ISSN 0973-3302

THE JOURNAL OF ACOUSTICAL SOCIETY OF INDIA

Volume 50 Number 2 April 2023

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

The Journal of Acoustical Society of India
Volume 50 Number 2 April 2023 pp 76-119

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.

ASI

The Journal of Acoustical Society of India

A quarterly publication of the Acoustical Society of India

Volume 50, Number 2, April 2023

ARTICLES

INFORMATION

FOREWORD

The works presented in this second Volume of the three Volume Series of the Special Issue of the Journal of Acoustical Society of India (JASI) look to present new technological advancements spanning the intertwined domains of speech, music, and languages, 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 2nd Volume is essentially a follow up of the 1st Volume, which presents several traditional and new researches in the domain of speech and music, while this volume lays emphasis on language identification, processing alongside presenting novel research on speech and music domains. 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 physical data in research in speech and music being of natural origin, a significant amount of mathematical modelling is non-deterministic in nature. Except for the primary mandatory signal processing aspects to extract parameters, the rest of the analysis has to rely largely on stochastic as well as non-stochastic random mathematical approaches like chaos and fractals. In fact, it often happens that new paradigms of mathematical approaches have to be resorted to, which often brings clarity to the non-deterministic mathematical approaches. The same is true for music. In fact, in a multi-categorical musical country like India, where one has classical music at one end and numerous ethnic and folk music at the other end, it is only the objective analysis of music which can bring out the cross-cultural intercourses taking place. Unfortunately, little attention has been paid to this subject by mainstream national knowledge institutions. The fact that in India, objective research in these two fields is never considered seriously by the academicians was also an important additional factor.

Music is said to be the language of emotions, and the activity of listening to music is indeed a part of everyday life. Language and music share certain important features. For e.g., both are finely structured systems of expression and communication based on perceptually discrete units organized into flowing acoustic sequences; both are universal in the human species and have ancient origins. Rhythm is a fundamental part of life. In humans, our ability to produce and perceive rhythmic behaviour, e.g., in music and language, might have evolved from primates or might be uniquely human. Nonverbal and paralinguistic cues provide a rich source of information about a speaker's emotions and social intentions when engaged in discourse. Language and music are also closely related in our minds. Does musical expertise enhance the recognition of emotions in speech prosody? Few works presented in this Edited Volume try to search for the answers to these questions in an eclectic manner.

To name a few aspects from the multifaceted nature of the study, viz, the purity of pitch, the purity of vowel articulation, the beauty of voice and that of rhythm, opens up a vast area of study. Music Information Retrieval (MIR) has been one of the most researched fields in the last decade, which deals with different aspects of searching and organising large collections of music, or music information, according to their relevance to specific queries, i.e., involving different classification and feature extraction algorithms. In this collection, we have a few chapters which lie on the periphery of MIR but cannot be said to be directly part of MIR. This has been done consciously as there already exists a number of books in the field of MIR and we wanted to keep the focus of this volume primarily on the classical studies of speech, music and their associated application-oriented fields.

Language identification is a human skill that has many practical applications. The necessity of automating language identification using a computer system is increasing as communicational affairs are growing daily. Musicological research has long existed since ancient times. However,

scientific investigations in Indian music have fallen far behind those in western music. The present state of science and technology can provide ample scope for quantitative and qualitative approaches to investigate shrutis, swaras, intervals, octaves (saptak), consonance (vaditya), musical quality, rhythm etc. In this Volume, we have few chapters dealing with Speaker recognition as well as spoken language identification.

The second volume contains six (6) chapters spanning over the following broad areas:

In the area of speech :

a. Speaker sentiment analysis involved in day-to-day conversations of customer support jobs.
b. Comparisons of the measures of speech intelligibility and naturalness before and after supportive speech therapy in siblings.

In the area of music :

a. Development of methods to select the correct audio descriptors, which enhances the application of the Hybrid Selection Algorithm for identifying appropriate audio descriptors for the Tabla strokes
b. Study of improvisations made by two eminent singers in two different gharanas in Indian classical music.

In the area of languages and linguistics :

a. Automatic spoken language identification for sixteen Indian languages using Relative Abundance Model (RAM) obtaining a recognition rate of 70%.
b. Speaker recognition in Bengali language from nonlinear features using SVM classifier with an accuracy of 96%

We have put special emphasis on collecting works related to language, speech, and musical paradigms of the Indian subcontinent. The Editors sincerely hope that serious researchers, academicians, interested personalities of this fast growing discipline would benefit from this Volume of JASI.

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

The Journal of Acoustical Society of India : Vol. 50, No. 2, 2023 (pp. 76-82)

Accuracy Test on 2D CNN on entire MFCC and MEL-spectrogram with and without data augmentation

Saswatika Joshi* and Subarna Chatterjee
Department of Computer Science & Engineering
MS Ramaiah University of Applied Science, Bengaluru, India
e-mail: subarna2000@gmail.com

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

ABSTRACT

Voice is an important part of communication; communication is the key to any business. Today most of business cannot run without interaction and organizations are struggling today to drive customer satisfaction, delight through their customer support and contact centres, this is due to lack of any kind of measurement, today with the help of AI and machine learning algorithms, it is possible to measure etiquettes through voice modulations and pitch. The goal of this proposed study is to find out how the various speakers in the conversation felt. We investigated several strategies for speaker discrimination and sentiment analysis, as well as accuracy tests on 1D and 2D CNN using the whole MFCC and Mel-Spectrogram with and without augmentation approaches, in order to find an efficient Model. We used (TESS, SAVEE, RAVDESS, CREAMA-D) datasets to train the model to classify gender and emotions, and we concentrated on balanced audio datasets including male and female speakers with diverse sentiments. We observed superior model accuracy on 2D CNN on the complete MFCC without augmentation (67.38 percent at 50 epoch) and used the f1-score as an evaluation metric, which yielded a weighted average of 69 on the test set and the best results on the "female pleasant surprise" and "angry" class with a score of 1.00 & 0.78.

1. INTRODUCTION

Sentiment Analysis exploits Natural Language Processing (NLP) methods to decipher and interpret human emotions[2]. This analysis is frequently employed in consumer fields to analyze client sentiment towards a product to appraise it and modify it.

The underlying technology that performs the basic function is Convolutional Neural Networks (CNN). CNNs employ convolution kernels and pooling methods to pull stationary features from datasets[4]. As technology advances and machines are created to perform more sophisticated and complex functions, NLP processes have advanced from text mining to Automatic Speech Recognition (ASR) that decode and discern audios through MFCCs. Mel Frequency Cepstral Coefficients (MFCCs) decode speech patterns and display the patterns in spectral bands called a mel-spectrogram. For instance, an MFCC evaluating echoes against a time signal generates a Frequency Spectrum with sharp distinctive peaks. Such patterns including pitch and tone are decoded to decipher the underlying sentiment[4,2]. To perform this function, MFCC mimics the distinctive human characteristics of sound generation[5]. The shape of an individual's

vocal tract and its arrangement which include the tongue, teeth determine the level of sound that is generated. Typical MFCC features include phenomes that are built to resemble the human vocal tract. This paper focuses on the application of MFCC and log-mel-spectrogram technology in sentiment analysis, specifically on its efficacy to decode the sentiments from speech of people with cognitive disabilities. We run the CREMA-D, RAVDESS, TESS and SAVEE datasets to test the accuracy of 1D and 2D CNN'S in deciphering the sentiments of augmented and non-augmented data.

The experiment was first run by building 1D and 2D CNN Models. Then, using the 1D CNN model, the datasets were run on MFCC and mel spectrogram without augmentation. The same datasets were run on MFCC and mel spectrogram with augmentation. This process was subsequently run for the 2D CNN model. In the CNN Modelling stage, the datasets were separated into 25% test data and 75% training data. The test data acted as a control representing accurate real-world data. The training data was used in the model to achieve correct interpretation between the features and the target.

The rest of the paper is organized as follows. In section 2 a brief review of relevant literature is provided. Section 3 and 4 explain the contributions and methodology respectively. Section 5 describes the experimental analysis. Finally, the conclusion is presented in Section 6.

2. RELATED WORK

Ferdiana *et al.*[6] show that increasing datasets and CNN layers during the training process develops the signal interpretation's accuracy. In the experimental, the dataset's range was increased to 595 and diversified into 5 classifications of cat sounds. The authors built a custom 2D CNN model, and used a Conv2D algorithm to add on a 2×2 Windows layer. Each convolution layer had an incremental number of nodes. The first layer had 16 nodes, the final layer comprised 128 nodes. This strategy recorded an accuracy of approximately 88.473254%.

The high accuracy evidences that training a CNN model with different nodes increases accuracy, and the data interpreted by MFCCs is improved over subsequent experiments. Further, testing this on RAVDESS, De Pinto *et al.*[1] ran a similar experiment on 1D CNNs, max-pooling operations and Dense Layers. The data was augmented in an FFMPEG library to the frequency of 44,1MH. The results reveal "an overall F_1 score of 0.91 with the best performances on the angry class (0.95) and worst on the sad class (0.87)."

This experiment shows good results of augmented data in MFCC, specifically in 1D CNNs. However, the authors do not diversify their experiments to include 2D CNNs or compare their results with non-augmented datasets or delineate the classification differences in deciphering gender datasets. Despite this, the authors perform an excellent experiment to reveal the capacity of sentiment analysis on the RAVDESS dataset. Further, the use of CNN has expanded in recent times, Jung *et al.*[7] employ CNN in medical diagnostic procedures to classify lung auscultation in COVID-19 cases. The dataset consisted of WAV audio sets which were 18 seconds long in 4 kHz frequencies. This study compared different spectrogram features against different CNN models. The outputs were segregated into 3 different categories, short-time Fourier transformed (STFT), MFCCS and a combination of both STFT and MFCC. The convoluting networks were DS-CNN and a standard CNN with conventional layers. Further, the DS-CNN was shrunk by altering the width and depth of the layers. Overall, the STFT documented accuracy of 82.27%, the MFCC was 73.02% and the fused MFCC / STFT recorded the highest accuracy score of 85.74%. This analysis[7] reveals the refinement in accuracy of fusing spectrogram features to achieve higher levels of precision. This is a notable discovery. However, the study only applied these features to lung sounds. Can the same results be achievable on complicated nuances in the cadences of human speech. Shen *et al.*, cite[8] apply fused spectrogram features to classify datasets of speech with similarly successful results. Here, they fused CNN and Bidirectional Long-Short-Term. (BLSTM) to analyze movie reviews datasets. The fused model had an accuracy rate of 89.7%. CNN only and BLTSM only classification recorded lower precision scores of 83.9% and 78.4%, respectively. Thus, it can be understood that fusing classifiers increase the exactness of detection. Dashtipour *et al.*[8] classify datasets of the Persian language. This study is unique as it assesses

non-English datasets. Thus, it is a test on CNN's training ability to decipher more complicated text with almost the same accuracy as English datasets. They discuss a study that utilized a Recurrent Neural Network (RNN) classifier to classify hateful content on Twitter in the English language. Here, RNN was superior to support vector machine (SVM) and recorded an accuracy of 95.33% compared to SVM's 75.22%. To compare these machine learning English results with Persian texts, the researchers conducted experiments to classify Persian datasets (movie reviews and hotel reviews) For shallow learning, Dashtipour *et al.*[8] employed the following classifiers logistic regression, support vector machine (SVM), and multi-layer perceptron (MLP). For deep learning, the authors employed 1D convolutional neural network (CNN), 2D-CNN, stacked long short-term memory (LSTM), and Bidirectional LSTM algorithms. The results disclosed that deep learning techniques outperformed shallow learning methods[8]. Further, stacked and augmented Bidirectional LSTM algorithms recorded the most outstanding accuracy scores of 95.61% for the movie reviews. However, 2D-CNN detected 89.76% of the hotel reviews. Lu *et al.*[3] studied a different approach to increase downstream sentiment accuracy. Instead of increasing convoluting layers to machine training like[5]. The authors[3] employed end-to-end (e2e) automatic speech recognition (ASR) as a pre-training strategy. The e2e model comprised an augmented RNN Transducer (RNN-T), which analyzed IEMOCAP and SWBD datasets. This performance strategy increased the accuracy of deciphering the IEMOCAP dataset from 66.6% to 71.7%[3]. The positive results of machine learning in deciphering English, Persian and its application in lung auscultation and cat sounds demonstrates that deep training can be applied to a wide variety of speech and sound, including datasets from people with cognitive disabilities. Further, fused, augmented and multi-layered CNN's consistently report better detection results.

3. CONTRIBUTIONS

We combined four audio emotions datasets to create a large dataset for training with a 1D/2D CNN model to recognize individual emotion and gender in this paper. Instead of raw audio data, we provide spectrograms to a Deep Learning model for audio emotion classification. The energy of an audio stream is described by spectrograms, which are time-frequency representations. They can be compared to a grid. The y axis shows frequency bins, while the x axis shows time. Each grid point is assigned a value that indicates the energy for a specific frequency band at a specific moment. Spectrograms can be represented graphically. For spectrograms, we used a similar method in which a pixel is substituted by a point in the frequency-time grid. We also used data augmentation techniques in conjunction with a CNN-based design. Putting labels on the data to create a balanced dataset of different gender, combine datasets.

The following are some of the contributions made by this study:

- Labelling the data from multiple datasets to create gender-balanced datasets.
- To expand the data volume, use data augmentation techniques.
- Using spectrogram to test CNN-based models for improved model accuracy.

4. METHODOLOGY

The experiment was first run by building 1D and 2D CNN Models. Then, using the 1D CNN model, the datasets were run on MFCC and mel spectrogram without augmentation. The same datasets were run on MFCC and mel spectrogram with augmentation This process was subsequently run for the 2D CNN model. In the CNN Modeling stage, the datasets were separated by a train_test_split function into 25% test data and 75% training data. The test data acted as a control representing accurate real-world

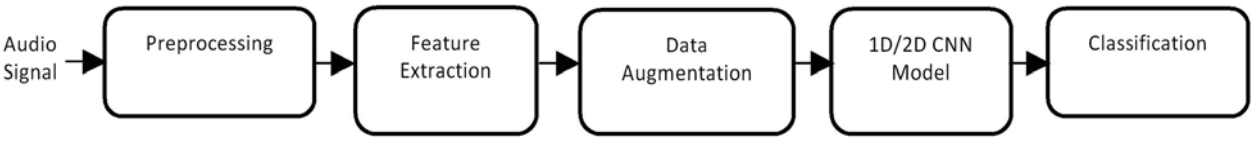

Fig. 1. Functional block of the proposed algorithm.

data. The training data was used in the model to achieve correct interpretation between the features and the target. The subsequent steps included Model serialization, model validation, prediction.

4.1 Dataset

a. ***Toronto emotional speech set (TESS) :*** It comprises 7 Ekman emotions, and is based on 2 speakers, a young and an older female. Thus, it is female dominated and balances out the male dominant speakers in SAVEE.
b. ***Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) :*** it has 24 speakers, categorized from 1-24. Here, male voices are odd numbered, while female voices are even numbered. Emotion Type: (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05= angry, 06 = fearful, 07 = disgust, 08 = surprised)
c. ***Surrey Audio-Visual Expressed Emotion (SAVEE) :*** It contains 4 folders each representing a male speaker. The female only TESS Dataset balances this imbalance.
d. ***Audio Classification :*** Prefix letters represent the emotion classes: a' = 'anger','d' = 'disgust','f' = 'fear','h' = 'happiness',n' = 'neutral','sa' = 'sadness','su' = 'surprise'
e. ***Crowd-sourced-Emotional-Multimodal-Actors Dataset (CREMA-D) :*** it has a good variety of emotions, with more intensity than RAVNESS. However, the "surprise" emotion was missing.

4.2 MFCC Algorithm

The first step of the MFCC converted the analogue speech to digital speech (A / D Conversion).

Pre-emphasis follows A / D Conversion. Here, higher frequency audios were boosted to frequencies of higher magnitudes. This frequency elevation increased the accuracy of speech detection[2]. Vowel sounds are primarily passed through this step. A vowel sound such as 'aa' has a low energy magnitude at a higher frequency. Thus, for accurate phonetic interpretation, its magnitude was increased by a first-order high-pass filter.

The third step was Windowing. The MFCC extracted elements in the dialogue then broke them down and characterized them into phones. As this segmentation[2] proceeded, it inevitably produced breaks in signals. This yielded signals that produced sound at high frequencies due to immediate collapses in amplitudes. To remedy and smooth this out, a Hamming / Hanning Window minced the signal breaks to prevent the production of high-frequency sounds.

In Fourth step, a Discrete Fourier Transform (DTF) converted the time spectrum signal into the frequency region signal which is easier to interpret. After the Filter Banks constructs Mel-Filters through conversion of the signal using the Mel frequency scale.

Data Augmentation method This was done by generating syntactic data for the audio set. by adding random values into the data, applying noise injection, shifting time, changing pitch and speed. The aspects that were altered included : (1) Static noise, (2) Shift, (3) Stretch, (4) Pitch, (5) Dynamic change and (6) Speed and pitch.

5. RESULTS

From the experiment: 2D CNN / MFCC without augmentation (accuracy: 67.38%) & with augmentation (accuracy: 62.8) 2D CNN / mel-spectrogram without augmentation (accuracy: 65.77%) & with augmentation (accuracy: 63.04%).1D CNN without augmentation (accuracy 18.09%) & with augmentation (accuracy: 47.29%)

From the 2D CNN model, MFCC with Augmentation had the least accuracy in classifying emotions (0.64). Augmentation features have lower accuracy levels compared to features without augmentation in classifying emotions. This finding aligns with other results from similar research findings[1,8] as augmentation resolves overfitting to increase accuracy. Thus, the model can interpret and differentiate the nuances in pitch, and cadences to distinguish emotions. MFCC without augmentation was the preferable output- it recorded the highest level of accuracy in classifying gender (0.98).

Table 1. f1-score: classification of emotions

Emotion	2CNN MFCC w-o Augmentation	2DCNN MFCC + Augmentation	2DCNN+ Mel Spectrogram w-o Augmentation	2DCNN + Mel Spectrogram + Augmentation	1DCNN w-o Augmentation	1DCNN + Augmentation
Angry	0.78	0.74	0.77	0.76	0.34	0.76
Disgust	0.66	0.57	0.65	0.64	0.22	0.64
Fear	0.59	0.57	0.59	0.58	0.26	0.58
Female	1.00	0.99	1.00	0.99	0.44	0.99
Happy	0.62	0.60	0.60	0.57	0.19	0.57
Neutral	0.71	0.66	0.68	0.67	0.19	0.67
Sad	0.67	0.63	0.63	0.62	0.22	0.62
Surprise	0.59	0.61	0.75	0.65	0.22	0.65
Accuracy	0.68	0.64	0.67	0.65	0.25	0.65
Macro avg	0.70	0.67	0.71	0.68	0.26	0.68
Weighted avg	0.68	0.64	0.67	0.65	0.25	0.65

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.

