Scientific Journal Impact Factor (SJIF): 1.711
International Journal of Modern Trends in Engineering and Research
IJMTER
ISSN (ONLINE) : 2349 - 9745
ISSN (PRINT) : 2393 - 8161

Speech/Music Differentiation and Male/Female Voice Diagnosis in Speech : A Hierarchical Approach

Arijit Ghosal[1], Suchibrota Dutta[2], Debanjan Banerjee[3]
[1]*Department of Computer Sc. & Engg., Neotia Inst. of Tech. Mgmt. And Sc.,*
[2]*Department of Computer Sc.& Application, Maharaja Manindra Chandra College,*
[3]*Department of Computer Sc.& Engg, Manipal University,*

Abstract— In this work, we have presented a simple hierarchical scheme for differentiating audio signals first into speech and music and further speech into male speech and female speech. In the first level, audio signal is classified into two sub-classes speech and music (music includes both instrumental and song) based on some popular salient low level time-domain acoustic features which are very closely related to the physical properties of source audio signal. Zero crossing rate (ZCR), short term energy (STE) and “delta-energy”, which is the very small change of energy of an audio signal, are used as feature in the first level of classification. The strength of the feature-set is further enhanced by incorporating features computed from the co-occurrence matrix of delta-energy. For the second phase of classification some popular low level frequency-domain acoustic features are used as male voice differs with female voice mostly in frequency domain. Along with low level frequency-domain features, low-level time-domain features are also used in male/female voice diagnosis in speech. We have used low level frequency-domain features based on spectral flux and spectral centroid and low level time-domain features based on ZCR and STE for this purpose. RANSAC and Neural-Net has been used at all stages as classifier. The experimental result denotes the efficiency of the proposed scheme.

Keywords— Speech/music differentiation, audio features, Delta-Energy plot, Male/female voice diagnosis, Spectral flux plot, Plot of co-occurrence matrix of delta-energy, RANSAC.

I. INTRODUCTION

With the huge growth of multimedia technology it is becoming quite necessary to create an audio library comprising of huge audio data. But managing of that library is becoming tougher and tougher day by day as these audio data needs to be classified properly into different categories such as speech, music etc for storing them in the library. Manual classification is just an impossible task due to the huge size of the audio data. So an automatic classification is obvious and it is becoming an active area of research. A lot of work has been done for the development of content-based image and video retrieval system. Comparatively, little work has been done on the audio domain [1].

An automatic audio classification system has two steps: *extraction of audio features from the input audio signal* and *classification of audio signal based on the extracted feature*. But, it is not easy to extract the features which represent the characteristics. It is observed that lot of efforts [2], [3], [4], [5], [6] have been made so far to classify audio signal.

Varieties of features for classification of audio data have been proposed by the researchers which may be categorized as low level features, perceptual/psychoacoustic features etc. Low level feature includes several time domain and frequency domain features. ZCR (zero crossing rate) [7], [8] and STE (short time energy) [9], [10] are the most widely used time domain features. Frequency domain approaches include features like signal bandwidth, spectral centroid, signal energy [11], [12], [13], [14], fundamental frequency [1], mel-frequency cepstral co-efficients (MFCC) [15], [16] etc. Perceptual/psychoacoustic features include measures for roughness [17], loudness [17], etc. In [18], a model representing the temporal envelope processing by the human auditory system has been proposed which yields 62 features describing the auditory filter bank temporal envelope (AFTE). Liu et al. [19] and Guo et al. [20] have dealt with sub band energy to describe audio data.

Audio data can be classified into different categories like speech, music, songs, or noise based on the feature vector describing the audio data. Lots of classification schemes having different complexity have been used for this purpose. El-Maleh [10] has proposed a two level speech-music classifier. A threshold based two level algorithm has been presented by J. Saunders [9]. Neural network based scheme has been also tried by Matiyaho and Furst [21]. Support Vector Machine [22], [23], [24] has also been used by many researchers [20], [25] for audio classification. Various classification schemes have deployed self organizing maps, k-nearest neighbor method, multivariate Gaussian models [4]. Hidden Markov model [26] has also been tried by Kimber and Wilcox.

