ISSN 0973-3302

THE JOURNAL OF ACOUSTICAL SOCIETY OF INDIA

Volume 47 Number 2-3 April-September 2020

ACOUSTICAL SOCIETY OF INDIA
1971
॥ नादाधीनमतोजगत् ॥

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

The Journal of Acoustical Society of India

Volume 46 Number 2-3 April-September 2020 pp 37-121

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.

B. CHAKRABORTY
Chief Editor

MAHAVIR SINGH
Managing Editor

A R MOHANTY
Associate Scientific Editor

Yudhishter Kumar Yadav
Devraj Singh
Kirti Soni
Assistant Editors

The Journal of Acoustical Society of India

A quarterly publication of the Acoustical Society of India

Volume 47, Number 2-3, April-September 2020

ARTICLES

INFORMATION

GUEST EDITOR

Editorial on the Special Issue on "Signal Processing in Acoustics"

Among the many facets of research in acoustics, the use of signal processing has been key in advancing both theory and applications of acoustics. The Journal of Acoustical Society of America has published many papers over the years that reflect the depth and diversity of signal processing techniques applied to acoustics. This special issue of the journal on the topic of "Signal Processing in Acoustics" is an endeavour to showcase the advanced research work being done in this field through a sample of papers from prominent researchers in the country and their research groups. A total of 8 papers are presented that provide insight into different signal processing techniques as applied to several areas of acoustics as follows: Audio Engineering (1), Animal Bioacoustics (1), Biomedical Acoustics (1), Signal Processing for Acoustic Source Detection (1), Speech Communication (3), and Underwater Acoustic Communication (1).

The paper on "Spatial Multi-Zone Sound Field Reproduction Using Differential Phase Constraint" presents a signal processing method to accurately reproduce the sound field in multiple zones for creating an immersive audio experience using an array of several loudspeakers. The synthesized sound field is compared with previous methods from the research literature. The next paper titled "Multi-component Oscillatory Model based Classification of Heart Sounds" describes a biomedical application of signal processing in which a multi-component oscillatory model is used for modeling the non-stationary heart sounds from phonocardiogram. Features are derived from the models that are then used for classifying the different heart sounds as normal or abnormal. The performance of different classifiers is compared amongst each other for the derived features on a standard database to draw inferences.

The next paper titled "An Overview of Techniques Developed for Bio-Sonar Characterization and Census of Ganges River Dolphin, India's National Aquatic Animal," presents the research work done to develop signal processing methods and hydrophone array systems for long-term passive acoustic monitoring of Ganges river dolphins using their echolocation acoustic signals. The paper also describes the development of an integrated visual and acoustic system to detect and find range of dolphins in the Ganges river. It describes how the system can be used for acoustics based census purposes that can be an alternative to a conventional human observer based method.

In the paper "Analysis of Source Signal and Vocal Tract for Detection of Out-of-breath Speech," a study of out-of-breath speech that results from physical activity is reported. The effect of this stressed condition of speech production on the excitation source signal and vocal tract resonance components of speech production is studied. Inferences are drawn from speech classification results based on features derived from the two speech components. The following paper titled "Significance of Excitation Source Information from Speech," describes the signal processing technique for extracting excitation source information from speech. It reviews how this is useful in various applications such as enhancement of noisy speech, speaker verification, spoof detection, natural quality speech synthesis, and detection of speech disorders. The next paper titled "An Interactive MATLAB based GUI for Speech Processing and Stress Detection" describes a training toolkit developed for assisting learners of speech processing fundamentals and applications. It allows a student interactivity with the software and allows processing of speech signals in real-time. It has three modules viz. learning module, signal processing module and stress detection module where the first module allows for review of theoretical concepts, the second module can be used to learn basic and advanced speech signal processing methods and the third module is an application learning environment of speech processing.

The paper "Multivariate Quadratic Regression based Direction Estimation of an Acoustic Source" presents a machine learning based technique for direction of arrival estimation of an acoustic source using a uniform linear array. A regression model is trained using correlation features and is then used to estimate the direction of arrival. The performance is compared with a conventional delay and sum beam former to show the potential advantage of such a signal processing approach to the problem of direction of arrival.

The following paper titled "Underwater Communications: An Open Challenge," describes the several challenges in underwater acoustic communications and how advanced signal processing algorithms are pushing the boundaries of performance. Several state-of-the-art communication modems are listed with their specifications. The paper describes channel distortion models and accurate noise models that are used to improve communication efficiency. The paper also gives results of experiments performed on underwater communications with a designed modem.

In conclusion, I would like to thank Editor-in-Chief Dr. Biswajit Chakraborty and the Editorial Board of JASI to give me the privilege to serve as Editor for this special issue. I also thank the reviewers who provided valuable and timely feedback on the manuscripts. Much as it was a pleasure for me to edit this special issue, I hope the reader will find the collection of papers to be interesting and insightful into the multi-faceted uses of signal processing in acoustics.

— Arun Kumar
Guest Editor - JASI
&
Professor and Head
Centre for Applied Research in Electronics
Indian Institute of Technology Delhi
New Delhi - 110016, India

The Journal of Acoustical Society of India : Vol. 47, No. 2-3, 2020 (pp. 37-48)

Spatial multi-zone sound field reproduction using differential phase constraint

Ajay Dagar and Rajesh M Hegde
Indian Institute of Technology Kanpur, India
e-mail: adagar@iitk.ac.in, rhegde@iitk.ac.in

[Received: 27-03-2020; Revised: 12-05-2020; Accepted: 12-05-2020]

ABSTRACT

Spatial multi-zone sound field reproduction obtained by an array of loudspeakers are capable of providing a personalized audio experience. The immersiveness of these spatial sounds depends on the accuracy of the method to reproduce the sound field in the zones of interest from the desired directions. The existing multi-zone methods either rely on improving the acoustic contrast between the zones or minimizes the error in the reproduced sound pressures. In this work, a differential phase constraint approach is developed that improves the directivity of the reproduced sound field while preserving the acoustic contrast and the reproduction error. An optimization problem is formulated that jointly minimizes the difference in the pressures and the phases in the zones of interest. The performance of the proposed method is evaluated in a simulated environment on the basis of obtained acoustic contrast, reproduction error and the directivity of the reproduced sound fields. Additionally, the statistical analysis is carried out to support the proposed framework.

1. INTRODUCTION

Multi-zone sound field reproduction with a set of loudspeaker arrays aims to deliver personalized sound zones to individual listeners in vicinity of multiple listeners in a shared environment. Private listening within a car environment, individualized listening in a home environment, personalized sound during video conferencing, multi- zone sound experience at various public places such as music stores and exhibition centres are the potential applications of multi-zone environment[1]. To meet the expectation of this increasing demand, extensive research has been carried-out over the last two decades. The initial methods[2]-[3] developed for the multi-zone environment, were based on controlling the acoustic energy in the desired zones of interest with an objective to maximize the acoustic contrast (AC) between the bright zones and the dark zones. Though, a significant improvement in acoustic contrast against 11dB of standard acceptable limit was achieved in[4], but the method lacks in accurately reproducing the sound fields in the bright zones within a car environment.

Later in[5]-[6], the multi-zone reproduction problem is addressed as a pressure matching (PM) approach to outperform AC methods in terms of improved reproduction error. Authors in[5], were able to find the optimal loudspeaker weights by solving the error minimization problem using least squares (LS). A modified-LS approach based on Tikhonov regularization and energy constraint on driving signals was also introduced in[7], [6]. To minimize the error and simultaneously improve the acoustic contrast, a joint of ACC and PM approach is used in[8]. Further, the robustness of these methods to uncertainties in acoustic environment is investigated in[9] using ACC and PM. In addition, some multi-zone sound field reproduction approaches are also developed in spatial and cylindrical harmonic domains considering mode matching between the local and global region of reproduction[10]-[11]. The drawback here is that, with increase in the reproduction area, the required number of loudspeakers also increases which may not be suitable for multi-zone sound field reproduction. In all of the above multi-zone approaches, authors were able to successfully reproduce the personal sound zone with a desired sound field and reduced reproduction error. Also, the corresponding loudspeaker gains were obtained by solving the formulated multi-zone problems. But it should be noted that none of authors tried to improve and preserve the directivity, referred as the spatial quality, of the reproduced sound fields. Few research articles addressing this aspect can only be found in literature[12]-[13]. Though in[13], with a planetary control method authors tried to preserve the structure of target fields but lacks in maintaining the accuracy in directional component in the desired direction. Thus, there is a scope of further improvement and a need of multi-zone methods that takes into account the spatial aspects of the reproduced sound fields.

In this paper, a spatial multi-zone reproduction problem is formulated using a differential phase constraint to enhance the spatial quality of the reproduced sound fields. The main contribution of this work is to formulate a framework that jointly minimizes the difference in the sound pressures and the phase in the zones of interest. Since the directivity pattern, is mainly present in the phase component of the signal, preserving the same enhances the immersiveness of the reproduced sound fields. Therefore, the present work is mainly focused on improving the directional component in the zones of interest without significant variation in the performance of the acoustic contrast. With the introduction of differential phase constraint, the proposed multi-zone problem becomes non-convex in nature, the optimization problem is re-formulated by introducing a phase constraint with certain relaxation. The performance of the proposed method is evaluated on the basis of obtained acoustic contrast, reproduction error and the directivity of the reproduced sound fields.

The remainder of paper is organised as follows: In Section-II, the spatial multi-zone sound field reproduction is discussed followed by modelling of target sound fields and problem formulation using proposed approach. Later in Section-III, the performance evaluation of the proposed method is carried out in terms of error analysis and an in-depth directivity analysis. Finally, the conclusion and future work are discussed in Section-IV.

2. MULTI-ZONE SOUND FIELD REPRODUCTION USING DIFFERENTIAL PHASE CONSTRAINT

Consider a multi-zone reproduction environment with Q listening zones where spatial sound is to be reproduced using an array of L loudspeakers, as shown in Figure 1. In this environment, the loudspeakers are assumed to be placed evenly over the surface of a sphere located at a point $r_l = (r_l, \Theta_l)$ with respect to the center of environment, considered as the origin. Also at each listening spot, consider a spherical region of radius R_q having polar coordinates (r_q, Θ_q) with respect to origin and I_q sample points evenly distributed over its surface. Here, $\Theta = (\theta, \phi)$ is considered a directional variable with θ and ϕ as elevation and azimuth angles in spherical domain, respectively.

Now, the complex pressure at i^{th} sample point $r_q^i = (r_q^i, \Theta_q^i)$ on the surface of q^{th} zone, due to the loudspeaker setup at a frequency w, is given by

$$p(k, r_q^i) = \sum_{l=1}^{L} a_l(k) \times g_q^i(k, r_l, r_q^i) \tag{1}$$

where, $a_l(k)$ is the driving signal for the l^{th} loudspeaker and $g_q^i(k, r_l, r_q^i)$ is the complex acoustic channel, between the i^{th} sample point in q^{th} zone and the lth loudspeaker. The complex channel $g_q^i(k, r_l, r_q^i)$ can be modelled on the basis of the type of the source considered at the loudspeaker position, *i.e.* either a plane wave source or a point source with spherical wave front, as given below:

$$g_q^i(k,\ r_l,\ r_q^i) = \begin{cases} e^{jkr_q^i \cos\Theta_l}, \text{ for a plane wave} \\ \dfrac{e^{jk|r_l - r_q^i|_2}}{\left\| r_l - r_q^i \right\|_2}, \text{ for a spherical wave} \end{cases} \tag{2}$$

where, k is the wave number corresponding to angular frequency ω and cos Θ_l corresponds to the unit projection of plane wave in microphone direction, given by[14]

$$\cos\Theta_l \ = \ \cos\ \theta_l\ \cos\ \theta_q^i + \ \cos(\phi_l\ -\phi_q^i)\sin\theta_l \sin\ \theta_q^i \tag{3}$$

In equation-(2), the channel gain for a plane wave source is defined by obtaining the far-field approximation of a point source as given in[14]. Now, irrespective of the state of a zone being a bright or a dark, the above equations are valid for all the zones of interest. Considering I_q sample points over the surface of q^{th} zone, the reproduced sound field, in matrix form, can be expressed as

$$P_q(k) \ = \ G_q(k)a(k) \tag{4}$$

where, $a(k) \in C^{L\times 1}$ is a vector corresponding to the loudspeaker gains and $G_q(k) \in C^{I_q\times L}$ is acoustic channel matrix with elements corresponding to the free field Green's function defined in equation-(2).

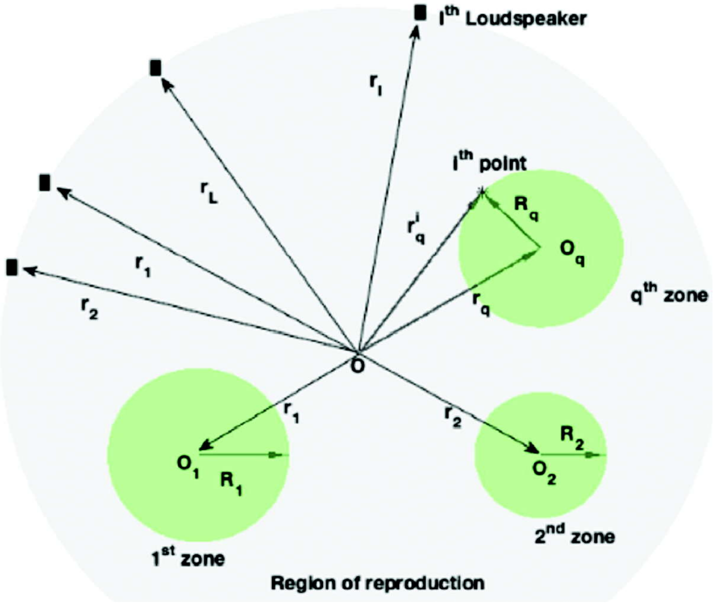


Fig. 1. Figure illustrating a multi-zone environment with Q zones of interest placed within the region of reproduction. The interior region enclosed by L loudspeakers placed over a sphere is considered as the region of reproduction.

In a multi-zone environment at a given frequency ω, the acoustic contrast between a bright zone and a dark zone is defined as the ratio of the average of the reproduced sound pressures given by[15]

$$AC_d^b(k) = \frac{I_D}{I_B} \times \frac{P_B^H(k)P_B(k)}{P_B^H(k)P_D(k)} = \frac{I_D}{I_B} \times \frac{a^H(k)G_B^H(k)G_B(k)a(k)}{a^H(k)G_B^H(k)G_D(k)a(k)} \tag{5}$$

where, $p_B(k) \in C^{I_B\times 1}$ and $p_D(k) \in C^{I_D\times 1}$ are the reproduced sound pressures in the bright and the dark zone, respectively. In a multi-zone environment where multiple bright zones and multiple dark zones are considered, the overall acoustic contrast $AC_{overall}$ is given as

$$AC_{overall}(k) \;=\; \frac{1}{(N_B \times N_D)}\sum_{b=0}^{N_B}\sum_{d=0}^{N_D} AC_d^b(k) \tag{6}$$

where, N_B and N_D are the number of bright zones and dark zones such that $(N_B+N_D) = Q$. Similar to[16], the other parameter that should be taken into consideration in each bright zone is the normalised reproduction error, which is given as

$$Error(k) \;=\; \frac{[p_B^t(k) - p_B(k)]^H [p_B^t(k) - p_B(k)]}{[p_B^t(k)]^H [p_B^t(k)]} \tag{7}$$

where, $p_B^t(k)$ is the target sound field to be reproduced in a bright zone. Similar to $AC_{overall}(k)$, the overall reproduction error Error$_{overall}(k)$ can be obtained by averaging the individual reproduction error in each zone.

In addition, from the perceptual point of view, the directivity of the desired sound must be preserved in the reproduced sound fields. To quantify the directivity, the directivity index (DI) is widely used and is obtained from the directivity factor (DF), given by[17]

$$DF(k) = \frac{\left\| y(k,\Theta_n) \right\|_2^2}{\dfrac{1}{4\pi}\displaystyle\int_0^{2\pi}\int_0^{\pi} \left\| y(k,\Theta_n) \right\|_2^2 d\theta d\phi} \tag{8}$$

where, $y(k,\Theta_n)$ is the response of the spatial filter in the zone of reproduction in the direction Θ_n, defined as[14]

$$y(k,\Theta_n) = \int_0^{2\pi}\int_0^{\pi} w^*(k,\, \Theta_n) p(k,\, r_q) d\theta d\phi \tag{9}$$

$$y(k,\Theta_n) = \sum_{i=1}^{Iq} w_i{}^*(k,\, \Theta_n) p(k,\, r_q^i) = \sum_{i=1}^{Iq} w^*(k,\, \Theta_n) p_q(k) \tag{10}$$

where, $w^*(k, \Theta_n)$ is the weight vector with elements defind as $w_i{}^*(k, \Theta_n)$ = $e^{-\overset{\square}{k}_n . r_q^i}$ obtained using a plane wave decomposition. In above, k_n is the wave vector in the direction Θ_n. Now, the objective here is to find the loudspeaker gains $a(k)$ that can accurately reproduce the target sound fields in the bright zones and simultaneously maximizes the overall acoustic contrast $AC_{overall}(k)$ while maintaining the required directivity pattern as defined by the target sound fields.

2.1 Modelling of Target Sound Fields

A multi-zone sound environment has multiple bright and dark zones with high fidelity spatial sound in bright zones and complete silence in dark zones. In each bright zone, the target sound field can be modelled using multiple plane waves arriving from different directions. Similar to[18], uniformly sampling the unitary sphere into N angular locations $\hat{v}_n$, the target sound field at i^{th} point on the surface of a bright zone can be expressed as

$$p^t(k, r_B^i) = \sum_{n=1}^{N} \varphi(\hat{v}_n, k)\, e^{jk\hat{v}_n . r_B^i} \tag{11}$$

where, $\varphi(\hat{v}_n, k)$ is the density function at frequency $\omega = kc_0$ in the direction of unit vector $\hat{v}_n$. For simplicity, we focus on the case when $\phi(\hat{v}_n, k) = \delta(\hat{v}_n - \hat{v}_o)$ which corresponds to a single plane wave arriving from the direction $\Psi_o = (\theta_o, \phi_o)$. Now, the equation-(11) can be re-written as

$$p^t(k, r_B^i) = e^{jk\hat{v}_o . r_B^i} \tag{12}$$

Therefore, the complex target pressures $p_B^t(k) \in C^{I_B \times 1}$ for a bright zone can be expressed as

$$p_B^t(k) \;=\; [p^t(k, r_B^1) p^t(k, r_B^2) \ldots p^t(k, r_B^{I_B})]^T \tag{13}$$

Also, it is assumed that each dark zone is completely a silent zone with minimum flow energy across it. Therefore, we can assume that target pressures $P_D^t(k)$ in each dark zone is a vector of size $I_B \times 1$ with element equal to zero.

2.2 Problem Formulation using Differential Phase Constraint

The multi-zone reproduction problem can be formulated by directly minimizing the error between the target sound fields and reproduced sound fields. The conventional pressure matching using Least Squares (LS) approach can be expressed as

$$\begin{aligned} &\min_{a(k)} \quad \frac{1}{2}\sum_{q=1}^{Q} \left| p_q(k) - p_q^t(k) \right|_2^2 \\ &s.t. \;\; p_q(k) \;=\; G_q(k) a(k) \;\; \forall q \;=\; 1, \ldots, Q \\ &\qquad a^H(k) a(k) \;\leq\; \beta \end{aligned} \tag{14}$$

The energy constraint on *a(k)* is applied to ensure that the array effort is always below a threshold β. The above problem is convex in nature and can be solved as such using[19]. The solution to this problem results in minimum error sound reproduction.

Furthermore, the performance of this conventional approach in terms of the spatial sound quality can be improved by introducing an additional phase constraint that preserves the directivity pattern. Since, it is necessary to maintain the accuracy of reproduction in bright zones, the phase constraint can only be applied on the bright zones. Now, the optimization problem can be formulated as

$$\begin{aligned} &\min_{a(k)} \quad \frac{1}{2}\sum_{q=1}^{Q} \left| p_q(k) - p_q^t(k) \right|_2^2 \\ &s.t. \;\; \mathrm{p}_q(k) \;=\; G_q(k) a(k) \;\; \forall q \;=\; 1, \ldots, Q \\ &\tan^{-1}\left(\frac{p_q^{im}(k)}{p_q^{re}(k)} \right) - \tan^{-1}\left(\frac{p_q^{t,im}(k)}{p_q^{t,re}(k)} \right) \leq \varepsilon, \; \forall q \in N_B \\ &\qquad a^H(k) a(k) \;\leq\; \beta \end{aligned} \tag{15}$$

where, *p^re(k)* and *p^im(k)* corresponds to the real and imaginary part of the complex pressures *p(k)*. The phase constraint applied here minimizes the phase error at each sample point by keeping it below a threshold ε but at the same time the formulated framework becomes non-convex in nature. Thus, to retain the convexity in the above framework, the phase constraint can be replaced by its relaxed approximate. Using the property of $\tan^{-1}(.)$, *i.e.* $\tan^{-1}(x) - \tan^{-1}(y) = \tan^{-1}\left(\frac{x-y}{1+xy} \right)$, the differential phase constraint can be replaced by its equivalent affine constraint. Therefore, the proposed framework can be re-formulated as

$$\begin{aligned} &\min_{a(k)} \quad \frac{1}{2}\sum_{q=1}^{Q} \left| p_q(k) - p_q^t(k) \right|_2^2 \\ &s.t. \;\; \mathrm{p}_q(k) \;=\; G_q(k) a(k) \qquad \forall q \;=\; 1, \ldots, Q \\ &\quad \mathrm{p}_q^{im}(k) = \hat{p}_q^t(k) \times p_q^{re}(k), \qquad \forall q \in N_B \\ &\qquad a^H(k) a(k) \;\leq\; \beta \end{aligned} \tag{16}$$

where $\hat{p}_q^t(k) = \dfrac{p_q^{t,im}(k)}{p_q^{t,re}(k)}$, $\forall q \in N_B$, is constant quantity. Now, the above formulated framework is convex in nature and can be solved by the standard cvx toolkit[19].

(1) Significance of Differential Phase Constraint : The performance of any multi-zone sound reproduction depends on how accurately the given loudspeaker setup is able to reproduce the sound fields close to the desired sounds of interest. In spatial sound field reproduction, it is required that the methodology used must preserve the directional content in the reproduced sound fields. Since, the directional information completely depends on the phase of the reproduced sound pressure, an additional constraint on phase will definitely improve the directivity of the reproduced sound fields. Additionally, in multi-zone sound reproduction, it is more necessary to maintain the accuracy of reproduction in bright zones, therefore, the differential phase constraint can only be applied on the bright zones.

3. PERFORMANCE EVALUATION

In this section, the experimental setup alongwith experimental conditions considered to evaluate the performance of the proposed methodology are discussed first. Subsequently, the analysis of reconstructed sound fields and the obtained results, conducted to show the effectiveness of proposed method, are discussed.

3.1 Experimental Setup

The proposed framework evaluated in a multi-zone environment where $Q = 4$ zones of interest, *i.e.* $N_B = 2$ bright zones and $N_D = 2$ dark zones, are considered as shown in Figure 2. The zones of interest are assume to be spherical in shape having a radius of $R_q = 0.1m$, $\forall_q$. In Figure 2(a), the points with coordinates (±0.75, ±0.45 0) defines the position of each listening zone, *i.e.* the center of each zone respectively. In the same figure, an array of 512 loudspeakers are assumed to be placed over a sphere of radius $R_l = 2m$ in icosahedron pattern. Similarly, it is assumed that the surface of each zone of interest is sampled icosahedronally and the reproduced sound pressures are analysed using these sample positions,

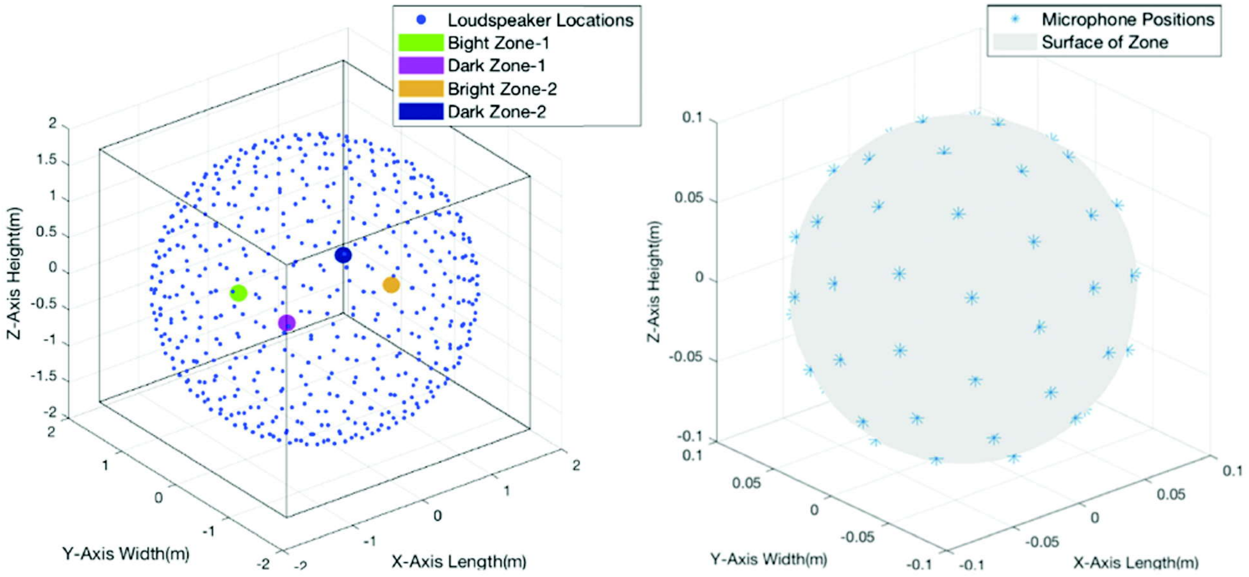


Fig. 2. Figure illustrating (a) the experimental setup with loudspeaker arrangement and zone positions (b) considered microphone positions in each zone.

i.e. I_q, $\forall_q$. Therefore, a total of 64 microphones are considered in each zone of reproduction, as shown in in Figure 2(b). In this experiment, among the available Q = 4 zone positions, bright zones are selected at diagonal positions. Though the choice of selection is completely application specific where the multi-zone scenario is considered, but in order to balance the distribution of bright and dark zones within the region of reproduction we evaluated the proposed framework with above assumptions.

3.2 Experimental Conditions

The proposed method is tested in a simulated multi-zone environment with the experimental setup as mentioned in Section-III-A. The simulated environment is considered with anechoic conditions. For simplicity, each loudspeaker is assumed to emit a plane wave with unit amplitude travelling in the direction of their position. Therefore, the complex acoustic channel matrices G_B, $\forall b \in N_B$ and G_D, $\forall d \in N_D$ are defined using equation-(2) corresponding to plane wave source. Also, in each bright zone, the target sound fields are considered to be a plane wave travelling from direction Ψ_1 = (150°, 0°) in bright zone-1 and Ψ_2 = (45°, 45°) in bright zone-2. The proposed method is tested for a broadband frequency ranging from F = 500 Hz to 3500 Hz.

3.3 Analysis of Reconstructed Sound Fields

In this work, the target sound field are reconstructed by two methods; the conventional-least square (LS) corre- sponding to pressure matching and the proposed method named as the least-square with differential phase constraint (LS-DPC). The reconstructed sound fields over the surface of both the bright and dark zones at F = 2000 Hz are illustrated in Figure 3. It can be observed that both LS and LS-DPC methods reproduce sound fields similar to target sound fields in both the bright zones. Also, both the methods minimize the energy flow significantly in both the darks zones. The performance of the reproduced sound field is also analysed using an average error distribution obtained in both the bright zones, Figure 4. In term of mean and variance of average error, both the methods show similar performance.

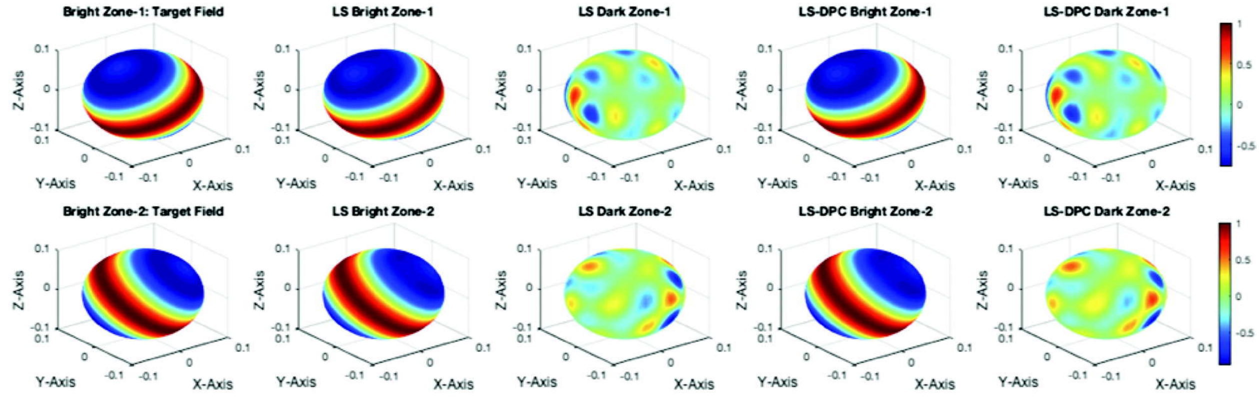


Fig. 3. Figure illustrating the target sound fields in both the bright zones along with the reproduced normalized sound fields using LS and LS-DPC methods in different zones obtained at F = 2000 Hz.

Additionally, a statistical analysis of average error distribution in both the bright zones for 2 different scenarios are listed in Table 1. In first scenario, the target field in bright zone-1 is fixed at Ψ_1 = (150°, 0°) and target field in bright zone-2 is rotated in 4 different directions as considered in the Table 1. In second scenario, the target field in bright zone-2 is fixed at Ψ_2 = (45°, 45°) and target field in bright zone-1 is rotated in 4 directions as considered earlier. From the Table, it can be observed that LS-DPC outperform in first scenario for case-1 and case-3 in both the bright zones and case-4 in bright zone-2.

Even in second scenario, LS-DPC shows similar performance as LS in both bright zones in case-1 and case-2. This results shows the equal performance of LS-DPC and LS in terms of average error distribution when averaged over multiple scenarios.

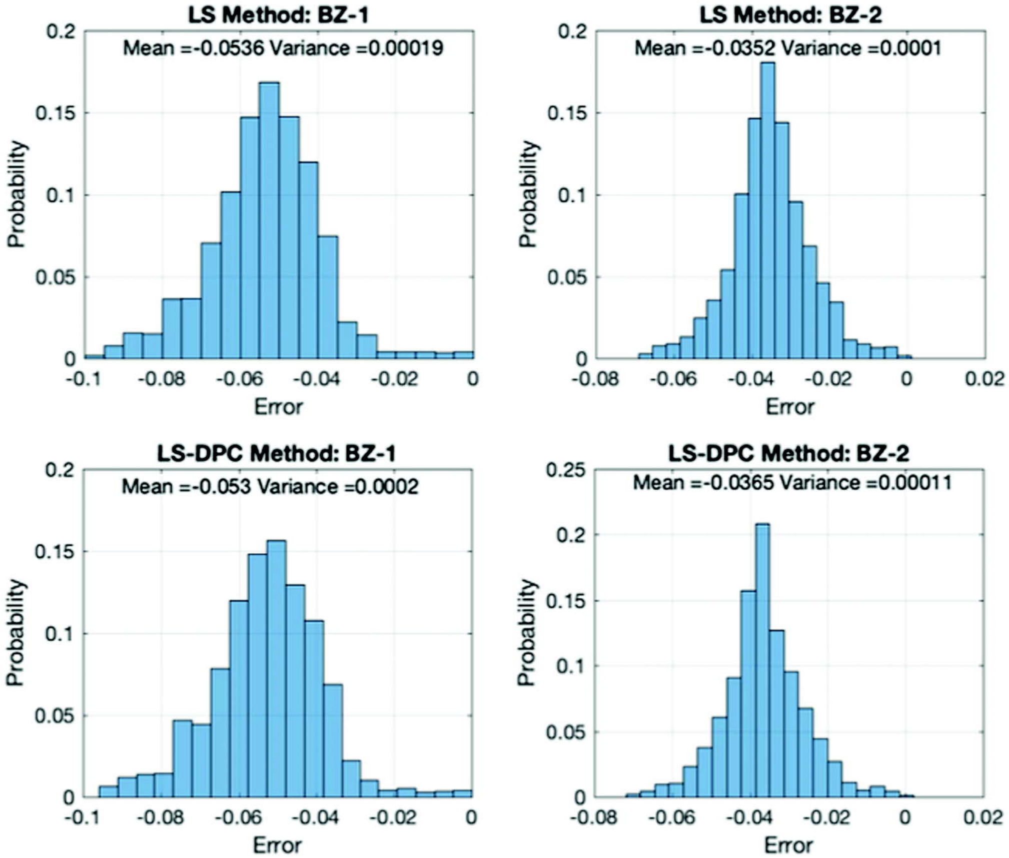


Fig. 4. Figure illustrating the average error distribution obtained by using LS and LS-DPC approaches obtained at frequency F = 2000 Hz.

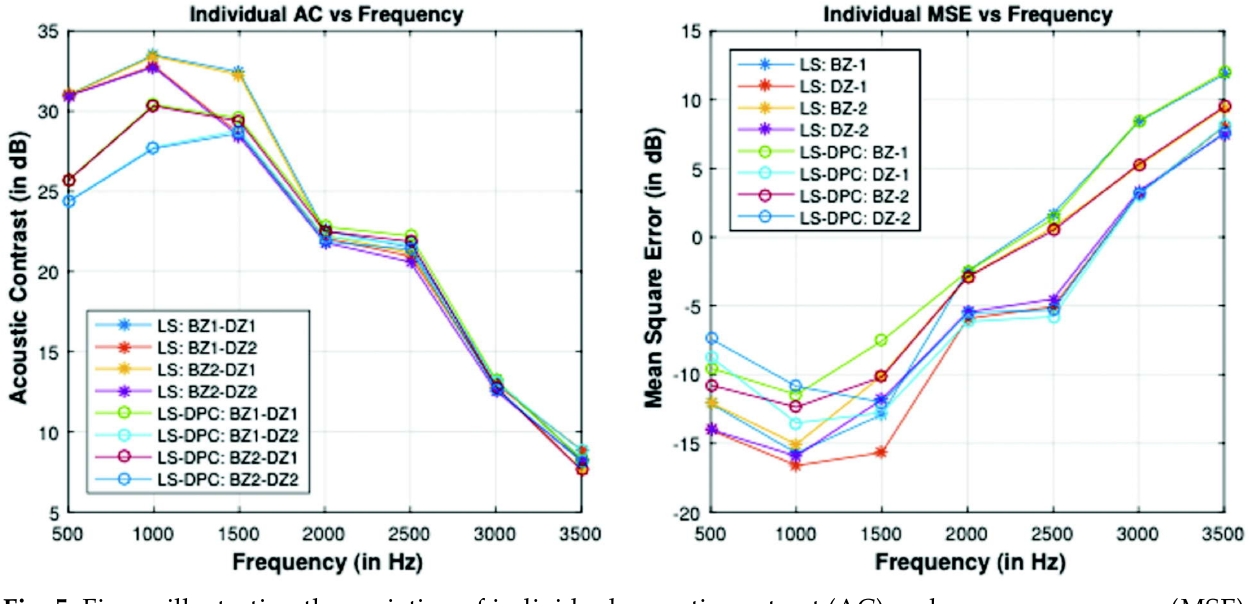


Fig. 5. Figure illustrating the variation of individual acoustic contrast (AC) and mean square error (MSE) obtained by using LS and LS-DPC approaches obtained at frequency F = 2000 Hz.

Table 1. Obtained statistical measures, mean (μ) and variance (σ^2), corresponding to average error distribution for both the bright zone (BZ) measured at F = 2000 Hz with fixed Ψ_1 = (150°, 0°) and fixed Ψ_2 = (450°, –45°) for (A) case-1 with Ψ = (0°, 36°) (B) case-2 with Ψ = (0°, 72°) (C) case-3 with and Ψ = (0°, 180°) (D) case-4 with Ψ = (0°, –36°).

Methods	Zones	Ψ_1 (Fixed)							
		Case-1		Case-2		Case-3		Case-4	
		μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
LS	BZ-1	-0.04356	0.0001556	-0.02868	0.0001163	-0.04114	0.0001432	-0.04105	0.0001382
	BZ-2	-0.02667	0.0000274	-0.03224	0.0000655	-0.01461	0.0000131	-0.02858	0.0000394
LS-DPC	BZ-1	-0.04110	0.0001411	-0.03167	0.0001065	-0.0328	0.0000933	-0.04234	0.0001471
	BZ-2	-0.02192	0.0000245	-0.03520	0.0000777	-0.01432	0.0000102	-0.02818	0.0000405
		Ψ_1 (Fixed)							
		Case-1		Case-2		Case-3		Case-4	
		μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
LS	BZ-1	-0.01527	0.0000220	-0.04666	0.0000974	-0.01513	0.0000166	-0.01804	0.0000476
	BZ-2	-0.03527	0.0001155	-0.05590	0.0001345	-0.03576	0.0001159	-0.03387	0.0001132
LS-DPC	BZ-1	-0.01542	0.0000185	-0.04632	0.0000974	-0.01644	0.0000263	-0.02139	0.0000537
	BZ-2	-0.03551	0.0001101	-0.05764	0.0001398	-0.03647	0.0001232	-0.03286	0.0001097

In Figure 5, the variation of individual acoustic contrast and individual mean square error (MSE) over the considered frequency range is presented. From Figure 5(a), it can be observed that the performance both LS and LS-DPC in terms of acoustic contrast are similar over the given frequency range. Also the performance in terms of MSE is observed similar for higher frequency ranges, as shown in Figure 5(b). It should be noted that the phase information is of lesser significance for the lower frequency ranges on a spherical region of reproduction having 20 cm of radius, *i.e.* equivalent to human head[20]-[21]. Therefore, an over-fitting of the phase degrades the performance of LS-DPC in lower frequency ranges.

Additionally, the placement of active number of loudspeakers and their corresponding weight are shown in Figure 6. Though, the conventional method has tendency to distribute reproduction error across all the loudspeaker, still the active number of loudspeakers are 510 when compared with 512 active loudspeakers in LS-DPC. The minimum threshold was set to $-20dB$ for a loudspeaker to be assumed in inactive state. From Figure 6, it can observed that the loudspeakers having lower energy in the range of 100–200 and 400–500 are boosted when the differential phase constraint is introduced. Further, the energy of the dominant loudspeakers becomes more evenly distributed after applying the phase constraint. Thus, it can be stated that LS-DPC method utilizes the loudspeaker setup more efficiently than conventional LS approach.

Although, both the LS and LS-DPC methods reproduce similar sound fields in both bright zones but it will be interesting to observe which method preserves the directivity close to the target sound fields. A detailed directivity error analysis comparing both the methods is presented in the next sub-section.