From the overall classification of emotions. 1D CNN without augmentation was unsuccessful; 2D MFCC without augmentation recorded the highest F_1 accuracy score of 0.68; 1D CNN without augmentation had the least F_1 accuracy score of 0.25. Both 2D CNN / mel-spectrogram had higher accuracy scores than 2D CNN / MFCC with augmentation. Mel-spectrogram with augmentation and mel-spectrogram without augmentation read 0.65 and 0.67 respectively; the least deciphered emotion across all the models with an average F_1 accuracy score of 0.52.

Table 2. f1-score for gender classification

Gender	2CNN MFCC w-o Augmentation	2DCNN MFCC + Augmentation	2DCNN+ Mel Spectrogram w-o Augmentation	2DCNN + Mel Spectrogram + Augmentation	1DCNN w-o Augmentation	1DCNN + Augmentation
Female	0.98	0.98	0.97	0.96	0.61	0.96
Male	0.98	0.97	0.96	0.95	0.65	0.95
Accuracy	0.98	0.97	0.97	0.96	0.63	0.96
Macro avg	0.98	0.98	0.97	0.97	0.57	0.97
Weighted avg	0.98	0.97	0.97	0.96	0.62	0.96

2D CNN was successful in distinguishing gender and had excellent F_1 scores. Mel-spectrogram had the same F_1 scores for augmented and non-augmented datasets. Here, the F_1 accuracy score was 0.97 which was similar to 2D MFCC with augmentation. 2D CNN / MFCC without augmentation had the highest accuracy score of 0.98. Thus, this is the best option for distinguishing gender.

5.1 Limitations

The focus of this research lies in analyzing datasets of people with cognitive disabilities. Thus, the scale of the project is limited to these types of speech and gathering numerous sounds that comprise of different categories by age and gender are difficult. Further, as a result of some technical challenges, the

project unsuccessfully strived to extract user sentiment through VoiceBot. However, it aims to mend this and successfully run this experiment.

5.2 Future Scope

To test the variation of results, the future goal of this project will be to run the exact experiment on other datasets with additional image recognition techniques such as VGG19 and ResNet50 on MFCC and Mel-spectrogram, and compare the results to detect the most applicable image recognition software. Further, to improve the accuracy of the project's model, we will consider to employ a training feature with fused filters and different number of nodes in the CNN layers. As recorded from other findings, we expect to increase the accuracy in classifying emotions up to 80%.

6. CONCLUSION

From this experiment, we run RAVDESS, TESS, SAVEE and CREMA augmented and non-augmented datasets on 1D, 2D CNN / MFCC and mel-spectrogram models. I tested the accuracy of each CNN model on MFCC and mel-spectrogram.

The 2D / MFCC without augmentation model was effective in distinguishing gender. 1D CNN with augmentation model done pretty well at distinguishing gender. This experiment was successfully and shows that 2D / MFCC and 2D mel-spectrogram models can decipher the sentiments of people with cognitive disabilities and shower superiority over 1D CNNs.

7. REFERENCES

[1] M.G. de Pinto, M. Polignano, P. Lops and G. Semeraro, 2000. "Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients." *In 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS),* pp. 1-5, doi: 10.1109 / EAIS48028.2020.9122698.

[2] H. Koo, S. Jeong, S. Yoon and W. Kim, 2020. "Development of speech emotion recognition algorithm using MFCC and prosody" *In 2020 International Conference on Electronics, Information and Communication (ICEIC),* pp. 1-4, doi: 10.1109 / ICEIC49074.2020.9051281.

[3] Z. Lu, L. Cao, Y. Zhang, C. -C. Chiu and J. Fan, 2020. "Speech sentiment analysis via pre-trained features from end-to-end asr models." *In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 7149-7153, doi: 10.1109 / ICASSP40776.2020.9052937.

[4] Z. Tüfekci, and G. Disken. 2019. "Scale-invariant MFCCs for speech / speaker recognition." *Turkish Journal of Electrical Engineering & Computer Sciences,* **27**(5), 3758-3762. doi: 10.3906 / elk-1901-231.

[5] A. Mertins and J. Rademacher, 2005. "Vocal tract length invariant features for automatic speech recognition." *In 2005 IEEE Workshop on Automatic Speech Recognition and Understanding,* pp. 308-312, doi: 10.1109 / ASRU.2005.1566473.

[6] R. Ferdiana, W. F. Dicka and A. Boediman. 2021. "Cat Sounds Classification with Convolutional Neural Network." *International Journal on Electrical Engineering and Informatics,* **13**(3), 755-765. doi: 10.15676 / ijeei.2021.13.3.15.

[7] S. Y. Jung, C. H. Liao, Y. S. Wu, S. M. Yuan and C. T. Sun, 2021. "Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features." *Diagnostics (Basel),* **11**(4), 732-744. doi: 10.3390 / diagnostics11040732.

[8] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani and A. Hussain. 2021. "Sentiment Analysis of Persian Movie Reviews Using Deep Learning." *Entropy,* **23**(5), 596, https: / / doi.org / 10.3390 / e23050596.

[9] N. Dey, ed., 2019. Intelligent speech signal processing: Academic Press.

[10] M. Telmem and Y. Ghanou, 2021. "The convolutional neural networks for Amazigh speech recognition system." *Telkomnika.* **19**(2), 515-522, http: / / doi.org / 10.12928 / telkomnika.v19i2.16793.

[11] K. D. Anadkat and H. M. Diwanji, 2021. "Effect Of Activation Function In Speech Emotion Recognition On The Ravdess Dataset." *Reliability: Theory & Applications,* **16**(3), 228-236, https://doi.org/10.24412/1932-2321-2021-363-228-236.

[12] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu and Y. B. Zikria, 2020. "Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network." *Sensors,* **20**(21), 6008, https://doi.org/10.3390/s20216008.

[13] P. Nantasri, E. Phaisangittisagul, J. Karnjana, S. Boonkla, S. Keerativittayanun and A. Rugchatjaroen, 2020. "A Light Weight Artificial Neural Network for Speech Emotion Recognition using Average Values of MFCCs and Their Derivatives." *In 2020 17th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON),* pp. 41-44, doi: 10.1109/ECTI-CON49241.2020.9158221.

[14] S. N. Zisad, M. S. Hossain and K. Andersson, 2020. Speech Emotion Recognition in Neurological Disorders Using Convolutional Neural Network. In: Brain Informatics. Lecture Notes in Computer Science, Springer, **12241,** 287-296, Cham. https://doi.org/10.1007/978-3-030-59277-6_26

[15] B.T. Atmaja and M. Akagi, 2020. "On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers." *In 2020 IEEE REGION 10 CONFERENCE (TENCON),* pp. 968-972, doi: 10.1109/TENCON50793.2020.9293852.

[16] T. J. Sefara, 2019. "The effects of normalisation methods on speech emotion recognition." *In International Multidisciplinary Information Technology and Engineering Conference (IMITEC),* pp. 1-8, doi: 10.1109/IMITEC45504.2019.9015895.

[17] D. Issa, M. F. Demirci and A. Yazici. 2020. "Speech emotion recognition with deep convolutional neural networks." *Biomedical Signal Processing and Control,* **59,** 101894, https://doi.org/10.1016/j.bspc.2020.101894.

[18] M. Xu, F. Zhang and W. Zhang, 2021. "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset." *In IEEE Access,* **9,** 74539-74549. doi: 10.1109/ACCESS.2021.3067460.

[19] M. Sajjad and S. Kwon, 2020. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM." *In IEEE Access,* **8,** 79861-79875. doi: 10.1109/ACCESS.2020.2990405.

[20] U. Kumaran, S. R. Rammohan, S. M. Nagarajan and A. Prathik. 2021. "Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN" *International Journal of Speech Technology,* **24**(2), 303-314. https://doi.org/10.1007/s10772-020-09792-x

[21] Z. Rzayeva and E. Alasgarov, 2019. "Facial emotion recognition using convolutional neural networks(FERC)." *SN Applied Sciences,* **2**(3), 1-8. https://doi.org/10.1007/s42452-020-2234-1.

[22] K. Venkataramanan and H. R. Rajamohan, 2019. "Emotion recognition from speech." arXiv preprint arXiv:1912.10458.

[23] F. A. Shaqra, R. Duwairi and M. Al-Ayyoub, 2019. "Recognizing emotion from speech based on age and gender using hierarchical models." *Procedia Computer Science,* **151,** 37-44. https://doi.org/10.1016/j.procs.2019.04.009

[24] A. Amjad, L. Khan and H. T. Chang, 2021 "Effect on speech emotion classification of a feature selection approach using a convolutional neural network." *Computer Science,* **7,** e766. https://doi.org/10.7717/peerj-cs.766.

[25] R. Rajak and R. Mall, 2019. "Emotion recognition from audio, dimensional and discrete categorization using CNNs" *In IEEE Region 10 Conference (TENCON),* pp. 301-305, doi: 10.1109/TENCON.2019.8929459.

The Journal of Acoustical Society of India : Vol. 50, No. 2, 2023 (pp. 83-90)

Impact of speech therapy on speech intelligibility and speech naturalness in siblings with ataxia telangiectasia like disorder

Subhiksha M. and Swapna N.

All India Institute of Speech and Hearing, Manasagangothri, Mysore-06, Karnataka, India
e-mail: subhiksha812@gmail.com

[Received: 10-03-2022; Revised: 15-06-2022; Accepted: 01-07-2022]

ABSTRACT

Certain conditions hinder participation in society by affecting the overall speech intelligibility and naturalness. This study aims to compare the measures of speech intelligibility and naturalness before and after supportive speech therapy in two female siblings, aged 30.2 years and 26 years, diagnosed with moderate developmental dysarthria secondary to Ataxia Telangiectasia Like Disorder (ATLD). Ataxia Telangiectasia, a variant of a spinocerebellar ataxia variety, is a rare autosomal recessive, progressive neurological impairment. The clinical manifestations include ataxia, dysarthria, chromosomal aberrations, telangiectasia, elevated alpha-fetoprotein levels, and other cerebellar signs. The siblings exhibited slurring, reduced speech rate, motor difficulties, and difficulty eating and swallowing. Speech therapy provided through the online mode focussed on improving the respiratory, phonatory, resonatory, articulatory subsystems, and prosody, using the "Programmed Subsystem" approach. The speech intelligibility and speech naturalness were assessed at the word, sentence, and conversation levels using a picture description task and sentence reading task for the siblings before the initiation and after the termination of online speech therapy. As a part of the speech naturalness assessment, the stress, intonation, pauses, rhythm, rate of speech, and articulatory proficiency were assessed on a 2-point rating scale in which higher scores indicated poorer naturalness and vice versa. The Quality of life (QoL) was also evaluated. The percent of speech intelligibility before initiation of speech therapy was 66.6%, which increased to 83.3% in the elder sibling and 70% in the younger sibling, which increased to 87.5%. Further, in the elder sibling, the total speech naturalness score was six(6) before therapy, which decreased to four and a half (4.5), and in the younger sibling, the scores decreased from 4 to 3. The QoL scores also decreased, which indicated better quality of life after therapy. It can be concluded from the study that speech intelligibility and naturalness improve with speech therapy in patients with ATLD. Future studies could assess speech intelligibility and naturalness in a larger sample, using spontaneous speech tasks that are more representative of speech used in daily contexts.

1. INTRODUCTION

The most crucial aspect of communication is to express oneself to the communication partner, ensuring no communication breakdown. The ability of the speaker to produce utterances that make the speech,

idea, and thought understandable to the listeners is referred to as Speech Intelligibility. On the parallel side, speech naturalness measures help us understand how speech matches the typical intonation patterns, rhythm, pitch, and stress, so that speech production sounds natural to the listener. Many neurological conditions can impair brain functioning leading to poor speech intelligibility and naturalness.

One such condition is Ataxia Telangiectasia (AT). AT is an autosomal recessive, complex, rare, progressive, neurodegenerative, and multisystem disease characterized by cerebellar ataxia, oculocutaneous telangiectasia, immunodeficiency, progressive pulmonary infections, elevated alpha-fetoprotein level, susceptibility to malignancy and increased sensitivity to ionizing radiations (Chen, Mucha and Oakes 2013). Individuals with AT present with staggering gait, tremors, other cerebellar signs, dysarthria, dysphagia, and telangiectasia. It is a rare condition and has been reported to occur worldwide in 1 in 40,000 and 1 in 1,00,000 live births (Rothblum-Oviatt, Wright, Lefton-Greif, McGrath-Morrow, *et al.* 2016).

The present case study focuses on two siblings who had dysarthria due to a lesion in the cerebellum and were diagnosed with Ataxia Telangiectasia Like disorder (ATLD), a variant of Ataxia Telangiectasia which falls under the broad hereditary category of spinocerebellar ataxias. Dysarthria is a group of motor speech disorders caused by neurological injury and characterized by abnormalities of breathing, phonation, articulation, and resonance patterns to irregularities of strength, rate, amplitude, firmness, tone, or precision of the speech articulators (Ludlow, 1994). A decrease in speech intelligibility and speech naturalness is associated with dysarthria. These aspects can cause challenges in participation in daily activities, changes in self-identity, social and emotional ruptures, and feelings of stigmatization, and thus reduce the quality of life.

Patients with progressive neurodegenerative disorders require long-term medical and supportive management, including speech therapy, physiotherapy, occupational therapy, behavioral therapy, or psychotherapy. The choice for a particular treatment depends on the clinical signs and symptoms the patient presents with.

Speech therapy needs to be provided to facilitate communication. Treatment approaches are either restorative, aimed at restoring the impaired function, or compensatory, aimed at compensating for deficits that are not responsive to retraining. The restorative treatment focuses on reducing the underlying impairment in the speech-production subsystems of respiration, phonation, articulation, resonation, and prosody. On the other hand, the compensatory approaches include implementing behavioral changes, such as decreasing speech rate, improving loudness, and modifying intonation to improve intelligibility. The programmed subsystem approach is restorative, and its effects on speech intelligibility and naturalness in individuals with AT have not been reported.

Further, according to the ICF model (World Health Organization 2001), treatment should focus on making an individual as independent as possible by facilitating the social use of speech and language and improving the quality of life. Therefore, it is also essential to assess the impact of therapy on the overall quality of life. There is a dearth of such studies, particularly in the Indian scenario.

2. AIM

This study compares speech intelligibility, naturalness, and overall quality of life before and after supportive speech therapy in siblings diagnosed with moderate developmental dysarthria secondary to ATLD.

3. METHOD

Two female siblings, who were native Malayalam speakers, 30.2 years and 26 years, respectively, diagnosed with moderate developmental dysarthria secondary to ATLD, were the participants of the study. They contacted the clinic through online mode with complaints of slurring, slow speech rate, difficulty walking, other motor difficulties, and eating and swallowing difficulties. The results of the MRI showed cerebellar atrophy involving superior and middle cerebellar hemispheres and superior vermis, indicating a spinocerebellar ataxia variety.

The Frenchay Dysarthria Assessment tool (FDA, Enderby 2011) was administered online to both siblings. This test includes eight subsections: reflex, respiration, lips, jaw, palate, laryngeal, tongue, and intelligibility. Speech therapy was delivered online, focusing on improving respiration, phonation, resonance, articulation, and prosody using the "Programmed Subsystem" approach. The siblings attended speech therapy for five months (3 therapy sessions per week) for 45 minutes each. During therapy, the focus was on improving each subsystem individually using a combination of facilitatory and compensatory strategies. Postural management through appropriate positioning was implemented, enhancing the capability of the respiratory system. Postural management was achieved using verbal and tactile reminders, clinician modeling, and prompting to maintain the correct posture. The therapy also focused on improving thoracic-abdominal breathing and obtaining good breath support for speech. With visual feedback, the participants were encouraged to place their hands on the abdomen to learn the inhalation and exhalation movements. The duration of exhalation, phonation, and speaking immediately on exhalation was also worked on. Oral-motor exercises to improve specific neuromuscular features like strength, steadiness, accuracy, range, and speed of movement of the articulators were carried out. Articulatory drills for imprecise sounds at syllable, word, phrase, and sentence levels were implemented using the Phonetic Placement method. The stress and intonation patterns were worked on using contrastive stress drills and pitch range exercises.

The FDA tool was re-administered on both siblings after five months of speech therapy to document the progress in the subsystems. The speech intelligibility and naturalness using picture description tasks were assessed for both siblings through pre and post-speech therapy using the Protocol for Assessment of Speech Intelligibility and Speech Naturalness in Dysarthria (D'Silva and Rajanna 2008). The picture was a line drawing about a 'market scene,' and the participants were asked to describe it. As a part of the speech naturalness assessment, stress, intonation, pauses, rhythm, rate of speech, and articulatory proficiency were assessed on a 2-point rating scale. Score '0' indicated normal, and '1' indicated that the naturalness parameter was affected.

The Quality of life (QoL) was evaluated using a scale called "Quality of life for dysarthric speakers"(Piacentini, Zuin, Cattaneo, Schundler 2011), which helps in understanding the perspective of the participant on the impact of speech therapy. The tool is a 40-item self-assessment tool that has four domains like Speech characteristics of a word (SC), Situational difficulty (SD), Compensatory strategies (CS), and Perceived reactions (PR). Participants had to respond with the following options: always (score = 4), often (score = 3), occasionally (score = 2), seldom (score = 1), or never (score = 0), depending on how frequently they experienced that situation in routine activities. The total score ranged from 0 to 160.

4. ANALYSIS

The progress seen in the speech subsystems through therapy was documented. The samples obtained through the picture description task were transcribed by postgraduate students in speech-language pathology who were native speakers of Malayalam. The narration intelligibility percentage was calculated using the formulae given below:

$$\frac{\textit{Number of intelligible words in narrated sample X 100}}{\textit{Number of words in the narrated sample}}$$

Higher scores in average intelligibility indicated better speech intelligibility. Naturalness measures were obtained by considering the six parameters mentioned in the protocol. A higher naturalness score indicated the inappropriateness of parameters contributing to speech naturalness. Finally, the scores specific for each domain and the total in QoL-DyS were calculated by adding the subdomain scores for each participant. Higher scores indicated severely compromised QoL.

