ISSN 0973-3302

THE JOURNAL OF ACOUSTICAL SOCIETY OF INDIA

Volume 50 Number 1 January 2023

ACOUSTICAL SOCIETY OF INDIA
1971
॥ नादाधीनमतोजगत् ॥

A Quarterly Publication of the ASI
https://acoustics.org.in

The Journal of Acoustical Society of India
Volume 50 Number 1 January 2023 pp 1-75

ASI

The Journal of Acoustical Society of India

The Refereed Journal of the Acoustical Society of India (JASI)

The Journal of Acoustical Society of India is a refereed journal of the Acoustical Society of India **(ASI)**. The **ASI** is a non-profit national society founded in 31st July, 1971. The primary objective of the society is to advance the science of acoustics by creating an organization that is responsive to the needs of scientists and engineers concerned with acoustics problems all around the world.

Manuscripts of articles, technical notes and letter to the editor should be submitted to the Chief Editor. Copies of articles on specific topics listed above should also be submitted to the respective Associate Scientific Editor. Manuscripts are refereed by at least two referees and are reviewed by Publication Committee (all editors) before acceptance. On acceptance, revised articles with the text and figures scanned as separate files on a diskette should be submitted to the Editor by express mail. Manuscripts of articles must be prepared in strict accordance with the author instructions.

All information concerning subscription, new books, journals, conferences, etc. should be submitted to Chief Editor:

B. Chakraborty, CSIR - National Institute of Oceanography, Dona Paula, Goa-403 004,
Tel: +91.832.2450.318, Fax: +91.832.2450.602, e-mail: bishwajit@nio.org

Annual subscription price including mail postage is Rs. 2500/= for institutions, companies and libraries and Rs. 2500/= for individuals who are not **ASI** members. The Journal of Acoustical Society of India will be sent to **ASI** members free of any extra charge. Requests for specimen copies and claims for missing issues as well as address changes should be sent to the Editorial Office:

ASI Secretariat, C/o Acoustics and Vibration Metrology, CSIR-National Physical Laboratory, Dr. KS Krishnan Road, New Delhi 110 012, Tel: +91.11.4560.8317, Fax: +91.11.4560.9310, e-mail: asisecretariat.india@gmail.com

ISSN 0973-3302

Printed at Alpha Printers, WZ-35/C, Naraina, Near Ring Road, New Delhi-110028 Tel.: 9810804196. JASI is sent to ASI members free of charge.

The Journal of Acoustical Society of India

A quarterly publication of the Acoustical Society of India

Volume 50, Number 1, January 2023

ARTICLES

INFORMATION

FOREWORD

The works presented in this first Volume of the three Volume Series of the Special Issue of the Journal of Acoustical Society of India (JASI) has the underlying motivation to look for objective data and corresponding categorization problems in different domains of speech, music, and corresponding interdisciplinary areas, which were all presented in the 26th International Symposium on Frontiers of Research on Speech and Music (FRSM - 2021), held at IIIT, Pune (virtual mode). The 1st Volume Volume attempts to consolidate the substances of different interesting pieces of investigations in this diverse field carried out by different researchers across the globe. It may quite be possible that in spite of taking utmost care in choosing the selected and thoroughly revised versions of the manuscript, some inadvertent errors might have cropped in, or there might be certain areas of research that we may have overlooked. If so, the Editors regret the same.

The quest for the origin of speech and music is enigmatic. holistic, manipulative, multimodal, musical and mimetic language of the hominids also known as 'hmmmm' or 'musilanguage,' refers to a pre-linguistic system of vocal communication from which both music and language are said to be derived later. Modern speech and music are forms of intentional emotional manipulation and, therefore, not possible until the onset of intentionality - the ability to reflect on the past and the future. Between 60,000 and 30,000 years ago, in the upper Palaeolithic age, humans started creating art in the form of paintings on cave walls, jewellery and so on. Speech and music as we know it now must also have emerged during that period. Speech and music are the two highest and most subtle and common forms of human communication. The understanding of these oldest and most fundamental faculties in human beings needs the scientific and technological resources of Language Processing.

The language we speak in our everyday lives propagates as sound waves through various media and allows communication or entertainment for us, humans. The music we hear or create can be perceived in different aspects as rhythm, melody, harmony, timbre, or mood. Like speech, singing can also be a way to communicate. Both speech and music can be described using the four basic parameters of sound: pitch (how high or low the note is), loudness, duration and timbre (the quality or tone of a sound; put simply, it is what makes one musical sound different from another). Speech and music use these parameters in different unique ways. In singing, the two are brought together. This may seem simple when we listen to music, but in fact, reaching an accommodation between speech and music is a complex human skill. The multifaceted nature of speech or music information requires algorithms and systems using sophisticated signal processing, machine learning and deep learning techniques to better extract useful information. A systematic scientific investigation into the production, perception and cognition of speech and music thus requires a multidisciplinary approach involving Physics, Computer Science, Phonetics, Physiology, Psychology, and Musicology, to name a few. Even if one puts aside aesthetic appreciation, the task becomes formidable.

It is this task that the concept of FRSM has been bringing forth since its inception in 1991 and this series of Edited Volumes of the Journal of Acoustical Society of India (JASI) is an attempt to bring together research works happening in this ever expanding field across the Indian subcontinent and abroad. The first volume contains seven (7) chapters spanning over the following broad areas:

In the area of speech :

a. Analysis of the intonation pattern of Odia, an Indian Language.

b. Proposal of a wavelet-based speech enhancement technique based on Bivariate Shrinkage function and Firm thresholding

In the area of music :

a. Analysis of the perceptual domain of emotional elicitation in humans evoked by Raga renditions played in two most popular classical string instruments.

b. Definition of Esraj's (Indian bowed string instrument) sound quality and its comparison with that of violin sounds, using timbre parameters.

In the area of inter-disciplinary research :

a. The problem of detecting mistakes and prescribing corrective feedback in pedagogy.

b. Study of the intermediality of musical emotions from the perspective of audio-visual and (AV) and audio-only (AO) stimulus and their corresponding neural manifestations.

c. Proposal of a novel method for bird species identification from the Timbral attributes of a sound produced by the bird.

The above-mentioned 7 (seven) selected works presented in this Special Issue have been taken from eminent researchers in the allied disciplines of Speech and Music across the Indian sub-continent. The Editors sincerely hope that serious researchers, academicians, interested personalities of this unique discipline would benefit from this Volume.

Shankha Sanyal, Archi Banerjee,
Sanjeev Sharma and Ranjan Sengupta
—Editors

The Journal of Acoustical Society of India : Vol. 50, No. 1, 2023 (pp. 1-5)

Intonation pattern of Odia

N. Jyoti Laxmi[1*] and Irfana M.[2]

[1]Speech-language pathologist, Institute of speech and audiology, Sweekaar Academy of Rehabilitation Sciences, Secunderabad-500 009, India

[2]Department of speech-language pathology,
All India Institute of Speech and Hearing (AIISH), Manasagangothri, Mysore-570 006, India
e-mail: laxmijyoti69@gmail.com

[Received: 12-03-2022; Revised: 05-06-2022; Accepted: 25-06-2022]

ABSTRACT

Intonation is an inevitable element of speech output and a mandatory part to convey the intent of the communication. Inappropriate usage of intonation causes misinterpretation by the listeners. Language-specific intonation patterns are observed and native speakers are tending to use them with great variations in their choices of intonation patterns. The present study aimed to analyze the intonation pattern of Odia, an Indian Language. There were 20 subjects including an equal number of male and female subjects for the study. The database consisted of four different types of sentences i.e. questions, exclamatory, declarative, and commands. A total of 20 samples were prepared with five sentences from each type of sentence. Subjects were asked to produce each sentence with appropriate intonation and the recorded database was analyzed for further understanding. Overall pitch contour and phrase-specific pitch patterns were extracted from the elicited samples. Peaking was noticed in an exclamatory sentence where commands had a greater occurrence of dipping patterns. Knowledge regarding the intonation of a specific language will facilitate understanding speech perception and computational linguistics.

1. INTRODUCTION

Speech not only expresses the strictly linguistic content of sentences but also the expression of attitudes and emotions of the speaker. Intonation is an inevitable element of speech output and a mandatory part to convey the intent of the communication. It is the melody of the sentence and is created by changes in the pitch of the voice, sentence stress, and rhythm. The most essential functions of intonation are to differentiate types of sentences such as questions, exclamatory, declarative, and commands and to segregate sentences into sense groups. Also, intonation allows speakers to express various emotions. Inappropriate usage of intonation causes misinterpretation by the listeners.

An association established by several studies on intonation attributes the high pitch levels to illustrate the attention of the listener, because it creates a contrast, and to emotions, such as anger, joy, anxiety, etc. Even though other parameters such as intensity, duration, rate of speech, accent, and vocal qualities have an impact to convey the meaning in the complexity of speech, the pitch variable solely affects the production and perception of speech. Hence the present study intended to explore the intonation pattern using pitch variation (Rodero, 2011; Mozziconacci, 2000; Pell, 2001). Sober emotions such as sadness and calmness had low pitch levels whereas emotions carrying a high level of activity, such as joy or fear, tend to be situated in the top end of the frequency spectrum of the speaker. Conversely, sadness and desire tend overall to be formed at the lower end (Waaramaa, Laukkanen, Alku & Väyrynen, 2008; Rodŕlguez, 2002; Johnstone & Scherer, 1999). An English question with a declarative sentence differs from its true declarative counterpart only in terms of tonal contour. The question is rising and the statement is falling (Gunlogson, 2001).

Several studies modeling the intonation patterns in different languages worldwide are observed in the literature (Benesty *et al.,* 2008; Cosi *et al.,* 2001; Vainio, 2001; Hwang and Chen, 1994; Buhmann et al., 2000). But, in the Indian context, there is a limited amount of data for intonation (Kumar *et al.,* 1993; Kumar, 1993). Language-specific intonation patterns are observed and native speakers are tending to use them with great variations in their choices of intonation patterns. A new language learner must organize the various sounds which he/she utters according to a system of intonation which the respective native speaker can recognize. A potentially more serious problem could arise if, instead of simply not being understood at all, or deemed offensive. That in itself is a telling argument for the teaching of the prosody of a foreign language. The importance of this communicative function has often been overlooked in other language teachings. Hence the present study aimed to analyze the intonation pattern of Odia. Odia formerly stated as Oriya is an Indo-Aryan language spoken in Odisha, an Indian state. It is the official language in Odisha formerly called Orissa (Mahapatra, 2002).

2. METHODOLOGY

2.1 *Participants*

There were 20 subjects with an equal number of males and females for the study. All of them were native speakers of Odia and consent was taken before the data collection. None of them reported any speech, language, or neurological deficits.

2.2 *Database*

The database consisted of four different types of sentences *i.e.* questions, exclamatory, declarative, and commands. A total of 20 samples were prepared with five sentences from each type of sentence. Each sentence was prepared with three words. The length of each word across the sentences was controlled where initial words were two syllables and following words were three syllables.

2.3 *Procedure*

Each participant was instructed regarding the emotions in which they need to say the respective stimuli. Trials were taken before the final data collection procedure to avoid confusion. Later, subjects were asked to produce each sentence with appropriate intonation and the recorded database was analyzed using Praat software for further understanding. F_0 of each word was extracted for each sentence where the temporal normalization was attained by taking each word as a separate segment and defining it as an individual interval in TextGrid. This was done distinctively for each gender to consider the predictable gender effect in F_0 and statistical analysis was administered to identify the significant effect of variables.

3. RESULTS

As seen in the Table 1 descriptive statistics were administered for the frequency of each word of the sentences. Variations between the mean F_0 of three words were less for questions and exclamations. However declarative and command sentences had comparatively higher differences between the words among the sentence group. There was an obvious gender difference in the F_0 and the variations between the words within each sentence type were greater in females. Standard deviations were less and it shows the data had homogeneity.

Further statistical analysis was done using Friedman, a nonparametric test to investigate the effect within each type of sentence. Results showed that there was significant difference seen in both gender

Table 1. Mean and standard deviation of the frequency of each word of the sentences

	Male		**Female**	
	Mean	**S.D**	**Mean**	**S.D**
QW1	120.57	1.771	245.81	9.149
QW2	126.44	2.551	240.57	5.216
QW3	124.54	1.301	242.06	5.754
EW1	119.60	0.996	258.77	10.162
EW2	117.76	0.648	236.26	4.113
EW3	114.22	0.765	256.52	9.124
DW1	126.25	0.714	262.21	4.281
DW2	118.14	1.365	244.23	3.661
DW3	117.79	0.771	238.73	6.799
CW1	125.05	0.630	354.17	2.970
CW2	116.63	1.340	235.31	9.388
CW3	111.82	0.749	234.78	7.863

*QW1= First word of Question; QW1= Second word of Question; QW3= Third word of Question; EW1= First word of Exclamatory; EW2= Second word of Exclamatory; EW3= Third word of Exclamatory; DW1= First word of Declarative; DW2= Second word of Declarative; DW3= Third word of Declarative; CW1= First word of Command; CW2= Second word of Command; CW3= Third word of Command

groups for all four types of sentences i.e. question (male-χ^2 = 16.800, p=0.000; female-χ^2 = 7.800, p=0.020), exclamatory (male-χ^2 = 18.200, p=0.000; female-χ^2 = 16.800, p=0.020), declarative (male-χ^2 = 15.200, p=0.001; female-χ^2 = 20.000, p=0.000), and command (male-χ^2 = 20.000, p=0.000; female-χ^2 = 15.200, p=0.001). Further, the Wilcoxon Signed Rank test was administered to see the differences in F_0 of words, and Z value and p values are given in Table 2. There was a significant difference between all the pairs of

Table 2. Results of Wilcoxon Signed Rank test

Pairwise comparison	**Male**		**Female**	
	/Z/	**p**	**/Z/**	**p**
QW2 - QW1	2.805	0.005**	2.296	0.022**
QW3 - QW1	2.805	0.005**	2.296	0.022**
QW3 - QW2	2.193	0.028**	1.786	0.074
EW2 - EW1	2.705	0.007**	2.807	0.005**
EW3 - EW1	2.812	0.005**	2.502	0.012**
EW3 - EW2	2.807	0.005**	2.807	0.005**
DW2 - DW1	2.809	0.005**	2.809	0.005**
DW3 - DW1	2.809	0.005**	2.809	0.005**
DW3 - DW2	0.562	0.574	2.809	0.005**
CW2 - CW1	2.809	0.005**	2.807	0.005**
CW3 - CW1	2.814	0.005**	2.807	0.005**
CW3 - CW2	2.809	0.005**	0.357	0.721

*QW1= First word of Question; QW1= Second word of Question; QW3= Third word of Question; EW1= First word of Exclamatory; EW2= Second word of Exclamatory; EW3= Third word of Exclamatory; DW1= First word of Declarative; DW2= Second word of Declarative; DW3= Third word of Declarative; CW1= First word of Command; CW2= Second word of Command; CW3= Third word of Command words of both genders except the second and third word pairs of questions and commands in females and the second and third-word pair of declarative in males.

Overall pitch contour was drawn from the elicited samples to visualize the pitch variation from word to word and shown in figure 1. For questions, dipping was seen in males whereas peaking was noticed in females. Similarly, for exclamation, dipping was noticed in males and progressive fall was seen in females. However, exclamatory and command sentence types had a similar pattern where both genders showed progressive fall.

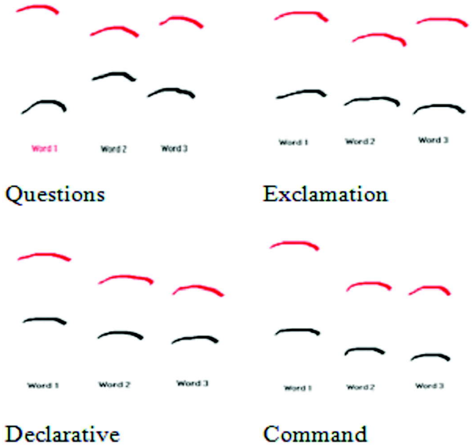


Fig. 1. Diagrammatic representation of F_0 changes across words in each type of sentence. Black lines represent male speakers and Red lines indicate Female speakers.

4. DISCUSSION AND CONCLUSION

The present study showed a reduction of the pitch for the middle word in comparison to the first word for most types of words. This is incongruent with Indian English and Hindi where there is a post-focal reduction in pitch range, duration, and RMS (Patil *et al.*, 2008). Ueyama and Jun (1998) reported similar trends in Japanese English and Korean English, but the post focal pitch reduction sustained till the second word which is not different from the results of the present study. This discrepancy can be because of the phonetic variation between the considered languages. Declarative and commands were showed similar trends for both gender, however, slight differences noticed in the case of question and exclamation in the present study. A similar gender effect in intonation development was reported in a previous study (Ferrand & Bloom, 1996). This can be reasoned as both physiological and socio-cultural factors appear to account for the changes that mark the intonation patterns in males and females.

To conclude the study, the present study considered four different types of sentences in Odia and cross-checked the intonation pattern using pitch variation. Results showed that for questions, dipping was seen in males whereas peaking was noticed in females. Similarly, for exclamation, dipping was noticed in males and progressive fall was seen in females. However, exclamatory and command sentence types had a similar pattern where both genders showed progressive fall. Knowledge regarding intonation will facilitate the understanding of speech perception of Odia and be helpful for computational linguistics. However, the present study considered only 3-word sentences to control the variable which can be a limitation of the study since it restricted the complexity of the sentences.

5. REFERENCES

[1] Benesty, Jacob, M. Mohan Sondhi and Yiteng Arden Huang, 2008. "Introduction to speech processing." In Springer Handbook of Speech Processing, *Springer, Berlin, Heidelberg,* pp. 1-4.

[2] Buhmann, Jeska, Halewijn Vereecken, Justin Fackrell, Jean-Pierre Martens and Bert van Coile, 2000. "Data-drove intonation modeling of 6 languages." *In Sixth International Conference on Spoken Language Processing.*

[3] Ferrand, Carole T. and Ronald L. Bloom, 1996. "Gender differences in children's intonational patterns." *Journal of Voice,* **10**(3), 284-291.

[4] Cosi, Piero, Fabio Tesser, Roberto Gretter, Cinzia Avesani and Mike Macon, 2001. "Festival speaks Italian!." *In Seventh European Conference on Speech Communication and Technology.*

[5] Gunlogson, Christine, 2001. True to form: Rising and falling declarative as questions in English. *Routledge.*

[6] Hwang S-H. and S-H. Chen, 1994. "Neural network-based F0 text-to-speech synthesizer for Mandarin." *IEE Proceedings-Vision, Image and Signal Processing,* **141**(6), 384-390.

[7] Johnstone, Tom and Klaus R. Scherer, 1999. "The effects of emotions on voice quality." *In Proceedings of the XIVth international congress of phonetic sciences,* Department of Linguistics, Univ. of California at Berkeley Berkeley, CA, pp. 2029-2032.

[8] Kumar A.S.M., 1993. "Intonation knowledge for speech systems for an Indian language." Ph.D. diss., Ph.D. thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology, Madras, Chennai, India.

[9] Madhukumar A.S., S. Rajendran and B. Yegnanarayana, 1993. "Intonation component of a text-to-speech system for Hindi." *Computer Speech & Language,* **7**(3), 283-301.

[10] Mahapatra B.P., 2002. Linguistic Survey of India: Orissa. Language Division, Office of the Registrar General, **1**.

[11] Moore Robert Ripley, 1965. A study of Hindi intonation. *University of Michigan.*

[12] Mozziconacci and Sylvie J.L., 2000. "The expression of emotion considered in the framework of an intonation model." *In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.*

[13] Patil, Umesh, Gerrit Kentner, Anja Gollrad, Frank Kügler, Caroline Féry and Shravan Vasishth, 2008. "Focus, word order and intonation in Hindi." *Journal of South Asian Linguistics,* **1**.

[14] Radiofonica and Rodero E. Locucion, 2003. "Voice on Radio." *Madrid, Spain: IORTV.*

[15] Rodrlguez A., 2002. "Propuestas para una modelizacio n del uso expresivo de la voz. [Proposals for a model of the expressive use of voice]." *Zer,* **13,** 157-175.

[16] Ueyama, Motoko and Sun-Ah Jun, 1996. "Focus realization of Japanese English and Korean English intonation." *UCLA Working Papers in Phonetics,* pp. 110-125.

[17] Vainio, Martti, 2001. "Artificial neural network based prosody models for Finnish text-to-speech synthesis".

[18] Waaramaa, Teija, Anne-Maria Laukkanen, Paavo Alku and Eero Väyrynen, 2008. "Monopitched expression of emotions in different vowels." *Folia Phoniatrica et Logopaedica,* **60**(5), 249-255.

The Journal of Acoustical Society of India : Vol. 50, No. 1, 2023 (pp. 6-20)

Speech enhancement based on bivariate shrinkage function and firm thresholding

Vasundhara Shukla* **and Preety D. Swami**
[1]*Department of Electronics and Communication Engineering*
Oriental College of Technology, RGPV, Bhopal, India
e-mail: shuklav2015@gmail.com

[Received: 30-03-2022; Revised: 17-06-2022; Accepted: 31-06-2022]

ABSTRACT

Speech enhancement is one of the basic requirements for mobile communication and speech recognition systems. Speech enhancement involves the extraction of clean signals from interfering noise. One of the methods to perform this is transforming the signal to an appropriate domain where these signals can be distinguished easily. In this work, we propose a wavelet-based speech enhancement technique based on the bivariate shrinkage function and firm thresholding. The bivariate shrinkage function assumes the inter-scale dependency between the wavelet coefficients, and firm thresholding overcomes the drawbacks of soft and hard thresholding, which makes the proposed technique unique in comparison to other thresholding techniques. Finally, the proposed approach is tested against different types of noise signals.

1. INTRODUCTION

Speech signals are often corrupted by noise during the capturing process, leading to severe problems in speech analysis, voice signature detection, speech recognition, and perception (Das *et al.* 2020). In principle, automating noise removal would be a critical first step in a large number of signal processing applications. Despite denoising's long history, there is always room for improvement. Originally, time-domain filtering was used to remove high-frequency noise from low-frequency signals, but it does not provide good results in real-world situations (Muralikrishnan and Raja 2009). A modern algorithm filters signals in some transform domains, such as Wavelet or Fourier, to improve performance (Fan *et al.* 2019).

A wavelet-based decomposition of a signal has been widely used in signal processing algorithms (Mallat 1999). The use of wavelet transforms in other fields has also increased after the community recognized that this could be used as an alternative to Fourier analysis.

Multi-resolution analysis has the advantage of being able to analyze non-stationary random processes, such as speech signals (Mallat 1999; Carnero and Drygajlo 1999). Most wavelet coefficients (WCs) of a speech signal are zero when transformed into a wavelet domain (Ahani, Ghaemmaghami, and Wang 2015). This implies that the wavelet is capable of representing signals with sparse nonzero coefficients. Considering that noise tends to have smaller values compared to speech components, the noise can be removed by shrinking wavelet coefficients. It is important to estimate a good threshold level to obtain a good noise reduction result (Daqrouq *et al.* 2010). Even though the wavelet transform can alone be used for speech denoising without machine learning, hence denoising can be performed without prior training. Using wavelets as a part of the feature extraction technique in machine learning based denoising algorithms can certainly improve their performances.

Wavelets have been used extensively for image processing applications; however, for speech processing, relatively little literature is available. Therefore in this work, we try to analyze the performances of different wavelets under different noises and noise levels to get a better insight into their behavior for speech enhancement applications. Additionally, knowledge of such characteristics can also help in adopting proper wavelet transforms for machine learning based speech denoising techniques.

The rest of the paper is organized as follows: In Section 2, a brief review of relevant literature is provided. Sections 3 and 4 explain the wavelet transform and Bivariate Shrinkage function respectively. Section 5 describes the proposed algorithm, followed by the experimental analysis in Section 6. Finally, Section 7 concludes the paper with possible future scopes.

2. RELATED WORK

Presently, several wavelet-based speech enhancement algorithms have been proposed. In this section, some of them are presented. J. W. Seok *et al.* (Seok and Bae 1997) used semi-soft thresholding to remove additive background noise from noisy speech in the wavelet domain. Furthermore, to prevent the quality degradation of the unvoiced sounds during the denoising process, the unvoiced region is classified first and then thresholding is applied differently.

M. Bahoura *et al.* (Bahoura and Rouat 2001) proposed a speech enhancement method based on the time adaptation of wavelet thresholds. The time dependence is introduced by approximating the Teager energy Wavelet of the coefficients. This technique does not require an explicit estimation of the noise level or priori knowledge of the SNR, which is usually needed in most of the popular enhancement methods.

S. H. Chen *et al.* (S.-H. Chen and Wang 2004) used the perceptual wavelet packet decomposition (PWPD) and the Teager energy operator (TEO) for speech enhancement. The main advantage of the proposed method is that the over thresholding of speech segments which usually occurs in conventional wavelet-based speech enhancement schemes can be avoided. In addition, the proposed method does not require a complicated estimation of the noise level or any knowledge of the SNR.

S. Badiezadegan *et al.* (Badiezadegan and Rose 2015) used the spectrographic masks for deriving thresholds for de-noising wavelet domain coefficients, making DWT-based de-noising more suitable for non-stationary noise conditions. The proposed approach reduces the impact of model mismatch associated with parametric approaches and exploits the robustness of the non-parametric wavelet de-noising approach.

M. Anouar *et al.* (Ben Messaoud, Bouzid, and Ellouze 2016) proposed a single-channel speech enhancement method based on the combination of the wavelet packet transform and an improved version of the principal component analysis (PCA). The method provides higher noise reduction and lower signal distortion even in highly noisy conditions without introducing artifacts. M. A. Oktar *et al.* (Oktar, Nibouche, and Baltaci 2016) proposed a discrete wavelet packet transform algorithm for speech signal denoising. Both hard and soft thresholding are applied. S. Mavaddaty *et al.* (Mavaddaty, Ahadi, and Seyedin 2017) presented a new learning-based speech enhancement algorithm via sparse representation in the wavelet packet transform domain. They proposed sparse dictionary learning procedures for training data of speech and noise signals based on a coherence criterion for each subband of the decomposition level. The speech enhancement algorithm is introduced in two scenarios, supervised and semi-supervised.

S. R. Chiluveru *et al.* (Chiluveru and Tripathy 2021) carried out the denoising of a noisy speech signal with a wavelet thresholding technique. The noisy signal was decomposed into different frequency bands, and this decomposition level (DL) was decided independently of non-stationary noise. In this work, a new DL detection procedure was presented, and it decides the decomposition level based on signal energy and speech dominance. Similar work was proposed in (Tohidypour and Ahadi 2016) by Tohidypour *et* *al.* However their work was based on Zhang thresholding function (Zhang and Desai 1998), whereas in this work we adopted the firm thresholding (Gao, Bruce, and Inc 1997) proposed by Gao *et al.*; another difference is that in the proposed work, multiple wavelet functions are considered and analyzed, which was not done in their work.

3. WAVELET TRANSFORM AND DENOISING

Wavelet analysis adopts a wavelet prototype function known as the mother wavelet, given in Eq. (1).

$$C_{j,k}=2^{-\frac{1}{2}}\int f(t)\phi(2^{-j}t-k)dt \tag{1}$$

This mother wavelet, in turn, generates a set of basis functions known as child wavelets through recursive scaling and translation. The variable *s* reflects the scale or width of a basis function, and the variable t is the translation that specifies its translated position on the time axis.

$$\psi(\tau,s) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right) \tag{2}$$

In Eq. (2), $\psi\left(\frac{t-\tau}{s}\right)$ is the mother wavelet, and the factor $\frac{1}{\sqrt{s}}$ is a normalization factor used to ensure that energy across different scales remains the same.

3.1 *Continuous Wavelet Transform*

Continuous wavelet transform (CWT) (Aguiar-Conraria and Soares 2014) of f(t), concerning the wavelet $\psi(t)$ is defined as:

$$\mathrm{CWT}(\tau, s) = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-\tau}{s}\right)dt \tag{3}$$

where τ is the translation coefficient, and s is the scaling coefficient. CWT analyzes the signal through the continuous shifts of a scalable function over a time plane. This technique results in redundancy, and it is numerically impossible to analyze an infinite number of wavelet sets.

3.2 *Discrete Wavelet Transform*

The Discrete Wavelet Transform (DWT) (Edwards 1992) was introduced to overcome the redundancy problem of CWT. The approach is to scale and translate the wavelets in discrete steps as given in Eq. (4).

$$\mathrm{DWT}\ (\tau_o, s_o) = \frac{1}{\sqrt{s_o^j}}\int f(t)\psi\left(\frac{t-k\tau_o s_o^j}{s_o^j}\right)dt \tag{4}$$

where s_o^j is the scaling factor, and τ_o is the translating factor, k and j are just integers. Usually s_o is a fixed dilation step size of 2 so that the sampling of the frequency corresponds to dyadic sampling. The translation factor τ_o is dependent on s_o and equal to 1 for a dyadic wavelet transform. Subsequently, we can represent the mother wavelet in terms of scaling and translation of a dyadic transform as:

$$\psi_{j,k}(t)=\frac{2^{-\frac{1}{2}}}{\phi(2^{-j}t-k)} \tag{5}$$

The coefficients of DWT can be represented as:

$$C_{j,k} = 2^{-\frac{1}{2}}\int f(t)\phi(2^{-j}t-k)dt \tag{6}$$

3.3 *Multiresolution Analysis*

The essential process for constructing a set of child wavelets from the mother wavelet depends on the basis dilation equation or the scaling function ϕ(t) given as:

$$\phi(t) = 2^{-\frac{1}{2}}\phi(2^{-j}t-k) \tag{7}$$

and the wavelet function is given as:

$$\psi(t) = 2^{-\frac{1}{2}}\psi(2^{-j}t-k) \tag{8}$$

where k is the scaling index and j is the translation index. These functions will form a filter bank or an approximate bandpass spectrum, which consists of a low-pass filter (scaling function) and a high-pass filter (wavelet function). The idea of multiresolution analysis, according to Mallat (Guo *et al.* 2000) is to iteratively break down a signal by passing it through this filter bank. The output of each filter stage will consist of the scaling coefficients.

$$C_j(k) = \int f(t)\phi_{j,k}(t)dt \tag{9}$$

and the detailed coefficients,

$$d_j(k) = \int f(t)\psi_{j,k}(t)dt \tag{10}$$

Consecutively, this will form a series of detailed coefficients and a set of scaling coefficients.

The decomposition process is shown in Figure 1, where S is the original signal, $\bar{S}$ is the estimated signal after 3 levels of wavelet decomposition, and A_n and D_n correspond to the scaling coefficients and detailed coefficients, respectively. The signal S can be approximated by $\bar{S}$ where, $\bar{S} = A_3 + D_3 + D_2 + D_1$. This expression shows that there will be a few large non-zero scaling coefficients and many nearly zero detail coefficients.

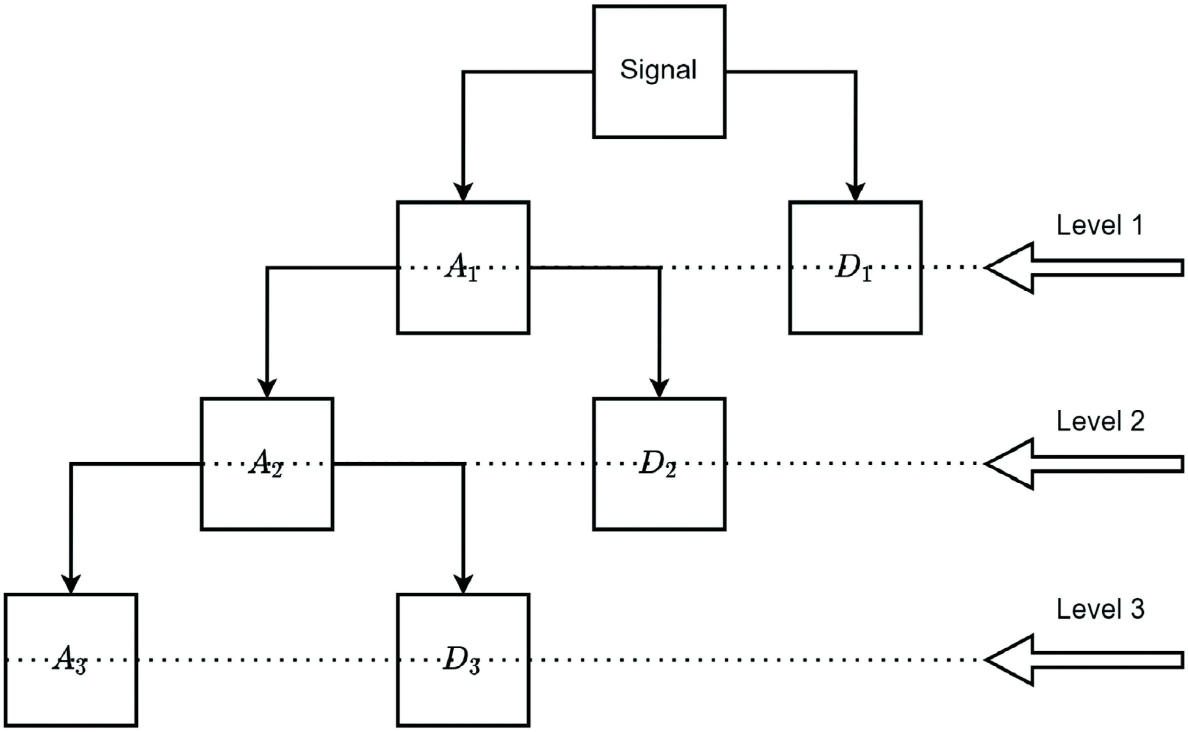


Fig. 1. Three Level Wavelet Decomposition.

