Co-occurrence Based Approach for Differentiation of Speech and Song

Arijit Ghosal*, Ranjit Ghoshal

Department of Information Technology, St. Thomas' College of Engineering and Technology, Kolkata, West Bengal, India

*Corresponding author
doi: https://doi.org/10.21467/proceedings.115.17

ABSTRACT

Discrimination of speech and song through auditory signal is an exciting topic of research. Preceding efforts were mainly discrimination of speech and non-speech but moderately fewer efforts were carried out to discriminate speech and song. Discrimination of speech and song is one of the noteworthy fragments of automatic sorting of audio signal because this is considered to be the fundamental step of hierarchical approach towards genre identification, audio archive generation. The previous efforts which were carried out to discriminate speech and song, have involved frequency domain and perceptual domain aural features. This work aims to propose an acoustic feature which is small dimensional as well as easy to compute. It is observed that energy level of speech signal and song signal differs largely due to absence of instrumental part as a background in case of speech signal. Short Time Energy (STE) is the best acoustic feature which can echo this scenario. For precise study of energy variation co-occurrence matrix of STE is generated and statistical features are extracted from it. For classification resolution, some well-known supervised classifiers have been engaged in this effort. Performance of proposed feature set has been compared with other efforts to mark the supremacy of the feature set.

Keywords: Speech song classification, acoustic features, STE, Co-occurrence matrix

1 Introduction

If a researcher wants to identify different types of genre of songs from general audio signals, the researcher may approach towards a hierarchical scheme for that purpose. General audio signal means audio signal which is comprised of different categories of audio signal like environment sounds, speech signals etc. The task of discrimination of speech signal and song signal finds its importance in the identification task of different genres being a fundamental step of the hierarchical approach for that purpose.

Moreover, size of multimedia resources is up surging day by day. Audio data is actually part of multimedia resource. To retrieve these audio data in future for different purposes including research purpose also they need to be stored proper way maintaining there categories properly. The task of categorization requires implementation of machine learning. To apply the concept of machine learning certain features which is specific to audio they need to be applied. These features are called as audio features. If these audio features are properly extracted entire audio data can be properly categorized. Speech and song are the two rudimentary categories of acoustic data. Hence, categorization of audio data may be started from the task of differentiation of speech and song. Differentiation of speech and song is very much exciting field of research due to its application in media services, search engines, and smart human-computer systems. Distinction of speech and song can also be recognized as an elementary phase towards implementation of automatic speech recognition scheme.

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

Continuous research in the domain of signal processing and data mining technologies has led improvement of research efforts in the range of audio data retrieval. Though speech and song discrimination task are very auspicious research area still surprisingly very less effort has been carried out in this area. Whatever the little bit efforts were there in this area they are mostly depended on frequency and perceptual domain acoustic features like rhythm, pitch, spectral flux and spectral roll-off etc. This effort proposes a simple statistical feature set to discriminate speech and song. The statistical features are computed from the co-occurrence matrix of STE.
This work is shaped as - portrayal of earlier works completed is discussed in section II. Approach of solution scheme is enlightened in section III. Section IV defines the investigational outcome with a comparative discussion followed by the conclusion section.