3.4 Directivity Analysis

In this section, the performance of both the methods LS and LS-DPC is evaluated on the basis of $y(k, \Theta_n)$, the directional component as defined in equation (9). The normalized distribution of directivity component, $\|y(k, \Theta_n)\|_2^2$, obtained in the target sound field and in the reproduced sound fields using both the methods are presented in Figure 7. It is difficult to distinguish the directivity patterns in both LS and LS-DPC methods by visual observations as both look alike. To quantify the difference, the corresponding

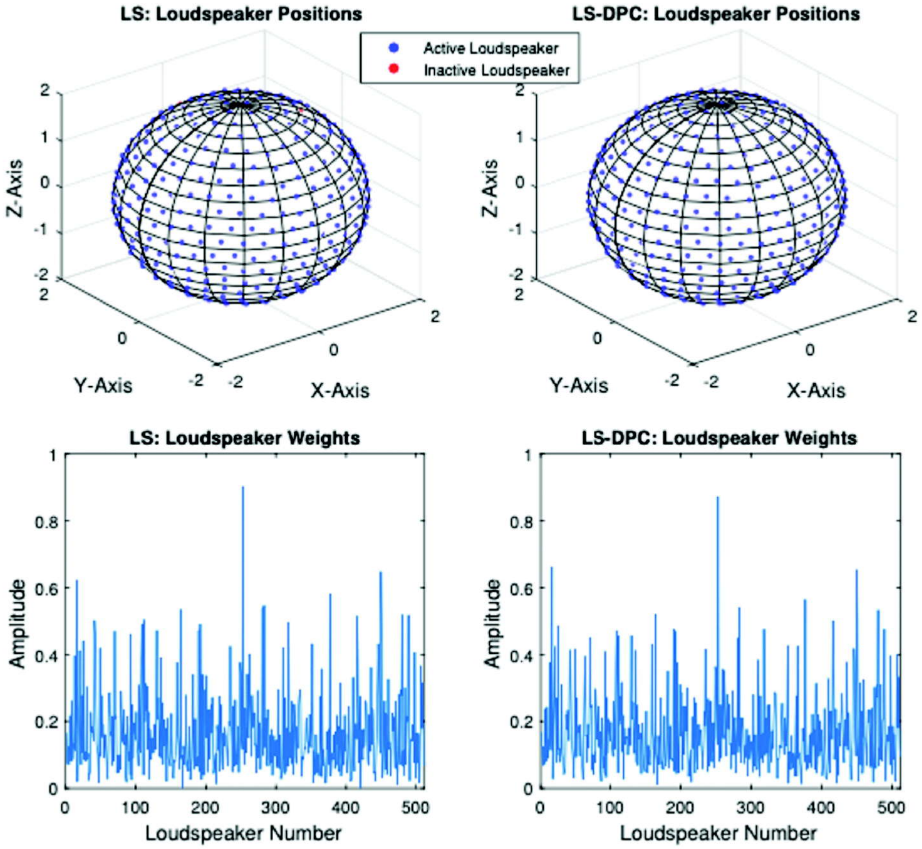


Fig. 6. Figure illustrating the active and inactive loudspeaker positions for (a) LS method and (b) LS-DPC method at frequency F = 2000 Hz. The corresponding loudspeaker weights thus obtained using LS and LS-DPC are shown in (c) and (d), respectively.

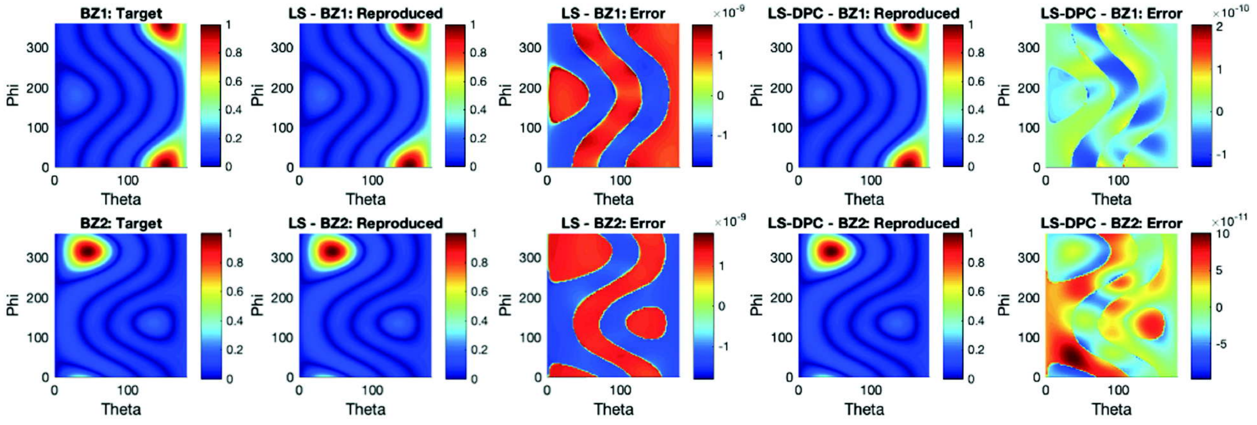


Fig. 7. Figure illustrating the normalized distribution of $|y(k, \Theta_n)|_2^2$ obtained in both the bright zones corresponding to (a) the target sound fields (b) reproduced sound fields using LS (d) reproduced sound field using LS-DPC and its corresponding error distribution (c) using LS and (e) using LS-DPC for F = 2000 Hz.

error distribution are obtained and shown in Figure 7(c) and Figure 7(e). Here, the error $E(k, \Theta_n)$ in the directional component $\|y(k, \Theta_n)\|_2^2$, is defined as

$$E(k,\Theta_n) = \|y(k, \Theta_n)\|_2^2 - \|y_{\text{target}}(k, \Theta_n)\|_2^2 \tag{17}$$

where, $y_{\text{target}}(k, \Theta_n)$ is the directivity component of the target sound field. From Figure 7, it can be clearly noticed that the range of error is lower in LS-DPC method when compared to LS method in both the bright zones. Thus, LS-DPC shows a significant improvement in preserving the directional component in both the bright zones when compared with LS approach.

Further, to show the effectiveness of LS-DPC method over LS method, the sum of absolute error (in dB) across multiple Θ_n for 2 different scenarios, as considered in Section-III-C, are listed in Table 2. From the Table, it can be observed that the sum of absolute error are better in LS-DPC method in both bright zones in case-1 and case-4 under first scenario where Ψ_1 = (150°, 0°) is fixed. Even in second scenario where Ψ_2 = (45°, 45°) is fixed, LS-DPC method outperform LS approach in most of the case scenarios.

Table 2. Obtained sum of absolute error (in dB) for both the bright zone (BZ) with fixed Ψ_1 = (150°, 0°) and fixed Ψ_2 = (45°, –45°) for (A) case-1 with Ψ = (0°, 36°) (B) case-2 with Ψ = (0°, 72°) (C) case-3 with and Ψ = (0°, 180°) (D) case-4 with Ψ = (0°, –36°).

		Ψ_1 (Fixed)				Ψ_2 (Fixed)			
Methods	Zones	Case-1	Case-2	Case-3	Case-4	Case-1	Case-2	Case-3	Case-4
LS	BZ-1	-59.93	-54.66	-54.91	-55.97	-41.99	-48.11	-48.63	-48.57
	BZ-2	-59.81	-54.73	-54.67	-55.83	-48.79	-47.96	-48.47	-48.45
LS-DPC	BZ-1	-57.86	-50.71	-52.43	-68.40	-55.40	-67.39	-55.97	-51.43
	BZ-2	-63.72	-55.38	-56.87	-66.55	-60.44	-69.30	-58.80	-53.63

4. CONCLUSIONS

In this work, a spatial multi-zone sound field reproduction method based on a differential phase constraint is presented. In-order to preserve the directional component in the reproduced sound fields, an additional phase constraint on the reproduced sound field is introduced in this framework. This convex optimization problem jointly minimizes the difference in acoustic pressure and the phase in the zones of interest. The multi-zone problem is thus formulated is generic in nature and can be applied to any multi-zone sound scenario. For the purpose of evaluation, the proposed framework is tested in a simulated environment to reproduce the desired sound in two different zones in the presence of two different dark zones. The performance of proposed method is compared with conventional pressure matching approach, formulated in least squares sense, in terms of acoustic contrast, reproduction error, loudspeaker performance and directivity analysis. The proposed approach outperforms the conventional method in terms of improved directivity and effective utilization of the given loudspeaker setup. In terms of acoustic contrast and reproduction error it provides similar performance when compared to state of the art approaches. An in-depth directivity analysis is also presented for various case scenarios of target sound fields in the sound zones of interest. Furthermore, the statistical results thus obtained supports the effectiveness of the proposed framework. As part of future work, the present framework can be extended to multi-zone sound reproduction in reverberant and noisy environments.

5. REFERENCES

[1] T. Betlehem, W. Zhang, M. A. Poletti and T. D. 2015. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Processing Magazine,* **32**(2): 81-91.

[2] J.-W. Choi and Y.-H. Kim, 2002. "Generation of an acoustically bright zone with an illuminated region using multiple sources," *The Journal of the Acoustical Society of America,* **111**(4), 1695-1700.

[3] J.-H. Chang, C.-H. Lee, J.-Y. Park and Y.-H. Kim, 2009. "A realization of sound focused personal audio system using acoustic contrast control," *The Journal of the Acoustical Society of America,* **125**(4), 2091-2097.

[4] J. Cheer, S. J. Elliott and M. F. S. Gálvez, 2013. "Design and implementation of a car cabin personal audio system," *Journal of the Audio Engineering Society,* **61**(6), 412-424.

[5] M. Poletti, 2008. "An investigation of 2-d multizone surround sound systems," in Audio Engineering Society Convention 125. *Audio Engineering Society.*

[6] T. Betlehem and C. Withers, 2012. "Sound field reproduction with energy constraint on loudspeaker weights," *IEEE Transactions on Audio, Speech and Language Processing,* **20**(8), 2388-2392.

[7] T. Betlehem and P. D. Teal, 2011. "A constrained optimization approach for multi-zone surround sound," in Acoustics, Speech and Signal Processing (ICASSP), *IEEE International Conference*, pp. 437-440.

[8] Y. Cai, M. Wu and J. Yang, 2014. "Sound reproduction in personal audio systems using the least-squares approach with acoustic contrast control constraint," The Journal of the Acoustical Society of America, **135**(2), 734-741.

[9] W. Jin, W. B. Kleijn and D. Virette, 2013. "Multizone soundfield reproduction using orthogonal basis expansion," in Acoustics, Speech and Signal Processing (ICASSP), *IEEE International Conference,* pp. 311-315.

[10] W. Zhang, T. D. Abhayapala, T. Betlehem and F. M. Fazi, 2016. "Analysis and control of multi-zone sound field reproduction using modal- domain approach," *The Journal of the Acoustical Society of America,* 140(3), 2134-2144.

[11] A. Dagar and R. M. Hegde, 2019. "Multi-zone sound field reproduction via sparse spherical harmonic expansion," in 1st *EAA Spatial Audio Signal Processing Symposium (EAA SASP-2019).*

[12] P. Coleman and P. J. Jackson, 2014. "Planarity panning for listener-centered spatial audio," in Audio Engineering Society Conference: 55th *International Conference: Spatial Audio. Audio Engineering Society.*

[13] P. Coleman, P. J. Jackson, M. Olik and J. Abildgaard Pedersen, 2014. "Personal audio with a planar bright zone," *The Journal of the Acoustical Society of America,* **136**(4), 1725-1735.

[14] B. Rafaely, 2015. Fundamentals of spherical array processing. *Springer,* **8**.

[15] S. J. Elliott, J. Cheer, J.-W. Choi and Y. Kim, 2012. "Robustness and regularization of personal audio systems," *IEEE Transactions on Audio, Speech, and Language Processing,* **20**(7), 2123-2133.

[16] X. Liao, J. Cheer, S. J. Elliott and S. Zheng, 2017. "Design of a loudspeaker array for personal audio in a car cabin," *Journal of the Audio Engineering Society,* **65**(3), 226-238.

[17] T. Van and L. Harry, 2002. "Optimum array processing: Detection estimation and modulation theory," in Part IV. *Wiley Interscience.*

[18] F. M. Fazi, T. Yamada, S. Kamdar, P. A. Nelson and P. Otto, 2010. "Surround sound panning technique based on a virtual microphone array," in Audio Engineering Society Convention 128. *Audio Engineering Society.*

[19] M. Grant, S. Boyd and Y. Ye, 2008. "Cvx: Matlab software for disciplined convex programming".

[20] P.-A. Gauthier, A. Berry and W. Woszczyk, 2005. "Sound-field reproduction in-room using optimal control techniques: Simulations in the frequency domain," *The Journal of the Acoustical Society of America,* **117**(2), 662-678.

[21] G. N. Lilis, D. Angelosante and G. B. Giannakis, 2010. "Sound field reproduction using the lasso," *IEEE Transactions on Audio, Speech, and Language Processing,* **18**(8), 1902-1912.

The Journal of Acoustical Society of India : Vol. 47, No. 2-3, 2020 (pp. 49-55)

Multi-component oscillatory model based classification of heart sounds

Samarjeet Das and Samarendra Dandapat
Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati, Guwahati, Assam-781 039
e-mail: samaren@iitg.ac.in

[Received: 30-03-2020; Revised: 17-05-2020; Accepted: 17-05-2020]

ABSTRACT

Automatic detection and classification of heart sounds (HSs) play a vital role in the diagnosis of cardiovascular diseases. In this paper, we propose a multi-component oscillatory model for the classification of HS segments of the PCG signal. A half-period sine wave is fitted between every two consecutive zero-crossing points to extract the proposed model parameters. The representation of the HS segments improved with the iterative use of multiple oscillations. The proposed method is tested and validated with a publicly available Physionet challenge 2016 database. The parameters of the model are deployed for the classification of normal and abnormal HS segments. The performance of the proposed method achieves a better average accuracy using a random forest classifier.

1. INTRODUCTION

Auscultation of heart sound (HS) is a primary and cost-effective method for the early detection of cardiovascular diseases (CVDs)[1]. HSs are of low-intensity acoustic vibrations in which bandwidth is ranging in between 10-1000 Hz[4]. Figure 1 shows the frequency range of HSs along with its intensity level. Phonocardiogram (PCG) signal records these sounds, and it is used widely to diagnose heart valve disorders (HVDs)[2]. It shows S1 and S2 sound patterns in healthy condition. However, in case of abnormalities along with these two sounds, other sounds, and murmurs, might occur[3]. Figure 2 shows one cycle of a standard PCG recording with its four segments, namely, S1 and S2 sound, Systole, and Diastole. The modeling of HSs is an essential tool in the automatic diagnosis process. The parameters of the model could be used as features to classify between normal and abnormal HS segment.

Numerous models have been proposed on the heart sounds for the analysis and classification of HSs. A pole-zero model has been proposed by Joo *et al.* to identify the prosthetic valve state[5]. Similarly, a chirp model is proposed by Xu *et al.* to extract the aortic and pulmonary components of S2 sound[6]. Damped-sinusoidal models are used widely for the analysis of heart sounds[7,8,9,10,11]. Rasmus *et al.* has classified S2 sound split using a windowed sinusoidal model[10]. Other approaches also used a matching pursuit model for the classification of valvular heart disorders[12]. The wavelet-based model is used by Maglogiannis *et al.* to classify heart sounds[13].

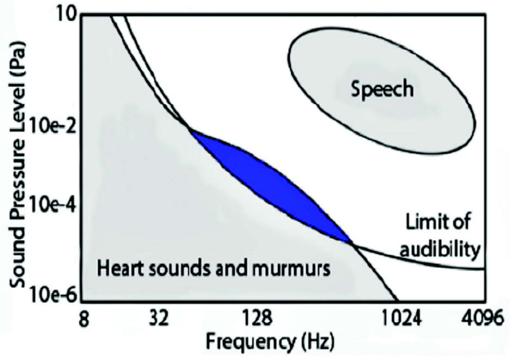


Fig. 1. Spectral characteristics of cardiac sound and their relation with the human audibility. The figure is adopted from Leatham *et al.*[4].

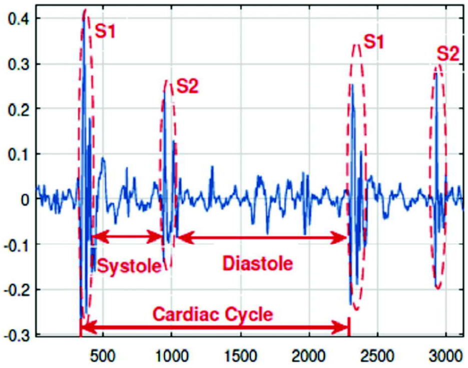


Fig. 2. Recording of a PCG signal. A cardiac cycle with its four segments i.e. S1, S2 Systole and Diastole.

Most of the existing models have been focused on the classification of a particular heart disorder. The performances of these models are not evaluated on a standardized database. In this work, we propose a multi-component oscillatory model that can capture the non-stationary behavior of HS segments properly. Further, the parameters of the model are deployed for the classification of normal and pathological HS segment using random forest (RF), support vector machine (SVM) and k-nearest neighbors (KNN) classifiers.

The rest of the paper is organized as follows: Section 2 presents the proposed method for the classification of the HS segment. Section 3 evaluates the model performance for the classification of HS. Section 4 ends with a few conclusive remarks.

2. PROPOSED METHODOLOGY

In this work, we propose a method for the classification of HS segments. The method consists of four sections, i.e., pre-processing of PCG signal, proposed multi-component oscillatory model-based feature extraction, and recognition of HS segments. Figure 3 shows the block schematic of the proposed method.

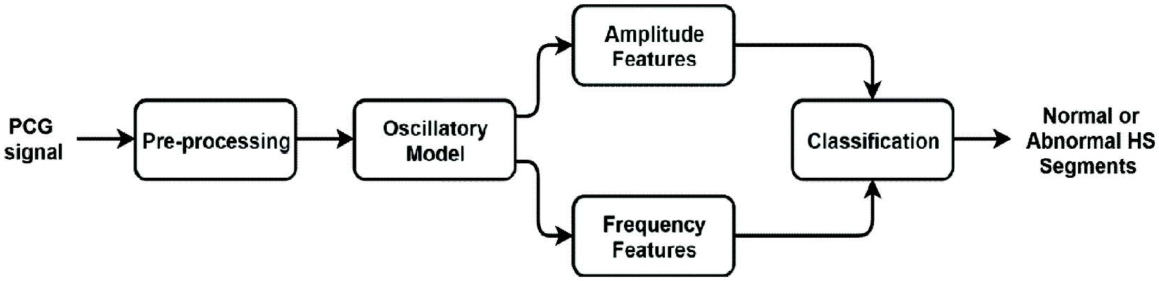


Fig. 3. Schematic outline of the proposed methodology.

2.1 Pre-processing

The unwanted signals such as ambient noise, lung sound, and surrounded speech affect the PCG signal recording[15]. Thus, the PCG signal is applied through a sixth-order Butterworth high pass and low pass filter with cut-off frequencies 10 Hz and 800 Hz, respectively. The friction spikes, which amplitudes are higher than the heart sounds, are removed using the Schmidt spike removal process[16]. The filtered PCG signal is segmented into cycles or frames. Each cycle is further segmented into four sections; namely, S1, S2, systole, and diastole correspond to each cardiac cycle. The segmentation of PCG plays a crucial role in localization and analysis of cardiovascular diseases, which affects those particular regions. Springer's segmentation algorithm is performed for the segmentation of these HSs[17].

2.2 Proposed Oscillatory Model

The proposed oscillatory model is employed upon both the normal and abnormal HS segments. A half period sine function is fitted between every two consecutive zero-crossing points to model the HS segment. Suppose S denotes a non-stationary signal with the number of zero-crossings *(N_z)* and sample points *(N_t)*. The parameters of the sine function between two consecutive zero-crossing points ti and ti+1 *i.e.*, amplitude, frequency, and phase are calculated as[14].

$$a^i = \frac{\pi}{2} mean(S)\Big|_{t_i}^{t_{i+1}},\ \omega^i = \frac{\pi}{t_{i+1}-t_i},\ \phi^i = \frac{\pi t_i}{t_{i+1}-t_i}$$

Sequentially, the model parameters for the entire signal(S) are formulated as:

$$a(t) = a^i,\ \omega(t) = \omega^i,\ \varnothing(t) = \varnothing^i,\ t_i < t \le t_{i+1} \qquad (1)$$

Where $i \in [1, N_z - 1]$, and $1 \le t \le N_t$. So, the representation of the signal (S) using the proposed model parameters of the Eq. (1) as follows:

$$S = a(t)[\sin[\omega(t).t + \varnothing(t)] \qquad (2)$$

It has been observed that HS segments are not mostly cycloidal shape, and in such cases, it isn't straightforward to represent the signal using Eq. (2). Therefore, to capture the variations of the HS segments properly, multiple oscillations or components can be modeled at different time instants. So, the above Eq. (2) can be rewritten as follows:

$$S = \sum\nolimits_{p=1}^{p} a^p(t)\sin[\omega^p(t).t + \varnothing^p(\mathrm{t})] \qquad (3)$$

Where $a^p(t)$, $\omega^p(t)$ and $\phi^p(t)$ are the corresponding amplitude, frequency and phase, respectively for p^{th} component sinusoid. Now, to compute for the p^{th} component parameters, firstly, the residual signal is calculated by finding the difference between the original signal and the reconstructed one (which is the

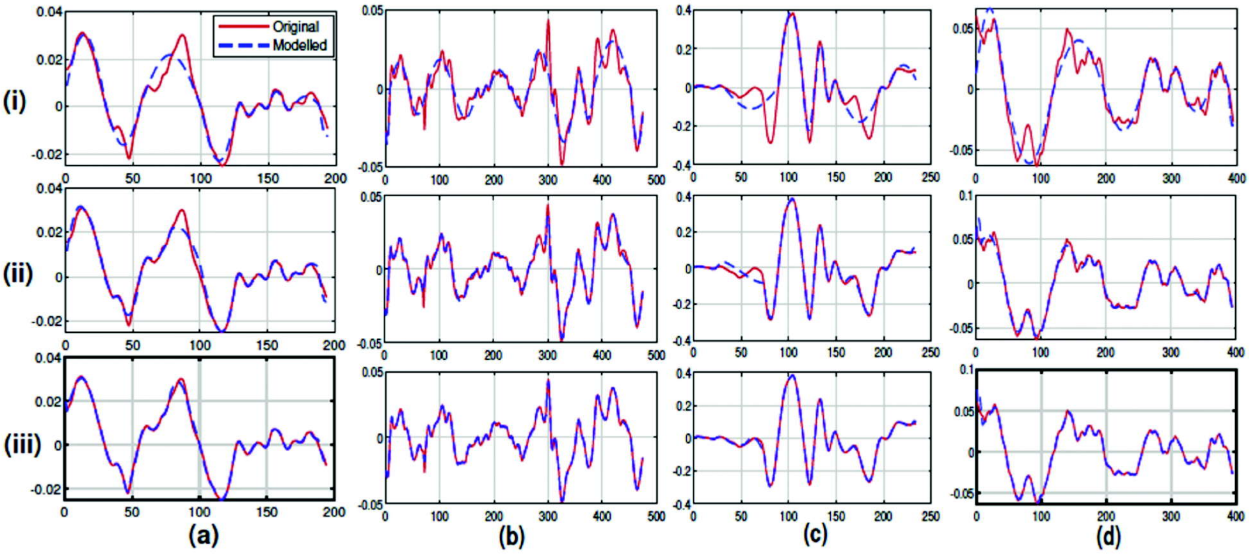


Fig. 4. Reconstruction of HS segments (a) S1 sound, (b) Systolic segment, (c) S2 sound, and (d) Diastolic segment. (i) The four subplots in the first row depict the first model component of all four HS segments. (ii) and (iii) depicts the second and third model components of the four HS segments.

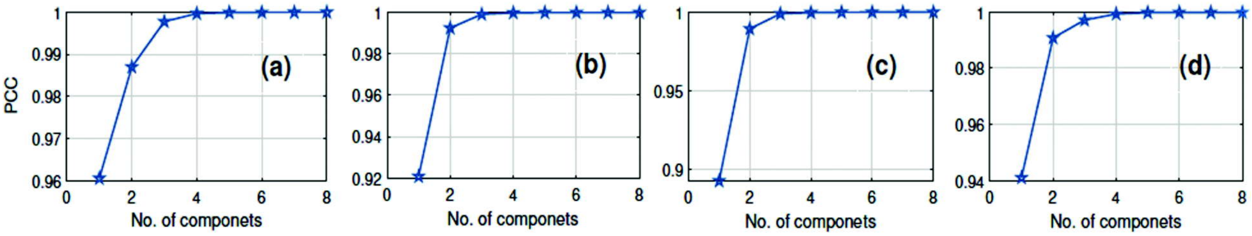


Fig. 5. PCC of normal HS segments up to model component 8.
(a) S1 sound, (b) Systolic segment, (c) S2 sound, and (d) Diastolic segment.

sum of the last p – 1 components) iteratively, by finding the model parameters of the Eq. (1). Figure 4 shows all the four HS segments of a normal PCG recording with its reconstruction for different model components. It is observed that the first model component is not able to capture all the morphological variations of the HS segments. So, with the increase in model components, it helps in extracting detail variations of the segment perfectly. Figure 5 shows the plot of the Pearson correlation coefficient (PCC) with the model components. It is observed that after the third model component, the PCC value reaches almost one.

2.3 Proposed model based features

The selective parameters of different model components of the HS segment are used as features for the classification of HS segments. It is tested that among parameters, amplitude, and frequency features perform better. Finally, eight prominent amplitudes and their corresponding frequencies from model component 1, while four prominent amplitudes and their corresponding frequencies from model component 2 are deployed to form the feature set for the proposed model. Statistical analysis is performed among these features. This shows that the distribution of these features has the capability to distinguish between normal and abnormal HS segment.

2.4 Classification

The selected feature sets are employed for the classification of normal and abnormal HS segment. The performance of the proposed method is evaluated using three classifiers, k-nearest neighbor (KNN), support vector machine (SVM) and random forest (RF). The classification performance is evaluated using a standard 10-fold cross-validation method.

3. RESULTS AND DISCUSSION

The proposed method is evaluated on a publicly available Physionet challenge 2016 database[18]. The database contains five datasets, namely, training set-a, b, c, d, e, and f. It contains a total of 3240 (2575- Normal and 665- Abnormal) PCG recordings from 764 subjects/patients. The sampling rate of each recording is 2 kHz, with a duration varied from 5 s to 120 s. Firstly, all the healthy and pathological PCG recordings are pre-processed. A total of 68647 numbers of normal and 18,636 numbers of abnormal segments of all the four HS segments are obtained. Each segmented HS is modeled by the proposed method. The evaluation of the proposed model is carried out with suitable measures for the recognition of HS segments.

The selected feature vector for different HS segments shows better potential for the classification. Figures 6 and 7 show the box plots of eight prominent amplitude features of first model component for S1 and S2 sound. It can be observed that the distribution characteristics differ between normal and abnormal cases for both S1 and S2 sound. Three supervised classifiers, RF, KNN, and SVM, respectively, are deployed to evaluate the proposed model performance. The configuration utilized for these classifiers, such as RF (estimators = 10, criterion = entropy), SVM (kernel function = RBF), KNN (k=10, distance measure = Euclidean). Each classifier's performance is 10-fold cross-validated. Table 1 shows the average

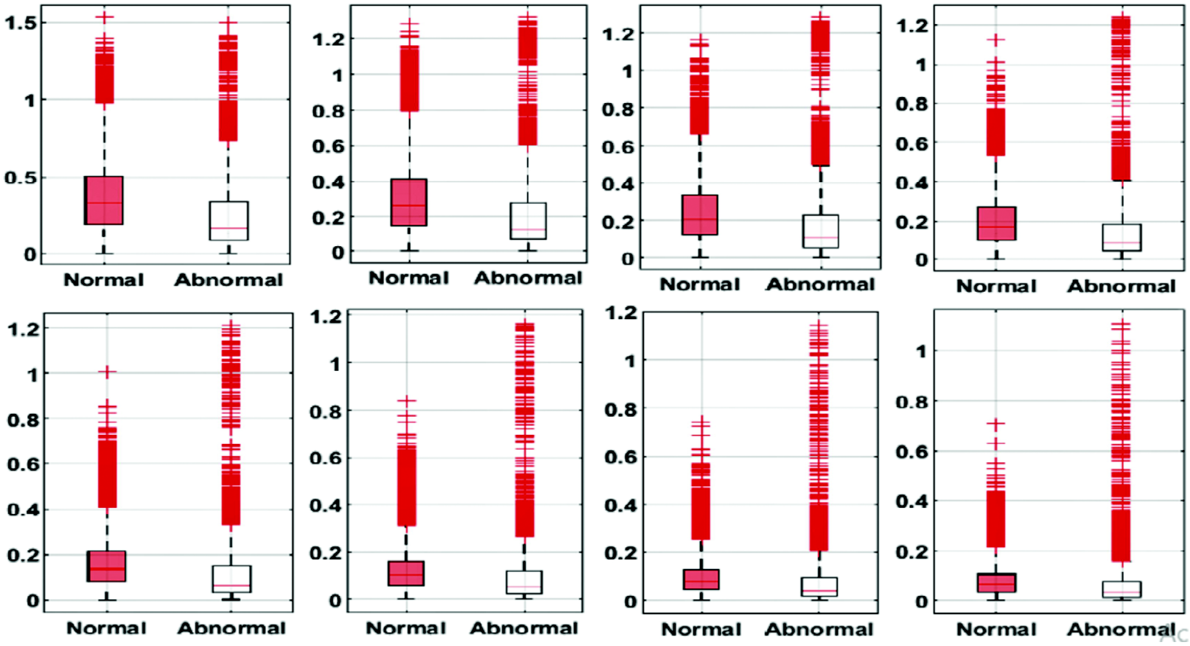


Fig. 6. Box plot of eight prominent amplitude features for the first model component of S1 sound.

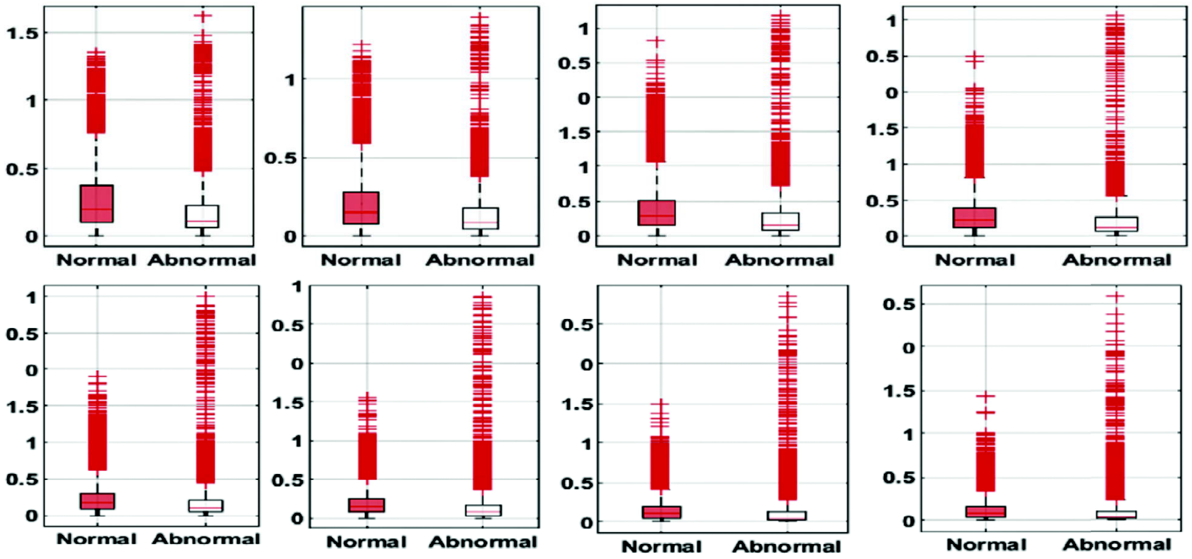


Fig. 7. Box plot of eight prominent amplitude features for the first model component of S2 sound.

Table 1. Average classification accuracy (%) for the HS segments.

Classification category	Classifier		
	RF	KNN	SVM
S1 sound	84.16	83.73	82.88
Systolic segment	83.25	82.63	82.75
S2 sound	84.85	84.74	83.84
Diastolic segment	82.81	81.66	80.61

classification accuracy for the HS segments. The result showed that the RF classifier has better average classification accuracy (%) for all the HS segments as compared to KNN and SVM. It produces average accuracies (%) of 84.16, 83.25, 84.85, and 82.81 for the S1 sound, systolic segment, S2 sound, and diastolic segment, respectively.

4. CONCLUSION

This paper presents a multi-component oscillatory model for the classification of HS segments as normal and abnormal. The parameters of the model are calculated between two consecutive zero-crossing points of an HS segment. The performance of the model is further improved with the addition of multi-component sine wave functions iteratively. The experiments were performed on the CinC challenge 2016 database, available in the Physionet archive. The classification results show that the average accuracy (%) of the RF classifier outperforms SVM and KNN for all the HS segments. In the future, proposed model features can be employed for real-time monitoring of heart valve disorders.

5. REFERENCES

[1] A.K. Bhoi *et al.,* 2015. Multidimensional analytical study of heart sounds: A review, *Int. J. Bioautomat.,* **19**(3), 351-376.

[2] P. Chandraratna *et al.,* 1975. Echocardiographic observations on the mechanism of production of the second heart sound, *Circulation,* **51**(2), 292-296.

[3] A.K. Dwivedi *et al.,* 2018. Algorithms for automatic analysis and classification of heart sounds-a systematic review, *IEEE Access,* **7**, 8316-8345.

[4] A. Leatham, 1975. Auscultation of the Heart and Phonocardiography, Churchill Livingstone, New York.

[5] T.H. Joo *et al.,* 1983. Pole-Zero Modeling and Classification of Phonocardiograms, *IEEE Trans. Biomed. Eng.,* **30**(2), 110-118.

[6] Jingping Xu *et al.,* 2001. Extraction of the aortic and pulmonary components of the second heart sound using a nonlinear transient chirp signal model, *IEEE Trans. Biomed. Eng.,* **48**(3), 277-283.

[7] H. Koymen *et al.,* 1987. A Study of Prosthetic Heart Valve Sounds, IEEE Trans. *Biomed. Eng.,* **34**(11), 853-863.

[8] T.S. Leung *et al.,* 1998. Analysis of the second heart sound for diagnosis of paediatric heart disease, IEE Proceedings - Science, *Measurement and Technology,* **145**, 285-290.

[9] A. Baykal *et al.,* 1995. Distribution of aortic mechanical prosthetic valve closure sound model parameters on the surface of the chest, *IEEE Trans. Biomed. Eng.,* **42**(4), 358-370.

[10] Saederup *et al.,* 2018. Estimation of the second heart sound split using windowed sinusoidal models, Biomed. *Signal Processing and Control,* **44**, 229-236.

[11] Tang *et al.,* 2016. Phonocardiogram signal compression using sound repetition and vector quantization, *Computers in biology and medicine,* **71**, 24-34.

[12] Jabbari *et al.,* 2011. Modeling of heart systolic murmurs based on multivariate matching pursuit for diagnosis of valvular disorders, Computers in biology and medicine, **41**(9).

[13] I. Maglogiannis *et al.,* 2009.Support Vectors Machine based identification of heart valve diseases using heart sounds, *Comput. Methods Prog. Biomed,* **95**, 47-61.

[14] G. Andre *et al.,* 2014. A parsimonious oscillatory model of handwriting, *Biol. Cybern,* **108**(3), 321-336.

[15] H. Tang *et al.,* 2010. Separation of Heart Sound Signal from Noise in Joint Cycle Frequency-Time-Frequency Domains Based on Fuzzy Detection, *IEEE Trans. Biomed. Eng.,* **57**(10), 2438-2447.

[16] Schmidt *et al.*, 2010. Segmentation of heart sound recordings by a duration dependent hidden Markov model, *Physiological measurement*, **31**.

[17] D.B. Springer *et al.*, 2016. Logistic Regression-HSMM-Based Heart Sound Segmentation, *IEEE Trans. Biomed. Eng.*, **63**(4), 822-832.

[18] Liu *et al.*, 2016. An open access database for the evaluation of heart sound algorithms, *Physiological Measurement,* **37**.

The Journal of Acoustical Society of India : Vol. 47, No. 2-3, 2020 (pp. 56-70)

An overview of techniques developed for Bio-SONAR characterization and census of Ganges river Dolphin, India's National aquatic animal

Rajendar Bahl

Centre for Applied Research in Electronics, Indian Institute of Technology Delhi,
Hauz Khas, New Delhi-110 016, India
e-mail: rbahl@care.iitd.ac.in

[Received: 20-03-2020; Revised: 23-05-2020; Accepted: 23-05-2020]

ABSTRACT

This paper provides an overview of major research activities since the year 2006 related to Passive Acoustic Monitoring (PAM) of the Ganges River Dolphin in India. The research has been coordinated by the author for the joint field expeditions funded by Japan in Odisha (at Budhabalanga River), in UP (at Karnavas, Bhitora) and a fully indigenous effort in UP (at Narora) enabled by CSR funding. The paper presents pioneering observations of bio-sonar characteristics including dolphin click pulse temporal properties, Source Level, and beam width using a hydrophone array consisting of a combination of linear SBL and a triangular SSBL array in Budhabalanga River. Subsequently, a specialised cross-array has been developed for habitat related studies in Ganges River at Karnavas that cover the variation in inter-click interval (ICI) depending on activity, and day-to-day temporal behaviour in multi-year studies. We present the field calibration test of an indigenous large triangular array for acoustic localization of dolphins up to over 250 metres in Narora. This has been followed up with a 2-hydrophone array for conducting acoustic census of dolphins from a moving boat. Several encounters with dolphins show the utility of this innovative technique which is also corroborated with visual recordings with on-board video cameras.

1. INTRODUCTION

The Ganges River Dolphin *Platanista gangetica* is the flagship marine species of the Ganges River system, being on top of the food chain in the river ecosystem. Therefore, it has been used as an indicator species for the health of the river system. It has been a matter of deep concern that the numbers of these animals have been dwindling due to various human activities. Due to its endangered status, it has been given strong protection by the government and has been declared the National Aquatic Animal in 2009. Monitoring the habitat and numbers of these animals is considered very important. The usual way is by human visual observation, but that is prone to various errors and constraints such as effects of weather and visibility. Since the dolphins use echolocation, one way to observe them is by passive acoustic monitoring (PAM). However, prior to the year 2006, very limited studies were available and little had been known about the acoustic characteristics of the bio-sonar of the Ganges River dolphin in the wild. This paper provides an overview of the research work done in developing PAM technologies for monitoring their habitat and also for counting their numbers. Techniques were first developed jointly with researchers from Japan, and subsequent CSR funding helped in developing indigenous technologies including Integrated Visual and Acoustic Survey (IVAS).

2. BIO-SONAR CHARACTERISTICS OF A LONE FREE-RANGING ANIMAL

The first systematic study of bio-sonar characteristics of Ganges dolphin in the wild environment was carried out during April 2006, when an isolated dolphin in the Budhabalanga river in Odisha (Orissa) was studied by a team of scientists. In this section, we discuss and present results on the various bio-acoustic aspects of the Ganges river dolphin.

2.1 Acoustic Observation System

Acoustic observation of echolocation clicks of marine mammals requires a specially designed array of hydrophones that can be effectively deployed in the animal's river habitat. Since the dolphin is manoeuvring in 3-dimensions in the water column, it is required to localize the sound source in 3-dimensions. A multi-hydrophone 3.2 metre long array composed of three hydro-phones (H1, H2, H3) forming an equispaced linear short base line (SBL) array and another two hydrophones (H4, H5) in conjunction with the central hydrophone (H2) forming a small super short base line (SSBL) triangular array in a plane perpendicular to the array axis shown in Fig. 1 has been used[1]. The array can be deployed both in horizontal and vertical orientation. The positional accuracy of the SBL system is better than 1 metre in range and less than 12 cm in cross range at a range of 50 metres.