3.4 *Denoising Procedure*

Given a noisy signal g = x + θ where x is the desired signal and θ is independent and identically distributed (i.i.d.) Gaussian noise N(0, σ^2), g is first decomposed into a set of wavelet coefficients y = W[g] consisting of the desired coefficient w and noise coefficient n. By applying a suitable threshold value T to the wavelet coefficients, the denoised desired coefficient $\hat{w} = T[y]$ can be obtained; Lastly, an inverse transform on the desired coefficient w will generate the denoised signal $\hat{x} = W^T[\hat{w}]$.

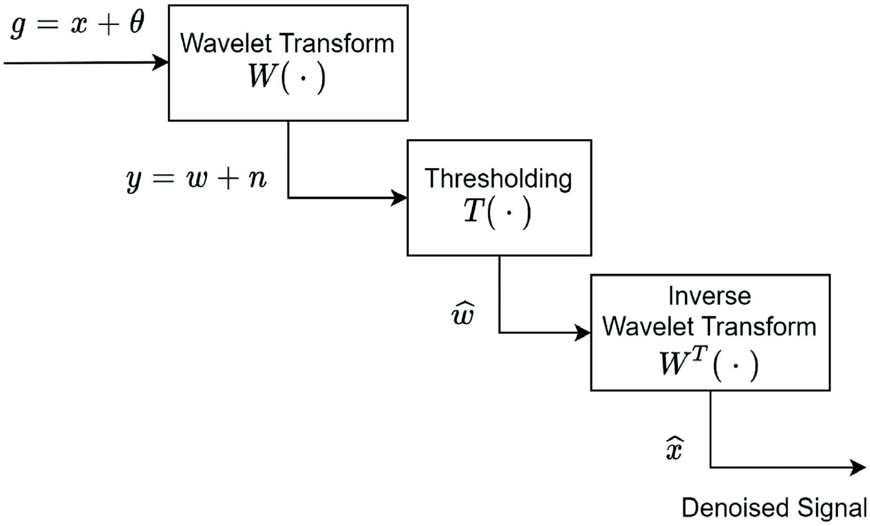


Fig. 2. Block diagram for DWT based denoising framework.

In general, a signal has its energy concentrated in a small number of coefficients, while noise has its energy spread across a large number of coefficients. Hence, through suitable thresholding or wavelet shrinkage of the higher spectral bands' components (where the noise predominantly lies), we can greatly reduce or remove the noise of the signal in the wavelet domain (G. Chen, Xie, and Zhao 2013).

4. BIVARIATE SHRINKAGE

4.1 *Denoising Procedure*

Suppose speech is corrupted by noise and presented as:

$$g = x + \theta \tag{11}$$

where, g = noisy signal, x = desired signal, θ = independent Gaussian noise. In the wavelet domain, we can write

$$y = w + n \tag{12}$$

where, y = noisy wavelet coefficient, w = true coefficient, n = noise coefficients.

In any estimation theory, the aim is to estimate w from noisy observation y. The maximum a posteriori (MAP) estimator is used for this purpose (Kazerouni *et al.* 2013). Marginal and Bivariate models are considered for this problem, but in the marginal model, statistical dependencies between wavelet coefficients are not considered; therefore, we use bivariate models.

4.2 *Bivariate Model*

The new shrinkage function which depends on both coefficient and its parent yield improved results for wavelet-based denoising. Here, we modify the Bayesian estimation problem to take into account the statistical dependency between a coefficient and its parent (Sendur and Selesnick 2002).

Let w_2 represent the parent of w_1 (w_2 is the wavelet coefficient at the same position as w_1, but at the next coarser scale.).

$$\begin{aligned} y_1 &= w_1 + n_1 \\ y_2 &= w_2 + n_2 \end{aligned} \tag{13}$$

where y_1 and y_2 are noisy observations of w_1 and w_2 and n_1 and w_2 are noise samples. Then we can write

$$y = w + n \tag{14}$$

$y = (y_1, y_2)$, $w = (w_1, w_2)$, $n = (n_1, n_2)$. The standard MAP estimator for w given corrupted y is

$$\hat{w}(y) = \arg \max_{w} P_{\frac{w}{y}}\left(\frac{w}{y}\right) \tag{15}$$

$$\hat{w}(y) = \arg \max_{w} P_{\frac{w}{y}}\left(\frac{w}{y}\right)P_w(w) \tag{16}$$

$$\hat{w}(y) = \arg \max_{w} P_N(y-w)P_w(w) \tag{17}$$

According to the bay's rule allows estimation of the coefficient can be found by probability densities of noise and prior density of wavelet coefficient. If the noise is assumed to be Gaussian, then its pdf can be written as:

$$P_n(n) = \frac{1}{2\pi\sigma_n^2},\ \exp\left(-\frac{n_1^2 + n_2^2}{2\sigma_n^2}\right) \tag{18}$$

Joint pdf of wavelet coefficients

$$P_w(w) = \frac{3}{2\pi\sigma^2},\ \exp\left(-\frac{\sqrt{3}}{\sigma} w_1^2 + w_2^2\right) \tag{19}$$

The $\hat{w}(y)$ defined in Eq. (17) can equivalently be defined as:

$$\hat{w}(y) = \arg \max_{w}[\log (P_N(y - w) + \log(P_w(w)) \tag{20}$$

Let us define $f(w) = \log(p_w(w))$ then using Eq. (18) and (19).

$$\hat{w}(y) = \arg \max_{w} \left[-\frac{(y_1-w_1)^z}{2\sigma_n^2} - \frac{(y_2-w_2)^z}{2\sigma_n^2} + f(w)\right] \tag{21}$$

This equation is equivalent to solving the following equations:

$$\frac{y_1-\hat{w}_1}{\sigma_n^2} + f_1(\hat{w}) = 0 \tag{22}$$

$$\frac{y_2-\hat{w}_2}{\sigma_n^2} + f_2(\hat{w}) = 0 \tag{23}$$

where f_1 and f_2 represent the derivative of f(w) with respect to w_1 and w_2 respectively. We know f(w) can be written as:

$$f(w) = \log (pw(w) = \log\left(\frac{3}{2\pi\sigma^2},\ \frac{\exp(-\sqrt{3}\sqrt{w_1^2 + w_2^2})}{\sigma}\right) = \log\left(\frac{3}{2\pi\sigma^2}\right) - \frac{-\sqrt{3}\sqrt{w_1^2 + w_2^2}}{\sigma} \tag{24}$$

From this

$$f_1(w) = \frac{\sqrt{3}w_1}{\sigma\sqrt{w_1^2 + w_2^2}} \tag{25}$$

$$f_2(w) = \frac{\sqrt{3}w_2}{\sigma\sqrt{w_1^2 + w_2^2}} \tag{26}$$

where σ^2 represents the marginal variance wavelet coefficients and calculated as:

$$\hat{\sigma}^2 = \left(\hat{\sigma}_y^2 - \hat{\sigma}_n^2\right)_+ \tag{27}$$

where, $(\cdot)_+$ defines as:

$$(\bullet)_+ = \begin{cases} 0, & (\bullet)_+ < 0 \\ (\bullet)_+, & (\bullet)_+ \geq 0 \end{cases} \tag{28}$$

σ_y^2 represents the noisy observations variance and calculated as:

$$\sigma_y^2 = \frac{1}{M}\sum_{y_i \in N(k)} y_i^2 \tag{29}$$

where M is the size of N(k), and N(k) denotes neighborhood coefficients. From equations (22) to (29) MAP estimator can be written as:

$$\hat{w}_1 = \frac{\left(y_1^2 + y_2^2 - \frac{\sqrt{3}\sigma_n^2}{\sigma}\right)_+ \bullet y_1}{\sqrt{y_1^2 + y_2^2}} \tag{30}$$

where the noise variance σ_n^2 is estimated from the noisy wavelet coefficients by applying a robust median estimator (Eq. (31)) on the finest scale wavelet coefficients.

$$\sigma_n^2 = \frac{\text{Median}\,(|y_i|)}{0.6745} \tag{31}$$

Bruce and Gao (BRUCE and GAO 1996) showed that hard thresholding tends to have a bigger variance and the soft shrink tends to have a bigger bias. To remedy the drawbacks they introduced a general firm (semisoft) shrinkage function given as:

$$f(w_1) = \begin{cases} 0 & \text{, if } |w_1| \leq \lambda_1 \\ \text{sgn}(w_1)\left(\frac{\lambda_2(|x| - \lambda_2)}{\lambda_2 - \lambda_1}\right) & \text{, if } \lambda_1 < |w_1| \leq \lambda_2 \\ w_1 & \text{, if } |w_1| \leq \lambda_2 \end{cases} \tag{32}$$

In this work, we also adopted the Firm thresholding and defined the λ_1 and λ_2 from the numerator of Eq. (30) as:

$$\lambda_1 = \frac{\sqrt{3}\sigma_n^2}{\sigma} \tag{33}$$

$$\lambda_2 = \sqrt{y_1^2 + y_2^2} \tag{34}$$

5. PROPOSED ALGORITHM

The flowchart shown in Fig. 3 describes the proposed algorithm. At the first step, the noisy speech signal is passed through the pre-processing block, where the signal is normalized.

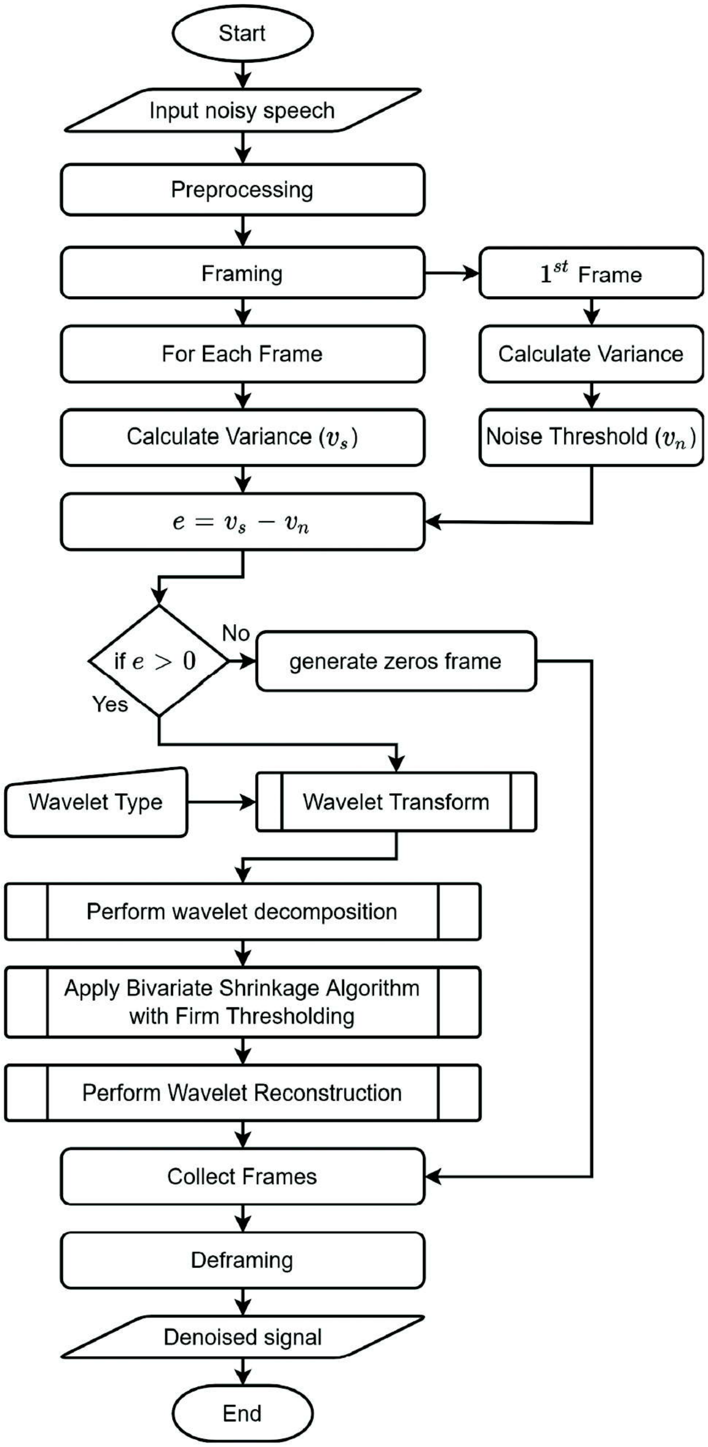


Fig. 3. Flow chart for proposed algorithm.

After pre-processing, the noisy signal is divided into blocks of fixed length samples called "frames". Here, the first frame is considered pure noise (silence before speech), does not contain any speech components, and is used to estimate the noise variance. This estimated noise variance is used to detect the presence of speech in each frame. The frame is considered useful (containing speech components) if its variance exceeds the first frame variance (noise threshold). If the frame contains the speech components, then it is considered for further processing; otherwise, the frame is dropped and a zero-filled frame is directly sent to the output. The noisy speech frame is transformed into a wavelet domain, where two-level decomposition is performed using the selected wavelet function. From the wavelet components firstly the noise variance σ_n^2 is estimated using Eq. (31), and then the λ_1 and λ_2 are calculated using Eq. (33), and Eq. (34). Finally, the denoised wavelet coefficients $(\hat{w}_1)$ are calculated using Eq. (32). Once the denoised coefficients are estimated, the denoised frame is created by taking the inverse wavelet transform of these coefficients. Finally, the denoised frames are converted into a single speech signal stream by applying the de-framing.

6. EXPERIMENTAL ANALYSIS

To test the performance of the proposed algorithm, clean speech signals are taken from the NOIZEUS dataset (Hu and Loizou 2007), whereas the noise signals are acquired from the Noisex92 dataset (Varga and Steeneken 1993). Three different noises White, F16, and Babble, at five different noise levels 0 dB, 2 dB, 5 dB, 7 dB, and 10 dB are considered. The performance of the proposed algorithm is evaluated using four quality measures: Signal to Noise Ratio (SNR), Segmental Signal to Noise Ratio (SSNR), Perceptual Evaluation of Speech Quality (PESQ) (Rix *et al.* 2001), and Short-Time Objective Intelligibility (STOI) (Taal *et al.* 2010), and also compared with the conventional wavelet denoising algorithms. The algorithm is also evaluated for five (db1 or haar, db3, db5, sym5, and coif5 where db, sym, and coif are denoting Daubechies, Symlets, and Coiflets respectively) different types of wavelets.

The experimental results obtained for White, F-16, and Babble noises are presented in Tables 1, 2, and 3 respectively. The results for white noise show that the proposed algorithm performs better for SNR and STOI measures, and the coif5 wavelet gives the best results for both algorithms and all measures.

Table 1. Results for White Noise.

		SNR (dB)									
SNR (dB)	**Noisy**	**DWT**					**Proposed**				
		haar	**db3**	**db5**	**sym5**	**coif5**	**haar**	**db3**	**db5**	**sym5**	**coif5**
0	0	3.08	3.80	3.97	3.77	4.13	4.07	4.23	4.33	4.24	4.36
2	2	3.62	4.58	4.81	4.49	5.07	4.70	4.99	5.12	5.03	5.17
5	5	4.24	5.58	5.87	5.54	6.25	5.45	5.96	6.13	6.02	6.20
7	7	4.59	6.09	6.50	6.06	6.85	5.91	6.49	6.73	6.61	6.83
10	10	4.91	6.53	7.20	6.68	7.58	6.55	7.41	7.71	7.58	7.83
		SSNR (dB)									
SNR (dB)	**Noisy**	**DWT**					**Proposed**				
		haar	**db3**	**db5**	**sym5**	**coif5**	**haar**	**db3**	**db5**	**sym5**	**coif5**
0	0.91	3.16	3.73	3.91	3.83	4.12	3.71	3.80	3.90	3.83	3.95
2	2.01	3.82	4.57	4.80	4.64	5.06	4.23	4.42	4.56	4.48	4.62
5	4.57	4.61	5.67	5.95	5.81	6.35	5.06	5.49	5.68	5.57	5.77
7	6.57	5.13	6.33	6.70	6.51	7.11	5.42	5.90	6.17	6.03	6.29
10	9.57	5.57	6.96	7.51	7.26	7.98	6.13	6.89	7.24	7.09	7.40

	PESQ										
SNR (dB)	Noisy	DWT					Proposed				
		haar	db3	db5	sym5	coif5	haar	db3	db5	sym5	coif5
0	1.58	1.45	1.62	1.75	1.75	1.85	1.67	1.64	1.62	1.65	1.66
2	1.65	1.59	1.85	1.99	1.97	2.08	1.81	1.85	1.83	1.80	1.84
5	1.79	1.59	1.98	2.08	2.05	2.20	1.87	1.94	1.96	1.89	1.95
7	1.91	1.64	2.05	2.15	2.11	2.26	1.96	2.01	2.04	2.00	2.10
10	2.09	1.70	2.12	2.23	2.21	2.35	2.09	2.20	2.24	2.17	2.17

	STOI										
SNR (dB)	Noisy	DWT					Proposed				
		haar	db3	db5	sym5	coif5	haar	db3	db5	sym5	coif5
0	0.68	0.63	0.61	0.59	0.62	0.59	0.69	0.69	0.69	0.67	0.69
2	0.70	0.67	0.66	0.64	0.66	0.64	0.74	0.74	0.74	0.73	0.73
5	0.74	0.74	0.71	0.69	0.70	0.68	0.78	0.79	0.77	0.77	0.77
7	0.77	0.72	0.74	0.71	0.73	0.71	0.79	0.79	0.78	0.78	0.78
10	0.81	0.73	0.76	0.74	0.75	0.74	0.79	0.80	0.79	0.79	0.79

Table 2. Results for F-16 Noise.

	SNR (dB)										
SNR (dB)	Noisy	DWT					Proposed				
		haar	db3	db5	sym5	coif5	haar	db3	db5	sym5	coif5
0	0	1.74	2.26	2.46	2.31	2.57	3.77	4.15	4.16	4.13	4.11
2	2	2.53	3.23	3.48	3.28	3.66	4.78	5.32	5.35	5.29	5.34
5	5	3.57	4.56	4.95	4.65	5.21	6.08	6.91	6.99	6.88	7.08
7	7	4.12	5.25	5.75	5.43	6.14	6.73	7.74	7.86	7.72	7.97
10	10	4.67	6.04	6.69	6.42	7.13	7.31	8.54	8.73	8.61	8.88

	SSNR (dB)										
SNR (dB)	Noisy	DWT					Proposed				
		haar	db3	db5	sym5	coif5	haar	db3	db5	sym5	coif5
0	0.99	1.50	1.83	1.98	1.91	2.13	3.15	3.49	3.45	3.46	3.37
2	2.15	2.44	2.94	3.14	3.05	3.34	4.32	4.81	4.80	4.78	4.77
5	4.63	3.59	4.41	4.77	4.57	5.06	5.77	6.57	6.62	6.56	6.72
7	6.62	4.28	5.22	5.62	5.44	5.98	6.39	7.28	7.38	7.29	7.49
10	9.62	5.11	6.18	6.75	6.60	7.19	7.04	8.15	8.36	8.28	8.88

	PESQ										
SNR (dB)	Noisy	DWT					Proposed				
		haar	db3	db5	sym5	coif5	haar	db3	db5	sym5	coif5
0	1.68	1.68	1.94	1.99	1.96	2.05	1.95	2.02	2.03	2.01	1.99
2	1.79	1.49	1.86	1.89	1.76	1.90	1.68	1.88	1.87	1.82	1.81
5	1.98	1.55	1.96	2.07	1.95	2.12	1.76	1.93	1.94	1.92	1.88
7	2.11	1.75	2.09	2.20	2.12	2.28	2.07	2.22	2.24	2.21	2.16
10	2.31	1.76	2.14	2.26	2.23	2.37	2.20	2.32	2.37	2.34	2.29

	STOI										
SNR (dB)	**Noisy**	**DWT**					**Proposed**				
		haar	**db3**	**db5**	**sym5**	**coif5**	**haar**	**db3**	**db5**	**sym5**	**coif5**
0	0.66	0.59	0.60	0.58	0.60	0.58	0.67	0.69	0.68	0.68	0.67
2	0.68	0.59	0.62	0.61	0.60	0.61	0.65	0.68	0.66	0.67	0.66
5	0.73	0.64	0.66	0.66	0.65	0.66	0.67	0.70	0.69	0.69	0.68
7	0.76	0.69	0.72	0.71	0.70	0.72	0.73	0.75	0.74	0.74	0.73
10	0.80	0.74	0.76	0.75	0.75	0.76	0.76	0.77	0.77	0.77	0.76

Table 3. Results for Babble Noise.

	SNR (dB)										
SNR (dB)	**Noisy**	**DWT**					**Proposed**				
		haar	**db3**	**db5**	**sym5**	**coif5**	**haar**	**db3**	**db5**	**sym5**	**coif5**
0	0	0.37	0.33	0.43	0.39	0.51	1.42	1.11	1.07	1.02	0.93
2	2	1.71	1.88	2.06	2.01	2.21	3.24	3.04	3.02	2.97	2.89
5	5	3.22	3.82	4.14	4.08	4.39	5.43	5.59	5.62	5.56	5.55
7	7	3.98	4.81	5.26	5.21	5.62	6.53	7.01	7.12	7.03	7.10
10	10	4.76	5.90	6.53	6.41	6.98	7.63	8.68	8.91	8.82	9.03

	SSNR (dB)										
SNR (dB)	**Noisy**	**DWT**					**Proposed**				
		haar	**db3**	**db5**	**sym5**	**coif5**	**haar**	**db3**	**db5**	**sym5**	**coif5**
0	1.14	0.99	1.14	1.27	1.21	1.36	1.98	1.86	1.88	1.82	1.78
2	2.43	1.90	2.16	2.35	2.30	2.51	3.34	3.30	3.34	3.27	3.24
5	5.15	3.32	3.84	4.14	4.15	4.42	5.24	5.47	5.57	5.48	5.54
7	7.12	4.19	4.90	5.28	5.31	5.68	6.26	6.79	6.97	6.87	7.03
10	10.12	5.17	6.09	6.62	6.66	7.13	7.31	8.31	8.64	8.53	8.83

	PESQ										
SNR (dB)	**Noisy**	**DWT**					**Proposed**				
		haar	**db3**	**db5**	**sym5**	**coif5**	**haar**	**db3**	**db5**	**sym5**	**coif5**
0	1.79	1.37	1.58	1.66	1.65	1.69	1.68	1.71	1.73	1.73	1.72
2	1.93	1.57	1.80	1.88	1.85	1.91	1.86	1.93	1.95	1.94	1.93
5	2.14	1.59	1.85	1.92	1.91	1.96	1.89	1.97	2.00	1.98	1.97
7	2.27	1.75	2.05	2.13	2.10	2.18	2.08	2.20	2.22	2.21	2.18
10	2.46	1.83	2.13	2.22	2.20	2.27	2.18	2.34	2.36	2.35	2.31

	STOI										
SNR (dB)	**Noisy**	**DWT**					**Proposed**				
		haar	**db3**	**db5**	**sym5**	**coif5**	**haar**	**db3**	**db5**	**sym5**	**coif5**
0	0.71	0.61	0.62	0.63	0.62	0.62	0.65	0.67	0.66	0.67	0.66
2	0.74	0.65	0.66	0.66	0.66	0.65	0.68	0.69	0.69	0.69	0.69
5	0.78	0.68	0.70	0.70	0.71	0.69	0.72	0.73	0.72	0.73	0.72
7	0.80	0.71	0.72	0.73	0.73	0.73	0.74	0.75	0.75	0.75	0.74
10	0.83	0.72	0.75	0.75	0.75	0.75	0.76	0.77	0.77	0.77	0.77

For F-16 noise, the proposed algorithm performs better for SNR, SSNR, and STOI. In all measurements, the db3 gives better results for STOI, whereas the coif5 gives better results for SNR and SSNR.

For Babble noise, the proposed algorithm outperforms the DWT for all measurements. In this scenario, the db5 wavelet performs better for most of the cases. Additionally, the improvements achieved by the proposed algorithm based on different evaluation measures are presented in Table 4.

Table 4. Improvement achieved by the proposed technique at different noises, noise levels and measurements.

Noise	SNR (dB)	Measure	DWT	Proposed	Improvement (%)
White	0	SNR (dB)	4.13 (coif5)	4.36 (coif5)	5.57
White	2	SNR (dB)	5.07 (coif5)	5.17 (coif5)	1.97
White	10	SNR (dB)	7.58 (coif5)	7.83 (coif5)	3.29
White	0	STOI	0.63 (haar)	0.69 (haar, db3, db5, coif5)	9.52
White	2	STOI	0.67 (haar)	0.74 (haar, db3, db5)	10..44
White	5	STOI	0.74 (haar)	0.79 (db3)	6.75
White	7	STOI	0.74 (db3)	0.79 (haar, db3)	6.75
White	10	STOI	0.76 (db3)	0.80 (db3)	5.26
F-16	0	SNR (dB)	2.57 (coif5)	4.16 (db5)	61.86
F-16	2	SNR (dB)	3.66 (coif5)	5.35 (db5)	46.17
F-16	5	SNR (dB)	5.21 (coif5)	7.08 (coif5)	35.89
F-16	7	SNR (dB)	6.14 (coif5)	7.97 (coif5)	29.80
F-16	10	SNR (dB)	7.13 (coif5)	8.88 (coif5)	24.54
F-16	0	SSNR (dB)	2.13 (coif5)	3.49 (db3)	63.84
F-16	2	SSNR (dB)	3.34 (coif5)	4.81 (db3)	44.01
F-16	5	SSNR (dB)	5.06 (coif5)	6.72 (coif5)	32.80
F-16	7	SSNR (dB)	5.98 (coif5)	7.49 (coif5)	25.25
F-16	10	SSNR (dB)	7.19 (coif5)	8.88 (coif5)	23.50
F-16	0	STOI	0.60 (db3, sym5)	0.69 (db3)	15.00
F-16	2	STOI	0.62 (db3)	0.68 (db3)	11.29
F-16	5	STOI	0.66 (db3, db5, coif5)	0.70 (db3)	6.06
F-16	7	STOI	0.72 (db3, coif5)	0.75 (db3)	4.16
F-16	10	STOI	0.76 (db3, coif5)	0.77 (db3)	1.31
Babble	0	SNR (dB)	0.51 (coif5)	1.42 (haar)	178.31
Babble	2	SNR (dB)	2.21 (coif5)	3.24 (haar)	46.60
Babble	5	SNR (dB)	4.39 (coif5)	5.62 (db5)	28.01
Babble	7	SNR (dB)	5.62 (coif5)	7.12 (db5)	26.69
Babble	10	SNR (dB)	6.98 (coif5)	9.03 (coif5)	29.36
Babble	0	SSNR (dB)	1.36 (coif5)	1.98 (haar)	45.58
Babble	2	SSNR (dB)	2.51 (coif5)	3.34 (haar, db5)	33.06
Babble	5	SSNR (dB)	4.42 (coif5)	5.57 (db5)	26.01
Babble	7	SSNR (dB)	5.68 (coif5)	7.03 (coif5)	23.76
Babble	10	SSNR (dB)	7.13 (coif5)	8.83 (coif5)	23.84
Babble	0	PESQ	1.69 (coif5)	1.73 (db5, sym5)	2.36
Babble	2	PESQ	1.91 (coif5)	1.95 (db5)	2.09
Babble	5	PESQ	1.96 (coif5)	2.00 (db5)	2.04
Babble	7	PESQ	2.18 (coif5)	2.22 (db5)	1.83

Conted.......

Babble	10	PESQ	2.27 (coif5)	2.36 (db5)	3.96
Babble	0	STOI	0.63 (db5)	0.67 (db3, sym5)	6.35
Babble	2	STOI	0.66 (db3, db5, sym5)	0.69 (db3, db5, sym5, coif5)	4.54
Babble	5	STOI	0.71 (sym5)	0.73 (db3, sym5)	2.81
Babble	7	STOI	0.73 (db5, sym5, coif5)	0.75 (db3, db5, sym5)	2.73
Babble	10	STOI	0.75 (db3, db5, sym5, coif5)	0.77 (db3, db5, sym5, coif5)	2.67

7. CONCLUSION AND FUTURE SCOPES

In this paper, a Bivariate Shrinkage based speech denoising technique is proposed. The proposed algorithm takes advantage of inter-scale dependencies between the wavelet coefficients to properly estimate the clean speech coefficients. The technique is tested for various types of noises with different strengths and also compared with the conventional DWT-based denoising technique. A comparison of the impact of different wavelet functions is also performed. Several speech quality measures are used to evaluate and compare the performance of the proposed algorithm. Finally, the overall analysis of results is also executed, which shows the better performance of the proposed algorithm in most of the scenarios.

In this work, the denoising is performed by utilizing the inter-scale dependencies of wavelet coefficients using an analytical process that requires an explicit mathematical model for the image and noise. However, in practical conditions, these models deviate significantly, which may result in inferior performance. Therefore, in the future, advanced machine learning algorithms or deep learning techniques can be utilized to obtain inter-scale dependencies of wavelet coefficients under practical conditions using empirical analysis.

8. REFERENCES

[1] Aguiar-Conraria Luís and Maria Joana Soares, 2014. "The Continuous Wavelet Transform: Moving Beyond Uni- and Bivariate Analysis." *Journal of Economic Surveys,* **28**(2), 344-75. https://doi.org/10.1111/joes.12012.

[2] Ahani Soodeh, Shahrokh Ghaemmaghami and Z. Jane Wang, 2015. "A Sparse Representation-Based Wavelet Domain Speech Steganography Method." *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **23**(1), 80-91. https://doi.org/10.1109/TASLP.2014.2372313.

[3] Badiezadegan Shirin and Richard C. Rose, 2015. "A Wavelet-Based Thresholding Approach to Reconstructing Unreliable Spectrogram Components." *Speech Communication,* **67**(March), 129-42. https://doi.org/10.1016/j.specom.2014.11.005.

[4] Bahoura M. and J. Rouat, 2001. "Wavelet Speech Enhancement Based on the Teager Energy Operator." *IEEE Signal Processing Letters,* **8**(1): 10-12. https://doi.org/10.1109/97.889636.

[5] Ben Messaoud, Mohamed anouar, Aïcha Bouzid and Noureddine Ellouze, 2016. "Speech Enhancement Based on Wavelet Packet of an Improved Principal Component Analysis." *Computer Speech & Language,* **35**(January), 58-72. https://doi.org/10.1016/j.csl.2015.06.001.

[6] Bruce Andrew G. and Hong-ye Gao, 1996. "Understanding WaveShrink: Variance and Bias Estimation." *Biometrika,* **83**(4), 727-45. https://doi.org/10.1093/biomet/83.4.727.

[7] Carnero B. and A. Drygajlo, 1999. "Perceptual Speech Coding and Enhancement Using Frame-Synchronized Fast Wavelet Packet Transform Algorithms." *IEEE Transactions on Signal Processing,* **47**(6), 1622-35. https://doi.org/10.1109/78.765133.

[8] Chen Guangyi, Wenfang Xie and Yongjia Zhao, 2013. "Wavelet-Based Denoising: A Brief Review." *In 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP),* pp. 570-74. https://doi.org/10.1109/ICICIP.2013.6568140.

[9] Chen Shi-Huang and Jhing-Fa Wang, 2004. "Speech Enhancement Using Perceptual Wavelet Packet Decomposition and Teager Energy Operator." *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology,* **36**(2), 125-39. https://doi.org/10.1023/B:VLSI.0000015092.19005.62.

[10] Chiluveru Samba Raju and Manoj Tripathy, 2021. "Speech Enhancement Using a Variable Level Decomposition DWT." *National Academy Science Letters,* **44**(3), 239-42. https://doi.org/10.1007/s40009-020-00983-3.

[11] Daqrouq Khaled, Ibrahim N. Abu-Isbeih, Omar Daoud and Emad Khalaf, 2010. "An Investigation of Speech Enhancement Using Wavelet Filtering Method." *International Journal of Speech Technology,* **13**(2): 101-15. https://doi.org/10.1007/s10772-010-9073-1.

[12] Das Nabanita, Sayan Chakraborty, Jyotismita Chaki, Neelamadhab Padhy and Nilanjan Dey. 2020. "Fundamentals, Present and Future Perspectives of Speech Enhancement." *International Journal of Speech Technology,* January. https://doi.org/10.1007/s10772-020-09674-2.

[13] Edwards Tim. 1992. "Discrete Wavelet Transforms: Theory and Implementation."

[14] Fan Linwei, Fan Zhang, Hui Fan and Caiming Zhang, 2019. "Brief Review of Image Denoising Techniques." *Visual Computing for Industry, Biomedicine, and Art,* **2**(1), 7. https://doi.org/10.1186/s42492-019-0016-7.

[15] Gao Hong-ye, Andrew G. Bruce, and Mathsoft Inc., 1997. "Waveshrink with Firm Shrinkage."

[16] Guo Dai-Fei, Wei-Hong Zhu, Zhen-Ming Gao and Jian-Qiang Zhang, 2000. "A Study of Wavelet Thresholding Denoising." In WCC 2000 - ICSP 2000. 2000 5th International Conference on Signal Processing Proceedings. 16th *World Computer Congress* 2000, **1**(1), 329-32. https://doi.org/10.1109/ICOSP.2000.894502.

[17] Hu Yi and Philipos C. Loizou, 2007. "Subjective Comparison and Evaluation of Speech Enhancement Algorithms." *Speech Communication, Speech Enhancement,* **49**(7), 588-601. https://doi.org/10.1016/j.specom.2006.12.006.