5. RESULTS

The FDA profiles revealed that the elder sibling had significant difficulty with the coughing and swallowing reflex subsection. She had difficulty with speech and non-speech tasks related to the respiratory

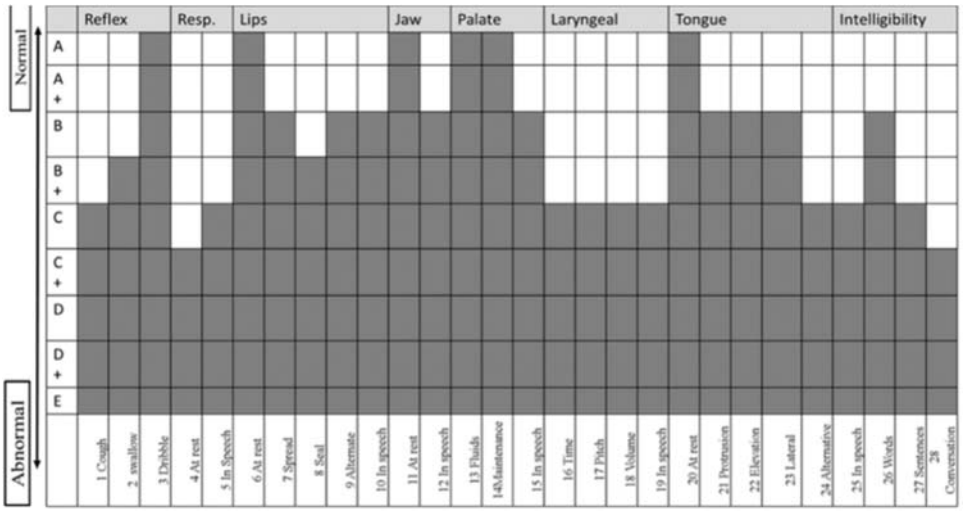

Fig. 1. FDA Profile of the elder sibling before therapy.

system, along with limited pitch and loudness ranges, indicating difficulty at the level of the phonatory system. The movement of articulators like the jaw, tongue, and lips was also restricted. She had better intelligibility at the word level relative to the sentence and conversation levels. The results of the elder sibling before the initiation of speech therapy have been graphically represented on the FDA profile in Fig. 1. The horizontal axis on the top depicts the sections, and the bottom represents the tasks. The vertical axis depicts the grade.

On the other hand, the FDA profile of the younger sibling revealed that the maximally affected system was the phonatory system with limited pitch and loudness range. There was some difficulty in the respiratory system and tongue movements in speech. In both the siblings, these respiratory difficulties manifested in the form of short utterance lengths, gasping for breath at the end of the utterances, reduction in intelligibility as the utterance length increased, and short rushes in speech occasionally. Also, the younger sibling had better intelligibility at the word level than at the sentence and conversation level. The results of the FDA of the younger sibling before the initiation of speech therapy are shown in Fig. 2.

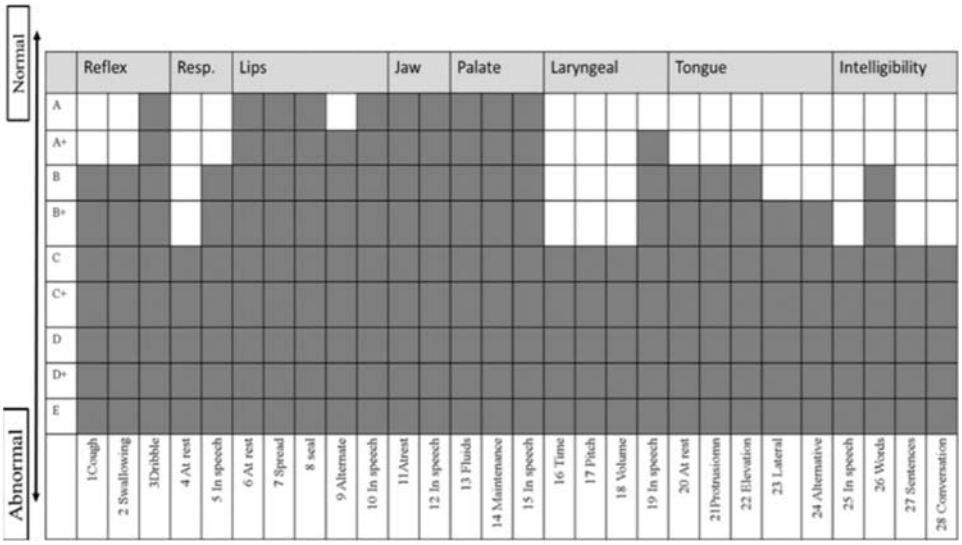

Fig. 2. FDA Profile of the younger sibling before therapy.

The siblings attended speech therapy online, after which a re-assessment was carried out. The pre-therapy and the post-therapy FDA profiles of the siblings have been depicted in Fig. 3. and Fig. 4. for easy comparison. The light gray areas that have been shaded depict the FDA profile before therapy initiation, and the darker shades show the progress after therapy. Progress was seen in both siblings on all the sections across a few tasks.

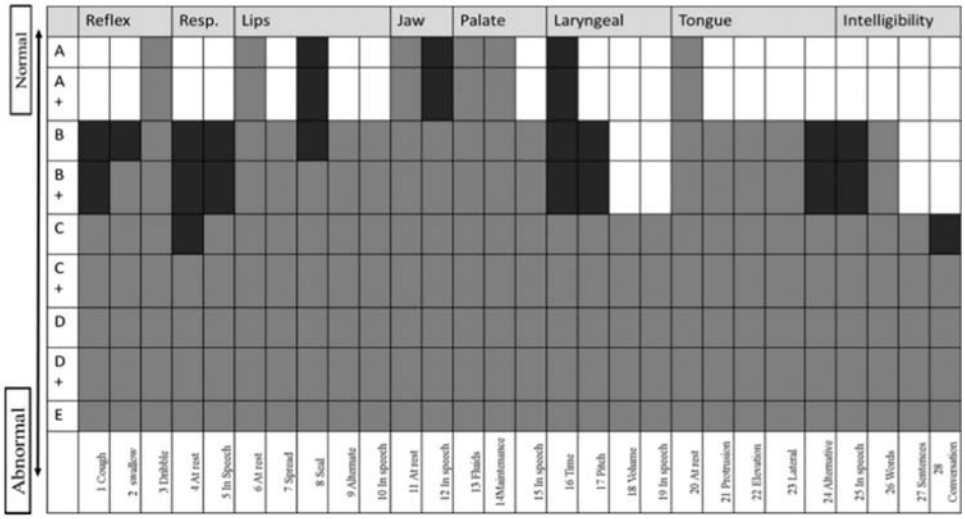

Fig. 3. FDA profile of the elder sibling before and after therapy.

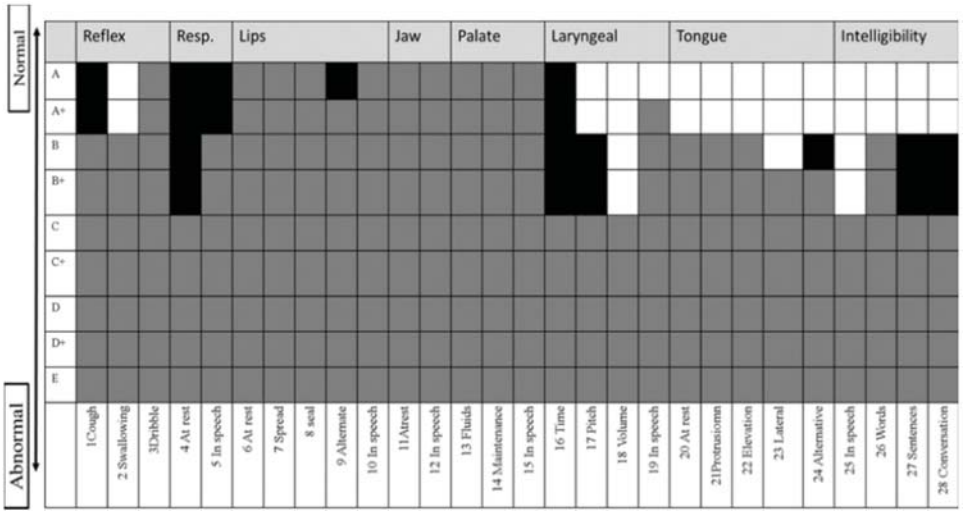

Fig. 4. FDA profile of the younger sibling before and after therapy.

The percentage of speech intelligibility before the initiation of speech therapy was 66.6%, which increased to 83.3% in the elder sibling and from 70% to 87.5% in the younger sibling. The total score of speech naturalness was '6' in the elder sibling before therapy, which decreased to '4.5' and in the younger sibling, the score decreased from '5' to '3'. The measures of intelligibility and naturalness are represented in the Table 1.

When the QoL scores were analyzed, both siblings had a score of 84 before therapy which was interpreted as "Moderately compromised QoL." The elder sibling's QoL score changed to 78, which

Table 1. Pre-therapy and Post-therapy scores on Speech Intelligibility and Speech Naturalness.

Speech outcome measures	Pre-therapy measures		Post-therapy measures	
	Elder sibling	Younger sibling	Elder sibling	Younger sibling
Speech intelligibility				
Number of words intelligible	12	28	15	35
Total number of words produced	18	40	18	40
Percentage of speech intelligibility	66.6%	70%	83.3%	87.5%
Speech Naturalness				
Use of stress	Inappropriate stress: 1	Reduced stress/ excess stress:1	Inappropriate stress: 1	Reduced stress/ excess stress:1
Use of intonation	Monotonous: 1	Monotonous:1	Monotonous: 0.5	Monotonous:0.5
Use of pauses	Inappropriate: 1	Inappropriate:1	Inappropriate: 0.5	Appropriate: 0
Use of rhythm	Dysrhythmic: 1	Appropriate: 0	Dysrhythmic: 1	Appropriate: 0
Rate of speech	Abnormal: (slow): 1	Abnormal (slow):1	Abnormal: (slow): 1	Abnormal:1
Articulatory proficiency	Affected: 1	Poor:1 (frequency of articulatory errors was more)	Improving: 0.5	Improving: 0.5 (the frequency of articulatory errors has reduced)
Total	6	5	5	3

indicated "Moderately compromised QoL" as per the norms given in the QoL measure. On the other hand, the younger sibling's post-therapy QoL score was 44, which indicated a "Mildly compromised QoL." The score reduction indicated an overall improvement in the quality of life in both siblings post-therapy.

6. DISCUSSION

The improvement in speech intelligibility and naturalness in the siblings, despite the progressive nature of the disorder, could be attributed to the speech therapy provided, which is supplemented with intensive home training and practice sessions. The therapy activities were initiated at the physiological level with the help of baseline measures. The activities neither created a fatigue effect on the participants nor was it less demanding on the physiological system. Progress was seen in both siblings on all the sections of the FDA, indicating improvement across all the subsystems of speech production.

Similar results were obtained in another study, where a subsystem approach was used on a 20-year-old with dysarthria secondary to traumatic brain injury (Netsell and Daniel 1979). The authors of this study use the term "component" to refer to the "subsystem." The approach emphasized the component-by-component analysis of the peripheral speech mechanism. The selection and sequencing of treatment procedures were based on the physiologic nature and severity of involvement in each component. They provided biofeedback during therapy. It was reported that speech intelligibility improved from approximately 5-10% to 95% during the rehabilitation. However, in the present study, biofeedback could not be provided as the therapy was conducted online.

Also, a 45-year-old person with flaccid dysarthria secondary to stroke had physiological responses similar to the present study. Therapy obeyed the hierarchy of motor speech treatment, initiating with respiration and resonance rehabilitation, followed by prosodic, phonation, and, lastly, articulatory treatment. The patient showed improvements in all motor speech bases. Some of the improvements reported by the authors include adequacy in respiratory support, resonance, prosody, articulatory precision, and vocal stability. The patient also reported improvement in quality of life (Portalete, Urrutia, Pagliarin, and Keske-Soares 2019).

The other factors which contributed to the improvement in the siblings in the current study were constant motivation, willingness to attend speech therapy, and family support. This result is consistent with the finding of the impact of focus and motivation on the prognosis of speech therapy (Ebert and Kohnert 2010). The participants also had a healthy competition to perform better, which served as a constant reinforcer for them, the caregiver, and the therapist. Regularity in attending the therapy sessions, trust in the clinician, and the shared professional bond also played an essential role in the success achieved in therapy.

The improvements in speech intelligibility and naturalness also lead to improved quality of life in both siblings. The quality of life improved to a greater extent in the younger sibling, which could be attributed to the difference in the extent of speech subsystems affected in both of them. The severity of speech problems was more significant in the elder sibling, as reflected through the FDA profiles (Fig. 1 and Fig. 2). More sections under each subsystem were affected in the elder sibling.

There are some shreds of evidence highlighting the feasibility of online mode in cases with neurological issues. A study for validating internet-based telehealth applications for assessing motor speech disorders in adults was conducted. A 70 to 100% level of agreement was achieved for online and offline assessment for 17 of the 23 FDA variables (Waite, Cahill, Theodares, and Busuttin 2006). These findings suggest that internet-based assessment has the potential as a reliable method for assessing motor speech disorders. Further, a review supported using the online practice as an appropriate service model in speech-language pathology for adults with chronic aphasia, apraxia, dysarthria, dysphagia, and Parkinson's disease (Weidner and Lowman 2020).

However, certain challenges had to be faced due to online services. First, the technical issues with connectivity and the internet were one of the major hurdles. Secondly, it was difficult to assess some neuromuscular features like the tone and strength of the muscles through the virtual mode. In addition, instrumental assessment of speech and swallowing was challenging. Nevertheless, the online mode of service delivery was helpful in the siblings considered in the present study.

7. CONCLUSION

This study documents the effect of speech therapy on speech intelligibility and naturalness in siblings with ATLD. The "Programmed Subsystem Approach," which targets the different speech subsystems in a hierarchy, contributed to improved speech intelligibility, naturalness, and better physiological support. This study also showed the impact of speech therapy on the quality of life in siblings with a progressive neurologic condition. Caution should be exercised while generalizing the results, as this is only a case report. Future studies should be carried out on a larger sample. Speech intelligibility and naturalness can be computed using spontaneous speech tasks more representative of speech used in daily contexts. The objective measures of each subsystem can also be derived and correlated with the perceptual assessment.

8. REFERENCES

[1] Chen, Yuxi, Mucha, Clayton and Oakes, Devin, 2013. "Ataxia Telangiectasia". Accessed December **8,** 2021. https://now.aapmr.org/ataxia-telangiectasia/

[2] Ebert, Kerry and Kohnert, Kathryn, 2010. "Common Factors in Speech-Language Treatment: An Exploratory Study of Effective Clinicians." *Journal of Communication Disorders*, **43**(2), 133-147. https://doi.org/10.1016/j.jcomdis.2009.12.002.

[3] Enderby, Pamela, 2011. "Frenchay Dysarthria Assessment." *International Journal of Language & Communication Disorders*, **15**(3), 165-173. https://doi.org/10.3109/13682828009112541.

[4] Ludlow, Christy, 1994. "Motor Speech Disorders." *Neurology*, **44.11**. https://doi.org/10.1212/wnl.44.11.2220-b.

[5] Netsell, Ronald and Daniel, Billie, 1979. "Dysarthria in Adults: Physiologic Approach to Rehabilitation." *Archives of Physical Medicine and Rehabilitation*, **60**(11), 502-508.

[6] Oliva, D'Silva, Rajanna and Manjula, 2008. "Protocol for assessment of speech intelligibility and speech naturalness in dysarthric in Kannada." MSc diss., University of Mysore.

[7] Piacentini, Valentina, Annalisa, Zuin, Cattaneo, Davide and Schindler Antonio, 2011. "Reliability and Validity of an Instrument to Measure Quality of Life in the Dysarthric Speaker." *Folia Phoniatrica et Logopaedica*, **63**(6), 289-295. https://doi.org/10.1159/000322800.

[8] Portalete, Caroline, Urrutia, Gabriel, Pagliarin, Karina and Marcia Keske-Soares, Marcia, 2019. "Motor speech treatment in flaccid dysarthria: a case report." *Audiology Communication Research*, **24,** n.pag. https://doi.org/10.1590/2317-6431-2018-2118.

[9] Rothblum-Oviatt, Cynthia, Wright, Jennifer, Lefton-Greif, Maureen, McGrath-Morrow, Sharon, Crawford, Thomas, and Lederman, Howard. 2016. "Ataxia Telangiectasia: A Review." *Orphanet Journal of Rare Diseases*, **11**(1), 159.

[10] Waite, Monique, Louise, Cahill, Theodaras, Deborah, Busuttin, Sarah, Russel, Trevor, 2006. "A pilot study of online assessment of childhood speech disorders." Journal of Telemedicine and Telecare, **12,** 92-94.

[11] Weidner, Kristen and Lowman, Joneen, 2020. "Telepractice for Adult Speech-Language Pathology Services: A Systematic Review." *Perspectives of the ASHA Special Interest Groups*, **5**(1), 326-338. https://doi.org/10.1044/2019_persp-19-00146.

[12] World Health Organization, 2001. ICF, International Classification of Functioning, Disability, and Health.Geneva: *World Health Organization.*

The Journal of Acoustical Society of India : Vol. 50, No. 2, 2023 (pp. 91-98)

Application of intelligent hybrid selection alogorithm for the identification of Tabla strokes in North Indian Classical Music

Shambhavi Shete, Saurabh Deshmukh and Anjali Burande

Maharashtra Institute of Technology, Aurangabad, India

e-mail: shambhavishete2020@gmail.com

[Received: 19-03-2022; Revised: 29-05-2022; Accepted: 17-06-2022]

ABSTRACT

Tabla is the most useful accompanying percussion rhythm instrument used in North Indian Classical Music. The homophonic sound produced by the Tabla instrument generates multiple harmonics which are damped by the rim of the Tabla. Therefore, Tabla stroke identification is a challenging task due to its homophonic texture. For the applications such as Automatic Tabla Stroke identification, Tabla Stroke classification, and Automatic Tabla Stroke Transcription, it is essential to correctly identify the Tabla stroke. Apart from solo strokes originating from the left and right drums, some strokes are considered to be basic Tabla Strokes that originated from both the drums simultaneously such as 'Dha', 'Dhin', and 'Tin'. Identification of such Tabla strokes is crucial and challenging due to many parameters such as the time instance at which left and right drums are hit by fingers to produce the stroke simultaneously, the location of the finger on each drum for each repetition of the same stroke, and the pressure with which the fingers hit both the drum membranes. The traditional Audio feature extractors such as Mel Frequency Cepstral Coefficient (MFCC) and Timbral Audio Descriptors are not sufficient to classify the Tabla strokes originating from both drums simultaneously in this context. Hence, an extensional method to select the correct audio descriptors is proposed here which enhances the application of the Hybrid Selection Algorithm for the selection of appropriate audio descriptors for the identification of Tabla strokes, especially for the strokes produced by the simultaneous hit on both the drums. The results show that the application of the proposed system of the Intelligent Hybrid Selection Algorithm, used to select the appropriate audio descriptors, for the identification of Tabla strokes supersedes the performance of the Hybrid Selection Algorithm. The parameters used to compare the performances are the accuracy of the Tabla Stroke identification, the complexity of the neural network, and the convergence of the learning algorithm due to the Intelligent pre-processing carried away before the application of the Hybrid Selection Algorithm. For a set of Timbral Audio Descriptors, the accuracy obtained for the Intelligent Hybrid Selection algorithm performance of the proposed system is found to be 89% as compared with the traditional Hybrid Selection algorithm with stroke identification accuracy of 81%. A Feed Forward Neural Network is used for the classification of the Tabla stroke source.

