of Science, Engineering and Technology
International Journal of Electrical and Computer Engineering
Vol:6, No:7, 2012

Musical Instrument Classification Using Embedded Hidden Markov Models

Ehsan Amid, *Student Member, IEEE,* Sina Rezaei Aghdam, *Student Member, IEEE*

Abstract—In this paper, a novel method for recognition of musical instruments in a polyphonic music is presented by using an embedded hidden Markov model (EHMM). EHMM is a doubly embedded HMM structure where each state of the external HMM is an independent HMM. The classification is accomplished for two different internal HMM structures where GMMs are used as likelihood estimators for the internal HMMs. The results are compared to those achieved by an artificial neural network with two hidden layers. Appropriate classification accuracies were achieved both for solo instrument performance and instrument combinations which demonstrates that the new approach outperforms the similar classification methods by means of the dynamic of the signal.

Keywords—hidden Markov model (HMM), embedded hidden Markov models (EHMM), MFCC, musical instrument.

I. Introduction

MUSIC has become an inseparable part of everyday human life. Moreover, with the advent of online music, extensive music databases have become accessible with a high variety of genres including a wide range of instruments. Meanwhile, this diversity has imposed additional problems for users as well as researchers. At the moment, the need for fast and efficient methods to perform content based search of music has substantially grown. For example, a common requirement for these engines is the ability to recognize different instruments presented in a polyphonic music.

Musical instrument recognition in a polyphonic music consists of a multiple pitch estimation process which includes the task of estimating fundamental frequencies and the onset times of notes associated with different instruments presented in a music signal. However, it is accounted as a challenging task while harmonics of different pitches overlap [4]. Therefore, this complexity has imposed a high demand for a fast and moderate search engine to perform a pitch independent analysis of instruments in a music signal.

Statistical models have been extensively used in speech recognition as well as musical information retrieval (MIR). For example, Marques and Moreno [27] used Mel-frequency cepstral coefficients as feature and attained %70 accuracy for distinguishing eight instruments using a support vector machines (SVM) classifier. Unlike deterministic models, statistical models aim to characterize statistical properties of a signal [1]. However, the main drawback associated with the aforementioned approaches is that the feature vectors associated with the signal are considered as i.i.d. random variables from an instance of a random process assumed to generate the signal without accounting for the dynamic behavior of the signal which is substantial for perception of acoustic phenomena by the human brain [5]. For this purpose, one can benefit from a more comprehensive model, namely hidden Markov model, which attempts to model frame-to-frame dynamic in music by means of state transition [2]. Such dynamic models have been considerably used in music analysis [6-10]. As a simple configuration, a single HMM can be trained in an unsupervised manner by a music signal and can be exploited to find music textures in a similar music [14]. In this case, each state is considered as a specific texture and by finding the state sequence which best matches the signal, one can find the sequence of textures representing the corresponding instrument combinations in the music. However, due to unsupervised training of the proposed method, feature vectors of distinct states may intermingle incorrectly leading to an inaccurate model. Another issue arises in finding the proper number of states for the model which needs additional process on the model to find the indistinguishable states. Additional techniques such as iHMM [25] can be used to infer the appropriate number of states or to find a more complex structure for the classification task [11]. Nevertheless, finding a supervised training approach is still a challenging task. In this paper, an alternative method based on embedded hidden Markov model is exploited to overcome these issues. An embedded hidden Markov model (EHMM) is an extension of an ordinary HMM where each state is composed of an independent HMM [26].

The remainder of this paper is organized as follows. Embedded hidden Markov model is described in section II. Section III provides feature extraction method and extraction of MFCC features used in this paper. The proposed method using EHMM is provided in Section IV. Experimental results are presented in Section V. Section VI concludes the paper and outlines the future work.