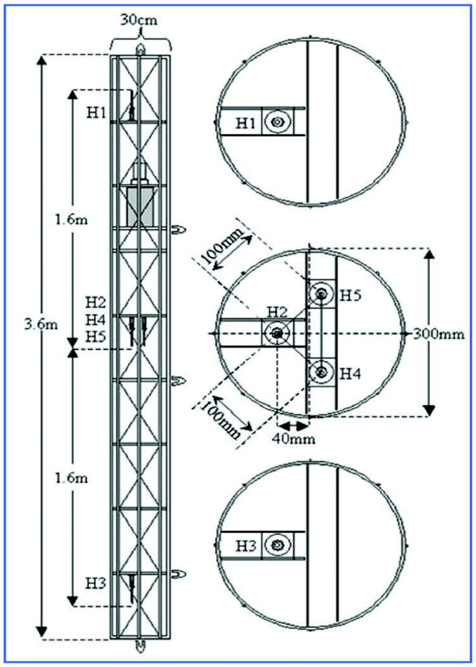


Fig. 1. Hydrophone array structure [1].

The recording system is shown in Fig. 2. The echolocation clicks are amplified over a sufficiently wide bandwidth of 30-180 kHz and sampled at a rate of 500 k-samples/second with a 16-bit A/D converter. The recording duration for each data file was set to 300 seconds.

2.2 Acoustic Characteristics of Clicks

A typical click (Fig. 3) is a short pulse of about 40 micro-seconds having a centre frequency of about 65 kHz. Figure 4 shows a sample of click data record. The ICI (inter click interval) of the click trains lies in the range of 20-160 milliseconds. The intermittent bursts of click trains indicate it is not possible for all emitted clicks to be recorded by the hydrophones due to the manoeuvring dolphin that is having a narrow sonar transmission beam.

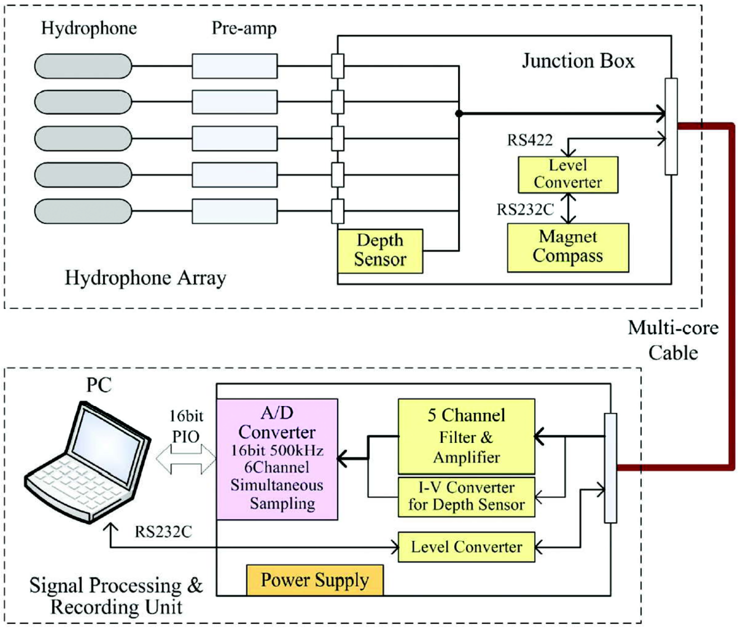


Fig. 2. Recording system.

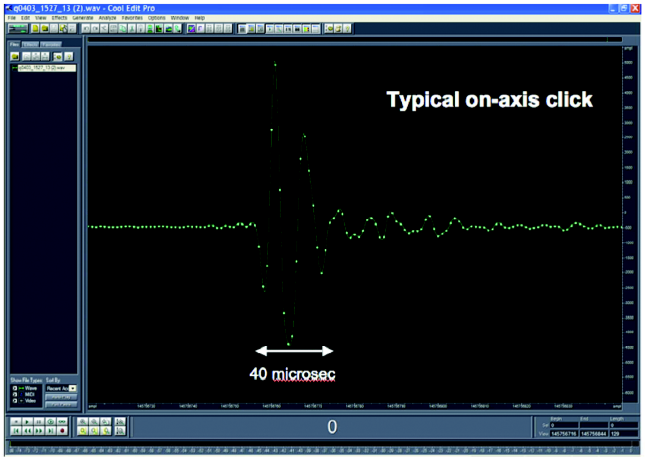


Fig. 3. Ganges river dolphin's typical click

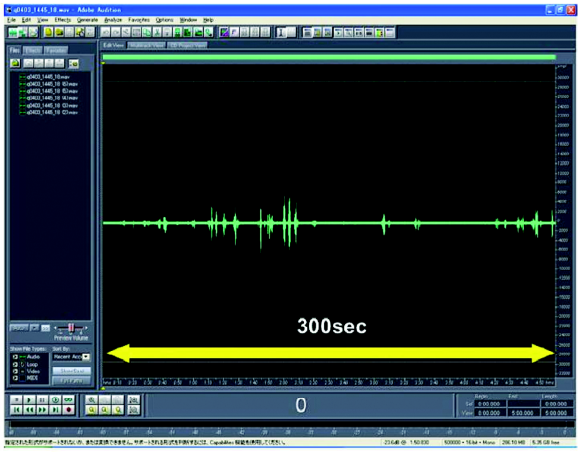


Fig. 4. Click trains observed over 300 seconds.

2.3 Source Level

Click source level is an important parameter to predict detection range of the dolphins using hydrophones. An estimate of the dolphin's 3-D location was first made from the relative delays of the click signal received at the three SBL hydrophones with the array deployed vertically. The signal level received at each hydrophone was then used to estimate the source level, called the apparent source level (ASL) using the hydrophone sensitivity and the range estimate of the source. The peak-to-peak SL of the dolphin click was obtained from those clicks that were directed towards the array, and was determined to average 172.5 dB re 1 μPa (see Fig. 5)[2]. The lone dolphin was usually seen swimming around the array within a range of approximately 30 metres during the experiment (see Fig. 6). The dolphin's trajectory indicated that clicks were well detected only when the dolphin swam generally toward the array. From this, we presume that the dolphin's transmission beam pattern is quite narrow so that the array would not record clicks when the dolphin was momentarily not facing the array.

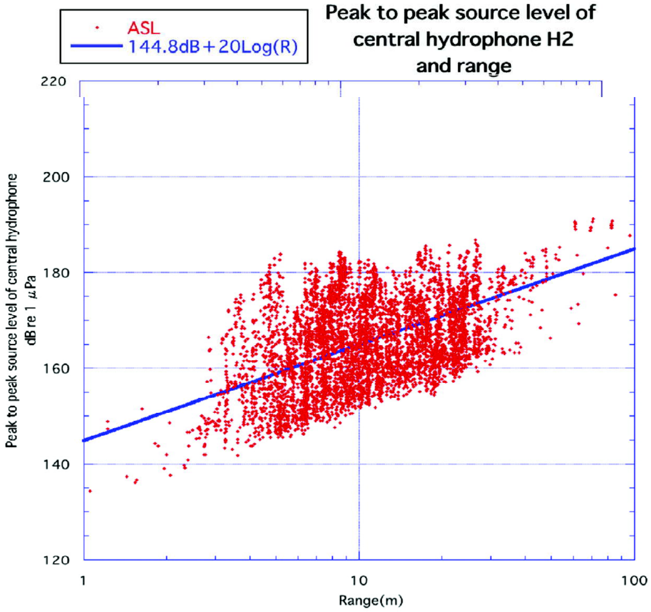


Fig. 5. Peak-to-Peak ASL on the central hydrophone versus range of the dolphin [2].

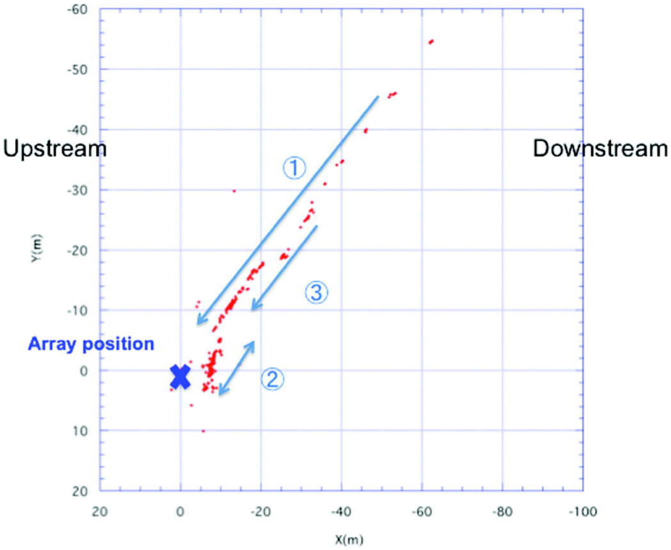


Fig. 6. An example of the dolphin's XY movement during 300 seconds. The dolphin was approaching towards the array (arrow no.1), then moving around the array area (arrow no. 2), and then again approaching the array (arrow no. 3) [2].

2.4 Beam Pattern

The beam pattern of the dolphin clicks in horizontal and vertical planes was innovatively obtained with the array deployed in horizontal and vertical configurations, respectively. The apparent source level (ASL) corresponding to the actual beam pointing direction (peak direction) was estimated by fitting a quadratic curve on measured values of ASL1, ASL2 and ASL3 (where the subscript refers to hydrophone number) on the 3 hydrophones of the SBL array. Using the location of the dolphin determined from the time delays, the angular separation of H1 and H3 from the peak direction were first estimated. The ASL1 and ASL3 were then plotted at these estimated angles to recreate the Ganges river dolphin's beam pattern in the appropriate plane, as shown in Fig. 7. This is the first time that a free-ranging Ganges dolphin's beam pattern was obtained.

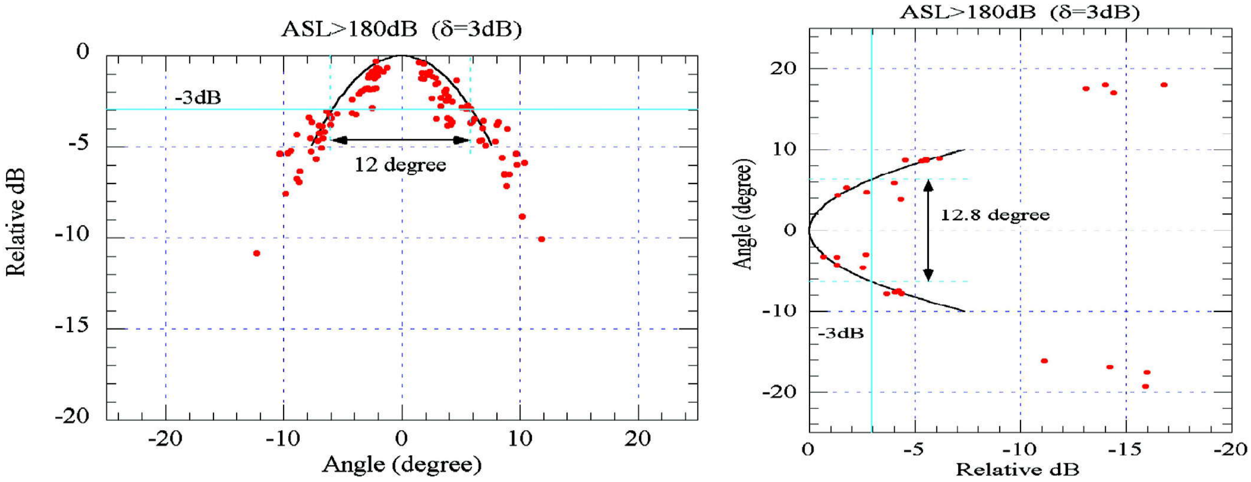


Fig. 7. The dolphin's the -3dB beamwidth is approximately 12 degrees in the horizontal plane and 12.8 degrees in the vertical plane [2].

3. HABITAT-RELATED STUDIES IN NARORA

Studies in the habitat of dolphins in a river system require a fixed array structure that does not require to be specifically oriented as in the previous case. Since the river habitat is mainly spread horizontally with relatively less depth, a horizontal cross array of 4 hydrophones with a 5th central hydrophone and a 6th vertically displaced hydrophone is adequate for localization.

3.1 Array system

A high frequency 6-hydrophone cross array system developed in January 2008 (Fig. 8) has been utilized for the long-term real-time monitoring at Narora. The inter-hydrophone spacing is 80 cm. The system calculates each dolphin click's real-time 3-D location and transmits the 3-D data to the host server through a GPRS modem. The system can be interfaced with sensors to monitor the river environment such as water quality (CTD), pH, dissolved oxygen, transparency *etc*. A computer with appropriate viewer program can connect with the host server anywhere by internet and a GUI can display the real-time location of the dolphins on the habitat map.

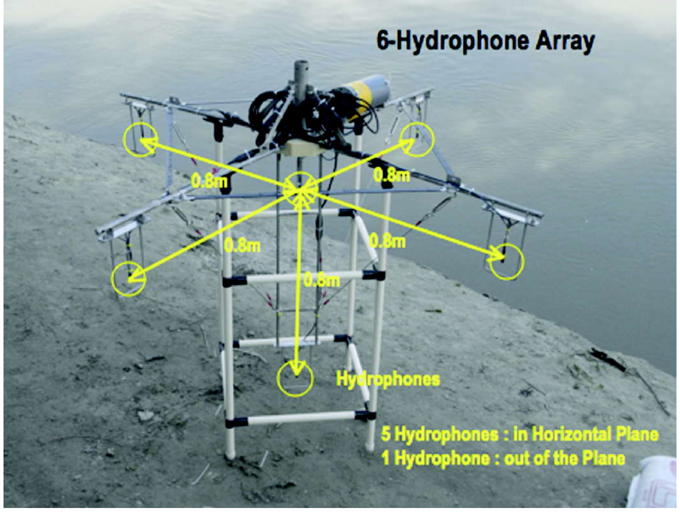

6-Hydrophone Array
0.8m
0.8m
0.8m
0.8m
0.8m
Hydrophones
5 Hydrophones : in Horizontal Plane
1 Hydrophone : out of the Plane

Fig. 8. High frequency 6-hydrophone array [3]

3.2 Typical behaviour of individual dolphins

The first long-term real-time habitat monitoring of the Ganges river dolphins was conducted from 12th November 2008 for 4 months in Karnavas village near Narora (UP). Several adults and a few calf dolphins are found to inhabit between Karnavas and Narora. It may be noted that Ganges river dolphin's underwater behaviour had been unknown until this time.

A particular data record was analyzed (Fig. 9), where one dolphin came to within approximately 10 meters from the array as shown in Fig. 10[3]. The ICI of the dolphin is found to vary during a data record perhaps due to its particular activity. It is normal for a man-made sonar to reduce ICI while approaching a particular object of interest, in order to get fast updates on the target. A similar situation has been observed in dolphin sonar. The vertical profile of the dolphin's position showed that during the bottom stay on 2-3 occasions lasting 0.4 to 3.7 seconds, ICI rapidly reduced to approximately 10 milliseconds from 30 milliseconds as it would have been investigating and approaching a specific target of interest.

From these results, it is estimated that the echo-location process of the dolphin changes between the normal search phase to a special interest phase corresponding to a reduction in ICI. This leads us to conjecture that the dolphin is foraging or looking at obstacles near the bottom during that period.

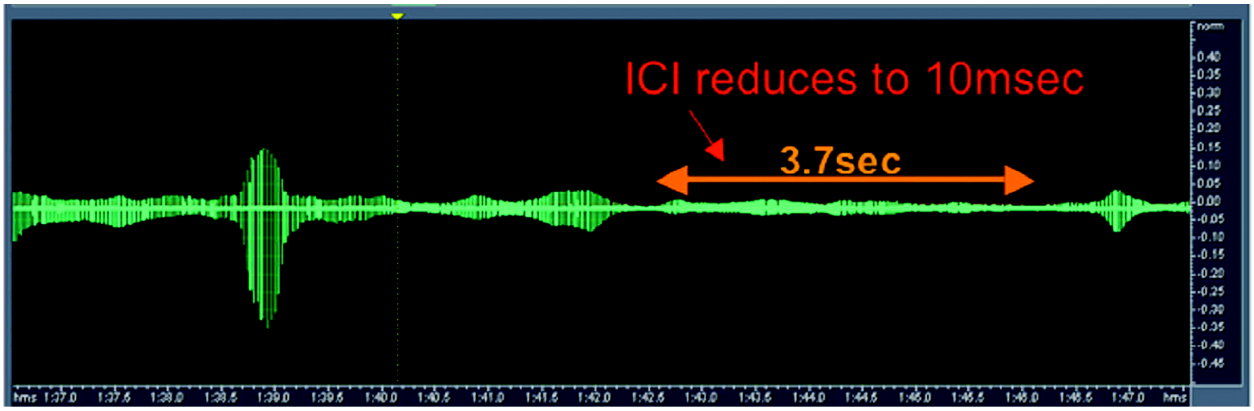


Fig. 9. Wave file data record of a click train [3]

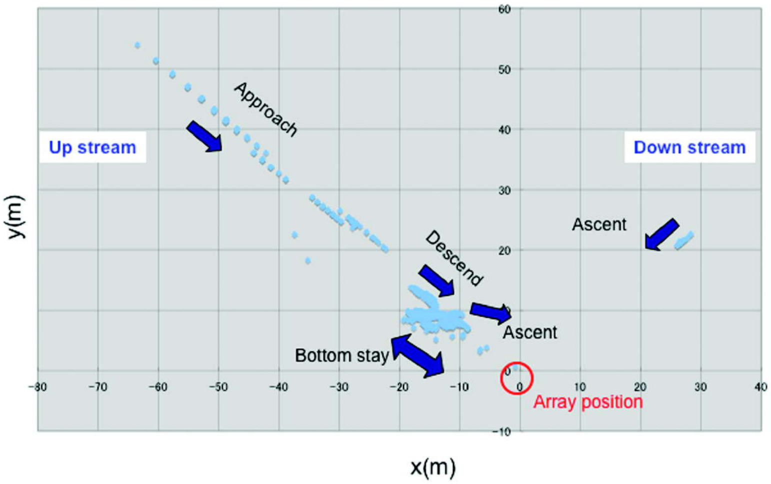


Fig. 10. 2-D trajectory of the dolphin

3.3 Temporal behaviour of the Narora population

Long term monitoring allowed us to understand temporal behavior of the dolphins in Narora. Temporal behaviour data for three consecutive years is presented here. The data is analyzed based on number of clicks number, average distance X (longitudinal position along the general river direction) and its standard deviation (SDX). Click number in a particular period is related to the dolphins' length of stay around the observation area, while average X and its SDX provides information of their habitat location along the river (upstream/downstream) and the region of frequent stay respectively[4]. Fig. 11 shows the monthly trends of average click number, average X and its standard deviation SDX on a particular day during three monitoring phases. Similar trends of averaged click number are seen in the first and second phases, *i.e.* the peak click number was reached in January in the first and second phase. Average X and its SDX also show similar trends of fluctuations in the first and second phases. However, the first phase shows larger average X (in the first half period) and smaller SDX (in the entire period) than the second phase. On the other hand, the trends of average click number, average X and its SDX in the third phase is different from the other two monitoring phases, *i.e.* the peak click number was reached in December, and average X and its SDX were almost stable in the third phase. From this analysis, we could say that the migration of the dolphins from the observation area began from the end of March to the beginning of April in the first and second phases, while the dolphins were mostly migrating or passing

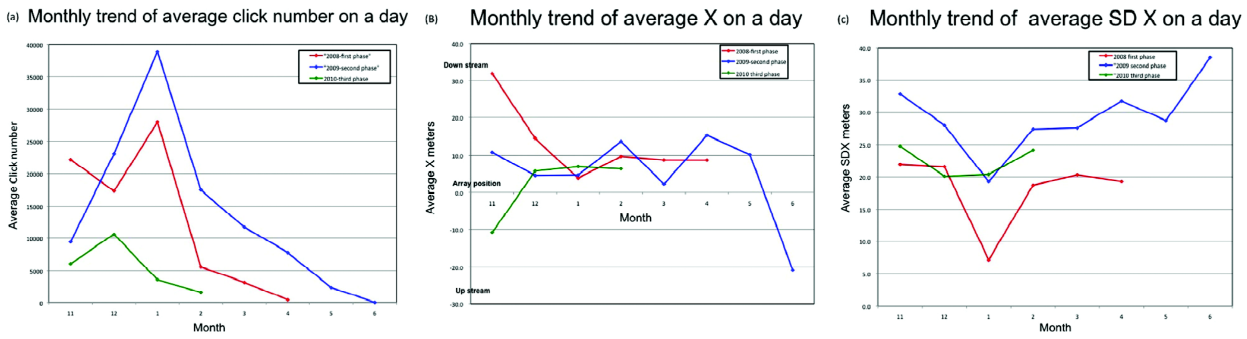


Fig. 11. Monthly trends of average click number changes, average X and its standard deviation SDX in three monitoring phases [4]

through the array area in the third phase during that period. It has also been noted by correlation that the Ganges river dolphins usually start migrating when the water level increases because of the seasonal changes (start of the monsoon season).

4. INDIGENOUS DESIGN OF AN INTEGRATED VISUAL AND ACOUSTIC SURVEY SYSTEM (IVAS)

4.1 Background

The conventional method of counting dolphins is to have trained dolphin watch personnel count the dolphins manually when they surface. Usually 3 persons looking in designated sectors ahead of the moving boat keep an account of dolphin sightings, taking special care not to have double counts. This is a tedious manual process that is prone to error, more so as the dolphins are underwater for more than 90% of the time and come to the surface only very briefly for breathing. This visual method may not provide us the exact population of dolphin in a particular area of the Ganges River due to human errors, weather conditions and visibility. Underwater detection of dolphin clicks, on the other hand can be done with hydrophones that can serve a major complementary role for obtaining more reliable dolphin counts.

IVAS addresses this problem by novel use of technology and unique design to accurately monitor Ganga dolphin numbers and habitat. IVAS is an integrated system of high definition cameras to survey the river surface, and hydrophones for underwater acoustic based click observation. The hydrophones are tuned to detect the bio-sonar clicks of dolphins as they forage for food. Figure 12 demonstrates IVAS concept with the help of 3 panels. The uppermost panel shows the situation where the dolphin is underwater, emerges out of water surface and re-submerges. The central panel shows how it would be captured by a video camera, only when it briefly surfaces. The lowermost panel shows the bio-sonar clicks in the frames prior to and after re-submergence. In this manner, IVAS may be considered to be a complementary use of cameras and hydrophones to detect dolphins.

The IVAS system consists of a processor that uses the signals from the hydrophones to determine direction and distance of the dolphin. To accomplish its task, the IVAS system consists of the following sensors:

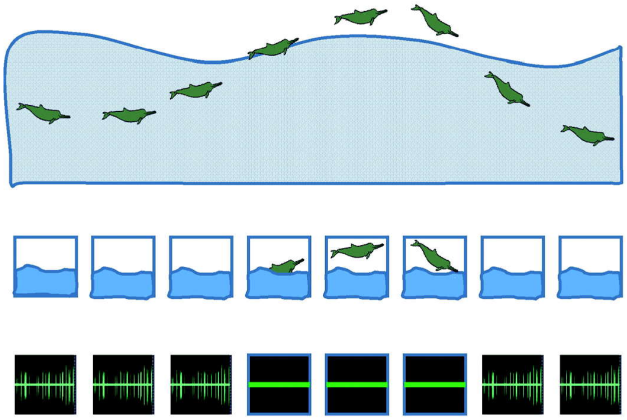


Fig. 12. 12 IVAS concept

- A set of 3 video cameras for surface surveillance that can also be expected to replace human observers for more consistent counts.
- Acoustic System of typically of 3 hydrophones (underwater sensors) for underwater detection of the dolphin clicks and to localize the dolphins.

4.2 Dolphin encounter with IVAS

We demonstrate the power of IVAS concept from an episode where an actual lone dolphin was observed both by the cameras and by the hydrophones. These results are discussed below.

The snapshots shown in the two panels in Figure 13 below are the observations from cameras and the hydrophone data. Clips in the upper panel are taken from the video cameras for the above-water surface surveillance and the clips in the lower panel are the simultaneous underwater hydrophone observations on an oscilloscope. They depict the following:-

- The first clip shows the situation of presence of a dolphin under the water surface which is confirmed by the presence of dolphin clicks recorded by underwater sensors.
- The second clip shows the presence of dolphin on the surface of water (seen inside the red circle) as observed and recorded by the camera, while dolphin clicks are no longer seen.
- The third clip shows the situation where the dolphin dives back into the water which can be seen by absence of it on the surface and re-appearance of the dolphin clicks recorded by underwater sensors.

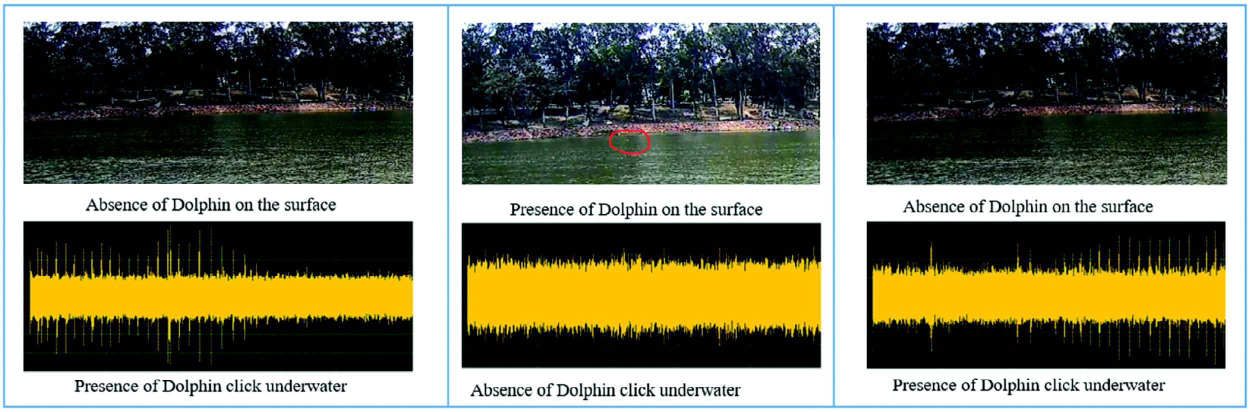

Absence of Dolphin on the surface
Presence of Dolphin click underwater
Presence of Dolphin on the surface
Absence of Dolphin click underwater
Absence of Dolphin on the surface
Presence of Dolphin click underwater

Fig. 13. Demonstration of IVAS concept with real data

4.3 Implementation of IVAS

The IVAS consists of a triangular array of 3 hydrophones that receive underwater dolphin clicks (Figure 14). The hydrophones are configured nominally as an isosceles triangular with base 3m and height 5m. It has been noted in our studies (section 2.2, 3.2) that the least ICI of the dolphins is about 10 msec. The maximum expected propagation time delay between hydrophones of the IVAS array for chosen dimensions is under 5 msec which ensures that two adjacent clicks on the hydrophones are unambiguously analysed without any overlap with the next click.

A given source will be observed at times T1, T2, T3 at hydrophones H1, H2 and H3 respectively (refer Figure 14). The observed inter-hydrophone delays are computed for localization of source: the source would lie on the equi-delay locus represented as a hyperbola. Thus, hyperbolas can be drawn for each pair of hydrophones, the intersection gives the actual location of source. Figure 14 shows two such hyperbolas H32 and H31 passing through the source.

The receiver (Fig. 15) processes the hydrophone signals which are amplified and then digitized in a data acquisition system (NI CDAQ-9134). The digitized data is transferred to a laptop PC for recording

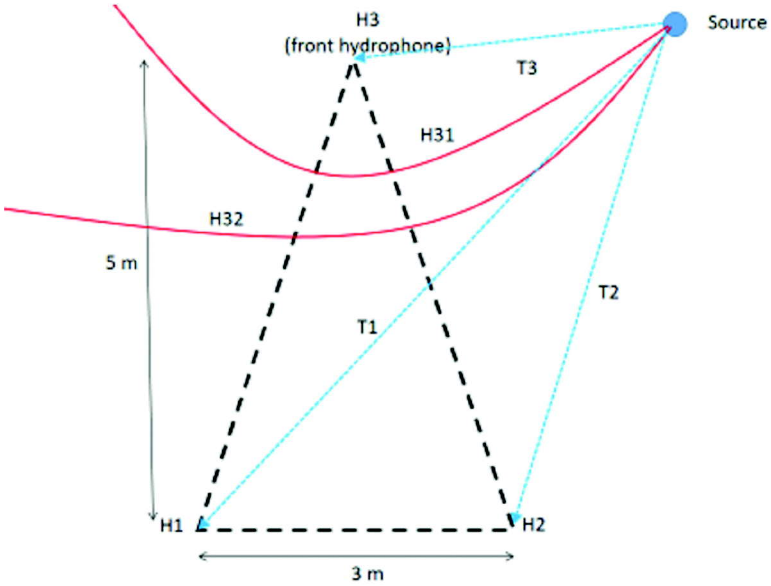


Fig. 14. IVAS triangular hydrophone array geometry showing intersecting equi-time delay hyperbolas for localization of source

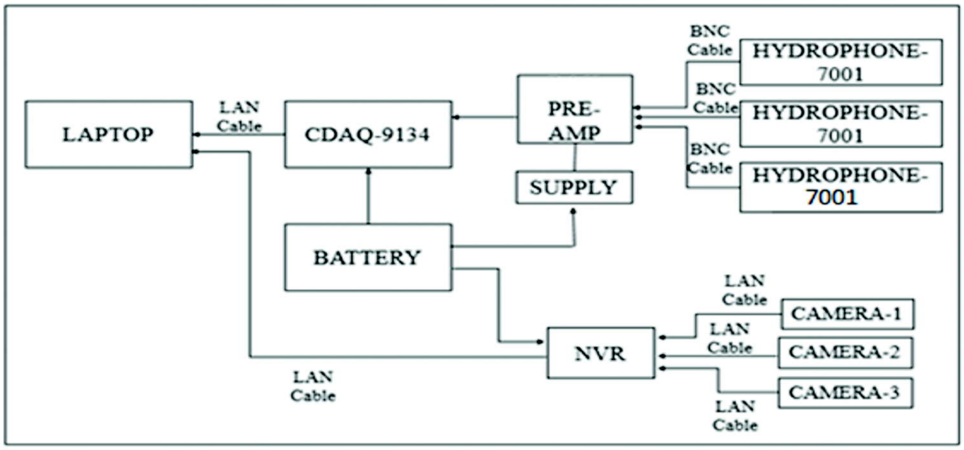
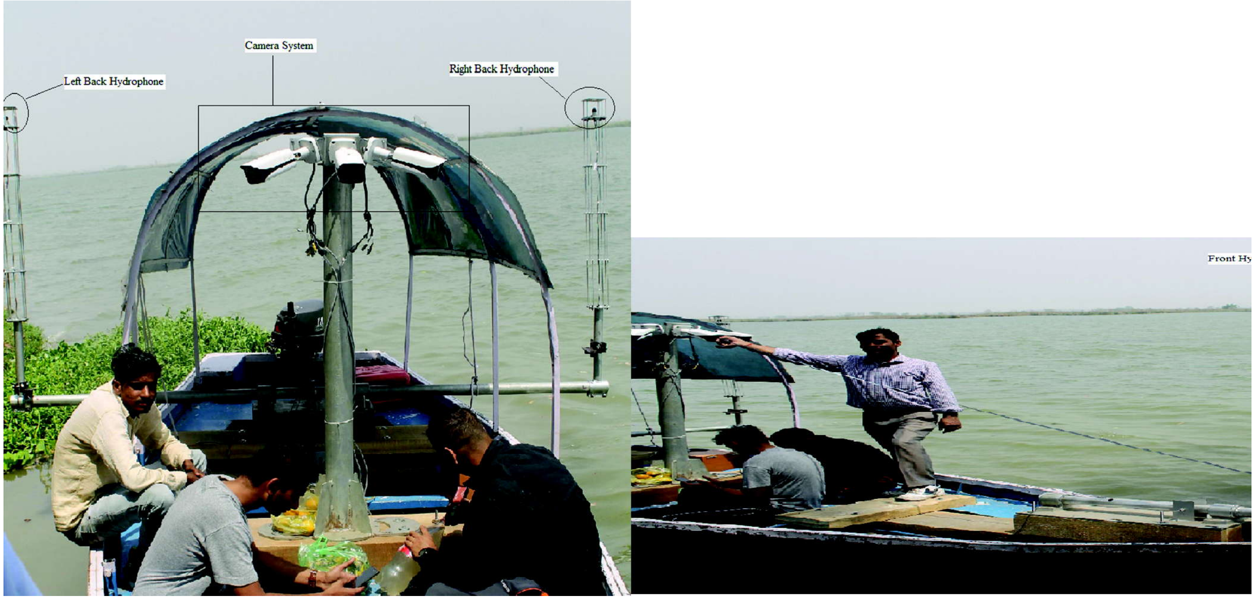


Fig. 15. IVAS receiver block diagram

and analysis. The three video cameras for observing the river surface are connected to a Networked Video Recorder (NVR) and also connected to the laptop PC for control and analysis. The laptop PC is used to monitor the audio and video signals and to store the data. The complete system is designed to run on battery.

Acoustic files from CDAQ are stored with time stamp with 1 microsecond resolution and are saved on every minute basis. The time delays of the click signals between the hydrophones are extracted that are then used to estimate of Range and Bearing of the dolphin clicks as explained in Figure 14. A database of time delays, range, bearing of the source up to the desired range was first created offline for the IVAS array. The extracted time delays from the actual click data are matched with this database to get the best estimate of range and bearing angle and hence the source location. The IVAS system as installed on a boat is shown in Figure 16. The three video cameras are set for identical FOV and pixel resolution.



(a) (b)

Fig. 16. IVAS hydrophone array: (a) Rear two hydrophones and cameras, (b) front hydrophone

Acquisition of videos and image data is done through NVR. The NVR clock is synchronized with that of the CDAQ to enable obtaining relevant frames from the video data to confirm the presence of a dolphin as and when required.

4.4 Field calibration of IVAS hydrophone array localization

Field calibration of IVAS was conducted to check its range and accuracy. In order to calibrate the system we used two boats so that we could transmit dolphin-like sounds from one and receive them at the other boat equipped with IVAS (Fig. 17). Transmission of signal pulses is done from various distances from the receiver boat that receives the incoming signal pulse on the three hydrophones. During this test, the precise distance of the transmitter boat was found using a Laser Rangefinder.

Fig. 17. IVAS Calibration trials with laser range-finder

The laser range finder gives the actual visual distance, while the underwater acoustic range from the acoustic data is estimated using the IVAS hydrophone array system as discussed above. The table below shows the compiled results of validation of distance between 20 metres to 260 metres, assuming a nominal speed of sound of 1500 m/sec.

Table 1. Validation of visual and acoustic range by IVAS.

Visual Range from Laser Range Finder (metres)	Acoustic Range estimated from IVAS hydrophone data (metres)
20	26.9
50	53.14
97	102.39
150	152.97
200	209.99
260	260.92

The relatively small errors of only a few meters between the laser range finder and acoustically calculated range are due to constant drift between source and transmitter in the river's current, and use of an approximate sound speed in water. This study has adequately demonstrated that the algorithm for determining dolphin range from acoustic data matches very well with the visual range even for distances as large as 250 metres. So, in principle, dolphins may be detected over as much as half a kilometer diameter of the river. This system can be deployed over extended periods to study dolphin movements at fixed locations, or even while moving the boat at low speeds. At higher speeds, large engine noise becomes a problem by obscuring dolphin clicks. In the next section, we show how we have modified the IVAS array to receive signals only from the front sector of the boat and hence avoid engine noise. This modification is well-suited for counting dolphins from a moving boat.

5. ACOUSTIC BASED DOLPHIN CENSUS USING IVAS

This section provides a summary of the dolphin census feasibility trial using a modified IVAS conducted on 21st December 2018 near Narora, UP. The field trial was conducted to validate the technique developed for acoustic based census with visual confirmation of dolphins. The experiment was conducted between Basi Ghat and Rajghat in Narora. Several instances were encountered where dolphin was briefly captured visually on the surface by the cameras while it was tracked acoustically for several minutes.

5.1 Principle of Acoustic Census

A pair of hydrophones is assembled and placed in the front of the boat as shown in Fig. 18. The separation between the two front in-line hydrophones (H1 and H2) is kept as 50 cm that would give a maximum inter-hydrophone propagation time delay of about 330 microseconds. Two air filled bottles are placed behind H1 and H2 hydrophones and act as an

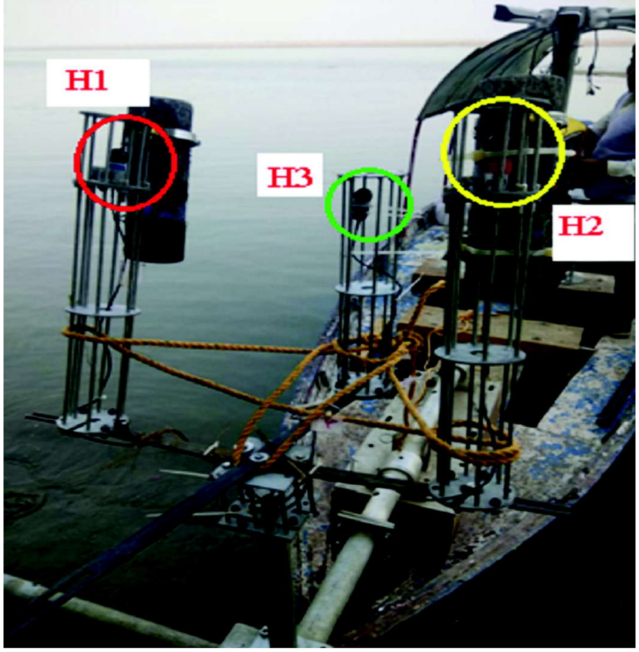


Fig. 18. Assembly of cameras and hydrophones with air-filled bottles (array lifted out of water)

acoustic discontinuity to prevent engine noise from reaching the hydrophones. In order to properly detect and count the dolphins in the front sector of the boat, the boat moved at a speed of 8-9 km per hour. A third hydrophone H3 without a backing bottle was used only as a reference hydrophone to demonstrate the effectiveness of the air bottles. In addition, three cameras were also deployed covering a forward 180 degree field view, 60 degree sector coverage of each camera, replacing three human observers used in a conventional visual census exercise.

The time delay between H1 and H2 represents the angle of the dolphin from the boat heading: 0 delay refers to the front of the boat, while negative and positive values refer to either side of the boat (Figure 19). A typical time delay plot is shown in Figure 20. Noise spikes mark as random dots, while dolphin clicks follow a pattern of delays (in red circle) corresponding to the consistent change in angle due to relative motion of boat and dolphin. The dolphin click region can be zoomed to study the delay track.