2 Related Works

Statistical feature extraction method was discussed by Haralick [1]. This thought is very much helpful in various fields of machine learning or pattern recognition. Gerhard [2] has exercised pitch-based features for speech and song classification. Bugatti et. al [3] has made a comparative analysis between statistical and neural approach in order to classify speech and music. Gerhard [4], in his other approach for speech song discrimination has applied perceptual features. This approach is fuzzy approach. Gerhard [5] has analysed the influence of silence and rhythm for differentiating speech and song. Tzanetakis [6] has noted the character of song specific singing accent. He has conducted a classification decision in every 20 milliseconds interval. Lin and Chen [7] have recommended a real time categorization plan to categorize audio signals into a number of kinds. A method for generation of co-occurrence matrix for a particular feature has been discussed by Umbaugh [8]. The concept of extracting features from co-occurrence matrix is extremely useful in various fields of pattern recognition. Concept of harmonic structure modelling in order to separate music signals into different categories has been applied by Zhang and Zhang [9]. Ruinskiy and Lavner [10] have recommended a way for recognition of sounds of breath both in speech and song signals.
For classification as well as segmentation of music and speech, Lavner and Ruinskiy [11] have advised a decision tree dependent algorithm. Gallardo-Antolín and Montero [12] have catalogued speech, song as well as music by employing histogram equalization dependent features. Salselas and Herrera [13] have also contributed in the research work of discrimination of speech and song. They have noticed the rhythmic and melodic outline for both speech and song. Spectral features for recognition of speech in case of varied aural signal has been applied by Sonnleitner et. al [14]. Bhavsar and Panchal [15] have conducted a review work on the applicability of Support Vector Machine (SVM) to categorize audio data. Velayatipour and Mosleh [16] have analysed diverse procedures for distinction of speech and music. Ramalingam and Dhanalakshmi [17] have paid an important contribution in this research field by proposing wavelet-based feature to discriminate speech and music types of audio signal. Zeng et. al [18] have proposed a multi-task scheme for performing many audio discriminations missions concurrently. They have applied Deep Neural Network (DNN) in their work.

3 Proposed Methodology

Earlier research activities reveal that frequency and perceptual domain acoustic features are capable to discriminate speech and song. But most of the previous efforts have involved frequency and perceptual domain acoustic features to discriminate speech and song. Features computed from pitch are widely used feature in case of voice signal processing. Hence, pitch based features are overused. Overuse of pitch and other perceptual domain acoustic features has motivated to search a good feature set for discriminating speech and song. In this

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

effort a Short Time Energy (STE) based time domain feature set has been proposed. The proposed scheme is depicted through Figure. 1.

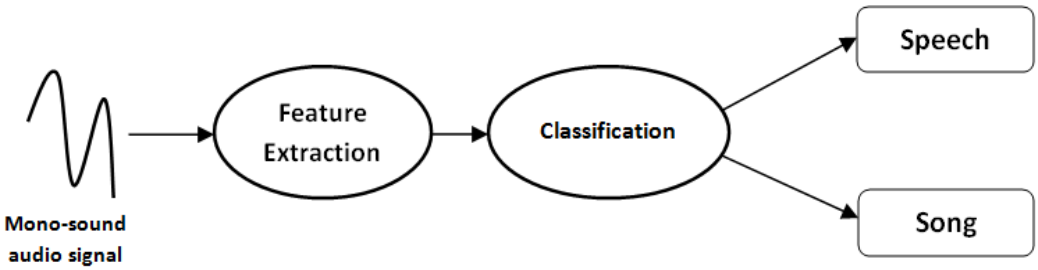

Figure 1: Outline of proposed scheme

3.1 Feature Computation

When we speak with solo voice then it is treated as speech but when speech is escorted with instrumental in background then it is considered as song. When we hear the sounds of speech and song, we can realise that energy circulation of song and speech signal is different. In case of speech signal there exists lots of silences while in case of song signal this silence does not exist at all. This happens due to the fact that maximum time instrumentals are being performed in the background of song. This observation has led to consider energy as feature. Energy is time domain feature. Short Time Energy (STE) is well capable to capture this energy variation nature of speech and song.

It is witnessed that energy level of speech signal is sensibly diverse than that of song signal. To be conversant with in what way energy level is deviating in regard to time for speech and song signal STE is necessitated to be computed. All the input audio files (both speech and song) are fragmented into numerous frames. STE is quantified for any input auditory signal divided into P frames {xi(m): $1 \leq i \leq P$} by equation (1) as:

$$ste_n = \frac{1}{leng} * \sum_{m=0}^{n-1} [x_n(st)]^2 \quad (1)$$

where, ste_n specifies energy matching to n^{th} frame, leng designates length of the frame, and $x_n(st)$ signifies st^{th} sample in n^{th} frame. All frames are formed in such a way that they are 50% overlapped to elude loss of any edge line nature for a frame.

