Review on Automatic Music Transcription System

Surekha B.Puri
Department of Electronics and telecommunication,
College of engineering, Pune, India
gosavi.12@gmail.com

S. P. Mahajan
Department of Electronics and telecommunication,
College of engineering, Pune, India
spm.extc@coep.ac.in

***Abstract*—In this paper, the literature survey of the automatic music transcription system have been presented. Now a day's most of the research work going on Music transcription and it is considered to be a most difficult problem even by human experts and current music transcription systems fail to match human performance. As compare to Monophonic AMT the Polyphonic AMT is a difficult problem because in polyphonic concurrently sounding notes from one or more instruments cause a complex interaction and overlap of harmonics in the acoustic signal. So we concentrate on all methods of polyphonic AMT. Most of the music transcription systems were developed for the instruments typically used in western music like Piano, Guitar etc. but very less paper/work has been publishing in the domain of harmonium note transcription which is widely used the instrument in Indian musical concerts.**

Keywords—Automatic Music Transcription, HMM, LPC, KNN, monophonic, Music Language Models, polyphonic, PLCA, RNN, SVM

I. Introduction

The process of conversion of the acoustic musical signal into its equivalent pitch, source of sound and onset time is called as a music transcription system. In western tradition, the piece of music is represented by the written notes as shown in Fig. 1.

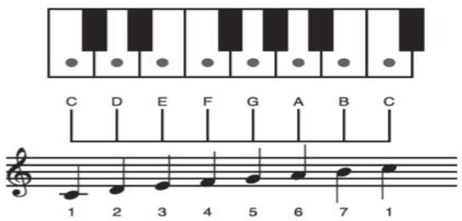

Fig. 1. Musical notation corresponding to the key

Music is the part of the human culture, it grows with human evaluation. The cultural music explosion had taken place between 60000 and 30000 years ago in Germany. The flute music was discovered 42000-43000 years ago using birds bone and mammoth ivory. This music is used for the oral tradition for many thousands of years. However, without recording the music notations, we haven't any idea about what type of music sounded like. The first development of (choral) musical notation was found in the church of Europe called "Plainchant" or "Gregorian chant". This music notation was based on whether the notes should higher or lower than previous notes.

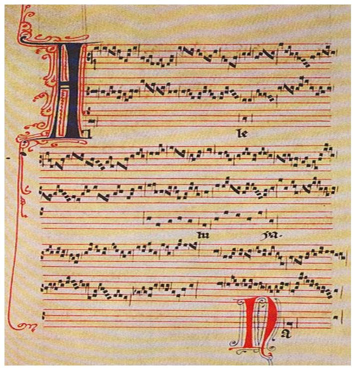

Fig. 2. Musical notation in church music

Firstly only one horizontal line was introduced, but later introduces stave of four horizontal lines.In the 16th century, the printed musical notations were attempted. But after the introduction of the printing press, the musical notations were printed by movable printers. In England, Queen Elizabeth granted to print and publish the music in the form of notation.

After the computer revolution, the music notation was recorded, edited, the process through the music software. Thismusic's are sounded like an original music. This is the musical revolution. The Notation software makes the musical field easier as we can make the correction in the middle of the piece, extraction of the piece of music from the music etc.In Indian music notation of Raga, Sargam is used. There are seven basic pitches of major scales i.e. SaReGaMaPaDhaNi (Shadja, Rishabh, Gandhar, Madhyam, Pancham, Dhaivat and Nishad). Sa and Pa are the known as 'achalaswar'. These are fixed notes. Another five notes (Re, Ga, Ma, Dha and Ni) are called 'Shudhha' pitch [1].

A. Monophony

Monophonic music texture is nothing but a one sound or single note and music only contains a melody line with no harmony. It is usually played by one person their own or by many people can play the same melody. Now a day's monophonic music is not generally used but some Middle Eastern music has a monophonic texture [2].