1. INTRODUCTION

Sound information retrieval is a branch of sound engineering that deals with information extraction from sound and its utilization for a given application. Indian classical music is divided into two types, North Indian Classical Music (NICM) and Carnatic Classical Music (South Indian Classical Music). The difference between the two traditions is in terms of style of singing, note presentation, variations in the structure of musical notes, and the accompanying instruments used. In NICM the music is structured around a group of musical notes called Raga (Chordia and Rae, Raag Recognition Using Pitch-Class and Pitch-Class Dyad Distributions 2007). The vocalist or an instrumentalist performs recitation of the raga in different tempos, namely Vilambit Laya (Slow Tempo), Madhya Laya (Mid-Tempo), and Drut Laya (Fast Tempo) (Singha 2018). The vocalist is accompanied by different types of musical rhythm instruments such as Tabla, Pakhawaj, Dholaki, Dhol, Mrudangam, and Dholak. In this research, musical performances of the Tabla instrument are considered.

Automatic Tabla stroke identification has various applications such as tempo detection (Bhaduri, Saha, and Majumdar 2014), Tala (rhythm) prediction (Shete and Deshmukh, North Indian Classical Music Tabla Tala (Rhythm) Prediction System Using Machine Learning 1 June 2021), sound source separation (Dittmar 2018), automatic music score generation (Gunawan, Iman and Suhartono 2020), and musical notes transcription (Bello and Plumbley 2004) (Duan and Benetos October 26, 2015). Automatic identification of Tabla stroke is accomplished using audio features. These features are called audio descriptors. The audio descriptors are the features that uniquely describe an audio (Peeters 2004). There are two major types of audio descriptors, temporal and spectral. Audio descriptors are also classified based on the time at which they are considered, e.g., instantaneous audio descriptors and audio descriptors of the entire audio signal (Peeters 2004). Mel Frequency Cepstral Coefficient (MFCC) is proven to be one of the best suitable techniques for the applications of speech and music (Jensen, et al. September 4-8, 2006) (Thiruvengatanadhan 2017). Timbre is the non-tangible, fourth dimension of a sound (Park 2004). This fourth dimension is multi-dimensional. Timbre is that attribute of sound that makes a human distinguish between two musical instrument sounds, such as flute and violin, producing a musical note with the same frequency, duration, and amplitude.

A novel method of identifying the Tabla strokes based on the pre-recognized Tabla stroke sources is proposed here. In this method, a machine learning technique is applied to identify the source of the Tabla stroke based on their Timbral attributes. To recognize an unknown Tabla stroke, the source of the Tabla stroke is first recognized. After recognizing the source of the Tabla strokes, respective Tabla stroke types are identified using a simple K-Nearest Neighbor (K-NN) classifier. Thus, before the application of the Hybrid Selection Algorithm, the source of the Tabla strokes is intelligently separated using machine learning, making the entire procedure a Tabla stroke identification using Intelligent Hybrid Selection Algorithm.

The Tabla strokes are identified using the proposed system and compared the performance of the system with the application of the traditional Hybrid Selection Algorithm used for the selection of appropriate audio descriptors and classification of the Tabla strokes.

This manuscript is organized as follows. Section 2 covers the state of the art related to the application of the Intelligent Hybrid Selection Algorithm for the identification of Tabla strokes. Section 3 proposes an effective implementation of an Intelligent Hybrid Selection Algorithm and methodology for calibrating the system accuracy. The experiments carried out and related conclusions are expressed in section 4. Section 5 concludes the research conclusions related to the application of the intelligent Hybrid Selection Algorithm for the identification of the Tabla strokes.

2. LITERATURE SURVEY

North Indian Classical Music (NICM) is based on melody, rhythm, and harmony (Agarwal, Karnick, and Bhiksha 2013). Tabla is one of the important accompanying percussion instruments which provide explicit rhythmic structure. Tabla is a set of two drums of different sizes and shapes. The sound of Tabla

is produced by tapping a finger on the drums with hands (left, right, or both). The methods with which the fingers are tapped produce different types of sounds called Tabla strokes (Datta, et al. 2017). Many times, the Tabla player produces a Tabla stroke by hitting the drum at a slightly different location. Although the sound produced by the stroke appears to be the same to a human ear, it makes a difference in the temporal and spectral descriptor values of the same stroke. Also, even for a professional Tabla player, it is difficult to apply the same pressure over the Tabla skin each time producing the same stroke. This introduces human error in the production of the Tabla strokes and makes a large difference in the Timbral aspects of the Tabla stroke to its source.

Automatic labeling of Tabla signals was accomplished using Hidden Markov Model (HMM). Mel Frequency Cepstral Coefficient (MFCC) technique was applied to extract the coefficient from the Tabla strokes segmented from 64 phrases of Tabla performance. The model executes similarly to the language models that were applied in large vocabulary speech recognition systems giving a lower error rate (Gillet and Gael 2003). For automatic labeling of the strokes, systems were implemented with or without using language modeling. Language modeling was useful to identify and separate acoustically similar strokes such as 'Ti' and 'Te'. The comparison showed that the system performance could be enhanced where language modeling was not used when using more sophisticated audio descriptors. The Timbral distinction between the similar stroke labels was the same. Therefore, an appropriate selection of audio descriptors is essential for automatic Tabla stroke identification (Chordia, Segmentation, and Recognition of Tabla Strokes September 11-15 2005). Previously Automatic transcription of percussion instruments has been done. The classification of percussion notes from nine different instruments has been carried out using a wide variety of spectral and temporal features (Herrera, Yeterian, and Gouyon 2002).

The Tabla strokes are categorized into three parts. Tabla strokes originated from the left drum, strokes originated from the right drum, and combined strokes were generated by hitting on both drums simultaneously. The basic stroke considered for the Tabla instrument also includes speedy alteration of both the drum strokes which are not considered in this research.

When the Tabla player strikes the skin of the Tabla, due to the presence of the ink on the surface, a resonating sound is produced. This generated sound has a homophonic texture. In contrast to western music, where all the musical instruments in an orchestra are played in different melody lines, simultaneously producing a polyphonic texture of the music. The drum set used in western music has different components tuned in different frequencies, that are simple to recognize and filter out, however, the resonating sound generated from the left drum, right drum, or both the drum Tabla produces a non-separable audio mixture (Raman 1934).

The acoustic sound produced by the Tabla instrument exhibits a blend of many frequencies together. The audio descriptors are useful to model the audio characteristics of the Tabla sound. There exist many audio descriptors broadly categorized as temporal and spectral, also in recent research Timbral audio descriptors are specified. The timbre of a sound is its fourth dimension other than amplitude, duration, and frequency, which is undefined. Various research has been done to model the Timbral aspect of a sound using Temporal and Spectral audio features (Park 2004). According to Music Information Retrieval toolbox developers, attack time, attack slope, zero-crossing rate (ZCR), roll-off, brightness, roughness, and irregularity are some of the audio descriptors that define Timbre (Lartillot n.d.).

Audio feature selection is a challenging task for any sound recognition application. Depending on the type of sound, the sound texture, and the complexity of the sound, different audio descriptors should be used carefully.

Traditionally, there exist two main approaches to selecting appropriate and rejecting the not useful audio descriptors. The filter and wrapper approaches are the two methods used to identify appropriate audio descriptors for the application of sound analysis (Nnamoko, et al. 2014). In the filter approach, based on a predefined criterion, audio descriptors are selected, and the remaining are discarded. In the wrapper approach, the output of a classifier is considered in the decision of selecting appropriate audio descriptors. The wrapper approach is of four types, namely, forward selection, backward selection, bidirectional, and

hybrid selection. The Hybrid Selection Algorithm (HSA) is an iterative procedure that considers the combination of audio descriptors and rejects the combination that decreases the accuracy of the system (Deshmukh 2012). The hybrid selection algorithm is purely based on accuracy considerations of the combination of audio descriptors but does not consider the Timbral attribute of the audio descriptor which may be useful for the applications.

To identify a Tabla stroke automatically, an appropriate selection of Timbral audio descriptors is essential based on the source of the production of the Tabla stroke. This research proposes an application of the Intelligent Hybrid Selection Algorithm that first separates the Tabla strokes based on their sources using machine learning and then selects appropriate audio descriptors for the identification of Tabla strokes using the supervised K-Nearest Neighbor algorithm.

3. PROPOSED SYSTEM

Automatic Tabla stroke identification is useful for many applications that include automatic Tabla stroke transcription (Chordia, Automatic Transcription of Tabla Music 2006), rhythm identification (Srinivasamurthy, et al. 2017), the discovery of percussion patterns (Gupta 20 September 2015), Tala prediction (Miron 2011), etc. As stated earlier a basic Tabla stroke is generated in three different ways, a stroke produced by hitting the left drum alone, the right drum alone, or both drums simultaneously. The specialty of NICM Tabla stroke is that although the stroke produced from both drums is treated as one stroke, it comprises basic strokes produced from the left and right drums. For example, a left drum stroke 'Ga', and a right drum stroke 'Ta', when produced simultaneously generate the basic stroke 'Dha' from the category of stroke produced from both drums simultaneously. This shows that, in the series of continuous Tabla strokes to be identified one after the other, it is crucial to clearly distinguish between single drum strokes (left or right) and basic drum strokes from both drums. Therefore, it is important before identifying the Tabla stroke to identify the source of the stroke to correctly identify the Tabla stroke type. Traditionally, various attempts were made to identify the Tabla strokes based on the audio attributes of each stroke. In each of these attempts, the strokes are classified merely based on the audio attributes of the stroke, where the source of the Tabla stroke is not considered anywhere. In this research we propose a novel method to enhance the capability of the Hybrid Selection Algorithm to correctly identify the audio descriptors useful for Tabla stroke identification by assigning a trained neural network that can correctly identify the source of the Tabla strokes.

Nine Tabla strokes are considered basic Tabla strokes out of which two strokes are produced from the left drum namely 'Ga' and 'Ka', and four strokes from the right drum namely, 'Na', 'Ta', 'Te', 'Tun' and three strokes from both the drums simultaneously, 'Dha', 'Dhin', 'Tin'. Traditionally, MFCC and spectral audio descriptors are popularly used as audio features in similar applications (Shete and Deshmukh, Automatic Tabla Stroke Source Separation Using Machine Learning 2021). The Tabla stroke identification accuracy is based on the total number of strokes correctly identified by the classifier. The performance of the classifier to correctly identify Tabla stroke depends on the audio features used. From the pool of many audio descriptors available, it is difficult to select appropriate audio descriptors. The Hybrid Selection Algorithm (HSA) iteratively attempts to finalize appropriate audio descriptors useful for stroke identification. Being the wrapper approach, HSA considers the output of the classifier to decide the usefulness of a descriptor. However, HSA does not consider the variations in the Timbral quality of the Tabla stroke sources.

The strokes produced by the Tabla instrument, due to the presence of the ink on the surface of the drums, generate homophonic sound texture. It is important to note that the sound texture of the Tabla stroke, considered one stroke, originating from both drums together, is complex, due to the non-separable mixture of the sound produced from both drums. Therefore, before selecting appropriate audio descriptors, using HSA, it is important to separate the strokes based on their source of production.

In the proposed system shown in Fig. 1, an Intelligent Hybrid Selection Algorithm (IHSA) is applied such that, each Tabla stroke is first recognized concerning its source of production, and then based on

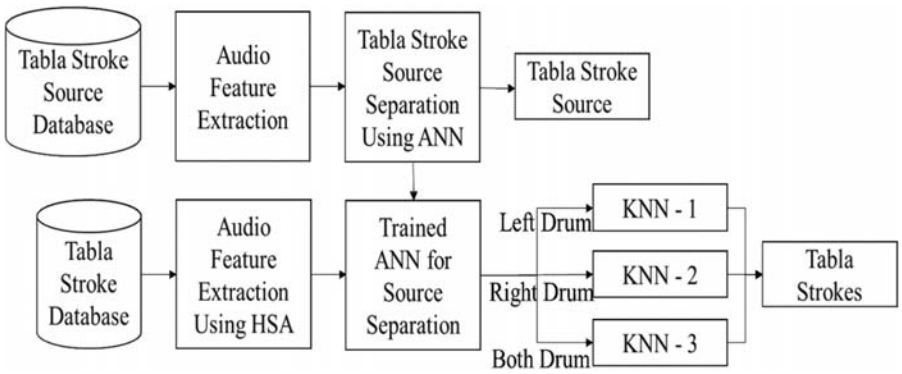

Fig. 1. Intelligent Hybrid Selection Algorithm for Tabla Stroke Identification

the type of the source a simple classifier is applied to identify the Tabla stroke. A Feed Forward Neural Network (FFNN) is trained separately using a Tabla stroke database which is categorized based on the source of the Tabla strokes. A total of nine basic Tabla strokes are included based on their categories of source in the database. The Neural Network is trained using 448 audio excerpts (left drum - 100 strokes, right drum - 168 strokes, both the drum - 180 strokes) that categorizes the Tabla strokes based on the three sources of their production.

Automatic onset detection is applied over the studio recording of a continuous series of Tabla strokes performed by a professional Tabla player for the duration of ~20 minutes. Based on the locations of the audio onsets, the Tabla strokes are segmented and separately saved in the .wav file format with a sampling frequency of 44,100 Hz. A total of 624 audio excerpts are extracted containing various instances of each of the nine basic Tabla strokes.

Three K-NN classifiers are separately deployed for the identification of the Tabla strokes such that each one of them is dedicatedly used to classify the Tabla stroke type based on its source. First K-NN is used only to identify Tabla strokes produced from the left drum. Similarly, the second and third K-NN classifiers are designed to identify Tabla strokes generated from the right drum and both drums respectively.

4. EXPERIMENTS AND RESULTS

An Intelligent Hybrid Selection Algorithm that utilizes a trained neural network to classify the Tabla strokes based on their source of production is proposed here. Audio descriptors play an important role in the extraction of useful information from an audio excerpt which is further used by the classifiers to identify the Tabla strokes. A comparison of the application of traditional audio descriptors versus Timbral audio descriptors for the identification of Tabla stroke shows that the Timbral audio descriptors are more useful for 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). The Timbral aspects of the Tabla strokes vary concerning their source of production.

The accuracy of Feed Forward Neural Network is calculated based on correctly identified Tabla stroke sources. The result shows that out of 134 Tabla strokes the neural network could correctly identify 123 stroke sources giving a sufficiently acceptable Tabla stroke source separation accuracy of 92%. The Intelligent Hybrid Selection Algorithm uses this neural network to identify the Tabla stroke source and further identify the Tabla stroke type using the appropriate K-NN classifier. The audio descriptors used for source separation and Tabla stroke identification contain Timbral audio descriptors along with spectral and Mel Frequency Cepstral Coefficients (MFCC).

In applications that use Artificial Neural Networks as a classifier, the performance of the network is improved with the increase in the dimensionality of the features, however, the performance degrades as

the total number of features used to go beyond the optimal number of features required for the classification. It also increases the convergence time and may not lead to the correct identification of Tabla stroke. Therefore, a balance of the total number of audio descriptors used and the neural network convergence time is necessary to be maintained. A separate audio database, containing strokes originating from three sources, used to train the neural network for source separation justifies this approach. Once the source of the Tabla stroke is identified, each new instance of an audio excerpt is identified using an appropriate K-NN classifier. The K-NN classifier dedicated to the left drum Tabla stroke has two classes, representing two stroke types produced from the left drum. Similarly, the K-NN classifier is used for the right drum, and both drums have four and three classes representing corresponding drum stroke types.

Table 1 shows Tabla stroke identification accuracy obtained at the end of each pass for the Hybrid Selection Algorithm and Intelligent Hybrid Selection Algorithm using Timbral audio descriptors. It is observed that for Hybrid Selection Algorithm, the combinations used in pass 2, 3, and 4 gradually increase the Tabla stroke identification accuracy with the increase in the audio descriptors. However, when source separation is performed, in advance, using Intelligent Hybrid Selection Algorithm, the number of descriptors required is minimized increasing the system's performance and accuracy. Table 1 shows in detail the combinations and corresponding Tabla stroke identification accuracies for only Timbral audio descriptors.

Table 2 shows the set of audio descriptors obtained at the end of all passes using Timbre and a combination of Timbre with other descriptors. The result shows that, although there is the overhead of extensively training a separate FFNN to identify the Tabla stroke source, the system performs better as compared with the traditional Hybrid Selection Algorithm. System performance using Timbral audio descriptors after source separation is found to be 89% by using ZCR, MFCC, and Brightness descriptors. For the combination of Timbral and other audio descriptors, the accuracy obtained using the Intelligent Hybrid Selection Algorithm is 85%. It is observed that the system performance starts degrading with the increase in a greater number of audio descriptors.

Table 1. Tabla Stroke Identification Accuracy Using Timbral Audio Descriptors.

Timbre (ZCR, Roll Off, Brightness, Roughness, Irregularity, MFCC, Attack Time, Attack Slope)				
Pass	Hybrid Selection Algorithm	Stroke Identification Accuracy (%)	Intelligent Hybrid Selection Algorithm	Stroke Identification Accuracy (%)
Pass 1	Roll Off, Brightness, MFCC	(37, 30, 62)	ZCR, Brightness, MFCC	(52, 48, 68)
Pass 2	(MFCC,Roll Off), (MFCC,Brightness), (MFCC,Irregularity)	(61, 63, 58)	(ZCR,MFCC), (ZCR,Roghness), (ZCR,Brightness)	(75, 69, 73)
Pass 3	(MFCC,Roll Off,ZCR), (MFCC,Roll Off,Brightness)	(66, 74)	(ZCR,MFCC,Roghness), (ZCR,MFCC,Brightness)	(78, 89)
Pass 4	(MFCC,Roll Off, Brightness, Irregularity)	81	(ZCR,MFCC,Brightness)	89

Table 2. Tabla Stroke Identification Accuracy Comparison.

Timbre and Other (Centroid, RMS, Low Energy, Spread, Skewness,				
Features	HSA	Accuracy (%)	IHSA	Accuracy (%)
Timbre	MFCC, Roll Off, Brightness, Irregularity	81	ZCR, MFCC, Brightness	89
Timbre + Other	MFCC, Brightness, ZCR, Centroid, Spread, Entropy, Skewness, Flatness,	79	ZCR, MFCC, Brightness, Entropy, Inharmonicity	85

5. CONCLUSION

To enhance the performance of the existing Hybrid Selection Algorithm for the selection of appropriate audio descriptors, automatic Tabla stroke source separation is proposed here. Tabla stroke identification accuracy is majorly dominated by the process of selecting audio descriptors. The hybrid Selection Algorithm uses an iterative wrapper approach that considers classifier performance for the selection of audio descriptors. The system performs well when single drum Tabla strokes are considered. However, when both the drums produce a combined stroke that is homophonic in texture requires careful bifurcation of the source of the Tabla sound.