It is acknowledged that mean and standard deviation of any feature are able to provide only an overall impression about its spreading. No precise information associated to it is available through mean and standard deviation. For detailed study about the characteristics of Short Time Energy (STE), thought of co-occurrence matrix has been exploited [8]. Thought of co-occurrence matrix has come from the domain of Image Processing where dissimilar intensity values contained by a neighbourhood generates a design and that is employed to parameterize the texture or appearance of an image. This concept is employed in this effort. Short Time Energy (STE) is generating a series of values for an input audio signal as the signal is broken into several frames. Occurrence of different STE values enclosed by a neighbourhood reverberates the pattern which actually characterizes the pattern of the audio signal. So, a matrix Co_m of dimension $B \times B$ (where, $B = \max\{ste_i\} + 1$) is produced. An component in

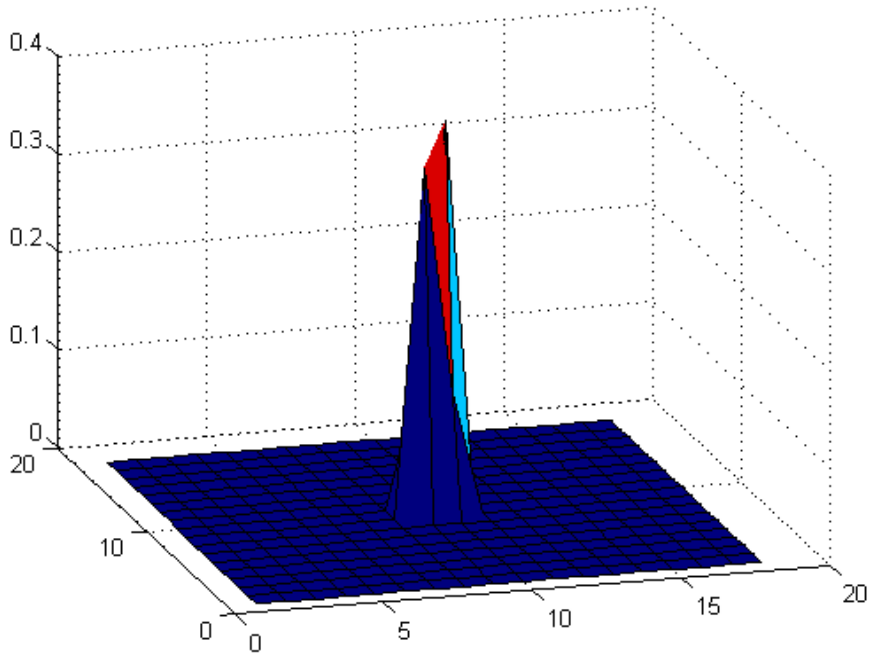

Figure 2: Co-occurrence matrix plot of STE for speech signal

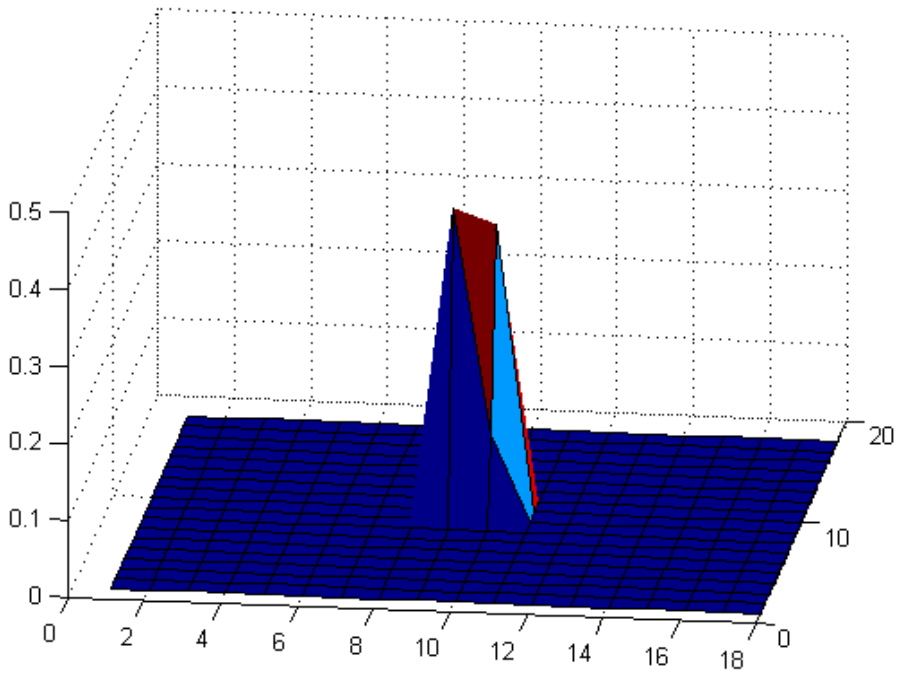

Figure 3: Co-occurrence matrix plot of STE for song signal

the matrix $Co_m(s, r)$ designates the number of occurrences of STE s and r in consecutive time illustrations. Finally, statistical features (entropy, energy, contrast, homogeneity and correlation) are calculated from the co-occurrence matrix [1]. Co-occurrence matrix plots of STE for speech and song are portrayed in Figure 2 and Figure 3 respectively. These two figures indicate that occurrence pattern of STE is quite dissimilar for speech and song.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

3.2 Classification

Simple besides popular supervised classifiers like Support Vector Machine (SVM), Neural Network and k-Nearest Neighbours (k-NN) have been enforced in this work as the main objective of this effort is to project the distinguishing capability of the advised feature set. Multi Layer Perceptron (MLP) model has been adopted in this effort to implement Neural Network. While performing classification, entire audio data set is divided into two equal parts – one part is used for training of the classifiers and the other part is used for testing the classifiers. The audio data set is comprising of total 200 audio files where speech and song files have equal number of files that is 100 files for each category. Now training and testing data set is formed by taking half of the files for every category. That is 50 files for each category has been considered for both training and testing purposes. SVM has been enforced here considering