In the context of audio classification system, at the first level it is necessary to classify an audio signal into speech and music. Music includes both instrumental that is music without voice and songs that is music with voice. This classification bears a significant importance as once an audio signal is classified into speech and music, the speech signal can further be classified into male and female categories and music signal can be categorized into instrument and song. We all know that automatically detecting the gender of a speaker has several important applications. These applications requires pure speech signal as input. Though song signal includes speech component within it but as it is mixed with instrumental component, it is very practical to go for a hierarchical approach where in the first level speech and music are discriminated and in the second level speech is classified into male and female speech.

One of the main application areas for speech recognition is voice input to computers for tasks like document creation (word processing) and financial transaction processing (telephone-banking). Voice recognition, by computer, is also used in access control and security systems. The need of gender classification of speech signal also arises in several situations such as sorting telephone calls by gender (eg. gender sensitive surveys). In content based multimedia indexing, the speaker's gender is a cue used in the annotation. Also, gender dependent speech codes are more accurate than gender independent ones (Marston, 1998; Potamitis et al., 2002). Therefore, automatic gender detection can be an important tool in multimedia signal analysis systems.

Speech processing is the study of speech signals and the various methods which are used to process them. In this process various applications such as speech coding, speech synthesis, speech recognition and speaker recognition technologies; speech processing is employed [32]. The main purpose of speech identification is to convert the acoustic signal obtained from a microphone or a telephone to generate a set of words [33, 34]. Rao *et al.* [35] have proposed that the different time-varying glottal excitation components of speech can be used for text independent gender recognition studies. They represented the excitation information in speech by using a linear prediction (LP) residual. They have used a Hidden Markov Models (HMMs) to capture the gender-specific information in the excitation of different voiced speech. A wavelet transform-based feature for the speech/music differentiation is proposed by T. Düzenli and N. Özkurt [36]. Bhandari *et al.* [41] have proposed an algorithm to segment audio and extract some features such as MFCC, Spectral Flux, SNR and ZCR.

In this work, we have focused on the efficiency of features. We have concentrated on the two most widely used time domain low level features, ZCR and STE. We have calculated some features which are based on ZCR and STE along with features based on delta-energy for speech/music differentiation. In most of the previous research works it is found that the male/female voice diagnosis is performed by considering *pitch* as feature. It is observed that the pitch value of female voice is different than that of male voice. This is the reason for using *pitch* as a feature by most of the researchers. In the second phase, for male/female voice diagnosis in speech we have relied on spectral flux and spectral centroid which are frequency domain low level features along with features based on ZCR & STE. Our motivation is to design a powerful set of features which will make the task of a classifier much easier.

The paper is organized as follows. Introduction is followed by the proposed methodology, where we have described the design of the features and classification scheme. Section 3 presents the experimental results and the concluding remarks are put into section 4.

II. PROPOSED METHODOLOGY

In this work, we have first discriminated speech and music. Computation of features and classification schemes are two major modules of the present work.

Computation of Features

In this work we have followed a hierarchical approach for the classification of audio signal into different categories. We have concentrated on such features which are capable to capture the suitable characteristics of the signal at different stages.

Features for Speech/Music Differentiation

It has been indicated in the past work that zero crossing rate (ZCR) and short time energy (STE) are two most important time domain, low level audio features which play major role in speech/music differentiation. This has motivated us to concentrate on these two time domain low level features. In the proposed methodology we have considered mean and standard deviation of ZCR and STE both as the discriminating audio features. We know that for speech signal major energy is confined in the lower frequency band in comparison to a music signal. Energy distribution of a music signal spreads over a wide a frequency range and considerable amount is confined in higher frequency bands. But in a speech signal, frequent occurrence of silence also forms a characteristic mark.