A Feed Forward Neural Network is trained for classifying the Tabla strokes based on their source of production. Before identifying the Tabla strokes their corresponding source of production is identified. Based on the source of the Tabla strokes, three K-NN algorithms are applied to identify the Tabla stroke type. The result shows that the Tabla stroke identification accuracy of 81% is obtained by using Hybrid Selection Algorithm and when applied to Feed Forward Neural Network for source separation using Intelligent Hybrid Selection Algorithm gives the Tabla stroke identification accuracy of 89%. The result also shows that a reduced set of audio descriptors is obtained with an increase in the overall system performance.

6. REFERENCES

[1] Agarwal, Parul, Harish Karnick and Raj Bhiksha, 2013. "A Comparative Study Of Indian And Western Music Forms." *International Society for Music Information Retrieval.*

[2] Bello J.P. and M Plumbley, 2004. "Fast Labeling of Notes in Music Signals." *5th International Conference on Music Information Retrieval, Barcelona.*

[3] Bhaduri, Susmita, Sanjoy Kumar Saha and Chandan Majumdar, 2014. "Matra and Tempo Detection for INDIC Tala-s." *Advanced Computing and Informatics Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014).* Switzerland, Springer.

[4] Chordia, Parag and Alex Rae, 2007. "Raag Recognition Using Pitch-Class and Pitch-Class Dyad Distributions." *International Conference on Music Information Retrieval, ISMIR 2007, Austria.*

[5] Chordia, Parag, 2006. Automatic Transcription of Tabla Music. Stanford, CA, USA: *Stanford University.*

[6] Chordia, Parag, 2005. "Segmentation and Recognition of Tabla Strokes." *6th International Conference on Music Information Retrieval. London. The UK.* pp. 107-114.

[7] Datta, Asoke Kumar, Sandeep Singh Solanki, Ranjan Sengupta, Soubhik Chakraborty, Kartik Mahto, and Anirban Patranabis, 2017. *Signal Analysis of Hindustani Classical Music. Singapore: Springer.* doi:10.1007/978-981-10-3959-1.

[8] 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.

[9] Dittmar, Christian, 2018. "Source Separation and Restoration of Drum Sounds in Music Recordings." Ph.D. Thesis.

[10] Duan, Zhiyao and Emmanouil Benetos, 2015. Automatic Music Transcription. Tutorial, Malaga, Spain: *ISMIR.*

[11] Gillet, Olivier and Richard Gael, 2003. "Automatic Labelling of Tabla Signals." *4th ISMIR Conference.*

[12] Gunawan, Alexander, Ananda Iman and Derwin Suhartono, 2020. "Automatic Music Generator Using Recurrent Neural Network." *International Journal of Computational Intelligence Systems,* **13**(1), 645-654. DOI: https://doi.org/10.2991/ijcis.d.200519.001.1

[13] Gupta, Swapnil, 2015. "Discovery of Percussion Patterns from Tabla Solo Recordings." *Master Thesis Report, Universitat Pompeu Fabra.*

[14] Herrera P., A. Yeterian and F. Gouyon, 2002. "Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques." Music and Artificial Intelligence, ICMAI 2002, *Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.* DOI:https://doi.org/10.1007/3-540-45722-4_8.

[15] Jensen, Jesper, Mads Christensen, Manohar Murthi and Soren Jenson, 2006. "Evaluation of MFCC estimation techniques for music similarity." Edited by EURASIP. *14th European Signal Processing Conference (EUSIPCO 2006). Florence, Italy.*

[16] Lartillot, Olivier. n.d. MIR ToolBox Manual. The University of Jyvaskyla, Finnish Centre of Excellence in Interdisciplinary Music Research.

[17] Miron, Marius, 2011. Automatic Detection of Hindustani Talas. Master Thesis, Barcelona: *Universitat Pompeu Fabra.* DOI:https://doi.org/10.5281/zenodo.1162292.

[18] Nnamoko, Nonso, Arshad Farath, David England, Jiten Vora and James Norman, 2014. "Evaluation of Filter and Wrapper Methods for Feature Selection in Supervised Machine Learning." *The 15th Annual Postgraduate Symposium on the convergence of Telecommunication, Networking, and Broadcasting. Liverpool.*

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

[20] Peeters, Geoffroy, 2004. "*A large set of audio features for sound description (similarity and classification) in the CUIDADO project"*.

[21] Raman C. V., 1934. "The Indian Musical Drums." *Proc. Indian Acad. Sci.* (Math. Sci.), **1,** 179-188. https://doi.org/10.1007/BF03035705.

[22] 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." Proceedings 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.

[23] 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.

[24] Shete, Shambhavi and Saurabh Deshmukh, 2021. "North Indian Classical Music Tabla Tala (Rhythm) Prediction System Using Machine Learning." *Advances in Intelligent Systems and Computing. Singapore: Springer.* pp. 187-197. DOI:https://doi.org/10.1007/978-981-33-6881-1_16.

[25] Singha, Vijay Prakash, 2018. An Introduction to Hindustani Classical Music: A Beginners Guide. India: *Roli Books Private Limited.*

[26] Srinivasamurthy A., A. Holzapfel, K.K. Ganguli and X. Serra, 2017. "Aspects of Tempo and Rhythmic Elaboration in Hindustani Music: A Corpus Study." *Frontiers of Digital Humanities.* doi:10.3389/fdigh.2017.00020.

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

The Journal of Acoustical Society of India : Vol. 50, No. 2, 2023 (pp. 99-104)

Measuring improvisation in Hindustani Vocal Music

Kaushik Banerjee*, Anirban Patranabis, Aditi Mondal,
Ranjan Sengupta, Argha Deb and Dipak Ghosh
Sir C. V. Raman Centre for Physics and Music, Jadavpur University, Kolkata-700 032, India
e-mail: sitar_kaushik@yahoo.com

[Received: 27-03-2022; Revised: 13-06-2022; Accepted: 11-07-2022]

ABSTRACT

In Hindustani music (HM), whether it is vocal or instrumental, improvisation is one of the key elements which prove its nobility and one of the causes of apotheosis (of HM) to the world of music. But it is neither so easy to improvise every rendition within the canonical frame of Hindustani classical raga nor every artist can do it. Here we took ten signals (5 each) of two eminent vocalists of two different generations and Gharanas. Our goal is to establish the acoustic cues that might be in the creativity and artistry of the musicians and expressed in their performances. To observe and understand the existence and execution of improvisation we took different parameters by which we knew the used notes, pause and its ratios; Meend or transition and phrasal patterns. In respect of time, the ratio of pause to note duration reflects the tempo of each rendition which is also an important cue of improvisation.

1. INTRODUCTION

A musician while performing expresses the raga (along with following the fundamentals) as per his mood and his surrounding ambiances. Thus there are differences from one rendition to another. Even if an artist sing or play same Raga and same Bandish (composition) twice (consecutive renditions) then there should be some dissimilarity in between two performances. These differences in the same raga renditions are generally called improvisation. Only expert music listeners can easily differentiate two different renditions of the same raga.

The wonders of HM lies within different musical elements, such as: Sruti or use of microtones, Meend or note to note transitions, Kwan swara or touching notes, Pakad or combination of notes which comprises the essence or main phrases of a raga and of course improvisation which includes all these elements along with Arohan - Abarohan, Vadi, Samvadi, Anubadi, Bivadi, Chalan (different note combination which represents a raga properly) and appropriate Talim (training), intense practice or Riaz and imagination [Raghava R. Menon, 1998.]. Artist's mood and the ambiance of any program are also included as important elements of improvisation in any rendering. It is said that Vilambit (slow tempo) or Maddya-vilambit (medium slow tempo) parts are best for improvisation and one can get the best impression of a raga in a rendition, [Dutta A.K.et. al. - 2019] so we took signals from medium slow tempo. This study might be suitable for differentiating Indian Gharana (style of music schools) music. To understand musical improvisation more clearly we took two sets of signals [Dutta A.K. *et al.*, 2017].

2. PROCEDURE

Five raga renderings in raga 'Sur-Malhar' or Surdasi Malhar sung by Pandit Kumar Gandharv and five raga renderings in raga Malkauns sung by Ustad Bade Ghulam Ali Khan. Both the two are great classical vocalists of the last century. Each of the raga renderings are of three minutes' duration. Raga clips of three minutes' duration were chosen in such a way that each raga renderings of all the five ragas of Sur-Malhar have same phrases and also all within the five raga renderings of Malkauns have some common phrases. All the ragas were collected from different archives in digitized form at a sampling rate of 44100Hz (16 bits per sample) in mono channel. So ten selected sound files thus constitute the database. From all signals, detection of the tonic 'Sa' position, extraction of other notes (with pitch value) and pitch profile and other respective and essential procedures regarding this experiment were done as per our previous research works [Datta A K. *et al.*, 2017, Banerjee. K. *et al.*, 2012].

Following are the key features which describe the improvisation made by the two artists in each of their performances: [a] frequency of uses of each notes - compared for all raga renderings for all notes used. This reflects the preferences of using touching notes (Kwan swara) and long durational notes. [b] Duration of notes - compared for all raga renderings for all notes used. In some rendering we found a very strong preference for a single note while in other an equal weightage was given to two or more notes. [c] Duration of silence - reflects the tempo and artist's mood during the performances. [d] Duration of transition between notes - unveiled the artist's mood and natural propensity of using Meend. In Indian classical music, transition between notes is one of the most important features of expertise and renowned artists' love using transition notes.

3. RESULTS AND DISCUSSION

At first we had measure the total number of times each note was sung and also the total duration of each note was sung by the artists. Finally, we took the ratio of duration of each note to the number of occurrence of that note, yields the average duration of each note, shown in figure 1. Average duration of notes may be a parameter for detecting the style of improvisation of an artist. Although all the ragas sung by the respective artist are of same phrase, but the average note duration was different from one raga to another. It is observed that none of the signals show equal note duration. So each performance of the artist was different from one another. It has been observed that the artist Pandit Kumar Gandharv sung the note 'Sa' for 15.19, 18.45, 12.2, 14.01 and 27.78 seconds and sung the note "pa' for 13.06, 21.57, 19.3, 16.8 and 22.04 seconds during the three minutes clips of five renderings and such differences were observed

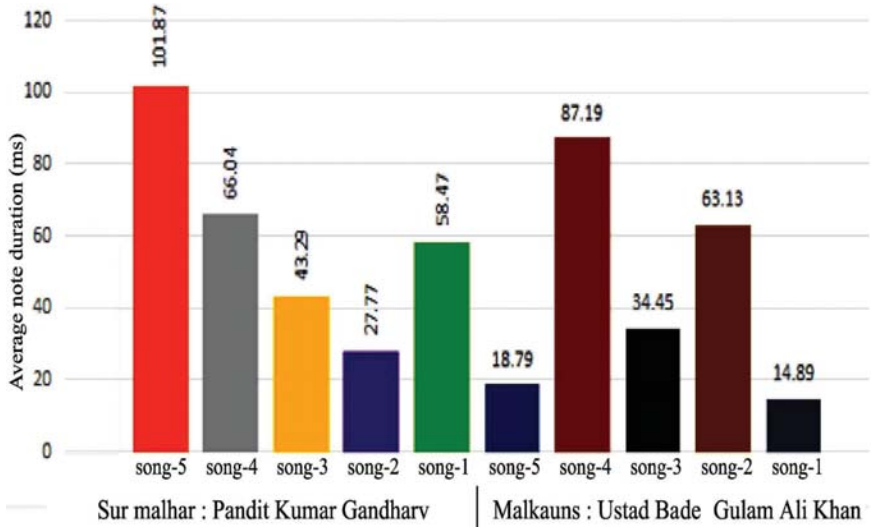

Fig. 1. Average note duration in millisecond for ten signals sung by two artists.

for all the other five notes. Similarly, it has been observed that the artist Ustad Bade Ghulam Ali Khan sung the note 'Ni' for 11.67, 9.78, 10.13, 23.34 and 8.11 seconds and sung the note 'sa' for 24.94, 14.38, 19.34, 60.06 and 11.17 seconds during the three minutes clips of five renderings and such differences were observed for all the other notes too. So the artists keep on improvising their every performance by changing the duration of notes and frequency of using notes, keeping the structure of the raga same. Apparently average note duration seems to be a cue for tempo but it is not so. In some of the renderings, artist used larger number of notes with smaller duration, while in some rendering they have used smaller number of notes with smaller duration. And the other combinations were also observed. Interestingly in every rendering artist had use different types of notes. Sometime an artist preferred using short notes while sometimes long notes. So it is artist's own choice as per the then mood of the artist as well as the the surrounding ambiances to use the notes in his own ways keeping the raga structure same. In contrast to western style of music, Indian classical music does not have fixed duration of notes and hence have no scope of improvisation. In contrast, improvisation in using notes in different ways is the key to Indian classical music. Such improvisation evokes different Rasas (emotions) in a raga among the listeners. No definite rule was found for improvisation of using notes in a raga rendering. Artist's frame of mind and audience's response may play an important role of improvisation during a raga performance by an artist.

We also identified the silence zones of the sound signals along with the time duration of the silence. It has been observed that the artist Pandit Kumar Gandharv silenced between notes for 67.22, 57.36, 24.11, 52.55 and 34.94 seconds during the three minutes clips of five renderings. Also, it has been observed that the artist Ustad Bade Ghulam Ali Khan silenced between notes for 127.22, 140.28, 137.13, 59.41 and 153.34 seconds during the three minutes clips of five renderings. Artists use the silence differently in different ragas. Different silence period in different raga renderings evoke different rasas (emotions). Although all the sung ragas had similar phrase pattern, silence in between two notes were found different. So this confirms the fact that these great maestros improvised their every performance every time they played a given raga. Now we took the ratio of silence duration to the note duration. Figure 2 shows the histogram of silence duration to the note duration of all the ten renderings. This measurement of silence to note ratio can signify certain information about the raga movement and also evoke the style of presenting a raga by an artist. It is observed that 2nd and 4th renderings are found almost similar and also the 3rd and 5th renderings were also found similar in silence to note duration of raga Sur Malhar sung by Pandit Kumar Gandharv. Also 2nd and 3rd renderings of raga Malkauns sung by Ustad Bade Ghulam Ali Khan are almost similar. So the feature like silence to note ratio may lead to identify the style of an artist [Patranabis A. *et al.*, 2021]. Another important finding from this measurement is to measure tempo of a raga rendering.

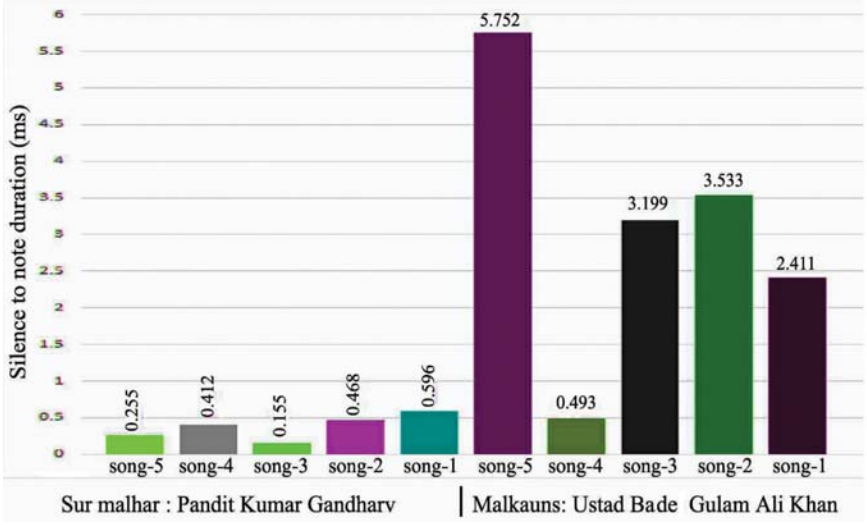

Fig. 2. Silence duration to the note duration of all the ten signals sung by two artists.

Tempo is one of the important musical features that control the rasa of Indian classical music [Patranabis A *et al.*, 2013]. Comparing the ratio of silence to note duration, we found that song 1 of Sur-Malhar is of slowest tempo while song 3 is of fastest tempo. If we arrange the other three renderings of Sur-Malhar in the ascending order of tempo we found it as song 2, song 4 and song 5. Song 5 of Malkauns is of slowest tempo while song 4 is of fastest tempo. If we arrange the other three renderings of Malkauns in the ascending order of tempo we found it as song 2, song 3 and song 1. Keeping the raga structure same artists are free to change the tempo of a raga from one rendering to another one. This is improvisation. In contrast to western style of music, Indian classical music does not have fixed silence period and improvisation in using silence period is very important in Indian classical music. It is artist's own choice to use silence in a raga rendering and hence it must play an important role of improvisation during a raga performance by an artist. Tempo of a raga rendering depends upon the frame of mind of the onset of singing.

Now the transitions between notes were identified and such transition time was measured. Such transition between two notes with a sliding tone (uninterrupted) is known as Meend in Indian classical music. Meend can be described as a mode of joining two or more notes by means of a graceful glide without any jerks. It is an essential and integral part of Indian classical music that gives the essence of Indian classical music which is strictly prohibited in western music [Datta *et al.*, 2009]. The Meend is a transition of notes that can range from a simple span of two notes to a whole octave. Meends are in general uncomplicated and smooth. Here in all the ragas the transition between the notes i.e. the basic meend is generally very slow in Pitch transition. These pitch transition and note sequences within the transitions are significant to the listeners in identifying the melodic similarity [Datta *et al.*, 2009]. We had identified the transition area between notes used by the artist and measured the total transition time for the whole signal. Figure 3 shows the total transition time for all the ten signals. Song-3 and song-4 of Malkauns by Ustad Bade Ghulam Ali Khan shows similar transition time but the other three renditions show wide variety of transition time. Song-1 and song-4 of Sur-Malhar by Pandit Kumar Gandharv shows nearly similar transition time. Song-2 and song-3 shows very large transition time while song-5 shows very low transition time. Such difference in transition time among notes and transition pattern keeping the phrasal pattern of the raga same proves the improvisation made by the artists in every raga performance. Different transition times along with transition patterns are important in evoking melodic essence of the raga. Meends were found between two notes, and more than two notes. So artists are free to use different meend patterns and no two signals were found to have uniform Meend patterns. It is found that in most of the Meend patterns for a given raga shows similar terminal note pairs but the tone movements are different. Meends were found to be linear or with up and down (complex type) and also in most of the cases in between two terminal notes, artists used very short touching notes. These touching notes evokes significant emotions among the listeners.

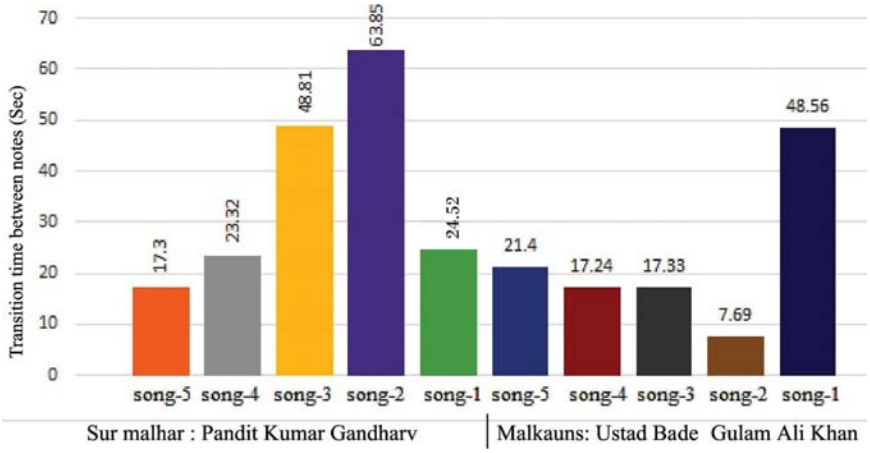

Fig. 3. Total transition time between notes of all the eleven signals sung by two artists.