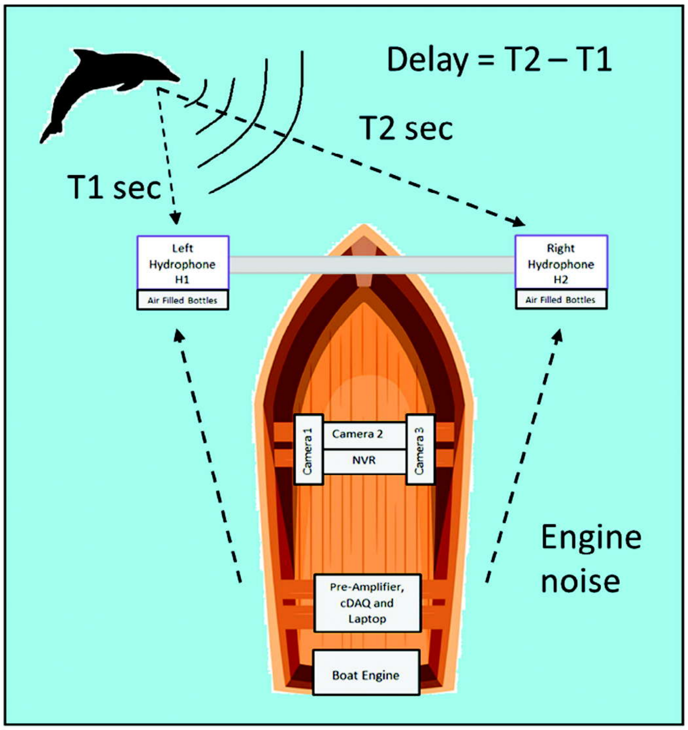


Fig. 19. Time Delay measurement gives direction of dolphin click

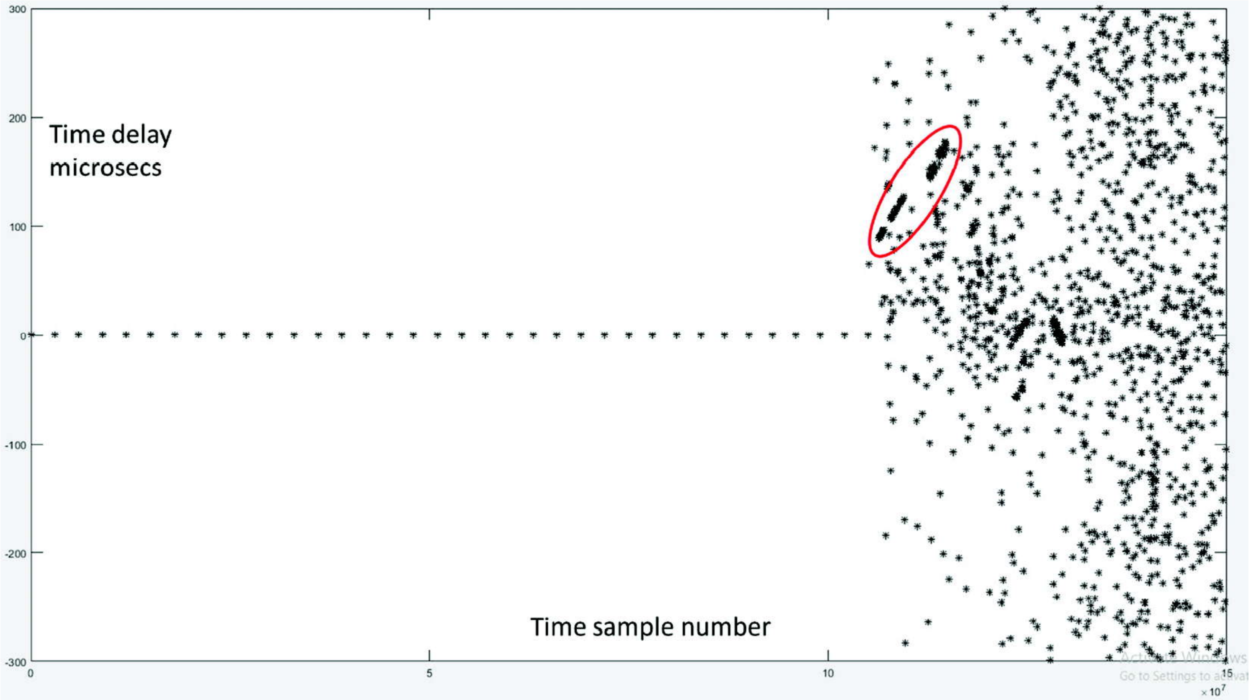


Fig. 20. Plot of observed time delay between H1 and H2 versus time

6. RESULTS AND OBSERVATIONS

The 3 hydrophone signals and video from 3 cameras were recorded on 21st December 2018 at Narora. There were several instances when dolphins were encountered both visually and acoustically. Two of these encounters are discussed below.

Acoustic recording of an instance where dolphin clicks were observed is given below in Figure 21 which shows signals from the three hydrophones.

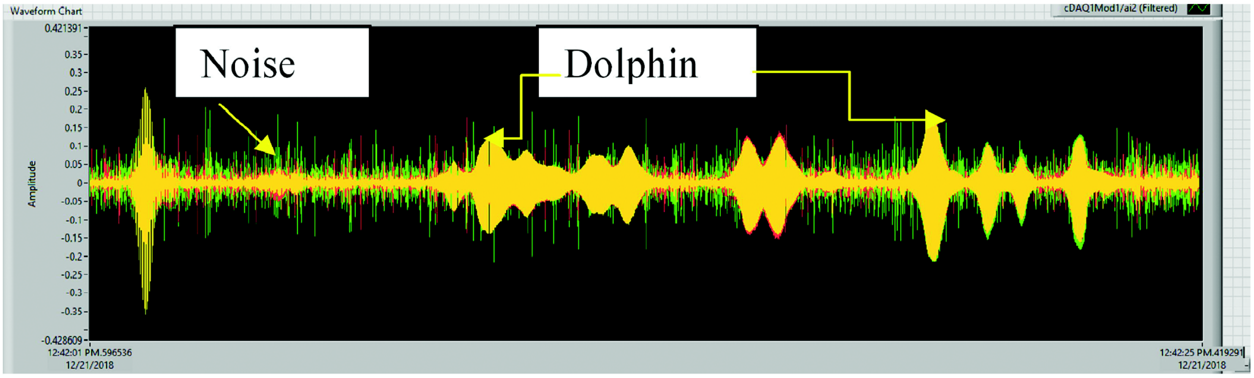


Fig. 21. Acoustic recording depicting dolphin clicks between 12:42:01pm and 12:42:25 pm

Here, the red, yellow (overlapping) and green waveforms represent H1, H2 and H3 hydrophone respectively. It can be observed from the above image that the signals recorded by H3 (green) have more noise compared to those recorded by H1 and H2. This shows that by using air-filled bottles, the engine noise is reduced substantially. The variation in the click envelope is due to the meandering movement of the narrow dolphin beam pattern relative to the hydrophone location.

While travelling downstream with a speed of 8-9 km/hr, at around 12:48 pm a dolphin was captured both acoustically and in the camera as shown below (Figures 22 and 23b). The surfacing time (when the dolphin comes out of the water surface to breathe) was between 12:48:30.211111 and 12:48:31.149289, which is about 1 second duration and indicated by absence of clicks. The time offset between the independent acoustic and video recordings was been obtained and corrected. The synchronized time is mentioned on each camera image (Fig 23b).

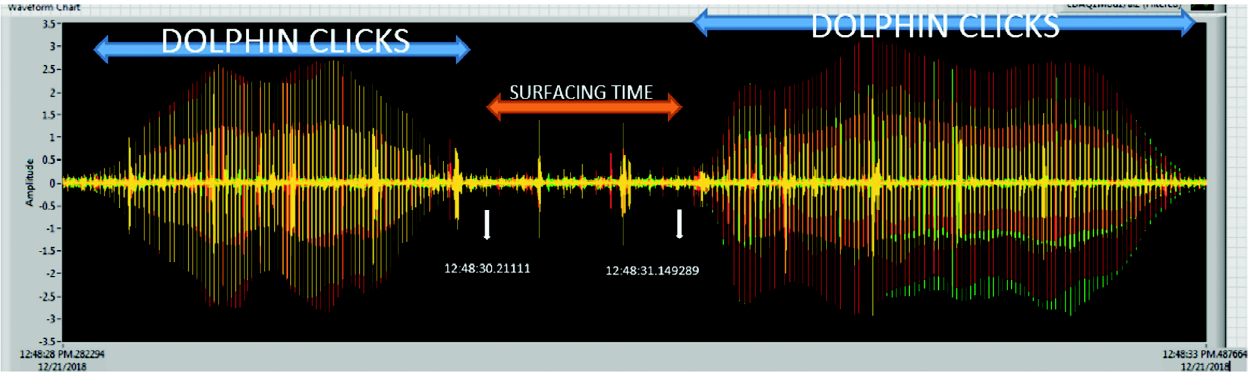
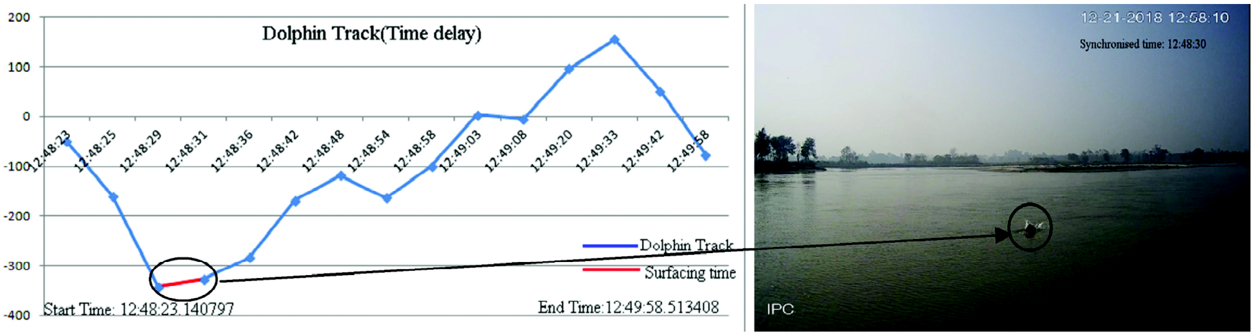


Fig. 22. Acoustic recording depicting dolphin click and surfacing time

The zoomed dolphin track (time delay, Figure 23a) shows a plot of delay vs time, between H1 and H2. X-axis represents time in seconds and Y-axis represents delay in microseconds. Not all clicks have

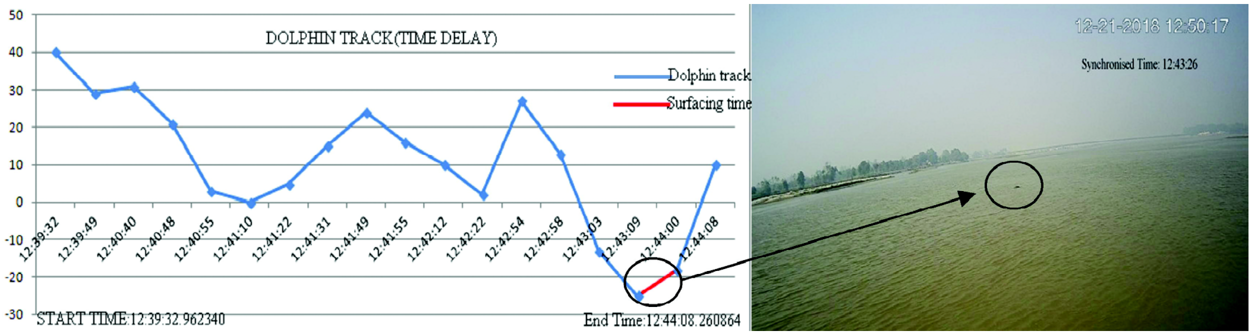


(a) (b)

Fig. 23. (a) Plot of the change in delay (track of 1 min 35 secs) and the highlighted red part here depicts the surfacing time starting from 12:48:30.211111 and ending on 12:48:31.149289 (b) Video grab of left camera showing dolphin in sight

been marked in the plot for clarity. Figure 23b shows the video grab of the surfaced dolphin corresponding to the time that clicks are absent, indicated by the red line in the plot. The clearly defined pattern in the time delay plot can be used to detect the dolphins. Such patterns of time delays will be seen for each dolphin that is encountered, thus we can estimate the number of dolphins by simply counting these dolphin tracks as the boat moves through the river. Double counting is avoided since the boat moves faster than the dolphin, which is soon left behind.

Another instance of dolphin encounter with corresponding delay plots is shown in figure 24 below. Once again, the dolphin surfaces at the time indicated by the red line in the plot.



(a) (b)

Fig. 24. (a) Plot of the change in delay (track of 4 mins 36 secs) and the highlighted part here depicts the surfacing time starting from 12:43:25 and ending on 12:43:28 (b) Video grab of front camera showing dolphin in sight

The trial has successfully demonstrated that IVAS can be used for dolphin census. The practical concern of blocking the engine noise using air-filled bottles adjacent to the hydrophones has been also addressed. The presence of dolphins was recorded both visually and acoustically. We have been able to validate the acoustic-based counting method with confirmatory dolphin sighting obtained in the video cameras while moving at a speed of about 8-9 km/hr. While acoustic tracks last over 1 minute, the sightings are very brief, under 3 seconds only. This method can, therefore, be effectively used for conducting dolphin census by non-experts since the dolphins can be more objectively counted from the recorded data, both visually and acoustically.

7. CONCLUSIONS AND ACKNOWLEDGMENTS

This paper has reviewed the pioneering work done to unravel the bio-Sonar characteristics of Ganges dolphin, their underwater movement behaviour and migration patterns. Innovative indigenous technology developed for the study and census of the dolphins using both acoustic and visual monitoring has been demonstrated. The initial part of the paper covered up to section 3 encompasses contributions of several organisations, namely The University of Tokyo, KDDI R&D labs Japan, World Wildlife Fund-India, and Chilika Development Authority. In this regard the author gratefully acknowledges prominent individual Japanese researchers: Professor Tamaki Ura, Ms Harumi Sugimatsu, Mr. Junichi Kojima, and Mr. Tetsuo Fukuchi. The second part of the paper on IVAS has been funded under CSR programme of NTT-Data (India) and was ably supported in the field by NGO M/s GANGA Sansthan headed by Dr. Vivek Mishra. Contribution of Dr. Kapil Dev Tyagi of Jaypee Institute of Technology is gratefully acknowledged.

8. REFERENCES

[1] T. Ura, R. Bahl, M. Yano, T. Inoue, T. Sakamaki and T. Fukuchi, 2006. Results from a high-resolution acoustic device for monitoring finless porpoises in coastal precincts off-Japan. *Proc IEEE Intl Conf on Oceanic Engg* (OCEANS ASIA-PACIFIC 2006), Singapore, 16-19 May 2006.

[2] R. Bahl, Harumi Sugimatsu, Junichi Kojima, Tamaki Ura, Sandeep Behera, Tomoki Inoue and Tetsuo Fukuchi, 2007. Beam pattern estimation of clicks of a free-ranging Ganges river dolphin. *Proceedings of IEEE Intl Conf on Oceanic Engg* (OCEANS 2007), Vancouver, Canada, September 29-October 04, 2007.

[3] H. Sugimatsu, T. Ura, J. Kojima, R. Bahl and S. Behera, 2008. Underwater behavioral study of Ganges river dolphins by using echolocation clicks recorded by 6-hydrophone array system. *Proceedings of IEEE Intl Conf on Oceanic Engg* (OCEAN'S 2008), Quebec, Canada, Sept 2008.

[4] H. Sugimatsu, J. Kojima, T. Ura, R. Bahl, Sandeep Behera and Vivek Sheel Sagar, 2011. Annual Behavioral Changes of the Ganges River Dolphins (platanista gangetica) Based on the Three Long-Term Monitoring Seasons using 6-Hydrophone Array System. *Proc. UT&SSC11*, Tokyo, Japan, 2011.

The Journal of Acoustical Society of India : Vol. 47, No. 2-3, 2020 (pp. 71-78)

Analysis of source signal and vocal tract for detection of out-of-breath speech

Sibasis Sahoo and Samarendra Dandapat
Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati, Guwahati, India
e-mail: sibasis2016@iitg.ac.in

[Received: 25-03-2020; Revised: 17-05-2020; Accepted: 17-05-2020]

ABSTRACT

Physical exercise induces fatigue on the human body. It leads to changes in the breathing pattern. The speech produced under this condition has been termed as out-of-breath speech. The air coming out of the lungs act as the source for generating sounds. Hence, the speech sound appears different from that of the normal state. In this work, the focus is on the analysis of the distinctive contribution of the source and the vocal tract (VT) under physical exertion. Harmonic peak to energy ratio (HPER) features are extracted from the integrated linear predicted residual (ILPR) signal, which is an approximation for the source signal. The frequency response of the VT filter is rendered in terms of Linear prediction (LP) coefficients. Both the source and the VT features are statistically evaluated under the normal and the out-of-breath states. The source signal is found to get significantly affected under physical exertion than the VT filter. A Gaussian Mixture Model (GMM) classifier validates the above observation where source-based HPER features give a better classification rate of 85.6% compared to 68.2% for the VT filter parameters.

1. INTRODUCTION

The speech signal is an acoustic signal that inherently captures the characteristics of both the system and the source that produce it. The system is the VT and the source being the air expelled from the lungs. The air, coming out of the lungs, gets modulated by periodical vibrating action of vocal folds which, in turn, causes the voiced sounds[1]. Similarly, the unvoiced sound is produced when the vocal folds are relaxed, and the turbulent air from lungs gushes through open vocal folds. A prior message, formulated at the human brain as a cognitive task, is communicated through a combination of voiced and unvoiced sounds. Hence, speech production is the result of a complex neuromuscular process. It requires a synchronised functioning among the cognition process, the respiratory process, phonatory process and articulatory process[2]. Any minor change in a speaker's physical and cognitive state can influence the speaker's ability to control the speech production system. It is this complexity that makes the speech signal a suitable marker for different health conditions[1] [3], emotional state[4], [5], speaker verification and recognition[6].

When a person performs physical exercise, the need for more oxygen makes the person breathe faster and deeper. In literature, this physical state of the person is called the out-of-breath state[7]. Inhalation period becomes shorter, and the exhalation period becomes longer. If any attempt is made to speak, the produced sound appears different from that of the normal state. Fig. 1 shows the waveform and the spectrogram of a sample speech segment for both the normal and the out-of-breath states. Some notable changes are: shortened speech duration, damped higher frequency components in the voiced region, more words are packed in a single breath cycle[8], a higher number of short breathing pauses, increased pitch frequency (F_0)[9], higher glottal open quotient, lower glottal close quotient, and waveshape is skewed positively[10]. These characteristic changes led researchers to detect physical stress level and physical exercise intensity by analysing the speech signal. Godin et al. used six different glottal features to recognise whether exertion caused by physical task stress[9]. In two separate studies, Egorow[11] and Thuong[12] attempted to estimate exercise intensity level under the influence of physical exercise. Deb *et al.*[7] used speech spectral features to determine the level of breathiness in a spoken sentence. In another work, they showed that different emotions accompany with different level of breathiness[5].

From the previous studies, it is evident that the speech signal is perceptually different under physical exertion. As the speech signal is the result of a filtration action of quasi-periodic impulse source by VT filter, both the source and the filter are expected to vary under stress. In this work, the aim is to assess the independent behaviour of the source signal and the VT filter under physical exertion. The integrated linear predicted residual (ILPR) signal[13] is used as an approximation to the source signal as it shows a higher correlation with the derivative of electroglottograph (DEGG) signal[14]. To represent the VT, the popular linear prediction (LP) coefficient representation is considered[15]. The details about ILPR and LP coefficients is elaborated in Section 2. In Section 3, the statistical analysis of the spectral characteristics is described, and finally, the conclusion is drawn in section 4.

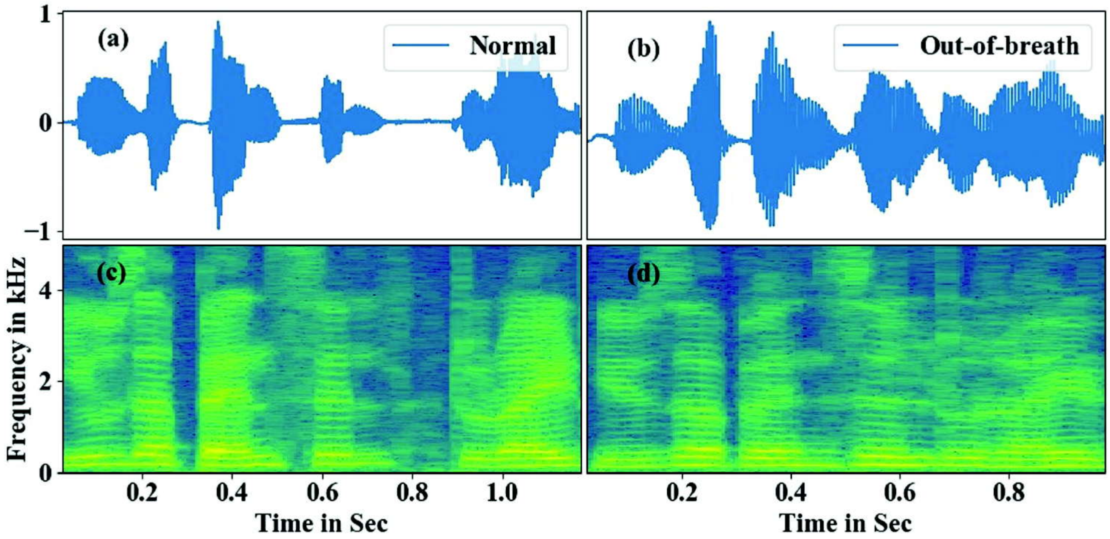


Fig. 1. Spectrogram of normal and out-of-breath speech.

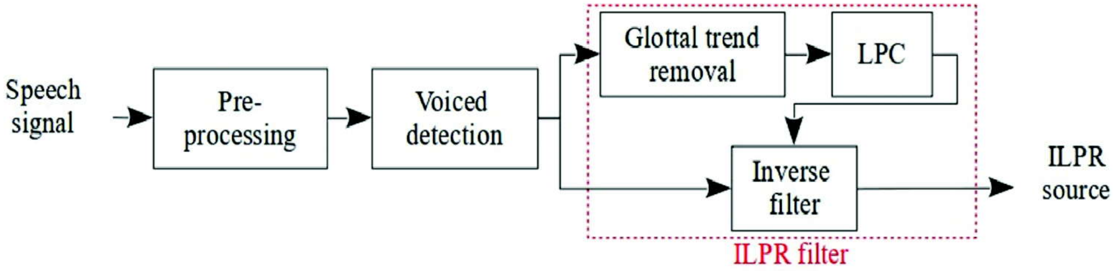


Fig. 2. Block diagram for extracting the ILPR source signal from its corresponding speech signal.

2. METHODOLOGY

Depending on the nature of the glottal source, *i.e.,* quasi-periodic impulse train or noise, the speech becomes voiced or unvoiced, respectively[15]. In this work, our focus is on the voiced regions. Using an energy-based threshold, Th = E_{avg}, the voiced regions are separated from the unvoiced and the silence regions, where E_{avg} is the average frame energy[7]. Now considering a 20ms voiced segment of speech, the ILPR source is estimated by LP-based inverse filtering. The corresponding filter coefficients are stored as representative of the VT that captures the frequency response of the VT filter. The spectral characteristic of the source signal is analysed in terms of the harmonic peaks of its magnitude spectrum. The detailed procedure for analysing the ILPR and the VT is described as follows.

2.1 ILPR Source extraction

ILPR is a source estimated from the speech signal that has a higher correlation with the DEGG signal[14]. ILPR filtering is performed to obtain this source signal. This filtering step involves inverse filtering of the speech signal where the filter coefficients are estimated by LP analysis on the glottal trend removed speech signal[13]. The advantage of ILPR over linear prediction residual (LPR) is that ILPR does not have sharp bipolar peaks at the glottal closing positions. Hence, it appears more like the natural source signal[14], [13]. Fig. 2 shows a block diagram of the extraction of the ILPR source signal from the speech signal. The steps for ILPR extraction begins with DC offset removal and amplitude normalisation belonging pre-processing block. The block following it detects the voiced region from the speech signal using an energy-based threshold. For a 20 ms segment of the voiced region, the glottal trend is removed by passing it through a pre-emphasis filter $H(z)$ = 1 – z^{-1}. The LPC block employs the autocorrelation method to compute linear prediction coefficients (LPC) using the filtered signal segment[15]. The initial speech segment is passed through an inverse filter which has the LPCs as the coefficients of the denominator polynomial. The output of the inverse filter is called ILPR signal.

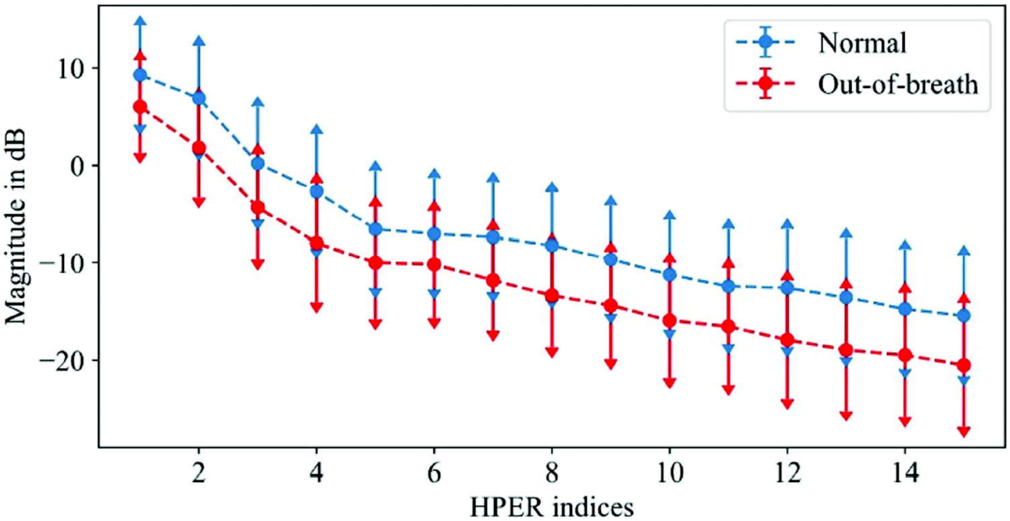


Fig. 3. For the speaker 'SK', the error bar plot shows mean and standard deviation of the HPER features evaluated on the ILPR signal.

2.2 Harmonic peak to energy ratio (HPER)

It is the ratio between the harmonic peak magnitudes and the energy of the segment of speech under consideration[5]. It has been shown that the harmonic structure of the speech signal gets influenced by cognitive load[5], [16]. For different emotions, the contribution of energy by the harmonic peaks are different. In this work, the first fifteen harmonic peaks are taken into consideration. The feature vector is constructed by appending the first fifteen HPER random variables HPER = [$HPER_1$, $HPER_{2,...,}$ $HPER_L$]. Here, L is set to 15. The following procedure is followed for computing HPER feature

(i) A 20 ms segment is considered from the voiced segment of the speech signal at every 10ms interval.

(ii) The segment is then passed through the ILPR filter for extracting the source signal.

(iii) Autocorrelation is then performed on the ILPR source to find out the fundamental period (T0).

(iv) The magnitude spectrum of the ILPR source is estimated using 1024 point discrete fourier transform (DFT).

(v) The first harmonic peak (H1) is obtained by picking the peak within ±5% range of fundamental frequency (f0) where

$$f_0 = \frac{1}{T_0} \tag{1}$$

(vi) Other harmonics peaks are obtained within ±5% range of f_0 at every ithmultiple of f_0, where i = 2, 3..., L.

(vii) Harmonic magnitudes are normalised by the total energy of that signal segment.

$$HPER_i = \frac{H_i}{E} \tag{2}$$

$$E = \sum\nolimits_{n=0}^{N} x_m^2(n) \tag{3}$$

where, i = 1, 2,..., L and n = 1, 2,..., N. $x_m(n)$ is the m^{th} segment of the speech signal of size N. E is the total energy in that segment.

2.3 Vocal tract information

The LP analysis based method has been used for extracting the vocal tract spectral information from the speech segment. It is a widely used method for parameterising the VT contribution of the speech signal[15], [17]. As per the LP procedure, each sample of a signal segment $x_m(n)$ is an approximation by a linear combination of a fixed number of past sample values. This approach allows the signal segment to be modelled by an all-pole filter. These coefficients of this filter are the manifestation of the VT in its parametric form. This kind of representation helps in portraying the spectral characteristics of the VT with infinite resolution in frequency. Here, the order of the all-pole filter is set to 14. It is as per the convention of sampling frequency fs in kHz plus four[13].

3. STATISTICAL ANALYSIS

The out-of-breath stress speech corpus[7] has been used for analysing the effect of physical stress. The features: LP coefficients, as well as the HPER features, are extracted as per the procedure given in Section 2. The mean values of these features for the normal and the out-of-breath states show their respective trend. Gaussian mixture model (GMM) classifier is used to determine separability of the two cases using the spectral features. The classification ability is quantified by three performance measures: specificity, sensitivity and accuracy. A five-fold cross-validation approach is followed where the whole data set is split into five sets of equal size; one set is used for testing and others for training. The procedure is continued for five times by changing the test set. The final result is expressed as the average of the measures for the five iterations.

3.1 Out-of-breath corpus

The out-of-breath corpus has been recorded by Deb *et al.* for assessment of breathiness in the speech signal under physical exercise condition[7]. It contains speech signal data recorded under three stress conditions: normal, out-of-breath and low-out-of-breath. Eight male and five female participants took part in the recording process. They were asked to utter a set of twenty English sentences under three stress conditions. The Out-of-breath speech was recorded after the speaker underwent jogging for six minutes. With a resting period of one minute, speakers repeated the sentences which were grouped as low-out-of-breath. The normal set of speech signal was recorded earlier when there was no physical effort by the speakers. The corpus contains 947 speech samples: 314 normal, 316 low-out-of-breath and 317 out-of-breath cases. In this work, speech samples under normal and out-of-breath conditions are taken into study. The speech samples are down-sampled to 10kHz for analysis.

3.2 Gaussian Mixture Model (GMM) Classifier

GMM treats each class of data as a mixture of Gaussians. Thus, it attempts to capture the variability in VT shape and glottal flow probabilistically[15]. Each class of stressed speech is represented by a gaussian mixture model $\lambda = \{\varphi_i, \mu_i, \Sigma_i\}$, where φ_i is the weight of the ith component in the mixture; μ_i and Σ_i are the component mean vector and covariance matrix respectively. Here, the mixture components are assumed to have full covariance. The model parameters are estimated by expectation-maximisation (EM) algorithm[18]. The classification performance is measured by computing *maximum a posterior probability (MAP)* estimate. An utterance belongs to that class for which it shows highest *MAP value*. The MAP is estimated as :

$$P(\lambda_j / X) = \frac{P(X / \lambda_j)P(\lambda_j)}{P(X)} \tag{4}$$

where, $j = 1,2,...,C$; $X = [X_0, X_{1,...}, X_{M-1}]$ is the collection of feature vectors; $P(\lambda_j)$ is the a priori probability; $P(\lambda_j | X)$ and $P(X | \lambda_j)$ are the posterior and likelihood probability densities respectively. In this work $C = 2$ forperforming binary classification.

4. RESULTS AND DISCUSSION

Fig. 3 shows the error plot corresponding to the HPER features extracted from the ILPR source signal. It plots the variation of mean and standard deviation values for the HPER features. For another two speakers JF and SC, the Fig. 4 shows probability density functions of the third and eighth HPER features extracted from the ILPR source. From the figure, it can be observed that the average contribution of energy by the harmonic peaks is lower in case of the out-of-breath state than that of the normal state. A similar result is seen when the ILPR based HPER features are considered on the whole corpus as given in Table 1. It shows that all the HPER features, except the HPER$_1$, have lower mean values. This characteristic indicates that the average energy contribution of the harmonics to the speech signal is lower for out-of-breath state than the normal state. A similar result can be observed for the HPER features extracted from the speech signal. Table 2 shows the mean for HPER features extracted directly from the speech signal

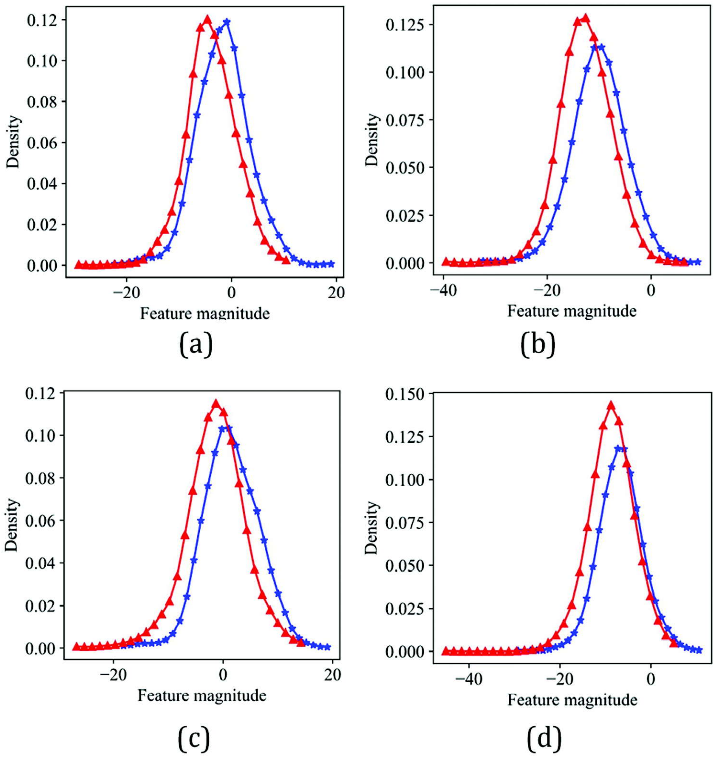


Fig. 4. Plot of probability distribution functions for the ILPR based HPER$_3$ and HPER$_8$ features for speaker JF in (a and b) and SC (c and d), respectively. The normal and the out-of-breath states are in blue and red color, respectively.

Table 1. Mean values (in dB) for the fifteen HPER features for ILPR source signal. Acronyms are N: Normal, OBS: out-of-breath.

Index →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
N	10.3	4.9	0.2	-2.4	-4.7	-6.5	-7.6	-9.0	-10.0	-11.1	-12.2	-13.1	-13.8	-14.5	-15.0
OBS	10.4	3.7	-0.6	-3.3	-5.5	-6.9	-8.0	-9.7	-11.0	-12.2	-13.0	-13.9	-14.5	-15.2	-15.8

Table 2. Mean values (in dB) for the fifteen HPER features for speech signal. Acronyms are N: Normal, OBS: Out-of-breath.

Index →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
N	-3.7	-3.1	-7.6	-13.0	-18.3	-22.5	-25.1	-27.2	-29.0	-30.4	-31.9	-33.2	-34.4	-35.8	-37.0
OBS	-3.2	-3.4	-8.7	-15.8	-21.5	-24.7	-26.7	-29.3	-31.3	-32.7	-33.7	-35.1	-36.6	-37.8	-38.6

considering the entire corpus. Observing the mean values for both the ILPR source and the speech signal, we can imply that the behaviour of the harmonics of the source is manifested in the harmonics of the speech signal. From table 4, it can be seen that ILPR based HPER features have a higher classification rate of 85.2% than that of the features extracted from the speech signal with an accuracy of 81.6%. The reduction in accuracy may be due to the spectrum colouration provided by the VT.

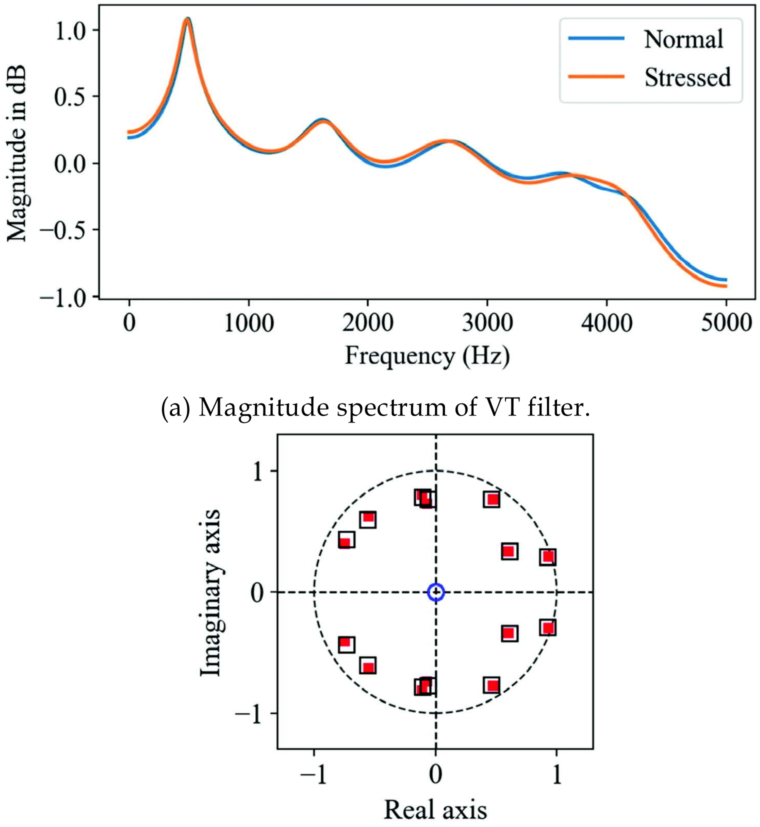
The above results give evidence that the spectral properties of the source signal vary from the normal to the out-of-breath state. On the other hand, to understand the behaviour of the VT under the out-of-breath and the normal states, LPC based representation of the VT is considered. Table 3 shows the mean values of the LP coefficients computed over the entire corpus. Here, the mean values are rounded off to the second decimal point. The corresponding magnitude spectrum and, the pole-zero plot is shown in Fig. 5. Both the plots show that the characteristics are quite similar, indicating that the VT is not much influenced in an average sense under the normal or the out-of-breath states. The mean square error *(mse)* is computed separately for the magnitude spectrum of the VT and the mean values of the HPER peaks for both the states. The *mse* is found to be 0.12 for the VT, which is five times smaller than the *mse* value of 0.61 computed for the fifteen source harmonics.

Table 3. Mean values for LP coefficients. The acronyms are N: normal, OBS: out-of-breath.

Index →	1	2	3	4	5	6	7	8	9	10	11	12	13	14
N	-1.01	0.72	-0.78	0.69	-0.60	0.60	-0.50	0.55	-0.18	0.29	-0.24	0.15	-0.09	0.07
OBS	-1.08	-0.80	-0.87	0.77	-0.72	0.76	-0.63	0.63	-0.26	0.30	-0.24	0.15	-0.09	0.07

Table 4. GMM classification result using LPC and HPER features. The HPER features are extracted from ILPR and speech signals, respectively. All values are in %.

Features	Specificity	Sensitivity	Accuracy
LPC	70.20	66.67	68.4
$\text{HPER}_{\text{ILPR}}$	91.84	78.82	85.2
$\text{HPER}_{\text{speech}}$	74.28	89.02	81.6
$\text{HPER}_{\text{speech}}$ + LPC	89.79	87.45	88.6
$\text{HPER}_{\text{ILPR}}$ + LPC	90.20	91.76	91.0



(a) Magnitude spectrum of VT filter.

(b) The symbols ○, ■ and □ represent the zeros, the poles for normal and the oles for out-of-breath cases respectively.

Fig. 5. Magnitude spectrum and pole-zero plot for VT filter with respect to LP coefficients averaged over normal and out-of-breath samples respectively.

The *mse* values suggest that the source experiences higher variation under physical exertion than the VT. Table 4 shows the classification accuracy using LPC. It is found to be 68.4%. This is quite low compared to the HPER features extracted from the ILPR source or directly from speech.