978-1-5386-4008-1/17/$31.00 ©2017 IEEE

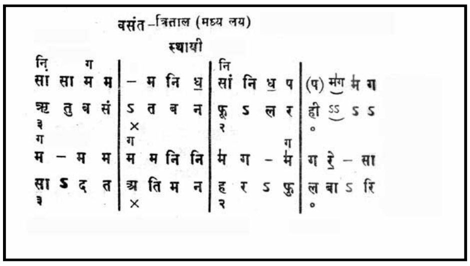

Fig. 3. Indian music notation

B. Heterophony

Heterophony music is formed by real-time recording by the different number of singers of different versions of the same tune. It contains parts of the different music of the same melody. The best example of the heterophony is the Jazz music [2].

C. Polyphony:

Polyphonic music is based on counterpoint which is the simultaneous performance of multiple melodies or tunes that are distinct from each other in notes and rhythm. The polyphony is nothing but the practice of controlling the relationship between the different melodies or tunes. It is one of musical structure the polyphony is typically described as thick or densely textured. In heterophonic texture, the part that all voices play is based on the same melody. The essence of polyphony was lost with the transition to harmony [2].

D. Homophony:

Homophony is defined as a texture in which we encounter most often multiple voices in which one it can consist of a single dominating melody and is accomplished by chords. There are the four different classical music textures in music [2].

II. AUTOMATIC MUSIC TRANSCRIPTION SYSTEM

The process of conversion of acoustic musical signals into another form of musical notation is called as Automatic music transcription (AMT). The musical notions are harmonium notes, sheet music, Musical Instrument Digital Interface (MIDI) file, piano rolls etc [3].

The transcription task is the translation from music to score. In the score, the notes are played in the form of time sequences which indicates the duration of pitches. The Fig. 4 shows the block diagram of the monophonic music transcription system.

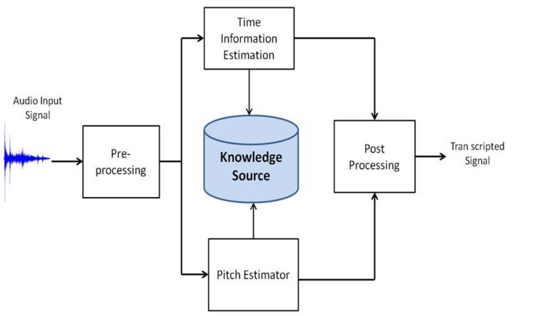

Fig. 4. General architecture of an automatic music transcription system

AMT is an important aspect of the Music Information Retrieval (MIR) aims to generate symbols. In Polyphonic AMT is a difficult problem because in which concurrently sounding notes from one or more instruments cause a complex interaction and overlap of harmonics in the acoustic signal [4].

Different authors define AMT in different ways. Emmanouil Benetos et al. [5] define "AMT is the process of converting an audio recording into a piano roll notation", while Klapuri et al. [6] define as the process of converting a recording into common music notation called Score. Likely, AMT system having unconstrained polyphony further complicates the modeling problem. Usually, model capture variability in the input signal and which aims to learn the timbral properties of the musical instrument [7] and issue related to large output space is solved by constructing the model which has maximum polyphony [8]. These issues with polyphonic music can be solved be unsupervised spectrogram factorization methods are addressed by incorporating harmonic constraints in the training algorithm [9][10].

III. MUSICAL TRANSCRIPTION MODEL

A. Acoustic Model

In the acoustic model file which contains the representation of the notation of the sounds that makes a meaningful word called acoustic model. Each label of the acoustic model is called 'Phoneme'. The English language is made up of 40 distinct sounds means 40 Phenoms which are useful for speech recognition. [17]

The acoustic model exist the relationship between the audio signal and the phonetic units in the language.

B. Music Language Model

In language models, Signal will capture the properties of a language and predict the next word in a speech sequence. The phonemes probability sequence are obtained by most similar word by lexical and language model. Language modeling is used in speech recognition, machine translation, part of speech tagging, parsing, handwriting recognition, information retrieval and other applications. From the review, most of the Language models are used in information retrieval in the query likelihood model.

C. Hybrid Model

The Hybrid model is the combination of acoustic and music language model. Most of the researchers used the convent for identifying pitches present in the input audio signal and compare their performance to various other acoustic models. The acoustic and language models are combined together under a single training objective using hybrid RNN architecture.