Table 1. ANOVA table for raga Sur-Malhar sung by Pandit Kumar Gandharv between the group silence duration and silence to note duration

Source of variation	Sum of Squares	d.f.	Mean Squares	F	P
Between	294.2	5	58.84	0.158	0.9759
Error	1.12E+04	30	372.5		
Total	1.15E+04	35			

Table 2. ANOVA table for raga Malkauns sung by Ustad Bade Ghulam Ali between the group silence duration and silence to note duration

Source of variation	Sum of Squares	d.f.	Mean Squares	F	P
Between	294.2	5	37.64	0.12	0.901
Error	1.15E+04	15	280.1		
Total	1.19E+04	19			

Now we have conducted an ANOVA test. A detail ANOVA table for raga Sur-Malhar (a group of 5) between the parameter silence and silence to note ratio is shown in table 1 and for raga Malkauns (a group of 5) between the parameter silence and silence to note ratio is shown in table 2. Since the F ratio is more than one in both the ragas, so it means that the variation among group means is more than the expected value. The P value is determined from the F ratio and the two values for degrees of freedom are also shown in the ANOVA table. In both the cases the P value found are large (> 0.05), conclude that the standard deviations of the populations are less different and hence the populations are similar. The number of degrees of freedom ("d.f.") for the numerator (found variation of group averages) is one less than the number of groups 5; the number of degrees of freedom for the denominator (so called "error" or variation within groups or expected variation) is the total number of notes minus the total number of groups. Plot of average absolute deviation from median with 95% Confidence Intervals for both the ragas are shown in the figure 4 and 5 where the dots represents the sample mean. For the raga Sur-Malhar, song-3 and song-5 have highest mean while song-1 has lowest mean. For the raga Malkauns, song-4 has highest mean while song-5 has lowest mean. Larger P value and overlapping in absolute deviation from median indicates no significant difference in the group. This leads to the style identification of the artists. So measurement of improvisation is an important cue of style analysis.

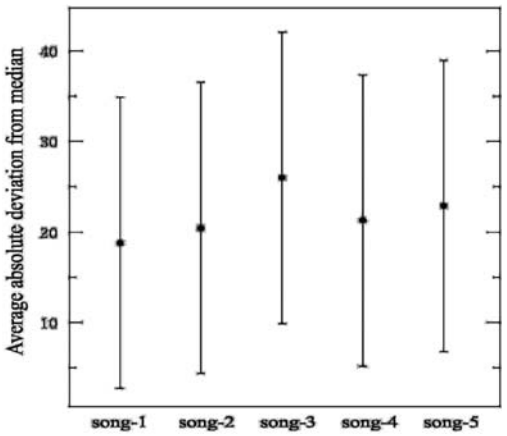

Fig. 4. Plot of average absolute deviation from median with 95% Confidence Intervals for raga Sur-Malhar.

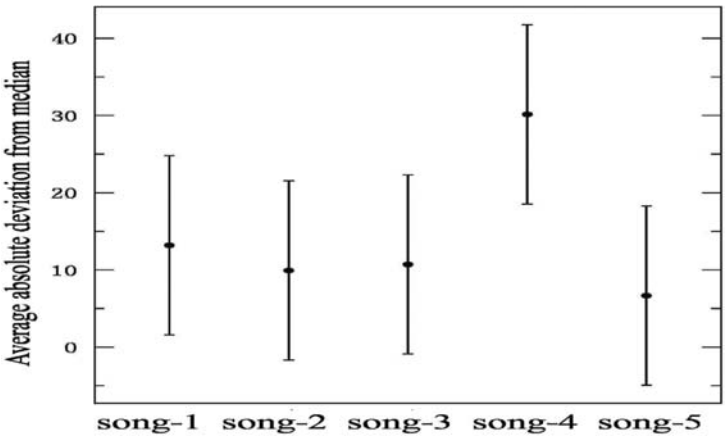

Fig. 5. Plot of average absolute deviation from median with 95% Confidence Intervals for raga Malkauns.

4. CONCLUSION

Above discussion proves that both the artists freely improvised their every performances keeping the basic of the raga fixed. It is artist's own choice to use the notes, silence and Meend in his/her own ways keeping the raga structure same. In contrast to western style of music, Indian classical music does not have fixed duration of notes, no fixed silence duration and improvisation in using these in different ways is the key to Indian classical music. Each improvised raga evokes different Rasas (emotions). No definite rule was found for improvisation, it is solely artist's discretion and it mainly depends upon the then mood and audience's response may play an important role of improvisation during a raga performance by an artist. Measuring improvisation will help the amateur musicians and young learner. This study can be used as in virtual learning tool.

5. REFERENCES

[1] Banerjee K., Patranabis A., Sengupta R. and Ghosh D., 2012. Search for spectral cues of emotion in Hindustani music. *Proceedings of the National Symposium on Acoustics-2012*, 5-7 December, KSR Institute for Engineering and Technology, KSR Kalvi Nagar, Tiruchengode-637215, Tamilnadu.

[2] Datta A. K., Sengupta R., Dey N. and Nag D., 2006. *Experimental Analysis of shrutis from performances in Hindustani music.* Monograph published by ITC Sangeet Research Academy, Kolkata, ISBN 81-903818-0-6.

[3] Datta A. K., Sengupta R., Dey N. and Mukherjee A. K., 2006. A methodology of note extraction from the song signals. *Proc. Int. Symposium- Frontiers of Research on Speech and Music, U.P. Technical University, Lucknow, India.*

[4] Datta A. K., Sengupta R. and Dey N., 2009. Objective Categorisation of Meends from Hindustani vocal performances. *J Acoust. Soc. India,* **36**(2), 92-95.

[5] Datta A. K., Solanky S. S., Chakraborty S., Sengupta R., Mahto. K. and Patranabis A., 2017. *Signal analysis of Hindustani classical music. Springer,* ISBN 978-981-10-3958-4.

[6] Datta A. K., Sengupta R., Banerjee K. and Ghosh D., 2019. *Acoustical Analysis of the Tanpura: Indian Plucked String Instrument. Springer,* ISBN 978-981-13-2609-7Raghava R. 1988. Menon. Indian Music; The Magic of the Raga. published by - Somaiya Publication PVT. LTD. Mumbai.

[7] Patranabis A., Banerjee K., Sengupta R. and Ghosh D., 2013. Measurement of emotion induced by Hindustani music- A human response and EEG study. Ninad, *Journal of the ITC Sangeet Research Academy,* **26-27.**

[8] Patranabis A., Banerjee K., Sengupta R. and Ghosh D., 2021. Identifying Style of Vocalist Using Silence in Hindustani Music. *The Journal of Acoustical Society of India,* **48**(1-2).

The Journal of Acoustical Society of India : Vol. 50, No. 2, 2023 (pp. 105-112)

Automatic spoken language identification for Indian languages using Relative Abundance Model (RAM)

Suparna Panchanan[1*], Saha Arup[2] and Datta Asoke Kr.[3]

[1]Department of Computer Science & Engineering, Brainware University, Kolkata-700125, India

[2]Sector-V, Saltlake Electronics Complex, Bidhannagar, Kolkata-700 091, India

[3]BOM Public Charitable Trust, 3/3 GirishGhosh Street, Kolkata 700 035, India

e-mail: suparna_mou2k@yahoo.co.in

[Received: 27-03-2022; Revised: 17-06-2022; Accepted: 11-07-2022]

ABSTRACT

Automatic Spoken Language Identification (ASLID) is an active research area now, particularly for India which is a multilingual country where 22 official languages are spoken. One major field of application of ASLID, is in security and intelligence service. Speech parameterization techniques LPC, MFCC, PLP and LPCC are extensively used in ASLID. India.

This paper presents the Relative Abundance Model (RAM) for ASLID. The model is based on four basic types of signal classes:

- Free voice
- Obstructed voice
- Quiescent and
- Quasi-random

The paper explores the potential of this model for ASLID with the details of methodology and the results. With the sixteen languages selected from eastern, northern, western and southern India namely Assamese, Bengal, Hindi, Marathi, Gujarati, Panjabi, Urdu, Malayalam, Odisha, Konkani, Maithili, Kannada, Manipuri, Nepali, Tamil and Telugu a recognition rate of 70 % is obtained. At the same time the EER of individual languages are also given.

1. INTRODUCTION

In today's world Man Machine Communication is one of the emerging technologies in the field of digital communication. In this speech based communication is the major attraction because speech is most natural means of communication and therefore we are more comfortable with it and also because this is the fastest mode in this domain of communication. Due to several real-life applications of automatic Spoken Language Identification (ASLID) like speech to speech translation systems, information retrieval from multilingual audio databases and multilingual speech recognition systems, it has become an active research area in present time. In a country like India where the literacy rate is low, the need of development of Man Machine Communication interfaces like Automatic Speech Recognition (ASR), Text-to-speech synthesis (TTS) etc., are of great importance as it empowers the common people to access the huge bank of knowledge for the improvement of their living standards. The another major field of application of ASLID, is in security and intelligence service where the information of the intercepted voice in vernacular has to be quickly interpreted by somebody who knows that vernacular so that necessary action can be taken promptly. In the period of 1973 to 1980, series of reports were documented by Texas

Instrument[1,2,3,4]. Some important review of ASLID have been reported by Muthusamy. Y.K. *et al.*[5] in 1994 and Matejka[6] in 2004.The different characteristics of spoken language such as articulatory parameters[7], acoustic features[8] prosody[9], phonotactic[10], lexical knowledge[11] *etc.* are used as the cues of ASLID. These features of the spoken language can be divided into two broad levels: Spoken level and Word level[12]. The spoken level features for human speech contain acoustic, phonetic, phonotactic and prosodic information and can be obtained directly from the raw speech, whereas the word level features contain the morphology, syntax and grammar information. The another important cue of ASLID is the word detection. But the word boundary detection is the major problem in this field. Linear prediction coefficients (LPC), Mel-frequency Cepstral coefficients (MFCC), and Perceptual Linear Prediction (PLP) and Linear prediction Cepstral coefficients (LPCC) are used speech parameterization techniques in ASLID. Gauvin *et al.*[13] proposed an approach of producing a multitude of streams with the use of phoneme lattices. It was observed that the use of phoneme lattices significantly improves the performance of Parallel Phone recognition and Language Modeling (PPRLM) systems when compared to single best performing recognizer developed by the MIT-LL team[14]. The PPRLM-lattice sub-system offered in[14] attained a 30s / primary condition EER of 4.87% making it the single best performing recognizer. The PPRLM based LID system by using the acoustic diversi?cation as an alternative acoustic modeling technique was proposed by Sim and Li[15]. In this study the combination of the phonetic and acoustic diversi?cation (PAD) was used to achieve EERs of 4.71 and 8.61% on the 2003 and 2005 NIST LRE data sets respectively. In 2006 Support vector machines (SVM) were suggested by W. Zhang *et al.*[16]. In the same year E. Noor and H. Aronowitz[17] have shown that better classification was achieved by combining anchor model and SVM. Naresh M. *et al.*[18] worked on Split and merge EM algorithm for four Indian languages. In 2010 two hybrid feature extraction methods, namely Bark Frequency Cepstral Coefficients (BFCC) and Revised Perceptual Linear Prediction Coefficients (RPLP) were applied by P. Kumar *et al.*[19]. A novel attention base recurrent neural network (RNN) was used for LID. Two attention approach, soft and hard approach were investigated in this study[20]. It was reported that 8.2% relative EER reduction was obtained compared with LSTM based frame level system by the soft approach and 34.33% performance improvement is observed compared to the conventional i-Vector system. Hence, from the recent study, we can say that EER is a standard metric of ASLID.

Indian languages belong to several language families. Most of them share common set of phonemes and also follow similar grammatical structure. Indo Aryan languages are mainly influenced by Sanskrit[24]. At the same time the influence of Sanskrit is also present in the Dravidian language namely Telugu and Malayalam. Hence the discrimination among the sixteen languages mainly from Indo Aryan, Dravidian and Tibeto Burman is really a challenging task. Taking this challenge, this paper presents Relative Abundance Model (RAM) for identification of sixteen Indian spoken languages automatically. The model is based on four basic types of signal classes mentioned below.

free voice *e.g.*, vowels and glides (quasi-periodic),
obstructed voice *e.g.*, murmurs, laterals (quasi-periodic)
silent periods, occlusions as well as silences caused by breath pause (quiescent) and
noise segments, sibilants, frictions in affricates (quasi-random)

2. DATA BASE

The database contains recordings of free speech (30 minutes) from sixteen languages as referred to in the last section. It is collected from the website www.youtube.com. "Format Factory" software is used to

Table 1. Data base

Name	Bengali	Assamese	Maithili	Odiya	Gujarati	Nepali	Punjabi	Konkani	Marathi	Hindi	Urdu	Tamil	Telugu	Malayalam	Kannada	Manipuri
Abbreviation	ben	asa	mat	odi	guj	nep	pun	kon	mrt	hin	urd	tam	tel	mll	kan	man
Language Family	Indo Aryan Eastern	Indo Aryan Eastern	Indo Aryan Eastern	Indo Aryan Eastern	Indo Aryan Western	Indo Aryan Northern	Indo Aryan North western	Indo Aryan Southern	Indo Aryan Southern	Indo Aryan	Indo Aryan Southern	Dravidian	Dravidian	Dravidian	Dravidian	Tibeto-Burman
Duration in Minutes	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30	30

convert the video files (*.flv) to *.wav files. The digitization is done at a sampling rate of 22050 samples / sec, 16-bit mono. A preliminary noise cleaning was done using the Cool Edit 2000 software.

3. EXPERIMENTAL DETAILS

State phase analysis was used to segment the speech signals into the three basic types namely quasi periodic, quasi random and quiescent[21]. State phase provides close to 98.6% accuracy in segmentation. Minimum length of a segment is 15 milliseconds for quasi-random and quiescent signals and a single pitch period for quasi periodic signals. Then using an algorithm[23] which uses primarily the dip in amplitude associated with obstruction of oral tract, partial or full, we again classify the quasi-periodic signal in two sub class's namely free voice and obstructed voice segments. We have thus extracted from the signal four types of basic segments for ASLID.

3.1 Model RAM

A continuous speech signal of a spoken language is segmented using state-phase algorithm into these four basic types of using state phase algorithm[21] with 98.6% accuracy. The ratio of the total duration of all segments of a particular type of signal class (such as obstructed voice) to the total duration of a language is the relative abundance of that particular signal type (say obstructed voice) for that language. The Relative abundance model (RAM) essentially describes the nature of the distribution of the durations of the four basic signal class normalised with respect to the duration in a particular language. It will be seen that the relative proportion of the aforesaid signal types are different for different languages. Figure 1 describes that the amount of variation of four basic signal classes for all the sixteen languages.

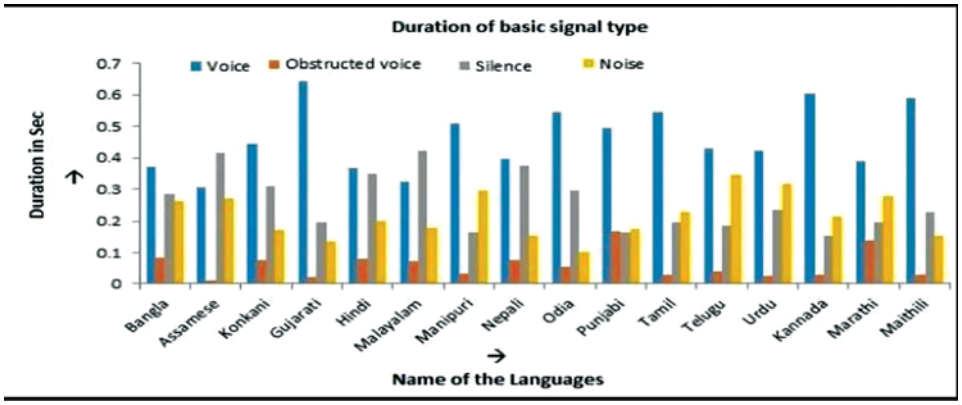

Fig. 1. RAM

3.2 ASLID Procedure

For testing the model for language identification we first divide the whole corpus of every language in 100 sets of 20 sec duration. Among the 100 sets the even numbered sets are used for training and odd numbered sets are used for testing purpose. In each set, average, standard deviation of four basic types of signals are calculated. Not only these but also the ratio of duration of four signal types with respect to the total duration of the set are calculated. These are treated as parameters for classification. Hence, for each signal type we are getting three parameters. The averages and standard deviations of these twelve (3×4) parameters are the representative a particular language. These parameters are used for recognition in the RAM model using a weighted Euclidian distance function.

$$D_k = \sum_{i=1}^{n} \frac{(x_i - \mu_{ki})^2}{\sigma^2_{ki}} \tag{1}$$

x_i = Mean of the unknown sample of i^{th} signal type

μ_{ki} = Mean of ith signal type for K^{th} language

σ_{ki} = Standard deviation of i^{th} signal type for k^{th} language

D_k = Euclidian Distance of the unknown sample from the k^{th} language.

3.3 EER

In this study two approaches are used. In one classification is done using minimum distance using Euclidian distance function weighted by inverse of variance and the other is the now popular ERR approach. Confusion matrices are used to represent the recognition rate. False rejection rate (FRR) and false acceptance rate (FAR) are two major components of ERR. The FRR is defined as the rate of a genuine language getting rejected from classification. FAR can be defined as the rate at which a wrong language is accepted as the required language. Normally FRR and FAR are calculated from the probability distribution curve[22] which is assumed to be Gaussian. But unfortunately the distribution of our parameters used in this study are not Gaussian. The normalized frequency distribution of voice parameter of Marathi and Urdu are given below.