E. Amid and S. R. Aghdam e-mail: {ehsan_amid, sina.rezaei}@aut.ac.ir

II. Embedded Hidden Markov Models

A. Hidden Markov Models

A traditional one-dimensional HMM is a doubly embedded stochastic process where the inner process includes a set of states which are not directly observable and the observation is performed through another set of stochastic processes that generates observable events [1]. At each time step, a transition occurs among the current state and the other states (possibly the same state) and an observation takes place regarding the new state. These observations are modeled by a set of probability distributions associated with each state. In case of a continuous observation HMM, these conditional probability

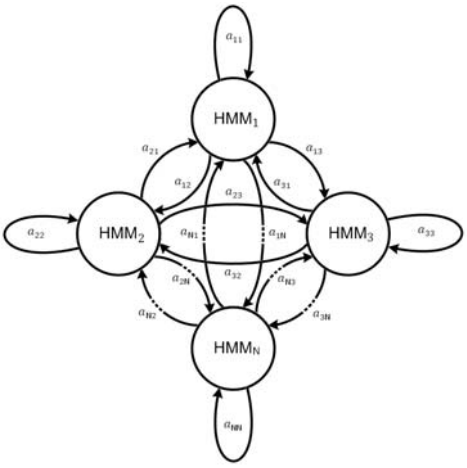

Fig. 1. A Typical EHMM Structure.

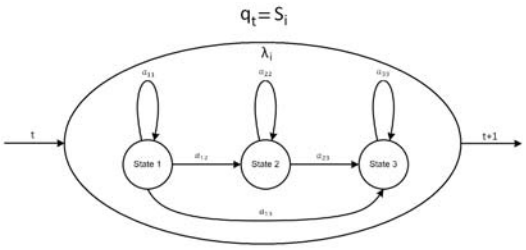

Fig. 2. State Transition in EHMM.

distributions are usually mixtures of Gaussian distributions called Gaussian Mixture Models (GMM) [2].
An HMM is denoted by $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ where $\mathbf{A}$, $\mathbf{B}$ and $\boldsymbol{\pi}$ are defined as follows:

- $\mathbf{A} = \{a_{ij}\}, a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq n$: State transition probability distribution
- $\mathbf{B} = \{b_i(O_t)\}, b_i(O_t) = P(O_t | q_t = S_i)$: State observation probability distribution
- $\boldsymbol{\pi} = \{\pi_i\}, \pi_i = P(q_1 = S_i)$: Initial state probability distribution

For a given model λ and observation sequence O, the joint probability of the observation sequence and the underlying state sequence is given by

$$P(O, Q|\lambda) = \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \prod_{t=1}^{T} b_{q_t}(O_t) \tag{1}$$

The probability of O (given the model) is obtained by summing over all possible state sequences Q

$$P(O|\lambda) = \sum_{\text{all Q}} \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \prod_{t=1}^{T} b_{q_t}(O_t). \tag{2}$$

B. Embedded Hidden Markov Models

The idea of a Hidden Markov Model can be extended to a more complex two-dimensional structure where each super state associated with an external HMM is composed of an independent HMM which includes a set of intrinsic states. This configuration is called an embedded hidden Markov model or simply an EHMM [15]. The structure of a typical EHMM is shown in Fig. 1.
Each observation related to the external HMM is called a super observation (or a super block). This super observation consists of a sequence of consecutive observations associated with the internal HMMs. So, between two state transition of the external HMM, many intrinsic state transitions occur between the states of the HMM corresponding to the current state. The idea of super observation is illustrated in Fig. 2.

Let O_t be the super observation at time t and the model be in state S_i, $i = 1, 2, \ldots, N$, at t i.e. $q_t = S_i$. O_t represents a sequence of observation vectors, namely

$$O_t = O_{t_1} O_{t_2} \cdots O_{t_l} \tag{3}$$

as well, where l is the length of the super observation. The probability of this observation, given the state S_i at time t and the EHMM, λ, is

$$P(O_t | q_t = S_i, \lambda) = P(O_t | \lambda_i) \tag{4}$$

Where λ_i denotes the internal HMM i. The internal HMMs in an EHMM act as likelihood estimators for the external HMM, similar to the GMMs in the one-dimensional case. But the difference with the traditional model lies in the ability of this model to split every super observation into a sequence of consecutive smaller observations which can be considered as a sequence of observations for the internal HMMs. Thus, by using this structure, it is possible to estimate the likelihood of every super observation while maintaining the temporal information of the observation.