IV. GENERALIZED AUTOMATIC MUSIC TRANSCRIPTION SYSTEM

The generalized block diagram for AMT is shown in Fig. 5.

A. Database selection

Database selection is the first step of automatic music transcription system. To evaluate the performance of the developed system, the annotated database is required. The different database for various music instruments is available publically. Some of them are explaining below in detailed.

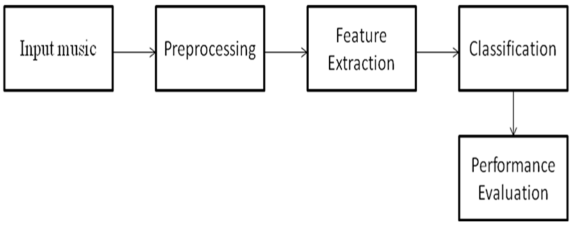

Fig. 5. Generalized Block diagram of automatic music transcription system

1. MAPS [18]

MAPS database provide 16 bit, 44 kHz sampled audio database for piano with ground truth. It contains about 65 hours of audio recording. The database is prepared using Disklavier (MIDI fied piano) and some high-quality synthesis software. In order to generalize the audio database, the musical instruments have been played in different recording conditions, different noise, different noise and using the different type of instrument.

The contents of MAPS are divided into four sets, which are detailed in section.

- The ISOL set: Isolated notes and monophonic excerpts.
- The RAND set: Chords with random pitch notes.
- The UCHO set: Usual chords from Western music.
- The MUS set: Pieces of piano music.

2. GTZAN [19]

George Tzanetak is collected the different piece of music for his research work to prepare database called GTZAN. It has been used to evaluate various genre classification systems. It contains 1000 song excerpts of 30 seconds, sampling rate 22050 Hz at 16 bit. All files are in .wav format. Its songs are distributed evenly into 10 different genres: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae and Rock.

3. RWC Musical Instrument Sound Database [20]

RWC (Real World Computing) is the copyright cleared, publically available music database aimed for research purpose. The database consists of 315 files of music which are broadly divided into popular music database (100 files), royalty free music (15 files), classical music (50 files), Jazz music (50 files), Music genre (100 files) and musical instrument (50 files). The database is prepared by MIDI files and its text files of lyrics.

4. MIR-1k [21]

Chao-Ling Hsu and Prof. Jyh-Shing Roger Jang work for preparing the Multimedia Information Retrieval lab (MIR-lab) dataset. It is mainly prepared for the singing voice separation. It contains 1000 song clips those are recorded with left and right channel. Data is annotated manually including pitch contours in semitone, indices and types for unvoiced frames, lyrics, and vocal/non-vocal segment. Duration of the database is 133 minutes having each music clip is of 4 to 13 second. The songs are selected from Chinese pop music which is sung by eight female and 12 males.

5. ENST-drums database [22]

Three Drummers namely Louis Cave, Bertrand Clouard and Frederic Rottier plays the drum to record the Drum music database. It is varied research database. It can mainly use for music transcription systems. The duration of recorded drum audio is 75 minutes.

6. ISMIR 2004 [23]

The idea of ISMIR database preparation has emerged at Music Technology Group of the Pompeu Fabra University. To accomplish a task 50 research group's works on audio analysis and synthesis. The database is prepared by three distinct sets of songs, two of them are training and development set and the third one are testing set. Total 729 tracks are recorded for classical (320), electronics (115), jazz (26), Metal (45), rock (101) and the world (122). All files are in .wav, 22.05 KHz, and mono format.

B. Feature

There are different types of features, such as the pitch, timbral features, rhythm features etc that are explained below. [24]

1. Pitch [25]

The perceived frequency of the musical note is called as pitch. In pitch frequency spacing of a harmonic series in the frequency domain representation of signal perceived logarithmically.

2. Timbral features [26]

The term of the auditory sensation, by which human can judge the sound called as timbre. In music, it is the quality measurement parameter of the music. The lowest frequency is called the fundamental frequency and the pitch produced by this frequency is used to name the note. The note is created by the number of frequencies in Hz. The lowest frequency is called the fundamental frequency.