[18] Kazerouni Abbas, Ulugbek S. Kamilov, Emrah Bostan and Michael Unser. 2013. "Bayesian Denoising: From MAP to MMSE Using Consistent Cycle Spinning." *IEEE Signal Processing Letters,* **20**(3), 249-52. https://doi.org/10.1109/LSP.2013.2242061.

[19] Mallat Stephane, 1999. A Wavelet Tour of Signal Processing. Elsevier.

[20] Mavaddaty Samira, Seyed Mohammad Ahadi and Sanaz Seyedin, 2017. "Speech Enhancement Using Sparse Dictionary Learning in Wavelet Packet Transform Domain." Computer Speech & Language **44**(July), 22-47. https://doi.org/10.1016/j.csl.2017.01.009.

[21] Muralikrishnan Bala and Jay Raja, eds., 2009. "Filtering in the Time Domain." *In Computational Surface and Roundness Metrology,* London: Springer, pp. 23-31. https://doi.org/10.1007/978-1-84800-297-5_4.

[22] Oktar Mehmet Alper, Mokhtar Nibouche and Yusuf Baltaci, 2016. "Speech Denoising Using Discrete Wavelet Packet Decomposition Technique." In 2016 24th *Signal Processing and Communication Application Conference (SIU),* pp. 817-20. https://doi.org/10.1109/SIU.2016.7495865.

[23] Rix A.W., J.G. Beerends, M.P. Hollier and A.P. Hekstra, 2001. "Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs." *In 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* (Cat. No.01CH37221), **2**(2), 749-52. https://doi.org/10.1109/ICASSP.2001.941023.

[24] Sendur L. and I.W. Selesnick, 2002. "Bivariate Shrinkage Functions for Wavelet-Based Denoising Exploiting Interscale Dependency." *IEEE Transactions on Signal Processing,* **50**(11): 2744-56. https://doi.org/10.1109/TSP.2002.804091.

[25] Seok Jong Won and Keun Sung Bae, 1997. "Speech Enhancement with Reduction of Noise Components in the Wavelet Domain." *In 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing,* **2**(2), 1323-26. https://doi.org/10.1109/ICASSP.1997.596190.

[26] Taal Cees H., Richard C. Hendriks, Richard Heusdens and Jesper Jensen, 2010. "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech." *In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 4214-17. https://doi.org/10.1109/ICASSP.2010.5495701.

[27] Tohidypour Hamid Reza and Seyed Mohammad Ahadi, 2016. "New Features for Speech Enhancement Using Bivariate Shrinkage Based on Redundant Wavelet Filter-Banks." *Computer Speech & Language,* **35**(January), 93-115. https://doi.org/10.1016/j.csl.2015.06.004.

[28] Varga Andrew and Herman J. M. Steeneken, 1993. "Assessment for Automatic Speech Recognition II: NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems." *Speech Communication,* **12**(3), 247-51. https://doi.org/10.1016/0167-6393(93)90095-3.

[29] Zhang Xiao-Ping and M.D. Desai, 1998. "Nonlinear Adaptive Noise Suppression Based on Wavelet Transform." *In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP* '98 (Cat. No.98CH36181), **3**(3), 1589-92. https://doi.org/10.1109/ICASSP.1998.681756.

The Journal of Acoustical Society of India : Vol. 50, No. 1, 2023 (pp. 21-31)

Classifying emotions from instrumental music : A psychological quest using Indian Classical Ragas

Medha Basu[1,2*], Archi Banerjee[2,3], Shankha Sanyal[2,4], Sayan Nag[5], Pijush Kanti Gayen[4], Kumardeb Banerjee[6] and Dipak Ghosh[2]

[1]Department of Physics, Jadavpur University, India

[2]Sir C.V. Raman Centre for Physics and Music, Jadavpur University, India

[3]Rekhi Centre of Excellence for the Science of Happiness, IIT Kharagpur, India

[4]School of Languages and Linguistics, Jadavpur University, India

[5]Department of Medical Biophysics, University of Toronto

[6]Department of Instrumentation and Electronics Engineering, Jadavpur University, India

e-mail: medhabasu1996@gmail.com

[Received: 26-02-2022; Revised: 07-04-2022; Accepted: 15-06-2022]

ABSTRACT

Indian classical music can be broadly classified into two groups, vocal music, and instrumental music. Both these categories have their own characteristic methods of musical presentations. Even though India has a big bank of classical instruments of different kinds, the three main categories are string (for example Sitar, sarod, santoor, violin etc.), wind (ex. flute, sehnai etc.) and percussion (ex. tabla, mridangam, pakhwaj etc.) instruments. Each instrument has its own unique acoustic features and performing techniques. The main pillar of Indian classical music is the Raga system, which are certain definite melodic frameworks. Ragas are believed to play a strong role in evoking emotions of different kinds. In this study, we have tried to analyse the perceptual domain of emotional elicitation in humans evoked by small parts of different Raga renditions played in two most popular classical string instruments, Sitar and Sarod. Live performances by eminent maestros have been taken for both these instruments, and clips of approximately 30 seconds duration have been selected. The chosen 36 clips were broadly divided into four primary emotional categories, namely happy, sad, calm and anxiety (with 9 clips belonging to each emotional category), using inter-coder reliability method taking the help from three experienced musicians of Indian Classical music. Using these Google forms, an audience response survey was conducted with 100 participants. For each of these chosen clips, the participants were asked to mark the appropriate emotions perceived by them from a given set of 11 emotions - Happy, sad, calm, anxiety, surprise, disgust, fear, anger, romantic, devotion and excitement, along with their corresponding intensities in a 5-point Likert scale (where ratings 1 to 5 mean very low, low, moderate, high, and very high intensities respectively corresponding to each emotion). From this pool of audience response, a "discreteness" parameter was evaluated for each of the chosen clips, with the help of which we have tried to observe the phenomenon of co-elicitation of these primary emotions with certain secondary emotions like excitement, romantic, devotion etc. Another parameter "success index" was also evaluated thereafter, to study and compare the levels of elicitation and co-elicitation of emotions in two kinds of audiences, musicians and non-musicians.

1. INTRODUCTION

Music is believed to be one of the strongest art forms to evoke emotions (Juslin 2008, 668) (Juslin 2013, 235-266) (Juslin and Sloboda 2001)(Hunter *et al.,* 2010, 129-164). Right from the evolution of mankind, sounds have been the first mode of communication. For different actions, different kinds of sounds were used. Music in later times has proved to be of great importance in emotional arousal in humans. Juslin and group (Juslin 2005) (Juslin *et al.,* 2010,605 - 642) (Juslin and Vastfjall 2008, 559-575) tried to explain a set of psychological mechanisms from emotional responses to music. Gabrielsson and Lindström (Gabrielsson *et al.,* 2010, 547-574) carried out extensive work regarding the strong musical experiences of different individuals. According to the neuro-cultural theory of Paul Ekman (Ekman *et al.,* 1971, 124), six basic human emotions are associated with separate autonomic activation patterns and facial expressions (Mohn et al., 2011, 503-517). Ekman, Friesen, and Tomkins worked extensively on facial expressions from emotional experiences (Ekman et al.,1971, 37-58). Certain works have also been done regarding the effect of music with and without lyrics (Brattico and Elvira 2011, 308) (Ali *et al.,* 2006, 4), which made the scientific community aware of the ability of instrumental music alone to evoke emotions. Thus, from the works mentioned above, it is well-established that music has a strong power of emotion elicitation. The tempo of a musical piece also affects the emotional response of the audience. A comparative study of this was done by Liu and Ying (Liu and Ying 2018, 2118). But, do emotions always get evoked in a singular manner, or do they show some kind of co-elicitation (Carrera *et al.,* 2007, 422-441) (Brehm *et al.,* 2006, 13-30)? By 'co-elicitation', we refer to the ability of a musical piece to evoke different emotions in a simultaneous manner. Indian Classical Music and its "*raga-rasa*" system has this unique property of evoking a number of emotions simultaneously in a particular rendition (Ramaprasad 2013, 153-156). Each raga in ICM is unique because of the combination of notes it uses, which gives it a distinct emotional fervor, which can be a superposition of a number of emotional states. In this manner, the diaspora of ICM is significantly different from its Western counterpart, as we have the concept of discrete emotions in Western classical music, while a particular raga in ICM can be a superposition of several emotional states. A number of previous studies have used psychometric, statistical as well as neural-network based approach to shed light into the classification of emotions from instrumental ICM clips (Valla *et al.,* 2017) (Roy *et al.,* 2021) (Mathur *et al.,* 2015) (Dandawate *et al.,* 2015, 725-729) (Nag *et al.,* 2022) (Sanyal *et al.,* 2020). If we turn our attention to audio-visual media, i.e., for film clips, emotional arousal can be seen prominently (Schaefer and Alexandre 2010, 1153-1172) (Gross *et al.,* 1995, 87-108) (Bartolini and Ellen 2011) (Gabert and Crystal 2015, 773-787). Gilman and Lee (Gilman and Lee 2017, 2061-2082) have tried to observe this phenomenon of emotional co-elicitation on film clips with the help of two new parameters, 'discreteness'and 'Success index'. In this particular work,we envisage to study if a similar kind of co-elicitation pattern is observed in the case of the instrumental section of Indian Classical music, which is devoid of any rhythmic accompaniment. The first part of the study is based on a detailed human response survey for selected audio clips (around 30 seconds' duration each). To keep the analysis simple, we have worked with the two most popular instruments of the string family of ICM, Sarod and Sitar (Miner and Allyn, 2004), and have studied the elicited emotions in clusters, with the help of detailed statistical algorithms. In the next part, we have tried to extend this study of emotional co-elicitation for musicians and non-musicians, to draw a comparison between emotion appraisals of both the category of participants.

2. EXPERIMENTAL PROCEDURE

As mentioned above, two classical Indian string instruments, Sitar and Sarod have been considered for this study. From different Raga renditions of varied emotions, 36 clips of approximately 30 seconds duration each have been selected with the software Wavepad (Vukovic and Jovana 2008, 451-456) from performances of eminent maestros of Indian classical music. Using these clips, an extensive audience response survey was carried out with 100 participants (M=44, F=56, average age=30.8 yrs, SD=13.83). None of the participants who took part in the survey had any reported hearing disorder or impairments. The survey was designed in the form of Google forms, where a five-point Likert scale rating (Harpe and Spencer, 2015, 836-850) (Joshi and Ankur2015, 396) was used for the emotional intensity rating (very low-

Table 1. Google form design for audience response survey

Clip no.	Very low (I1)-1	Low (I2)-2	Moderate (I3)-3	High (I4)-4	Very high (I5)-5
HAPPY					
SAD					
CALM					
ANXIETY					
SURPRISE					
DISGUST					
EXCITEMENT					
ANGER					
ROMANTIC					
DEVOTION					
FEAR					

1, low-2, moderate-3, high-4, very high-5) and the participants were asked to mark in a set of 11 emotions (happy, sad, anxiety, calm, surprise, disgust, excitement, anger, devotion, romantic, fear) for every clip. An example of response design corresponding to one clip is given below in Table 1.

From the total response, the weighted average intensity was calculated for every clip. From the result of the average intensity values, two new parameters,'discreteness' or 'hit-rate' and 'success-index' of elicitation were introduced to get a detailed understanding of emotional arousal corresponding to each musical clip. From the detailed analysis of all three parameters, clusters of emotions showing close values of appreciable elicitation were studied. From these clusters, we tried to explain the phenomenon of co-elicitation of emotions from string instruments like Sitar and Sarod clips of ICM. To support the phenomenon of co-elicitation in a more mathematical method, ANOVA test (Miller and Rupert 1997) (Girden and Ellen 1992) was carried out on the computed parameter values and studied. These elicitation levels were also studied separately for musicians (N_M=51) and non-musicians (N_{NM}=49) and a comparative picture between both categories was drawn.

3. METHODOLOGY

As discussed above, from the extensive audience response survey, the following parameters were calculated.

2.1 *Average intensity*

Considering a single clip at first, the following procedure was followed to calculate the weighted average intensity. Out of 11 emotions provided (each with 5 intensities), one emotion was taken at a time. The various intensity levels for a particular emotion k were denoted by different values of variable ik (*i.e.* very low is represented by i_k=1, low-2, moderate-3, high-4, and very high-5). The number of participants who marked any particular intensity (i_k) for that particular emotion (k) was noted as nik. This number (n_{ik}) was then multiplied by the weight of the corresponding intensity (i_k). The total intensity value of a particular emotion (k) was obtained by doing a summation over these calculated weighted values for each intensity (i_k). Finally, dividing this total intensity by the total number of responses (N) (combining all emotions) recorded for the clip under consideration, we get the weighted average intensity $(AI)_k$ for the particular emotion (k) for that selected clip. In essence, following mathematical formula was used to calculate weighted average intensity value for each of the 11 emotions provided in the form:

$$(AI)_k = \Sigma_i(i_k n_{ik})/N \tag{1}$$

Where, i_k = Intensity rating for emotion 'k',
i_k € 1,2,3,4,5

k = h, s_1, c, a_1, s_2, d_1, e, a_2, r, d_2, f, which signify each of the emotions under survey:

(h=happy, s_1=sad, c=calm, a_1=anxiety, s_2=surprise, d_1=disgust, e=excitement, a_2=anger, r=romantic, d_2=devotion,f=fear)

$(AI)_k$ = weighted average intensity of emotion 'k' (for any particular clip)

n_{ik} = no. of participants who marked for any intensity i_k of emotion k

N = total no. of participants who responded for the particular clip

Average intensity values for all the 11 emotions were calculated individually by the same method as explained in Eqn. (1). This gave an idea about how strongly different emotions were evoked for that particular clip. This procedure was followed for all the clips to get the total set of average intensity values.

For this study, the emotions 'happy' and 'sad' have been considered as the primary emotions. From the total set of average intensity values for all 36 clips, the clips showing appreciable values of AI *i.e.* ? 0.3 for the primary emotions 'happy' and 'sad' emotions were selected for further detailed study. For the chosen clips, the emotions having values of appreciable elicitation (*i.e.* AI $\geq$ 0.3) under each category of target emotions 'happy' and 'sad' were identified as the secondary or co-elicited emotions. For a better understanding of the co-elicited emotions, the parameter 'discreteness' was calculated for the selected clips, which is discussed in details below.

2.2 Discreteness

The parameter 'discreteness' or 'hit-rate'was used to understand the idea of emotional co-elicitation in a more quantitative manner. This gives a measure of how discretely an emotion has been evoked in a cluster. After calculation of the average intensity values, only one particular emotion was considered. For a single participant, if the intensity of that emotion was ranked at least one point higher than the intensities of all other emotions, then it was given a marking of 1. *Example:* For a single participant response of a clip, discreteness for happiness (d_h) was calculated in the following way-

$$d_h = 1 \text{ only if } i_h > i_k \tag{2}$$

where, i_h = Intensity rating for emotion happy
i € 1,2,3,4,5

i_k = Intensity rating for any emotion other than happy

k = s_1, c, a_1, s_2, d_1, e, a_2, r, d_2, f

(s_1=sad, c=calm, a_1=anxiety, s_2=surprise, d_1=disgust, e=excitement, a_2=anger, r=romantic, d_2=devotion, f=fear)

This method was carried out for all the participants who had responded to that particular clip. The total number obtained by adding all the '1 markings' (the number of participants marking that emotion at least one point higher than the other emotions) was then divided by the total no of participants responding to that clip. This final value obtained was a measure of how discretely the emotion under study here, was evoked amongst all the participants when this particular clip was concerned. The following mathematical expression was used to express the total discreteness value of the emotion happy for a single clip (D_h):

$$D_h = \Sigma(d_h)n/N \tag{3}$$

where, n = 1-100 (participant no.)
$(d_h)n$ = discreteness of happiness for participant no. 'n'

N = total no. of responding participants

This method was carried out for all the emotions for a single clip, and then similarly all other clips were also analyzed. Comparing the discreteness values of different emotions within the clusters of 'happy' and 'sad' separately, the levels of co-elicitation of different emotions were studied.

Next, to get the effect of both average intensity and discreteness, the parameter success index was calculated, and compared for two participant categories - musicians (51) and non-musicians (49). Thus we have an almost 50-50 ratio of the number of musicians and non musicians (the two experimental class of audience) who have participated in the human response survey.

2.3 *Success index*

This parameter takes into account both of the above-mentioned parameters 'average intensity' and 'discreteness', and completely portrays the emotional quotient of any clip. The values of average intensity and discreteness were normalized and given the terms z-centers. The normalized averaged intensity and discreteness values were represented as $(AI_k)_z$ and $(D_k)_z$ respectively in Eqn. (4). These were then added to get the parameter called 'success-index'(S), which was calculated by the formula given below:

$$S_k= (AI_k)_z+ (D_k)_z \tag{4}$$

where, S_k = Success index for emotion 'k'

$(AI_k)_z$ = Normalized total weighted average intensity for emotion 'k'

$(D_k)_z$ = Normalized total discreteness for emotion 'k'

k = h, s_1, c, a1, s_2, d, e, a2, r, d, f, which signify each of the emotions under survey (h=happy, s_1=sad, c=calm, a_1=anxiety, s_2=surprise, d_1=disgust, e=excitement, a_2=anger, r=romantic, d_2=devotion, f=fear)

Success index thus takes into account both the facts that how intensely and discretely an emotion has been evoked. This index was calculated for all the selected clips under emotions 'happy' and 'sad', separately for the categories of musicians and non-musicians, to understand the overall elicitation level of emotions in both categories of participants, as well as their different levels of co-elicitation.

3. RESULTS AND DISCUSSIONS

3.1 *From weighted average intensity values*

As discussed above, from the total survey of 36 clips, the clips showing maximum average intensity rating for emotions 'happy' and 'sad', for both the instruments Sitar and Sarod were identified. For the emotion 'happy', 4 clips of Sitar and 4 clips of Sarod were selected, which showed appreciably high average intensity values (≥0.3) and were marked as the 'happy' clips out of all 36 clips in the survey, or the clips evoking happiness primarily. Similarly, for the emotion 'sad', 4 clips of Sitar and 7 clips of Sarod were selected, which showed appreciably high average intensity values for 'sad' (≥0.3) out of all 36 clips and were marked as the 'sad' clips, or the clips evoking sadness primarily. The mean results for both these kinds of clips are shown below in bar graphs.

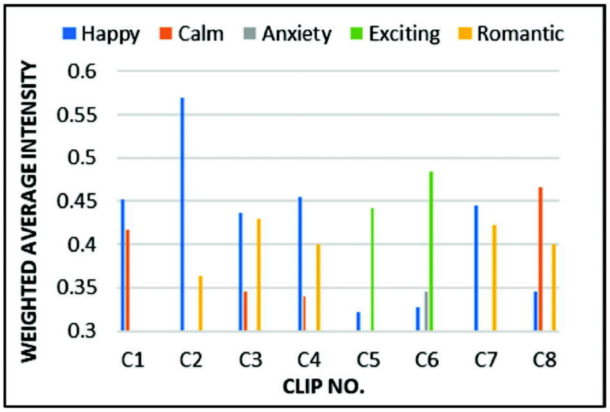


Fig. 1. Plot for Weighted average intensity of 'happy' clip

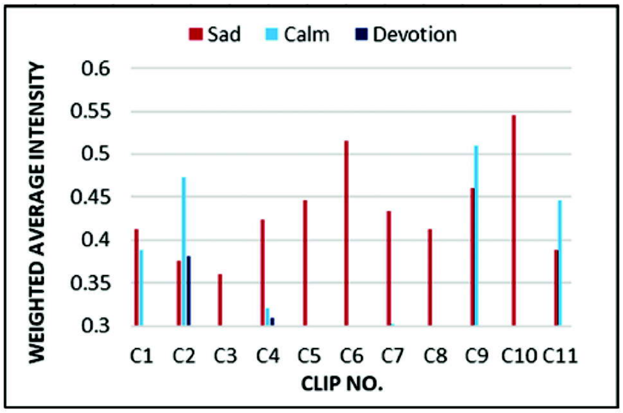


Fig. 2. Plot for Weighted average intensity of 'sad'clip.

Figure 1 shows the total emotional response for the clips belonging to the target class 'happy' of both instruments (C1, C2, C3, C4 for Sitar, and C5, C6, C7, C8 for Sarod) of the same string family. Average intensity values either approximately equal to or greater than 0.3 were considered to correspond to appreciable elicitation of the respective emotion, and thus were taken as the lowest y-axis value for the average intensity plotting. As seen in Fig 1, out of the 8 clips, C1 (average intensity: 0.4514), C2 (0.5697), C3 (0.4364), C4 (0.4545), and C7 (0.4457) show very high elicitation values (>0.4) of happy. Clips C5 (0.3212), C6 (0.3273),and C8 (0.3455) show lower, but appreciable elicitation levels (>0.3) for happy. In clips C3 (0.4303), C4 (0.4), C7 (0.4229), and C8 (0.4), romantic show very high values of elicitation (?0.4). C2 (0.3636) also shows appreciable elicitation value for romantic, but not as high as the other 4 clips. Clips C1 (0.4171) and C8 (0.4667) show very high elicitation values for calm (?0.4) whereas C3 (0.3455), C4 (0.3393) also show moderate elicitation values for calm. Clips C5 (0.4424) and C6 (0.4848) show very high elicitation values for excitement, but this emotion elicitation is not found in any other clips. Clip C6 (0.3455) shows some elicitation of anxiety but as no other clip show this emotion, anxiety has been omitted from the cluster 'happy'. So, from the average intensity values, it can be inferred that happiness being the primary emotion, the secondary co-elicited emotions are found to be romantic, calm, and exciting.

The same study was carried out with the clips which belonged to the target class 'sad', shown in Fig 2, depicting total emotional response for the 'sad' clips of both instruments (C1, C2, C3, C4 for Sitar, and C5, C6, C7, C8, C9, C10, C11 for Sarod).Out of the 11 clips, C1 (0.4121), C4 (0.4242), C5 (0.4457), C6 (0.5151), C7 (0.4343), C8 (0.4121), C9 (0.5091), and C10 (0.5455) show very high elicitation values of sadness (> 0.4), whereas the clips C2 (0.3758), C3 (0.36) and C11 (0.0.3886) show low but appreciable values of elicitation. In clips C2 (0.4727), C9 (0.5091), and C11 (0.4457), calm show very high values of elicitation (?0.4), and a low but appreciable elicitation level of calm is seen in C1 (0.3879). Clips C2 (0.3818) and C4 (0.3091) show some elicitation of devotion. So, it can be concluded, that sadness being the primary emotion, the secondary co-elicited emotions are found to be calm and devotion.

As is evident from the average intensity values, the emotions mostly co-elicited along with happy are found to be romantic, calm, and exciting and the emotions mostly co-elicited along with sad are found to be calm and devotion. To understand this phenomenon in a more quantitative manner, the next parameter discreteness was studied for both the 'happy' and 'sad' clusters.

3.2 *From discreteness values*

For each of the emotions under these two emotion clusters, discreteness was computed, which is shown below in two scatter plots.

Fig 3 shows the discreteness values of the emotions happy, romantic, calm and exciting. The lowest value for the y-axis was taken 0.3, to only plot the emotions with significant elicitation. As can be seen

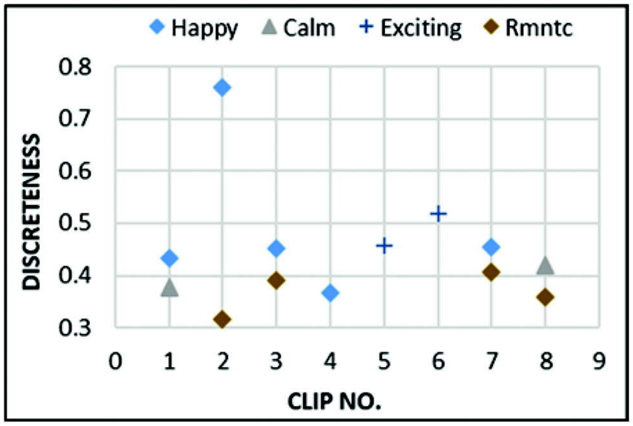


Fig. 3. Plot for Discreteness values of 'happy' cluster

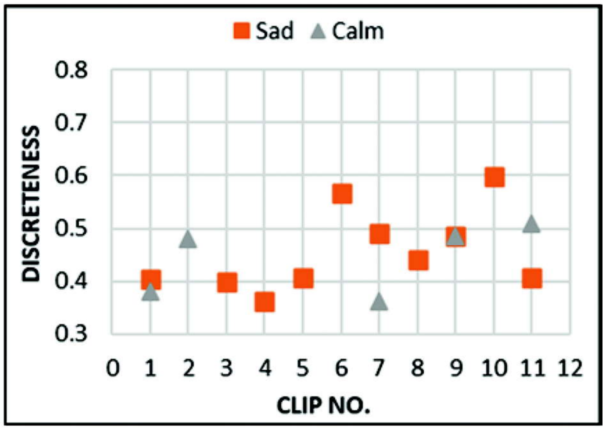


Fig. 4. Plot of Discreteness values of 'sad' cluster

from the plot, happy is the most elicited emotion (5 points: C1, C2, C3, C4, C7), followed by romantic (4 points: C2, C3, C7, C8), calm (2 points: C1, C8) and exciting (2 points: C5, C6). Fig 4 shows the discreteness values of the emotions sad, calm and devotion, which were plotted following the same procedure explained above. As can be seen, the most elicited emotion is sad (10 points: C1, C3, C4, C5, C6, C7, C8, C9, C10, C11) followed by calm (5 points: C1, C2, C7, C9, C11). Devotion is not found as a plot point for any of the clips, which indicates its low co-elicitation level with sad. To statistically study the findings from discreteness values and also to understand the level of co-elicitations of different emotions within a cluster, the ANOVA test was applied to the values, with a significance level of p=0.05. For the ANOVA test, the discreteness values for the closely elicited emotions (as found in the sections discussed above) were taken as the input parameters. For both 'happy' and 'sad' clusters, the ANOVA test results between the co-elicited emotions are shown below.

Table 2a. 'Happy' cluster ANOVA.

Emotion pair	F value	p-value	Significance
Happy-Romantic	3.0966	0.1003	0
Happy-Calm	7.5930	0.0155	1
Happy-Exciting	5.3744	0.0361	1

Table 2b. 'Sad' cluster ANOVA.

Emotion pair	F value	p-value	Significance
Sad-Calm	7.1480	0.0146	1
Sad-Devotion	54.5891	0.0001	1

Table 2a shows the ANOVA values for all the emotions within the cluster 'happy'. It is seen that other than romantic, all other emotions (calm and exciting) within the cluster show significance value 1 with happy, *i.e.,* can be considered to have significantly different elicitation level from happy. So, it can be inferred that the level of co-elicitation of romantic with happy is highest. Table 2b shows the ANOVA values for all the emotions within the cluster 'sad'. Here, all emotions within-cluster (calm and devotion) show significance value 1 with sad. Thus, for the 'sad' clips, the phenomenon of co-elicitation is much lower and sadness thus appears to be a more dominant emotion than happiness.

3.3 *From success index values*

In the next level, another parameter 'success index' was studied. Adding the normalized values of both average intensity and discreteness, the success indices of the emotions were computed. Success index takes into account the effect of both the parameters - 'average intensity' and 'discreteness', *i.e.* for a particular clip, this parameter considers both how intensely and how discretely a particular emotion is evoked within a groups of emotions provided. Using this feature, we look to categorize and qualify the emotional response obtained from similar clips across the two categories of audience who have participated in this study - musicians and non-musicians. Since we have an almost equal distribution of musicians and non-musicians, SI proves to be a very interesting parameter to study the differential perceptual response of emotions for the two categories of participants. The success indices for the happy and sad clips are represented below in two scatter plots, separately for musicians and non-musicians, to draw a comparison between elicitation and co-elicitation levels of emotions in both categories of audiences.

The lowest value for the y-axis was taken as 0.8, to only plot the success indices with appreciable values, which are high enough to contribute to the total emotional appraisal in audience. As can be seen from the plots in Fig 5a and Fig 5b, musicians show much greater no of plot points of success index as

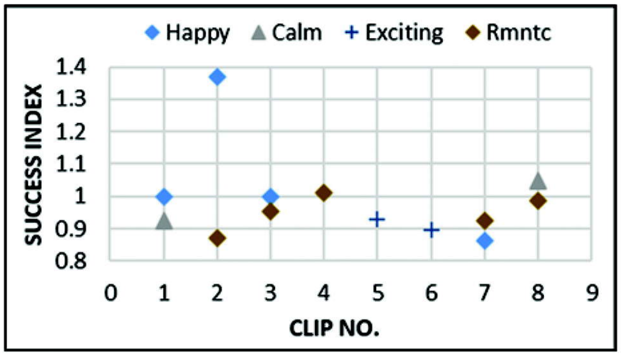


Fig. 5a. Plot for success index values for 'happy' clips in musicians

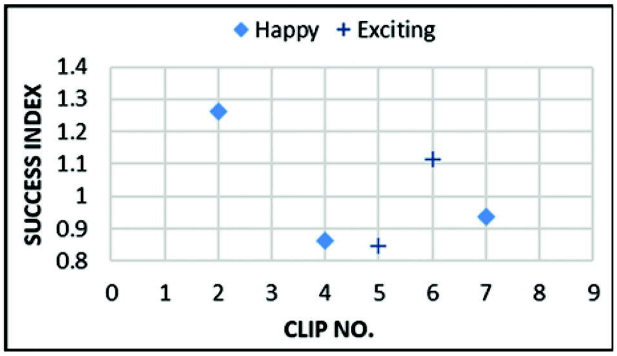


Fig. 5b. Plot for success index values for 'happy' clips in non-musicians

compared to non-musicians. The plots for musicians include total 13 points, ranging over points for emotions happy (4), romantic (5), calm (2), as well as exciting (2). Plots for non-musicians on the other hand show only 5 plot points, consisting of only three for happy and two for exciting. In the plot for musicians, 13 points show values either equal or greater than 0.8, but in the plot for non-musicians, only 5 points show values equal or greater than 0.8. So, it can be concluded that, the overall level of emotional elicitation is much more in musicians than in non-musicians, and also the phenomenon of emotional co-elicitation is much more prominently observed in musicians.

Next, the same technique was applied for clips belonging to target class 'sad', shown in the scatter plots below.

The plots Fig 6a and Fig 6b were drawn following the same method as followed in happy clips, taking the same lowest y-axis value. It is seen that musicians show much greater no of plot points of success index with respect to non-musicians. The plots for musicians include total of 15 points, ranging over points for sad (9), calm (3), and devotion (3). Plots for non-musicians on the other hand show only 7 plot points, consisting of only 3 for sad and 4 for calm. In the plot for musicians, 15 points show values either equal to or greater than 0.8. But in the plot for non-musicians, only 7 points show values equal to or greater than 0.8. Another interesting observation is that for the target class 'sad', non-musicians' SI for the emotion 'calm' is greater than that of emotion 'sad'. The same is not true for musicians, where the SI of target class 'sad' is much higher (both in number of clips as well as in quantitative value) than other co-elicited emotions 'calm' and 'devotion'. Again, it can be concluded that both the overall level of emotional elicitation, and the levels of co-elicitation are much more in musicians than in non-musicians.

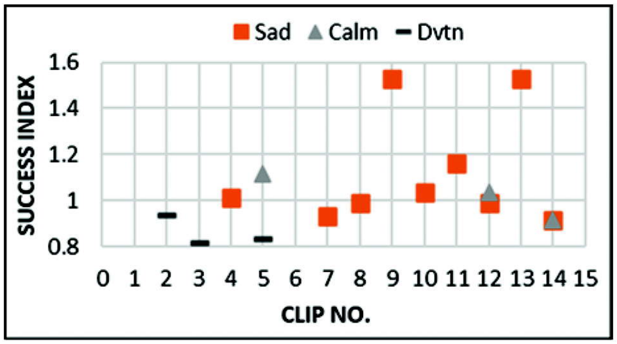


Fig. 6a. Plot for success index values for 'sad' clips in musicians

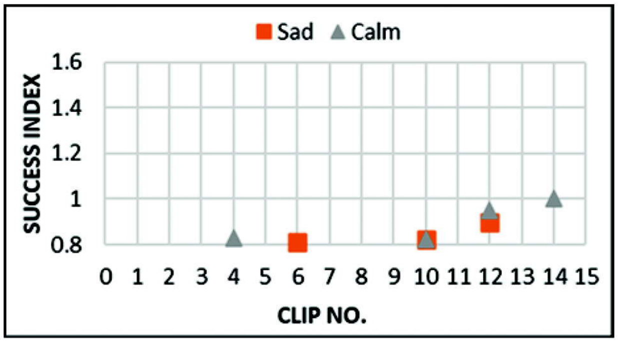


Fig. 6b. Plot for success index values for 'sad' clips in non-musicians

4. CONCLUSION

In the present study, an extensive audience response survey was carried out on several clips of two instruments of the string family-Sitar and Sarod. From the analysis of parameters like average intensity, discreteness, and success index, we have tried to understand the phenomenon of co-elicitation of emotions in Indian Classical Music. The responses of musicians and non-musicians have been studied separately which have led to the following conclusions:

- The emotions mostly co-elicited along with happy are romantic, calm, and exciting, with romantic having the highest level of co-elicitation.
- The emotions mostly co-elicited along with sad are calm and devotion. But levels of co-elicitation of these emotions are low as sadness as a discrete emotion was found to have high dominance over the other emotions.
- The overall emotional arousal/ elicitation levels were found to be much higher in musicians than that in the non-musician category.
- Different emotions which were co-elicited for musicians were dropped for non musicians' response. For *e.g.,* 'devotion' was co-elicited with 'sad' for musicians and was dropped for non musicians. Similarly 'calm' and 'romantic' were co-elicited with 'happy' for musicians, but both were absent for non-musicians.Thus, the extent of co-elicitation or simultaneous arousal of emotional activities were also found to be higher in musicians than in non-musicians.