III. Feature Extraction

Different features including temporal [17], spectral or cepstral [18] as well as power spectra [19] have been extensively used in music content analysis. In this paper, Mel-frequency cepstral coefficients and their first and second derivatives were used as features.

A. Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCC) provide a compact representation of the spectral envelope of the signal based on a STFT by using a model of human auditory system which has a higher resolution in lower frequencies. They can provide an effective pitch-independent feature to model transfer function of the auditory filter, regardless of the excitation source; even though they do not represent a homomorphic transform in the same way as complex and real cepstrum which are inverse Fourier transform of the logarithm or logarithm magnitude of the Fourier transform, respectively [2].
In addition to their reliability in speech recognition tasks, it has been shown that MFCC can be effectively used in music analysis [20-21]. However, they may be insufficient in case of a complex texture including many musical instruments playing simultaneously.

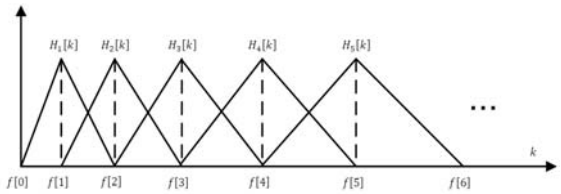

Fig. 3. Mel-frequency Filterbank.

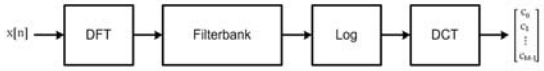

Fig. 4. Extraction of MFCC.

To obtain MFCC, the signal is first pre-emphasized by an all-zero filter of the form $H(z) = 1 - \alpha z^{-1}$ in which α typically ranges from 0.93 to 0.97. In this paper, α is chosen equal to 0.95. Then, the signal is framed into overlapping windows with a proper duration through which the signal is assumed to be quasi-stationary. Let $x[n]$ and $X[k]$ be the windowed signal and its DFT, respectively. A Mel-scale filterbank is imposed on the FFT of each window to obtain log-energy in every sub-band.

$$E_m = \ln\left(\left[\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right]\right) \tag{5}$$

where E_m is the log-energy of the signal in the m_{th} sub-band, $0 \leq m < M$. H_m denotes the m_{th} filter of the filterbank

$$H_m[k] = \begin{cases} 0 & k \geq f[m-1] \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])} & f[m-1] \leq k \leq f[m] \\ \frac{(f[m+1]-k)}{(f[m+1]-f[m])} & f[m] \leq k \leq f[m+1] \\ 0 & f[m+1] > k. \end{cases} \tag{6}$$

In (6), $f[m-1]$ and $f[m+1]$ represent the edges of the m_{th} filter and $f[m]$ denotes its center, as shown in Fig. 3. This filterbank is developed on a non-linear scale called Mel-frequency scale which approximates the behavior of the human auditory system [28]. Finally, a discrete cosine transform is applied to decorrelate the resulting coefficients. It is shown in [22] that DCT is appropriate both for speech and music spectra to achieve decorrelated vectors. The overall procedure to evaluate MFCC for an audio signal is shown in Fig. 4.
Fig. 5 shows feature extraction process where an additional Hamming window is multiplied to each frame to smooth the edges of the signal at both sides. A Hamming window is defined by

$$w[n] = 0.54 - 0.46\cos(\frac{2\pi n}{N}), \;\; 0 \leq n \leq N \tag{7}$$

where $N+1$ is the length of the window [2]. This would be effective to reduce the effect of discontinuity of the signal in each segment at both sides [23].

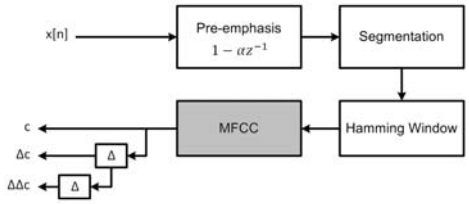

Fig. 5. Feature Extraction.