The accuracy results in Table 4 suggest the spectral characteristics of the ILPR signal have larger changes under physical exertion. Thus, the source signal is more capable of separating the normal and the out-of-breath speech with an accuracy of 85.6%. The spectral features of the VT represented by LP coefficients seem to have moderately affected due to physical exertion. Thus, classification accuracy is found to be 68.4%. However, the highest classification accuracy of 91% is obtained by combining the spectral features derived from both the source signal and the VT. This result is higher than the spectral features derived directly from the speech signal.

5. CONCLUSION

In this work, the effect of physical exertion on the source signal and the VT is analysed. The source signal is estimated from the speech signal by the ILPR filtering, and the LP coefficients parameterise the VT. The spectral analysis of the ILPR signal showed that the energy contributed by harmonic peaks is different for the normal and the out-of-breath states. The harmonics at higher frequency have a lower average magnitude in case of the out-of-breath state. Using GMM classifier, the ILPR based HPER features showed an accuracy of 85.6% for classifying the normal speech from the out-of-breath speech. This is better than the speech-based HPER features. On the other hand, the analysis of VT using LP coefficients does not show any major changes. The GMM classifier shows a moderate accuracy of 68.4% for distinguishing the state of a speech signal using LP coefficients. This implies that the effect of physical exertion does not have any significant influence on VT. For performing classification and determining the state of a speech sample, the combination of source and VT features are found to give the highest accuracy of 91%.

6. REFERENCES

[1] N. Cummins, A. Baird and B. W. Schuller, 2018. "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*.

[2] B. Conrad and P. Schönle, 1979. "Speech and respiration," *Arch. f¨or Psychiatr. und Nervenkrankheiten,* **226**, 251-268.

[3] S. Deb, S. Dandapat and J. Krajewski, 2017. "Analysis and Classification of Cold Speech using Variational Mode Decomposition," *IEEE Trans. Affect. Comput.,* pp. 1-1.

[4] C. T. Ishi, H. Ishiguro and N. Hagita, 2010. "Analysis of the Roles and the Dynamics of Breathy and Whispery Voice Qualities in Dialogue Speech," *EURASIP J. Audio, Speech, Music Process.,* **2010**, pp. 1-12, jan 2010.

[5] S. Deb and S. Dandapat, 2016. "Classification of speech under stress using harmonic peak to energy ratio," *Comput. Electr. Eng.,* **55**, 12-23.

[6] G. Senthil Raja and S. Dandapat, 2010. "Speaker recognition under stressed condition," *Int. J. Speech Technol.,* **13**, 141-161.

[7] S. Deb and S. Dandapat, 2017. "Fourier model based features for analysis and classification of out-of-breath speech," *Speech Commun.,* **90**, 1-14.

[8] J. Trouvain and K. P. Truong, 2015. "Prosodic characteristics of read speech before and after treadmill running," in 16th *Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 3700-3704.

[9] K. W. Godin, T. Hasan and J. H. L. Hansen, "Glottal Waveform Analysis of Physical Task Stress Speech," tech. rep.

[10] S. Sahoo and S. Dandapat, 2019. "Analysis of Speech Source Signals for Detection of Out-of-breath Condition," in Int. Conf. Pattern Recognit. Mach. Intell., *Springer*, pp. 418-426.

[11] O. Egorow, T. Mrech, N. W. P. I. and U. 2019, "Employing Bottleneck and Convolutional Features for Speech-Based Physical Load Detection on Limited Data Amounts," in *INTERSPEECH*, pp. 1666-1670.

[12] K. P. Truong, A. Nieuwenhuys, P. Beek and V. Evers, 2015. "A database for analysis of speech under physical stress: detection of exercise intensity while running and talking," in *Interspeech*, pp. 3705-3709.

[13] A. P. Prathosh, T. V. Ananthapadmanabha and A. G. Ramakrishnan, 2013. "Epoch Extraction Based on Integrated Linear Prediction Residual Using Plosion Index," *IEEE Trans. Audio. Speech. Lang. Processing,* 21, 2471-2480.

[14] N. Adiga and S. R. M. Prasanna, 2015. "Detection of Glottal Activity Using Different Attributes of Source Information," *IEEE Signal Process. Lett.,* **22**, 2107-2111.

[15] T. F. Quatieri, 2006. Discrete-time speech signal processing: principles and practice. *Pearson Education India*.

[16] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang and Lian Li, 2015. "Speech Emotion Recognition Using Fourier Parameters," *IEEE Trans. Affect. Comput.,* **6**, 69-75.

[17] T. Drugman, P. Alku, A. Alwan and B. Yegnanarayana, 2014. "Glottal source processing: From analysis to applications," *Comput. Speech Lang.,* **28**, 1117-1138.

[18] D. G. Duda, Richard O and Hart, Peter E and Stork, 2012. Pattern classification. *John Wiley & Sons.*

The Journal of Acoustical Society of India : Vol. 47, No. 2-3, 2020 (pp. 79-89)

Significance of excitation source information from speech

Vikram C. Mathad[1], Sishir Kalita[2] and S. R. Mahadeva Prasanna[3]
[1]Department of Speech and Hearing Sciences, Arizona State University, Tempe, USA
[2]Armsoftech.air, Chennai, India
[3]Department of Electrical Engineering, IIT Dharwad, Dharwad, India
e-mail: cmvikramshiva@gmail.com

[Received: 28-04-2020; Revised: 19-05-2020; Accepted: 19-05-2020]

ABSTRACT

Quasi-periodic glottal vibrations are considered as the primary source of excitation during speech production. Excitation (glottal) source signal can be estimated by using linear prediction-based inverse filtering operation. Epochs, fundamental frequency, periodicity, the strength of excitation (SoE), and the shape of the glottal pulse are considered as the important attributes of excitation source and used in the analysis of speech. In this article, the importance of excitation source information in various applications of speech processing, namely, enhancement of noisy speech, speaker verification, speech synthesis, and detection of speech disorders are briefly presented. Speech regions anchored around the epochs can be characterized as high signal-to-noise ratio regions, and processed for speech enhancement. The features derived from the estimated excitation source signal are found to be rich in speaker-specific information and used in speaker verification tasks. Inclusion of voicing decision, SoE, aperiodicity in statistical speech synthesizers showed a significant improvement in the synthesized speech quality. The SoE, fundamental frequency, glottal activity regions, and glottis landmarks are used in the clinical applications of speech.

1. INTRODUCTION

Speech is produced by the excitation of a time-varying vocal tract system[1]. The excitation to the vocal tract system may be produced due to the quasi-periodic vibration of vocal folds (voiced sounds), the formation of narrow constriction (fricatives), and abrupt release of the completely constricted vocal tract (plosives)[2]. Glottal vibrations are considered as the primary source of excitation in the speech production system[1],[3]. The extraction of features or parameters from the time-varying speech signal is one of the significant objectives of speech signal processing. Most of the speech-based applications like speaker verification, speech recognition, text-to-speech synthesis, use conventional short-time spectral features, *e.g.,* Mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients. These features are computed by the short-time processing of speech signals using a window size of 10-30 ms[4]. These features capture the vocal tract characteristics; however, the excitation source information is not being explicitly captured.

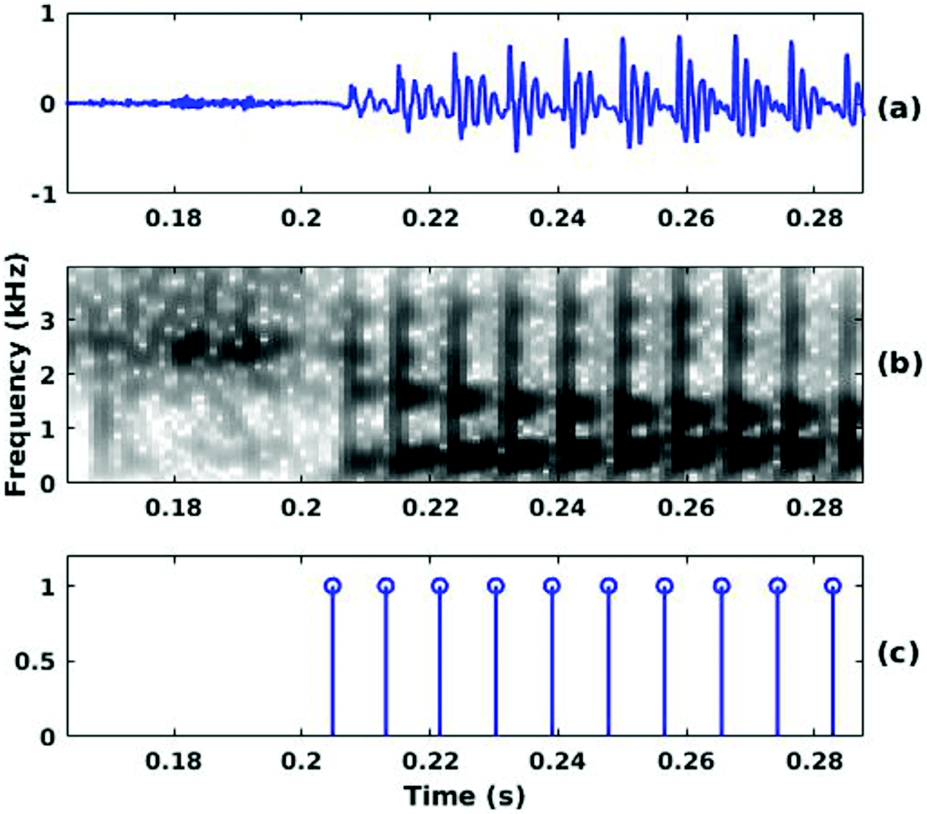


Fig. 1. Significance of excitation source information. **(a)** Speech waveform, **(b)** wideband spectrogram and **(c)** impulse train separated by pitch period.

Mathematically, excitation source information is represented as the train of impulses, which are separated by the pitch period[4]. This impulse train is generally used in speech synthesizers. Fig. 1 demonstrates the importance of excitation source information. The speech signal and spectrogram are shown in Fig. 1(a) and (b), respectively. The impulse train corresponding to the glottal excitation is plotted in Fig. 1(c), where the impulses are separated by pitch period (T_0). This impulse train is considered as the approximation of source in most of the speech synthesizers. However, apart from this, the explicit use of excitation source information is not widely used in speech-based applications.

In the spectrogram (Fig. 1(b)), the vertical striations correspond to the glottal closure instants (GCIs) or epochs[5]. In the spectrogram (Fig. 1(b)), regions around the epochs indicate the presence of a high signal-to-noise ratio (SNR) in terms of relatively darker intensities. The processing of such high SNR regions around the epochs is carried out for the enhancement of noisy speech signals[6] and formant estimation[7]. The anatomy of vocal folds and its vibrating pattern vary among the speakers. Hence, the speaker-specific information derived from the excitation source is used for speaker verification[8],[9]. The strength of excitation (SoE) is another important parameter and its applications are demonstrated in voicing detection[10],[11], pathological voice detection[12] and speech synthesis[13],[14].

This article reviews the importance of excitation source information in different speech processing applications. The paper is organized as follows: The extraction of excitation source information from the speech signal is explained in Section II. The usefulness of excitation source information in speech enhancement and speaker verification systems is described in Section III and IV, respectively. The applications in speech synthesis and clinical systems are mentioned in Section V and VI, respectively. Finally, the paper is concluded in Section VII.

2. EXTRACTION OF EXCITATION SOURCE INFORMATION

Speech is considered as the output of the vocal tract system excited by the excitation source signal. Since both source and system information is embedded in a speech signal, estimation of the source signal is a primary step in the excitation source-based processing of speech. According to the linear source-filter theory of speech production[15], the speech signal *s[n]* resulted due to the convolution of excitation source *e[n]* and vocal tract response *h[n]*. Thus, *s[n]* can be written as,

$$s[n] = e[n] * h[n] \tag{1}$$

Linear prediction (LP) analysis is the most widely used approach for the estimation of *h[n]*. The vocal tract system's transfer function *H[z]* is modeled by an all-pole filter[4]. *H[z]* is given by

$$H[z] = \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

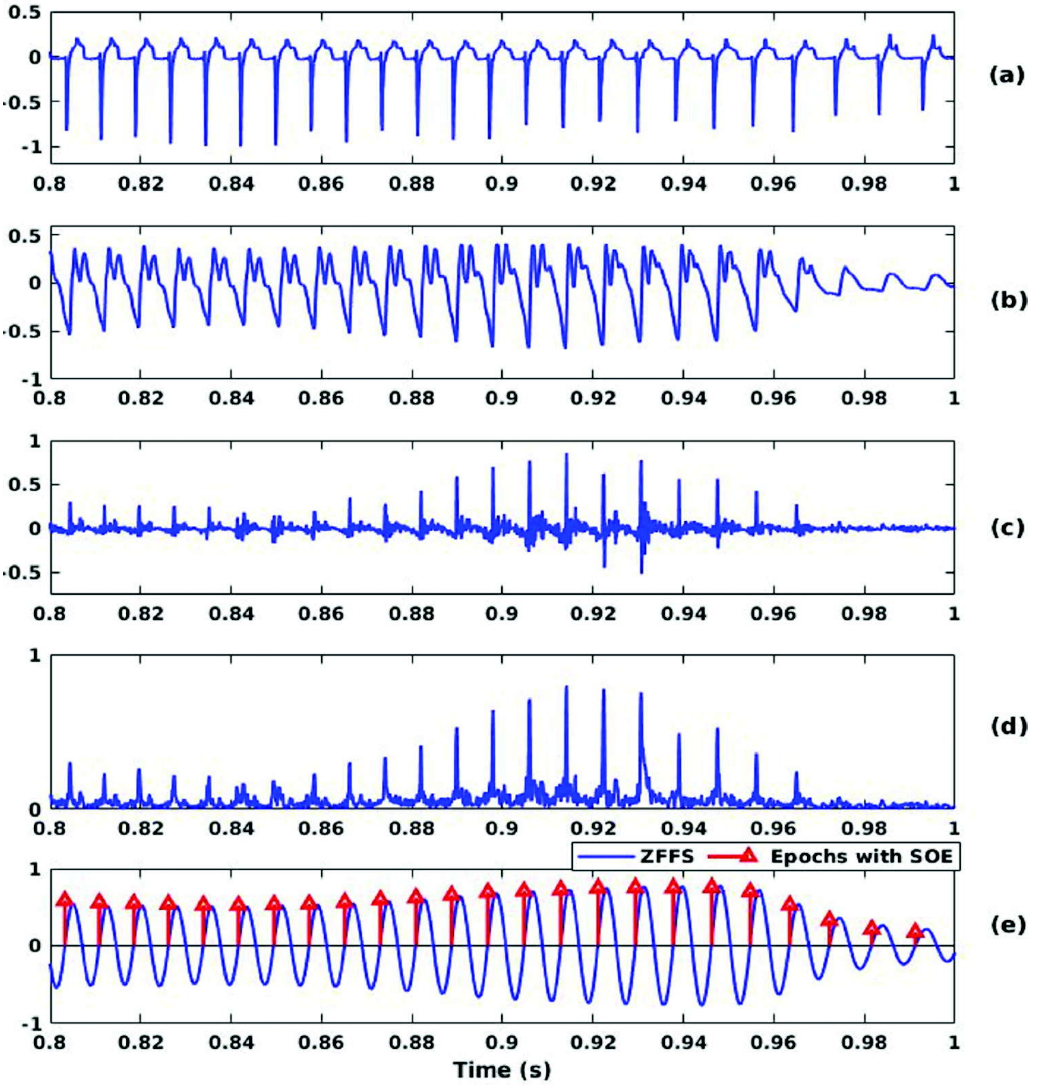


Fig. 2. (a) DEGG, **(b)** speech waveform, **(c)** LP residual, **(d)** Hilbert envelope of LP residual and **(e)** ZFFS superimposed with epochs and their strength of excitation (SoE).

where, a_k, $k = [1, p]$ are the LP coefficients and p is the order of the LP model. The LP residual *e[n] i.e.*, the inverse filtered signal is given by

$$e[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k] \tag{3}$$

The LP residual is widely used as the representation of the excitation source signal[16]. Fig. 2(a)-(c) demonstrates the differenced electroglottograph (DEGG) signal, speech signal, and LP residual signal, respectively. The large negative peaks of DEGG correspond to the GCIs, around which the LP residual shows sharp peaks. The LP residual is bipolar. Hence, Hilbert envelope of LP residual is computed to emphasize the peaks around epochs[16] and it is shown in Fig. 2(d).

Estimation of source features from the inverse filtered speech critically relies on the accuracy of LP modeling. However, under noisy conditions and high-pitched speech, the LP modeling may not be accurate due to the presence of non-stationarity. To alleviate such problems, methods those estimate the glottal source features directly from the speech signal are proposed[17]. Zero frequency filtering is one such method, which estimates the epoch locations and their strength directly from speech signal[17]. Zero frequency filter (ZFF) is realized by the cascading of two ideal resonators, whose poles are located on the unit circle in the Z-plane. The trend of the filtered signal is removed using a local mean subtraction method. The trend removed filtered signal is referred to as zero-frequency filtered signal (ZFFS). The slope of ZFFS computed at each epoch location is referred to as SoE[17]. ZFFS and estimated epoch locations along with the SoE are shown in Fig. 2(e).

The estimated source features from LP residual and ZFFS are widely used in several applications, such as speech enhancement, speaker verification, voicing decision, TTS, and pathological speech analysis. These applications are described in the subsequent sections.

3. ENHANCEMENT OF NOISY SPEECH

Conventional speech enhancement methods involve the spectral modeling of noise from non-speech regions. The noise characteristics are estimated from the non-speech regions and subtracted from the degraded speech signal to get the enhanced signal. The performance of such methods critically depends on the accuracy of noise estimation. However, noise estimation may be difficult in real-time, as the noise characteristics randomly vary over time. Alternatively, temporal enhancement algorithms are proposed in the literature[6],[18]. In[7], processing of the excitation source signal, such as LP residual, is carried out for the noisy speech enhancement. The basis for the temporal enhancement method is that human beings perceive the speech by capturing the regions from high SNR regions and then extrapolates the low SNR regions. Hence, the temporal enhancement methods involve the identification and enhancement of high SNR regions in noisy speech. In the temporal enhancement, the LP residual is weighted at two levels, namely, gross and fine levels. Gross level weight highlights the high SNR regions of 40-100 ms, whereas fine level weight emphasizes the regions of 2-3 ms. Fine-level processing involves the enhancement of regions around the epochs. Because, around the epochs, impulse-like energy is delivered to the vocal tract. Further, the LP residual is multiplied with the weight function derived by emphasizing high SNR regions. The weighted LP residual is used to excite the time-varying all-pole filter derived from the noisy speech to generate the enhanced speech signal.

Based on the concept of the excitation source-based speech enhancement approach proposed in[18], various temporal enhancement algorithms are proposed in the literature. These algorithms involve the enhancement of speech in the presence of environmental noise [18],[19], reverberant noise[6],[20], multi-speaker noise[21] and multi-channel speech[22]. A concept of foreground speech enhancement is proposed in[23]. Authors of[23],[24], applied the concept of temporal enhancement in fore-ground speech enhancement. In [24], a zero band filtering method (a modified version of ZFFS) is used to detect the epochs, and later, these epochs are used to identify the fine-level weight function. The gross level weight is derived using excitation strength, normalized autocorrelation peak strength, and modulation spectrum features. The significance of foreground speech enhancement is demonstrated in the spoken query detection system[25].

4. SPEAKER VERIFICATION AND SPOOF DETECTION

The importance of excitation source information is also demonstrated for speaker verification applications. Experiments in[26] showed that humans can recognize people by listening to the LP residual signal. This statement motivated the researchers to explore the excitation source for the extraction of speaker-specific information. In[27], authors showed that synthesized speech using random noise, instead of LP residual contains less speaker-specific information. A combination of source feature, i.e., the energy of the LP residual with the vocal tract feature, *i.e.,* linear prediction cepstral coefficients (LPCCs) resulted in an improved speaker recognition performance over LPCCs alone[28]. Authors of[8] carried out a detailed experiment to extract speaker-specific information from the LP residual using the auto-associative neural network (AANN) framework. The experimental results showed that the LP residual computed using the LP order of 8-20 can represent the speaker-specific information best. Also, the authors showed that AANN requires significantly less amount of data for training the speaker model.

Motivated by the presence of speaker-specific information in LP residual[8], a framework for the speaker verification using short-utterance is demonstrated in[9],[29]. The work utilized different attributes of the excitation source, *i.e.,* Mel power difference of spectrum in sub-band, residual Mel-frequency cepstral coefficient, and discrete cosine transform of the integrated linear prediction residual (ILPR). These three features emphasize the different attributes of source, namely, periodicity, smoothed spectrum information, and shape of the glottal signal, respectively. The performance of the speaker verification system is evaluated using the NIST SRE 2003 database. The experimental evaluation showed that for the 2s duration of test speech, a combination of excitation source features resulted in better performance than the MFCCs. Further, the fusion of source features with MFCCs significantly improves the performance of the Automatic speaker verification (ASV) system for 2 sec of test duration. The details of the experimental results are given in Table 1.

Table 1. Performance of mfccs, source fusion, and source features with MFCCs to show the importance of source features for short utterance ASV system [9].

Test duration	MFCCs		Source fusion		Source fusion + MFCCs	
	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
10 sec	5.81	0.109	10.57	0.1964	5.1	0.0965
5 sec	10.52	0.1977	11.97	0.2252	8.18	0.1524
3 sec	16.94	0.31	15.85	0.2854	11.47	0.2148
2 sec	22.31	0.4128	20.19	0.3759	16.08	0.3025

Speaker verification systems are highly vulnerable to spoofing attacks, and it has been observed that their performances get severely degraded when subjected to these attacks[30]. Researchers have explored and showed the importance of excitation source features to detect one of the spoofing attacks, called a replay attack. In one of the works[30], two source-based features, namely epoch feature and mean and skewness of peak to sidelobe ratio of the Hilbert envelope of LP residual are explored. These features characterize the excitation source behavior around the GCIs. After performing a score-level fusion of source and state-of-the-art spectral features (constant Q cepstral coefficients (CQCCs) and MFCCs), the combined system significantly outperforms the individual systems by a significant margin. However, in[30], authors only utilize the information around GCIs and do not explore the dynamic characteristics of the source signal between two GCIs. This information can be extracted with the help of an ILPR-based source signal, which models the temporal shape of the voice source signal between two adjacent GCIs[31]. Fig. 3 shows four glottal cycles of a speech signal and the corresponding ILPR for the original and spoofed signal. It can be seen from the figure that the dynamics of the ILPR signal between two GCIs are distorted as that of the original. It is expected that characterizing the source temporal dynamics between two GCIs may give an improvement in the spoof detection system. The experiments are performed on the ASVSpoof

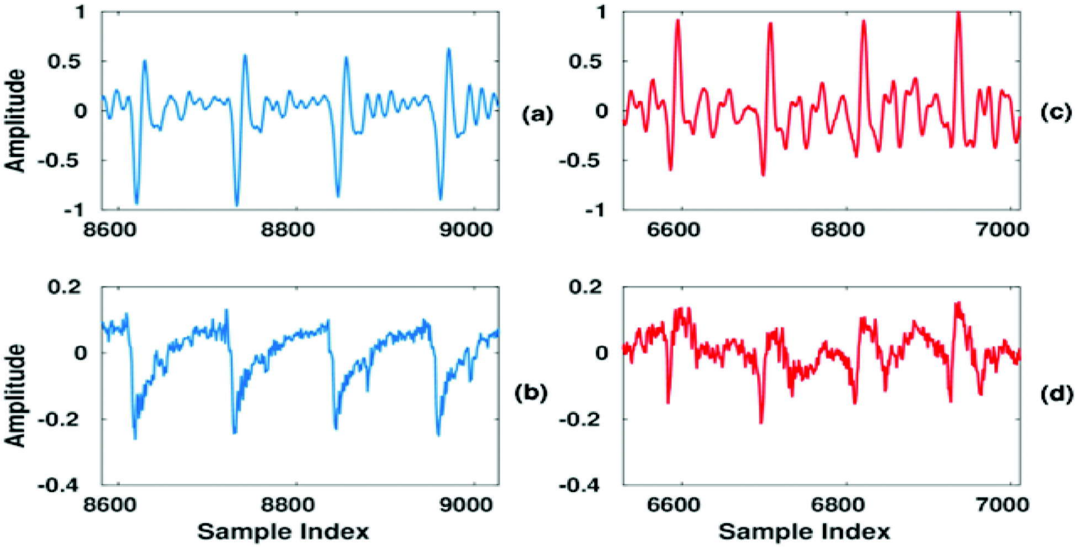


Fig. 3. *ILPR* signals for segments of genuine and spoofed speech signals. (a)-(b) and (c)-(d) represent the speech signal and its corresponding ILPR signal for genuine and spoofed signals, respectively.

2017 Version 2.0 database. On fusing the systems developed using CQCC features and proposed source features, an equal error rate of 9:41% is achieved on the evaluation set.

5. SPEECH SYNTHESIS

The naturalness of synthesized speech highly depends on the modeling of excitation source components. In the work[32], the importance of epochs in speech synthesis is demonstrated. Synthesized speech by modeling the components anchored around the epochs showed perceptually more significant than analyzing the entire speech signal[32]. Generally, the excitation source is modeled as random noise in unvoiced sounds and periodic impulse train in voiced sounds. Due to this, the synthesized speech is perceived as buzzy, monotonous, and unnatural[13]. Efforts have been made towards the modeling of source parameters, such as periodicity, voicing decision, phase of excitation component, etc. to improve the naturalness of synthesized speech. The use of excitation source parameters in speech synthesis is explained in this section.

Different attributes of the excitation source, i.e., periodicity, energy, and asymmetry nature of excitation source, are used for the voicing decision in[10]. The normalized autocorrelation peak strength and SoE computed from ZFFS are used to represent the periodicity and energy aspects, respectively. Skewness to kurtosis ratio (SKR) computed from the ILPR signal is used to represent the asymmetric nature of the excitation source. Fig. 4(a) and (b) represent the speech waveform and ZFFS, respectively. NAPS and SoE computed from ZFFS are plotted in Fig. 4(c) and (d), respectively. ILPR and SKR are shown in Fig. 4(e) and (f), respectively. Fig. 4 shows that different attributes of excitation source highlight the voiced region of speech. Hence, these three pieces of evidence are used to train the SVM, k-means, and deep belief network classifiers for the voicing decision[11]. The voicing decision is incorporated in the HMM-based statistical speech synthesizer[11]. Quality of the synthesized speech is evaluated using subjective and objective measures and compared with the STRAIGHT, REAPER, and TEMP based voicing decision methods. The excitation source-based voicing decision method showed better performance when compared to the state-of-the-art methods.

Generally, the glottal source is characterized as periodic in nature. However, there is a noise component present along with the periodic component. Researchers showed that the modeling of noise component, *i.e.*, aperiodic component improves the naturalness of synthesized speech. In[13], periodic and aperiodic components are estimated using the ILPR signal. The signal components below and above 4000 Hz are separated. The low-frequency components of ILPR contain periodic, whereas high-frequency components contain aperiodic components. Mel-cepstral coefficients (MCEPs) computed from the filtered signal are used in the HMM-based speech synthesizer. The incorporation of the aperiodic component significantly improved the naturalness of synthesized speech[13].

The strength of glottal vibration is relatively higher in the noise environment than the quiet one. In [14], the peaks of the Hilbert envelope of LP residual are used for the characterization of Lombard speech. The strength of peaks is relatively higher for Lombard speech than normal. Based on this analysis, the post-processing of synthesized speech is carried out to improve its intelligibility in a noisy environment.

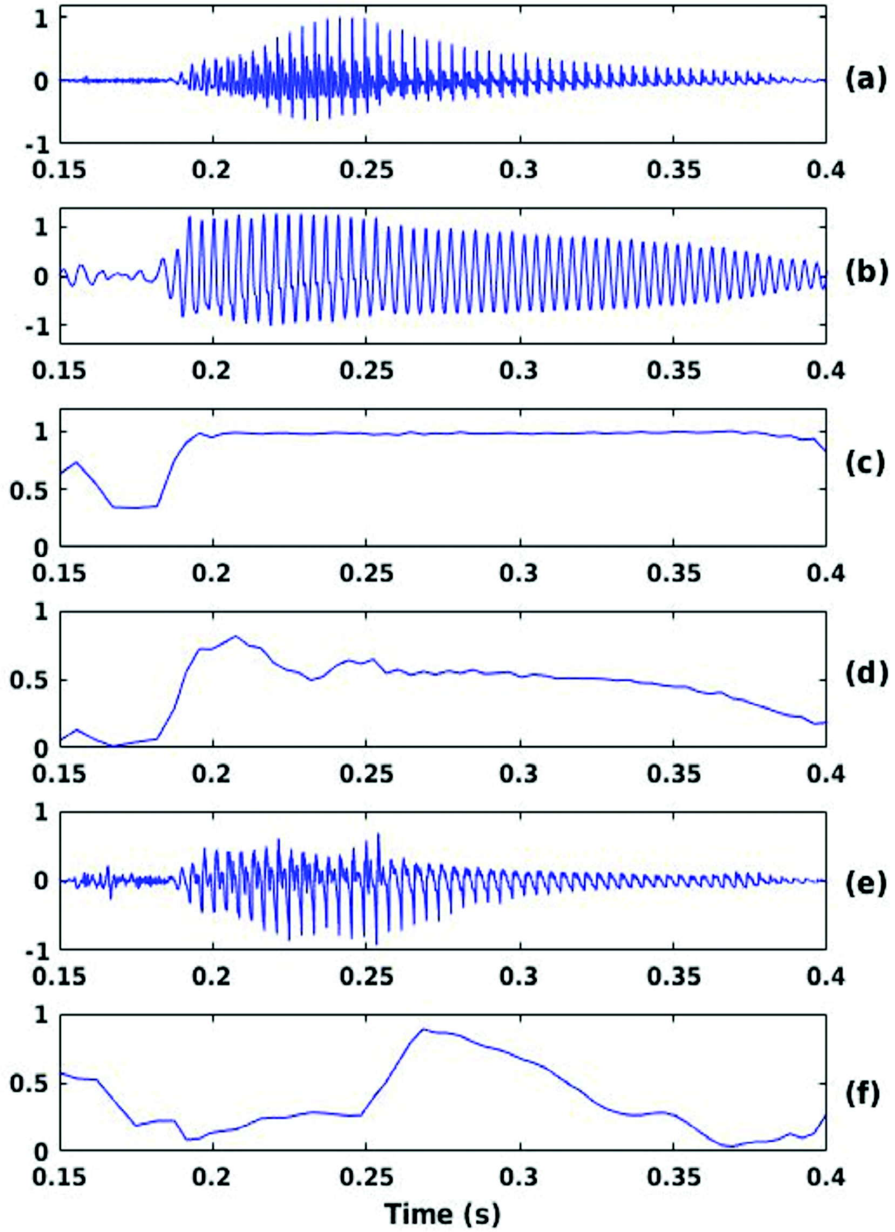


Fig. 4. Different attributes of the excitation source for voicing decision.
(a) Speech signal, **(b)** ZFFS, **(c)** NAPS, **(d)** SoE, **(e)** ILPR and **(f)** SKR feature.

6. DETECTION OF SPEECH DISORDERS

The presence of organic vocal fold pathology or neurological disorders affects the periodicity characteristics of glottal source. The period and amplitude of glottal vibrations are analyzed to characterize the voice disorders. Conventionally, the speech signal of 20-30 ms is analyzed for the detection of voice disorders. The knowledge of epochs can be utilized for the characterization of voice disorders[12]. Authors in[12] compute the jitter and shimmer using the epochs and strength of excitation to analyze the normal and pathological speech[12]. Fig. 5 demonstrates the importance of instantaneous pitch period and SoE in the discrimination of normal and pathological voice signals. Variation of the instantaneous pitch period (jitter) is high in the case of pathological speech than normal (Fig. 5(d) and (j)). SoE captures the strength of glottal vibrations. Variation of SoE (shimmer) also shows higher values for pathological speech than normal (Fig. 5(f) and (l)). The usage of excitation source-based features showed improvement over PRAAT-based features[33] in the classification of normal and voiced disordered speech[12].

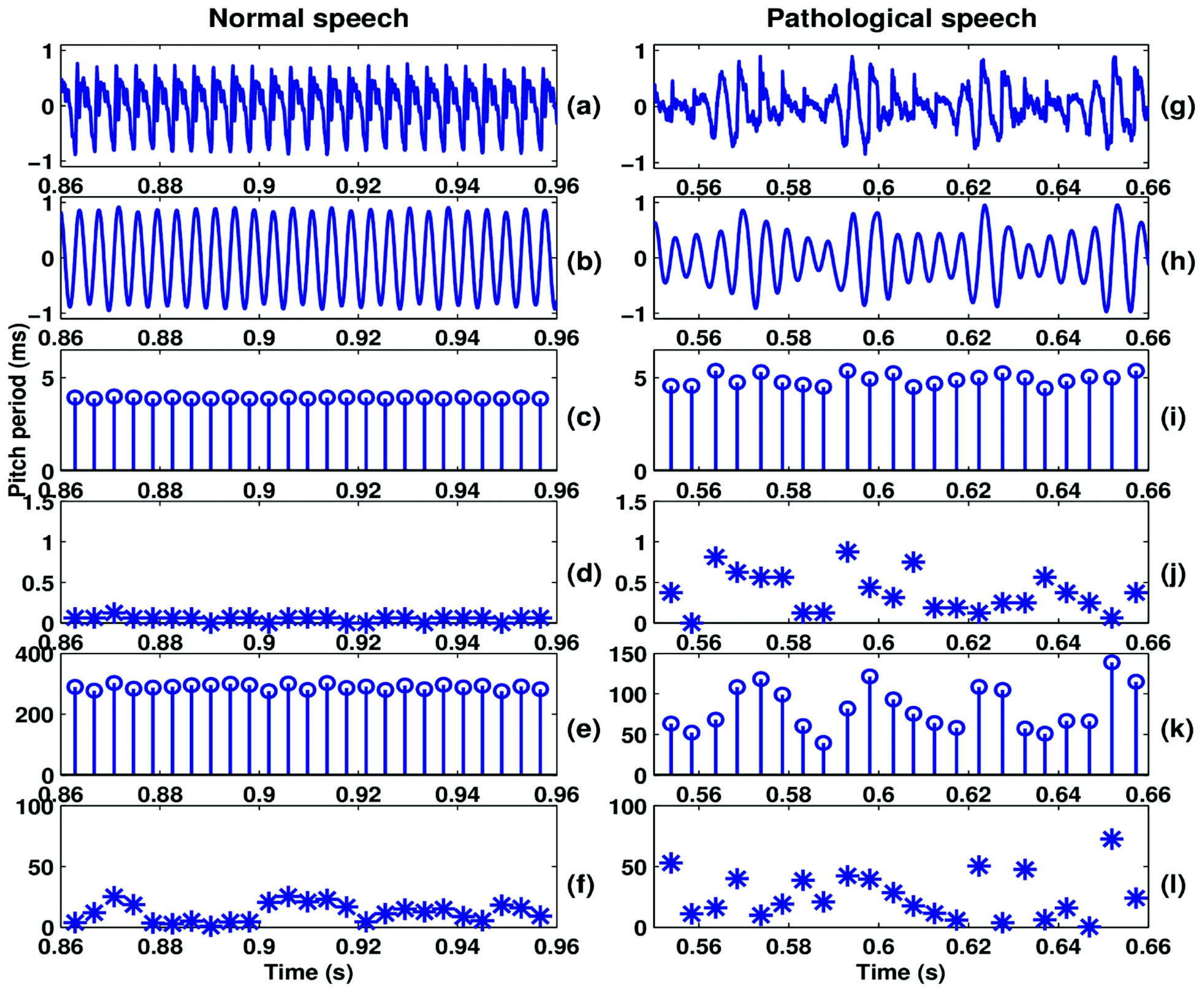


Fig. 5. Analysis of the instantaneous pitch period and SoE derived from ZFFS for normal and pathological speech [12]. **(a)-(f)** and **(g)-(l)** represent the speech signal, ZFFS, instantaneous pitch period (T_0) contour, a first-order absolute difference of T_0, SoE contour, the first-order absolute difference of SoE, for normal and pathological voice signals, respectively.

The glottal activity regions are analyzed for the estimation of hypernasality severity [34]. Hypernasality refers to the perception of abnormal nasal resonances in voiced sounds[35]. Voiced regions are detected using the different attributes of the source proposed in [10]. The estimated hypernasality scores from glottal activity regions showed a better correlation concerning the clinical ratings when compared to the processing of the entire speech signal[34]. The presence of glottal vibrations is necessary for the discrimination of voiced and unvoiced plosives. The presence/absence of glottal vibrations in the production of unvoiced/voiced stops refers to the erroneous of glottal vibrations. These errors are referred to as glottal activity errors[36]. The excitation strength feature computed using ZFFS is used for the automatic detection of glottal activity errors in cleft lip and palate (CLP) speech [36].

The speech segments anchored around the onset and offset of glottal vibrations (glottis-landmarks) are associated with the formant transitions. Processing of speech signals around the glottis (g) landmarks is carried out to analyze the intelligibility of cleft palate speech[37]. Joint spectro-temporal features (2-dimensional discrete cosine transform coefficients) extracted from the speech signal around the g-landmarks showed better correlation concerning the clinical ratings of intelligibility. The effect of supraglottal constriction on glottal vibrations is analyzed for the detection of misarticulated sounds in[38],[39]. The excitation source characteristics are not only decided by the vocal fold structure, but also by the vocal tract constriction. An increase in the vocal tract constriction inhibits the glottal vibrations. The effect of supra glottal constriction on the glottal source is analyzed for the detection of misarticulated trills and nasalized voiced stops in CLP speech[38],[39].

Voice onset time (VOT) is an important cue used in the diagnosis of dysarthria. VOT refers to the interval between the onset of burst and the onset of glottal vibrations during plosives production [2]. VOT values signify the control and coordination between the glottal source and vocal tract articulators. The knowledge of the excitation source is necessary for the measurement of VOT. Algorithms have been proposed by exploiting the periodicity and energy attributes of the excitation source signal [40],[41]. The applications of VOT in the automatic classification of normal vs. Parkinson's disorders are demonstrated in[42],[43].

7. CONCLUSION

In this article, the significance of the excitation source information in various speech-based applications, *i.e.,* speech enhancement, speaker verification, spoof detection, speech synthesis, and detection of speech disorders are reviewed. The performance of these systems depends on the accurate estimation of source parameters from the speech signal. In these applications, the excitation source information is derived from the LP residual, and ZFFS. However, the LP residual is highly sensitive to noisy conditions and pitch variations. Hence, methods for the robust estimation of source parameters from the speech signal need to be developed.

8. REFERENCES

[1] B. Yegnanarayana and S. V. Gangashetty, 2011. "Epoch-based analysis of speech signals," *Sadhana,* **36**(5), 651-697.

[2] K. N. Stevens, 2000. Acoustic phonetics. *MIT Press,* **30**.

[3] K. S. R. Murty and B. Yegnanarayana, 2008. "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.,* **16**(8), 1602-1613.

[4] L. R. Rabiner, R. W. Schafer *et al.*, 2007. "Introduction to digital speech processing," Foundations and Trends R in Signal Processing, **1**(1-2), 1-194.

[5] C. Vikram and S. R. M. Prasanna, 2017. "Epoch extraction from telephone quality speech using a single-pole filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**(3), 624-636.

[6] B. Yegnanarayana and P. S. Murthy, 2000. "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. Speech, Audio Process.,* **8**(3), 267-281.