5. SCOPE OF FUTURE WORK

This study can be extended to other families of Indian instruments like flute and sehnai (wind instruments) along with a comparative study of the co-elicitation based emotional parameters from both string and wind family of instruments can be carried out. This study can be made even more exhaustive, by comparing emotional appraisals for different instruments within the same family, e.g. - comparison between Sitar, Sarod, Santoor and Violin. Vocal music also would form a large section of analysis and would add a completely new dimension to the whole study. This domain, if explored in details might lead to a new algorithm connecting different aspects of Indian Classical Music, to different primary and secondary layers of emotional elicitations and co-elicitations. Furthermore, the neuro-cognitive manifestations of these co-elicitation based effects in musicians and non-musicians would also be an interesting case study. This pilot study would act as a precursor to these frontier explorations in the nascent domain of Indian Classical Music.

6. ACKNOWLEDGEMENT

MB (Ref No. - D-7/Sc/231/21; Index No. - 54/21/Phys./27) acknowledges the Department of Science and Technology (DST), Govt. of India, for providing the DST Inspire Fellowship for this study.AB (SR/CSRI/PDF-34/2018) and SS (DST/CSRI/2018/78) acknowledge DST, Govt. of India for providing the DST-CSRI (Cognitive Science Research Initiative) Post-Doctoral Fellowship to pursue this research work.

7. REFERENCES

[1] Ali S. Omar and Zehra F. Peynircioğlu, 2006. "Songs and emotions: are lyrics and melodies equal partners?" *Psychology of music,* **34**(4), 511-534.https://doi.org/10.1177/0305735606067168

[2] Bartolini Ellen Elizabeth, 2011. "Eliciting emotion with film: Development of a stimulus set.".

[3] Brattico Elvira, VinooAlluri, Brigitte Bogert, Thomas Jacobsen, Nuutti Vartiainen, Sirke Nieminen, and Mari Tervaniemi, 2011. "A functional MRI study of happy and sad emotions in music with and without lyrics." *Frontiers in psychology,* **2,** 308.https://doi.org/10.3389/fpsyg.2011.00308

[4] Brehm Jack W. and Anca M. Miron, 2006. "Can the simultaneous experience of opposing emotions really occur?" *Motivation and Emotion,* **30**(1), 13-30.

[5] Carrera Pilar and Luis Oceja, 2007. "Drawing mixed emotions: Sequential or simultaneous experiences?" *Cognition and emotion,* **21**(2), 422-441.https://doi.org/10.1080/02699930600557904.

[6] Dandawate Yogesh H., PrabhaKumari and Anagha Bidkar, 2015. "Indian instrumental music: Raga analysis and classification." In 2015 1st international conference on next generation computing technologies (NGCT), *IEEE,* pp. 725-729. 10.1109/NGCT.2015.7375216.

[7] Ekman Paul and Wallace V. Friesen, 1971. "Constants across cultures in the face and emotion." *Journal of personality and social psychology,* **17**(2). https://doi.org/10.1037/h0030377.

[8] Ekman Paul, Wallace V. Friesen and Silvan S. Tomkins, 1971. "Facial affect scoring technique: A first validity study." pp. 37-58. https://doi.org/10.1515/semi.1971.3.1.37

[9] Gabert-Quillen, Crystal A., Ellen E. Bartolini, Benjamin T. Abravanel and Charles A. Sanislow, 2015. "Ratings for emotion film clips." *Behavior research methods,* **47**(3), 773-787.

[10] Gabrielsson Alf and Siv Lindström, 2010. "Strong experiences with music." *Handbook of music and emotion: Theory, research, applications,* pp. 547-574.

[11] Gilman T. Lee, Razan Shaheen, K. Maria Nylocks, Danielle Halachoff, Jessica Chapman, Jessica J. Flynn, Lindsey M. Matt and Karin G. Coifman, 2017. "A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation." *Behavior research methods,* **49**(6), 2061-2082.

[12] Girden Ellen R., 1992. ANOVA: Repeated measures. No. 84. sage.

[13] Gross James J. and Robert W. Levenson, 1995. "Emotion elicitation using films." *Cognition & emotion* **9**(1) 87-108.https://doi.org/10.1080/02699939508408966

[14] Harpe Spencer E., 2015. "How to analyze Likert and other rating scale data." Currents in pharmacy teaching and learning, **7**(6), 836-850.https://doi.org/10.1016/j.cptl.2015.08.001

[15] Jones Mari Riess, 2010. "Music perception: Current research and future directions." *Music perception* pp. 1-12.

[16] Joshi Ankur, Saket Kale, Satish Chandel and D. Kumar Pal, 2015. "Likert scale: Explored and explained." *British journal of applied science & technology,* **7**(4), 396.DOI: 10.9734/BJAST/2015/14975

[17] Juslin P.N., 2005. "How does music arouse emotions." *In Thirteenth Conference of the International Society for Research on Emotions.* Bari, Italy.

[18] Juslin Patrik N., Simon Liljeström, Daniel Västfjäll and Lars-OlovLundqvist, 2010. "How does music evoke emotions? *Exploring the underlying mechanisms.*"

[19] Juslin Patrik N., Simon Liljeström, Daniel Västfjäll, Gonçalo Barradas and Ana Silva, 2008. "An experience sampling study of emotional reactions to music: listener, music, and situation." *Emotion,* **8**(5), 668. https://doi.org/10.1037/a0013505

[20] Juslin Patrik N. and Daniel Västfjäll, 2008. "Emotional responses to music: The need to consider underlying mechanisms." *Behavioral and brain sciences,* **31**(5), 559-575. doi:10.1017/S0140525X08005293

[21] Juslin Patrik N., 2013. "From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions." *Physics of life reviews,* **10**(3), 235-266. https://doi.org/10.1016/j.plrev.2013.05.008

[22] Juslin Patrik N. and John A. Sloboda, 2001. "Music and emotion." *Theory and research.*

[23] Liu Ying, Guangyuan Liu, Dongtao Wei, Qiang Li, Guangjie Yuan, Shifu Wu, Gaoyuan Wang and Xingcong Zhao, 2018. "Effects of musical tempo on musicians' and non-musicians' emotional experience when listening to music." *Frontiers in Psychology,* **9,** 2118. https://doi.org/10.3389/fpsyg.2018.02118

[24] Mathur Avantika, Suhas H. Vijayakumar, Bhismadev Chakrabarti and Nandini C. Singh, 2015. "Emotional responses to Hindustani raga music: the role of musical structure." *Frontiers in psychology,* **6,** 513. https://doi.org/10.3389/fpsyg.2015.00513

[25] Miller Jr and Rupert G. Beyond, 1997. ANOVA: basics of applied statistics. *CRC press.*

[26] Miner Allyn. Sitar and Sarod, 2004. In the 18th and 19th Centuries. *Motilal Banarsidass Publ.,* **7**.

[27] Mohn Christine, Heike Argstatter and Friedrich-Wilhelm Wilker, 2011. "Perception of six basic emotions in music." *Psychology of Music,* **39**(4), 503-517.https://doi.org/10.1177/0305735610378183

[28] Nag Sayan, Medha Basu, Shankha Sanyal, Archi Banerjee and Dipak Ghosh, 2022. "On the application of deep learning and multifractal techniques to classify emotions and instruments using Indian Classical Music." *Physica A: Statistical Mechanics and its Applications,* **597,** 127261.

[29] Ramaprasad Dharitri, 2013. "Emotions: an Indian perspective." *Indian Journal of Psychiatry,* **55**(2), S153.10.4103/0019-5545.105514

[30] Roy S., A. Banerjee, S. Sanyal, D. Ghosh and R. Sengupta, 2021. "A study on Raga characterization in Indian classical music in the light of MB and BE distribution." *In Journal of Physics: Conference Series,* **1896**(1), 012007. *IOP Publishing*.

[31] Sanyal Shankha, Archi Banerjee, Medha Basu, Sayan Nag, Dipak Ghosh and Samir Karmakar, 2020. "Do musical notes correlate with emotions? A neuro-acoustical study with Indian classical music." In Proceedings of Meetings on Acoustics 179ASA, *Acoustical Society of America*, **42**(1), 035005. https://doi.org/10.1121/2.0001397.

[32] Schaefer Alexandre, Frédéric Nils, Xavier Sanchez and Pierre Philippot, 2010. "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers." *Cognition and emotion,* **24**(7), 1153-1172. https://doi.org/10.1080/02699930903274322.

[33] Valla Jeffrey M., Jacob A. Alappatt, Avantika Mathur and Nandini C. Singh, 2017. "Music and emotion-a case for north indian classical music." *Frontiers in psychology,* **8,** 2115. https://doi.org/10.3389/fpsyg.2017.02115.

[34] Vukovic Jovana, David R. Feinberg, Benedict C. Jones, Lisa M. DeBruine, Lisa LM Welling, Anthony C. Little and Finlay G. Smith, 2008. "Self-rated attractiveness predicts individual differences in women's preferences for masculine men's voices." *Personality and Individual Differences,* **45**(6), 451-456.https://doi.org/10.1016/j.paid.2008.05.013

The Journal of Acoustical Society of India : Vol. 50, No. 1, 2023 (pp. 32-46)

Timbre space of Esraj: A comparison with violin

Anirban Patranabis*, Kaushik Banerjee, Tarit Guhathakurta,
Ranjan sengupta and Dipak Ghosh
Sir C.V. Raman Centre for Physics and Music, Jadavpur University, Kolkata-700 032, India
e-mail: anir.thikana@gmail.com

[Received: 01-03-2022; Revised: 05-05-2022; Accepted: 31-05-2022]

ABSTRACT

Esraj, a sweet and pleasant sounded Indian musical instrument (MI) is under Indian bowing category. It can produce a variety of sounds depending on the size of the instrument and gauge of strings. Unfortunately, it never became as popular as Violin. As per size and shape, there are a few types of Esraj available and its timbre quality also differs from one to another. Sound of some high pitch Esraj are generally perceived similar with violin. Using timbre parameters, this study attempts to define the Esraj's sound quality and compare with that of violin sounds. We have four sound signals each of four different strings from two different Esraj and also two different Violins. We have observed that they have different timbre space and produces different sound quality. The tone of Violin is of much better quality than the tone of Esraj.

1. INTRODUCTION

Esraj, an Indian form of chordophones and bowed string musical instrument (MI) is not as old as sitar and sarod. Since 19th century the practice and development of Esraj in Indian musical scenario; both in semi-classical and commercial purpose was flourished. According to its size, shape (2-3 shapes are available) and gauge of strings, Esraj can produce variety of sweet sounds. It was mostly popular in the eastern part of the country, namely: Darvanga, Gaya and specially Kolkata and Santiniketan. In North India it was mainly used for accompanying in religious songs. At Santiniketan it was very popular only because of Rabindranath Tagore, who introduced Esraj in Visva Bharati University. He introduced Esraj mainly for the accompaniment of Rabindra sangeet. During late 19th century to early 20th century the culture of Esraj actually flourished. But somehow after that, the trend of using Esraj for accompanying Rabindrasangeet started declining. Now Santiniketan is the only famous centre of this sweet sounded MI [Banerjee Kaushik, 2017]. This musical instrument has some uniqueness: easy to play, very sweet and pleasant to hear and can produce variety of sounds depending on the size of the instrument and gauge of strings but unfortunately it never got much popularity like sitar, sarod. The violin is a string instrument which has four strings and is played with a bow. The strings are usually tuned to the notes G, D, A, and E. Though Violin, a foreign bowing instrument won the heart of Indian music lovers but Esraj failed to do so. This striking feature increased our inquisitiveness to study the timbral quality and spectral features of these two MI.

Basic comparison between esraj and violin are as follows. Esraj has a long fingerboard with frets while violin has a fretless fingerboard. Esraj has parched goat skin sound box while sound box of violin is made of pine. Other than these differences it is also noticeable that - neck of Esraj is hollow like other Indian string musical instruments. Whereas like other western string musical instruments neck of the violin is solid. Bridge of esraj is either wooden made or made of bones or dear's horn. Bridge of violin is made of maple wood. Sound post is observed in violin but not in esraj. Bow size of violin is more than that of esraj. Violin strings are made from a variety of materials including catgut (sheep intestine), nylon, and steel while main string of esraj is made of steel and other strings are made of steel or brass. The main structure of esraj is made of toon or segun. Main structure of violin is made of spruce, willow, maple, ebony and rosewood.

All the measurements and tuning of violin are fixed and standard for a long period of time; whereas, in case of Esraj there are not so many verities available. For our experiment we took the two mostly accepted versions among those Esrajs. One is of standard Indian scale that tunes as - F# (M□a), C# (S□a), C# (S□a), G# (Pa□) or F# (M□a), C# (S□a), G# (Pa□), C# (S□a) etc. Indian musicians usually mention tonic note or 'Sa' as 'scale', because tonic note or 'Sa' is the key note. On the foundation that is 'Sa', other notes are determined by relative ratios. So here the scale is said as C#. This Esraj uses for both classical and accompaniment (normally for Rabindra Sangeet) purpose. On the contrary, high pitch Esraj is only uses for accompaniment (normally for Rabindra Sangeet and preferably female voices). Though tuning patterns are same but notes are of higher octave. Such as: Ma (D#), Sa (A#), Pa (F), Sa (A#) etc.

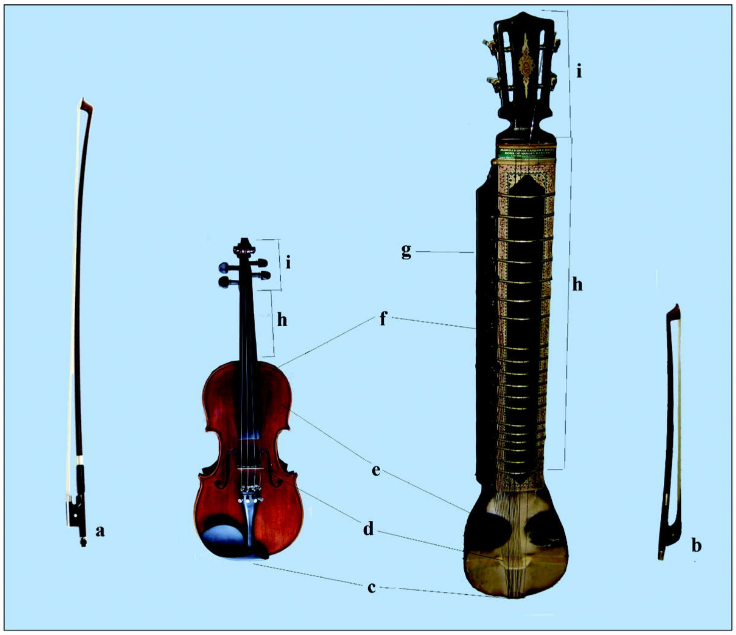


Fig. 1. (a) violin bow, (b) Esraj bow, (c) tail peace, (d) bridge, (e) belly, (f) fingerboard (violin) & fret board (Esraj), (g) nistir kath or sympathetic peg's panel, (h) neck, (i) peg box.

Although the playing technique mainly bowing and fingering technique of these two instruments are similar, sound quality is different. As a result, esraj failed to gain popularity among the listeners as well as musicians but violin is highly popular to the mass. This is only due to the difference in sound quality or difference in timbre. This striking feature leads to encourage this comparative study of the sound signals of these two musical instruments. So the objective of this research work is to assess the timbral and spectral quality of Esraj and Violin sounds and then to compare the timbral and spectral quality of Esraj with Indian made violin. By analyzing the signals from both instruments, this study attempts to define the Esraj's sound quality and compare with that of violin sounds. Physical characteristics and resonance patterns was also focused to study their complete timbre space. The most important thing of this paper is that this is comparing two musical instruments of two different cultures: violin - is basically western musical instrument; whereas, Esraj is an Indian musical instrument. but both are bowing musical instrument. Although Esraj and Violin belongs to bowing instruments, we have observed that they have different timbre space and produce different sound quality.

This study includes a brief historical background about the musical instrument esraj. Here we have ignored the historical background of violin as it is well known and highly popular musical instruments. Lots of research work has been done which studied the sound quality of violin [Houssay A. M. H., 2007 and Hutchins C. M. A., 1983]. This is followed by methodology which describes the procedure of finding the timbre space and to compare that for the two instruments. Next section consists of the result and discussion. Here we compare the two instruments from the spectrogram and LTAS view of the sound signals of each strings of both the instruments. This also includes the comparative study of the sound signals of the two instruments based on spectral irregularities and harmonic shift. Lastly the timbre space of the two instruments is compared. Finally, a conclusive remark of rich violin sound over esraj is made.

2. PREVIOUS WORK

Many studies for stringed instruments exist. Most analyses have been made on violins, guitars, and pianos, but comparative studies on these and less popular instruments have not been done. A lot of research work has been done on the feature recognition of musical instruments but a very few work has been done on the comparative study of the sound quality of musical instruments. In the journal article [Sengupta R. *et al.,* 2003] Comparative Studies of Musical Instruments, by Michael Ramey where the author had made a comparative study on the manifold music cultures of the world in terms of the music styles involved as well as the socio-cultural context. In the paper "A comparative study of stringed instruments" by Gordon Ramsey [Sengupta R. *et al.,* 2007] made an in depth analysis of a variety of stringed instruments. The experiments included spectrum analysis, body resonances using pattern analysis and high-speed videos to visually observe the string oscillation modes. The spectral analysis was done on all instruments with the strings plucked or picked at different locations. Corresponding high-speed video was taken on many to observe how the waves propagate along the string. String resonances were compared to the body resonances to see the synthesis between the instruments. Comparisons of the spectrum, body resonances, and string oscillations have been made between these instruments to gain a better understanding of how they operate and why each emits the unique sounds that it does. No research work has been done so far to compare two similar type (according to playing technique) viz violin and esraj (a less popular musical instrument). No study has been made on the comparison of the sound quality of esraj and violin sounds based on timbre analysis.

3. HISTORICAL BACKGROUND ABOUT ESRAJ

Esraj is comparatively not much older musical instrument in respect of other bowing musical instruments like Sarengi. Its nearest ancestor was Mauri or Mauryvina which is now almost obsolete musical instrument. The life of Esraj is not more than 250 to 350 years. And it had never been highly popular musical instruments in respect of sitar or sarod. It is said that - Esraj had been created with the combination of sitar and Sarengi [Banarsidas M, 1997]. But it is also true for Dilruba. However, there were some other variations of Esraj like - Taus, Tar-Sahenai etc.[Banarsidas M, 1997] but those are not much different from Esraj. For instance - Taus is an enlarged Esraj whereas, Tar-Sahenai is smaller than Esraj, only a small funnel like gramophone speaker attached on the bottom of the belly. C.R. Day described Taus or Esrar, sometimes this instrument is also called Mohur; [Day C. R., 1974] that is similar to Mauri or Mauryvina of other contemporary interpretation [Ghosh L. N., 1975]. Basically, other than Sarengi and to some extent Dilruba, all the above mentioned bowing musical instruments are nearly same in shape but the sizes and string arrangements are different. Obviously the sound quality or timbres have to be different.

Practically, though it was nurtured by several learned and intellectual personalities including Tagore family members of Kolkata (then it was Calcutta), but still now it has not been standardized. Perhaps easy playing technique of Esraj made it popular among the ladies of the elite society of Kolkata during late 19th century. Though musicians of other parts of India also nurtured this musical instrument but most of the evolutions were happened in the eastern part of the country. Kanailal Dhendi of Gaya and his disciple Chondrika Prasad Dubey of Darbhanga were some of the legendary persons of 19th and early 20th century's Esraj playing [Mukhopadhyay D., 1976], [Ramsey G., 2014]. Though both of them were from Gaya Gharana (Hanumandasji) but they were familiar in Kolkata musical arena. Since early days, Vishnupur Gharana has had some major contribution in Esraj playing [Mukhopadhyay D., 1976]. Along with Dhrupad singing and sitar playing, they also used to play Esraj; Anantala Bandopadhyay, Gopeshwar Bandopadhyay, Surendranath Bandopadhyay were some of them [Tagore S. M., 1983]. Perhaps, due to some limitations this bowing musical instrument could not able to be a major attraction of the then music lovers. As a result, its popularity declined day by day. Fortunately, there was Rabinranath Tagore who played the major role in the patronization of this sweet sounded simple musical instrument. He introduced Esraj as the accompanying musical instrument for his songs in Visva-Bharati [RoyChowdhury H. K., 1929]. Disciple and nephew of Surendranath Bandopadhyay, Ashesh Bandopadhyay was one of the early teachers of Sangeet Bhaban (Visva-Bharati) who did some experiments with the inspiration from Surendranath Bandopadhyay. He started playing it in perpendicular position; while, earlier, Esraj was played in inclined position. Later, Ranadhir Roy disciple of Ashesh Bandopadhyay improvised the Esraj which is now popular to the world of music lovers [RoyChowdhury H. K., 1929].

Hence, to observe all possibilities we took two different sets of sound signals of two different Esrajs; classical Esraj - that is comparatively larger in size and high pitch Esraj, which is smaller in size and use for accompaniments. Violin signals are also taken from a good quality Italian violin. The basic tuning difference is that Indian instruments (here Esraj) are tune in relative (ratio) pitch whereas, violin and other western musical instruments tune in fixed pitch. For instance, if the four strings of classical Esraj are tuned as: Ma (4th) - Sa (tonic) - Pa (5th) - Sa (bass), then Sa or tonic maybe C, C# or D. in respect of tonic Ma and Pa would be F, F# or G (Ma) and G, G# or A (Pa). But along with its fixed shape and size all violins are usually tuned in G (Ma), D (Sa), A (Pa), E (Re).

4. METHODOLOGY

Since we are concentrating on the sound quality of the musical instruments, we have recorded the sound of open strings of well- tuned instruments i.e. the sound of the four main strings. Instead of comparing all 12 or 24 notes, it is sufficient to analyse only 4 notes (only the open strings) to distinguish the timbral quality of the musical instruments. These open string sounds are sufficient to study the source characteristics of the instrument. Sound of each notes may be considered as fundamental for that particular string. Indian string musical instruments are generally tunes in three main notes (Indian), namely - Sa or Sardj (tonic), Ma or Madhyam (4th from tonic), Pa or Pancham (5th from tonic). In this research work we have considered eight sound signals from two different Esraj and also eight signals from two different Violins. Since esraj become so unpopular among the musicians, it is very difficult to get esraj player presently. Also only two types of esraj are recognized and widely accepted by the renowned musicians and scholars. So we considered two recognized esraj. So we stick to two instruments only. Moreover, since we are aimed to study the source characteristics of the instrument and then to compare with violin, it is sufficient to make the comparative study with four instruments only.

One of the Esraj was tuned at high pitch and the sound signals of four open strings comprising the notes Ma (D#), Sa (A#), Pa (F), Sa (A#). The other Esraj was tuned at normal pitch F# (M?a), C# (S?a), C# (S?a), G# (Pa?) or F# (M?a), C# (S?a), G# (Pa?), C# (S?a). The sound signals of four open strings of two Violins comprising the notes E, A, D and G were considered. Tuning of violin is fixed pitched and standard but earaj is tuned with varied pitch.

All the sixteen signals were monophonic and were digitized at a sampling rate of 44100Hz (16 bits per sample) in mono channel. Each signal space was divided into non-overlapping windows of 1024 sample points. In the signals, timbre variation is solely due to difference in construction of the MI. Each signal represents a note and the Fast Fourier transform (FFT) provides the harmonic contents of the note. Amplitude (dB) and frequency (Hz) of each partial were measured from the Long Term Average Spectra (LTAS).

Timbre parameters were calculated using amplitude and frequency data from those partials. We have analysed several timbral and spectral parameters viz. tristimulus 1,2 and 3, spectral brightness, spectral centroid, spectral irregularity, odd and even parameters, harmonic shift, variation of pitch and resonant frequency of each of the sound signals [Ciglar M, 2009]. [Datta Ashok Kumar et. al, 2017], [Jensen K et. al., 2001] and [Jensen K., 2002]. All these parameters are measured from the long term average spectra of FFT. These parameters were used to analyse sound source characteristics or the resonance characteristics of both the instruments and are helpful in comparing with the sound quality of esraj and Indian made violin. A complete timbre space of Esraj reveals some interesting characteristics and defines the structural map of Esraj. This study aimed at comparison of esraj and violin sound based on timbre space only and is a unique study.

5. RESULTS AND DISCUSSIONS

At first waveform and spectrogram of all the sixteen signals needed to be studied thoroughly. Out of these sixteen signals we have shown only two spectrogram and waveform views in fig. 2 and fig. 3.

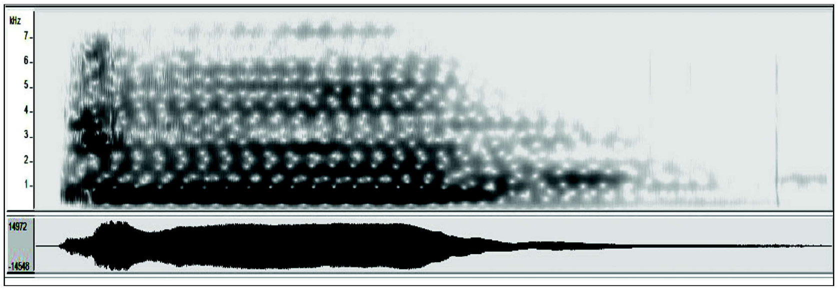


Fig. 2a. Spectrogram view and waveform view of D# (Ma) string of Esraj.

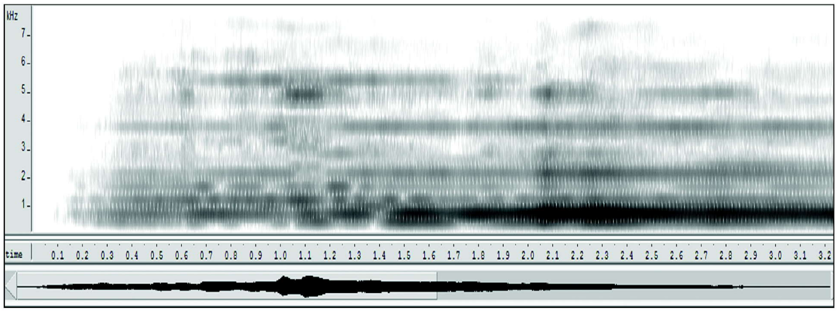


Fig. 2b. Spectrogram view and waveform view of A# (Sa) string of Esraj.

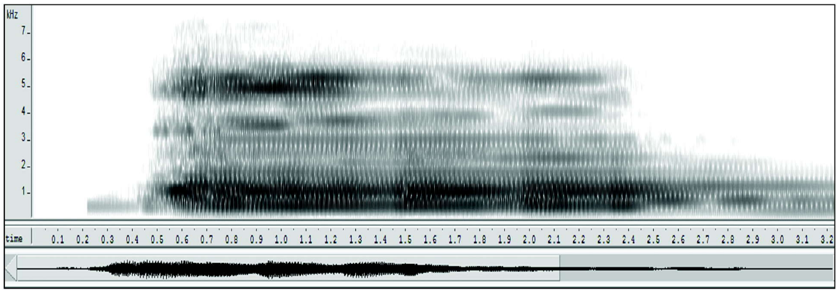


Fig. 2c. Spectrogram view and waveform view of F (Pa) string of Esraj.

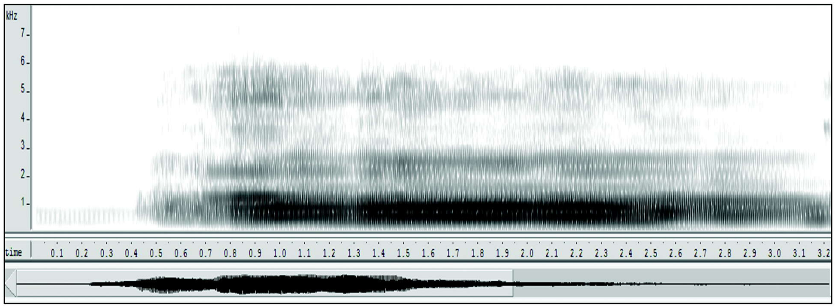


Fig. 2d. Spectrogram view and waveform view of A# (Sa) string of Esraj.

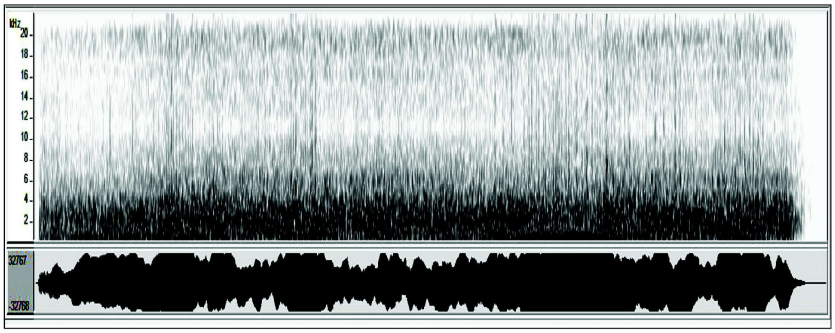


Fig. 3a. Spectrogram view and waveform view of D string of Violin.

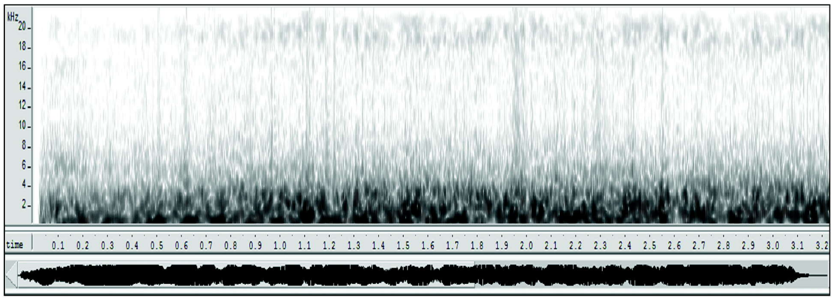


Fig. 3b. Spectrogram view and waveform view of G string of Violin.

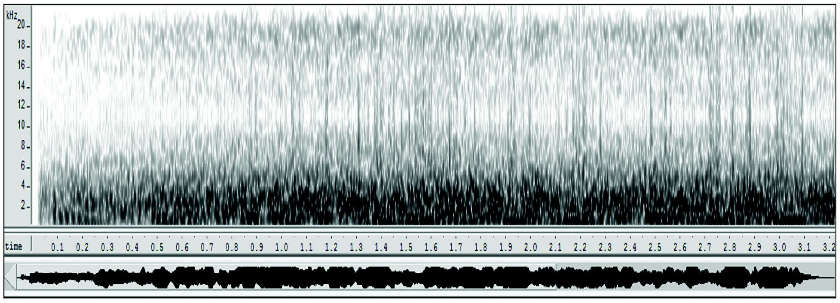


Fig. 3c. Spectrogram view and waveform view of A string of Violin.

Spectrogram is a spectro-temporal representation of the sound. The spectral representation of a sound signal is nothing but to decompose the stochastic process of sound production in a musical instrument into its periodic components. The spectrum density function is the periodic behaviour of realization from the stochastic process.

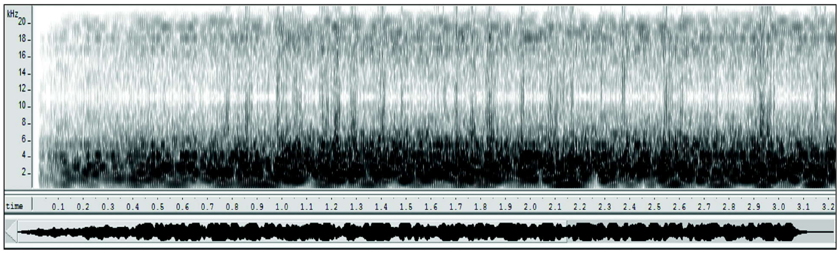


Fig. 3d. Spectrogram view and waveform view of E string of Violin.

Fig 2(a) to 2(d) shows the spectral view of esraj and fig. 3(a) to 3(d) shows the spectral view of violin. From the figures it is observed that the sound of thick strings shows an average frequency below 1kHz while thin strings show frequency greater than 1kHz while frequencies of violin strings are above 2kHz. Maximum energy density of all violin string sound signals are maximum at 2.5 kHz to 4 kHz. Maximum energy density of all the Esraj string sound signals are found to be below 1 kHz. So the resonance occurs in violin at higher frequency ranges than that in Esraj. Average pitch of thick string of esraj is 117 Hz with loudness 55 dB and that for violin is 188 Hz and 56 dB. Average pitch of thin string of esraj is 312 Hz with loudness 52 dB and that for violin is 329 Hz and 68 dB. For other two strings average pitch is 234 Hz and 177 Hz with loudness 46 dB and 47 dB respectively for esraj. While average pitch is 294 Hz and 220 Hz with loudness 58 dB and 64 dB respectively for other two strings of violin. From this data it may be concluded that the violin sounds are in general louder with higher pitch compared to esraj sound. Thick strings of violin produce higher pitch sound in comparison to thick strings of esraj. But the loudness of thick strings of both the instruments remain same. Again thin strings of violin produce louder sound than the thin strings of esraj.