Considering audio data as discrete signals, it is said that a zero crossing has occurred whenever two successive samples have different signs. Rate of zero crossing provides an impression regarding the frequency content. Audio signal is divided into N frames $\{x_i(m): 1 \leq i \leq N\}$. Then, for i^{th} frame, zero crossing rates are computed as follows:

$$z_i = \sum_{m=1}^{n-1} \mathrm{sign}[x_i(m-1) * x_i(m)] \qquad (1)$$

where, n is the number of samples in the i^{th} frame and

$$\mathrm{sign}[v] = \begin{cases} 1, \text{ if } v > 0 \\ 0, \text{ otherwise} \end{cases} \qquad (2)$$

The short term energy (STE) for i^{th} is defined as:

$$E_i = 1/n * \sum_{m=0}^{n-1} [x_i(m)]^2 \qquad (3)$$

The audio signal is broken into frames of specific size. The frames are overlapped to avoid the missing of any characteristics of a particular frame. For each frame ZCR and STE is computed. Finally, mean and standard deviation of ZCR and STE over all the frames are considered.

As energy variation of speech signal is completely different from that of music signal, so, change of energy variation of a speech signal between two consecutive frames will obviously differ from music signal. We are denoting the change of energy variation for two consecutive frames as "delta-energy'. For speech signal it will show some peaks only when there will not be any silence, otherwise it will come down near to zero because of presence of silence. But for music signal such kind of pattern will not be observed due to absence of silence.

Let, for i^{th} frame, energy is calculated as E_i, for $i+1^{th}$ frame energy is calculated as E_{i+1}. Then delta-energy is calculated as:

$$\Delta En = E_{i+1} - E_i \qquad (4)$$

We have observed that mean and standard deviation of a feature gives only an overall idea about the distribution. So to precisely study the characteristics of delta-energy, we have we have utilized the concept of co-occurrence matrix [28]. For an image, the occurrence of the different intensity values within a neighborhood reflects a pattern and it is utilized to parameterize the appearance/texture of an image. We have used the same concept here. Delta-energy is computed using equation (4). Thus, $\{\Delta EN_i\}$, a sequence of delta-energy is obtained for the signal. Occurrence of different delta-energy values within a neighborhood reflects the pattern and characterizes the behavior of the signal. Thus, a matrix, C of $L \times L$ dimension (where, $L = max\{\Delta EN_i\} + 1$) is formed. In our case, an element in the matrix C(i, j) stands for number of occurrences of delta-energy i and j in consecutive time instances. Finally, co-occurrence matrix based statistical measures [31] namely energy, entropy, homogeneity, contrast, correlation are computed which represent the pattern of the co-efficients.

Calculating all the features, we have now developed a 9- dimensional feature vector (mean and standard deviation of ZCR and STE both, energy, entropy, homogeneity, contrast, correlation) which will act as the feature vector for an audio signal.