[7] B. Yegnanarayana and R. N. Veldhuis, 1998. "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. on Speech, Audio Process.* **6**(4), 313-327.

[8] S. R. M. Prasanna, C. S. Gupta and B. Yegnanarayana, 2006. "Extraction of speaker-specific excitation information from linear prediction residual of speech," Speech Communication, **48**(10), 1243-1261.

[9] R. K. Das and S. R. M. Prasanna, 2016. "Exploring different attributes of source information for speaker verification with limited test data," *The Journal of the Acoustical Society of America,* 140(1), 184-190.

[10] N. Adiga and S. R. M. Prasanna, 2015. "Detection of glottal activity using different attributes of source information," *IEEE Signal Processing Letters,* **22**(11), 2107-2111.

[11] N. Adiga, B. K. Khonglah and S. M. Prasanna, 2017. "Improved voicing decision using glottal activity features for statistical parametric speech synthesis," *Digital Signal Processing,* **71**, 131-143.

[12] N. Adiga, C. Vikram, K. Pullela and S. M. Prasanna, 2017. "Zero frequency filter-based analysis of voice disorders.".

[13] N. Adiga and S. M. Prasanna, 2016. "Source modeling for hmm based speech synthesis using integrated lp residual," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5590-5594.

[14] B. Sharma and S. M. Prasanna, 2017. "Enhancement of spectral tilt in synthesized speech," *IEEE Signal Processing Letters,* **24**(4), 382-386.

[15] G. Fant, 1981. "The source filter concept in voice production," *STL-QPSR,* **1**, 21-37.

[16] T. V. Ananthapadmanabha and B. Yegnanarayana, 1979. "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust. Speech, Signal Process.,* **27**(4), 309-319.

[17] K. S. R. Murty, B. Yegnanarayana and M. A. Joseph, 2009. "Characterization of glottal activity from speech signals," *IEEE Signal processing letters,* **16**(6), 469-472.

[18] B. Yegnanarayana, C. Avendano, H. Hermansky and P. S. Murthy, 1999. "Speech enhancement using linear prediction residual," *Speech communication,* **28**(1), 25-42.

[19] P. Krishnamoorthy and S. R. M. Prasanna, 2011. "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication,* **53**(2), 154-174.

[20] P. Krishnamoorthy and S. R. M. Prasanna, 2009. "Reverberant speech enhancement by temporal and spectral processing," *IEEE transactions on audio, speech, and language processing,* **17**(2), 253-266.

[21] 2009. "Application of combined temporal and spectral processing methods for speaker recognition under noisy, reverberant or multi-speaker environments," *Sadhana,* **34**(5), 729.

[22] B. Yegnanarayana, S. M. Prasanna and K. S. Rao, 2002. "Speech enhancement using excitation source information," *IEEE International Conference on Acoustics, Speech and Signal Processing,* **1**, 1-541.

[23] K. Deepak, B. D. Sarma and S. M. Prasanna, 2012. "Foreground speech segmentation using zero frequency filtered signal," *Thirteenth Annual Conference of the International Speech Communication Association*.

[24] K. T. Deepak and S. R. M. Prasanna, 2016. "Foreground speech segmentation and enhancement using glottal closure instants and Mel cepstral coefficients," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **24**(7), 1205-1219.

[25] A. Dey, S. Shahnawazuddin, Deepak K.T., S. Imani, S. R. M. Prasanna and R. Sinha, 2016. "Enhancements in Assamese spoken query system: Enabling background noise suppression and flexible queries," *Twenty-Second National Conference on Communication (NCC)*, pp. 1-6.

[26] T. C. Feustel, R. J. Logan and G. A., 1988. Velius, "Human and machine performance on speaker identity verification," *The Journal of the Acoustical Society of America,* **83**(S1), S55-S55.

[27] K. S. R. Murty, S. M. Prasanna and B. Yegnanarayana, 2004. "Speaker-specific information from the residual phase," *International Conference on Signal Processing and Communications, 2004. SPCOM'04.* pp. 516-519.

[28] H. Wakita, 1976. "Residual energy of linear prediction applied to vowel and speaker recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing,* **24**(3), 270-271.

[29] A. Pandey, R. K. Das, N. Adiga, N. Gupta and S. M. Prasanna, 2015. "Significance of glottal activity detection for speaker verification in degraded and limited data condition," in TENCON 2015-2015 *IEEE Region 10 Conference.* pp. 1-6.

[30] S. Jelil, R. K. Das, S. M. Prasanna and R. Sinha, 2017. "Spoof detection using source, instantaneous frequency and cepstral features." *Interspeech.*

[31] S. Jelil, S. Kalita, S. M. Prasanna and R. Sinha, 2018. "Exploration of compressed ilpr features for replay attack detection." *Interspeech.*

[32] N. Adiga and S. R. M. Prasanna, 2013. "Significance of instants of significant excitation for source modeling." *Interspeech,* pp. 1677-1681.

[33] P. Boersma and D. Weenink, 2018. "Praat: Doing phonetics by computer [computer program]. version 6.0. 37," Retrieved February, **3**, 2018.

[34] C. M. Vikram, A. Tripathi, S. Kalita and S. R. M. Prasanna, 2018. "Estimation of hypernasality scores from cleft lip and palate speech." *Interspeech,* pp. 1701-1705.

[35] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone and T. L. Whitehill, 2008. "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal,* **45**(1), 1-17.

[36] C. M. Vikram, S. R. M. Prasanna, A. K. Abraham, M. Pushpavathi and K. Girish, 2018. "Detection of glottal activity errors in the production of stop consonants in children with cleft lip and palate." *Interspeech*, pp. 382-386.

[37] S. Kalita, S. R. Mahadeva Prasanna and S. Dandapat, 2018. "Importance of glottis landmarks for the assessment of cleft lip and palate speech intelligibility," *The Journal of the Acoustical Society of America,* **144**(5), 2656-2661.

[38] C. M. Vikram, S. K. Macha, S. Kalita and S. R. Mahadeva Prasanna, 2018. "Acoustic analysis of misarticulated trills in cleft lip and palate children," *The Journal of the Acoustical Society of America,* **143**(6), EL474-EL480.

[39] C. M. Vikram, N. Adiga and S. R. M. Prasanna, 2019. "Detection of nasalized voiced stops in cleft palate speech using epoch-synchronous features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **27**(7), 1189-1200.

[40] A. Prathosh, A. Ramakrishnan and T. Ananthapadmanabha, 2014. "Estimation of voice-onset time in continuous speech using temporal measures," *The Journal of the Acoustical Society of America,* **136**(2), EL122-EL128.

[41] V. Stouten *et al.*, 2009. "Automatic voice onset time estimation from reassignment spectra," *Speech Communication,* **51**(12), 1194-1205.

[42] A. Bayestehtashk, M. Asgari, I. Shafran and J. McNames, 2015. "Fully automated assessment of the severity of parkinson's disease from speech," *Computer speech & language,* **29**(1), 172-185.

[43] D. Montana,˜ Y. Campos-Roca and C. J. Perez,´ 2018. "A diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of parkinson's disease," *Computer methods and programs in biomedicine,* **154**, 89-97.

The Journal of Acoustical Society of India : Vol. 47, No. 2-3, 2020 (pp. 90-101)

An interactive MATLAB based GUI for speech processing and stress detection

Amit Abhishek, Sibasis Sahoo and Samarendra Dandapat
Electronics & Electrical Engineering
Indian Institute of Technology Guwahati-781039, India
e-mail: amit.abhishek74@gmail.com

[Received: 25-03-2020; Revised: 23-05-2020; Accepted: 23-05-2020]

ABSTRACT

In this work, a MATLAB based graphical user interface (GUI) has been developed for assisting students in learning concepts of speech signal processing. It is capable of real-time interactive speech signal analysis. The study of speech processing and its application requires the understanding of signal processing as well as machine learning concepts. A software having the flexibility to demonstrate these concepts will be highly helpful for students as well as teachers. It helps in complementing the classroom teaching of speech signal processing and its applications. The GUI consists of three modules: Learning module, signal processing module, and stress detection module. The learner module helps the user by giving a detailed walk-through of the GUI. The real-time stress detection module is a real-time application of speech signal processing and machine learning concepts. It can identify four stress classes: urgency, work-load (breathy), pathological and, normal cases. It allows students to learn concepts of speech signal processing, analyse the signals as well as experiment on them. Finally, it shows an application of the knowledge acquired in the form of a real-time stress detection module where it demonstrates the detection of stress from the speech signal.

1. INTRODUCTION

Speech signal processing is an active area of research, and it is attracting a lot of attention from students and researchers. Current classroom teaching takes help of presentations, printed materials and programming assignments. Except for the programming assignments, students lack interactive sessions which may help them to understand the concepts better. Therefore, there is a need for a unified system consisting of all relevant speech signal processing tools and sample applications. This will help instructors to demonstrate the concepts with minute details. At the same time, students can also use it on their own to understand the concepts better. There are several software available on the internet for speech signal analysis. But, most of them focus on a preliminary level of speech processing. The University College of London has designed two software WASP and Speech Filing System (SFS). WASP is designed for performing introductory acoustic analysis. SFS is a more powerful tool for a similar purpose. This software performs standard operations like sound acquisition, replay, annotate, formant estimation, spectrographic analysis and advanced operations like speech synthesis, recognition and supports software development at user's end. Praat[1,2], and wave surfer[3] are two other software that are quite popular among the researcher community. These application software share the common tools like pitch extraction, pitch variation, formant tracking, spectrographic study, signal annotation etc. However, these have not much been used for classroom study by beginners. They also lack a proper speech processing application which will excite students to pursue the domain further.

Stress recognition is a process of identifying stress class from the user's speech[4]. The characteristics of the speech signal vary under different stress conditions which affect the performance of a machine in case of human-machine interaction[5-7]. The causes of this stressed speech can be due to emergency conditions, fatigue/physical environmental factor, pathological condition (disease), sleep deprivation, perceived threat, glottal abnormalities, work-load, noisy environments (Lombard effect)[8]. Stress classification can help improve speech and speaker recognition. It can have applications in (i) prioritising emergency situations, (ii) medical situations, (iii) analysis of breathing pattern of a sportsperson, (iv) assessing the quality of customer satisfaction in telecommunication industry and (v) forensic analysis of the caller by the law enforcement agencies[9,10].

In this work, a new GUI has been developed. It aims at complementing the classroom teaching of concepts of speech signal processing. It permits users to analyse the speech signal in real-time and to experiment on it by providing different temporal and spectral signal processing. Features like signal acquisition, resampling, playback, partial selection, autocorrelation, pitch estimation help a user in processing the speech signal in time domain. At the same time, it has tools that can process the signal in the frequency domain to produce spectrum using different windowing methods, spectrogram, estimate formant structure etc. It also has some advanced functionalities such as linear prediction (LP) analysis, LPCC, MFCC, residual signal generation, vowel generation and many more. At any time, in case of any doubts regarding the concepts, the user can access the learner module to get the theoretical background. A stress detection module is also included in this GUI as a real-time application for addressing the challenges of detecting states of the stress of a person using the speech signal. This module can identify stress classes like emergency, work-load, pathological and normal states. A new stressed speech database *IITG-stress Database* is recorded consisting of above four stress classes, unlike the existing database which contains styled stress classes like angry, sad, happy, anxiety etc.

The organisation of the paper is summarised as follows. The detail explanation of the proposed architecture is given in Section 2. Feature extraction, statistical analysis and stress detection are carried out in section 2.3. Finally, we conclude the work in section 3.

2. PROPOSED GUI ARCHITECTURE

The flow diagram of the proposed GUI is shown in Fig. 1. It consists of three sections: (i) Speech Processing (ii) Learner and, (iii) Stress detection. An itemised list of the applications present in this GUI is shown in Fig. 2. Speech processing module encapsulates different speech signal processing tools ranging from standard ones like audio recording, replaying, formant extraction, pitch estimation, spectrographic study etc., to advanced tools like LPC, MFCC, cepstral analysis, vowel generation *etc*. Speech signal researchers can find this module handy at their disposal. Learner module gives the user a walk-through of the GUI. In addition to that, it also contains the basics of speech signal processing concepts and their references. The stress detection section is basically a sample

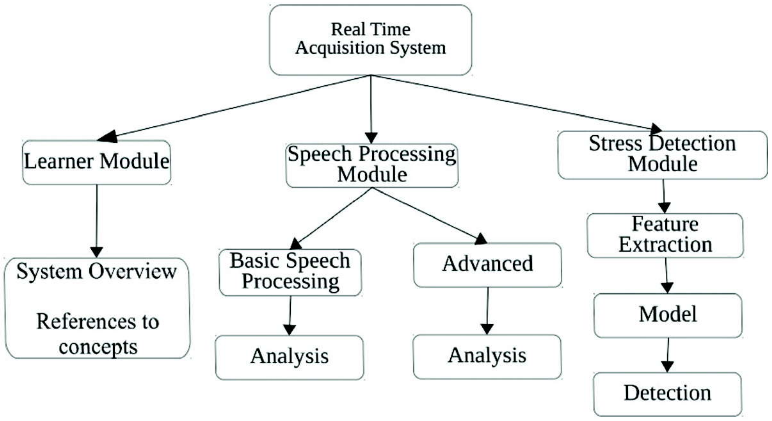
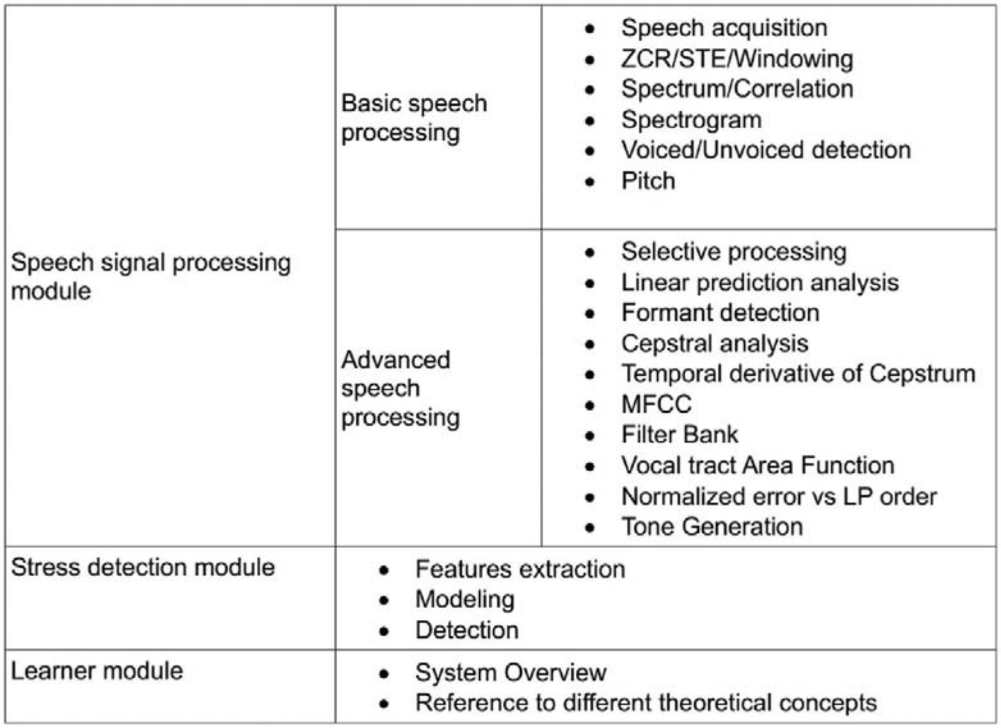


Fig. 1. GUI flow diagram.



Speech signal processing module	Basic speech processing	- Speech acquisition - ZCR/STE/Windowing - Spectrum/Correlation - Spectrogram - Voiced/Unvoiced detection - Pitch
	Advanced speech processing	- Selective processing - Linear prediction analysis - Formant detection - Cepstral analysis - Temporal derivative of Cepstrum - MFCC - Filter Bank - Vocal tract Area Function - Normalized error vs LP order - Tone Generation
Stress detection module	- Features extraction - Modeling - Detection	
Learner module	- System Overview - Reference to different theoretical concepts	

Fig. 2. Brief Layout of the modules included in the proposed system.

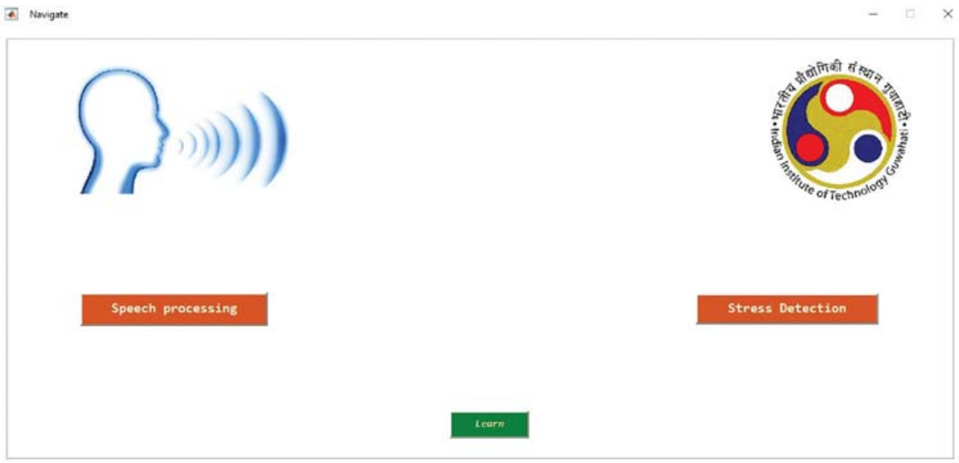
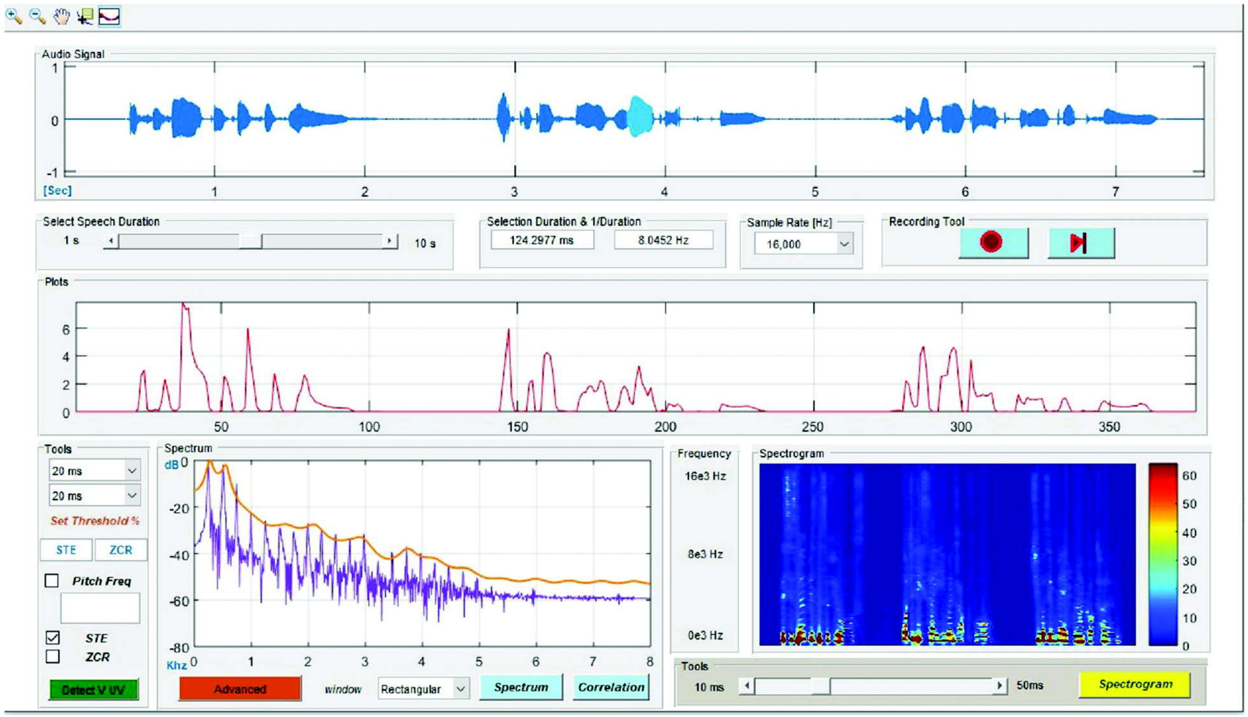


Fig. 3. Home page of the proposed GUI

application showing the detection of stress conditions from a speech signal. The home page of this GUI is shown in Fig. 3, which provides buttons to enter into the individual sections.

2.1 Speech Processing Module

Speech Processing module can be accessed from the Home screen by clicking the button 'Speech processing', its user interface (UI) is shown in Fig. 4. It mostly consists of standard signal processing tools like signal recording, replaying, sampling frequency variation, pitch calculation, normalised short-term zero-crossing calculation, normalised short-term signal energy, voiced-unvoiced detection, windowing and spectrographic analysis.



Select Speech Duration: 1 s – 10 s

Selection Duration & 1/Duration: 124.2977 ms 8.0452 Hz

Sample Rate [Hz]: 16,000

Tools: 20 ms, 20 ms

window: Rectangular

Tools: 10 ms – 50ms

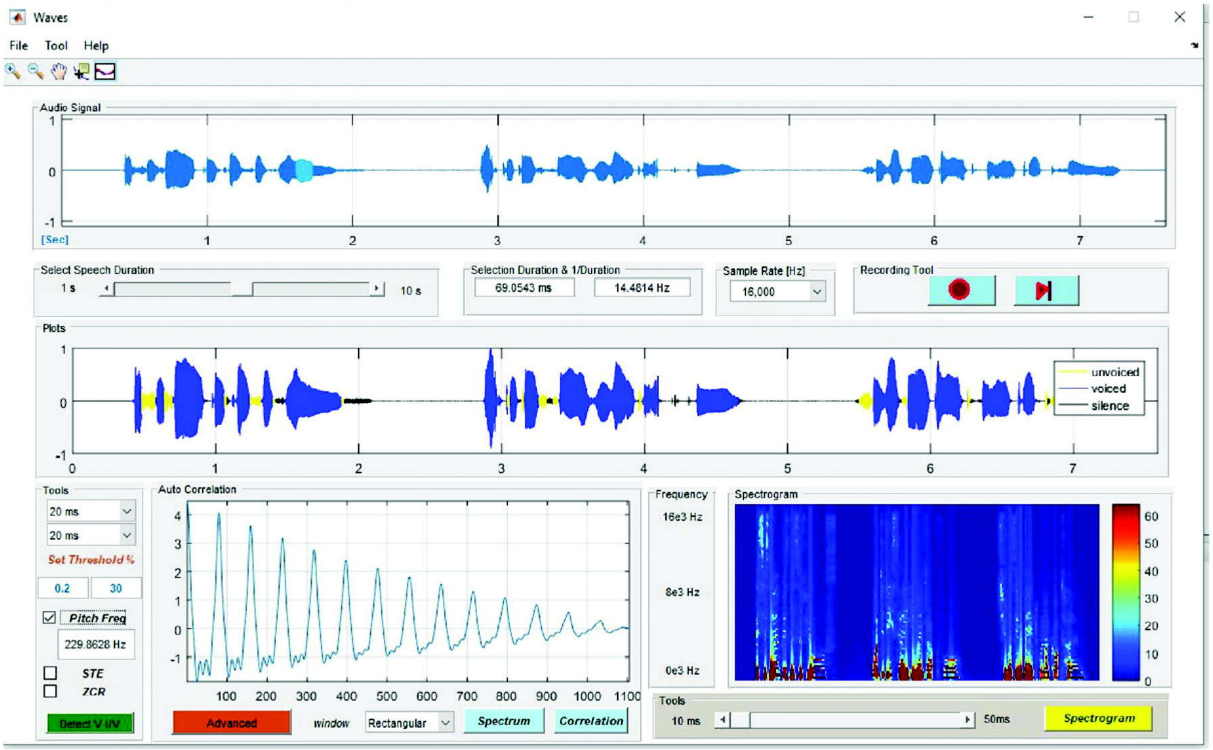
Fig. 4. Basic Speech Processing module indicating short time energy in the middle panel and Fourier magnitude spectrum in the lower-left panel.

It mainly consists of four display screens. The top screen is used for displaying the amplitude normalised audio signal with the capabilities of zooming, panning and segment selection. A user has the option to record a new audio signal using the in-built recorder or open an existing audio file having a .wav extension. For recording, we can set different sampling frequency and duration. The duration can be varied from 1 sec to 10 sec. There exists a playback button which can be used to make a perceptual judgement about the audio content and its quality.

The short-time energy and the zero-crossing rate are two basic characteristics of speech signal for identifying voiced and unvoiced regions. The module has two check-boxes: STE and, ZCR for showing short-time energy and zero-crossing rate plots in the central display panel. Using different threshold values for energy and zero-crossing rate, it is possible to detect voiced, unvoiced and silent regions. The different colour scheme is used for better identification of the plots: yellow for voiced, black for silent and, blue for unvoiced regions respectively. Generally, speech signal processing is carried out in overlapping segments of 20-40ms frames where the signal is assumed stationary. The UI takes care of this concept by having two dropdown lists to select the frame size and frameshift values, which in turn are used for computing short-time energy and zero-crossing rate.

There are two display panels located at the lower half of the UI. The left side panel shows the magnitude spectrum of the signal selected in the top panel. It can employ different windowing techniques by choosing an appropriate window from the dropdown list: rectangular, Hamming, Hanning and Barlette. Its corresponding spectrum is displayed after clicking button Spectrum. By default rectangular window is applied. The right side display panel is reserved for displaying spectrogram of the whole signal. It is equipped with a slider for choosing the frame size ranging from 10 ms to 50 ms in order to make the spectrogram wideband to narrowband.

The advanced speech processing tools can be accessed by pressing the Advanced button located at the bottom of the Speech processing UI. User needs to select a segment of the speech signal in the top



Waves

Selection Duration & 1/Duration: 69.0543 ms, 14.4814 Hz

Sample Rate [Hz]: 16,000

Tools: 20 ms, 20 ms

Set Threshold %: 0.2, 30

☑ Pitch Freq: 229.8628 Hz

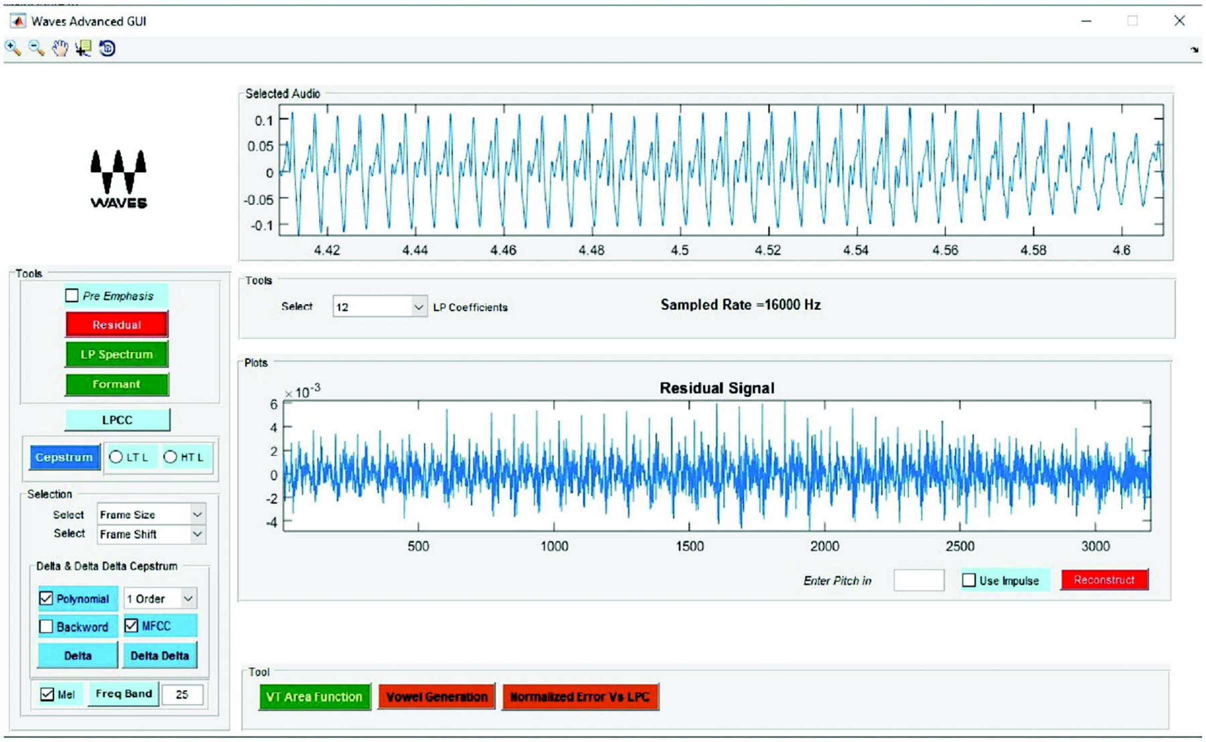
□ STE

□ ZCR

window: Rectangular

Tools: 10 ms, 50ms

Fig. 5. Speech Processing module indicating voiced-unvoiced detection and autocorrelation



Waves Advanced GUI

Select: 12 LP Coefficients

Sampled Rate =16000 Hz

□ Pre Emphasis

○ LT L ○ HT L

Select: Frame Size

Select: Frame Shift

☑ Polynomial: 1 Order

□ Backword

☑ MFCC

☑ Mel

Freq Band: 25

Enter Pitch in

□ Use Impulse

Fig. 6. Advanced Speech Processing module

displaypanel of Speech processing window before pressing the Advanced button. The advanced GUI is shown in Fig. 6. This GUI contains signal processing tools like residual signal computation using Linear Prediction (LP), LP spectrum, formant detection, cepstral analysis using liftering methods, MFCC etc. Each tool is represented by its respective button located at the left part of the GUI. It has two display panels. The job of the upper display panel is to show the selected signal, whereas the lower display panel has multiple purposes. It shows either a processed signal or spectrum of the signal depending upon the button pressed.

At the bottom part of GUI, there exist four buttons such as Freq Band, VT area Function, Vowel generation and Normalised error vs LPC. Upon setting the number of filter banks and pressing Freq Band, lower display-panel shows the triangular filter banks non-linearly spaced in frequency scale. It relates the linear frequency scale to non-linear mel scale. Clicking VT area Function will produce a pattern that approximates the variation of the vocal tract from the glottis to lips for the selected signal.

Vowel Generation button leads to a new GUI which can be used for generating vowel signals artificially. GUI has a list of five vowels when any one of them is selected, and upon pressing the Resonate button, the corresponding vowel sound is created using a predefined set of formants for that vowel. The Play button can be used to hear the sound. There is a slider named Add Noise which can be used to experiment with naturalness of artificially created vowel sound. GUI also provides a user with the flexibility to set the formants in *Hz* with their corresponding amplitude and hear the sound thus generated.

2.2 Learner module

The learner module is designed to familiarise users with the various functionalities present in this application software. The screen 'Home' contains a button named Learn that leads to a dialogue box that asks for a choice between Speech processing and Stress detection. Choice Speech processing gives a GUI for speech processing learner module. This GUI gives walk-through of the speech processing module. It also holds various references to theoretical concepts of different signal processing tools implemented in this application software. Similarly, Stress Detection choice gives another GUI which contains walk-through details and references for different machine learning tools implemented in stress detection module.

2.3 Stress detection Module

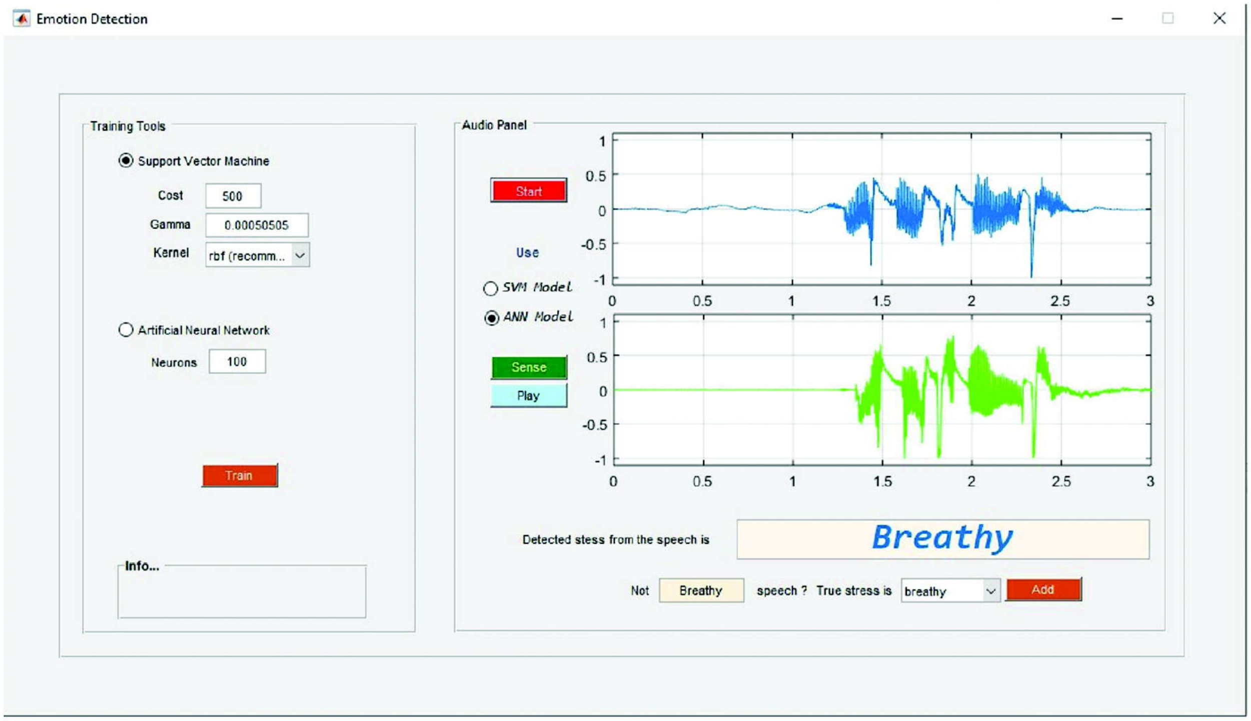
The GUI for stress detection module is shown in Fig. 7. This module is divided into two sections: (1) training, (2) detection. Training section provides two machine learning tools: Support Vector Machine (SVM) and Artificial Neural Network (ANN) to train on the stressed speech database. In the detection section, the user needs to choose between SVM model or the ANN model before going for stress class detection. Upon clicking the Start button, the application starts capturing an audio signal. Now, the user can give audio input and hit the Sense button. The stress module now will sense the stress class and will display class name inside the lower display panel. In case of a wrong detection, users can choose to add the test speech signal to the database by providing actual stress class followed by the training process. Stress detection unit has been trained on a new stress database that contains four stress classes such as breathy, pathological, emergency and normal. More description of the database is given in Section 2.3.2.

2.3.1 *Stress Analysis and Detection*

Most of the stress databases that are currently in use contain styled emotion or speech samples with Lombard effect. Few studies have been reported on stress conditions like emergency, breathy, work-load, sleep-deprivation and pathological condition. In this work, a new stressed speech database is considered as described by Amit *et al.*[11]. The stress detection module has been trained on four different stress classes like Breathy, Pathology, Normal and Urgency. The MFCC and the Fourier parameters have been used as the features for training and testing the machine learning models based on SVM and ANN.

2.3.2 *IITG-Stress Database*

The IITG stress database classifies the stress into four classes as urgency, breathy, pathological and, normal. It is a new database consisting of seventeen non-professional speakers (3 female, 14 male), from



● Support Vector Machine

Cost 500

Gamma 0.00050505

Kernel rbf (recomm...

○ Artificial Neural Network

Neurons 100

Use ○ SVM Model ● ANN Model

Detected stess from the speech is **Breathy**

Not Breathy speech ? True stress is breathy

Fig. 7. Stress Detection module

different parts of India, took part in recording the database[11]. The speakers belong to the age group of 23 to 30 years, and they are research students of the Indian Institute of Technology Guwahati. Each speaker uttered five distinct sentences for the four stress cases, respectively. The database has 85 utterances for each stress class. In total, the database has 340 speech samples. The recording task was performed inside a closed room with a sampling frequency of 11025 Hz. The sentences used in this work are listed in Table. 1.

Table 1. Recorded Sentences

1	The fire is spreading
2	Give me some water
3	The storm is coming
4	Call the ambulance
5	Hurry up; there is an accident on the highway

2.3.3 *Features Extraction and analysis*

Two sets of features: the Fourier parameters (FP) and the Mel frequency cepstral coefficients (MFCC) are used for analysing the speech signals under stress[13], [12]. MFCCs are a set of features that are widely used for tasks like emotion recognition, speech recognition and speaker recognition. These features imitate the perceptual behaviour of the human auditory system. On the other hand, the Fourier parameters are basically the Discrete Fourier coefficients computed from the Discrete Fourier transform. The Fourier parameters have been found to be affected by speech production under different stress conditions[13].

In this work, 13-dimensional MFCC, along with its velocity and acceleration parameters, are taken into consideration. Similarly, in the case of FPs, a 110-dimensional feature vector is taken considering the symmetry of the Fourier spectrum of the 20ms frame of the speech signal. For each frame of duration 20 ms, a feature vector of length 149 is extracted. The statistical analysis is performed using machine learning tools like artificial neural network (ANN) and support vector machine (SVM)[14].

A speaker-independent approach is used for performing the stress classification to understand the person independence nature of the stress classes. In this approach, the utterances of four speakers are considered for testing while the utterances of the rest of the speakers are taken for training the machine learning model. This procedure is followed for five times, and each time a different set of speakers are grouped for testing while the rest are used for training.

Classification using SVM One-vs-One

To perform the multi-class classification, the "One-vs-one" classification approach is carried out. The SVC tool, an SVM implementation present inside the Python package Sklearn, has been used for the speech stress classification. The radial basis function (RBF) has been used as the kernel function. The hyperparameters 'gamma' and 'C' were not touched upon and were left with their default values 0.076 and 1 respectively.

Classification using Neural Network

A feed-forward neural network with a single hidden layer has been tested in this work. This neural network is fully connected and consists of 200 neurons in the hidden layer. The 'ReLU' is used as the activation function of the hidden layer, and 'SoftMax' is used as the activation function of the output layer. Overall, the network consists of 149, 200 and 4 nodes at the input, the hidden and the output layers, respectively. The neural network is trained using a fixed size of 200 of epochs.

3. RESULTS

In this work, the multi-class stress classification task has been performed using two different machine learning tools. The classification results have been computed by following a speaker-independent approach. Five sets of train and test cluster pairs are created where the speakers present in each test cluster are independent of the speakers present in train cluster. For every train and test cluster pair, the ratio of the speech samples is kept at 80 to 20.

Table 2 and Table 3 show the confusion matrices for the SVM and the ANN classifiers respectively. It can be seen from both the tables that, at the classifier level, both the classifiers are showing a comparable performance for multi-class stress classification. The classification accuracy is found to be 50% and 49% for SVM and ANN classifiers, respectively. However, at the feature level, both the classifiers indicate that the Fourier parameter features show a better overall classification performance than MFCC features. Fourier parameters with SVM and ANN give 51% and 50% compared to 47% and 44% respectively.