Formant frequencies for the thick string is 563 Hz with loudness 44 dB for esraj and 675 Hz and 50 dB for violin. Again formant frequencies for the thin string is 788 Hz with loudness 50 dB for esraj and 1350 Hz and 50 dB for violin. Resonance occurs in thick string sound much earlier than the thin string sound for both the instruments. Formant frequencies for the thick strings of violin are little higher than that compared to the thick strings of esraj. But the resonance occurs at high frequencies in thin strings of violin while the resonance occurs at lower frequencies in thin strings of esraj.

So we may conclude that the resonance frequencies and hence the formant frequencies of violin string sounds are always higher than the esraj string sound. Here lies the difference in sound quality of the two instruments. This shows a distinguishable difference in timbre space of sound of two musical instruments.

Additive sinusoidal components that are obtained from a harmonic sound is known as partial. Partials must have time-varying amplitude and frequency. Each partial has frequency which are most likely the multiples of the fundamental frequency. The fundamental frequency is measured in the frequency domain using FFT. Now in order to study partial to partial characteristics of a sound signal we need to study mainly the frequency and amplitude of each partial and were studied from LTAS of each signals. From these data we compute all the timbre parameters.

In the fig. 4 and fig. 5 we can observe the LTAS curves in which intensity (power) has been plotted along Y axis and frequency has been plotted along X axis for Violin (D string) and Esraj (D string). Only two out of sixteen signals are shown here. Violin signals shows a regular waxing and waning along with a regular decay quasi periodically but sound signals of Esraj shows no such trend. For the violin sounds only the fundamental frequency has the highest amplitude while this is not so for esraj sounds. Even though both are bowed string instruments, intensity of the fundamental frequency of violin sounds are much higher than that of esraj sounds. Interestingly it is found that both the instruments produce frequencies beyond the upper audible limit for human being.

Damping is another important characteristic feature for violin sounds. At the onset and offset of the fundamental frequency a uniform decay is observed in all violin string sound. From these power spectra

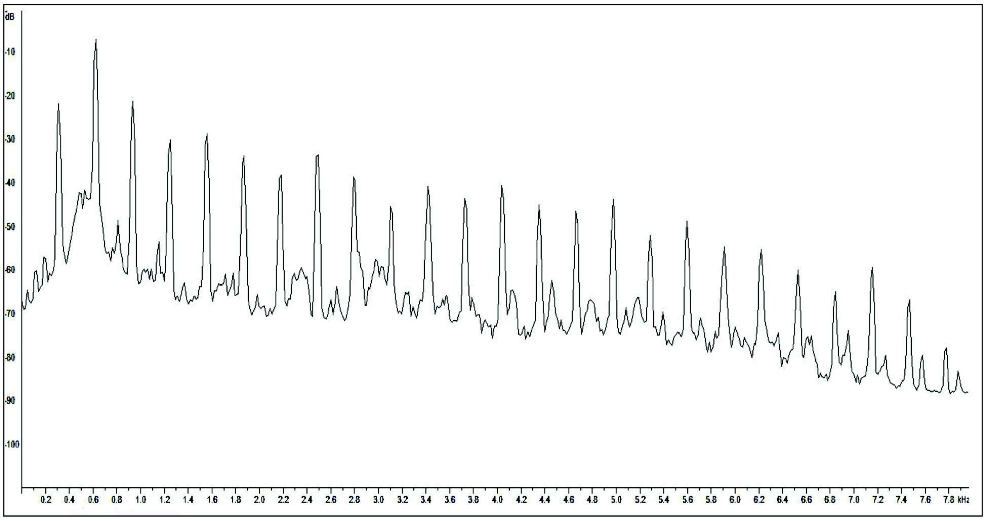


Fig. 4. LTAS view intensity (power) along Y axis vs frequency plot along X axis for Esraj sound (D string).

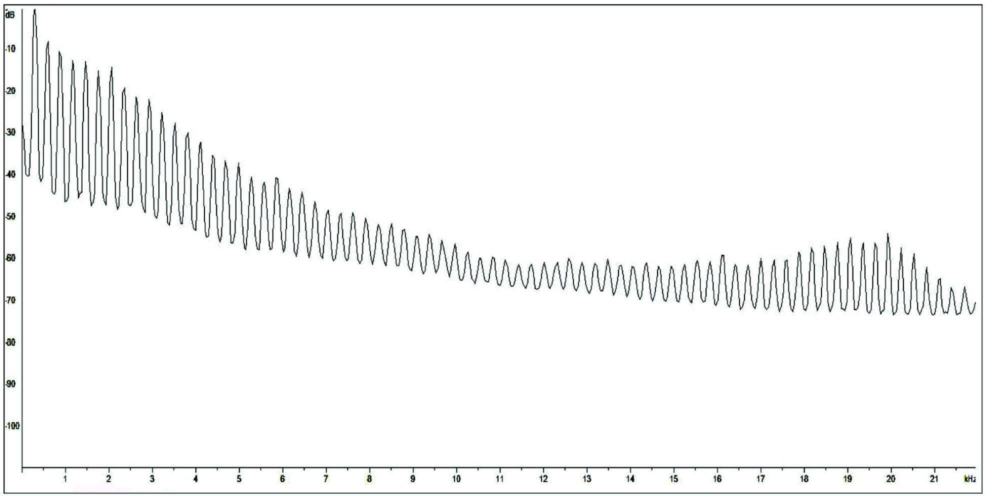


Fig. 5. LTAS view intensity (power) along Y axis vs frequency plot along X axis for Violin (D string).

we have identified the frequency and amplitude data for each of the partials and proceeded for calculating various timbre parameters.

Measure of spectral irregularity of a signal is actually the measure of spectral smoothness. It is the measure of the irregularity of a signal where the local mean is compared with the current amplitude value. Smoothness of a spectrum is an indicator for partials belonging to a same sound source and a single higher intensity partial is more likely to be perceived as an independent sound [Tagore S. M., 1983]. It is also useful in revealing complex resonant structures of string instruments [Ramsey G., 2014].

We have initiated the original pitch data around the onset and offset of a steady segment to result in a smooth transition into and away from the steady note. Then from the LTAS curves we have measured the timbre parameters. Among the timbre parameters, spectral irregularity is important. We have measured the Irregularity by the deviation of amplitude of a given partial from the average amplitude of that partial, partial before and after it. Irregularity in amplitude of the partials of Esraj and Violin are shown in fig. 6, fig. 7, fig. 8 and fig. 9. From the graphs it is observed that the irregularity among partials are in general much less in esraj sounds than the violin sounds. More detailed observation reveals the following.

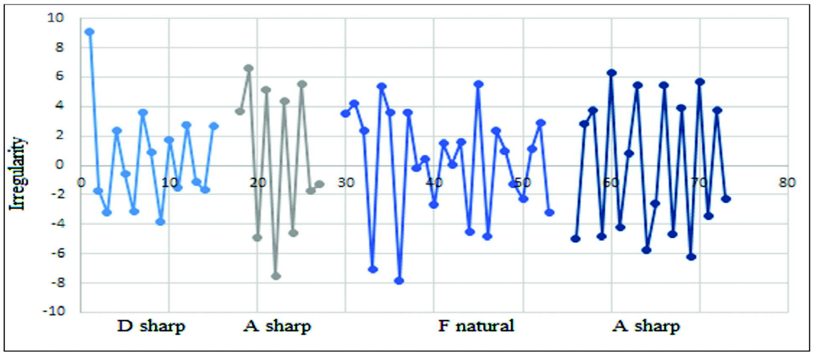


Fig. 6. Irregularity of four strings of high pitched Esraj.

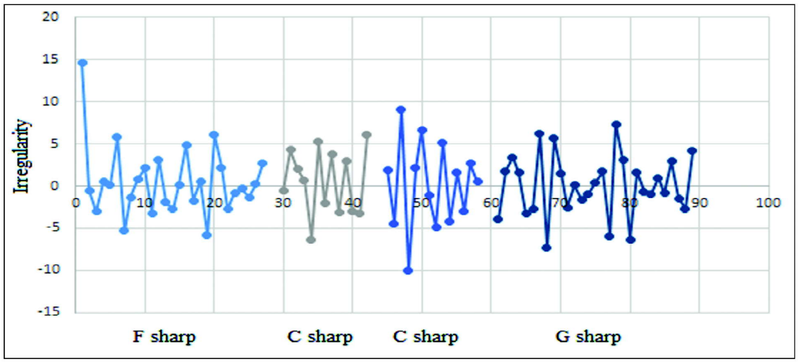


Fig. 7. Irregularity of four strings of normal pitched Esraj

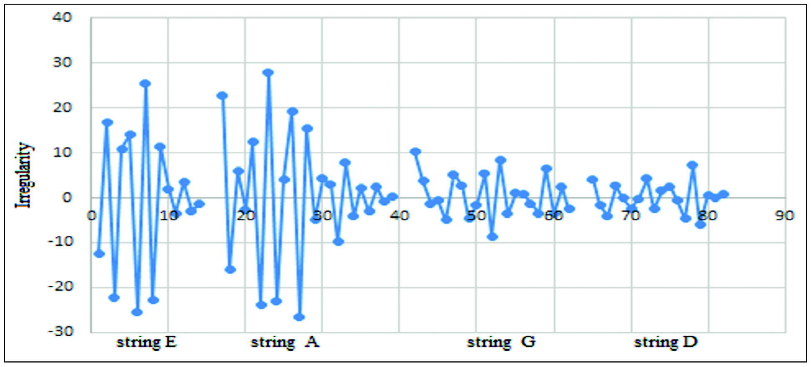


Fig. 8. Irregularity of four strings of Indian made Violin 1.

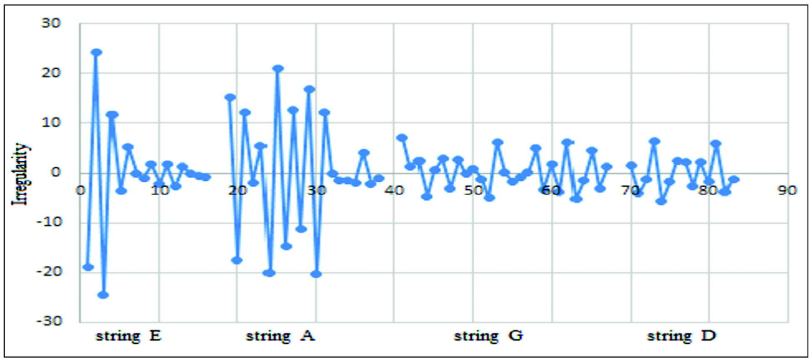


Fig. 9. Irregularity of four strings of Indian made Violin 2

String D sharp is the main string of high pitched Esraj but the Irregularity in amplitude among partials of string D sharp (thin string) is the least while it is more and similar for rest of the other three strings. Variation of amplitude shows regular waxing and waning for all the three strings except F natural (thick) string. Hence the sound quality of these three strings are not lively. Amplitude variation of String F natural little high and the sound quality show little liveliness and with some additive noise. In general, a string of high amplitude variation produces a tone of high quality [Datta Ashok Kumar et. al, 2017]. Despite their differences in thickness variation in irregularity is small. So the timbre quality of this Esraj is weak. In comparison to the High pitched Esraj, strings of the normal pitched Esraj show little more irregularity in amplitude. String F sharp is the thinnest string and string C natural is the thickest string among four strings. String F sharp is the main string of this normal pitched Esraj. Although being thin strings, Strings F sharp and G sharp shows larger irregularity with little more amplitude fluctuation within the time frame than the rest two strings. Amplitude fluctuation of string F sharp and G sharp is more and hence it produces lively sound. So the timbre quality of this Esraj is better than the other one.

All the four strings of Violin are main string. All the four strings of both the Violin shows higher Irregularity in amplitude among partials. Variation of amplitude shows irregular waxing and waning irrespective of thick and thin strings. Much greater fluctuation in the irregularity of partials are observed in the two strings E and A while the fluctuations in irregularity in partial is quite lesser in the two thicker strings D and G. So both the frequency and amplitude perturbation is higher in violin sounds than the esraj sounds. So the timbre quality of violin is more vivid than esraj. Hence the sound quality of Violin is very lively.

Comparing fig. 2, fig. 3, fig. 4 and fig. 5, it is obvious that the range of irregularity of both the Esraj is very low compared to the range of irregularity of Indian made Violin. Hence, the fluctuation in amplitude of violin is much higher than Esraj. So the tone of Violin is of much better quality than the tone of Esraj.

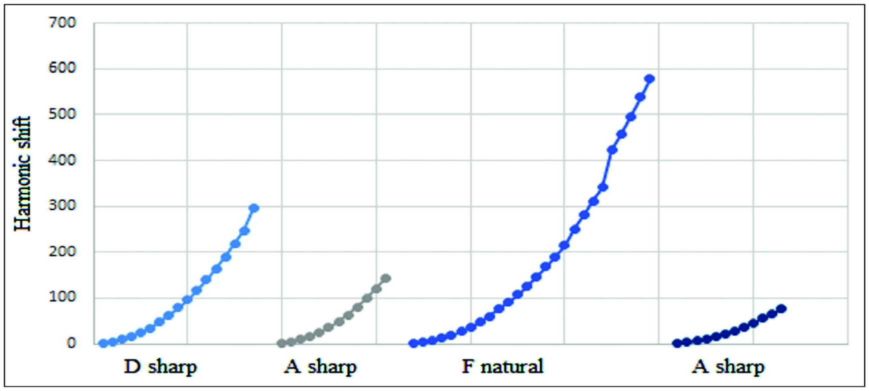


Fig. 10. Harmonic shift of four strings of high pitched Esraj.

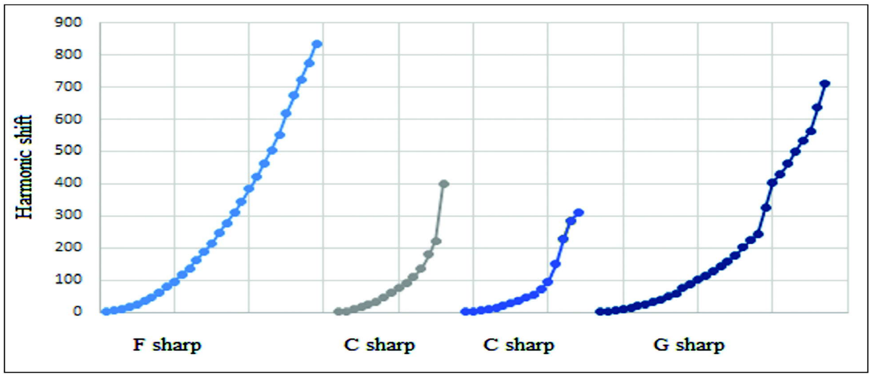


Fig. 11. Harmonic shift of four strings of normal pitched Esraj.

Analysis was also done with pitch pattern behavior/harmonic behaviour of each signal, in which we have made a distribution of pitch/harmonic shift of each peaks of LTAS (Long Term Average Spectra). Harmonic shift is measured by the ratio of frequency of a partial to the fundamental frequency and is shown in fig. 10, fig. 11, fig. 12 and fig. 13. All the distributions, the best fit polynomial curve is of highest order with R2 values 0.9999. From the figures we can observe the harmonic behaviour among partials. Good resonance structure of the MI shows strong harmonic behaviour among partials.

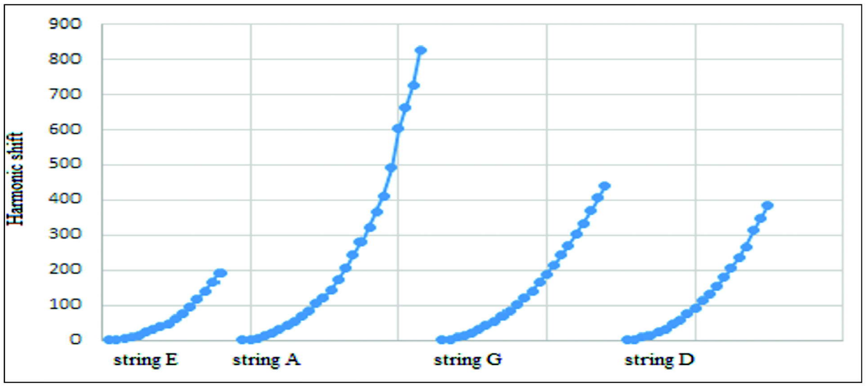


Fig. 12. Harmonic shift of four strings of Violin 1

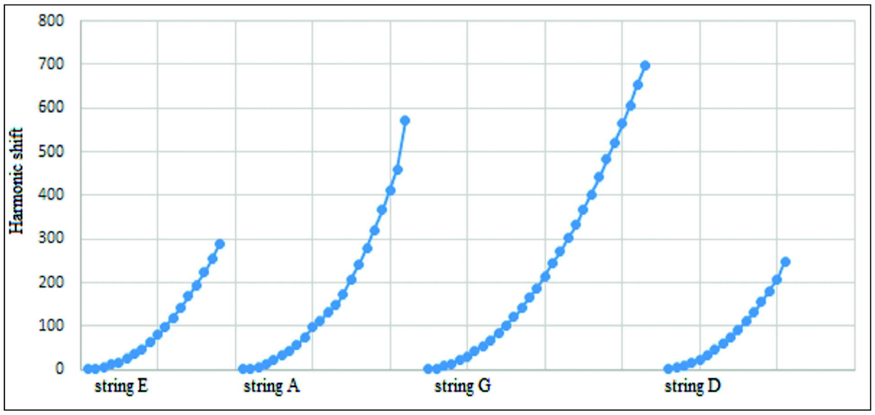


Fig. 13. Harmonic shift of four strings of Violin 2

All the graphs are quasi harmonic and hence inharmonicity of sound of both the MI is less. The thin string A# shows minimum harmonic variations and another thin (main) string (D sharp) of high pitched Esraj shows some little harmonic variations while the string F natural shows maximum harmonic variations. The thin (main) string (F sharp) and G# of normal pitched Esraj show larger harmonic variation while C# shows little harmonic variations. So the string of normal pitched Esraj has better coupling with the resonating chamber than compared to the high pitched Esraj. The thick string (F natural) of high pitched Esraj shows larger harmonic variations while the thick string (C sharp) of normal pitched Esraj show a little harmonic variation.

Thin string E of violin shows least harmonic variations among partials, while Maximum harmonic variations are observed in strings A and G. Both the violin shows similar type of harmonic behavior. So the timbre quality of violin does not differ much. Comparing harmonic shift of esraj sounds from fig. 10 and fig. 11 with the harmonic shift of Violin signals from fig. 12 and fig. 13, it is found that the harmonic behavior of Violin is more linear than Esraj.

Since the sound of a MI can be qualified by its timbre, defining the identity and expression of the instrument, timbral characteristics such as brightness, tristimulus T1, T2, and T3, and the odd and even

Table 1. Timbre parameters of four MI.

	Strings	Brightness	T1	T2	T3	ODD/EVEN	Centroid
High pitched Esraj	D sharp	10.307	0.037	0.098	0.865	1.153	6.365
	A sharp	7.338	0.081	0.143	0.777	0.869	3.382
	F natural	14.827	0.045	0.06	0.895	0.987	5.457
	A sharp	12.339	0.037	0.079	0.885	1.065	3.222
Normal pitched Esraj	F sharp	16.714	0.04	0.041	0.92	1.078	6.141
	C sharp	8.66	0.084	0.145	0.771	1.131	2.838
	C sharp	9.372	0.054	0.142	0.804	1.077	4.421
	G sharp	17.501	0.025	0.06	0.916	1.078	4.869
Violin 1	A	14.257	0.058	0.093	0.849	1.382	9.62
	D	12.572	0.031	0.121	0.849	0.886	7.453
	E	10.601	0.034	0.193	0.772	1.003	8.603
	G	13.124	0.078	0.066	0.856	1.067	5.091
Violin 2	A	12.69	0.052	0.105	0.843	1.455	7.868
	D	10.491	0.031	0.176	0.793	0.844	5.571
	E	10.941	0.03	0.179	0.79	0.751	11.943
	G	13.661	0.05	0.057	0.894	1.001	6.824

parameters have been extracted, along with spectral brightness and spectral centroid for each string (Table 1).

Brightness of all the strings of violin shows consistency indicate a good timbre quality. High value of spectral brightness indicates a strong fundamental for both the MI. The thick string A sharp of high pitched Esraj and both the thick strings C sharp of normal pitched Esraj, is lacking in brightness and hence sound of these two strings have weak fundamental. Both T1 and T2 for A# of high pitched esraj and C# for normal pitched esraj is high while other strings shows relatively consistent T1 and T2 values. T3 of all the strings of both the esraj shows relatively similar values. Inconsistency in the values of T1 and T2 is an indication of poor timbre quality and a possible maladjustment of the strings with the bridge. T1 of all the violins are almost similar and hence all the violin shows dominant presence of fundamental in the sound of all strings. T2 of G strings of both the violins is low and hence the mid frequency ranges of G string sounds are weaker than other strings. T3 of all the strings of both the esraj shows relatively similar values. Consistency in the values of T1, T2 and T3 is an indication of strong timbre quality and a good coupling of the strings with the bridge causing a good resonance structure. Low value of T1 and T2 indicates that the energy of the lower frequency partials is too low, energy shoots up at the higher frequency partials. Such characteristics is observed in both the MI. So the process of gaining energy by the partial remains same for the MI. Centroid of Esraj is lower than the centroid of violin. Violin sounds show a strong decay with a dominant fundamental while Esraj sound shows a weak decay with strong fundamental. Consistency in T3 of all the strings of violin proves the strong coupling between the resonating chamber with the bridge and string, hence produces better sound quality. Also higher partials of violin sound pump up energy which causes the violin sound to have a good timbral quality. Odd to even ratio values are close to one for both the instruments, hence no dominance of either even or odd harmonics are found in the sound signals.

From table 2 it is clear that the two violins have uniform timbre space while it is different for the two Esraj. Lower standard deviation in brightness of violin sounds indicates the existence of strong fundamental while higher standard deviation of brightness of esraj sounds indicates the existence of weaker fundamental. Variation in standard deviation of T1, T2 and T3 of violin shows consistency. This is an important finding for violin sound that the energy is being distributed among partials uniformly in violin

Table 2. Standard deviation of timbre parameters for the four instruments.

	Brightness	T1	T2	T3	ODD/EVEN	Centroied
High pitched Esraj	3.171	0.021	0.035	0.448	0.121	1.553
Normal pitched Esraj	4.692	0.025	0.054	0.559	0.027	1.364
Violin 1	1.529	0.022	0.055	0.04	0.212	1.947
Violin 2	1.486	0.012	0.059	0.049	0.313	2.759

sounds. Violin sounds have strong fundamental and gradually weakens at higher harmonics causing a lower T3 values. This is not so for esraj sounds. Esraj sounds have weaker fundamental and gradually stronger at higher harmonics causing a higher T3 values. Since Esraj is mainly played by a single string (main string), so it needs to have different tension in the string in order to get a definite pitch value of tonic 'Sa'. This causes different timbre space for this musical instrument. In the fig. 16 we observe the KNN plot of two instruments based on timbre parameters. The graph is self-explanatory; it clearly differentiates the two instruments. Ten data points from the timbre parameters are found to be closer. The sound of violin strings is shown by "small line in pink" and are marked as negative while that for esraj strings is shown by "blue square". It is clear that the timbre data of violin sounds are more closer than that of esraj sound.

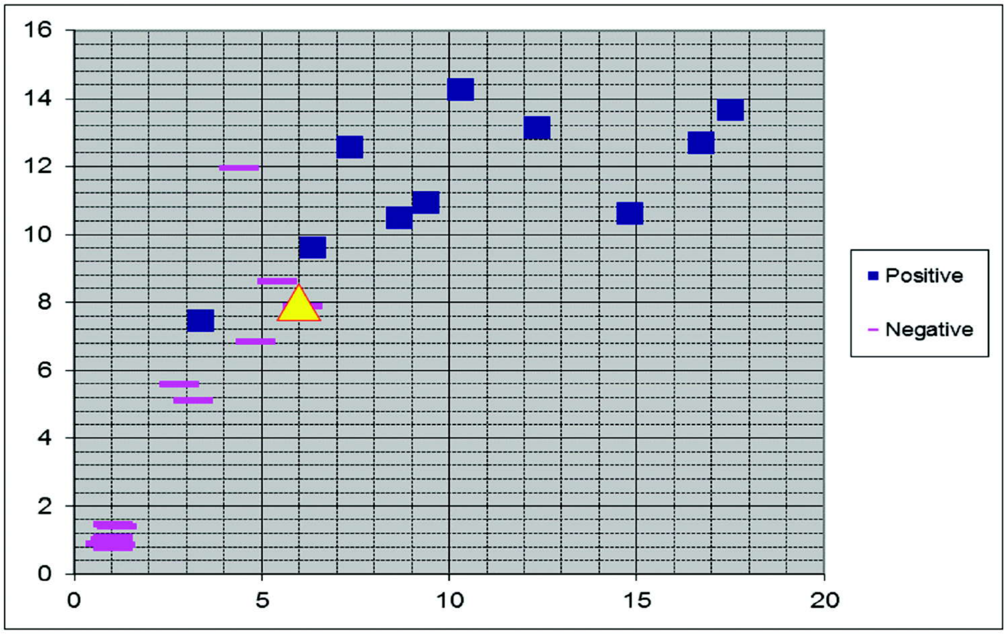


Fig. 14. KNN of two instruments based on the timbre parameters

6. CONCLUSION

Brightness and Centroid of Esraj sounds among four strings do not have any uniqueness while that of Violin sounds have. So the intensity and hence the energy distribution in the spectrum of Esraj sounds varies from one string to another. Sound of Esraj strings show very strong fundamental and a weak decay, but sound of violin strings shows strong fundamental and strong decay. Also signal from some strings of Esraj show equal distribution of energy during fundamental, mid frequency range and at higher frequency ranges, while some signal shows a low energy during fundamental but energy pumps up at higher frequency range. Thus a wider variation of energy distribution was observed in sound signals of strings of Esraj. In all sound signals of violin a strong decay and higher energy distribution was observed at mid and higher frequency ranges. Analyzing the sound from both the A sharp strings of high pitched Esraj we may conclude that there is a weak coupling of strings with the bridge and the resonating chamber. Similar observation was found for both the C sharp strings of the normal pitched Esraj. Pitch pattern was uniform for signal of two thin strings while it shows wider variation for other two thick strings. This was observed for both the MI. Irregularity graphs for Esraj sound shows regular waxing and waning of amplitude among partials while the pitch shift graph shows inharmonicity. That is why the Esraj sounds are sweet to hear. However, violin sounds are rich in overtones and resonance and show a rich timbre space due to strong coupling between bridge, wooden body and the resonant cavity. Esraj sounds are not rich in overtones and resonance and do not have rich timbre space due to loose coupling among bridge and the body and much smaller resonant cavity. The reason for weak resonance is due to the leather cover wounded over the wooden structure in Esraj. We may assume that the difference in sound quality between Esraj and Violin, as observed by us, may be due to difference in quality of strings, type of wood and bridge, the size and shape of the resonating chamber and finally the pressure applied on the strings. It is also found that the violin sounds are in general louder with higher pitch compared to esraj sound. Also the formants of the sound of violin strings are higher than that of the esraj strings. Vibrating strings of violin produces sound that is rich in harmonics. Bowing the string not only allows a range of expressive techniques, but also supplies energy continuously and so maintains the harmonic richness. So due to poor timbre quality of esraj, it is not as popular as violin in Hindustani music.

7. REFERENCES

[1] Banerjee Kaushik, 2017. Indian string musical instruments - making & makers. Kolkata, *Parul Prokashani Pvt. Ltd.,* ISBN 978-93-86186-61-4.

[2] Banarsidas M., 1997. Sitar and Sarod in the 18th and 19th Centuries: Allyn Miner. *Delhi, Publishers Pvt. Ltd.,* First Edition.

[3] Ciglar M., 2009. "The temporal character of timbre". Diploma Thesis. IEM - Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz. *Slovenia.* www.iem.at.

[4] Datta Ashok Kumar, Solanki Sanjib Singh, Sengupta Ranjan, Chakraborty Soubhik, Mahato Kartik and Patranabis Anirban, 2017. Signal Analysis of Hindustani Classical Music. *Springer Nature,* ISBN 978-981-10-3959-1.

[5] Day C.R., 1974. Music and musical instruments of Southern India and the Deccan: (1891). *Delhi reprint by B.R. Publication Corporation.*

[6] Ghosh L.N., 1975. Geet - Baddyam (part-1). *Pratap Narayan Ghosh*, 1st edition, Calcutta.

[7] Houssay A.M.H., 2007. Violin Strings, Set up, Bows and Performance Techniques, Conference Paper, September 2007 https://www.researchgate.net/publication/272173135.

[8] Hutchins C.M.A., 1983. History of Violin Research, The Journal of the Acoustical Society of America **73,** 1421, https://doi.org/10.1121/1.389430)

[9] Jensen K. and Georgios M., 2001. "Hybrid Perception". 1st Seminar on Auditory Models, Lyngby, Datalogisk Institute, University of Copenhagen, Universitetsparken1, Copenhagen, Denmark.

[10] Jensen K., 2002. Perceptual and physical aspects of musical sounds. University of Copenhagen, Universitetsparken1, Copenhagen, Denmark.

[11] Mukhopadhyay D., 1976. Bangalir Ragsangeet Charcha: Pharma K.L.M. (P) LTD. 1st edition , Calcutta Roy A. 2003. Esrajer Ranadhir. Collection of Alpana Roy, Published by - Papyrus, Kolkata. ISBN-81-8175-002-0.

[12] Roy Chowdhury H.K., 1929. The Musicians of India (illustrated) Part-1 (reprint): Kuntalin Press, Calcutta.

[13] Mukhopadhyay D., 1976. Bangalir Ragsangeet Charcha: Pharma K.L.M. (P) Ltd. 1st edition Calcutta

[14] Michael R.1976. Comparative Studies of Musical Instruments, *Computers and the Humanities,* **10**(2), 93-100, Springer, https://www.jstor.org/stable/30199786.

[15] Ramsey G., 2014. A comparative study of stringed instruments, *The Journal of the Acoustical Society of America,* **135,** 2184, https://doi.org/10.1121/1.4877112.

[16] Sengupta R., Datta A.K., Dey N. and Nag D 2003. Acoustic Cues for the Timbral Goodness of Tanpura- *J. Acoust. Soc. India,* **31**.

[17] Sengupta R. *et.al.,* 2007. Random perturbations in harmonium signals, *Journal Acoustical Society of India,* 34(1), 53-59.

[18] Tagore S.M., 1983 Yantra Kosha or A treasury of the musical Instruments of Ancient and Modern India, and of various other countries: Calcutta , Sarmila Prakasani 1983, Calcutta

The Journal of Acoustical Society of India : Vol. 50, No. 1, 2023 (pp. 47-54)

AI for education: Language and music learning

Vipul Arora
Department of Electrical Engineering, IIT, Kanpur, India
e-mail: vipular@iitk.ac.in

[Received: 17-02-2022; Revised: 31-03-2022; Accepted: 05-05-2022]

ABSTRACT

Artificial Intelligence (AI) has revolutionized various fields by enabling machines to perform tasks such as image identification, speech recognition, and text understanding. The impact of these AI algorithms extends beyond academia into industry and public use, influencing sectors like healthcare, business analytics, smart homes, and governance. Education has also benefited from AI advancements. Web-based technologies, including search engines and online videos, have made education more accessible and engaging. Massive open online courses (MOOCs) exemplify how learning can be available to anyone at any time. Enhanced learning management systems now facilitate closer teacher-student and peer interactions. The next frontier in education involves smart tutors that assist teachers by organizing content, grading submissions, recommending lessons, and identifying areas needing attention. AI technologies, particularly in text and image processing, are already integrated into many educational tools. However, speech or audio processing technologies remain underutilized in education despite their potential impact. Vocal capabilities are crucial in education, and leveraging technology in this area can significantly enhance learning. This article explores AI applications in pronunciation training and music teaching. Pronunciation training is vital for language learners to achieve fluency and correct phoneme production, prosody, and intonation. In music education, AI can address components such as melody, rhythm, and pronunciation. The article emphasizes imitation-based learning, where AI helps learners correct deviations from reference sounds provided by expert teachers. Additionally, it highlights the importance of explainable AI (XAI) in education, where providing insights into mistakes and corrective feedback is more beneficial than merely identifying errors. Integrating AI with domain knowledge can create explainable systems that offer valuable feedback to learners.

1. INTRODUCTION

Artificial Intelligence (AI) has brought unprecedented success in recent years by enabling machines to do tasks such as identifying images, recognising speech and understanding text. These algorithms' success is not confined to academia anymore but extends to industry, and the outcomes are available for public use. AI-based products are impacting diverse areas, such as healthcare, business analytics, smart homes and governance. The education sector is not behind. Web-based technologies such as search engines and online videos have been among the primary tools for educators and learners. They have not only made education accessible but also engaging. Massive open online courses are one example of making learning available to anyone at any time. This technology for one-way instruction has been improved in the form of learning management systems that allow closer teacher-pupil and peer-to-peer interactions.

The next step could be smart tutors that not only assist the teachers by organising their content but also bear some of the teaching load themselves. This could be done by automatically grading student submissions, recommending lessons to the learners, and detecting areas where the teacher’s attention is needed.

Text and image processing technologies have found many applications in the educational tools available commercially. On the other hand, speech or audio processing technologies, which are highly impactful in general, are yet to receive the due traction for education. The vocal faculty plays an essential role in education, and tapping it with technology can have an unprecedented impact.

This article discusses the applications of AI in pronunciation training and music teaching. Pronunciation training is essential for the learners of a new language. Apart from memorising vocabulary, learning syntax and constructing sentences, speaking fluently is vital to language learning. Correct pronunciation, at a basic level, entails producing the phonemes correctly in isolation and running speech. It also includes prosody and intonation, learnt at an advanced level. Music is an interplay of various components such as melody, rhythm and lyrics, which involve pitch, time and pronunciation, respectively. At a basic level, these components could be dealt with separately, while they need a holistic treatment at advanced levels of pedagogy.