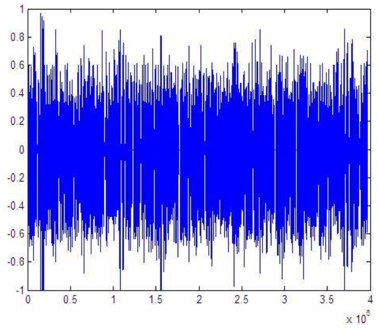

Fig 1. Original speech signal

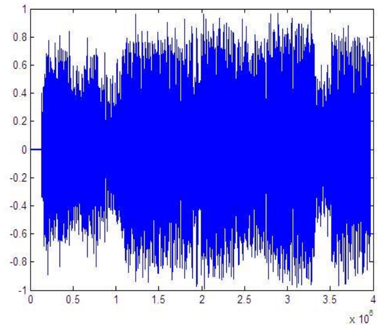

Fig 2. Original music signal

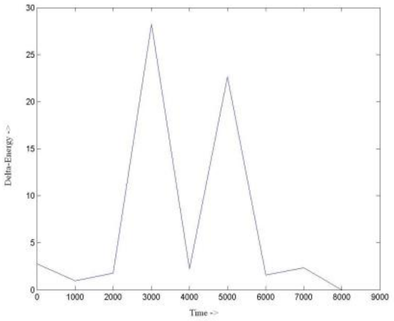

Fig 3. Delta-energy plot of speech signal

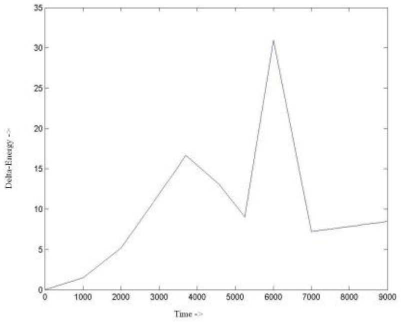

Fig 4. Delta-energy plot of music signal

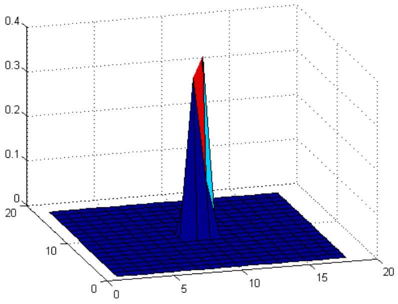

Fig 5. Plot of co-occurrence matrix of delta-energy for speech signal

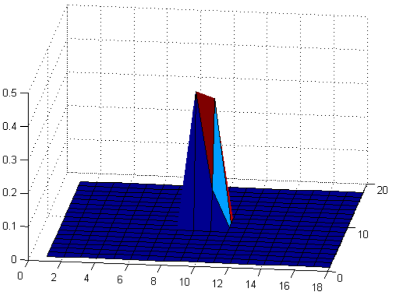

Fig 6. Plot of co-occurrence matrix of delta-energy for music signal

Features for Male/Female Voice Diagnosis in Speech

The speech signal is recorded in a variety of acoustic conditions like indoor, outdoor, and telephone speech etc. Therefore, a gender identifier for multimedia applications have to be robust to channel and acoustic condition changes.

The voice consists of sound made by a human being using the vocal folds. The human voice is specifically that part of human sound production in which the vocal folds or vocal cords are the primary sound source. The vocal folds or vocal cords are a vibrating valve that chops up the airflow from the lungs into audible pulses that form the laryngeal sound source. Male and female human beings have different sizes of vocal fold; thus reflecting the male-female differences in larynx size. The difference in vocal folds size between male and female means that they have differently pitched voices. Male voices are usually lower-pitched and have larger folds compared to female voice. The sound of each individual's voice is unique not only because of the actual shape and size of an individual's vocal cords but also due to the size and shape of the rest of that person's body, especially the vocal tract, and the manner in which the speech sounds are habitually formed and articulated. Humans have vocal folds that can loosen, tighten, or change their thickness, and over which breath can be transferred at varying pressures. The shape of chest and neck, the position of the tongue, and the tightness of otherwise unrelated muscles can be altered. Any one of these actions results in a change in pitch, volume, timbre, or tone of the sound produced. Though voice of each person is unique but still all male voices produce certain speech patterns and all female voices also produce certain speech patterns which are different from the pattern produced by male voices. Though we know that the ZCR value for male voice and female voice is different but we have also observed that in frequency domain male voice and female voice is completely different. So it is obvious that only ZCR will not be enough to discriminate male voice with female voice compared to what low level frequency domain features can. It is noticed that there is phonetic differences between male and female voices also. This difference cannot be measured by low level time domain features rather it can be measured by low level frequency domain features. These have motivated us to concentrate on low level frequency domain features based on spectral flux and spectral centroid.

Spectral flux is also called as *Spectral Variation*. Spectral flux of an audio signal measures how quickly the power spectrum of that signal changes. Spectral flux is defined as the variation value of spectrum between the adjacent two frames in a short-time analyze window. Spectral flux can be used to determine the timbre of an audio signal. Spectral flux is calculated as:

$$SF(k) = \sum_{i=0}^{n-1} s(k,i) - s(k-1,i) \qquad (5)$$

SF(k) is the spectral flux of the k^{th} spectrum. *s(k,i)* is the value of the i^{th} bin in the k^{th} spectrum, *s(k-1,i)* is analogous for the spectrum before *k*. We subtract the values of each bin of the previous spectrum from the values of the corresponding bin in the current spectrum and sum those differences up to arrive at a final value which is the spectral flux of spectrum *k*.