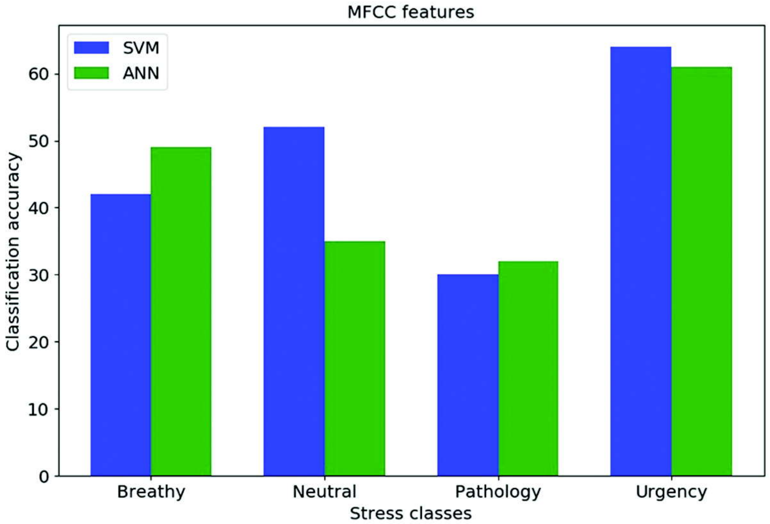


Fig. 8. Accuracy comparison with MFCC

Table 2. Confusion Matrix of SVM Classification

		MFCC features		
	Breathy	Neutral	Pathological	Urgency
Breathy	42	28	8	22
Neutral	17	52	12	19
Pathological	18	30	30	22
Urgency	22	4	10	64
		Average accuracy = 47%		
		Fourier parameters		
	Breathy	Neutral	Pathological	Urgency
Breathy	45	27	8	20
Neutral	14	56	17	13
Pathological	13	39	28	20
Urgency	5	13	8	74
		Average accuracy = 51%		
		Combined MFCC and Fourier parameters		
	Breathy	Neutral	Pathological	Urgency
Breathy	45	25	7	23
Neutral	10	54	19	17
Pathological	13	38	26	23
Urgency	5	7	12	76
		Average accuracy = 50%		

Table 3. Confusion Matrix of ANN Classification

		MFCC features		
	Breathy	Neutral	Pathological	Urgency
Breathy	49	16	15	20
Neutral	18	35	22	25
Pathological	21	27	32	20
Urgency	14	5	20	61
		Average accuracy = 44%		
		Fourier parameter features		
	Breathy	Neutral	Pathological	Urgency
Breathy	52	20	13	15
Neutral	12	51	27	10
Pathological	18	38	33	11
Urgency	2	14	18	66
		Average accuracy = 50%		
		Combined MFCC and fourier features		
	Breathy	Neutral	Pathological	Urgency
Breathy	45	20	13	22
Neutral	8	49	29	14
Pathological	14	35	35	16
Urgency	3	6	18	73
		Average accuracy = 49%		

From another point of view, we looked at the class level performance of the classifiers. Fig. 8 shows the bar plots for stress level accuracies for the MFCC features using SVM and ANN classifiers. Similarly, Fig. 9 and Fig. 10 show the bar plots using Fourier parameter and the combination of the MFCC and the Fourier parameters, respectively. These figures show that the *Urgency* class has the highest rate of classification for both the classifiers, whereas the class *pathology* performed the least. From Table 2 and

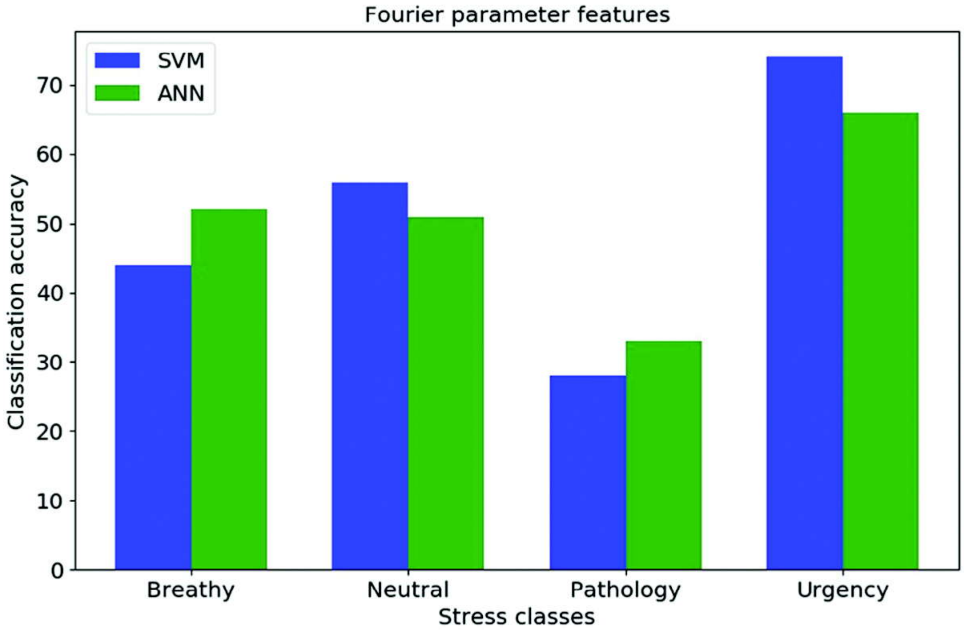


Fig. 9. Accuracy comparison with Fourier parameters

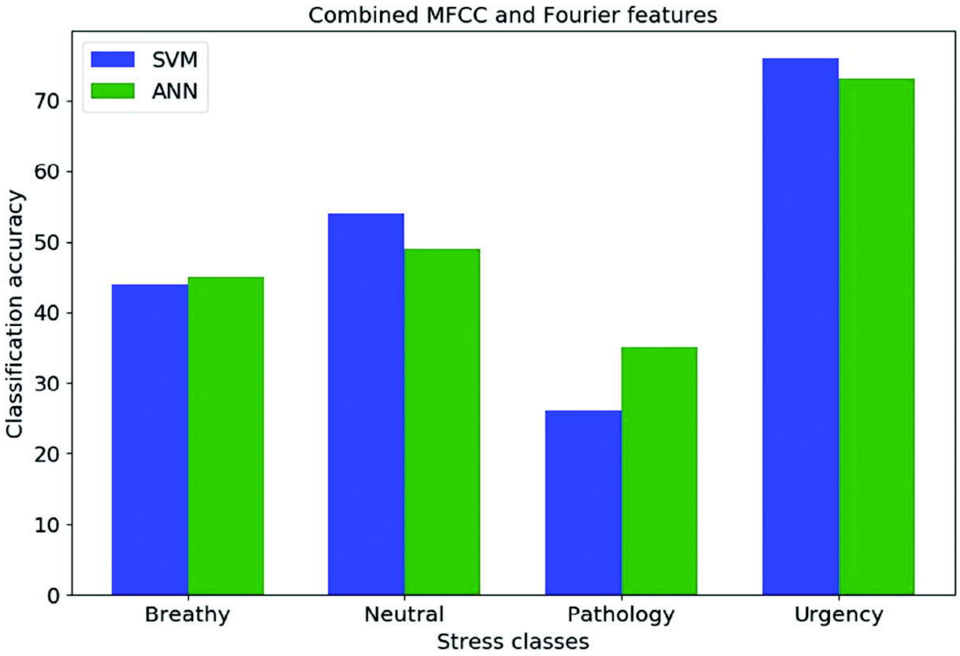


Fig. 10. Accuracy comparison with the combination of MFCC and fourier parameter features.

Table 3, it is seen that the class *Urgency* shows 76% and 73% accuracy with SVM and ANN respectively using the combination of MFCC and Fourier parameter features. This performance is the highest of all classes. The stress classes *Breathy* and *Neutral* showed better accuracy with Fourier parameter features for both the classifiers. The class *Pathological* gave the least performance in terms of accuracy as the decision of both the classifiers conflicting with the class Neutral.

4. CONCLUSION

In this work, a real-time speech analysis and stress detection system is designed. The architecture is designed to be helpful in learning practical aspects of speech signal processing techniques. The learner module assists the user to familiarise with speech signal processing concepts, algorithms as well as the application itself. The user can exploit the signal processing module to analyse the speech signal interactively. It allows both temporal and spectral processing of the signal, which will enable a user to have a better understanding of the characteristics of the speech signal. As a sample application of the learned concepts, a speech-based stress detection module is integrated into the GUI. It is capable of recognising the stress classes like neutral, emergency, breathy and pathological using speech signal in real-time. The performance of the task of stress detection is evaluated using statistical tools like SVM and ANN. The analysis showed that Fourier parameters are more effective in characterising and classifying stress in speech signals than MFCC. This analysis also indicates that the *Urgency* class is best classified using speech signal that the stress class *Pathology*.

5. REFERENCES

[1] M. Kumbhakarn and B. Sathe-Pathak, 2015. "Analysis of emotional state of a person and its effect on speech features using praat software," pp. 763-767.

[2] V.M. Ramesh and S.H.V., 2008. "Exploring data analysis in music using tool praat," pp. 508-509.

[3] N. Dhananjaya and B. Yegnanarayana, 2010. "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters,* **17**, 273-276.

[4] S. Ramamohan and S. Dandapat, 2006. "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Transactions on Audio, Speech, and Language Processing,* **14**, 737-746.

[5] S. Shukla, S. Dandapat and S.R.M. Prasanna, 2016. "A Subspace Projection Approach for Analysis of Speech Under Stressed Condition," *Circuits, Systems, and Signal Processing,* **35**, 4486-4500.

[6] G. Senthil Raja and S. Dandapat, 2010. "Speaker recognition under stressed condition," *International Journal of Speech Technology,* **13**, 141-161.

[7] B. Priya and S. Dandapat, 2016. "Subspace filtering approach based on orthogonal projection for better analysis of stressed speech under clean and noisy environments," *International Journal of Speech Technology,* 19, 731-742.

[8] M.E. Ayadi, M.S. Kamel and F. Karray, 2011. "Survey on speech emotion recognition : Features; classification schemes; and databases," *Pattern Recog.,* **44**(3), 572-587.

[9] M. Kotti and F. Paterno, 2012. "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *Int. J. Speech Technol.,* **15**, 131-150.

[10] S. Ntalampiras, I. Potamitis and N. Fakotakis, 2009. "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP J. Audio; Speech; Music Process.,* **2009**(13), 1-15.

[11] A. Abhishek, S. Deb and S. Dandapat, 2019. "Analysis of breathy, emergency and pathological stress classes," in Machine Intelligence and Signal Analysis, *Springer,* pp. 497-508.

[12] S. Deb and S. Dandapat, 2019. "Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions," *IEEE Trans. Affect. Comput.,* **10**, 360-373.

[13] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang and Lian Li, 2015. "Speech Emotion Recognition Using Fourier Parameters," *IEEE Trans. Affect. Comput.*, **6**, 69-75.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 2011. "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* **12**: 2825-2830.

The Journal of Acoustical Society of India : Vol. 47, No. 2-3, 2020 (pp. 102-111)

Multivariate quadratic regression based direction estimation of an acoustic source

Mohd Wajid[1*], Shardul Yadav[1] and Mohammed Usman[2]
[1]Department of Electronics Engineering, Z.H.C.E.T., Aligarh Muslim University, Aligarh, India
[2]Department of Electrical Engineering, King Khalid University, Abha, Saudi Arabia
e-mail: mohdwajid@zhcet.ac.in

[Received: 22-04-2020; Revised: 23-05-2020; Accepted: 23-05-2020]

ABSTRACT

This paper proposes a direction-of-arrival (DOA) estimation technique through regression machine learning model applied on the signals acquired with Uniform Linear Array (ULA) of microphones. The regression model has been trained using correlation features of the different microphone signals at the ULA of microphones. The root-mean-square angular error (RMSAE) of the multivariate-linear and multivariate-curvilinear regression model has been compared with the standard algorithm i.e. delay and sum (DAS) beam former method and it has been observed that the regression model outperforms the conventional DAS beam former under different levels of sensor noise.

1. INTRODUCTION

Estimation of Direction-of-arrival (DOA) of an acoustic signal finds utility in several applications such as sensors in robots for smooth and unhindered movement in an unknown environment, automatic camera steering towards the speaker in a conference room, smart home automation, hands-free mobiles and hearing aids, speech enhancement in a reverberant environment, tracking and surveillance of aerial/ underwater targets, *etc.*[1-19]. In the presence of impairments such as sensor noise, ambient noise, reverberation and interference, the requirements for the DOA of an acoustic source becomes very critical. Accurate estimation of DOA in the presence of such impairments can be done by employing techniques such as microphone-array and acoustic vector sensors (AVS)[20-25]. Different algorithms have been developed for DOA estimation of an acoustic source using signals received at ULA such as time difference of arrival (TDOA), subspace, beam forming, maximum likelihood, compressed sensing, *etc.*[26-31].

Readily available high-performance computing devices have popularized the use of machine-learning techniques in a wide range of application domains. The regression-based machine learning model tackles this problem by realizing a non-linear relation between the input features and the output DOA. In this article, DOA has been estimated using regression machine learning models and comparison of results with delay and sum (DAS) beam forming method is presented for different signal to noise ratio (SNR) values.

The rest of the article is organized as follows. In Section 2, DAS beam forming with ULA is presented. Section 3 discusses the DOA estimation using regression models. Details of simulation parameters and discussion of results are presented in Section 4 and Section 5 concludes the paper.

2. DELAY AND SUM BEAMFORMING METHOD

A source transmitting an exponential narrow band signal *s(t)* of wave length λ arriving at angle θ with the ordinate. A linear combination of *M* acoustic sensors (microphones or hydrophones) separated by a distance d between any two consecutive microphones placed along the abscissahas been used as a ULA receiver. The received signal at the $z^{th}(1 \le z \le M)$ acoustic sensor can be written as

$$y_z(t) = s(t)\exp\left(-j2\frac{\pi}{\lambda}(z-1)\,d\sin\theta\right) + n_z(t) \tag{1}$$

where, $n_z(t)$ represents the additive white Gaussian noise (AWGN) at the z^{th} microphone. The received signal *y(t)* at the ULA receiver can be expressed in vector form as

$$\mathbf{y}(t) = [y_1(t) \quad y_2(t)\ldots \quad \ldots y_M(t)]^T = \mathbf{A}(\theta)s(t) + \mathbf{n}(t) \tag{2}$$

where, $A(\theta)$ represents the steering vector of the uniform linear array, *n(t)* is the AWGN vector and $[.]^T$ denotes the transpose operator. The correlation matrix C_{yy} (of order *M* × *M*) of the received signal vector *y(t)* is

$$\mathbf{C_{yy}} = E\left[\mathbf{y}(t)\mathbf{y}^H(t)\right] = \mathbf{A}(\theta)\mathbf{S}\mathbf{A}^H(\theta) + \mathbf{C_n} \tag{3}$$

where $[.]^H$ denotes the Hermitian transpose and *E*[.] denotes the ensemble average. The correlation matrices *S* of the signal and that of noise C_n are expressed as

$$\mathbf{S} = E\left[\mathbf{s}(t)\mathbf{s}^H(t)\right] \tag{4}$$

$$\mathbf{C_n} = E[\mathbf{n}(t)\mathbf{n}^H(t)] \tag{5}$$

Since the noise is 'white', the correlation between any two noise components is zero and all the noise components have the same variance. Therefore, the noise correlation matrix can be expressed as

$$\mathbf{C_n} = \sigma^2\mathbf{I} \tag{6}$$

where, *I* is the identity matrix and the noise variance σ^2 represents the average noise power for zero mean Gaussian noise. Equation (3) and (6) can be combined to get

$$\mathbf{C_{yy}} = \mathbf{A}(\theta)\mathbf{S}\mathbf{A}^H(\theta) + \sigma^2\mathbf{I} \tag{7}$$

The DOA estimation in DAS beam forming method is done by calculating the signal power, $P(\phi)$ for each of the possible arrival angles and the estimated DOA angle is the argument of $P(\phi)$ that maximizes $P(\phi)$[32-35].

$$P(\phi) = \mathbf{A}^H(\phi)\mathbf{C_{yy}}\mathbf{A}(\phi) \tag{8}$$

where, $A(\phi)$ denotes the look-for-direction vector which scans for all possible arrival angles to determine the direction of arrival angle ϕ *i.e.* the value of ϕ at which $P(\phi)$ becomes maximum.

3. DOA ESTIMATION USING REGRESSION ANALYSIS

The regression model is a machine learning model which estimates the relationship between the input features (one or more independent variables, ρ_{ij}) and the output value (dependent variable, θ). A regression model is generally used where the prediction values belong to a continuous range and could also be in floating point. The basic idea behind a regression model is to trace a curve or hyper plane in multi-dimensional space based on the data points given during the process of training and then mark the new input point on the same curve to predict the output value[36-39].

The first step is to extract the features from the signals impinging on the microphones of ULA. Then these features along with the correct angle values are used for training the regression model. The features used in training the regression model are correlation co-efficient between the signals captured using the M microphones of ULA. For Mmicrophones' signals there will be $\frac{M!}{2!(M-2)!}$ numbers of correlation efficient. The correlation coefficient measures the linear dependency between two signals. For n^{th} observation, if each microphone signal has K samples then the correlation coefficient is defined as

$$\rho_{ij}^{n} = \frac{1}{K-1}\sum_{k=1}^{K}\left(\frac{\overline{m_i^n(k)-\mu_{m_i^n}}}{\sigma_{m_i^n}}\right)\left(\frac{\overline{m_j^n(k)-\mu_{m_j^n}}}{\sigma_{m_j^n}}\right) 1 \le i, j \le M \text{ and } i \ne j \tag{9}$$

Where, $\mu_{m_i^n}$ and $\sigma_{m_i^n}$ are the mean and standard deviation of i^{th} microphone signal $m_i^n(n)$, respectively, and $\mu_{m_j^n}$ and $\sigma_{m_j^n}$ are the mean and standard deviation of j^{th} microphone signal $m_j^n(n)$ [1-3]. The superscript n indexes the specific observation and $1 \le n \le N$. The multivariate quadratic regression model can be expressed as

$$\theta^{n} = \alpha_{o} + \sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\beta_{ij}\rho_{ij}^{n} + \sum_{\substack{i=1\\ i\ne j}}^{M}\sum_{\substack{j=1\\ i\ne j}}^{M}\sum_{\substack{p=1\\ p\ne q}}^{M}\sum_{\substack{q=1\\ p\ne q}}^{M}\mu\gamma_{ij,\,pq}\rho_{ij}^{n}\rho_{pq}^{n} + \varepsilon^{n}\,,\ \mu = 0 \text{ } or \text{ } 1 \tag{10}$$

which is quadratic polynomial (if $\mu = 1$) in M numbers of variable, where α_o, β_{ij}, $\gamma_{ij,pq}$ and ε are the bias parameter, linear effect parameter, quadratic effect parameter and error term respectively. These parameters will be determined during the training of the regression model. The above regression model is linear as the model function is linear in the model parameters. FromN number of observations, we estimate the regression model parameters

$$\hat{\theta}^{n} = \hat{\alpha}_{o} + \sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\hat{\beta}_{ij}\rho_{ij}^{n} + \sum_{\substack{i=1\\ i\ne j}}^{M}\sum_{\substack{j=1\\ i\ne j}}^{M}\sum_{\substack{p=1\\ p\ne q}}^{M}\sum_{\substack{q=1\\ p\ne q}}^{M}\mu\hat{\gamma}_{ij,\,pq}\rho_{ij}^{n}\rho_{pq}^{n} \tag{11}$$

The error for the n^{th} observation, $\varepsilon^n=\theta^n - \hat{\theta}^n$, is the difference between the true value and the predicted value of the dependent variable. Then ordinary least squares method obtains regression model parameter estimates which minimize the objective function, which is the sum of squared errors (SSE), with respect to the model parameters,

$$SSE = \sum_{n=1}^{N}\left[\varepsilon^{n}\right]^{2} \tag{12}$$

The minimization of this objective function yield the normal equations in the model parameters, which can be solved to estimate the parameters. If $\mu = 0$, then the regression model is named as *polynomial regression of order 1* (PR1) and else if $\mu = 1$, then the regression model is named as *polynomial regression of order 2* (PR2)

4. SIMULATION PARAMETERS AND RESULTS

It has been assumed that the medium of acoustic wave propagation is quiescent, homogeneous and isotropic. A ULA (as shown in Fig. 1) consisting of four microphones (point size) with an Omni-directional beam-pattern that has been placed along the x-axis. The separation d between each of the consecutive microphones has is 10cm. A point size acoustic source in the far-field, transmitting a 1 kHz sinusoidal signal which is traveling at the speed of sound i.e. 343 m/s is considered. The relative attenuation of the signals impinging on the microphones is neglected. All estimates of DOA are considered to be in the

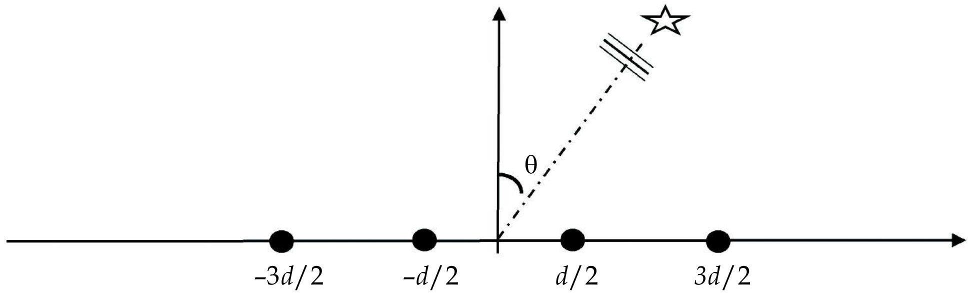


Fig. 1. Uniform linear array of microphones (circle in indicates the microphone), the angle θ is measured with respect to the positive y-axis and the sound source in the far-field is represented with a 6-point star symbol.

clockwise direction with respect to the positive y-axis. The received signals have a duration of 25 ms with a sampling rate of 48 kHz.

The noisy training signals vectors have been generated from ULA's received signal vector after addition of independent zero-mean Gaussian noise (at each microphone)with a range of SNR values. For each DOA, a set of 1400 independent noisy signal vectors have been used for training the regression model and 600 independent noisy signal vectors have been used for testing the regression model for SNR values from 26dB to 10dB decremented in steps of 4dB.

In the regression models, the 46-ary system has been used, where the possible DOA range from 0° to 90°, with a fixed increment of 2° for training the regression model. Regression models have been trained with 46-ary system, where the training DOA angles are from 0° to 90°, incremented in steps of 2°. While, in the 91-ary system, the DOA angles ranging from 0° to 90°, incremented in steps of 1°are used for testing. The features used for the training of the regression model are the cross-correlation coefficient of the signals of every two microphones in the array of four microphones. In the process of testing, the model marks the points for inputs on the curve or hyper plane and predict the output. The performance of the regression models has been tested with the performance of a standard DOA estimation algorithm known as *delay and sum*. DAS method required to search in the space, that searches the maxima of a function for different DOA. During the simulation, we have searched the space from 0° to 90°, with a fixed increment of 1°. The performance of the trained model has been evaluated in terms of the root-mean-square angular error *(RMSAE)* and average RMSAE $(\overline{RMSAE})$, the *RMSAE* is defined as

$$RMSAE(\theta) = \sqrt{\frac{\sum_{i=1}^{N}(\theta - \theta_i)^2}{N}} \tag{13}$$

where, θ_i is the i^{th} prediction for the true angle θ and *N* is the number of predictions from different realization of observed noisy vector for each true angle θ.

The expression of $(\overline{RMSAE})$ is given below

$$\overline{RMSAE} = \frac{1}{NOA}\sum_{\theta=0^\circ}^{90^\circ} RMSAE(\theta) \tag{14}$$

where, *NOA* is the number of true angles, and *RMSAE* (θ) is the root-mean-square angular error at true angle θ.

Figures 2 to 6 shows the RMSAE versus the true DOAs for the two regression models (PR1 and PR2) and DAS for different SNR. For most of the DOA values the regression model gives lower *RMSAE* value, also for the end-fire (*i.e.* DOA close to 90°) the relative performance of regression models improves with the decrease in the SNR. For broadside (*i.e.* DOA close to 90°) the DOA performs better than the regression

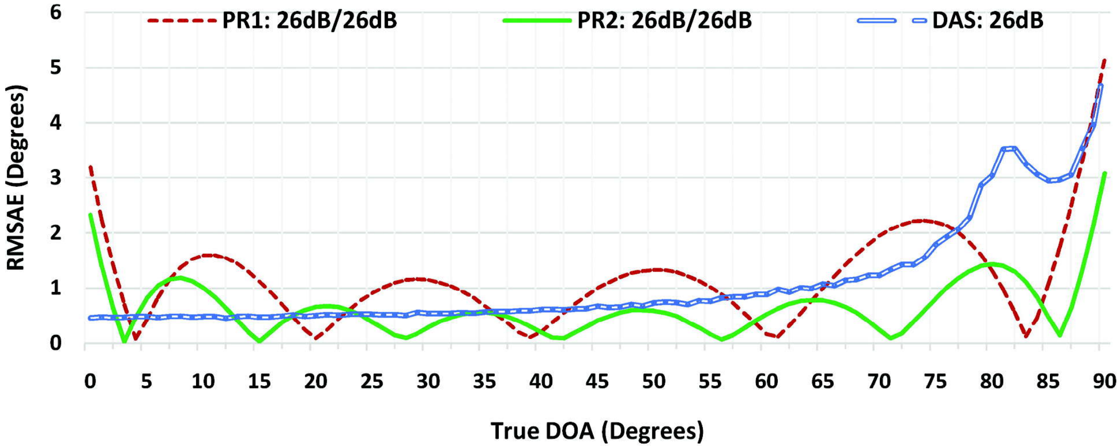


Fig. 2. *RMSAE* versus True DOA for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 26 dB SNR and testing is done 26dB SNR.

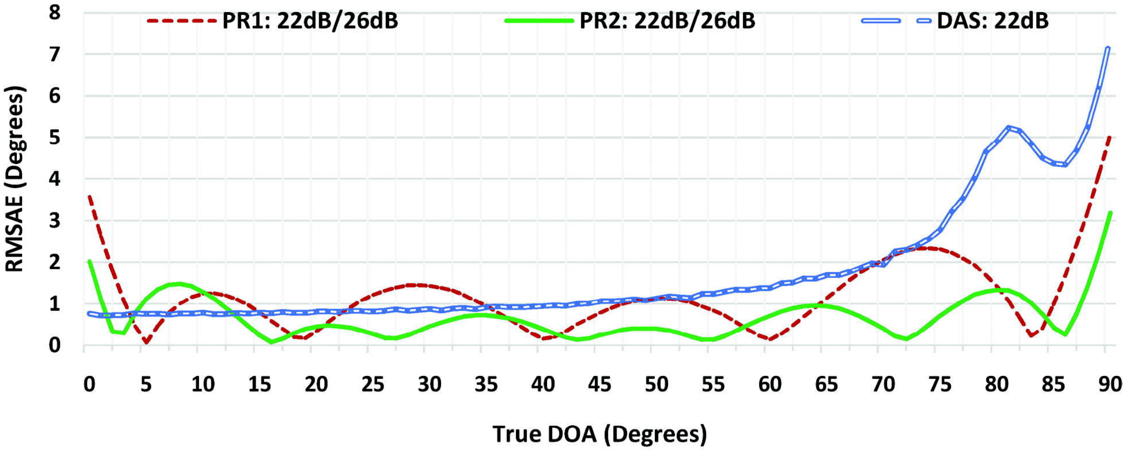


Fig. 3. *RMSAE* versus True DOA for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 26 dB SNR and testing is done 22 dB SNR.

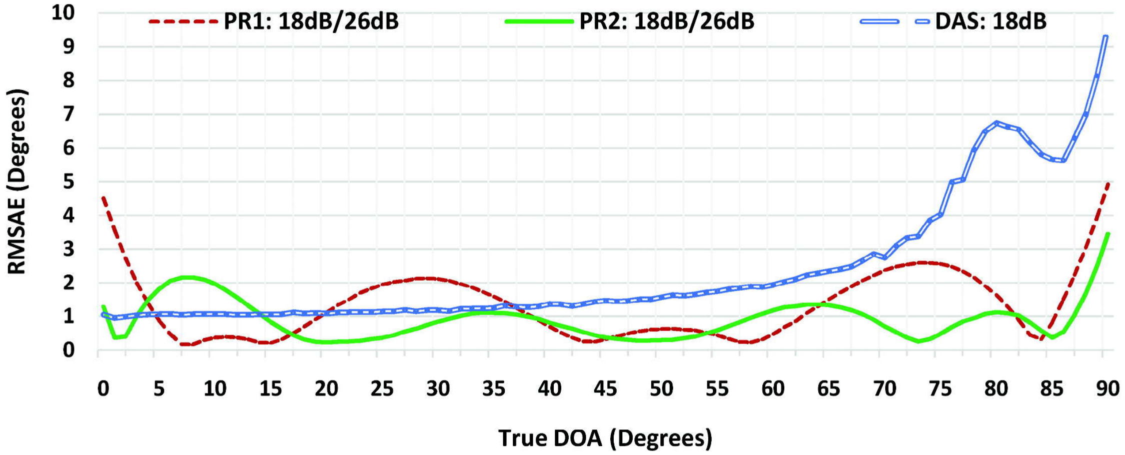


Fig. 4. *RMSAE* versus True DOA for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 26 dB SNR and testing is done 18 dB SNR.

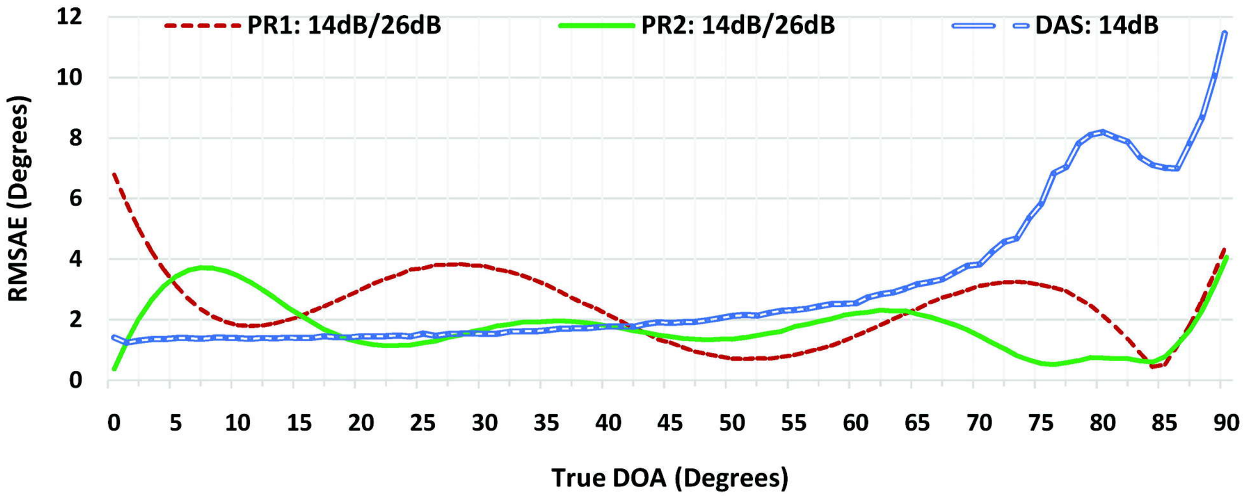


Fig. 5. *RMSAE* versus True DOA for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 26 dB SNR and testing is done 14 dB SNR.

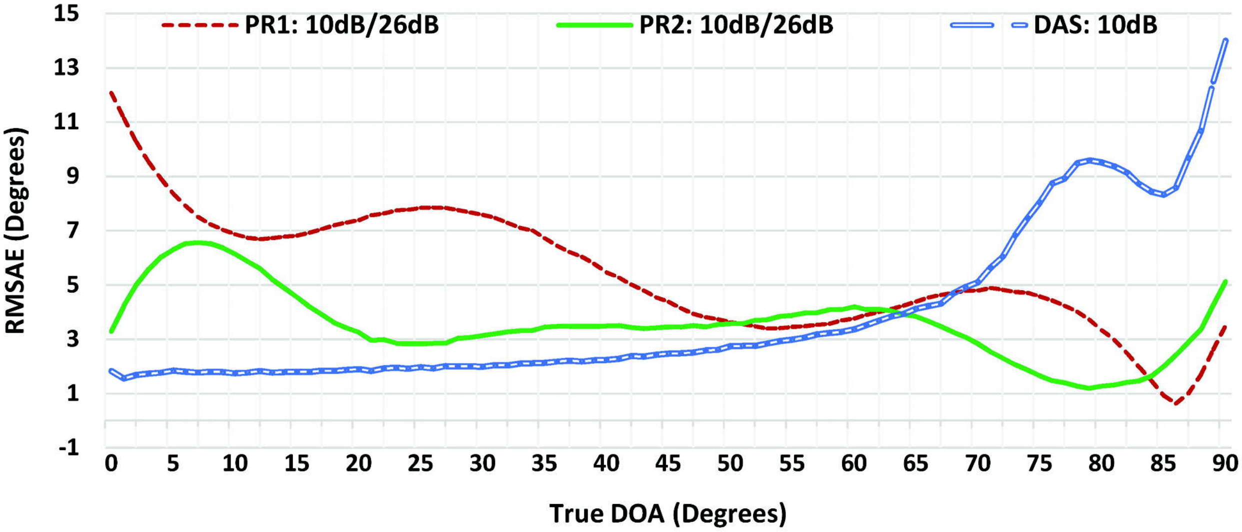


Fig. 6. *RMSAE* versus True DOA for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 26 dB SNR and testing is done 10 dB SNR.

models. The average performance for all DOAs is measured using $(\overline{RMSAE})$, Fig. 7 shows the $\overline{RMSAE}$ versus SNR for regression models and DAS. It has been observed that multivariate quadratic regression model (PR2) is consistently better than the PR1 and DAS for all SNR. However, PR1 is better than the DAS for the SNR values 22 dB, 18 dB and 14 dB, and DAS is better than PR1 for SNR value of 10 dB.

Also, it has been observed that the testing performed for the regression models is better at the SNR value 26 dB at which it is trained, as we decrease the SNR the performance falls down, but PR2 still performs better than PR1 and DAS.

Further, to test the robustness of the regression models we have trained the models with the same parameters as discussed above but at SNR of value 10 dB instead of 26 dB. The results are shown in Fig. 8 and Table 1. These results show that RMSAE is still lower for the PR1 and PR2 than the DAS beam former and relatively PR2 is better than PR1.

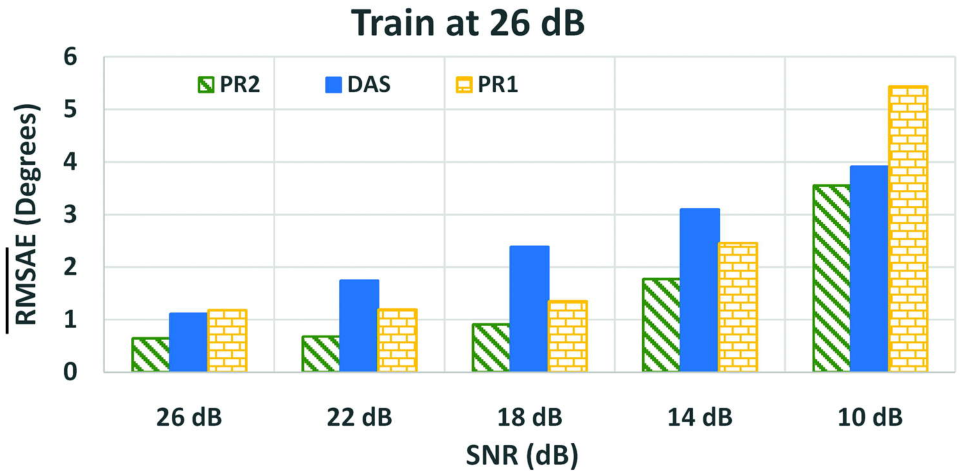


Fig. 7. $\overline{RMSAE}$ versus True DOA for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 26 dB SNR and testing is done for SNR values 26 dB to 10 dB with a decrement 4 dB.

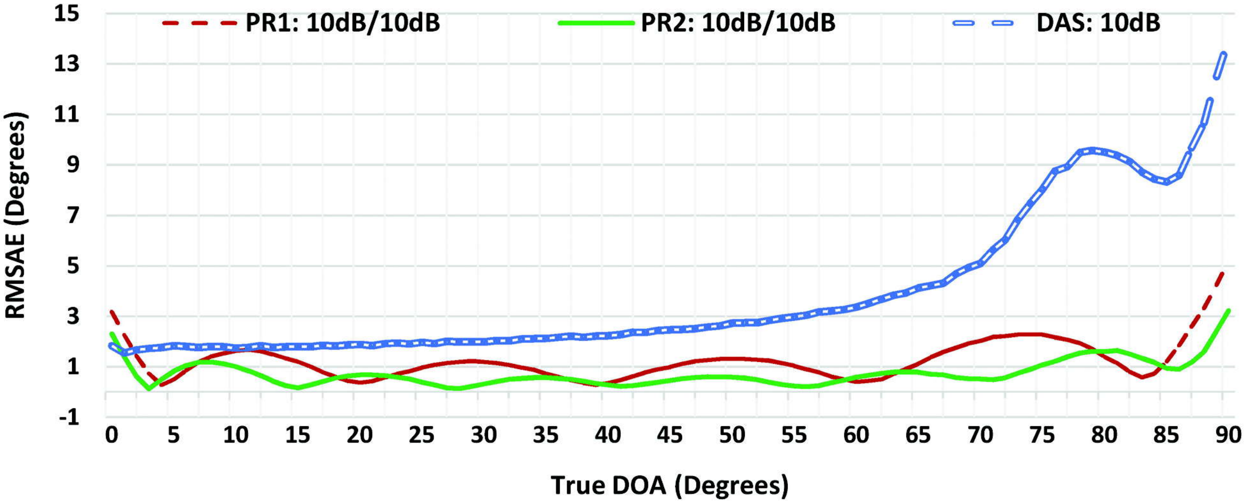


Fig. 8. *RMSAE* versus True DOA for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 10 dB SNR and testing is done 10 dB SNR.

Table 1. $\overline{RMSAE}$ for the two regression models (PR1 and PR2) and DAS. The regression models are trained at 10 dB SNR and testing is done 10 dB SNR.

SNR (dB)	Average RMSAE ($\overline{RMSAE}$) (in degrees)		
	DAS	PR1	PR2
10	3.910	1.273	0.740

5. CONCLUSION

A regression model based approach has been presented for the DOA estimation of an acoustic source using a ULA. For two different regression models, training and testing have been performed for the DOA estimation in the presence of sensor noise. It has been shown that the DOA estimation using the multivariate quadratic regression model with ULA of Omni-directional microphones perform (in terms of *RMSAE*) better than conventional algorithm *i.e.* DAS beam forming. The main advantage of these regression models is that they can predict instantaneous DOA and does not require to search for space as in DAS. Future work can be done on a regression model for range estimation for the near field sources and it can also include practical issues like reverberation and implementation.

6. REFERENCES

[1] X. Zheng, C. Ritz and J. Xi. 2016. Encoding and communicating navigable speech soundfields. *Multimedia Tools and Applications,* **75**(9), 5183-5204.

[2] A. Asaei, M. Taghizadeh, S. Haghighatshoar, B. Raj, H. Bourlard and V. Cevher. 2016. Binary sparse coding of convolutive mixtures for sound localization and separation via spatialization. *IEEE Transactions on Signal Processing,* **64**(3), 567-579.

[3] I. Bekkerman and J. Tabrikian. 2006. Target detection and localization using MIMO radars and sonars. *IEEE Transactions on Signal Processing,* **54**(10), 3873-3883.