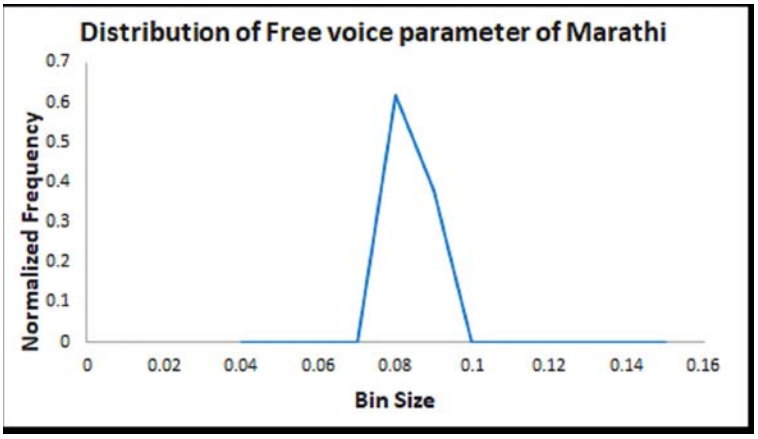

Fig. 2. Distribution of Marathi

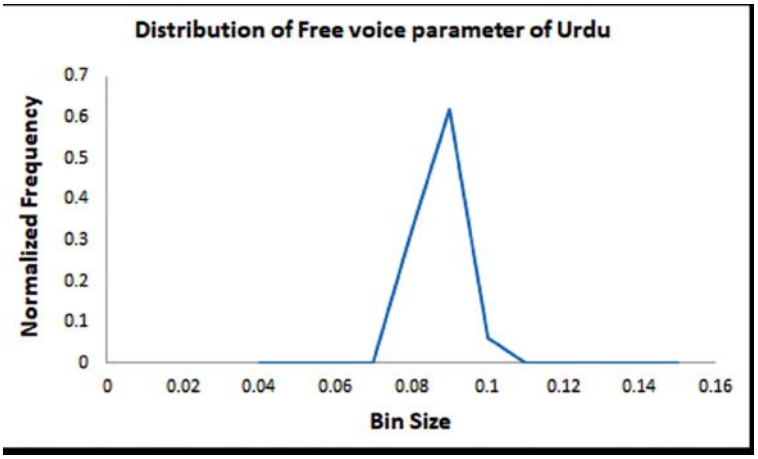

Fig. 3. Distribution of Urdu

Hence, it is difficult to fit a proper mathematical formula which gives the probability distribution of these skewed curves. So, instead of taking probability we have to consider the minimum distance to calculate the FRR and FAR. Here the weighted Euclidian distance classifier is used to classify unknown data to classes which minimize the distance between the data and the class in multi-feature space. Therefore, for the purpose of ERR a new approach is used. For this a threshold distance is considered, if the unknown sample data is less than the threshold distance then it will go to next process i.e. the classification process. In the classification process the unknown data can be correctly identified or it can be falsely included in a different language class i.e. FAR. On the other hand, if the distance is greater than the threshold value it will be rejected, i.e. FRR. The two circumstances can be more clearly visualized form the figure 4.

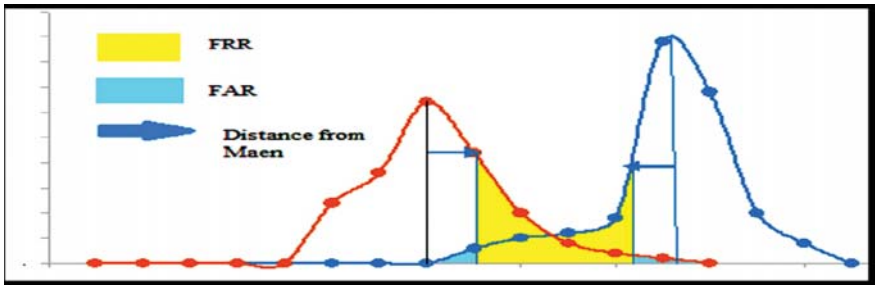

Fig. 4. Distribution of voice parameter for two language

From the figure 4, it is clear that FRR decreases with the larger value of threshold distance and the opposite is true for the FAR. Now from the definition of EER we can find out the EER from the intersection point of the two curves.

4. RESULT

A visual perusal of figure 1 indicates that the RAM models are somewhat different for different languages. Particularly Assamese, Punjabi and Malayalam stand out distinctively. One could see (Figure

Table 2. Confusion Matrix of RAM without rejection

Name	ben	asa	guj	kon	mll	Man	Nep	Odi	Pun	Tam	Tel	Urd	Hin	Kan	Mrt	Mat
ben	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
asa	0	36	2	0	0	0	0	0	0	0	0	0	4	0	0	8
guj	0	0	39	1	5	0	1	0	1	3	0	0	0	0	0	0
kon	0	0	14	15	10	0	8	3	0	0	0	0	0	0	0	0
mll	0	1	0	14	33	0	0	0	0	0	0	0	0	0	0	2
Man	0	6	0	0	0	44	0	0	0	0	0	0	0	0	0	0
Nep	0	0	0	0	0	0	44	0	0	0	0	0	0	0	0	6
Odi	0	12	0	0	0	0	0	38	0	0	0	0	0	0	0	0
Pun	0	0	7	13	0	0	0	0	30	0	0	0	0	0	0	0
Tam	0	0	13	0	0	0	0	0	0	28	0	0	0	0	0	9
Tel	0	3	0	0	0	0	0	0	0	0	47	0	0	0	0	0
Urd	0	0	6	7	0	0	0	3	0	2	0	16	0	0	10	6
Hin	0	0	0	2	0	0	0	0	0	0	0	0	48	0	0	0
Kan	0	0	0	0	0	0	0	3	0	1	0	0	0	46	0	0
Mrt	0	0	4	0	0	0	0	0	0	0	0	0	18	0	28	0
Mat	0	0	1	10	0	0	3	12	0	0	0	0	0	0	0	24
Fa	0	22	47	47	15	0	12	21	1	6	0	0	22	0	10	31
Error	0	0.0275	0.05875	0.05875	0.01875	0	0.015	0.02625	0.00125	0.0075	0	0	0.0275	0	0.0125	0.03875

Table 3. Confusion Matrix of RAM with rejection

Name	ben	asa	guj	kon	mll	mon	nep	odi	pun	tam	tel	urd	hin	kan	mrt	mat	rejection
ben	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
asa	0	27	6	0	0	1	0	0	0	0	0	0	10	0	0	6	0
guj	0	0	33	1	11	0	1	0	1	3	0	0	0	0	0	0	0
kon	0	0	11	20	7	0	9	3	0	0	0	0	0	0	0	0	0
mll	0	1	0	4	40	0	0	0	0	0	0	0	1	0	0	4	0
mon	0	8	0	0	0	38	0	0	0	0	0	0	4	0	0	0	0
nep	0	0	0	0	0	0	44	0	0	0	0	0	0	0	0	6	0
odi	0	7	0	0	0	0	0	41	0	0	0	0	2	0	0	0	0
pun	0	0	6	11	0	0	0	0	33	0	0	0	0	0	0	0	0
tam	0	0	11	0	0	0	0	0	0	28	0	0	0	0	0	11	0
tel	0	2	0	0	0	0	0	0	0	0	48	0	0	0	0	0	0
urd	0	0	4	5	0	0	0	2	0	3	0	21	0	0	8	7	0
hin	0	0	5	3	0	0	0	0	0	0	0	0	42	0	0	0	0
kan	0	0	0	0	0	0	0	2	0	6	0	0	0	42	0	0	0
mrt	0	0	4	0	0	0	0	0	0	0	0	0	9	0	37	0	0
mat	0	0	11	1	0	0	0	10	0	0	0	0	0	0	0	28	0
Fa	0	18	58	25	18	1	10	17	1	12	0	0	26	0	8	34	

1) Telugu, Tamil, Panjabi and Urdu to be quite similar. The confusion matrices for classification of fifty sets of each language are given in Table 2 & Table 3. The confusion matrix in Table 2 represents the identification rate without any rejection and in this case recognition rate is 70.75%. The last row of Table 2 indicates false acceptance (Fa). Here we want to mention that 16 languages with 50 test sets each are the input of the classifier. Hence, if we divide the Fa of particular language by total number of input sets (16 x 50) we will get an idea of error rate of that particular language. Table 3 describes the recognition rate at EER which forces the rejection of some samples and in this case identification rate is 70%. Both the confusion matrices are given for 6 min data set. In the first case we are getting only false acceptance (Table 2) but in the second case (Table 3) we are getting false acceptance as well as false rejection. Recognition rate is very good for Bengali, Manipuri, Nepali, Telugu, Hindi and Kannada. On the other hand, identification is very poor for Konkani, Urdu, Marathi due to which overall recognition rate falls. Table 3 describes the variation of EER for 20 sec, 40 sec 360 sec duration of the test sets.

It is observed from Table 4 that the EER cannot be found out for Telugu, Urdu, Kannada, Bengali and Manipuri for 360 sec and 40 sec durations respectively as the FAR is zero. But it is very much clear that EER decrease with the increase of signal duration.

Table 4. EER in %

Time	ben	asa	guj	kon	mll	man	nep	odi	pun	tam	tel	urd	hin	kan	mrt	mat
20 sec	5	5	12	6	2	0.2	2	3	1.2	6.4	1	2	3	1	2.5	10
40 sec	2.1	2.7	12	6	2	--	2	2	1.1	4.4	1	2	5	1	2	6.5
360 sec	--	3	7	4	3	0.2	1.2	2	0.1	1.8	--	--	3	--	1.2	3.8

5. DISCUSSION

In this paper we have discussed a model for ASLID which is mainly based on time domain parameters. The results of are quite noteworthy because of two reasons. First is the practically usable high recognition rate achieved for such a huge number of languages of the same language family. Secondly, we are getting quite impressive result without any arsenal of linguistic feature. Here we want to point out that though EER is a widely accepted standard in ASLID, its contribution towards the goodness of a classificatory analysis is not clear. We are not sure that the condition of equal error rate gives any extra merit to the classificatory analysis. Of course it gives a confidence measure to the recognition rate. Here we want to point out that Table 2 also gives the error which in turn gives the confidence of recognition for individual language. It may also be worthy to mention that the parameters related to the supra segmental features of a language generally used in human SLID has not been used here. This is an important area for our future work.

6. REFERENCES

[1] R. G. Leonard and G. R. Doddington, 1974. Automatic language identification. Technical Report RADC-TR-74-200, *Air Force Rome Air Development Center*.

[2] R. G. Leonard and G. R. Doddington, 1975. Automatic language identification. Technical Report RADC- TR-75-264, *Air Force Rome Air Development Center.*

[3] R. G. Leonard and G. R. Doddington, 1978. Automatic language discrimination. Technical Report RADC-TR-78-5, *Air Force Rome Air Development Center*.

[4] R. G. Leonard, 1980. Language recognition test and evaluation. Technical Report RADC-TR-80-83, *Air Force Rome Air Development Center.*

[5] Muthusamy Y.K., E. Barnard and R.A. Cole, 1994. "Reviewing Automatic Language Identification", in *IEEE Signal Processing Magazine,* pp. 33-41.

[6] Matejka P., 2003. Review of automatic language identification. *Proceedings of 10th Conference and Competition STUDENT EEICTT 2004,* Brno, CZ, **2,** 5, ISBN 80214-2635-7.

[7] Kirchhoff K., Parandekar S. and Bilmes J., 2002. Mixed-memory Markov models for Automatic Language Identification. *In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing,* pp. 761-764.

[8] Shivakumar P.G., Chakravarthula S.N. and Georgiou P., 2016. Multimodal Fusion of Multirate Acoustic, Prosodic, and Lexical Speaker Characteristics for Native Language Identification. *In: Proceedings of the Interspeech*, pp. 2408-2412.

[9] Adda-decker M., Antoine F., Vasilescu Loana, Lamel Lori, Vaissiere J. and Geoffrois Edouard, 2003. Liénard. Jean-sylvain., Phonetic knowledge, phonotactics and perceptual validation for automatic language identification. *In: Proceedings of the ICPhS*, pp. 747-750.

[10] Zissman M.A. and Berkling K.M., 2001. Automatic language identification. *Speech Communication*, **35**(1), 115-124

[11] Matrouf D., Adda-Decker M., L.F., Lamel L. and Gauvain J., 1998. Language identification incorporating lexical information. *In: Proceedings of the ICSLP*, pp. 181-184.

[12] Tong R., Ma B., Zhu D., Li H. and Chng E.S., 2006. Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification. *In: Proceedings of the IEEE International Conference on Acoustics Speed and Signal Processing Proceedings,* pp. 205-208.

[13] Gauvain J., Messaoudi A. and Schwenk H., 2004. Language recognition using phone lattices. *In: Proceedings of the INTERSPEECH*, pp. 25-28

[14] Shen W., Campbell W., Gleason T., Reynolds D. and Singer E., 2006. Experiments with Lattice-based PPRLM Language Identification. *In: Proceedings of the IEEE Odyssey*. The Speaker and Language Recognition Workshop.

[15] Sim Chai Khe and Li Haizhou, 2008. On acoustic diversification front-end for spoken language identification. *IEEE Trans Audio, Speech, Language Processing,* **16**(5), 1029-1037.

[16] Zhang W., Li B., Qu D. and Wang B., 2006. Automatic Language Identification using Support Vector Machines. *In: Proceedings of the 8th international Conference on Signal Processing.*

[17] Noor E. and Aronowitz H., 2006. Efficient Language Identification using Anchor Models and Support Vector Machines. *In: Proceedings of the IEEE Odyssey*. The Speaker and Language Recognition Workshop.

[18] Manwani N., Mitra S.K. and Joshi M.V., n.d. Spoken Language Identification for Indian Languages Using Split and Merge EM Algorithm. *Lecture Notes in Computer Science Pattern Recognition and Machine Intelligence,* pp. 463-468.

[19] Kumar P., Biswas A., Mishra A.N. and M Chandra, 2010. Spoken language identification using hybrid feature extraction methods. *Journal of Telecommunication,* **1,** 11-15.

[20] Geng W., Wang W., Zhao Y., Cai X. and Xu B., 2016. End-to-End Language Identification Using Attention-Based Recurrent Neural Networks. *In: Proceedings of the Interspeech,* pp. 2944-2948.

[21] Chowdhury Soumen, Datta K. A. and Chaudhuri B. B., 2000. Pitch detection algorithm using state phase analysis. *Journal of the Acoustical Society of India,* **28**(1-4), 247.

[22] Li K. P. and Porter J. E., 1988 Normalizations and selection of speech segments for speaker recognition scoring, *ICASSP,* p. 595.

[23] Saha A. and Datta K. A., 2011. A System for Analysis of Large Scale Speech Data for the Development of Rules for Speech Synthesis, *LNCS 7172,* publisher Springer, ISBN: 978-3-642-31979-2, e-ISBN: 978-3-642-31980-8, pp. 197-206.

[24] Vanishree V., 2011. Provision for Linguistic Diversity and Linguistic Minorities in India. Master's thesis. *Applied Linguistics, St. Mary's University College, Strawberry Hill, London.*

The Journal of Acoustical Society of India : Vol. 50, No. 2, 2023 (pp. 113-119)

Speaker recognition in Bengali language from nonlinear features

Uddalok Sarkar[1], Sayan Nag[1,5], Chirayata Bhattacharya[4], Shankha Sanyal[2,3*], Archi Banerjee[2,3], Ranjan Sengupta[2] and Dipak Ghosh[2]

[1]Department of Electrical Engineering

[2]Sir C.V. Raman Centre for Physics and Music

[3]Department of Physics

[4]Department of Electronics & Telecommunication Engineering, Jadavpur University

[5]Department of Medical Biophysics, University of Toront, Canada

e-mail: ssanyal.ling@jadavpuruniversity.in

[Received: 01-04-2022; Revised: 12-06-2022; Accepted: 29-06-2022]

ABSTRACT

At present Automatic Speaker Recognition system is a very important issue due to its diverse applications. Hence, it becomes absolutely necessary to obtain models that take into consideration the speaking style of a person, vocal tract information, timbral qualities of his voice and other congenital information regarding his voice. The study of Bengali speech recognition and speaker identification is scarce in the literature. Hence the need arises for involving Bengali subjects in modelling our speaker identification engine. In this work, we have extracted some acoustic features of speech using non linear multifractal analysis. The Multifractal Detrended Fluctuation Analysis reveals essentially the complexity associated with the speech signals taken. The source characteristics have been quantified with the help of different techniques like Correlation Matrix, skewness of MFDFA spectrum etc. The Results obtained from this study gives a good recognition rate for Bengali Speakers.

1. INTRODUCTION

Automatic Speaker recognition refers to identification of "who is speaking" on the basis of features of their respective speech waves. There are two major aspects of Automatic speaker recognition, speaker verification and speaker identification[1]. In Identification task the system has to identify the speaker (class) from a set of known voices (training dataset); whereas the speaker verification task involves accepting or rejecting identity claim of a speaker, where Speaker can be from open set or closed set. ASR can also be classified into two type: text dependent Automatic Speaker Recognition, text independent Automatic Speaker Recognition. In text dependent ASR the same phrase or key word is used for both enrollment and verification. Hence, the system knows the phrase to be uttered by the speaker beforehand. This type of system is much accurate. But text independent ASR is much robust and flexible but hard to design and less accurate, as the utterance of the speakers is different in enrollment and verification[2].

Most of the previous studies on Automatic Speaker recognition uses statistical characteristics related to temporal, spectral and cepstral features of each frame of the whole speech wave. Some temporal features for speaker identification are Zero crossing rate, Signal Energy, Maximum Amplitude and some useful spectral features are pitch contour, spectral Centroid, spectral flux, Perceptive Linear Prediction (PLP). Mel frequency Cepstral Coefficients is most frequently used cepstral feature in the field of Automatic Speaker Recognition[3]. Hence, most of the popular features that are used for speaker identification are primarily based on temporal to frequency domain transformation of speech wave. All these frequency domain techniques primarily use Fourier Transformation for transforming the signal into frequency domain. This method is strongly questioned for nonstationary aspect of signal. Numerous high frequency harmonics are left unattended in Fourier spectral analysis[4].

Non-linear dynamical modeling for source clearly indicates the relevance of non-deterministic / chaotic approaches in understanding the speech signals[5-7]. In this context fractal analysis of the speech signal which reveals the geometry embedded in signal assumes significance. Fractal analysis of audio signals was first performed by Voss and Clarke[8], who analyzed amplitude spectra of audio signals to find out a characteristic frequency f c , which separates white noise (which is a measure of flatness of the power spectrum) at frequencies much lower than f c from very correlated behavior ($1/f^2$) at frequencies much higher than f c . Speech data is essentially a quantitative record of variations of a particular quality over a period of time. However, it is well-established experience that naturally evolving geometries and phenomena are rarely characterized by a single scaling ratio; different parts of a system may be scaling differently. That is, the clustering pattern is not uniform over the whole system. Such a system is better characterized as 'multifractal'[9]. A multifractal can be loosely thought of as an interwoven set constructed from subsets with different local fractal dimensions. Real world systems are mostly multifractal in nature. Speech too, has nonuniform property in its movement[10,11]. In a number of recent studies[11,12], Multifractal Detrended Fluctuation Analysis (MFDFA)[9] have been applied to extract specific features of different music clips.

In this study, for the first time, we have applied MFDFA on the Bengali speech corpus of 5 participants to extract specific features which can help in development of an efficient speaker recognition algorithm in respect to Bengali speech. The multifractal spectral width generated from the self-similar speech signals have been used as a parameter to extract robust features from the different speech signals recorded from the speakers. An attempt has been made here to model a Multifractal Detrended Fluctuation Analysis based text independent Automatic Speaker recognition system. The study reveals new and interesting information which can pave the way for a number of new avenues in the domain of nonlinear approach to automatic speaker recognition. The same can be further expanded in regard to different languages.