In this article, we focus on imitation-based learning, where the expert teacher provides the reference sounds, which the learner listens to and tries to imitate. The goal of the AI system is to detect mistakes or deviations in the sounds of the learner as compared to the reference sounds and help the learner correct them.

While AI technology is advancing in different directions, such as transfer learning, domain adaptation, generative modelling, reinforcement learning, privacy-preserving and ethical AI, some of these directions are incredibly vital for education. This article would like to highlight the utility of explainable AI (XAI). While XAI is highly desirable in applications involving critical decisions, it is also a great asset for education. Just telling a learner, ″you are wrong″, has little utility compared to telling her why she is wrong and how she can correct it. Modern AI tools often are black boxes, giving little insight into the underlying structure. Hence, the right blend of AI with domain knowledge is required to make explainable systems that could offer corrective feedback to the learner.

2. AI AIDED PRONUNCIATION TRAINING

The articulated sound recorded by a microphone is modelled as an outcome of many linguistic processes occurring at different levels. Generally, speech is represented as text or graphemes, *e.g.,* consider the word ″APPLE″. However, the graphemic form is mostly conventional and may not convey the underlying relationships between sounds qua linguistic units. Hence, phonological forms are commonly used as the underlying representation of speech. *E.g.,* the phonological representation of ″APPLE″ is / æp∂l/. Phonology studies relationships between sounds. Many of these relationships are invariant across languages. To model these relationships, phonological features have been formulated[JFH51].

The problem can be formulated as follows:

1. A ″target″ sentence is given in the L2 to speak. The corresponding sequence of phonemes is called the ″canonical″ phoneme sequence. The teacher’s audio can also be provided for the learner to listen to.
2. The learner speaks the sentence into the microphone. It is recorded as the ″spoken″ audio.
3. Now, the goal is to find mispronunciations in the spoken audio and to provide corrective feedback to the learner.

A simple method to compare two audio signals is to use dynamic time warping (DTW)[LG12]. Each audio signal is converted into a sequence of feature vectors, which serve as the units to be compared. If the spoken utterance matches the target, the matching score is high. A low matching score indicates a mistake. The feature vectors should contain linguistic information to make the match more meaningful.

They may not include other information, *e.g.,* the gender of the speaker, that is irrelevant to the task at hand. Simple DTW has several limitations; for instance, it uses hand-designed features and matches with only one target audio, which is just one sample of possible pronunciations. To overcome these shortcomings, and to further incorporate more explainability into the system, automatic speech recognition (ASR) technologies are used.

2.1 ASR based Mispronunciation Detection

ASR technologies use sequence models to convert audio signals into a sequence of linguistic units, which could be phonemes or graphemes. Primarily two kinds of ASR technologies are prevalent these days: i. HMM-based[Rab89] and ii. end-to-end learning based[Gra12]. In this article, we focus on HMMbased ASR for mispronunciation detection, while we acknowledge some recent works on end-to-end mispronunciation detection[YC21, GCC+22].

The HMM-based ASR system[PGB+11] comprises two models: an acoustic model and a language model. The acoustic model takes the short-time spectral features extracted from the audio and generates probability scores for the underlying units (phonemes, or sub-phonemic states, called Senones). The language model, a finite state transducer, uses the output of the acoustic model to string together these units along a single path to generate the final sequence of words.

The ASR system described above can be modified for the task of pronunciation detection in the following different ways:

- Use the probability scores generated by the acoustic model as the representative sequences for the reference and the learner's utterances and compare them using DTW[LG12].
- Train the ASR system using reference utterances, *i.e.,* experts' speech. Force align the learner's utterance with the corresponding reference phoneme sequence (assuming it is available). A measure known as Goodness of Pronunciation (GOP) is defined to quantify the quality of each phoneme. Mainly, it involves comparing the reference phoneme probability to the probability of any other phoneme in the same time window [WY00].
- Train the ASR system using reference utterances, *i.e.,* experts' speech. Force align the learner's utterance with the corresponding canonical phoneme sequence (assuming it is available). Feed the probability scores generated by the acoustic model into a mispronunciation detection model, trained explicitly for this purpose[ALR18]. This approach is known as the classifier-based approach.
- Force align the learner's utterance with the corresponding canonical phoneme sequence (assuming it is available). However, during forced alignment, the language model provides different paths that model the possible mispronunciations. If there is a mispronunciation, decoding will prefer the deviated path and hence, will detect the mispronunciation[QMS16]. This approach is called the Error Recognition Network (ERN) based approach. It is illustrated in Fig. 1.

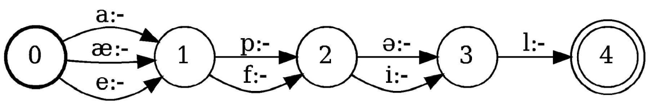


Fig. 1. An error recognition network.

2.2 Phonology for Explainability

The above methods detect mistakes but do not provide corrective feedback. If the system is implemented for non-native language (L2) learning, the learners are already familiar with a native language (L1). The L1 influences the type of mispronunciations and can be used to construct corrective feedback that the learner could easily comprehend. Phonological features provide a neat framework for modelling linguistic phenomena that cut across languages. Phonological features can be detected from the audio signals using AI-based methods[AR16] and are found to be effective in cross-language knowledge transfer[ALR16]. A set of phonological features characterises each phoneme. These features include voc for vowels and cons for consonants; cor, dor, high and low for tongue positions; lab for lip involvement; and so on. Each phoneme is mapped to a set of these features by a linguist.

A phonological feature extractor network is trained explicitly. This network can have some layers shared with the acoustic model.[ALR18] estimates the probabilities of phonological features at each time frame. For every phoneme detected as mispronounced, the phonological feature probabilities are averaged and compared with the phonological features of the target phoneme. The mismatch in phonological features provides an insight into how the mispronunciation is to be corrected. For example, a missing lab feature while pronouncing the German vowel ¨o shows that the lips were not rounded. A subsequent module is needed to translate the feedback in terms of phonological features into a language understandable to the learner. This module could be L1-specific so the user can comprehend the feedback message. *E.g.,* in case of a missing lab, while trying to pronounce ¨o, the feedback to a Hindi speaker could be ″round your lips as in mor (peacock)″. A detailed block diagram is shown in Fig. 2.

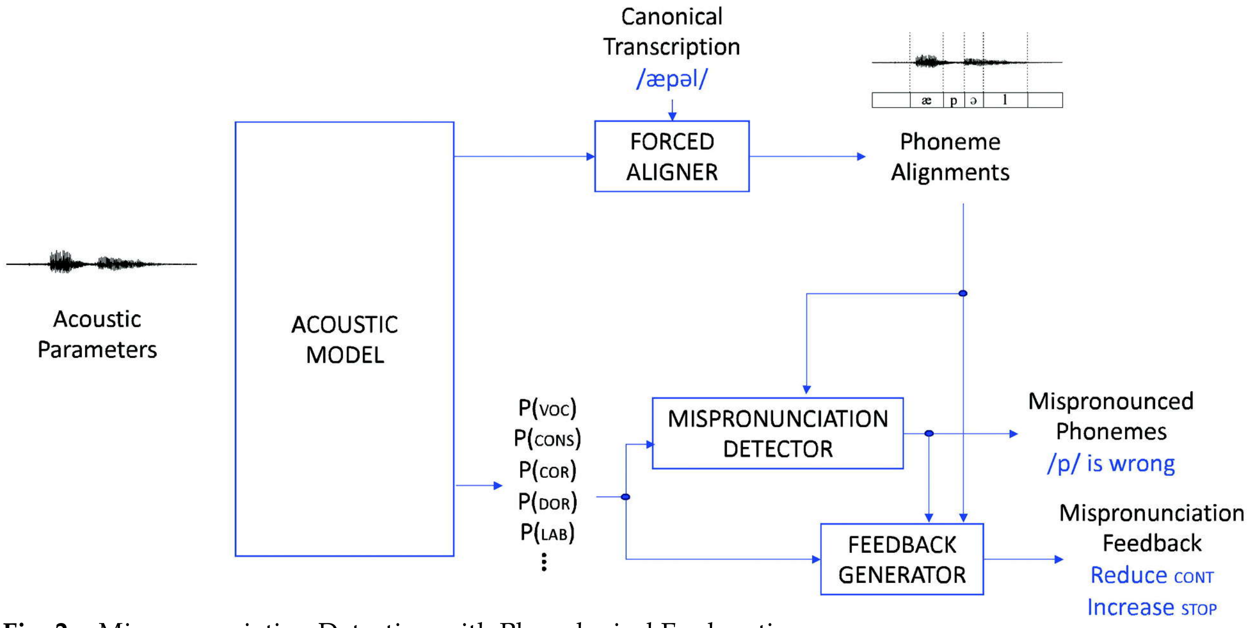


Fig. 2. Mispronunciation Detection with Phonological Explanations.

3. AI AIDED MUSIC TEACHING

Music involves tonal and rhythmic aspects. Music can be produced by a singer or by instruments. However, singing music also involves linguistic aspects due to the lyrics. A music learner may focus on learning any of these aspects. To develop a computer-based assistant for any aspect, one has to be able to analyse that aspect computationally. Decades of research have given rise to various algorithms to examine many of these aspects[KD07, MEKR11, HRS+18]. Some of the works done in the context of Western music are reviewed here[JMC21].

In this article, we focus on singing in Indian classical music. Singing involves melody, rhythm, breathing patterns and pronunciation, for which pedagogical tools could be developed. However, we focus only on melodic aspects. Again, we focus on learning by imitation and not on improvisation.

3.1 Melody Extraction

Melody is a sequence of musical notes. *E.g.,* Indian classical music uses 12 notes denoted as S, R, R, G, G, M, M, P, D, D, N, N. The property of sound that distinguishes different notes in music is called pitch. Melody is one of the critical aspects of Indian music and is of particular interest to learners.

Mathematically, pitch mostly (if not always) corresponds to the fundamental frequency (F0) of the audio waveform recorded via microphone. F0 is the frequency of the smallest repeating unit in audio. Generally, F0 is accompanied by its harmonics, *i.e.,* integer multiples of F0. A typical spectrogram of a sound is shown in Fig. 3. Sounds could be voiced or unvoiced. Voiced sounds have a periodic waveform, and hence, they have an F0. Examples of voiced sounds include the vowels (/a A i I u U/). Unvoiced sounds have a non-periodic waveform and are exemplified by fricatives /f s S/ and plosives /k c t t p/. For melody, only voiced sounds are of interest. A melody extraction algorithm aims to extract the F0 contour from the audio signal. Since singing voice is quasi-periodic, i.e., its pitch changes with time, F0 is extracted from short time windows of the audio signal, resulting in an F0 contour. Typically, these windows are of length 50 ms and consecutive windows are shifted by 10 ms.

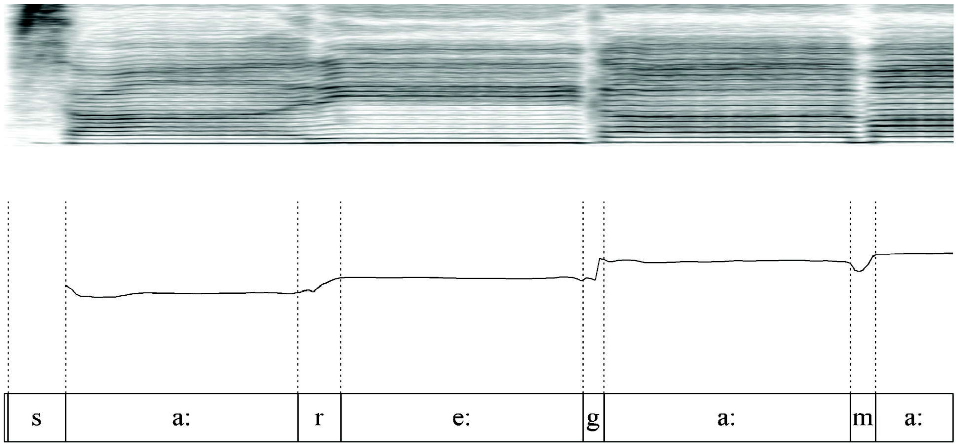


Fig. 3. Spectrogram, F0 contour and time alignment of phonemes in a monophonic audio.

When the singing voice is present alone, with no background instruments playing, estimating F0 is easier compared to polyphonic audio, i.e., the audio containing a mixture of sounds from multiple sources (voice and instruments). A typical spectrogram of the same sound as in Fig. 3, when accompanied by other instruments, is shown in Fig. 4.

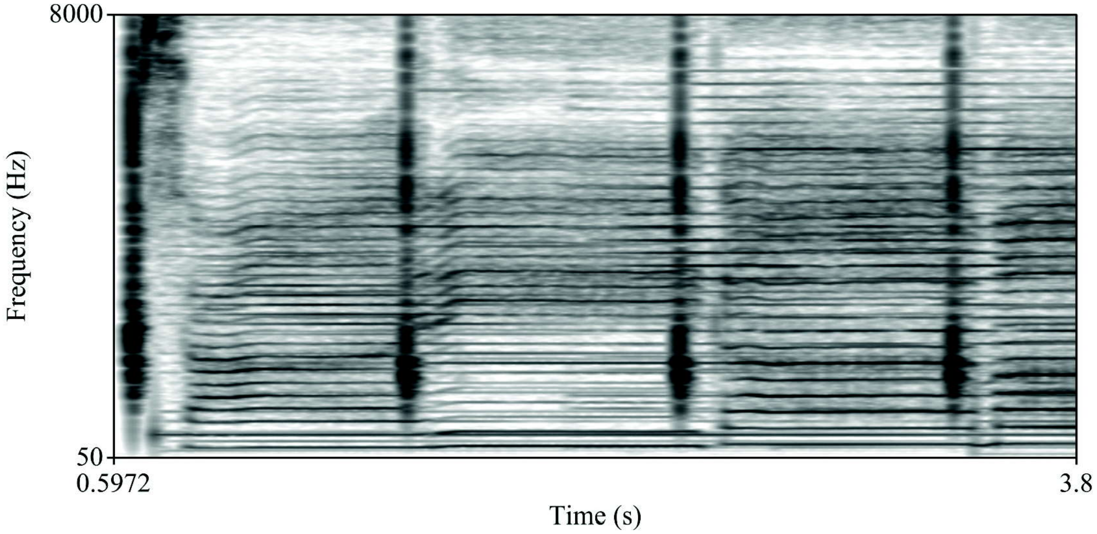


Fig. 4. Spectrogram of polyphonic audio. Due to multiple overlapping sounds, melody extraction becomes difficult.

Numerous signal processing based methods have been developed for melody extraction from polyphonic music[AB13, SGER14]. Recently, deep learning based approaches are being developed for more robust and reliable melody estimation[SSA21, Su18, RRD22].

3.2 *Melodic Similarity*

To check the correctness of the learner's singing, her voice must be compared with the teacher's. Since the F0 contour is a scalar time series, comparing the F0 contours of the learner with that of the teacher is simple. While there have been limited works on assessing melodic similarity for teaching, there has been ample work on other related applications. Query-by-humming is one example that involves matching the melodic contour of the query audio with that of the target song in the database[MF17, RA20].

Music learning can be carried out in two modes: synchronous and asynchronous. In synchronous mode, the learner's singing is synchronised with the teacher's. This can be achieved in two ways: either the learner sings with a percussion accompaniment with the same score as the teacher, or the learner listens to the teacher while singing herself and tries to match the teacher's singing. The latter way could be without a percussive accompaniment too. On the other hand, in the asynchronous mode, the learner sings the target notes or musical entities as sung by the teacher, but they both need not be in exact time synchrony.

In the synchronous mode, a point-wise matching of F0 values of the teacher and the learner, sung simultaneously, can help find mistakes in singing. They are considered matched if the two F0 values are less than half semitone apart. This simple rule can help find errors in both pitch and timing (rhythm), as a mismatch will occur when the learner is unable to hit the same note as the teacher or when she fails to hit the note at precisely the right time. Fig. 5 illustrates this for the F0 contours from the teacher audio (as blue ribbon) and the student audio (as green and red dots). The teacher audio here is the same as that of Fig. 3. This simple rule could form a baseline method and is effective for primary learners. But advanced levels involve learning various musical embellishments and ornaments, *e.g., gamaka, kaṇa, murki*. Finding mistakes here can potentially be accomplished by AI-based methods. The asynchronous mode is similar to the pronunciation training problem, where the system detects the uttered canonical sequence, which in the case of music, will be a sequence of musical entities, namely notes and various embellishments.

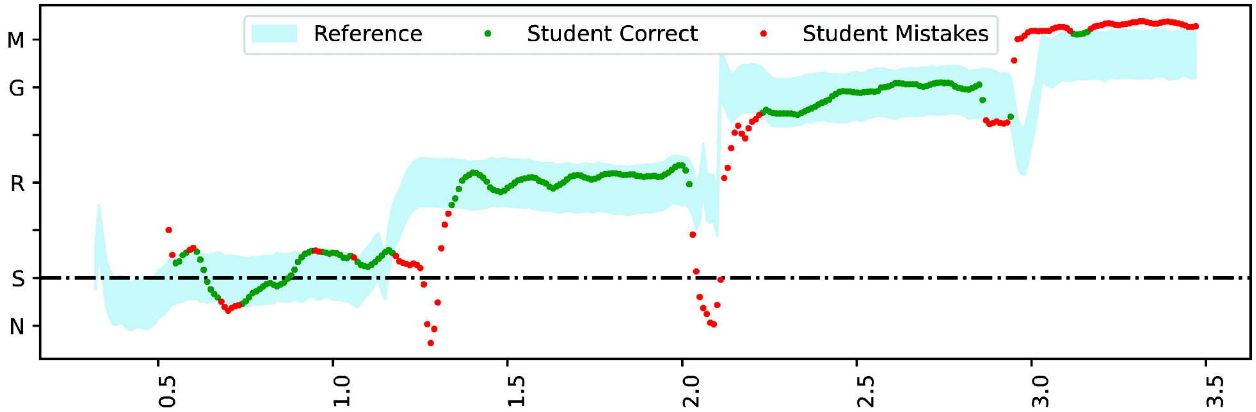


Fig. 5. Comparing teacher's F0 contour (blue ribbon) with that of the learner (green and red dots). Green dots are marked as correct, and red dots are marked as incorrect by the model.

4. CONCLUSION

The problem of detecting mistakes and prescribing corrective feedback in pedagogy is an exciting new area. Great potential lies in deep learning technologies to contribute to this area. While several issues in AI, such as privacy, trust and fairness, need to be considered while deploying AI-based solutions, the technology for language and music is mature enough to be explored for innovating pedagogical applications.

5. REFERENCES

[AB13] Vipul Arora and Laxmidhar Behera, 2013. On-line melody extraction from polyphonic audio using harmonic cluster tracking. *IEEE Transactions on Audio, Speech and Language Processing,* **21**(3), 520–530.

[ALR16] Vipul Arora, Aditi Lahiri and Henning Reetz, 2016. Attribute based shared hidden layers for cross-language knowledge transfer. *In IEEE Spoken Language Technology Workshop (SLT),* pp. 617–623.

[ALR18] Vipul Arora, Aditi Lahiri and Henning Reetz, 2018. Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America,* **143**(1), 98–108.

[AR16] Vipul Arora and Henning Reetz, 2016. Automatic Speech Recognition : What Phonology Can Offer. *In The Speech Processing Lexicon: Neurocognitive and Behavioural Approaches*. Edited by Aditi Lahiri and Sandra Kotzor.

[GCC+22] Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang and James Glass, 2022. Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE,* pp. 7262–7266.

[Gra12] Alex Graves, 2012. Connectionist temporal classification. In Supervised sequence labelling with recurrent neural networks, *Springer,* pp. 61–93.

[HRS+18] Eric J Humphrey, Sravana Reddy, Prem Seetharaman, Aparna Kumar, Rachel M Bittner, Andrew Demetriou, Sankalp Gulati, Andreas Jansson, Tristan Jehan, Bernhard Lehner, *et al.*, 2018. An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music. *IEEE Signal Processing Magazine,* **36**(1), 82–94.

[JFH51] Roman Jakobson, C Gunnar Fant and Morris Halle, 1951. Preliminaries to speech analysis: The distinctive features and their correlates.

[JMC21] Fatemeh Jamshidi, Daniela Marghitu and Richard Chapman, 2021. Developing an online music teaching and practicing platform via machine learning: A review paper. *In International Conference on Human-Computer Interaction, Springer,* pp. 95–108.

[KD07] Anssi Klapuri and Manuel Davy, 2007. *Signal processing methods for music transcription*.

[LG12] Ann Lee and James Glass, 2012. A comparison-based approach to mispronunciation detection. *In 2012 IEEE Spoken Language Technology Workshop (SLT), IEEE,* pp. 382–387.

[MEKR11] Meinard Muller, Daniel PW Ellis, Anssi Klapuri, and Gaël Richard, 2011. Signal processing for music analysis. *IEEE Journal of selected topics in signal processing,* **5**(6), 1088–1110.

[MF17] Naziba Mostafa and Pascale Fung, 2017. A note based query by humming system using convolutional neural network. *In INTERSPEECH,* pp. 3102–3106.

[PGB+11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, *et al.,* 2011. The kaldi speech recognition toolkit. *In IEEE 2011 workshop on automatic speech recognition and understanding, number CONF. IEEE Signal Processing Society.*

[QMS16] Xiaojun Qian, Helen Meng, and Frank Soong, 2016. A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **24**(6), 1020–1028.

[RA20] Shivangi Ranjan and Vipul Arora, 2020. A bioinformatic method of semi-global alignment for query-by-humming. *In 2020 IEEE 4th Conference on Information & Communication Technology (CICT), IEEE* pp. 1–5.

[Rab89] Lawrence R Rabiner, 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE,* **77**(2), 257–286.

[RRD22] Gurunath Reddy, K Sreenivasa Rao and Partha Pratim Das, 2022. Melody extraction from polyphonic music by deep learning approaches: A review. *arXiv preprint arXiv:2202.01078.*

[SGER14] Justin Salamon, Emilia Gómez, Daniel PW Ellis and Gaël Richard, 2014. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine,* **31**(2), 118–134.

[SSA21] Aman Kumar Sharma, Kavya Ranjan Saxena and Vipul Arora, 2021. Frequency-anchored deep networks for polyphonic melody extraction. *In 2021 National Conference on Communications (NCC), IEEE,* pp. 1–5.

[Su18] Li Su, 2018. Vocal melody extraction using patch-based cnn. *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE,* pp. 371–375.

[WY00] Silke M Witt and Steve J Young, 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication,* **30**(2-3), 95–108.

[YC21] Bi-Cheng Yan and Berlin Chen, 2021. End-to-end mispronunciation detection and diagnosis from raw waveforms. *In 2021 29th European Signal Processing Conference (EUSIPCO), IEEE,* pp. 61–65.

The Journal of Acoustical Society of India : Vol. 50, No. 1, 2023 (pp. 55-66)

Intermediality of musical emotions in a multimodal scenario: A deep learning aided EEG correlation study

Shankha Sanyal[1,6*], Archi Banerjee[1,2,3], Sayan Nag[1,5], Medha Basu[1,4], Madhuparna Gangopadhyay[1] and Dipak Ghosh[1]

[1]Sir C.V. Raman Centre for Physics and Music, Jadavpur University
[2]Rekhi Centre of Excellence for the Science of Happiness, IIT Kharagpur
[3]Shrutinandan School of Music, Kolkata
[4]Department of Physics, Jadavpur University
[5]Department of Medical Biophysics, University of Toronto
[6]School of Languages and Linguistics, Jadavpur University
e-mail: ssanyal.sanyal2@gmail.com

[Received: 15-03-2022; Revised: 03-06-2022; Accepted: 23-06-2022]

ABSTRACT

Multimodality of emotional arousal from different input stimulus has intrigued researchers of varied disciplines for the last two or three decades. The present study looks to study the intermediality of musical emotions from the perspective of audio visual and (AV) and audio-only (AO) stimulus and their corresponding neural manifestations. How does the perceptual arousal to a music clip vary when the music video of the same audio content is presented to them, i.e. if the participants hear the same audio stimulus with and without the video element, how would their emotional associations and arousals vary? The study has been designed to analyze the same through extensive audience responses and corroborate the results with nonlinear EEG time series analysis. A psychological experiment was conducted on 50 non-musician participants using 8 AO and 8 AV clips representing two clips from each of the four intended emotional areas - happy, sad, calm, anxiety respectively. The audio clips chosen for this study consisted of the AO versions of the chosen AV clips and were asked to mark the appropriate emotions corresponding to each AV and AO clip along with their respective intensities in a 5 point Likert scale. The average intensity of emotional elicitation from each clip was evaluated and the emotional associations were compared for each pair consisting of the audio and video versions of the same musical content. Next, an EEG study was conducted on 5 participants, who were presented the same set of AV and AO clips as in the psychological experiment and time series data from electrodes belonging to frontal, parietal, occipital and temporal electrodes were extracted for further analysis. Convolutional Autoencoders were used to convert the AV samples to an 1-D time series data just like the intensity waveform of AO data. a robust nonlinear cross-correlation analysis technique Multifractal Detrended Cross Correlation Analysis (MFDXA) was then used to find the degree of cross correlation between the EEG time series data extracted from the different electrodes and the 1-D time series data obtained from the AO and AV clips. Thus, we have a direct correlation between the source stimuli and the output neural response. The study provides interesting new information about multi-modal perception and difference of arousal based activities related to musical emotion processing.

1. INTRODUCTION

Emotions are experienced and realized in our everyday lives, being generated from a variety of different sensory inputs, but the literature lacks one clear definition for the same. The general consensus among the community is that there is an associated physiological or a neurological response that generally accompanies an emotional appraisal. Under the influence of an external emotional stimulus, the specific physiological changes include fluttering and pounding heart, a sweating palm and the tension - relaxation cycles of muscle (Bradley, 2000). In the neurological domain, a number of EEG studies use several features including PSD, fractal dimension, including alpha, theta and gamma power to quantify/ characterize emotional responses generated from a multitude of input stimulus (Lang, Bradley & Cuthbert, 2010, 437-450). These physical, behavioural and neurological changes are very much subjective for specific emotions and are used for classification of one emotion from another.

Multimodality of emotional response has intrigued researchers over the past few decades, and there have been a number of different approaches which deals with stimulus representation from audio signals, facial/visual stimulus as well as from audio-visual responses. In the domain of musical emotions, visual cues including gestures, body postures and facial expressions mediate the arousal based activities (Chen *et al.*, 2014, 14-20; Chapados *et al.*, 2008, 639-651). When audio and visual information were provided to the listeners in tandem, studies provide conflicting information as to whether the emotional information generated from the two different modalities would be synchronous with each other or not. Vines *et al.* (2011, 157-170) and Vuoskoski *et al.*. (2016, 179) found that AO (audio-only presentations) elicit greater response in the GSR (galvanic skin response) based study while no remarkable difference were found for AV (audio-visual presentation) and AO. While these results point to the direction that multi-modal emotional input does not elicit a very strong response as compared to the unimodal input (i.e. either only audio, or only video), there are several other studies also which point in the other direction. Chapados and Levitin (2008) discovered that there was a strong emotional valence based response present in between AV and AO, and also between AV and VO (visual-only presentation) in the galvanic skin response results. These suggested that the multimodal emotional input has evoked stronger response as compared to the unimodal (i.e. only audio or only visual) input (Chapados *et al.*, 2008, 639-651). Another interesting study by Platz & Kopiez (2012, 71-83) reported that audio-visual presentations actually in a way support the musical presentation and hence lead to enhanced musical appreciation. In the domain of EEG signal processing, multimodal emotion recognition have been attempted by several researchers, including some where EEG data is coupled with other bio-sensors like eye tracking (Zheng *et al.*, 2014, 5040-5043), auto-encoder based techniques (Zhang, 2020, 164130-164143), applying brain functional connectivity measures from EEG data (Wu *et al.*, 2022, 016012. 103-126) and several other tools of machine learning algorithms (Zhang *et al.*, 2020,). Now most of these studies focus on the application of different EEG features for optimal classification of emotional appraisal obtained from multi-modal stimulus generated from video/ movie clips taken from different publicly available database. A comparative study dealing with the degree of emotional arousal generated from uni-modal and multi-modal stimuli using robust nonlinear correlation based techniques applied on EEG signals is attempted in this work.

The term correlation in general has been used conventionally to determine the degree of similarity between two signals. In signal processing algorithms, cross-correlation measures the similarity between two nonlinear time series as a function of the lag of one relative to the other. A robust technique called Multifractal Detrended Cross correlation Analysis (MFDXA) (Zhou, 2008, 066211) has been used to analyze the multifractal behaviors in the power-law cross-correlations between any two non-linear time series data (in our case music/ video signals and sonified EEG signals). With this technique, all segments of the music/ video clips and the sonified EEG signals were analyzed to find out a cross correlation coefficient (?x) which gives the degree of correlation between these two categories of signals. For uncorrelated data, ?x has a value 1 and the lower the value of ?x more correlated is the data (Podobnik *et al.*, 2011, 066118). MFDXA as a tool has found applications in different areas of research, starting from stock markets, wind direction assessment, health risks monitoring, literary text analysis to music signal analysis and emotion analysis from EEG signals (Lin *et al.*, 2017, 751-760; Chakraborty *et al.*., 2022; Ghosh *et al.*., 2018, 392-403; Sanyal *et al.*., 2019, 13-31; Ghosh *et al.*., 2019, 1343-1354; Roy *et al.*., 2021, 1023-1053; Sanyal *et al.*., 2016, 1-7). Sonification of the acquired EEG signals were done following the methodology employed by Sanyal *et al.*, (2019), while convolutional autoencoders (Kingma *et al.*., 2013; Huang *et al.*., 2017, 1551-1561; Mukherjee *et al.*., 2019, 2027-2031; Castrejon *et al.*., 2019, 7608-7617; Liu *et al.*., 2021, 701-710) were used for conversion of video signals to 1-D time series.

As mentioned earlier, researchers have tried to explore multi-modality of emotional integration in humans from different perspectives using various physiological response based measures. But how does the perceptual arousal to a music clip vary when the music video of the same audio content is presented to them, i.e. if the participants hear the same audio stimulus with and without the video element, how would their emotional associations and arousals vary? The present study has been designed to analyze the same through extensive audience responses and corroborate the results with nonlinear acoustic and time series analysis of the audio and video clips coupled with their EEG responses. A psychological experiment was conducted on 50 non-musician participants using 8 audio and 8 video clips representing two clips from each of the four intended emotional areas - happy, sad, calm, anxiety respectively. The audio clips chosen for this study consisted of the audio-only versions of the chosen video clips and were designed by replacing the visual parts of the video with black screen. All the participants were presented with a response evaluation sheet consisting of a given set of 8 conventional emotions belonging to both the positive and negative valence halves of Russell's 2D emotional sphere (Russell, 2003) and were asked to mark the appropriate emotions corresponding to each audio/video clip along with their respective intensities in a 5 point Likert scale. The audio and video clips were of about 30 seconds each and were presented to the audience in a shuffled manner with a 30 second gap in between two consecutive clips to minimize any remnant emotional elicitation from the previous clips. The resultant average intensity of emotional elicitation from each audio/video clip was evaluated and the emotional associations as well as their intensities were compared for each pair consisting of the audio and video versions of the same musical content. The psychological experiment was followed by an Electroencephalography (EEG) experiment performed on 5 participants, who were made to listen to the audio and video clips using the same protocol as the human response experiment. EEG measures the neuro-electrical responses originating from different lobes of the brain as a result of the neuronal interactions occurring due to presentation of multimodal and unimodal stimuli to the participants (Sanyal *et al.*, 2021, 81). The 1-D acoustic waveforms obtained from the audio signals were subjected to a cross correlation analysis with the upsampled 1-D EEG waveform obtained from the different electrodes using the conventional MFDXA technique (Sanyal *et al.*, 2019). The MFDXA technique provides a parameter λ_x which is essentially the degree of correlation between the source audio and the output EEG signals. Next, the video clips were converted to individual 1-D time series signals using Convolutional Autoencoders (Kingma *et al.*, 2013; Huang *et al.*, 2017, 1551-1561; Mukherjee *et al.*, 2019, 2027-2031; Castrejon *et al.*, 2019, 7608-7617; Liu *et al.*, 2021, 701-710) and the same MFDXA technique was applied to obtain the degree of emotional association and transfer between the video clips and the EEG signals. Thus, using perceptual and robust nonlinear methodology, this study attempts to provide new interesting information about the multimodality of emotional perception and cognition in the audio and visual domains, especially in the context of music clips. The study has been designed in such a way that it compares and analyzes the human emotional response to audio and video stimulus through extensive audience responses and corroborates the obtained results with nonlinear cross correlation analysis of the audio and video clips with neural data from EEG response.