[4] K. Wong and M. Zoltowski. 1997. Closed-form underwater acoustic direction-finding with arbitrarily spaced vector hydrophones at unknown locations. *IEEE Journal of Oceanic Engineering,* **22**(3), 566-575.

[5] X. Sheng and Y.H. Hu. 2005. Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. *IEEE Transactions on Signal Processing,* **53**(1), 44-53.

[6] S. Zhao, S. Ahmed, Y. Liang, K. Rupnow, D. Chen and D. Jones. 2012. A real-time 3D sound localization system with miniature microphone array for virtual reality. *In Industrial Electronics and Applications (ICIEA),* 7th *IEEE Conference,* pp. 1853-1857).

[7] J. Clark and G. Tarasek. 2006. Localization of radiating sources along the hull of a submarine using a vector sensor array. *IEEE(OCEANS).* pp. 1-3

[8] R. Carpenter, B. Cray and E. Levine. 2006. Broadband Ocean Acoustic (BOA) Laboratory in Narragansett Bay: preliminary *in-situ* harbor security measurements. *In Photonics for Port and Harbor Security II,* pp. 620409.

[9] J. DiBiase, H. Silverman and M. Brandstein. 2001. Robust localization in reverberant rooms. *Springer,* pp. 157-180.

[10] D. Bechler, M. Schlosser and K. Kroschel. 2004. System for robust 3D speaker tracking using microphone array measurements. *In Intelligent Robots and Systems (IROS 2004). IEEE/RSJ International Conference*, pp. 2117-2122.

[11] S. Argentieri and P. Danes. 2007. Broadband variations of the MUSIC high-resolution method for sound source localization in robotics. *In Intelligent Robots and Systems (IROS 2007). IEEE/RSJ International Conference,* pp. 2009-2014.

[12] K. Nakadai, D. Matsuura, H. Okuno and H. Kitano. 2003. Applying scattering theory to robot audition system: Robust sound source localization and extraction. *In Intelligent Robots and Systems* (IROS 2003). *IEEE/RSJ International Conference*, pp. 1147-1152.

[13] S. Zhao, E. Chng, N. Hieu and H. Li. 2010. A robust real-time sound source localization system for olivia robot. *APSIPA Annual Summit and Conference.*

[14] X. Xiao, S. Zhao, D. Nguyen, X. Zhong, D. Jones, E.S. Chng and H. Li. 2014. The NTU-ADSC systems for reverberation challenge 2014. *REVERB challenge workshop*.

[15] S. Delikaris-Manias, J. Vilkamo and V. Pulkki. 2016. Signal-dependent spatial filtering based on weighted-orthogonal beam formers in the spherical harmonic domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **24**(9), 1511-1523.

[16] S. Delikaris-Manias and V. Pulkki. 2013. Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing,* **21**(11), 2356-2367.

[17] C. Zhang, D. Flor\^encio, D. Ba and Z. Zhang. 2008. Maximum likelihood sound source localization and beam forming for directional microphone arrays in distributed meetings. *IEEE Transactions on Multimedia,* **10**(3), 538-548.

[18] T. Bogaert, E. Carette and J. Wouters. 2011. Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna. *International Journal of Audiology,* **50**(3), 164-176.

[19] B. Widrow. 2000. A microphone array for hearing aids. *In Adaptive Systems for Signal Processing, Communications and Control Symposium* 2000. AS-SPCC. *The IEEE* 2000, pp. 7-11.

[20] M. Wajid, A. Kumar and R. Bahl. 2016. Design and analysis of air acoustic vector-sensor configurations for two-dimensional geometry. *The Journal of the Acoustical Society of America*, **139**(5), 2815-2832.

[21] M. Wajid, A. Kumar and R. Bahl. 2016. Bearing Estimation in a Noisy and Reverberant Environment Using an Air Acoustic Vector Sensor. *IUP Journal of Electrical and Electronics Engineering,* **9**(2), 53.

[22] M. Wajid, A. Kumar and R. Bahl. 2017. Direction-finding accuracy of an air acoustic vector sensor in correlated noise field. In 2017 4th *International Conference on Signal Processing, Computing and Control (ISPCC),* pp. 21-25.

[23] M. Wajid, A. Kumar and R. Bahl. 2017. Direction-of-arrival estimation algorithms using single acoustic vector-sensor. In 2017 *International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT),* pp. 84-88.

[24] S. Yadav, M. Wajid and M. Usman. 2020. Support Vector Machine-Based Direction of Arrival Estimation with Uniform Linear Array. *Springer, Singapore.*

[25] M. Wajid, B. Kumar, A. Goel, A. Kumar and R. Bahl. 2019. Direction of Arrival Estimation with Uniform Linear Array based on Recurrent Neural Network. In 5th *International Conference on Signal Processing, Computing and Control (ISPCC).*

[26] A. Liu, D. Yang, S. Shi, A. Zhu and Y. Li. 2019. Augmented subspace MUSIC method for DOA estimation using acoustic vector sensor array. *IET Radar, Sonar & Navigation,* **13**(6), 969-975.

[27] F. Shi. 2019. Two Dimensional Direction-of-Arrival Estimation Using Compressive Measurements. *IEEE Access,* **7**, 20863-20868.

[28] X. Cui, K. Yu, S. Zhang and H. Wang. 2019. Azimuth-Only Estimation for TDOA-based Direction Finding with Three-Dimensional Acoustic Array. *IEEE Transactions on Instrumentation and Measurement.*

[29] Z. Meng and W. Zhou. 2019. Direction-of-arrival estimation in coprime array using the ESPRIT-based method. *Sensors,* **19**(3), 707.

[30] K. Varma, 2002. Time delay estimate based direction of arrival estimation for speech in reverberant environments. *(Doctoral dissertation, Virginia Tech).*

[31] C. Zhou, Y. Gu, Y. Zhang, Z. Shi, T. Jin and X. Wu. 2017. Compressive sensing-based coprime array direction-of-arrival estimation. *IET Communications,* **11**(11), 1719-1724.

[32] B. Van Veen and K. Buckley. 1988. Beam forming: A versatile approach to spatial filtering. *IEEE assp magazine,* **5**(2), 4-24.

[33] S. Haykin. 1985. Array signal processing. *Englewood Cliffs, NJ, Prentice-Hall*, p. 493.

[34] D. Manolakis, V. Ingle, S. Kogon and others. 2000. Statistical and adaptive signal processing: spectral estimation, signal modeling, adaptive filtering, and array processing. *McGraw-Hill Boston.*

[35] S. Haykin. 1985. Radar array processing for angle of arrival estimation. IN: Array signal processing (A85-43960 21-32). *Englewood Cliffs, NJ, Prentice-Hall,* pp. 194-292. *Research supported by the Canadian Department of Communications and NSERC*, pp. 194-292.

[36] D. Freedman. 2009. Statistical models: theory and practice. Cambridge University Press.

[37] A. Rencher and W. Christensen. 2012. Chapter 10, Multivariate regression-Section 10.1, Introduction. *Methods of multivariate analysis, Wiley Series in Probability and Statistics*, **709**, p. 19.

[38] H. Seal. 1967. Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. *Biometrika,* **54**(1-2), 1-24.

[39] X. Yan and X. Su. 2009. Linear regression analysis: theory and computing. *World Scientific.*

The Journal of Acoustical Society of India : Vol. 47, No. 1, 2020 (pp. 76-85)

Underwater communications: An open challenge

Monika Agrawal
CARE IIT Delhi
e-mail: maggarwal@care.iitd.ernet.in

[Received: 28-10-2019; Revised: 03-03-2020; Accepted: 17-03-2020]

ABSTRACT

Almost seventy percentage of earth is water, mainly comprising of sea/ocean. These giant oceans are unexplored because of various reasons. Recent technology advancement can play a crucial role in unearthing the deep sea. Underwater wireless communication plays a significant role in understanding marine life, water pollution, oil and gas rig exploration, surveillance, naval tactical operations for coastal securities, to observe the changes in the underwater environment and many more. The underwater medium is quite challenging, therefore, realizing even a low data rate, low range communication system is also very challenging. The objective of this paper is to understand these challenges associated with underwater communication and some explore some methods to combat them.

1. INTRODUCTION

Little did Aristotle imagine when he discovered the propagation of acoustic waves through water, in around 400BC, that it would lead to an era of undersea wireless communications. Underwater Wireless Sensor Network (UWSN) is a network of unmanned, unwired, heterogeneously distributed sensors/systems for complete surveillance, sea profiling and many other applications. This completely submerged system is becoming increasingly important not only for defense but also for other commercial and social activities[21]. Physical layer link along with the deployment and maintenance of these completely submerged underwater sensors, limits the range, capacity, applications etc. of these system,

The challenge of the physical layer starts from the selection of carrier waves. Electro-Magnetic waves get absorbed even at a small distance of less than a meter. Light waves provide another possible carrier to carry signal in water though they can carry lots of data but then again the range is very minimal, much lesser than a km. Acoustic waves are the only waves which can travel in water to a reasonable distance.

It was during World War II when underwater acoustic (UWA) communications suddenly became a requirement. 'Gertrude', the first underwater telephone, was developed in 1945 at Naval Underwater Systems Centre, USA.

The modern age of UWA communication initiated with the evolution of digital acoustic modulator-demodulators (modems). The Digital Acoustic Telemetry System (DATS) proposed by the scientists from Massachusetts Institute of Technology (MIT) and Woods Hole Oceanographic Institution (WHOI) was the first one of its kind, and, eventually led to the first generation of commercial acoustic modems[2, 3]. The acoustic telemetry modem (ATM-845) was one of the earliest commercially available modems developed in a collaborative program by WHOI and Datasonics in the late 1980s. It offered a power

Table 1. Off the shelf Underwater acoustic Modems

Underwater Acoustic Modem	Carrier Frequency (kHz)	Bandwidth (kHz)	Maximum Data Rate	Maximum Distance
Aquatec AQUAModem 1000	9.75	4.5	2kbps	5km
DSPComm Aquacomm Marlin	23	14	480bps	1km
DSPComm AquacommMako	23	14	240bps	100m
Evologics S2CR 48/78USBL	48-78	30	31.2kbps	1km
Evologics S2CR 40/80USBL	38-64	26	27.7kbps	1km
Evologics S2CR 18/34USBL	18-34	16	13.9kbps	3.5km
Evologics S2CR 12/24USBL	13-24	11	9.2kbps	6km
Evologics S2CR 7/17USBL	7-17	10	6.9kbps	8km
AM- OFDM-S	21-27	n/a	1.6kbps	4km
LinkQuest UWM 2200	71.4	35.7	35.7kbps	1km
LinkQuest UWM 3000	10	5	5kbps	3km
LinkQuest UWM 3000H	10	5	5kbps	3km
LinkQuest UWM 4000	17	8.5	8.5kbps	4km
LinkQuest UWM 10000	10	5	5kbps	1km
Teledyne Benthos Atm9xx	18.5,24.5,11.5	5	15.36kbps	6km
Teledyne Benthos Atm88x	18.5,24.5,11.5	5	15.36kbps	6km
TriTechMicroModem	22	4	40bps	500m

efficient unit capable of 1,200 bit per sec (bps) data transmission and 80 bps data reception. The Datasonics ATM-850 was an extension to the ATM-845 and was capable of transmitting and receiving data at 2,400 bps. These modems eventually became a basis for many commercial modems. In the last two decades, several commercial modems along with several research modems[11-17] have been launched, Table 1 presents the brief of the same. But even now the area of underwater acoustic communications is challenging and has many open research issue and also implementation problems which need attention. Addressing these challenges will definitely provide a better system for underwater communication.

2. ACOUSTIC WAVES AS CARRIER

Information carrying acoustic waves are pressure waves. Low frequency pressure waves are capable of propagating over long distances in water, but as the frequency increases, they get attenuated/absorbed very fast in this aqueous medium. Therefore, low frequency are best candidate acoustic waves to serve as carrier for communication but these low frequency waves provide very small band width, therefore, designing any reliable long-range high speed underwater communication systems using acoustic signals as carrier is quite a challenging task.

To understand better consider a very generic communication scenario of L transmitters and M receivers. Let $s_l(.)$ denotes an independent and identically distributed (IID) input symbol sequence from the given constellation corresponding to l^{th} transmitter. x(.) denotes the pulses carrying the information. Correspondingly the signal received at the m^{th} receiver at time instant k is given by,

$$r_m(k) = \sum_{l=0}^{L-1} \sum_{n} h_{m,l}(k,n) s_l(n) x(k-nZ) + n_m(k) \tag{1}$$

where $h_{m,l}$ (.)is the channel impulse response between the 1^{th} transmitter and m^{th} receiver, Z is the oversampling rate and n_m (.) denotes underwater channel noise.

2.1 Underwater Channel

The propagation of acoustic signal through underwater channel can be modelled as Linear Time Varying system. This is due to the continuous wave motion and continuous agitation of the reflection points on the sea surface as well as at uneven bathy. This makes the incident acoustic waves scatter in a random fashion. The multipath structure, thus formed, is not stationary and changes rapidly resulting in time-varying impulse response of the channel. Channel structure also depends upon the transmission link configuration *i.e.* horizontal or vertical, the ocean depth, type of sea bathy, range, *etc*. It also depends upon environmental parameters, sound speed profile, and many more.

In horizontal UWA links, multipath spread extends upto several tens or even hundreds of symbol intervals, whereas in vertical link spread is much limited. This is much different from terrestrial RF link, which does not experience such high degree of multipath spread.

The propagating acoustic waves in an UWA channel undergo spreading, refraction, reverberation, dispersion, absorption scattering, etc. These are mainly frequency-dependent losses and also are function of ranges. Therefore, for efficient short-range communications one uses different bandwidth than the corresponding long-range communications. Infact bandwidth available for short range is much more than long range link. Mathematically, the path loss corresponding to signal of frequency f at a distance l is given by

$$A\left(l,f\right)=A_0 l^k a(f)^l$$

In decibels (dB) it is given as,

$$10logA(l,f)\,/\,A_0=k.10logl+l.10loga(f) \qquad (2)$$

The first term on RHS of (2) symbolizes the spreading losses, k represents the geometry of propagation, k=2 for spherical spreading and k=1 for cylindrical spreading and the second term provides the absorption losses. *a(f)* is the absorption coefficient, according to Thorp's formula it is given as[2],

$$10loga\left(f\right)=0.11f^2\,/\,(f^2+1)+44f^2\,/\,(4100+f^2)+2.7510^{-4}f^2+0.003$$

As the medium is bounded, reflections at the boundaries will give rise to multipath and signal will reach to the receiver from the sources through different paths. Ray theory provides the skeleton for determining multipath structure where the overall channel transfer function for P multipath of the channel can be given as,

$$H\left(l,f\right)=\sum_{P=0}^{P-1} r_P\sqrt{A(l_P,f)}e^{-j2\pi f\tau_P}$$

where r_p caters for all other losses and τ_P is the time delay corresponding to the p^{th} path.

Further the low speed of sound in water causes severe Doppler distortion that can either be viewed as a shift in frequency or a scaling of time or both. The effect of Doppler can spread over a few milliseconds (ms). It results in inter-carrier interference (ICI) also known as frequency spreading. Further systems involving moving platforms give adverse Doppler shifts resulting from relative motion between the transmitter and receiver.

A generic channel model is required encompassing all the irregularities, so that one can design ways to handle them which is very much required to design a robust underwater communication system. The underwater channel is very dynamic and many a times it is difficult to incorporate all these effect in a mathematical model. Therefore, experimental data is also used to manifest different stochastic distributions suitable to model the UWA multipath and Doppler structure.

2.2 Channel Noises

The studies related to UWA noise were initiated at the very time when underwater bell-and-hydrophone systems came into use in the early 1900's. Though deployed for picking up the bell's sound to ensure navigational safety, the ship-mounted hydrophones used to pick up background noise as well which made it difficult to detect the actual signal of interest.

In communication background noise plays a very important role, Traditionally, this channel noise is assumed to be additive white Gaussian (AWGN). This assumption of Gaussianity is motivated by the classical central limit theorem (CLT). But noises in an underwater acoustic (UWA) channel often carries impulsive components from various site-specific sporadic sources such as, biological sources, shipping traffic, ice-cracking, earthquakes, underwater explosives, off shore oil exploration-production etc. Impulsive samples from such sources, punctuate the continuous background noise arising from the ocean waves, surface agitation, turbulence, thermal noise *etc*. This often leads to a noise probability density function (pdf) with heavy-tail. Moreover, this infinite variance pdf disobeys the classical CLT. The underwater acoustic channel noise, thus, can no longer be appropriately approximated by traditional Gaussian statistics.

The statistical understanding of these noises is very important for designing any solution to combat this. Based on the estimated density function (pdf) of the noise better receivers can be designed. This impairments of impulsive noise on signal detection can be mitigated either by designing algorithms that can suppress the impulsive behavior and/or alter the noise characteristics to a Gaussian-like behavior so that the standard optimal Gaussian receiver can be reused. One can also develop optimal receivers that can adapt to an impulsive noise environment so as to recover the transmitted information from the noise corrupted signal without changing its statistical characteristics.

Billions and billions of noise data samples are available which can be used to understand the noise statistics, All the open source data form NOAA and Venus site have been used to understand the noise properties[22,23]. The data has been used to estimate the probability density function using histogram. Further refinement to these estimates is achieved using Kernel density, orthogonal polynomial based approaches. The estimated pdf is approximated to popularly known pdfs for better understanding of noise statistics. Table 2 summaries these findings.

Table 2. Shows the probability of the noise samples collected at various location having the stated statistical distribution. These noise samples were collected by the hydrophones having different resonant frequency (mentioned there). Here Very Low (probability<.25), Moderate (probability[.25,.5]), High (probability [.5,.75]) and very high (probability>.75)

Noise Samples	Gaussian	Cauchy	Middletone	Gaussian Mixture of 2	Gaussian mixture of 3	Generalized Gaussian
Gulf of Mexico (5kHz) [22]	Very low	Moderate	Moderate	Moderate	High	High
Gulf of Mexico (5kHz) [22]	Very low	Moderate	Moderate	High	High	High
Barkley Canyon (128kHz) [23]	Very low	Moderate	Moderate	Moderate	High	Moderate
ClayoquotSlop (128kHz) [23]	Very low	Moderate	Moderate	High	High	Moderate
Indian Ocean (100kHz)	Very low	Moderate	Moderate	Moderate	Moderate	High

The results show that underwater noises cannot be modelled as Gaussian noise with very high probability and these non-Gaussian noises require special treatment.

A robust and rugged underwater communication system should be able to overcome all the above mentioned hurdles. In brief the challenges in underwater communication are enormous in number and are of various genre, therefore, a viable solution is only possible by fusing various signal processing, information theoretic, wave propagation based algorithms/concepts, to cater for impulsive noise, multipath spread, Doppler shift *etc*.

3. UNDERWATER ACOUSTIC COMMUNICATION SYSTEM

The challenges are vast and to handle them very robust techniques are required because the medium is changing dynamically.

3.1 Multipath Combating Techniques

The traditional approach for combating ISI (inter symbol interferences) is to use an adaptive equalizer whose tap length is defined by the degree of multipath compensation. In a profoundly dispersive UWA channel, obviously number of equalizer taps inflates. Decision feedback (DFE), Fractionally Spaced equalizer (FSE) have been suggested and are used but still the dynamic variability of underwater channel is difficult to model.

Time reversal (TR) is a simpler and effective technique to handle underwater multipath. It provides spatio-temporal focusing of transmitted energy at the required receiver position. This Spatial focusing improves signal-to-noise-ratio (SNR) and thus, abates fading. Temporal focusing reduces delay spread of the channel, which, in turn, minimizes resultant ISI. These double benefits provide beneficial gains of using Time Reversal (TR) technique in UWA communication.

TR can be implemented at the transmitter in the form of a pre-coder filter, whose impulse response is time-reversed conjugated version of channel impulse response. Effective/Virtual TR can also be implemented at receiver having the time reversed version of impulse response as matched filter at receiver to provide almost similar gain.

Another way to handle this frequency selective underwater channel is by converting the complete underwater channel into large number of orthogonal flat fading channels. Orthogonal Frequency Division Multiplexing scheme (OFDM) partitions the given frequency band into constant magnitude sub-bands. Each sub-band in OFDM techniques is an independent orthogonal sinusoidal carrier. The frequency selectivity of channel is handled by dividing the broadband data into parallel narrowband channels. But the time selectivity is an issue. Doppler introduces large ICI, which introduces more practical hurdle.

Many techniques to combat Doppler such as resampling, Phased Lock Loop (PLL) based compensators etc. have been suggested.

Another novel way is to use frequency sweep signal which are resistant to the detrimental effects of Doppler. Chirp spread spectrum (CSS) based technique offers robust performance with very simple matched filtering based decoder. It offers a preferred solution, which can particularly be adapted for the difficult UWA channel. Recently, OCDM, Orthogonal Chirp division multiplexing, based upon multiplexing chirp signals within the same time slot and bandwidth has been suggested and they provide several performance gains.

Compressive sensing based greedy algorithms like Matching pursuit, Orthogonal Matching pursuit, Compressive Sensing Matching Pursuit, etc. are being tried out to handle this difficult channel.

As the problem quite complex, therefore, one single solution is not sufficient. Various combinations and new insight are very much required for an optimum solution.

3.2 Noise Handling Techniques

Numerous quantitative and qualitative studies support the Non-Gaussian nature of underwater noises. This leads to simple fact that unanimously adopted minimum Euclidean distance AWGN receiver does not perform optimally especially if the impulsive noise component becomes heavy.

This impairment of impulsive noise on signal detection can be mitigated in two ways. One is by developing signal processing algorithms that can suppress the impulsive behavior and/or alter the noise characteristics to a Gaussian-like behavior so that standard signal processing techniques, optimal or sub-optimal in a Gaussian noise environment, can be reused. On the other hand, one can develop optimal or sub-optimal receivers adapted to an impulsive noise environment so as to recover the transmitted information from the noise corrupted signal without changing its statistical characteristics.

Under the first category many filters like Median filter, Laplace filter, etc. have been designed for the purpose of smoothening impulsive noise from a signal[20]. However, many of them filter not only falls short in its ability to smoothen the impulsive noise, but also might remove some significant portions of the signal, in case if the noise parameters deviate from the standard. One class of impulsive noise suppression filter is the myriad filter. It is a non-linear filter which provides an ML estimate of the location of an IID random sequence, called the sample myriad. It is found to be very useful in canceling impulsive noise while designing wireless receivers particularly when the noise is modeled as Symmetric α stable.

Another class of techniques for mitigating non Gaussian noise is by designing the optimal or sub-optimal receivers for the prevailing noise conditions. Designing optimal receivers for non Gaussian noise is highly cumbersome, especially when the underlying noise process is modeled as a stochastic process but does not possess a closed form PDF. The term 'optimal decision rule' refers to a decision criterion which minimizes the probability of wrong decision at the receiver. If the transmitted signal is denoted as *s(k)*, and the received signal is denoted by *r(k)* as *r(k)* = *s(k)* + *n(k)* in the presence of noise *n(k)*.

let $p_n(.)$ be the pdf of the noise. Then the Maximum likelihood (*ML*) optimum detector decides the transmitted signal $\hat{s}_{ML}(k)$ by maximizing

Table 3. Probability of Error for different Noise scenario for M-PAM.

Noise Statistics	Probability of Error
GG	$P_{E-GG}^{M-PAM} = \frac{(M-1)}{M\Gamma(\frac{1}{\beta})}\gamma_u\left[\frac{1}{\beta},\left\{\frac{\varepsilon_b}{\sigma_n^2}\frac{3\Gamma\left(\frac{1}{\beta}\right)\log_2 M}{\left(M^2-1\right)\Gamma\left(\frac{1}{\beta}\right)}\right\}^{\frac{\beta}{2}}\right]$
GM	$P_{E-GM}^{M-PAM} = \frac{2\left(M-1\right)}{M}\sum_{j=1}^{K}\epsilon_j\mathbb{Q}\left(\sqrt{\frac{3\log_2 \mathrm{M}}{\mathrm{M}^2-1}.\frac{\epsilon_\mathrm{b}}{\sigma_\mathrm{nj}^2}}\right)$
BCGM	$P_{E\text{-}BCGM}^{M\text{-}PAM} = \frac{2\left(M-1\right)}{M}(1-\rho)Q\left(\sqrt{\frac{3\log_2 M}{2(M^2-1)}.\frac{\grave{o}_b}{\sigma_{nj}^2}}\right)+\frac{\rho}{2}-\frac{\rho}{\pi}\tan^{-1}\tan^{-1}\left(\sqrt{\frac{3\log_2 M}{2(M^2-1)}.\frac{\grave{o}_b}{\sigma_{nj}^2}}\right)$

Table 4. Probability of Error for different Noise scenario for M-QAM.

Noise Statistics	Probability of Error
GG	$P_{E-GG}^{M-QAM} = \frac{2(\sqrt{M}-1)}{\sqrt{M}\Gamma(\frac{1}{\beta})}\gamma_u\left[\frac{1}{\beta},\left\{\frac{\varepsilon_b}{\sigma_n^2}\frac{3\Gamma\left(\frac{3}{\beta}\right)\log_2\sqrt{M}}{\left(M^2-1\right)\Gamma\left(\frac{1}{\beta}\right)}\right\}^{\frac{\beta}{2}}\right]\left(1-\frac{(\sqrt{M}-1)}{\sqrt{M}\Gamma(\frac{1}{\beta})}\gamma_u\left[\frac{1}{\beta},\left\{\frac{\varepsilon_b}{\sigma_n^2}\frac{3\Gamma\left(\frac{3}{\beta}\right)\log_2\sqrt{M}}{\left(M^2-1\right)\Gamma\left(\frac{1}{\beta}\right)}\right\}^{\frac{\beta}{2}}\right]\right)$
GM	$P_{E-GM}^{M-QAM} = \frac{4\left(\sqrt{M}-1\right)}{\sqrt{M}}\sum_{j=1}^{K}\epsilon_j\mathbb{Q}\left(\sqrt{\frac{3\log_2\sqrt{\mathrm{M}}}{2(\mathrm{M}-1)}.\frac{\epsilon_\mathrm{b}}{\sigma_\mathrm{nj}^2}}\right)\left\{1-\frac{\left(\sqrt{M}-1\right)}{\sqrt{M}}\sum_{j=1}^{K}\epsilon_j\mathbb{Q}\left(\sqrt{\frac{3\log_2\sqrt{\mathrm{M}}}{2(\mathrm{M}-1)}.\frac{\epsilon_\mathrm{b}}{\sigma_\mathrm{nj}^2}}\right)\right\}$
BCGM	$P_{E\text{-}BCGM}^{M\text{-}QAM} = \frac{4\left(\sqrt{M}-1\right)}{\sqrt{M}}(1-\rho)Q\left(\sqrt{\frac{3\log_2 M}{4(M^2-1)}.\frac{\grave{o}_b}{\sigma_{nj}^2}}\right)+\frac{\rho}{2}-\frac{\rho}{\pi}\tan^{-1}\tan^{-1}\left(\sqrt{\frac{3\log_2 M}{2(M-1)}.\frac{\grave{o}_b}{\sigma_{nj}^2}}\right)$

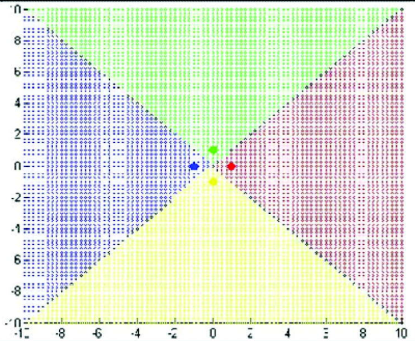
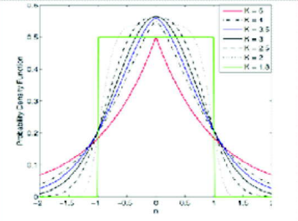
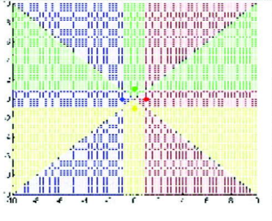
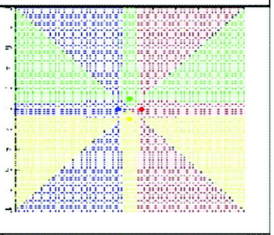
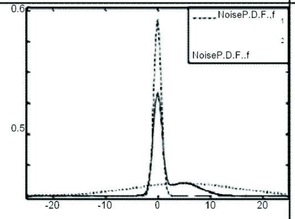
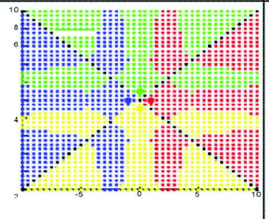
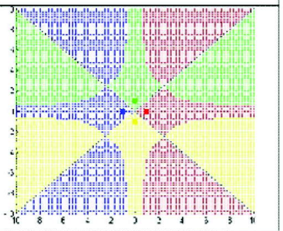
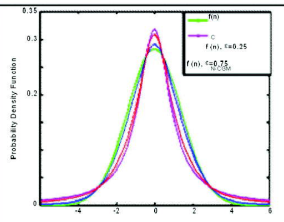
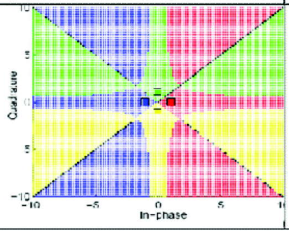
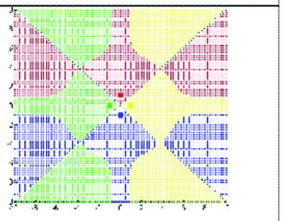
$$\hat{s}_{ML}(k) = \arg\{\max\max\left\{p_n\left(\frac{r(k)}{s(k)}\right)\right\}\}$$

The above expression cannot be simply maximized, it depends on noise statistics and many more parameters. Table 3 and Table 4 summarize the probability of error in underwater communication for the prevalent underwater noises for M-PAM and M-QAM constellation. All the decision rules will depend upon the type of signal constellation used, and noise statistics. Similarly, the MAP decision device is obtained by maximizing

$$\hat{s}_{MAP}(k) = \arg\{\max\max\left\{p_n\left(\frac{s(k)}{r(k)}\right)\right\}\}$$

The decision device is very much dependent upon noise and signal constellation. Table 5 summarizes some optimum decision devices for some typical scenarios. It is clear that the optimum decision rule very much depends upon the noise statistics, different noise pdf results in different decision devices.

Table 5. Decision device for different constellation

Transmitted Constellation			
Noise Statistics	Noise probability density function	Received Constellation-1	Received Constellation-2
GG			
GM			
BCGM			

This requires some adaptive model in designing the receiver to combat the noise.

4. RELIABLE WIRELESS UNDERWATER LINK

Along with the crucial and unique issues of underwater acoustic communication, *i.e.* the impulsive noise, and, the intensive multipath, there are other deterrent issues like timing and carrier synchronization in designing a robust underwater communication system. Therefore, in order to establish reliable wireless UWA link, the system has to cater many other communication impairments also. Handling synchronization becomes much more challenging in the presence of highly dispersive multipath channel and non Gaussian noise. Many algorithms have been suggested for the same.

Channel error correction coding is another important block of this complete reliable system. Many channel coding methods are borrowed from RF communication systems and used here to get required performance. The block diagram of the complete system to provide an acoustic underwater communication link is shown in figure.

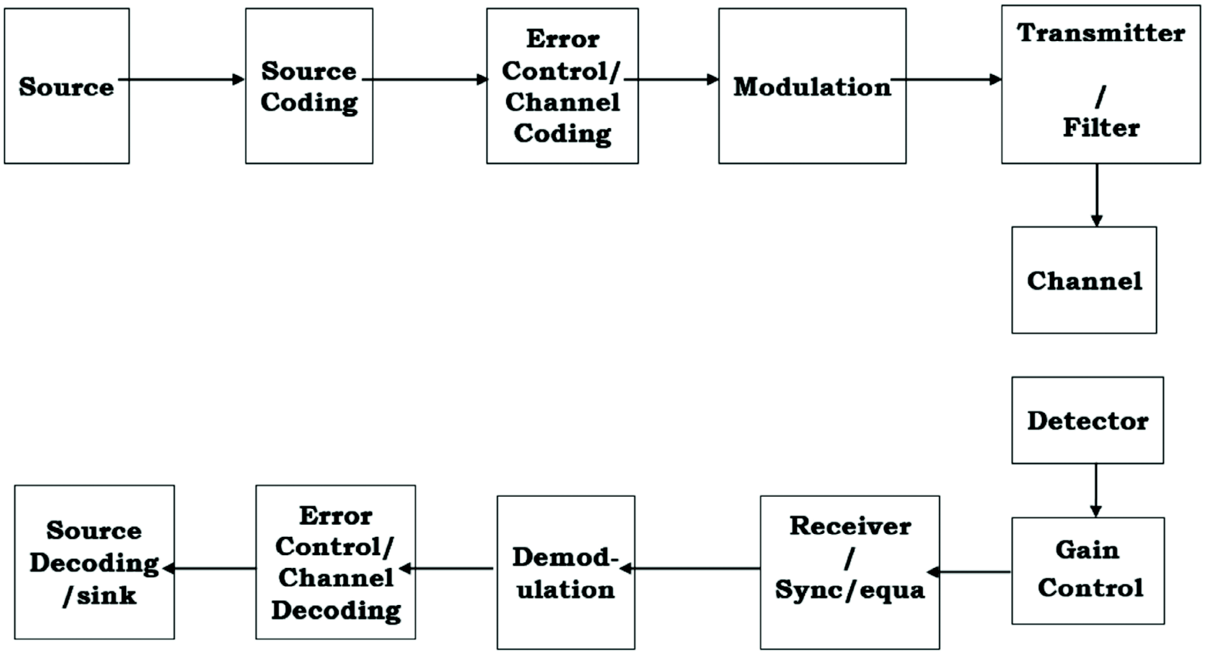


Fig. 1. Block Diagram of complete Communication system

This system has been realized and tested in the fields[24]. It has been designed to achieve the maximum data rate of 5000 bits/second (with Constellation of BPSK/QPSK) for a range if 5km. The system was tested at Chennai, India in Jan 2018 using two moored fishing boats. The transducers were lowered using cables from the boats which are kept stationary. The summary of the results is given table 6.

Table 6. Results of the experiment conducted in Bay or Bengal near Chennai.

Data rate (symbols/s)	Distance (km)	Bit error rate	Modulation
100	3	0	BPSK
500	3	0	BPSK
3000	3	0	BPSK/QPSK
3000	5	0	8PSK
4000	5	.0025	QPSK

5. CONCLUSIONS

The practical UWA channels is quite complex, the techniques have been suggested to combat the combined effect of impulsive noise, multipath spread, Doppler, etc. in real time. But the dynamic nature of Oceans/Sea makes underwater acoustic system more complex. Non-stationary, non-Gaussian underwater noises requires dynamic learning of environment which is very much essential for a viable solution. Recent advances in Artificial Intelligence and big data have opened new opportunities here also, which can be used to design smart and reliable system.

6. REFERENCES

[1] Stojanovic Milica and James Preisig. 2009. "Underwater acoustic communication channels: Propagation models and statistical characterization," *IEEE communications magazine,* **47**(1): 84-89.

[2] Khan Muhammad Waqas, Yimin Zhou and Guoqing Xu. 2014. "Modeling of acoustic propagation channel in underwater wireless sensor networks," 2nd *International Conference on Systems and Informatics (ICSAI 2014).*

[3] Ross Donald. 2013. Mechanics of underwater noise. *Elsevier.*

[4] Bhadouria Vijay Singh, *et al.* 2017. "Quantitative analysis of the effect Time Reversal Mirror on coherence bandwidth and delay spread," *OCEANS 2017-Aberdeen. IEEE.*

[5] S.S. Lokesh, Arun Kumar and Monika Agrawal. 2008. "Structure of an optimum linear precoder and its application to ML equalizer," *IEEE Transactions on Signal Processing,* **56**(8): 3690-3701.

[6] Nee, Richard van and Ramjee Prasad. 2000. OFDM for wireless multimedia communications. *Artech House, Inc.*

[7] Ouyang Xing and Jian Zhao. 2016. "Orthogonal chirp division multiplexing," *IEEE Transactions on Communications,* **64**(9): 3946-3957.

[8] Banerjee, Sharbari, and Monika Agrawal, "A Simple Analytical Design Approach to Space Time Trellis Codes," *Wireless personal communications.* **75**(2): 1141-1154.

[9] Chitre Mandar, John Potter and Ong Sim Heng. 2004. "Underwater acoustic channel characterisation for medium-range shallow water communications," *Oceans' 04 MTS/IEEE Techno-Ocean'0*4 (IEEE Cat. No. 04CH37600). **1**.

[10] Reynolds Douglas, 2015. "Gaussian mixture models," *Encyclopedia of biometrics,* pp. 827-832.

[11] F. Fauziya, Brejesh Lall and Monika Agrawal, 2018. "Impact of vector sensor on underwater acoustic communications system," *IET Radar, Sonar & amp*; Navigation, 2018.

[12] A. Goel, P. Gupta and M. Agrawal, 2013. â€œSER Analysis of PTS Based Techniques for PAPR Reduction in OFDM Systems, â€. *Digital Signal Processing, Elsevier,* **23**(1): 302-313.

[13] Sharbari Banerjee, Monika Agrawal and F. Fauziya, 2017. "A Generalized Gaussian Noise Receiver for Improved Underwater Communication in Leptokurtic Noise," *OCEANS 2017 MTS/IEEE Aberdeen,* pp. 1â€"8.

[14] S. Banerjee and M. Agrawal, 2014. "On the performance of underwater communication system in noise with Gaussian Mixture statistics," *Communications (NCC), 2014 Twentieth National Conference,* **1**: 6.

[15] S. Banerjee and M. Agrawal, 2013. "Underwater Acoustic Noise with Generalized Gaussian Statistics: Effects on Error Performance," *Oceans 2013, Norway.*

[16] S. Banerjee and M. Agrawal, 2013. "Underwater Acoustic Noise with Generalized Gaussian Statistics: Effects on Error Performance, â€œA Time Reversal Technique for minimizing Equalizer complexity in High Rate Multi-antenna UWA Link, â€. *NCC* 2013.

[17] S. Banerjee and M. Agrawal, 2013. "On the Performance of Underwater Communication System in Noise with Gaussian Mixture Statistics," *Oceans 2013, San Diego.*

[18] S. Banerjee and M. Agrawal, 2012. "Time Reversal Precoder: An Efficient Tool for more Reliable Underwater Acoustic Communication," *Oceans 2012, Korea.*

[19]. Divya Pratap Singh Parihar, Ankit Agarwal and Monika Agrawal, 2010. â€œ Time reversal mirror: Temporal and spatial focusing tool, â€. *Oceans* 2010.

[20] Santosh Biradar, Monika Agrawal, Arun Kumar and Rajendar Bahl, 2007. â€œ Design of Data packet for Time delay and Doppler estimation for Underwater Acoustic communication, â€. *Sympol* 2007.

[21] A. Song, M. Stojanovic and M. Chitre, "Editorial Underwater Acoustic Communications: Where We Stand and What Is Next?," *IEEE Journal of Oceanic Engineering,* **44**(1): 1-6.

[22] National Centers for Environmental Information, https://ngdc.noaa.gov/

[23] Ocean Data Network, https://www.oceannetworks.ca/venus-data-now-available-oceans-20

[24] Monika Agrawal *et.al.*, Development of Underwater Acoustic Communication System based on Time Reversal Mirror Phase-II, sponsored by NIOT, Chennai India.

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)