2. EXPERIMENTAL DETAILS

Five (5) paragraphs selected from widely popular Bengali texts by eminent novelists were taken for analysis. The recordings were done from 5 (five) speakers who were asked to read the text normally without any emotional content being imbibed in them. The signals are digitized at the rate of 22050 samples / sec 16 bit format. The readings of each script have been kept within time duration of more or less 2 minutes and similar phrases (which were about 10 seconds duration) have been extracted then after for the ease of analysis. The fractal analysis of different segments has been carried out separately to get the necessary multifractal measures.

3. METHOD OF ANALYSIS

Method of multifractal analysis of sound signals : The time series data obtained from the sound signals are analyzed using MATLAB[13] and for each step an equivalent mathematical representation is given which is taken from the prescription of Kantelhardt *et al.*[9].

The complete procedure is divided into the following steps:

Step 1 : Converting the noise like structure of the signal into a random walk like signal. It can be represented as:

$$Y(i) = \sum (x_k - \bar{x}) \tag{1}$$

Where $\bar{x}$ is the mean value of the signal.

Step 2 : The local RMS variation for any sample size s is the function $F(s,v)$. This function can be written as follows:

$$F^2(s,v) = \frac{1}{s}\sum_{i=1}^{s}\{Y\ [(v-1)s+i] - y_v(i)\}^2 \tag{2}$$

Step 3 : The q-order overall RMS variation for various scale sizes can be obtained by the use of following equation

$$F_q(s) = \left\{\frac{1}{Ns}\sum_{v=1}^{Ns}[F^2(s,v)]^{\frac{q}{2}}\right\}^{\left(\frac{1}{q}\right)} \tag{3}$$

Step 4 : The scaling behaviour of the fluctuation function is obtained by drawing the log-log plot of $F_q(s)$ vs. s for each value of q.

$$F_q(s) \sim s^{h\,(q)} \tag{4}$$

The $h(q)$ is called the generalized Hurst exponent. The Hurst exponent is measure of self-similarity and correlation properties of time series produced by fractal. The presence or absence of long range correlation can be determined using Hurst exponent. A monofractal time series is characterized by unique $h(q)$ for all values of q.

The generalized Hurst exponent $h(q)$ of MFDFA is related to the classical scaling exponent $\tau(q)$ by the relation

$$\tau(q) = qh(q) - 1 \tag{5}$$

A monofractal series with long range correlation is characterized by linearly dependent q order exponent $\tau(q)$ with a single Hurst exponent H. Multifractal signal on the other hand, possess multiple Hurst exponent and in this case, $\tau(q)$ depends non-linearly on q[9].

The singularity spectrum $f(\alpha)$ is related to $h(q)$ by

$$a = h(q) + qh'(q) \tag{6}$$

$$f(a) = q[a - h(q)] + 1 \tag{7}$$

Where α denoting the singularity strength and $f(\alpha)$, the dimension of subset series that is characterized by α. The width of the multifractal spectrum essentially denotes the range of exponents. The spectra can be characterized quantitatively by fitting a quadratic function with the help of least square method[9] in the neighborhood of maximum α_0,

$$f(\alpha) = A(\alpha - \alpha_0)^2 + B(\alpha - \alpha_0) + C \tag{8}$$

Here C is an additive constant $C = f(\alpha_0)$ = land B is a measure of asymmetry of the spectrum. So obviously it is zero for a perfectly symmetric spectrum. We can obtain the width of the spectrum very easily by extrapolating the fitted quadratic curve to zero.

Width W is defined as,

$$W = \alpha_1 - \alpha_2 \text{ with } f(\alpha_1) = f(\alpha_2) = 0 \tag{9}$$

The width of the spectrum gives a measure of the multifractality of the spectrum. Greater is the value of the width W greater will be the multifractality of the spectrum. For a monofractal time series, the width will be zero as $h(q)$ is independent of q. The spectral width has been considered as a parameter to evaluate

how the features of speech of a particular speaker varies from another. Different features have been extracted using the multifractal curve as the input parameter.

Statistical Features of MFDFA Spectrum : To extract statistical features from the MFDFA spectrum we had to normalize the singularity spectrum $f(\alpha)$ by,-

$$f_n(\alpha) = \frac{f(\alpha)}{\sum_\alpha f(\alpha)} \tag{10}$$

Then we treated MFDFA spectrum as with normalized singularity spectrum as a probability distribution. We have taken two statistical features of this distribution for classification purpose. These are:

1. Feature-1 : Difference between median and mode- The mathematical expression we used to determine this value is,-

$$\text{Feature 1} = \text{median}\,(\alpha) - \underset{\alpha}{\text{argmax}}\, f_n(\alpha) \tag{11}$$

2. Feature-2 : Skewness - Mathematical Expression used for Skewness is,-

$$\text{Feature 2} = \text{E}\left(\frac{\alpha-\mu}{\sigma}\right) \tag{12}$$

Where, normalized $f_n(\alpha)$ is treated as pmf of discrete random variable α.

3. RESULTS AND DISCUSSION

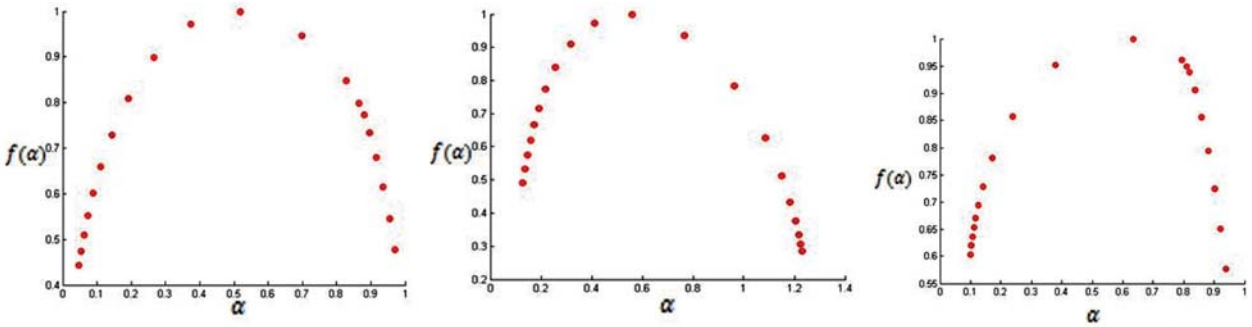

Fig. 1(a) Fig. 1(b) Fig. 1(c)

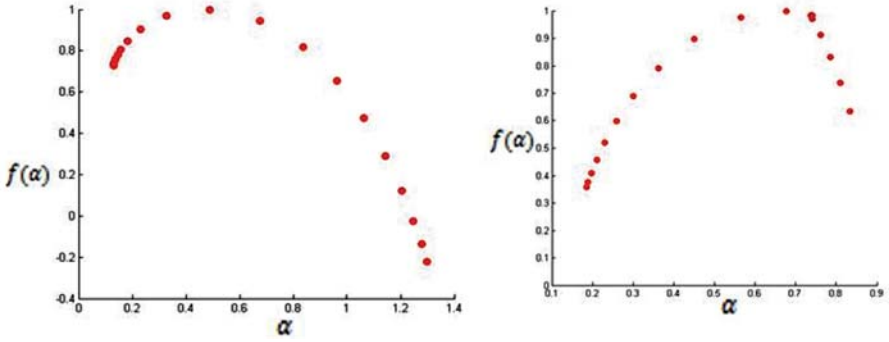

Fig. 1(d) Fig. 1(e)

Fig. 1. MFDFA spectrum instances: (a) Speaker 1 (b) Speaker 2 (c) Speaker 3 (d) Speaker 4 (e) Speaker 5

The following figure (Fig. 1) is a representative display of the multifractal spectrum obtained for five different speakers

From Fig. 1 We have an intuition of how the MFDFA spectrum varies from speaker to speaker. To emphasis more on this fact we tried to extract cross correlation coefficients between MFDFA spectrums of different speech waves. To obtain these cross correlation coefficients we followed protocol of Fig. 2.

In our Experiment we had 5 subject speakers and 20 Speech waves of each Speaker. Hence we came up with 100 total speech waves and hence a Cross Correlation Matrix of size 100 x 100. In Table 1 the Speech Data and its corresponding speaker information are tabulated. In Fig. 3 the Correlation matrix is shown which is of size 100 x 100.

Table 1. Speech data index versus speaker Identity.

Speech Index in Correlation Matrix	Speaker Identity
0 to 20	Speaker1
21 to 40	Speaker2
41 to 60	Speaker3
61 to 80	Speaker4
81 to 100	Speaker5

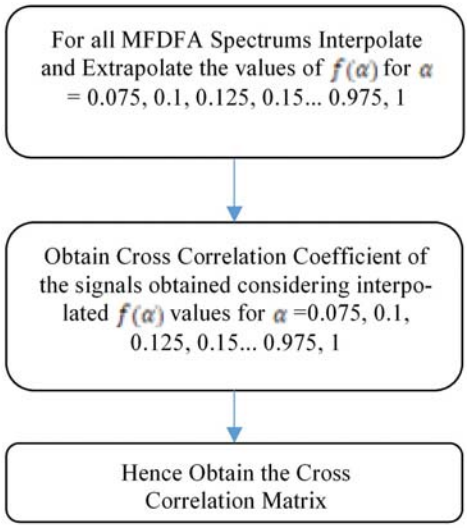

Fig. 2. Protocol for determining cross correlation coefficient of MFDFA spectrum.

In Fig. 3 the Correlation matrix clearly shows that the Cross Correlation coefficients between MFDFA spectrums of speech waves in case of same speakers are very high but for different speakers these coefficients are very low. For instance, the Correlation Matrix shows that MFDFA spectrum for speaker4 and speaker3 are very much dissimilar (having correlation coefficient in the range -0.2 - 0.2), almost the same response can be seen in case of speaker4 and speaker5. But, in case of speaker2 and speaker4, we find that the values of cross-correlation coefficient are higher than the previous case, *i.e.* in the range 0.4-0.6. Hence we can safely assume that there exist certain similarities in the spectra of speaker 2 and 4, as well as in

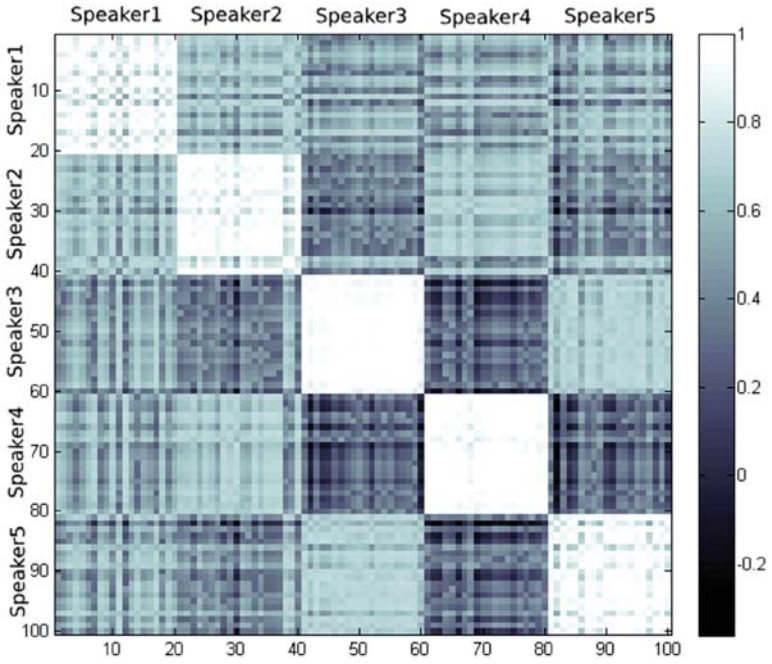

Fig. 3. Correlation Matrix.

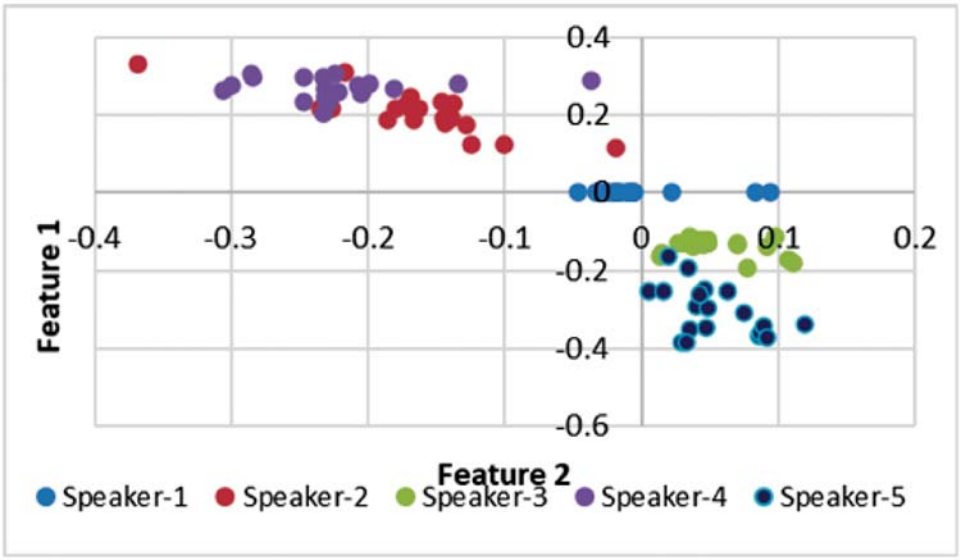

Fig. 4. Position of Speakers on the 2-dimensional feature space.

Speaker3 and 5. The feature plot of Fig. 4 gives a complete validation in support to the inferences drawn from this correlation matrix.

Fig. 4 shows a feature plot where the 5 speakers have been shown as a scatter diagram in 2-D feature space. In this plot, the X-direction is Feature 1 and Y-direction is Feature 2.

We can infer from Fig. 4 that for speaker2 and speaker4 the MFDFA spectrum is skewed in same direction with higher Skewness measure in case of speaker4. And for speaker2 and speaker4 $\underset{\alpha}{\text{argmax}}\, f_n(\alpha) < \text{median}\,(\alpha)$. Similarly, for speaker3 and speaker5 the spectrum is skewed in same direction and speaker4 $\underset{\alpha}{\text{argmax}}\, f_n(\alpha) < \text{median}\,(\alpha)$. For speaker1 the Skewness measure is almost zero and $\underset{\alpha}{\text{argmax}}\, f_n(\alpha) < \text{median}\,(\alpha)$. We have seen the very same instances in Fig. 1(a), Fig. 1(b), Fig. 1(c), Fig. 1(d) and Fig. 1(f).

After extracting these features for the whole speech database we fitted it with support vector machine. Using RBF kernel for SVM and with a holdout ratio of ¼ we got the overall accuracy of 96% for speaker identification.

In Table 2 the confusion matrix between classifier prediction and true classes is shown. Prediction accuracy of speaker2 is seen to be 80%. one of the speaker 2 instances is seen to be predicted as speaker 4

Table 2. Confusion Matrix of speaker identification.

		Classifier Prediction					
Actual Value		Speaker1	Speaker2	Speaker3	Speaker4	Speaker5	Total
	Speaker1	5	0	0	0	0	100%
	Speaker2	0	4	0	0	0	100%
	Speaker3	0	0	5	0	0	100%
	Speaker4	0	1	0	5	0	83.3%
	Speaker5	0	0	0	0	5	100%
	Total	100%	80%	100%	100%	100%	96%

instance. In most of the cases, it is seen that the prediction accuracy is 100% using this technique, which makes it a effective and robust method for speaker recognition in the nonlinear scenario.

4. CONCLUSION

Speaker identification through different acoustic features has been the domain of extensive research in the field of speech signal processing. Till date, most of the studies dealt with linear features which characterize the speaker. In this work, for the first time robust nonlinear techniques have been applied to characterize the speech samples corresponding to Bengali language from 5 speakers. Several features like correlation matrix, skewness have been extracted from the nonlinear multifractal spectrum to classify the speakers. The findings have been put through SVM classifier which resulted in 96% classification accuracy in the features applied. This pilot study has immense potential to be applied in different languages to provide a robust algorithm for effective speaker recognition.

5. REFERENCES

[1] Zhu Qifeng and Abeer Alwan, 2003. "Nonlinear feature extraction for robust speech recognition in stationary and non-stationary noise." *Computer speech & language,* **17.4,** 381-402.

[2] Chen Wen-Shiung and Jr-Feng Huang, 2009. "Speaker recognition using spectral dimension features." Computing in the Global Information Technology, 2009. ICCGI'09. *Fourth International Multi-Conference on IEEE.*

[3] Md Afzal Hossan. Thesis on "Automatic Speaker Recognition Dynamic Feature Identification and Classification using Distributed Discrete Cosine Transform Based Mel Frequency Cepstral Coefficients and Fuzzy Vector Quantization", RMIT

[4] Bhaduri S. and D. Ghosh, 2016. "Non-invasive detection of alzheimer's disease-multifractality of emotional speech". *J. Neurol. Neurosci.*

[5] Behrman A., 1999. Global and local dimensions of vocal dynamics. *Journal-Acoustical Society of America,* **105,** 432-443.

[6] Kumar A. and Mullick S. K., 1996. Nonlinear dynamical analysis of speech. *The Journal of the Acoustical Society of America,* **100**(1), 615-629.

[7] Sengupta R., Dey N., Nag D. and Datta A. K., 2001. Comparative study of fractal behavior in quasi-random and quasi-periodic speech wave map. *Fractals,* **9**(04), 403-414.

[8] Voss R. F. and J. Clarke, 1975. "1/f noise in speech and music." *Nature.* **258,** 317-318.

[9] Kantelhardt J. W., Zschiegner S. A., Koscielny-Bunde E., Havlin S., Bunde A. and Stanley H. E., 2002. Multi-fractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications,* **316**(1-4), 87-114.

[10] Lopes R. and Betrouni N., 2009. Fractal and multifractal analysis: a review. *Medical image analysis,* **13**(4), 634-649.

[11] Sanyal S., Banerjee A., Patranabis A., Banerjee K., Sengupta R. and Ghosh D., 2016. A study on Improvisation in a Musical performance using Multifractal Detrended Cross Correlation Analysis. *Physica A: Statistical Mechanics and its Applications,* **462,** 67-83.

[12] Banerjee A., Sanyal S., Mukherjee S., Guhathakurata T., Sengupta R. and Ghosh D., 2016. How Do the Singing Styles vary over Generations in different Gharanas of Hindustani Classical Music A Comparative Non Linear Study. arXiv preprint arXiv:1604.02250.

[13] Ihlen E. A. F. E., 2012. Introduction to multifractal detrended fluctuation analysis in Matlab. *Frontiers in physiology,* **3,** 141.

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)