2. EXPERIMENTAL DETAILS

2.1 *Selection of Clips :*

The sample data consisted of 8 popular movie clips of around 30 second each and correspondingly 8 audio clips generated from the video clips itself, of which 2 each belonged to the 4 opposite axes of Russell's 2-D emotional sphere (predicted class). Thus, for the predicted emotional classes, we have 2 clips each from Anxiety, Calm, Sad and Happy. These have been illustrated in the Figure 1 below:

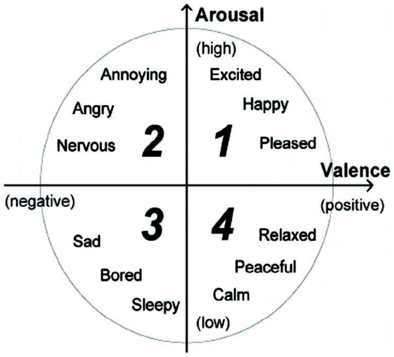
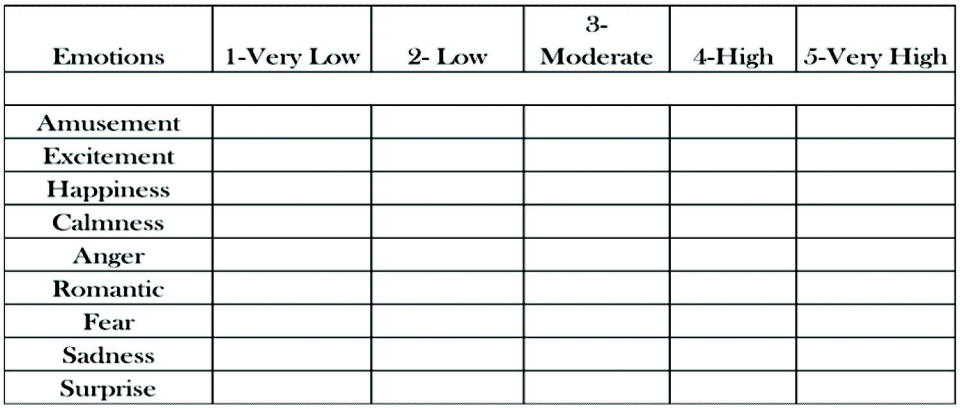


Fig. 1. Russell's emotional sphere with the 2 clips from each of opposite valence marked.

The audio-visual clips were chosen in such a manner that there was no conversations going on in the clip and a background score coupled with strong visuals were the main driving force.

2.2 *Perceptual Response Study :*

A human response survey was conducted on 50 participants (23 M, 27 F; average age = 25.7 years SD=6.25 years) who were asked to observe and listen to the AV (Audio Visual) and AO (Audio only) clips presented to them in a random manner. During the AO clips, a fixation cross was presented on a black screen, while the audio played in the background. The entire study was conducted in Online Mode using Google Forms and an informed consent was obtained from all the participants along with demographics prior to the study. Participants were asked to mark the emotional intensity of 8 audio and video clips in a scale of 1-5 (where 1 denoted lowest and 5 denoted highest intensity of a particular emotion) on a Response Sheet like the one shown in Fig. 2 below.



Emotions	1-Very Low	2- Low	3- Moderate	4-High	5-Very High
Amusement					
Excitement					
Happiness					
Calmness					
Anger					
Romantic					
Fear					
Sadness					
Surprise					

Fig. 2. Response sheet presented to the participants of human response study.

The participants were provided a palette of 9 standard musical emotions to choose from, the averaged intensity corresponding to each clip was calculated from them. 3 participants were rejected from the entire study design as their responses were detected as outliers from the complete set. 47 participants were finally considered for the complete survey.

3. EEG STUDY

3.1 *Subjects Chosen for EEG study :*

5 right handed adults (3 male and 2 female) voluntarily participated in this study. None of them had any conventional musical training. Neither reported any neurological disorders or auditory impairment. Their ages were between 25 to 30 years (SD=2.25 years). All experiments were performed at the Sir C.V. Raman Centre for Physics and Music, Jadavpur University, Kolkata. An informed consent obtained from each of them as per the guidelines set by Ethical Committee, Jadavpur University (Approval 3/2013).

3.2 *Experimental Protocol :*

Each subject was prepared with an EEG recording cap with 19 electrodes (Fig. 3) (Ag/AgCl sintered ring electrodes) placed in the international 10/20 system. Impedances were checked below 5 k Ohms. The EEG recording system (Recorders and Medicare Systems) was operated at 256 samples/s recording on customized software of RMS. The data was band-pass-filtered between 0 and 50 Hz to remove DC drifts. Each subject was seated comfortably in a relaxed condition in a chair in a shielded measurement cabin and 120 cm in front of a computer screen. They were asked to keep their eyes open all the time with gaze on a fixation cross on the computer screen.

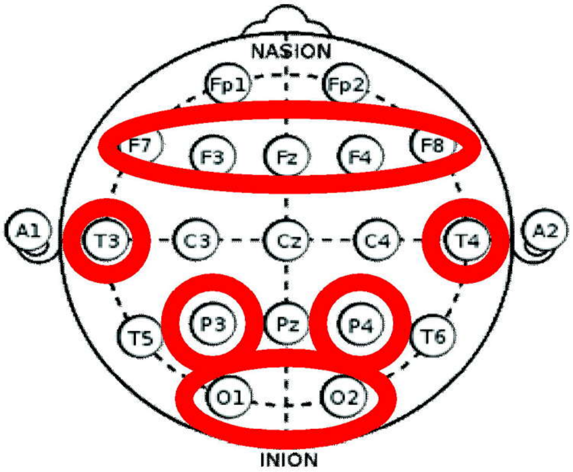


Fig. 3. The position of electrodes according to the 10-20 international system.

After initialization, an approx. 16 min recording period was started with the chosen audio and video clips in the following playing order

2 minutes (Resting/Baseline Condition)   30 seconds Clip 1 (Audio)   30 seconds Rest
↓
30 seconds Clip 1 (Audio)   30 seconds Rest   30 seconds Clip 2 (Video)
↓

… Same protocol being repeated for rest of the audio and video clips being presented to the participants in a random manner.

We identified four different lobes of the brain namely frontal, parietal, temporal and occipital whose functions tally with our work. From these lobes, electrodes F3/F4 (frontal), P3/P4 (parietal), O1/O2 (occipital) and T3/T4 (temporal) were chosen for our study and further analysis.

4. METHODOLOGY

4.1 *Average Intensity Values (from perceptual response) :*

The following procedure was followed to calculate the weighted average intensity of the 8 AO and AV clips provided to the participants. Out of 9 emotions provided (each with 5 intensity levels), one emotion was taken at a time. The various intensity levels for a particular emotion k were denoted by different values of variable ik (i.e. very low is represented by ik=1, low-2, moderate-3, high-4, and very high-5). The number of participants who marked any particular intensity (ik) for that particular emotion (k) was noted as nik. This number (nik) was then multiplied by the weight of the corresponding intensity (ik). The total intensity value of a particular emotion (k) was obtained by doing a summation over these calculated weighted values for each intensity (ik) . Finally, dividing this total intensity by the total number of responses (N) (combining all emotions) recorded for the clip under consideration, we get the weighted average intensity (AI)k for the particular emotion (k) for that selected clip. In essence, the following mathematical formula was used to calculate weighted average intensity value for each of the 9 emotions provided in the form:

$$(AI)_k = \Sigma_i(i_k n_{ik})/N \quad (1)$$

Where, i_k = Intensity rating for emotion 'k',
i_k € 1,2,3,4,5

k = h, s_1, c, a_1, s_2, d_1, e, a_2, r, d_2, f, which signify each of the emotions under survey:

(h=happy, s_1=sad, c=calm, a_1=anxiety, s_2=surprise, d_1=disgust, e=excitement, a_2=anger, r=romantic, d_2=devotion, f=fear)

$(AI)_k$ = weighted average intensity of emotion 'k' (for any particular clip)

n_{ik} = no. of participants who marked for any intensity i_k of emotion k

N = total no. of participants who responded for the particular clip

For each clip, average intensity values for all the 9 emotions were calculated individually by the same method as explained in Eqn. (1). This gave us an idea about how strongly different emotions were evoked for that particular clip. This procedure was followed for all the AO and AV clips to get the total set of average intensity values.

4.2 *Convolutional Autoencoders :*

Comprising of CNN based encoders and decoders, convolutional autoencoders have been widely used in the recent past for dimension reduction and noise reduction (Kingma *et al.*, 2013; Liu *et al.*, 2021, 701-710). In our case, each of the frames of a movie was compressed into respective low-dimensional representations by employing a convolutional auto-encoder. We chose 32 to be the low-dimensional representation. Therefore, for an entire movie consisting of N frames we have obtained a N x 32 representation. Finally, for each electrode, we perform Multifractal Detrended Cross Correlation analysis (please refer to next section for more details) between the corresponding EEG electrode time-series (of length N) and each of the 32 representations (of length N) thereby obtaining 32 cross-correlation values for each electrode. We have reported the mean of these values (please refer to Results section). The methodology employed has been shown in Fig. 4 in a nutshell.

4.3 *Multifractal Detrended Cross Correlation Analysis (MFDXA) :*

The MFDXA algorithm used here was proposed first by Zhou (2008,) and later applied in several works to compute the degree of correlation between sonified EEG signals and the input audio signals (Sanyal *et al.*, 2019, 13-31; 2021, 81; Nag *et al.*, 2021;). From the MFDXA technique, we have a scaling exponent λ(q), which essentially describes the amount of long range temporal correlations (LRTCs) present in the cross-correlated signal spectrum, q being the scale dependence present due to multifractality of the spectrum. A representative figure (Fig. 5) below reports the variation of cross correlation exponent λ(q) with q for two particular samples (Video Clip 1 and F4 electrode), also the variation of h (q) with q

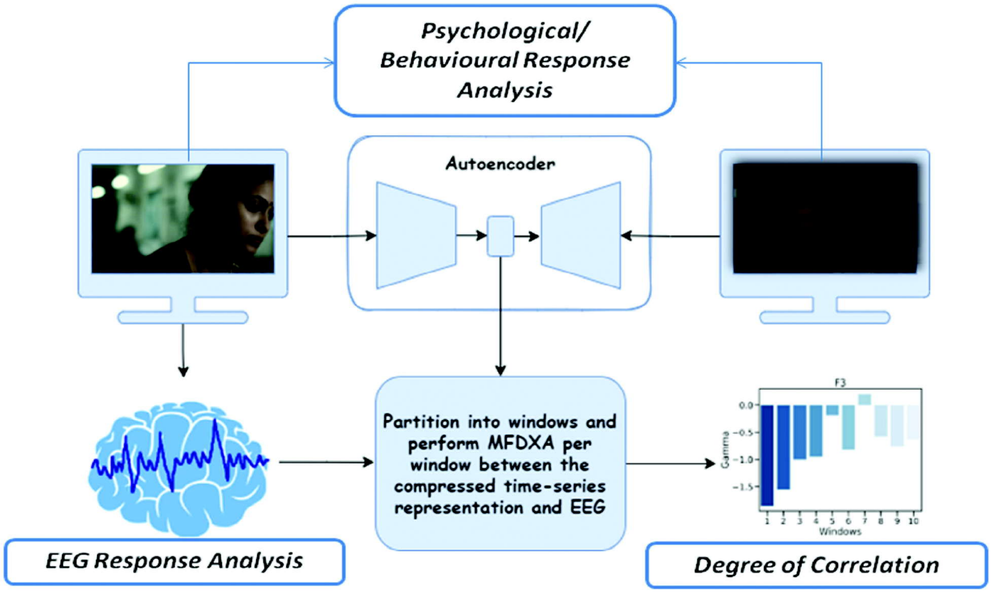

Psychological/ Behavioural Response Analysis
Autoencoder
Partition into windows and perform MFDXA per window between the compressed time-series representation and EEG
F3
0.0
−0.5
−1.0
−1.5
Gamma
1
2
3
4
5
6
7
8
9
10
Windows
EEG Response Analysis
Degree of Correlation

Fig. 4. Methodology employed for Autoencoders followed by MFDXA techniques.

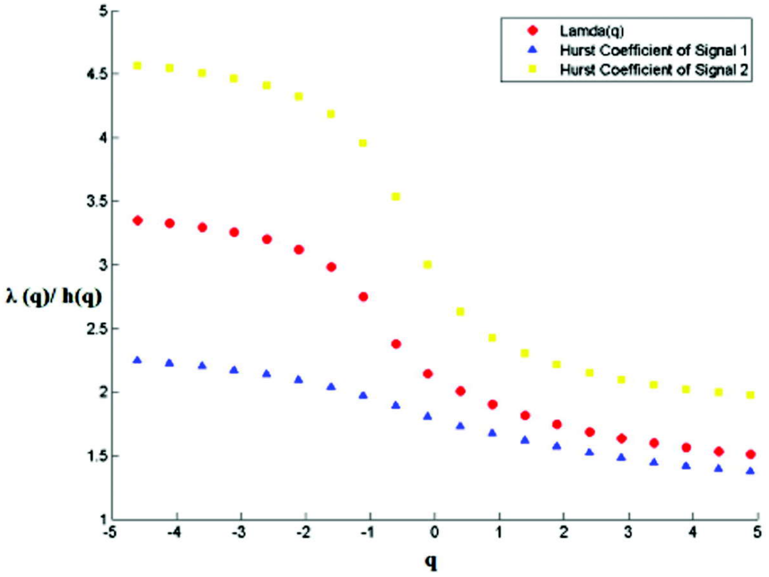


Fig. 5. Variation of λ (q) and h (q) for an EEG signal and video clip

for those two samples obtained from MFDFA technique are also shown in the same figure for comparison. Here h(q) being the generalized Hurst exponent of individual samples. Podobnik & Stanley (2008) have demonstrated the relation between cross-correlation exponent, γ_x and scaling exponent $\lambda(q)$ derived from $\gamma_x = 2 - 2\lambda(q = 2)$. For uncorrelated data, λ_x has a value 1 and the lower the value of γ and γ_x more correlated is the data. We know that $h(q) = 0.5$ indicates that the series is an independent random process, and for $h(q) < 0.5$ it is characterized by long-range anti-correlations while for $0.5 < h(q) < 1$, it is featured by long-term correlations (Ihlen, 2012, 141). In this case the signal is stationary. We computed the cross-correlation exponent, λ_x, between the different EEG electrodes and AO, AV samples belonging to the varied emotional classes.

The following figures (Fig. 6 a-d) represent the perceptual emotional response of the 8 pairs of AV and AO clips rated by 50 participants in a human response analysis test. Apart from the target class of emotions, all the other sets of emotions which were presented to the participants in the response sheet have been plotted in the graphs to look for the co-elicitation of other emotions along with the target emotion.

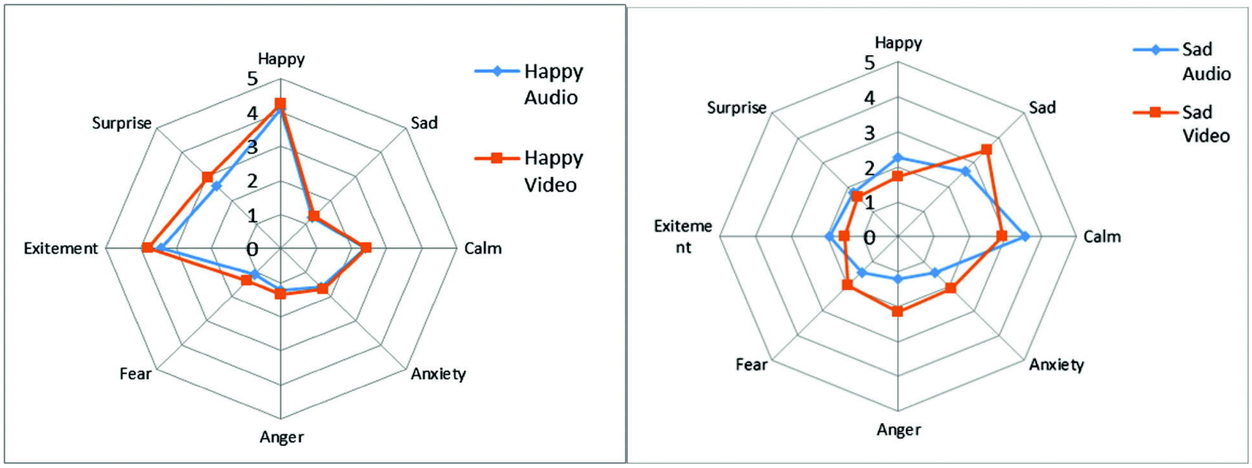


Fig. 6(a,b). Plot for clips belonging to the target class 'happy' and 'sad'

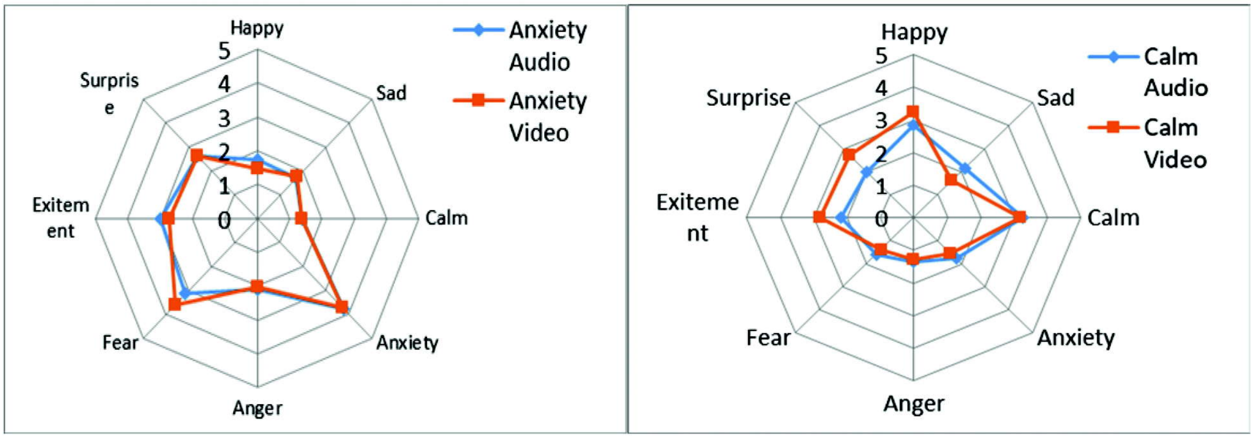


Fig. 6(c,d). Plot for clips belonging to the target class 'anxiety' and 'calm'

From the plots, we see that in general, a significant difference between the visual and auditory modalities is not observed from the perceptual study. The arousal based response corresponding to the target set of AV and AO clips are more or less comparable in both the domains. Another interesting response obtained from the plots is the co-elicitation of emotions, *i.e.* apart from the target class, there are a number of other emotions which have been simultaneously aroused in the AO-AV pairs. While for the 'happy' clips, the emotion 'excitement' is also found to predominate, for the 'anxiety' clips, the emotions 'fear', 'excitement' and 'surprise' are co-elicited. For the target class 'sad', the emotional class 'calm' coexists profoundly, while for the target class 'calm', 'sad' and 'anxiety' is found to co-exist. Thus, although from the perceptual study, we do not have a clear distinction in terms of the arousal based effects in AV and AO clips, we gather interesting information regarding co-elicitation of emotional classes simultaneously. To look further into the multi-modality of emotional appraisal and to corroborate these results, we next move on to analyze the results of the EEG study conducted using a similar protocol.

The nonlinear correlation between the EEG alpha wave and the time series generated from audio and movie clips were assessed with the help of Multifractal Detrended Cross Correlation (MFDXA) method. As already discussed in the methodology section, MFDXA gives a cross correlation exponent ?x, which is essentially the degree of correlation between the time series of movie clips/ audio clips and the EEG signals obtained from the 8 electrodes chosen. In this way, we look forward to establish the differential neural response corresponding to the use of multi-modal stimuli. The following figures (Fig. 7 a-d) report the variation of averaged degree of correlation in the chosen eight (8) electrodes for the four (4) target emotional class, while the participants were subjected to the AO and AV clips in the aforementioned experimental protocol.

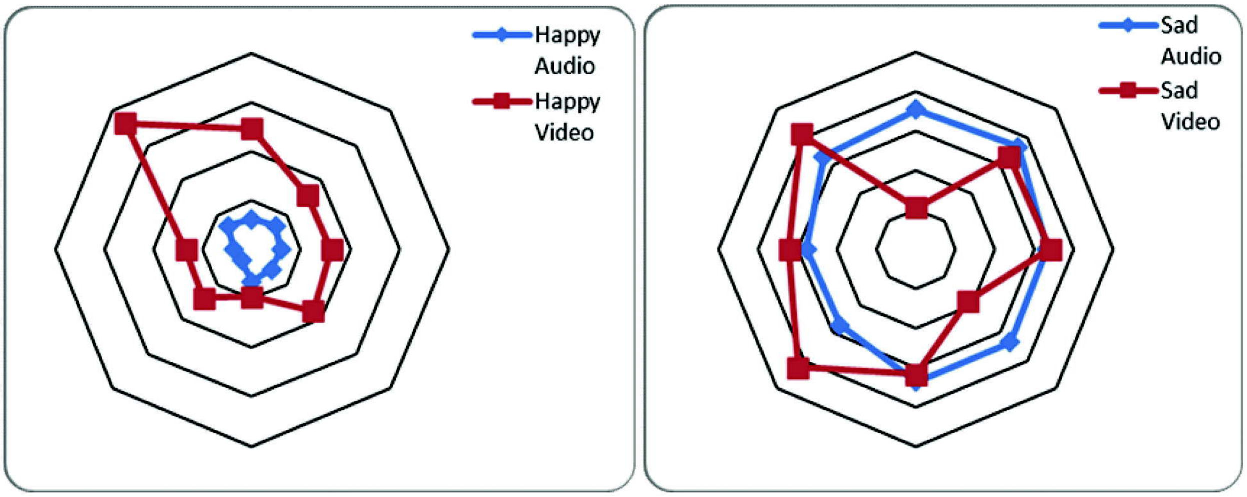


Fig. 7(a,b). Degree of correlation in the 8 electrodes for clips belonging to the target class 'happy' and 'sad'

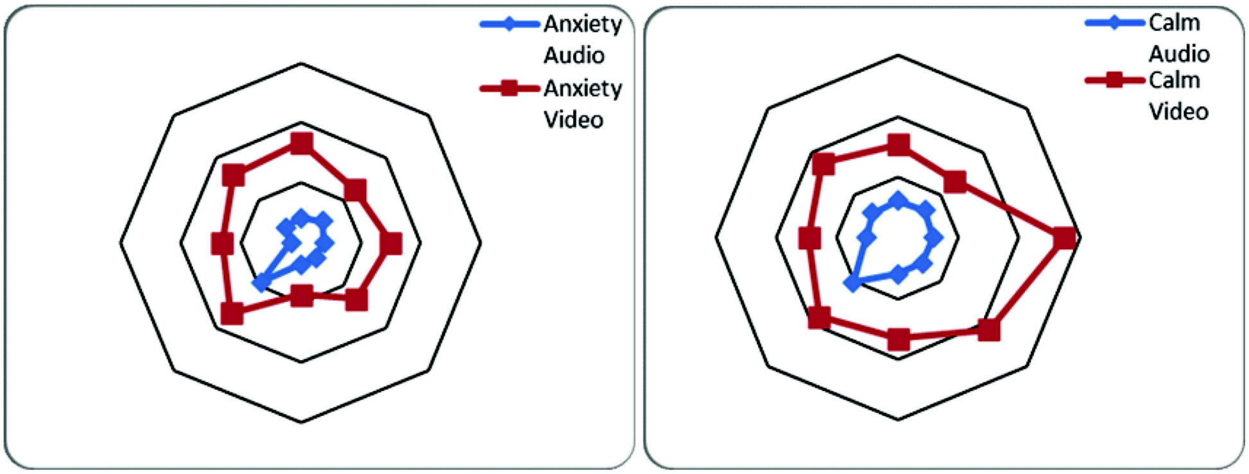


Fig. 7(c,d). Degree of correlation in the 8 electrodes for clips belonging to the target class 'happy' and 'sad'

From the figures, it is evident that for all the emotional valences, the cross-correlation coefficient is found to be higher for audio clips as compared to the video clips corresponding to all electrodes in general. Since a greater negative value implies a stronger correlation, hence the AO clips are the ones which show the strongest degree of correlation in most of the electrodes. In case of the target class 'happy', for both AV and AO clips, a strong dip in correlation is noticed for right temporal electrode. Lower correlation is observed in occipital electrodes corresponding to happy AV clips. In sad emotional valence, it is seen that the degree of correlation is higher for AV clips, indicating higher appraisal of negative valence for visual domain. A strong correlation is observed in the frontal and occipital electrodes indicating the processing of negative valence. Interestingly, the AV clip belonging to the target class 'calm' shows a dip in cross correlation scaling corresponding to occipital electrodes, while the calm audio clips have very high correlation for all the electrodes. For clips belonging to target class 'Anxiety', AO clips show strong correlation in temporal electrodes, while the AV clips report a dip in correlation corresponding to frontal and parietal electrode. In this manner, we have tried to establish a differential response corresponding to how musical emotions are correlated with different lobes of brain when we have audio visual clips and audio clips extracted from the same video.

5. CONCLUSION AND FUTURE DIRECTIONS :

The study presents the following interesting conclusions:

1. In the perceptual domain, the cognition of multi-modal emotion from audio and audio-visual domain do not provide significant difference, although there are significant signs of co-elicitation of emotions, where several other emotional classes are co-elicited with the target class.
2. In the neural domain, the cross correlation exponent generated from the MFDXA technique computed the degree of correlation of the source audio/ video clips and the output EEG data. This shows significant difference in arousal based activities corresponding to the two modalities.
3. While the degree of correlation is significantly high for the audio clips, the video clips provide a lower value. The audio and video clips for the target class "happy", which co-elicited with "exciting" show significant correlation in the frontal and occipital lobes. For the target class "anxiety", strong correlation is seen in the temporal lobe for AO clips, while for AV clips, the parietal electrodes have high degree of correlation
4. The "sad valence" clips provide a unique exception from this phenomenon where the cross-correlation is higher for the AV clips in specific electrodes

Thus, for the first time we report a direct correlation between the input audio/visual stimuli and the output EEG response obtained from different lobes. The differential processing of emotional appraisal generated from audio and visual modalities is reported here. In the neural domain, distinct correlation characteristics is obtained corresponding to the audio and video clips, indicating more robust studies need to be done to establish unique brain response corresponding to the two domains

Future works in this direction include nonlinear fractal analysis of the source characteristics, i.e. quantification of the time series data obtained from the source audio and video signals. This would help in gaining more insights into the time series evolution of the two types of input modalities. Also, a quantification of the 1-D time series data obtained from convolutional autoencoders would be done in this manner. Cross-correlation analysis of the two modalities, i.e. MFDXA has to be performed also between the time series data of similar audio and video clips. This would lead to better understanding of how the two are correlated and better interpretation of the neural observations in this light can be made. The experiment can be repeated with greater number of participants as well as more variety of emotions taken from Russell's 2D emotional sphere. This pilot study would act as a precursor to all these fascinating studies in the domain of automated emotion identification using brain computer interface (BCI) systems.

6. ACKNOWLEDGEMENTS

SS (DST/CSRI/2018/78) and AB (SR/CSRI/PDF-34/2018) acknowledge DST, Govt of India for providing the DST-CSRI (Cognitive Science Research Initiative) Post Doctoral Fellowship to pursue this research. MB (Ref No.- D-7/Sc/231/21 ; Index No.- 54/21/Phys./27) thankfully acknowledges the Department of Science and Technology (DST), Govt. of India, for providing the DST Inspire Fellowship to pursue this research work.

7. REFERENCES

[1] Bradley, Dinah. *Dynamic Breathing*. Hachette UK, 2011.

[2] Castrejon Lluis, Nicolas Ballas and Aaron Courville, 2019. "Improved conditional vrnns for video prediction." *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7608-7617.

[3] Chakraborty Sayantan, Sourav Samanta, Shukla Samanta, Dipak Ghosh and Kumardeb Banerjee, 2022. "Complexity Analysis of Wind Energy, Wind Speed and Wind Direction in the light of nonlinear technique." *arXiv preprint arXiv:2206.14582.*

[4] Chapados Catherine and Daniel J. Levitin, 2008. "Cross-modal interactions in the experience of musical performances: Physiological correlates." *Cognition,* **108**(3), 639-651.

[5] Chen Xuhai, Lingzi Han, Zhihui Pan, Yangmei Luo and Ping Wang, 2016. "Influence of attention on bimodal integration during emotional change decoding: ERP evidence." *International Journal of Psychophysiology,* **106,** 14-20.

[6] Ghosh Dipak, Sayantan Chakraborty and Shukla Samanta, 2019. "Study of translational effect in Tagore's Gitanjali using Chaos based Multifractal analysis technique." *Physica A: Statistical Mechanics and its Applications,* **523,** 1343-1354.

[7] Ghosh Dipak, Srimonti Dutta, Sayantan Chakraborty and Shukla Samanta, 2018. "Chaos based nonlinear analysis to study cardiovascular responses to changes in posture." *Physica A: Statistical Mechanics and its Applications,* **512,** 392-403.

[8] Huang Heng, Xintao Hu, Yu Zhao, Milad Makkie, Qinglin Dong, Shijie Zhao, Lei Guo and Tianming Liu, 2017. "Modeling task fMRI data via deep convolutional autoencoder." *IEEE transactions on medical imaging,* **37**(7), 1551-1561.

[9] Ihlen Espen AF, 2012. "Introduction to multifractal detrended fluctuation analysis in Matlab." *Frontiers in physiology,* **3,** 141.

[10] Kingma Diederik P. and Max Welling, 2013. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114*.

[11] Lang Peter J. and Margaret M. Bradley, 2010. "Emotion and the motivational brain." *Biological psychology,* **84**(3), 437-450.

[12] Lin Tzu-Kang and Yi-Hsiu Chien, 2017. "A structural health monitoring system based on multifractal detrended cross-correlation analysis." *Structural engineering and mechanics: An international journal* **63**(6), 751-760.

[13] Liu Bowen, Yu Chen, Shiyu Liu and Hun-Seok Kim, 2021. "Deep learning in latent space for video prediction and compression." *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 701-710.

[14] Mukherjee Subham, Spandan Ghosh, Souvik Ghosh, Pradeep Kumar and Partha Pratim Roy, 2019. "Predicting video-frames using encoder-convlstm combination." In ICASSP 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 2027-2031.

[15] Nag Sayan, Uddalok Sarkar, Shankha Sanyal, Archi Banerjee, Souparno Roy, Samir Karmakar, Ranjan Sengupta and Dipak Ghosh, 2021. "A Fractal Approach to Characterize Emotions in Audio and Visual Domain: A Study on Cross-Modal Interaction." *arXiv preprint arXiv:2102.06038*.

[16] Platz Friedrich and Reinhard Kopiez, 2012. "When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance." *Music Perception: An Interdisciplinary Journal,* **30**(1), 71-83.

[17] Podobnik Boris and H. Eugene Stanley, 2008. "Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series." *Physical review letters,* **100**(8), 084102.

[18] Podobnik Boris, Zhi-Qiang Jiang, Wei-Xing Zhou and H. Eugene Stanley, 2011. "Statistical tests for power-law cross-correlated processes." *Physical Review E* **84**(6), 066118.

[19] Roy Souparno, Archi Banerjee, Chandrima Roy, Sayan Nag, Shankha Sanyal, Ranjan Sengupta and Dipak Ghosh, 2021. "Brain response to color stimuli: an EEG study with nonlinear approach." *Cognitive Neurodynamics,* **15**(6), 1023-1053.

[20] Sanyal Shankha, Archi Banerjee, Ranjan Sengupta and Dipak Ghosh, 2016. "Chaotic Brain, Musical Mind-A Non-Linear eurocognitive Physics Based Study." *Journal of Neurology and Neuroscience* pp. 1-7.

[21] Sanyal Shankha, Sayan Nag, Archi Banerjee and Dipak Ghosh, 2021. "Now You See Me: Identifying non-linear cross-correlation between audio-visual and EEG signals." *In perception,* **50**(1), 81-81. 1 Olivers Yard, 55 City Road, London EC1Y 1SP, England: Sage Publications Ltd.

[22] Sanyal Shankha, Sayan Nag, Archi Banerjee, Ranjan Sengupta and Dipak Ghosh, 2019. "Music of brain and music on brain: a novel EEG sonification approach." *Cognitive neurodynamics,* 13(1), 13-31.

[23] Vines Bradley W., Carol L. Krumhansl, Marcelo M. Wanderley, Ioana M. Dalca and Daniel J. Levitin, 2011. "Music to my eyes: Cross-modal interactions in the perception of emotions in musical performance." *Cognition,* **118**(2), 157-170.

[24] Vuoskoski Jonna K., Elia Gatti, Charles Spence and Eric F. Clarke, 2016. "Do visual cues intensify the emotional responses evoked by musical performance? A psychophysiological investigation." *Psychomusicology: Music, mind and brain,* **26**(2), 179.

[25] Wu Xun, Wei-Long Zheng, Ziyi Li and Bao-Liang Lu, 2022. "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition." *Journal of neural engineering,* **19**(1), 016012.

[26] Zhang H., 2020. Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder. *IEEE Access,* **8,** 164130-164143.

[27] Zhang Jianhua, Zhong Yin, Peng Chen and Stefano Nichele, 2020. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review." *Information Fusion,* **59,** 103-126.

[28] Zheng Wei-Long, Bo-Nan Dong and Bao-Liang Lu, 2014. "Multimodal emotion recognition using EEG and eye tracking data." In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, *IEEE,* pp. 5040-5043.

[29] Zhou Wei-Xing, 2008. "Multifractal detrended cross-correlation analysis for two nonstationary signals." *Physical Review E* **77**(6), 066211.