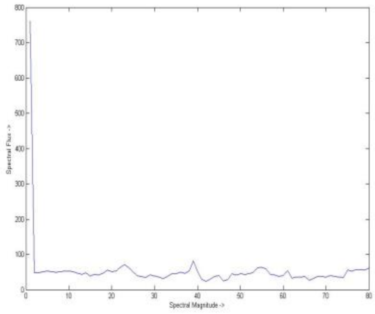

Fig 7. Spectral flux plot of male voice

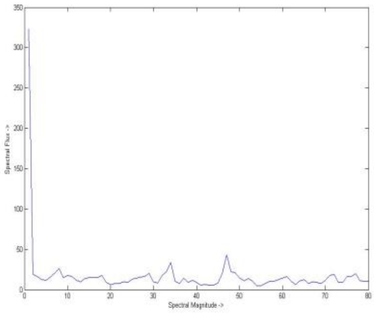

Fig 8. Spectral flux plot of female voice

We know that spectral flux variation of male voice is different compared to that of female voice. In case of male voice, initial value of spectral flux is much more than that of female voice. There is a sharp fall of spectral flux with the increment of spectral magnitude from zero for both male voice and female voice. In case of male voice spectral flux does not vary much that is it never comes towards zero value but for female voice, spectral flux varies much more than that of male voice (fig 7 and fig 8) that is it comes near to zero many times. So for spectral flux plot of female voice much more peaks can be observed compared to spectral flux plot of male voice. We have considered both mean and standard deviation of spectral flux.

Spectral centroid is a measure used to characterize a spectrum. It denotes the position where the “center of mass” of the spectrum is situated. It is calculated as the weighted mean of the frequencies present in the input signal, with their magnitudes as the weights.

We know that, the ZCR and STE value for male voice and female voice is different. This observation has motivated us to consider features based on ZCR & STE also. We have considered mean and standard deviation of ZCR and STE both. Finally we have considered mean and standard deviation of ZCR, STE and spectral flux along with spectral centroid as our 7-dimmensional feature-set.

Classification

For classification we had keep in mind that the main objective of this work is to judge the capacity of the proposed feature for discriminating speech and music and male/female voice diagnosis. SVM based classifier is robust but the crucial task of it is the parameter tuning for optimal performance. For that reason we have used two simple classifiers: RANdom Sample And Consensus (RANSAC) [29] and Neural-Net.

RANSAC is a re-sampling technique that generates solutions by using the minimum number observations (data points) required to estimate the underlying model parameters. Unlike conventional sampling techniques that use as much of the data as possible to obtain an initial solution and then proceed to prune outliers, RANSAC uses the smallest set possible and proceeds to enlarge this set with consistent data points [29]. The major strength of RANSAC over other estimators is that the estimation is made based on inlier that is whose distribution can be explained by a set of model parameters.

Initially the minimum number of points required to determine the model parameters are selected randomly. Then the parameters of the model are solved. Next, how many points from the set of all points fit with a predefined tolerance € are determined. Now, If the fraction of the number of inliers over the total number points in the set exceeds a predefined threshold €, re-estimation of the model parameters using all the identified inliers are done and then it is terminated. If not, these steps are repeated maximum of N times, where N, is chosen high enough to ensure that the probability p (usually set to 0.99) that at least one of the sets of random samples does not include an outlier. RANSAC estimates the model relying on the inliers, unlike other technique; it is less affected by the noisy data. So RANSAC is suitable for our purpose.

An artificial neural network (ANN), usually called neural-net (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. An ANN is typically defined by three types of parameters:

The interconnection pattern between different layers of neurons

The learning process for updating the weights of the interconnections

The activation function that converts a neuron's weighted input to its output activation.

The input of the network is the feature vectors. The dimensions of the feature vectors is N (N = no of features), corresponding to the N parameters of features extracted from the audio content. The network contains layer of hidden neurons that are not part of the input or output of the network. These hidden neurons enable the network to learn complex tasks by extracting progressively more meaningful features from the input vectors. The number of neurons in each hidden layer can be specified in the experiments. The number of the neurons in output layer is determined by the number of audio classes we want to classify. For example, if we want to classify the audio into music and speech, we have two neurons in output layer corresponding to the music and speech respectively.