Quadratic kernel type. Neural Network which has been applied here considering MLP model of it. The MLP model has been configured by putting 5 neurons in the input layer reflecting 5 statistical feature values, 2 neurons in the output layer reflecting 2 classes of audio files - speech besides song. There is only one hidden layer in the MLP model where there exist 4 neurons. k-NN has been configured in this effort by considering 'City Block' as distance metric, k=3 and Nearest Neighbour rule for breaking the tie (if any occurs).

4 Experimental Results

A custom audio dataset has been formed for performing the speech and song discrimination experiment. The data set is containing a total 200 audio files – 100 audio files for each category (speech and song). These audio files are mono types – for easy analysis of different features of audio. Duration of these audio files are restricted to 90 seconds. During formation of the audio data set, voices of different languages as well as voices of dissimilar age groups are reflected. The data set is formed by collecting audio files from various sources like speech soundtracks of various live programs, CD footage. Some audio files are collected from Internet also. Few audio files in this audio data set are kept noisy intentionally to reflect real life scenario.

To avoid chances of some incidents like missing of any border appearances for a certain frame, all the audio files are fragmented into various frames. These frames are 50% overlapped with adjacent previous frames. Half (50%) of data set has been utilized for training purpose and the residual 50% has been utilized for testing. The classification outcomes are tabularized in Table 1.

Table 1: *Classification accuracy for discrimination of speech and song*

Classification Scheme	Classification Accuracy (in %) for proposed work
SVM	94.5
Neural Network (MLP)	92.0
k-NN	91.5
SVM	94.5

4.1 Comparative Analysis

Discrimination performance of advised feature set has been compared with some previous efforts aiming to differentiate speech and song. The custom audio data set which is used in this exertion has also been used to implement their schemes. This effort has been compared with the two schemes proposed by David Gerhard

[2], [4]. In the first effort [2] Gerhard has applied features computed from pitch. In his second effort [4], he has made use of pitch, rhythm based perceptual features. Fuzzy classification scheme has been adopted in his second effort. The comparative analysis is tabularized in Table 2. The comparative analysis reveals that STE based proposed feature set accomplishes better result aiming to differentiate speech and song.