The Journal of Acoustical Society of India : Vol. 50, No. 1, 2023 (pp. 67-75)

Application of timbral audio descriptor models for the identification of bird species

Shambhavi Shete, Saurabh Deshmukh, Anjali Burande,
Saurav Padghan, Amit Nitnaware and Piyush Sonkamble
MIT Aurangabad, India
e-mail: shambhavishete2020@gmail.com

[Received: 19-03-2022; Revised: 18-05-2022; Accepted: 06-06-2022]

ABSTRACT

Bird species identification is one of the popular applications of Sound Information Retrieval systems. The process of identifying the specific bird from the recorded sounds of similar or different bird species is a crucial task. Birds are good indicators of environmental health. Bird Species identification helps us to understand more about environmental health and regional biodiversity. Manual analysis of bird species through the recorded sound of the birds may be accurate but it is very time-consuming and ambiguous to identify different birds from the same species. The automatic bird detection system suffers from ubiquitous characteristics of bird vocalizations. This includes variations of the bird sound vocalizations within and across species, the variability of the recordings, etc. The variability of the bird's sound depends on different levels of noise, the environment, and the simultaneous vocalizations. Due to the growing size of the environmental recordings, it is necessary to develop an accurate and efficient approach to analysing the bird species based on various attributes of the sound produced by the bird. In this paper, we have proposed a novel method to identify a bird from the Timbral attributes of the sound produced by the bird. The timber of a bird sound is its dimension which is considered as a non-tangible fourth dimension of the sound that has no MKS, CGS, or SI units assigned. We have used a novel method to categorize the bird's sound based on the timbral attributes of the sound produced by the birds. In contrast with the existing system, we have implemented a model of Timbral attributes of the sound for each bird to uniquely categorize the bird from not only the other birds but also from the other similar species of the same birds. The application of the unique timbral model for each bird helps identify not only the type of bird but also individual birds from the same species. The results show that with the application of a feed-forward neural network and proposed Timbral Audio Descriptor modelling of each bird species an accuracy of 94 % is achieved for the Identification of bird species.

1. INTRODUCTION

Birds are living beings that beautify nature. They play important role in maintaining the natural balance of the cycle of nature. Birds maintain the forest cover through seed distribution and pollination. Thus, making them an important link in the natural food chain. However, due to climate change and habitat destruction, many bird species are facing the threat of population decline. Ornithologists faced many problems in the identification of bird species due to several climate changes and the encroachment of human beings over the jungles. The Ornithologist studies all the details of birds such as their presence in the environment, their biology, distribution, impact on the environment, various bird species, etc. The identification of birds is usually done by bird specialists based on the divisions proposed by Linnaeus: Kingdom, Phylum, Class, Order, Family, and Animals (CK-12 2021). Manual surveying of all the birds living in their natural habitat is difficult as the birds occupy a different range of habitats and frequently change their habitat based on the season. The manual survey is time-consuming, expensive, and requires expertise to classify the birds.

To automatically recognize a bird based on its chirp requires a great extent of signal processing and machine learning techniques. There is a fundamental difference and similarity between sound produced by a bird and a human voice. Syrinx is the organ, located at the caudal end of the trachea, through which all the birds produce sound. It uses all the air that passes through it. On the other hand, a human creates sound using only 2% of the air exhaled through the larynx (Note 2021).

Similar to syrinx in birds which is part of its respiratory system, the larynx is a part of the human respiratory system through which when air is passed the sound is produced. The vocal cord in the human throat vibrates as the air passes over it. In short, the sound production system in humans and birds is similar in one aspect yet has a difference in them. The functional and morphological differences in the physical process of sound generated by birds and humans are at the peripheral level.

On a larger scale, there are similarities in the observations of bird sounds and human speech. For the detection of human speech various analysis of the human voice is made using audio descriptors and analyzers such as Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding Coefficient (LPCC), and different temporal and spectral features of the sound. Similar to the human speech recognition system, sounds of the birds could be identified by applying signal processing techniques for feature extraction and classification.

The birds communicate by creating various audio chirps.

These chirps are different according to the situation. Some of them are high pitch and some of them are low pitch. Birds generate sound according to the situation in which they are. Birds exchange a variety of information for each voice. Birds can transmit various warnings of impending danger using the sounds they produce.

Automatic identification of Bird species from their chirps involves audio preprocessing, audio feature extraction, and training of machine learning algorithms. Other than temporal and spectral audio features, in recent years a trend of using Timbral audio descriptors has been established for the applications of music and speech. This research proposes the application of Timbral audio descriptors for the extraction of features from bird sound for the identification of bird species using the Feed Forward Neural Network (FFNN). The results are cross-validated using 10-fold cross-validation using Multi-Support Vector Machine (Multi-SVM).

This manuscript is organized in various sections. The literature survey related to automatic bird species identification from voice is emphasized in section number 2. Section 3 elaborates on the proposed automatic bird identification system using machine learning. This section explains the step-by-step execution of the process along with the proposed way of calibrating the performance of the system. Various experiments with different audio descriptor sets and various classifiers along with the analysis of the results are explained in section 4. Section 5 concludes the research conclusions related to the application of Timbral audio descriptor models for the identification of bird species.

2. LITERATURE SURVEY

Birds are very useful ecological indicators since they respond very quickly to changes in the environment. Bird species identification and monitoring are challenging even after advancements in technology because of the presence of dynamic variations in the environment, seasons, and behavior of birds in different atmospheric changes. There have been many popular challenges given to the technocrats to identify a bird species from their environmental audio recordings. The popular challenges are MLSP 2013 (Forrest, et al. 2013), NIPS4B 2013 (Lasseck 2013), and BirdCLEF 2016 (Goeau, et al. 2016). In the Avesound project, more than 24600 bird sound samples containing sounds of 999 different bird species have been analyzed using spectrogram and classified using a Convolutional Neural Network. The author concluded that a deep residual neural network can be used to classify bird species based on acoustical data recordings. The results show that multiple width frequency delta data augmentations may not increase the classification accuracy. Also, the usage of additional metadata increases the predictability of the model (Martinsson 2017).

The sound of the birds could be divided into hypothetical classes such as bird songs and bird calls. The bird songs are generally produced by male birds and calls are produced by both male and female birds throughout the year. Typically, a syllable is a suitable unit for the recognition of the bird species. The sound of the bird is inharmonic. The segmentation is difficult because concurrent syllables are extracted from the raw recordings. The sound of the bird is segmented based on the onsets into different syllables using short-time signal energy and a short-time maximum of the spectrum. A total of nineteen low-level acoustical features have been used for the identification of 300 bird species. Short-time behavior is measured by the seven spectral features, means, and variance and five features useful to describe the static properties of the birds. The bird species identification has been done with K- Nearest Neighbor (K-NN) classifier. Euclidean and Mahala Nobis distances are calculated which gave the system accuracy of 53% and 82% respectively for the identification of the bird species (Fagerlund 2004).

Nowadays, wildlife biologists face the public policy issue of the interaction of birds with wind turbines. To handle this problem, the acoustical behavior and characteristics of the birds need to be monitored. To solve this issue nocturnal bird flight calls species are identified to prevent the negative effects of the wind farms. The audio feature extraction included Spectrogram-based Image Frequency Statistics (SIFS) feature extraction algorithm with the comparison of the features extracted from Discrete Wavelet Transform (DWT) and Mel Frequency Cepstral Coefficients (MFCC). Four different classifiers namely, K-NN, Multilayer Perceptron (MLP), Hidden Markov Models (HMM), and Evolutionary Neural Network (ENN) were deployed to identify the bird species. The results show that with a mixture of MFCC and SISF features, the MLP gives 92% accuracy for bird species identification (Bastas 2011).

The application of the Hidden Markov Model (HMM) for bird species identification was also one of the most promising types of research. The system uses a recognition system that uses two HMMs. A modified Dynamic Time Wrapping (DTW) technique was applied to extract audio features from the audio of over 38 hours of natural field recordings, consisting of 48 bird species. The results showed that bird species identification accuracy of over 93% could be achieved using the proposed two sets of the Hidden Markov Model system (Zakeri 2017).

The bird songs and bird calls are confusing and usually misinterpreted. The singing is limited to songbirds, but non-songbirds can also sing. The diversity of the sound produced by birds is large and in the range of fundamental frequency between 100 Hz to 1 kHz. With the presence of two independent vibrating membranes in the syrinx, birds can produce two independent carrier waves. The characteristics of simple voice sound are fundamental frequency and harmonics. Various bird species have two sound sources to produce the sound. The sound can be produced by the birds in different situations using either of the membranes alone, using two membranes together, or by switching between sound sources from one membrane to another. Therefore, a single feature extraction system cannot become a thumb rule to catch hold of the audio attributes of a bird sound.

In this research, we proposed a feature extraction system based on the Timbre of the sound produced by the bird. Timbre is a nontangible attribute of sound that uniquely describes it (Park 2004). Various audio descriptors have been categorized and applied for different sound information retrieval applications such as speech and speaker identification (Thiruvengatanadhan 2017) (Serizel and Giuliani 2017), and music information retrieval applications such as automatic singer identification (Deshmukh and Bhirud, North Indian classical music's singer identification by timbre recognition using MIR toolbox 2014), gender detection (Deshmukh and Bhirud, A Novel Method to Identify Audio Descriptors, Useful in Gender Identification from North Indian Classical Music Vocal 2014), automatic Tabla stroke identification (Shete and Deshmukh, Analysis and Comparison of Timbral Audio Descriptors with Traditional Audio Descriptors Used in Automatic Tabla Bol Identification of North Indian Classical Music 2019) (Shete and Deshmukh, Automatic Tabla Stroke Source Separation Using Machine Learning 2021), etc. Typically, sound features are extracted in the time and frequency domain, however, to identify minute differences between sound produced by birds of different categories and from the same category, special audio features are to be uniquely characterized to identify the bird species.

3. PROPOSED SYSTEM

The bird species identification system is implemented here using twelve different species of birds. The birds are conceptually categorized into two groups, singing birds that produce prolonged sounds and birds that produce chirping sounds. The purpose of categorizing the bird species into these groups is to relate the usefulness of Timbral audio descriptors for the identification of the bird species. By considering the diversity in the sound produced by a bird in different situations and environments and distinguishing between different birds with the help of sound produced by them, a careful approach to selecting appropriate audio descriptors is proposed here.

Figure 1. shows the Timbral audio descriptor model used to identify the bird species. The birds sound database is preprocessed using onset detection and segmentation to create individual samples of each bird. A total of sixty samples of twelve birds are segmented and saved using the Audacity software tool giving a total of 720 audio excerpts used for training. A total of 358 different audio excerpts are used to test the bird species identification system. Each audio sample is a mono audio recording represented using 16-bit pulse code modulation (PCM) .wav file format. The sampling frequency used is 44100 Hz.

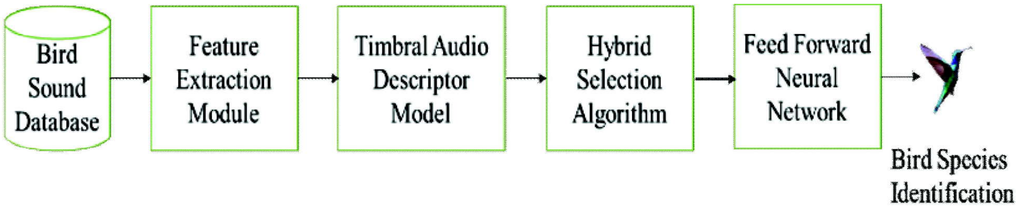

Bird Sound Database
Feature Extraction Module
Timbral Audio Descriptor Model
Hybrid Selection Algorithm
Feed Forward Neural Network
Bird Species Identification

Fig. 1. Timbral Audio Descriptor Models for the Identification of Bird Species.

From the existing many audio descriptors that are categorized into temporal and spectral audio features, there exist many approaches to select appropriate audio descriptors that uniquely define the Timbre. The taxonomy used to declare Timbral audio descriptors namely, Attack Time, Attack Slope, Zero Crossing Rate (ZCR), Roll-off, Roughness, Brightness, and Irregularity along with MFCC is applied here (Lartillot n.d.). A Feed Forward Neural Network (FFNN) is trained using extracted audio descriptors for the classification of sound produced by twelve birds. The sound of each bird is modeled by applying an appropriate set of audio descriptors useful for the identification of the sound based on its Timbral attribute. The Hybrid Selection Algorithm with its wrapper approach is applied here that considers the selection of appropriate audio descriptors based on the accuracy obtained from the classifier (S. H. Deshmukh 2012). The Neural Network is trained using the Levenberg-Marquardt algorithm (LM) and the mean squared error (MSE) is calculated.

The performance of the proposed system is calibrated utilizing percentage accuracy obtained in the identification of bird species based on the Timbral model of the sound. The results are compared with the results of the Multi-Support Vector Machine (Multi-SVM) Classifier. The Multi-SVM is used to classify the birds by using one versus the rest of all policies.

4. RESULTS AND DISCUSSIONS

The audio excerpts are segmented and separately saved generating the training audio data. The saved audio excerpts are preprocessed to remove noise and normalize amplitude. Each audio excerpt containing bird sound is passed through the onset detection and segmentation process to separate each chirp of the bird's sound. Fig. 2. shows the onsets detected from two different sound samples of the same bird. Segmentation of audio excerpt is done based on the amplitude of the onset value.

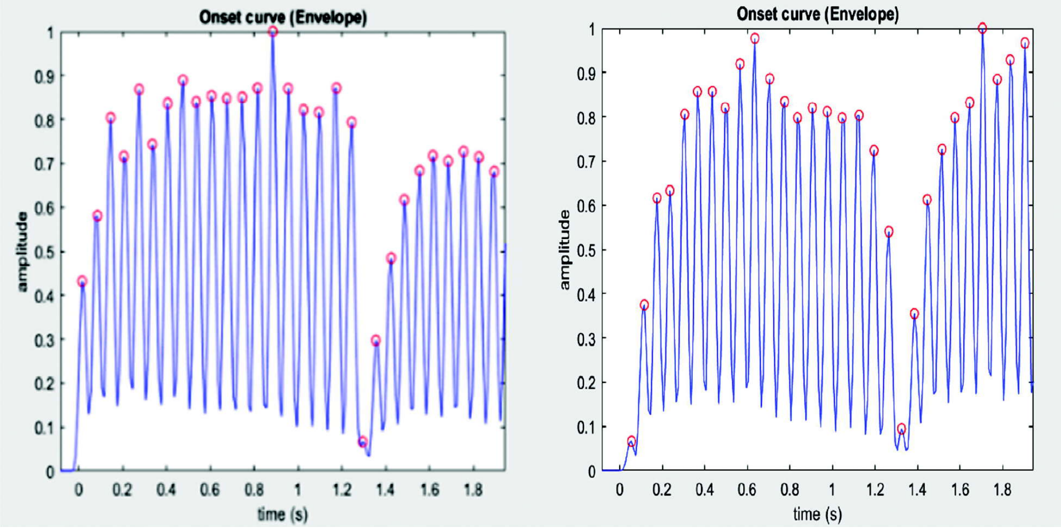


Fig. 2. Onsets of Audio Samples of the Same Bird.

It is observed that the onsets of the sounds produced by the birds (sparrow) which make the chirps of short duration with a sharp note and high pitch sound are consecutively placed near to each other. On the other hand, there are birds (nightingale) that produce long notes of sound which give long distant locations of the onsets. Fig. 3. shows onsets detected from sound excerpts belonging to two different birds, out of which one bird is producing a series of short chirps and the other is producing comparatively long high-pitched sounds.

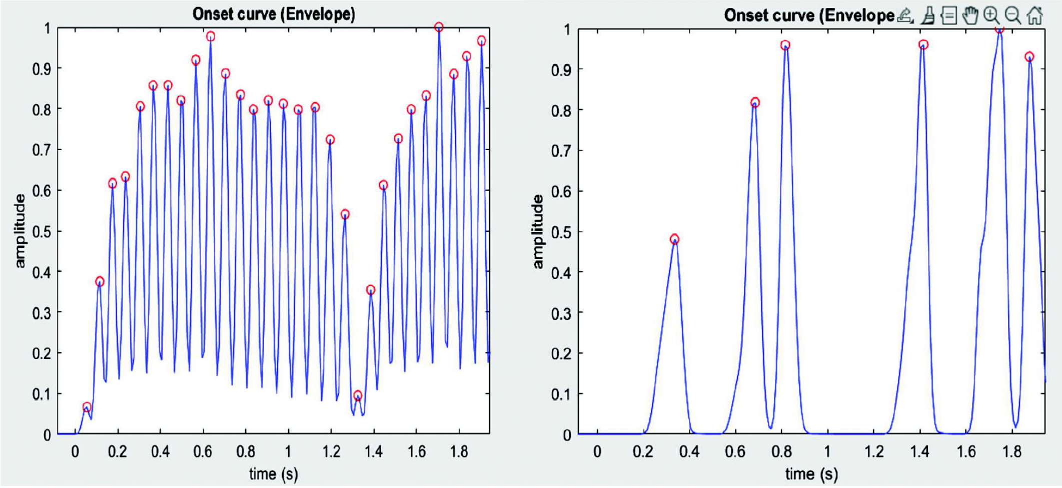


Fig. 3. Onsets of Audio Samples of Different Birds.

Audio segmentation is an algorithm designed to automatically reveal semantically meaningful temporal segments present in an audio signal. The visual representation of segmented audio of the same bird from a long recording of natural bird sounds is shown in Fig. 4.

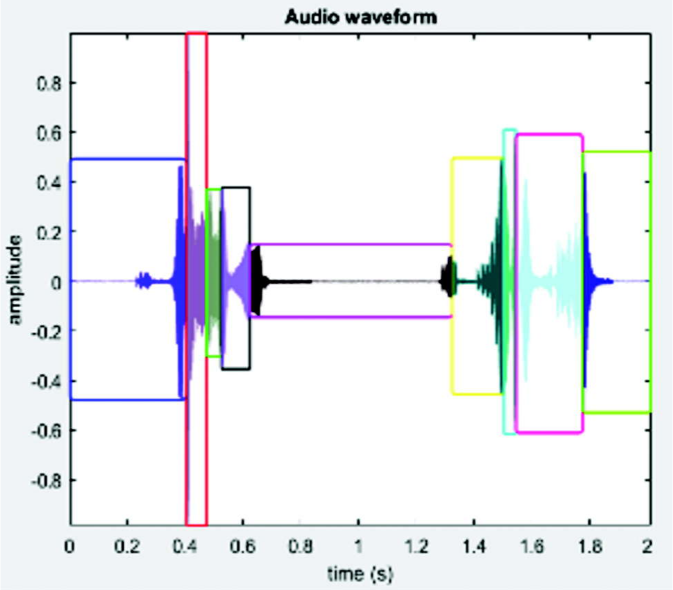


Fig. 4. Visual Segments of Same Bird Sound Samples.

Timbral audio features are extracted from all the audio excerpts and Hybrid Selection Algorithm is applied to select a set of most relevant audio descriptors for the identification of bird species. During the training of the neural network, a total of twenty audio descriptors is given as input to the Feed Forward Neural Network (FFNN) and with the help of one hidden layer and twelve neurons in the output layer corresponding to the twelve birds. A 10-fold cross-validation method is applied to cross-verify the performance of the trained neural network for the audio excerpts not given during training. The system is evaluated using percentage accuracy obtained from the classifier to correctly identify the bird species. Multi-SVM classifier is also used to compare the bird species identification accuracy, where one versus the rest all binary classification is used for all twelve bird species.

The system performance error is calculated based on misclassified audio excerpts in percentage. The results are calculated using the traditional MFCC audio feature and using a set of timbral audio descriptors. A comparison of % errors calculated for Singing and Chirping birds using Timbral and MFCC features obtained from FFNN and Multi-SVM is shown in Table 1. The results show that with the application of the Hybrid Selection Algorithm, a reduced set of Timbral audio descriptors is derived for each category of the twelve bird species. The Timbral audio descriptor modeling for each category gives the lowest error rate using FFNN.

For the singing bird category, where a single prolonged voice is produced by the bird has reduced values of zero-crossing rate, on the other hand, the brightness audio feature plays a prominent role in identifying the bird species of this category. Similarly, when the bird produces repetitive chirps, the values of zero-crossing rate and irregularity play a prominent role in distinguishing the birds that produce chirps. The application of one versus rest all strategy for Multi-SVM using Timbral audio descriptor supersedes its performance over MFCC coefficients used as audio features. Fig. 5. shows the graph of bird species identification errors obtained from FFNN and Multi-SVM with and without using Timbral audio descriptors.

Table 1. Comparison of bird species identification errors.

Type of Birds		Bird Species Identification % Errors			
	Name of Birds	FFNN		Multi-SVM	
		Timbre	MFCC	Timbre	MFCC
Singing Birds	Nightingale	2.70	10.81	18.92	21.62
	Mockingbird	6.67	10.00	26.67	23.33
	Asian Koel	7.41	22.22	18.52	14.81
	Peacock	5.56	19.44	11.11	16.67
	American Robin	4.76	19.05	14.29	16.67
	Canary	6.25	25.00	21.88	25.00
Chirping Birds	Sparrow	9.09	22.73	13.64	9.09
	Ostrich	7.14	21.43	21.43	32.14
	Chaffinch	0.00	20.00	12.00	16.00
	Parakeet	3.23	12.90	19.35	22.58
	Parrot	8.70	21.74	8.70	17.39
	Woodpecker	12.00	16.00	20.00	24.00

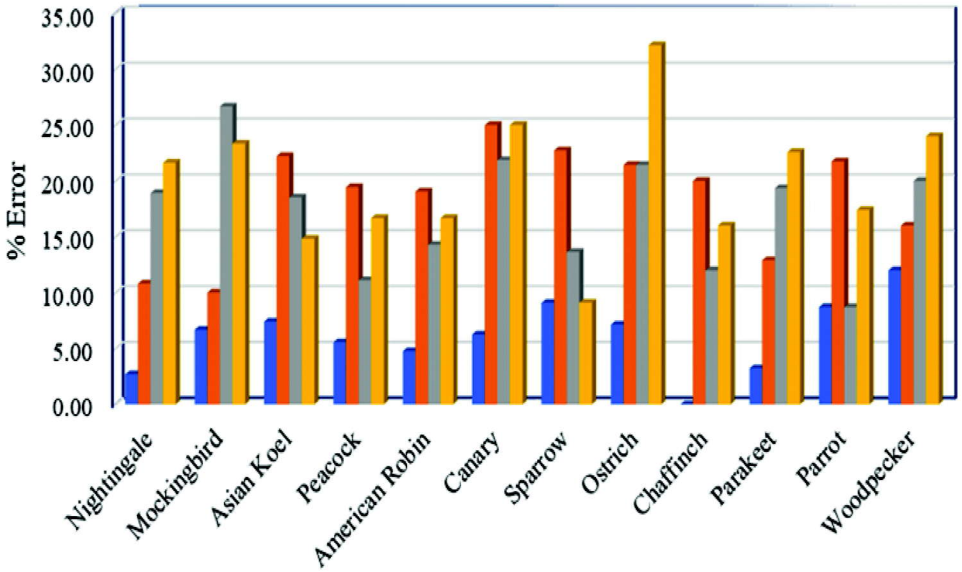


Fig. 5. Comparison of Bird Species Identification Error.

It is observed that for the singing bird category Timbral audio descriptor model containing roll-off, brightness, roughness, and MFCC using the FFNN classifier gives 94.61% accuracy and for the chirping bird category the Timbral audio descriptor model containing ZCR, irregularity, and MFCC gives 93.51% accuracy. The overall average accuracy of the system is found to be 94% using the Timbral audio descriptor model.

5. CONCLUSION

Application of Timbral Audio Descriptors models for Identification of Bird Species is presented here. Automatic Bird Species Identification is of great interest to ornithologists who face many challenges to keep track of the bird population due to climate changes. The birds, like humans, produce sounds using syrinx. Broadly, birds could be classified as singing birds and chirp birds. The singing bird's sounds are prolonged monophonic melodies while the chirps are of short duration and repetitive. Timber is the fourth dimension of the sound which is non-tangible and unique for each source of the sound. In the existing system of automatic bird species identification, popularly, Mel-Frequency Cepstral Coefficients (MFCC) are used as audio features. For a wrapper audio feature selection approach, the application of the Hybrid Selection Algorithm for the selection of the appropriate set of Audio descriptors which represent timbral attributes of the sound enhances the system performance in terms of bird species identification accuracy. The feature extraction process extracts a total of 20 audio descriptors (7 Timbral+13 MFCC) from the audio excerpts. An audio dataset designed using audio recordings of twelve birds (six singing and six chirping birds) that contains 1078 audio excerpts is used here (780 Training samples and 358 test samples). Feedforward Neural Network (FFNN) and Multi-SVM classifiers are used to identify the effect of the application of Timbral Audio descriptor modeling on the Bird species identification system. The results show that for the singing bird category, the Timbral audio descriptors roll-off, brightness, roughness, and MFCC give the highest bird species identification accuracy of 94.61% and 93.51% for the chirping bird category using ZCR, irregularity, and MFCC audio descriptors. These audio descriptor models are not only useful for the classification of interclass bird species but also helpful to correctly identify intraclass bird species. The average Bird species Identification accuracy using Timbral Audio Descriptor modeling is obtained as 94%. In the future, a greater number of spectral and temporal Audio descriptors could be combined with Timbral audio descriptors to make the system more generalized for any type of bird species identification.

6. REFERENCES

[1] Bastas, Selin A., 2011. Nocturnal Bird Call Recognition System for Wind Farm Applications. *Toledo: The University of Toledo.*

[2] CK-12. 2021. "Linnaean Classification." In Introductory Biology, *LibreTexts.* p. 6535. https://bio.libretexts.org/@go/page/6535.

[3] Deshmukh, Saurabh Harish, 2012. "A Hybrid Selection Method of Audio Descriptors for Singer Identification in North Indian Classical Music." *IEEE Explorer. Himeji, Japan.* pp. 224-227.

[4] Deshmukh, Saurabh and Sunil Bhirud, 2014. "A Novel Method to Identify Audio Descriptors, Useful in Gender Identification from North Indian Classical Music Vocal." *International Journal of Computer Science and Information Technologies,* **5**(2), 1139-1143.

[5] Deshmukh, Saurabh and Sunil Bhirud, 2014. "North Indian classical music's singer identification by timbre recognition using MIR toolbox." *International Journal of Computer Applications,* **91**(4), 1-5.

[6] Fagerlund, Seppo, 2004. Automatic Recognition of Bird Species by Their Sounds. *Espoo: Helsinki University of Technology.*

[7] Forrest, Briggs, Yonghong Huang, Raviv Raich, Konstantinos Eftaxias, Zhong Lei, William Cukierski, Sarah Frey Hadley, *et al.*, 2013. "The 9th Annual MLSP Competition: New Methods for Acoustic Classification of Multiple Simultaneous Bird Species in a Noisy Environment." *IEEE International Workshop on Machine Learning for Signal Processing. Southampton, UK.*

[8] Goeau, Harve, Harve Glotin, Willem-Pier Vellinga, Robert Planque and Alexis Joly, 2016. "LifeCLEF Bird Identification Task 2016: The arrival of Deep Learning." *CLEF: Conference and Labs of the Evaluation Forum. Evora, Portugal.* https://hal.archives-ouvertes.fr/hal-01373779.

[9] Lartillot, Olivier. n.d. MIR ToolBox Manual. The University of Jyvaskyla, Finnish Centre of Exce! ence in Interdisciplinary Music Research.

[10] Lasseck, Mario, 2013. "Bird song classification in field recordings: Winning solution." *Proceedings of international symposium Neural Information Scaled for Bioacoustics. Nevada.*

[11] Martinsson, John, 2017. Bird Species Identification using Convolutional Neural Networks. *Gothenburg, Sweden: University of Gothenburg.*

[12] Note, Bird, 2021. How Birds Produce Sound. Bird Note. April 1. Accessed Jan 29, 2022. https://www.birdnote.org/listen/shows/how-birds-produce-sound.

[13] Park, Tae Hong, 2004. Towards Automatic Musical Instrument Timbre Recognition. Ph.D. Thesis, *Candidacy: Princeton University.*

[14] Serizel, R. and D. Giuliani, 2017. "Deep Neural Network Approaches for Speech Recognition with Heterogeneous Groups of Speakers Including Children." *Natural Language Engineering,* **23**(3), 325-350. doi:10.1017/S135132491600005X.

[15] Shete, Shambhavi and Saurabh Deshmukh, 2019. "Analysis and Comparison of Timbral Audio Descriptors with Traditional Audio Descriptors Used in Automatic Tabla Bol Identification of North Indian Classical Music." *Proceeding of International Conference on Computational Science and Applications, Algorithms for Intelligent Systems. Pune: Springer.* pp. 295-307. https://doi.org/10.1007/978-981-15-0790-8_29.

[16] Shete, Shambhavi and Saurabh Deshmukh, 2021. "Automatic Tabla Stroke Source Separation Using Machine Learning." Advances in Computing and Data Sciences, ICACDS 2021. *Communications in Computer and Information Science. Cham: Springer.* DOI:https://doi.org/10.1007/978-3-030-81462-5_22.

[17] Thiruvengatanadhan R., 2017. "Speech/Music Classification using MFCC and KNN." *International Journal of Computational Intelligence Research,* **13**(10), 2449-2452.

[18] Zakeri Masoud, 2017. Automatic Bird Species Identification Employing an Unsupervised Discovery of Vocalisation Units. *Birmingham: The University of Birmingham.*

INFORMATION FOR AUTHORS

ARTICLES

The Journal of Acoustical Society of India (JASI) is a refereed publication published quarterly by the Acoustical Society of India (ASI). JASI includes refereed articles, technical notes, letters-to-the-editor, book review and announcements of general interest to readers.

Articles may be theoretical or experimental in nature. But those which combine theoretical and experimental approaches to solve acoustics problems are particularly welcome. Technical notes, letters-to-the-editor and announcements may also be submitted. Articles must not have been published previously in other engineering or scientific journals. Articles in the following are particularly encouraged: applied acoustics, acoustical materials, active noise & vibration control, bioacoustics, communication acoustics including speech, computational acoustics, electro-acoustics and audio engineering, environmental acoustics, musical acoustics, non-linear acoustics, noise, physical acoustics, physiological and psychological acoustics, quieter technologies, room and building acoustics, structural acoustics and vibration, ultrasonics, underwater acoustics.

Authors whose articles are accepted for publication must transfer copyright of their articles to the ASI. This transfer involves publication only and does not in any way alter the author's traditional right regarding his/her articles.

PREPARATION OF MANUSCRIPTS

All manuscripts are refereed by at least two referees and are reviewed by the Publication Committee (all editors) before acceptance. Manuscripts of articles and technical notes should be submitted for review electronically to the Chief Editor by e-mail or by express mail on a disc. JASI maintains a high standard in the reviewing process and only accept papers of high quality. On acceptance, revised articles of all authors should be submitted to the Chief Editor by e-mail or by express mail.

Text of the manuscript should be double-spaced on A4 size paper, subdivided by main headings-typed in upper and lower case flush centre, with one line of space above and below and sub-headings within a section-typed in upper and lower case understood, flush left, followed by a period. Sub-sub headings should be italic. Articles should be written so that readers in different fields of acoustics can understand them easily. Manuscripts are only published if not normally exceeding twenty double-spaced text pages. If figures and illustrations are included then normally they should be restricted to no more than twelve-fifteen.

The first page of manuscripts should include on separate lines, the title of article, the names, of authors, affiliations and mailing addresses of authors in upper and lowers case. Do not include the author's title, position or degrees. Give an adequate post office address including pin or other postal code and the name of the city. An abstract of not more than 200 words should be included with each article. References should be numbered consecutively throughout the article with the number appearing as a superscript at the end of the sentence unless such placement causes ambiguity. The references should be grouped together, double spaced at the end of the article on a separate page. Footnotes are discouraged. Abbreviations and special terms must be defined if used.

EQUATIONS

Mathematical expressions should be typewritten as completely as possible. Equation should be numbered consecutively throughout the body of the article at the right hand margin in parentheses. Use letters and numbers for any equations in an appendix: Appendix A: (A1, (A2), etc. Equation numbers in the running text should be enclosed in parentheses, i.e., Eq. (1), Eqs. (1a) and (2a). Figures should be referred to as Fig. 1, Fig. 2, etc. Reference to table is in full: Table 1, Table 2, etc. Metric units should be used: the preferred from of metric unit is the System International (SI).

REFERENCES

The order and style of information differs slightly between periodical and book references and between published and unpublished references, depending on the available publication entries. A few examples are shown below.

Periodicals:

[1] S.R. Pride and M.W. Haartsen, 1996. Electroseismic wave properties, *J. Acoust. Soc. Am.*, **100** (3), 1301-1315.

[2] S.-H. Kim and I. Lee, 1996. Aeroelastic analysis of a flexible airfoil with free play non-linearity, *J. Sound Vib.*, **193** (4), 823-846.

Books:

[1] E.S. Skudzryk, 1968. *Simple and Comlex Vibratory Systems*, the Pennsylvania State University Press, London.

[2] E.H. Dowell, 1975. *Aeroelasticity of plates and shells*, Nordhoff, Leyden.

Others:

[1] J.N. Yang and A. Akbarpour, 1987. Technical Report NCEER-87-0007, Instantaneous Optimal Control Law For Tall Buildings Under Seismic Excitations.

SUMISSIONS

All materials from authors should be submitted in electronic form to the JASI Chief Editor: B. Chakraborty, CSIR - National Institute of Oceanography, Dona Paula, Goa-403 004, Tel: +91.832.2450.318, Fax: +91.832.2450.602,(e-mail: bishwajit@nio.org) For the item to be published in a given issue of a journal, the manuscript must reach the Chief Editor at least twelve week before the publication date.

SUMISSION OF ACCEPTED MANUSCRIPT

On acceptance, revised articles should be submitted in electronic form to the JASI Chief Editor (bishwajit@nio.org)