IV. EHMM in Musical Instrument Classification in a Polyphonic Music

In a polyphonic music, different instruments are played individually or simultaneously. Assuming the total number of instruments in a music signal to be c, there exist $N = 2^c - 1$ distinct combinations of instruments (Considering silence as a combination yields 2^c combinations). Each combination is considered as a music texture. These textures are performed successively, based on a specific principle which is highly dependent on music genre as well as types of the instruments of the track [24]. This may be a common repetition of one particular texture after another [12]. These principles can be exploited to recognize different textures in a polyphonic music by using an EHMM as follows.
First, for every possible texture, an individual HMM is trained by using any conventional re-estimation algorithm (e.g. Baum-Welch Algorithm [29]). Since, there may be numerous amount of instruments in different music tracks, we restrain ourselves to a special genre with only a limited number of textures. Generalization of the proposed method to consider different genres will be discussed later. Additionally, textures with trivial chance of occurrence can be ignored since they may not appear in practical music performances.
Let λ_i, $i = 1, 2, \ldots, 2^c - 1$ be the model trained for texture i and θ_t be the texture at time t. These models are considered as likelihood estimators of states in an ergodic EHMM structure. For any given music signal, the corresponding observation sequence, $O = O_1 O_2 \cdots O_T$ is extracted and then, the state sequence which best explains the observation is found using a dynamic programming method called Viterbi algorithm [16] by defining

$$\delta_t(i) = \underbrace{max}_{q_1, q_2, \ldots, q_{t-1}} P[q_1\, q_2 \cdots q_t = S_i, O_1 O_2 \cdots O_t|\lambda] \tag{8}$$

and maximizing over all possible paths at time T. $\delta_t(i)$ denotes the highest probability of a particular state sequence, at time t, which accounts for the first t observations of the observation sequence and ends in state S_i. Viterbi algorithm is described in the appendix.
By assuming O_t as stated in (3) to be super observation of l consecutive MFCC vectors of an observation sequence, every time step of the external HMM is equal to l times the time step of every internal HMM. This requires the assumption that each texture in the track is played at least for l consecutive blocks. This is not an inappropriate hypothesis since in a real-world music performance, a texture usually lasts at least for a considerable portion of a second [24].

TABLE I
CLASSIFICATION RESULTS FOR THE FIRST EXPERIMENT WITH HMMS INCLUDING 3 STATES AND 2 GAUSSIAN MIXTURE MODELS.

	Violin	Piano	Duet
Violin	**%92.87**	%0.00	%7.13
Piano	%0.82	**%95.08**	%4.10
Duet	%0.00	%5.42	**%94.58**

TABLE II
CLASSIFICATION RESULTS FOR THE SECOND EXPERIMENT WITH HMMS INCLUDING 4 STATES AND 5 GAUSSIAN MIXTURE MODELS.

	Violin	Piano	Duet
Violin	**%95.47**	%0.00	%4.53
Piano	%0.60	**%96.02**	%3.38
Duet	%0.00	%3.28	**%96.72**

TABLE III
CLASSIFICATION RESULTS FOR NEURAL NETWORK CLASSIFIER WITH 2 HIDDEN LAYERS.

	Violin	Piano	Duet
Violin	**%92.98**	%0.00	%7.02
Piano	%0.95	**%97.74**	%1.31
Duet	%30.36	%3.81	**%65.83**

TABLE IV
CLASSIFICATION RESULTS FOR SUPER OBSERVATIONS EACH EQUAL TO 1 SECOND OF THE SIGNAL.

	Violin	Piano	Duet
Violin	**%93.33**	%0.00	%6.67
Piano	%0.00	**%99.86**	%0.14
Duet	%20.00	%0.00	**%80.00**

Therefore, an EHMM having HMMs corresponding to different combinations as likelihood estimators and with a set of proper transition distribution among the states can be used to find the best sequence of states that matches any given music signal, namely,

$$Q = q_1\, q_2 \cdots q_T. \tag{9}$$

Since $q_t = S_i$ corresponds to i_{th} state of the EHMM, λ_i, finding the best sequence is equal to finding the sequence of textures, Θ, played in that track,

$$\Theta = \theta_1\, \theta_2 \cdots \theta_T. \tag{10}$$

Using the proposed structure in finding the sequence of musical instrument combinations in a polyphonic music is appropriate in two manners: First, it takes into account more transitional properties of the music signal while it determines the probability of occurrence of any given observation which consists of a super block by using an intrinsic HMM that is capable of performing several transitions within its states which reflects temporal characteristics of the signal. Second, by means of state transition probability distribution of the EHMM,