Table 2: *Comparative analysis of proposed feature set with other works*

Precedent Method	Classification Accuracy (in %)
David Gerhard [2]	88.5
David Gerhard [4]	89.0
STE based proposed feature set (SVM)	94.5

5 Conclusion

For discrimination of speech and song, statistical feature based modest feature set has been suggested in this work. Statistical feature set are derived from the co-occurrence matrix of STE. Experimental outcome shows that proposed feature set is able to discriminate speech and song better than previous efforts of discrimination of speech and song. The suggested feature set is very simple in computation. Also dimension of the proposed feature set is also small. In future some other audio features may be explored to improve the discrimination accuracy. Some advanced methods like deep learning may also be explored to carry out the said discrimination task in a better way.

References

[1] R. M., Haralick, L. G. Shapiro, "*Computer and Robot Vision*", Addison-Wesley Longman Publishing Co., Inc., Boston, Vol. I, 1992.

[2] D. Gerhard, "Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing", *Canadian Acoustics*, 30(3), pp. 152-153, 2002.

[3] A. Bugatti, A. Flammini, P. Migliorati, "Audio classification in speech and music: a comparison between a statistical and a neural approach", *EURASIP Journal on Advances in Signal Processing*, 4, 2002.

[4] D. Gerhard, "Perceptual features for a fuzzy speech-song classification", *IEEE International Conference on Acoustics Speech And Signal Processing*, 4, pp. 4160-4160, 2002.

[5] D. Gerhard, "Silence as a Cue to Rhythm in the Analysis of Speech and Song", *Canadian Acoustics*, 31(3), pp. 22-23, 2003.

[6] G. Tzanetakis, "Song-specific bootstrapping of singing voice structure", *IEEE International Conference on Multimedia and Expo, ICME'04*, 3, pp. 2027-2030, IEEE, 2004.

[7] R. S. Lin, L. H. Chen, "A new approach for classification of generic audio data", *International Journal of Pattern Recognition and Artificial Intelligence*, 19(01), pp. 63-78, 2005.

[8] S. E. Umbaugh, *Computer imaging: digital image analysis and processing*. CRC press, 2005.

[9] Y. G., Zhang, C. S. Zhang, "Separation of music signals by harmonic structure modeling", *Advances in Neural Information Processing Systems*, pp. 1617-1624, 2006.

[10] D. Ruinskiy, Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals", *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), pp. 838-850, 2007.

[11] Y. Lavner, D. Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation", *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1-14, 2009.

[12] A. Gallardo-Antolín, J. M. Montero, "Histogram equalization-based features for speech, music, and song discrimination", *IEEE Signal processing letters*, 17(7), pp. 659-662, 2010.

[13] I. Salselas, B. Herrera, "Music and speech in early development: automatic analysis and classification of prosodic features from two Portuguese variants", *Journal of Portuguese Linguistics*, 10(1), pp. 11-35, 2011.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

[14] R. Sonnleitner, B. Niedermayer, G. Widmer, J. Schlüter, "A simple and effective spectral feature for speech detection in mixed audio signals", *Proceedings of the 15th International Conference on Digital Audio Effects*, DAFX-1 – 7, 2012.

[15] H. Bhavsar, M. H. Panchal, "A review on support vector machine for data classification", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1(10), pp. 185-189, 2012.

[16] M. Velayatipour, M. Mosleh, "A review on speech-music discrimination methods", *International Journal of Computer Science and Network Solution*, 2(2), pp. 67-78, 2014.

[17] T. Ramalingam, P. Dhanalakshmi, "Speech/music classification using wavelet based feature extraction techniques", Journal of Computer Science, 10(1), pp. 34-44, 2014.

[18] Y. Zeng, H. Mao, D. Peng, Z. Yi, "Spectrogram based multi-task audio classification". *Multimedia Tools and Applications*, 78(3), pp. 3705-3722, 2019.