3. Zero crossings [27]

This feature is used to measure the voice detection rate and counts the number of times that the sign of the signal amplitude changes in the time domain in one frame. For single-voiced signals, mostly the zero crossings are used to make a rough estimation of the fundamental frequency.

4. Centroid

The center of gravity of the spectrum is called centroid. It is calculated as the weighted mean of the frequencies present in the music signal.

$$C_r = \frac{\sum_{k=1}^{N/2} f[k] X_r^k}{\sum_{k=1}^{N/2} |X_r^K|} \quad (1)$$

Where, f [k] is the frequency at k. The centroid is the measure of higher and lower frequency in the spectra. Higher the centroid, more become the higher frequency and brighter the textures. Due to its effectiveness to describe spectral shape and centroid measures are used in audio classification tasks.

5. Roll off [28]

The frequencies in the spectra in which 85% magnitude distribution are determined. Similar the centroid, it measure of spectral shape and higher values for high frequencies. So there exists a strong correlation between both the features.

$$\sum_{K=1}^{M} X_r[K] = 0.85 \sum_{K=1}^{N/2} X_r[K] \quad (2)$$

Where, M is rolled off

6. Flux [29]

The squared difference between the normalized magnitude and that of the signal frame of the spectra is called as flux. The equation for flux is

$$F_r = \sum_{K=1}^{N/2} (|X_r[K]| - |X_{r-1}[K]|)^2 \quad (3)$$

Flux is an important feature for the separation of music from speech

C. Feature Extraction techniques

1. Mel Frequency Cepstral Coefficient (MFCC)

Mel-Frequency Cepstrum Coefficient feature is used broadly in acoustic, sound, and speech-related research areas due to its compatibility to represent MFC which becomes the short span of the spectrum of an audio frame. In MFCC used 13 cepstral coefficients to represents MFC, Hamming weighting window to apply before FFT, 40 Mel Filter Banks of 130 and 6854 Hz, 1KB block size and 512B step size [30].

MFCC is the used mostly in the audio recognition systems. It is calculated by the combination of the forty groups of coefficients. Then coefficients are scaled by the logarithmic scale and finally, DCT is applied to decorrelate.

2. Fast Fourier Transform (FFT)

For detecting pitches, the FFT and the STFT are the traditional feature extraction techniques in the frequency domain in signal analysis. However, the time-frequency resolutions are linear but human perception is logarithmic [31].

3. The Short-Time Fourier transform (STFT)

Fourier transform is an important mathematic tool for converting time dependent signal into frequency dependent signal. When Fourier transform is applied to the local sections of the signal, STFT plays an important role in feature extraction.The audio signal of music instruments are non-stationary signals, it means the spectrum of the signal changes with respect to time. The Time-Frequency representation of discrete STFT is given by Eq. (4)

$$X_{STFT}[m,n] = \Sigma_{L-1\, k=0}\, x[k] w[k-m]\, e_{-j2\pi nk/L} \quad (4)$$

Where, X[k] is the signal and w[k] is the window function to be applied to the signal. The STFT is given by the product of signal x[k] and windowing function w [k-m].

4. Spectral Shape Statistics

The spectral statistical analysis includes below attributes: [30]

$$\mu_i = \frac{\sum_{n=1}^{N} f_k^i * a_k}{\sum_{n=1}^{N} a_k} \quad (5)$$

$$\text{Centroid} = \mu_1 \quad (6)$$

$$\text{Spread} = \sqrt{\mu_2 - \mu_1^2} \quad (7)$$

$$\text{Skewness} = \frac{2\mu_1^3 - 3\mu_1\mu_2 + \mu_3}{S_w^3} \quad (8)$$

$$\text{Kurtosis} = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{S_4} - 3$$

(9)

5. Wavelet Transform (WT)

The wavelet feature extraction technique is as below

a) Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) is a special case of the WT which provides a compressed representation of a signal in time and frequency that can be computed efficiently. The DWT performs the fast analysis using multi-rate filter banks [10]. The multi-rate filter banks can be viewed as constant Q transform filter banks which having the octave spacing in between centers of the filters [19].

b) Wavelet Packet Decomposition (WPD) [31]