III. EXPERIMENTAL RESULT

In order to carry out the experiment, a database is prepared consisting of 200 speech files and 200 music files. All are of around 40-45 seconds duration. File are obtained from CD recording, recording of live programs and downloaded from various sites in Internet. Sampling frequency is 22050 Hz, 16-bit per sample and of type mono. For speech both the voice of male and female are considered, different languages are also present. Some are noisy also. Music files consists of wide variety of instruments like flute, piano, guitar, drum etc as well as different types of songs like bhangra (an Indian genre), classical, jazz, rock etc.

To compute the features, an audio file is divided into frames. Each frame consists of 150 samples of which there is an overlap of 50 samples between two consecutive frames. In our experiment, we have relied on MATLAB for implementation of Neural-Net. For Neural-Net, we have considered nine input nodes and two output nodes with five hidden nodes in the one and only hidden layer for first phase of classification and for second phase of classification we have considered seven input nodes and two output nodes. There is only one hidden layer with four hidden nodes for second phase of classification.

TABLE I: ACCURACY OF SPEECH/MUSIC DIFFERENTIATION

Classifier	Classification Accuracy (in %)					
	Proposed Features			Features Based on Mean and Std. Dev. of ZCR & STE		
	Speech	Music	Overall	Speech	Music	Overall
Neural Network	91%	93%	92%	86.5%	90.5%	88.5%
RANSAC	97.5%	93.5%	95.5%	95%	93%	94%

As mean and standard deviation of ZCR and STE are traditionally used audio features for speech/music differentiation we have compared the strength of proposed feature for speech/music differentiation with that also. Table I shows the performance of the proposed features for speech/music differentiation.

TABLE II: ACCURACY OF MALE/FEMALE VOICE DIAGNOSIS IN SPEECH

Classifier	Classification Accuracy (in %)		
	Proposed Features		
	Male Speech	Female Speech	Overall
Neural Network	90.5%	88.5%	89.5%
RANSAC	95.5%	93.5%	94.5%

Table II shows the performance of the proposed features for male/female voice diagnosis in speech.
It is evident that the proposed features perform well in discriminating the speech and music signals as well as male/female voice diagnosis in speech. It has been also found that RANSAC performs better for the data set with considerable variation.
In our experiment, we have used 50% of each type of data as training set for determining the model and rest of the data have been used for testing. Once again, the experiment is done reversing the training and test data set. Average testing accuracy is considered and is shown in Table I and Table II.

IV. CONCLUSION

In this work, we have presented a hierarchical scheme to classify the audio signals in various categories. In the first phase that is speech/music differentiation, we have proposed new set of low level time-domain features based on Mean of ZCR, Standard deviation of ZCR, Mean of STE, and Standard deviation of STE, and energy, entropy, homogeneity, contrast, correlation computed from the co-occurrence matrix of delta-energy. Experimental result indicates the potentiality of such features for speech/music differentiation. The classification performance of the proposed features is much better than that of traditional features based only on ZCR and STE. In the second phase that is male/female voice diagnosis in speech, we have proposed a new set of feature-set comprising of low level frequency-domain features based on spectral centroid along with mean and standard deviation of spectral flux and low-level time-domain features based on mean and standard deviation of both ZCR and STE. Experimental result denotes the potentiality of such features for diagnosis of male/female voice in speech signal. In this work, to highlight the strength of the proposed features, simple classification scheme based on Neural-Net and RANSAC has been adopted. Comparing the results generated by these classifiers, it is found that RANSAC gives better result compared to Neural-Net. In future, further sub-classification of the data in the individual class may be carried out.

REFERENCES

[1] T. Zhang, C. C. J. Kuo, Content-based classification and retrieval of audio. SPIE's International Symposium on Optical Science, Engineering, and Instrumentation. International Society for Optics and Photonics, 1998.

[2] S. Davis, P. Mermelstein, Comparison of parametric representations monosyllabic word recognition in continously spoken sentences. IEEE Transactions on acoustics, speech and signal processing, vol. 28, pp. 357–366, 1980.