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i), 1 \leq i, j \leq N \tag{11}$$

one can impose an a priori knowledge about the musical events while finding the best state sequence for the input observation. This can be done by training the external HMM (finding the state transition probability distribution and the initial state probability distribution for EHMM) using several music tracks which results in probabilities which reflect a statistical analysis of the musical structure in a specific genre. This means those successive combinations which have more chance of occurrence in a real-world performance will be given a high probability while those being scarce will be ignored with a lower probability of incidence. Moreover, additional knowledge from various sources including music experts may be effective on assigning these transition probabilities. As a result, the state transition probability distribution will impact on finding the best state sequence of a given observation and enhance the accuracy.

Given the state transition probability distribution, a_{ij}, and $\delta_t(i)$, $i = 1, 2, \ldots, n$, $\delta_{t+1}(j)$ can be determined using the inductive equation

$$\delta_{t+1}(j) = \underbrace{max}_{i}[\delta_t(i) a_{ij}].P(O_t|\lambda_j). \tag{12}$$

This eliminates the chance of occurrence of invalid state sequences where there exists two specific states that can not succeed ($a_{ij} = 0$). Additionally, state transition probability distribution is applied in finding the best path by means of maximizing over all the states that can precede the current state.

V. EXPERIMENTAL RESULTS

The database used in this paper consists of 50 classical music tracks each comprising two different instruments, i.e. piano and violin. Consequently, it requires three separate HMMs to be trained for the EHMM. These HMMS were trained using music clips of one second duration by a multiple-observation Baum-welch algorithm [30]. To extract MFCC features, each clip was segmented into 25 ms windows with %66 overlap. A left-right model was used for each HMM with three states and two Gaussian mixture models for the first experiment and four states and five mixtures for the second one. After training these HMMs, they were used in an ergodic EHMM structure with three states. To train the EHMM, the same training music pieces were used with each super block consisting of $l = 120$ consecutive frames of the internal HMMs, equal to one second of music duration. For testing purpose, the test set including 30 music pieces of the same genre was applied to the model. The results are shown in Table I and Table II for the first and second experiment, respectively.

It can be seen that the highest accuracy was achieved for the duet class while the lowest was attained for the violin. Moreover, most of the errors occur in misclassification of single instruments as combination of instruments. This may have two main reasons. First, the resonance of the instruments are ignored in the labeling phase of the music pieces leading to some incorrect labels i.e. the other instrument which actually is not playing is still resonating and affecting the signal. The second reason is that MFCC are not sufficient features for the purpose of multiple-instrument representation. This can be

made up for by augmenting other features such as temporal and spectral features which are capable of characterizing important properties of music such as music timbre [31].

To compare the results, a MLP neural network classifier with two hidden layers consisting of 20 and 15 neurons in the first and second hidden layer, respectively, was used to classify different instrument combinations. Same MFCC vectors were used as features. The results are shown in Table III.

As shown in table, an NN classifier achieves higher accuracy in case of solo instrument performance classification, but it considerably underperforms when classifying instrument combination. The main reason for this low accuracy is due to the fact that it is impractical to label musical clips of very short durations since it is not possible to distinguish between musical instruments or to detect sound source activity. Another approach would be to consider longer time intervals and assign the most frequently occurring label to the corresponding clip. Table IV shows the classification results for time intervals of one second, equivalent to intervals used in the proposed method for EHMM.

An enhancement in classification accuracy is obtained both for the case of solo and instrument combination performance. However, the proposed method outperforms in instrument combination classification, in all cases. This higher accuracy is due to EHMM's dynamic structure which incorporates a priori knowledge about the musical events by means of a state transition probability distribution. This transition coefficients modify the probability of occurrence of different combinations in the next step, given the internal models and the current state.