WPD is the called as the generalizing version of the DWT. Wavelet packet decomposition gives good time-frequency resolution hence it is used in a field of audio and speech processing. The difference between DWT and WPD is the DWT is applied to the LPF filter but WPD is applied to both LPF and HPF filter output.

c) Hybrid Algorithm DWPD [31]

Hybrid algorithm DWPD is the combination of the DWT and WPD algorithm. When the high-frequency components are removed from the signal, it retains the features of the signal and thus it reduces the noise but sometimes high-frequency component also contains the important information. This is the main drawback of DWT. To overcome the disadvantages of the previous method, the hybrid method is developed.

i. The audio signal is decomposed into the low and high-frequency band.
ii. DWT is applied over low-frequency component and WPD is applied over high-frequency signal.
iii. Low-frequency and high-frequency feature are combined to form the hybrid features

D. Classification techniques/ Machine Learning Algorithms

After the feature selection process it is important to classify the signal. Classification is the process by which a particular label is assigned to a particular audio format. It is this label that would define the signal and its origin. A classifier defines decision boundaries in the feature space, which separate different sample classes from each other.

1. Support Vector Machine (SVM)

SVM is widely used in the sound classification task as well as in AMT systems. In supervised learning, we provided the audio is the piano signal or not. For example, if the data are linearly separable, then there will be some hyperplanes available that can separate those data into two classes without error. From all those hyperplanes, choose the one that has the most maximum margin, hence we called it the maximum-margin hyperplane [6].

These mapping can be conducted by using the Kernel function. Kernel function was applied to transform input samples into a variable product. There are two kinds of Kernel function that mostly used for non-linear mapping: Polynomial and Gaussian Kernel [3].

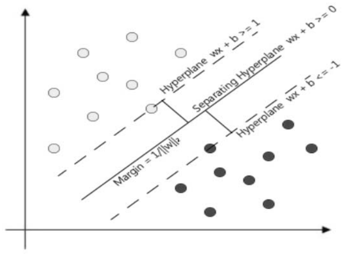

Fig. 6. Maximum Margin Support Vector Machine

2. k-nearest neighbor (KNN) classifier

KNN is a supervised classifier. The testing data is classified on the basis of the majority of k nearest neighbors. This may not be allowed incorrect placement of database but it is analogues to the human would process.

In KNN, Training set T is used to label the unlabeled testing data. First of all, the mean of the maximum value of training data and test data is calculated then distance is calculated between the nearest k samples closest to test data. The label of maximum nearest neighbor is assigned to the testing sample.

In Fig. 8, each training samples are marked with * and testing sample is marked with ●. The 5 k nearest neighbor is considering. The nearest 5 labels are shown in the circle. The number of samples having more number of labels will assign to the unlabeled B.

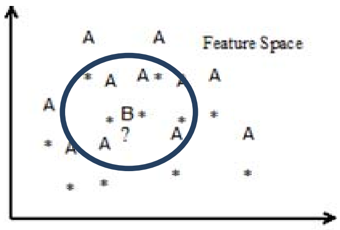

Fig. 7. An example of k nearest neighbor rule.

3. Hidden markov Model(HMM)

The Markov properties are those properties where the next state of the process depends on the present state. The system which uses Markov properties is called as Markov Process. When the Markov processes with hidden states are converted into statistical Markov Model, it is known as Hidden Markov Model..

4. Linear Predictive Coding (LPC) [32]

Linear Predictive Coding produces coefficients minimizing the difference between the actual speech samples and the linearly predicted ones. It is a very reliable method. Mostly Auto Regression model is used for speech.

5. PLCA [31]

Probabilistic Latent Component Analysis (PLCA) is dealing with an arbitrary number of dimensions and can exhibit various features such as sparsity or shift-invariance. It is defined as

$$P(x) = P(z)\prod_{j=1}^{N} P(x_j|z) \qquad (10)$$

Where, P(x) is the N dimensional distribution of the variable

$x = x_1, x_2, x_3 \ldots x_N$.

Z = latent variable

P (x_j | z) = 1D distribution

This model represents a mixture of marginal distribution products to approximate an N-dimensional distribution.