[3] E. Wold, T. Blum, D. Keislar, J. Wheaton, Content-based classification, search, and retrieval of audio. IEEE Transactions on Multimedia, vol. 28, pp. 27–36, 1996.

[4] E. Scheirer, M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1997, pp. 1331–1334.

[5] H. Wang, A. Divakaran, A. Vetro, S. F. Chang, H. Sun, Survey on compressed-domain features used in video/audio indexing and analysis. Technical report, Department of electrical engineering, Columbia University, New York, 2000.

[6] Y. Wang, Z. Liu, J.C. Huang, Multimedia content analysis using both audion and visual cues. IEEE signal processing magazine, vol. 17, pp. 12–36, 2000.

[7] C. West, S. Cox, Features and classifiers for the automatic classification of musical audio signals. International Conference on Music Information Retrieval, 2004, pp. 531–537.

[8] J. Downie, The scientific evaluation of music information retrieval systems: Foundations and future. Computer Music Journal, vol. 28, no. 2, pp. 12–33, 2004.

[9] J. Saunders, Real-time discrimination of broadcast speech/music. IEEE International. Conference on Acoustics, Speech, Signal Processing, 1996, pp. 993–996.

[10] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, Speech/music discriminator for multimedia application. IEEE International Conference on Acoustics, Speech and Signal Processing, 2000.

[11] H. Beigi, S. Maes, J. Sorensen, U. Chaudhari, A hierarchical approach to large-scale speaker recognition. Proceeding of the International Computer Music Conference, 1999.

[12] M.A. Cohen, S. Grossberg, L.L. Wyse, A spectral network model of pitch perception. Journal of the Acoustical Society of America, vol. 498, pp. 862–879, 1995.

[13] C. Mckay, I. Fujinaga, Automatic genre classification using large high-level musical feature sets. In the Proceeding of International Conference MIR, 2004.

[14] C. West, S. Cox, Finding an optimal segmentation for audio genre classification. International Symposium on Music Information Retrieval, 2005.

[15] A. Eronen, A. Klapuri, Musical instrument recognition using ceptral coefficients and temporal features. IEEE International Conference on Acoustics, Speech and Signal Processing, 2000, pp. 753–756.

[16] J.T. Foote, Content-based retrieval of music and audio. SPIE, 1997, pp. 138–147.

[17] E. Zwicker, H. Fastl, Psichoacoustics: Facts and models. Springer Series on Information Science, 1999.

[18] J. Breebaart, M. McKinney, Feature for audio classification. International Conference on MIR, 2003.

[19] Z. Liu, J.H.A. Wang, T. Chen, Audio feature extraction and analysis for scene classification. IEEE Workshop on Multimedia Signal Processing, 1997.

[20] G. Guo, S. Z. Li, Content-based audio classification and retrieval by support vector machines. IEEE Transactions on Neural Networks, vol. 14, no. 1, pp. 209–215, 2003.

[21] B. Matityaho, M. Furst, Classification of music type by a multilayer neural network. Journal of the Acoustical Society of America, vol. 95, 1994.

[22] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, Support vector clustering. Journal of Machine Learning Research, vol. 2, pp. 125–137, 2001.

[23] B. M.J. Tax, R.T.W. Duin, Support vector domain description. Pattern Recongintion Letters, vol. 20, pp. 1191–1199, 1999.

[24] F. Camastra, A. Verri, A novel kernel method for clustering. IEEE Transactions on PAMI, vol. 27, pp. 801–805, 2005.

[25] S.O. Sadjadi, S.M. Ahadi, O. Hazrati, Unsupervised speech/music classification using one-class support vector machines. In the Proceeding of the ICICS, 2007.

[26] D. Kimber, L. Wilcox, Acoustic segmentation for audio browsers. Computing Science and Statistics, pp. 295-304, 1997.

[27] A. Ghosal, R. Chakraborty, R. Chakraborty, S. Haty, B. C. Dhara, S. K. Saha, Speech/music classification using occurrence pattern of zcr and ste. In 3rd International Symposium on Intelligent Information Technology Application, pp 435–438, China, 2009. IEEE CS Press.