VI. Conclusion

In this paper, the ability of EHMM structure in musical instrument classification in a polyphonic music was investigated. The results proved that this structure can be effectively used for classification of musical instruments in a specific genre, including a limited number of instruments. To extend the proposed method for the purpose of the classification of instruments in music pieces from different genres, first a typical genre classification technique as presented in [32] and [33] can be used to detect the genre of a specific music piece and then the corresponding EHMM can be imposed for instrument classification. Comparison of the results with those achieved by a MLP neural network indicates that the proposed method outperforms the NN classifier in musical combination classification while attaining comparable accuracy in case of solo instrument performance. Furthermore, different temporal and spectral features can be utilized along with the MFCC to improve the performance.

Appendix
Viterbi Algorithm

To find the single best state sequence, $Q = q_1\, q_2 \cdots q_T$ for a given observation sequence, $O = O_1\, O_2 \cdots O_T$, $\delta_t(i)$ is defined by (8). It can be evaluated by the inductive equation

$$\delta_{t+1}(j) = \underbrace{max}_{1 \le i \le n} [\delta_t(i) a_{ij}].b_j(O_t|\lambda). \tag{13}$$

The overall procedure can be stated as follow where an additional parameter, $\psi_t(j)$, is included to keep track of the argument which maximizes (eq) for each t and j.

1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1), 1 \le i \le n \tag{14}$$
$$\psi_1(i) = 0. \tag{15}$$

2) Recursion:

$$\delta_t(j) = \underbrace{max}_{1 \le i \le n} [\delta_{t-1}(i) a_{ij}].b_j(O_t|\lambda), \tag{16}$$
$$2 \le t \le T,\ 1 \le j \le n$$
$$\psi_t(j) = arg\ \underbrace{max}_{1 \le i \le n} [\delta_t(i) a_{ij}], \tag{17}$$
$$2 \le t \le T,\ 1 \le j \le n$$

3) Termination:

$$P^* = \underbrace{max}_{1 \le i \le n} [\delta_T(i)]$$
$$q_T^* = arg\ \underbrace{max}_{1 \le i \le n} [\delta_T(i)], \tag{18}$$

4) Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, \cdots, 1. \tag{19}$$

Acknowledgment

The authors would like to thank Hamed Dilish for his support and assistance.

References

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE,* vol. 77, no. 2, pp. 257-286, Feb. 1989.

[2] Xuedong Huang, Alejandro Acero, Alex Acero and Hsiao-Wuen Hon, *Spoken language processing: a guide to theory, algorithm, and system development*, Prentice Hall PTR, 2001.

[3] Lawrence R. Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Pearson Education, 1993.

[4] Jun Wu, E. Vincent, S. A. Raczynski, T. Nishimoto, N. Ono and S. Sagayama, "Polyphonic Pitch Estimation and Instrument Identification by Joint Modeling of Sustained and Attack Sounds", *Selected Topics in Signal Processing, IEEE Journal of* , vol.5, no.6, pp.1124-1132, Oct. 2011

[5] J. J. Aucouturier and M. Sandler, "Segmentation of musical signals using hidden Markov models", *presented at the 110th Conv. Audio Eng. Soc.*, May 2001.

[6] T. Virtanen and T. Heittola, "Interpolating hidden Markov model and its application to automatic instrument recognition", *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* , vol., no., pp.49-52, 19-24 April 2009.

[7] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden Markov models", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 4, pp. 360370, Apr. 1999.

[8] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs", *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on* , vol.2, no., pp. 133- 136 vol.2, 1-4 July 2003.

[9] Jonghyun Lee and Joohwan Chun, "Musical instruments recognition using hidden Markov model", *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on* , vol.1, no., pp.196-199 vol.1, 3-6 Nov. 2002.