6. Dynamic Bayesian Network (DBN)

The Bayesian networks are an expert system that captures all existing knowledge is static which specify certain points in time. In term of speech and audio signal processing, the Bayesian need of the extended. It includes direct edge pointing in the direction of time. The Bayesian networks are represented by direct acyclic graphs.

7. RNN

A recurrent neural network (RNN) is an artificial neural network. In this network, the directed cycle is formed between the connections. RNN are the powerful temporal model. It captured the dependencies between the inputs. RNNs and their more complex variants have just been applied successfully to the problem of symbolic music prediction. Hence it creates the interest of researchers in the symbolic knowledge problem to improve AMT [4].

E. Performance measurement

The performance of the music transcription system is calculated using different performance parameters. To evaluate the performance some metrics have to be calculated as explained below for 'Sa' note.

TP (True Positive): Sa is present and correctly classified as Sa
TN (True Negative): Sa is not present and not detected Sa
FP (False Positive): Sa is present and not detected Sa
FN (False Negative): Sa is absent and detected as Sa

1. Sensitivity/ Recall

Sensitivity is the ratio of True Positive (TP) to the summation of TP and FN. It is also called as the true positive rate (TPR). Mathematically it is given by-

$$Sensitivity = \frac{TP}{TP+FN} \qquad (11)$$

2. Specificity

Specificity is the ratio of True Negative (TN) to the summation of TN and FN. It is also called as True Negative Rate (TNR). Mathematically it is given by,

$$Specificity = \frac{TN}{TN+FP} \qquad (12)$$

3. Accuracy

It is the ratio of correct assessment to the number of all negative assessment. Mathematically it is given by,

$$Accuracy = \frac{TP+TN}{TP+TN++FP+FN} \qquad (13)$$

V. Conclusion

This paper reviews and presents the different methods and features for automatic transcription system. Automatic music transcription system in necessary to automate the process of manual transcription as it may cause the mistake because of human error. AMT provides the fast, robust, reliable and accurate solution for the transcription of musical instrumental notes. The different databases are explained in the paper. The databases are made based on the musical instruments with different sampling rate. To represent the musical signal features plays the important role. In this paper, we explain various features and feature extraction techniques. For the small dataset with large variation in features, KNN classifier works well, for moderate features size, SVM classifier works well while large dataset deep neural network like RNN gives good results.

References

[1] Ashwini S. Deo, "The metrical organization of Classical Sanskrit verse", Journal of Linguistics, March 2007, pp. 1-58.

[2] Bethanie Hansen, David Whitehouse, Cathy Silverman, "Introduction to Music Appreciation", American Public University System *e*Press, revised edition 2014.

[3] A. Cogliati and Z. Duan, "Piano music transcription modeling note temporal evolution," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, 2015, pp. 429-433.

[4] S. Sigtia, E. Benetos and S. Dixon, "An End-to-End Neural Network for Polyphonic Piano Music Transcription," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, May 2016, pp. 927-939.

[5] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, Anssi Klapuri, "Automatic Music Transcription: Challenges and Future Directions", Journal of Intelligent Information Systems, Vol. 41, Issue 3, pp 407-434, December 2013.

[6] A. Klapuri and M. Davy, Signal processing methods for music transcription. Springer Science & Business Media, ISBN-10: 0-387-30667-6, 2007.

[7] T. Berg-Kirkpatrick, J. Andreas, and D. Klein, "Unsupervised transcription of piano music," in Advances in Neural information Processing Systems (NIPS), pp. 1538–1546,2014.

[8] A. P. Klapuri, "Multiple fundamental frequency estimations based on harmonicity and spectral smoothness," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pp. 804–816, 2003.

[9] P. Smaragdis and J. C. Brown, "Nonnegative matrix factorization for polyphonic music transcription," in 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. IEEE, 2003, pp.177-180.

[10] S. A. Abdallah and M. D. Plumbley, "Polyphonic music transcription by the non-negative sparse coding of power spectra," in Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR), 2004.pp.318-325.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015

[12] G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," EURASIP Journal on Applied Signal Processing, vol. 2007, no. 1, pp. 154–154, 2007.

[13] J. Nam, J. Ngiam, H. Lee, and M. Slaney, "A classification-based polyphonic a Piano transcription approach using learned feature representations." in Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), 2011.pp.175-180.