[28] S. E. Umbaugh, Computer Imaging: Digital Image Analysis and Processing. CRC Press, 2005.

[29] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model for model fitting with applications to image analysis and automated cartography. ACM Communications, vol. 24, pp. 381–395, 1981.

[30] M. Zuliani, C. S. Kenney, B.S. Manjunath, The multiransac algorithm and its application to detect planar homographies. IEEE Conf. on Image Processing, 2005.

[31] R. M. Haralick, L.G. Shapiro, Computer and Robot Vision (Vol-I). Addision-Wesley, 1992.

[32] F. Zanuy, S. McLaughlin, A. Esposito, A. Hussain, J. Schoentgen, G. Kubin, W. Kleijn, P. Maragos, Non-linear Speech Processing: Overview and Applications, Control & Intelligent Systems. ACTA Press, vol.30, no.1, pp. 1-10, 2002.

[33] A. Othman, M. Riadh, Speech Recognition Using Scaly Neural Networks. World Academy of Science, Engineering and Technology, vol. 38, pp. 253-258, 2008.

[34] G. Singh, A. Junghare, P. Chokhani, Multi Utility E-Controlled cum Voice Operated Farm Vehicle. International Journal of Computer Applications, vol. 1, no. 13, pp. 109-113, 2010.

[35] R. Rao, A. Prasad, Glottal Excitation Feature based Gender Identification System using Ergodic HMM. International Journal of Computer Applications, vol. 17, no.3, pp-31-36, March 2011.

[36] T. Düzenli, N. Özkurt, Discrete and Dual Tree Wavelet Features for Real-Time Speech/Music Discrimination. International Scholarly Research Notices, 2011.

[37] A.P. Simpson, Phonetic differences between male and female speech. Language and Linguistics Compass, Volume 3, Issue 2, pp. 621–640, March 2009.

[38] Y. Zengi, Z. Wu, T. Falk, W. Chan, Robust GMM based Gender Classification using Pitch and Rasta-PLP Parameters of Speech. in proceeding of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp. 13-16 Aug 2006.

[39] M. Sigmund, Gender Distinction Using Short Segments of Speech Signal. International Journal of Computer Science and Network Security, vol.8, no.10, pp. 159-162, Oct 2008.

[40] J. Silovsky, J. Nouza, Speech, Speaker and Speaker's Gender Identification in Automatically Processed Broadcast Stream. Radio Engineering Journal, vol.15, no.3, pp. 42-48, Sep 2006.

[41] G. M. Bhandari, R.S. Kawitkar, M.C. Borawake, Audio Segmentation for Speech Recognition Using Segment Features. International Journal of Computer Technology & Applications, vol. 4, no. 2, pp. 182-186, 2013.

[42] Y. Shue, M. Iseli, The role of voice source measures on automatic gender classification. in proceeding of IEEE International Conference on acoustics, Speech and Signal Processing, Las Vegas, pp. 4493-4496, 2008.

[43] A. Ghosal, S. Dutta, Automatic male-female voice discrimination. International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014. IEEE, 2014.

[44] M. Sedaaghi, A Comparative Study of Gender and Age Classification in Speech Signals. Iranian Journal of Electrical & Electronic Engineering, vol. 5, no. 1, pp. 1-12, March 2009.

[45] J. Rodger, P. Pendharkar, A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. International Journal of Human-Computer Studies, vol.60, pp. 529–544, 2004.

[46] K. Rakesh, S. Dutta, K. Shama, Gender Recognition using speech processing techniques in LABVIEW. International Journal of Advances in Engineering & Technology, vol. 1, no. 2, pp. 51-63, May 2011.

[47] S. Dutta, A. Ghosal, S. K. Saha, Speech/Music Classification Using Delta-Energy and RANSAC. International Conference on Computer Applications (ICCA), ASDF, 2012.

[48] M. A. Haque, J.M. Kim, An analysis of content-based classification of audio signals using a fuzzy c-means algorithm. Multimedia tools and applications vol. 63, no. 1, pp. 77-92, 2013.