[10] N. Degara, M. E. P. Davies, A. Pena and M. D. Plumbley, "Onset Event Decoding Exploiting the Rhythmic Structure of Polyphonic Music", *Selected Topics in Signal Processing, IEEE Journal of* , vol.5, no.6, pp.1228-1239, Oct. 2011.
[11] Yuting Qi, J. W. Paisley, L. Carin, "Music Analysis Using Hidden Markov Mixture Models", *Signal Processing, IEEE Transactions on* , vol.55, no.11, pp.5209-5224, Nov. 2007.
[12] R. J. Weiss and J. P. Bello, "Unsupervised Discovery of Temporal Structure in Music", *Selected Topics in Signal Processing, IEEE Journal of* , vol.5, no.6, pp.1240-1251, Oct. 2011.
[13] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Classification of musical patterns using variable duration hidden Markov models", *IEEE Trans. Audio, Speech, Lang. Process.* ,voI.14, pp.1795-1807, 2006.
[14] Jean-Julien Aucouturier and Mark Sandler, "Segmentation of Musical Signals Using Hidden Markov Models", *Presented at the 110th Convention*, Amsterdam, The Netherlands, 12-15 May 2001.
[15] Kai Shen, Sheng Gao, Peiqi Chai and Q. Sun, "Music Identification Using Embedded HMM", *Multimedia Signal Processing, 2005 IEEE 7th Workshop on* , vol., no., pp.1-4, Oct. 30 2005-Nov. 2 2005.
[16] G. D. Forney, "The Viterbi algorithm", *Proc. IEEE*, vol.61, pp. 268-278, Mar. 1973.
[17] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefcients and temporal features", *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000, vol. 2, pp. 753-756.
[18] J. C. Brown, "Computer identication of musical instruments using pattern recognition with cepstral coefcients as features", *J. Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 19331941, 1999.
[19] E. Vincent and X. Rodet, "Instrument identication in solo and ensemble music using independent subspace analysis", *in Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2004, pp. 576-581.
[20] A. Eronen, "Comparison of features for musical instrument recognition", *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the* , vol., no., pp.19-22, 2001.
[21] A. Eronen and A. Klapuri; , "Musical instrument recognition using cepstral coefficients and temporal features", *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on* , vol.2, no., pp.II753-II756 vol.2, 2000.
[22] Beth Logan, "Mel frequency cepstral coefficients for music modeling", *In International Symposium on Music Information Retrieval*, 2000.
[23] Monson H. Hayes, *Statistical digital signal processing and modeling*, John Wiley & Sons, Inc., 1996.
[24] N. C. Maddage, "Automatic structure detection for popular music", *Multimedia, IEEE* , vol.13, no.1, pp. 65- 77, Jan.-March 2006.
[25] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes", *J. Amer. Statist. Assoc.*, vol. 101, pp. 15661581, 2006.
[26] Katrin Weber, "HMM Mixtures (HMM2) for Robust Speech Recognition", http://www.idiap.ch/publications/weberrr-0334.bib.abs.html, 2003.
[27] J. Marques and P. Moreno, "A study of musical instrument classification using gaussian mixture models and support vector machines", *Compaq Computer Corporation*, Tech. Rep. CRL 99/4, 1999.
[28] S. S. Stevens and J. Volkman, "The Relation of Pitch to Frequency", *Journal of Psychology*, 1940, 53, pp. 329.
[29] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes", *Inequalities*, vol. 3, pp. 1-8, 1972.
[30] S. E. Levinson, L. R. Rabiner and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035-1074, Apr. 1983.
[31] Xin Zhang and Z. W. Ras, "Analysis of Sound Features for Music Timbre Recognition", *Multimedia and Ubiquitous Engineering*, 2007. MUE '07. International Conference on , vol., no., pp.3-8, 26-28 April 2007.
[32] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", *Speech and Audio Processing, IEEE Transactions on* , vol.10, no.5, pp. 293- 302, Jul 2002.
[33] Changsheng Xu, N. C. Maddage and Xi Shao, "Automatic music classification and summarization", *Speech and Audio Processing, IEEE Transactions on* , vol.13, no.3, pp. 441- 450, May 2005.

Ehsan Amid received the B.Sc. degree in electrical engineering from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2012. His research interests include speech processing, pattern recognition and machine learning.

Sina Rezaei Aghdam received B.Sc. degree in electrical engineering from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2011. He is currently working toward the M.Sc. degree at the Amirkabir University of Technology (Tehran Polytechnic).
His research interests range from wireless communications to digital signal processing.