[14] S. B¨ock and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 121–124.IEEE,2012.

[15] Gustavo Reis, Member, IEEE, Francisco Fernandéz De Vega, Senior Member, Ieee, And Aníbal Ferreira, Member, IEEE. 2013"Automatic Transcription Of Polyphonic Piano Music Using Genetic Algorithms, Adaptive Spectral Envelope Modeling, And Dynamic Noise Level Estimation" IEEE Transactions On Audio, Speech, And Language Processing, 1558-7916, Vol. 20, No. 8, October 2012.

[16] Gowrishankar B. S. and Dr. Nagappa U Bhajantri "An Exhaustive Review of Automatic Music Transcription Techniques" International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016.

[17] E. Eide, J. R. Rohlicek, H. Gish and S. Mitter, "A linguistic feature representation of the speech waveform," 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 1993, pp. 483-486.

[18] Valentin Emiya, Nancy Bertin, Bertrand David, Roland Bade au. MAPS - A piano database for multi-pitch estimation and automatic transcription of music. [Research Report], 2010, pp.11.

[19] George Tzanetakis and Perry Cook, "Musical Genre Classification of Audio Signals", IEEE transactions on speech and audio processing, Vol. 10, No. 5, July 2002.

[20] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, Ryuichi Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database, 4th International Conference on Music Information Retrieval (ISMIR 2003), October 2003, pp.229-230.

[21] Chao-Ling Hsu and Jyh-Shing Roger Jang. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. IEEE Transactions on Audio, Speech, and Language Processing, February 2010, pp. 310–319.

[22] Olivier Gillet and Gaël Richard, *"ENST-Drums: an extensive audio-visual database for drum signals processing*, In Proc of ISMIR'06, Victoria, Canada, 2006.

[23] Gomez, E., Gouyon, F., and Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., Cano, P., and Wack, N., "ISMIR 2004 Audio Description Contest", Technical Report, Music Technology Group – Universitat Pompeu Fabra, 2006.

[24] V. S. Shelar, D. G. Bhalke, "Musical Instrument Transcription Using Neural Network", International Journal of Scientific Research Engineering &Technology (IJSRET).Vol. 1, Issue 2, May 2012, pp 011-015.

[25] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification", J. Audio Eng. Soc, Vol. 52, pp. 724-739, July/August 2004.

[26] Babu Kaji Baniya, Deepak Ghimire and Joonwhoan Lee, "Automatic Music Genre Classification Using Timbral Texture and Rhythmic Content Features", ICACT Transactions on Advanced Communications Technology (TACT) Vol. 3, Issue 3, May 2014, pp 434-443.

[27] Park S. K., Kil R.M., Jung Y. G., Han MS. (2007) Zero-Crossing-Based Feature Extraction for Voice Command Systems Using Neck-Microphones. In: Liu D., Fei S., Hou ZG., Zhang H., Sun C. (eds) Advances in Neural Networks – ISNN 2007. ISNN 2007. Lecture Notes in Computer Science, vol 4491. Springer, Berlin, Heidelberg

[28] Marko Kos, Zdravko Kacic, Damjan Vlaj, "Acoustic classification and segmentation using modified spectral roll-off and variance-based features", Digital Signal Processing, Vol. 23, Issue 2, March 2013, Pages 659-674

[29] Yuxin Peng, Chong-Wah Ngo, Cuihua Fang, Xiaoou Chen, and Jianguo Xiao, "Audio Similarity Measure by Graph modeling and Matching" MM'06, October 23–27, 2006, pp603-606.

[30] Sang Hyun Park," Musical Instrument Extraction through Timbre Classification", NVIDIA Corporation Santa Clara, CA 95050.

[31] Rekha Hibare, Anup Vibhute, "Feature Extraction Techniques in Speech Processing: A Survey", International Journal of Computer Applications (0975 – 8887) Volume 107 – No 5, December 2014, pp.1-8.

[32] Shweta Tripathy, Neha Baranwal, G.C.Nandi, "An MFCC based Hindi Speech Recognition Technique using HTK Toolkit", IEEE Second International Conference on Image Information Processing (ICIIP-2013), pp. 539-544.