International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0279

e ISSN: 2584-2854
Volume: 03
Issue: 05 May 2025
Page No: 1765 - 1773

A Review on Speaker Diarization for Whispered Speech Audio

Mr. Chaitanya Pampana[1], Dr. M. Vijay Reddy[2], Dr. K. Jhansi Rani[3]
[1]Research Scholar, Gandhi Institute of Engineering and Technology, Gunupur, Odisha, India.
[2]Professor, Gandhi Institute of Engineering and Technology, Gunupur, Odisha, India.
[3]Assistant Professor, Jawaharlal Nehru Technological University, Kakinada, AP, India.
***Emails:** chaitanya.pampana@giet.edu[1], mvijayreddy@giet.edu[2], jhansikaka@jntucek.ac.in[3]*

Abstract

Speaker diarization, the process of partitioning an audio stream into segments according to the speaker identity, is crucial for various applications in speech processing and analysis. Whispered speech, characterized by its low amplitude and altered spectral properties, presents unique challenges for conventional diarization algorithms designed for clear, normal speech. In this study, I propose a novel approach for supervised speaker diarization specifically tailored to whispered speech audio streams. Supervised learning techniques, utilizing annotated data to train models capable of accurately distinguishing between speakers in whispered speech recordings. The design incorporates extraction techniques that effectively capture the faint spectral cues present in whispered speech, hence augmenting the diarization system's discriminative ability. Furthermore, I investigate the combination of acoustic modeling and domain-specific knowledge to enhance diarization performance in whispered speech scenarios. The suggested strategy on a variety of whispered voice datasets, contrasting its effectiveness with cutting-edge diarization techniques. The precision with which whispered speech can be divided into speaker-specific intervals using a supervised technique. Analyze the effects of various variables on diarization performance, including feature representations and dataset properties. The findings of this research contribute to advancing speaker diarization technology, particularly in challenging acoustic environments characterized by whispered speech. The proposed supervised approach holds promise for practical applications in surveillance, forensic analysis, and human-computer interaction, where accurate speaker segmentation in whispered speech recordings is essential.

***Keywords:** Speaker diarization; feature extraction; Voice activity detection; Deep neural network; Speaker clustering; Diarization Error Rate.*

1. Introduction

Speaker diarization plays a pivotal role in speech processing by segmenting audio recordings based on speaker labels, addressing the question of "who spoke when." This process is distinct from speech recognition, as it focuses on identifying speakers rather than transcribing speech content. Prior to speaker diarization, knowledge about "what is spoken" exists, but the speaker's identity is unknown [1]. Integrating speaker diarization with a speech recognition system enriches transcriptions by providing speaker-specific labels, offering a comprehensive understanding of both spoken content and the associated speakers [2]. The technology is widely applied in transcription services, call center analytics, meeting analysis, and other domains where discerning speakers and understanding conversational dynamics are crucial for extracting meaningful insights from audio data. Speaker diarization models play a pivotal role in identifying regions of homogeneous speakers, employing both unsupervised and supervised techniques. These models are particularly designed for digital representation, making them well-suited for the distinctive nature of speech signals [3]. The process of speech recognition involves converting acoustic signals from devices like telephones or microphones into coherent word sequences (Figure 1).

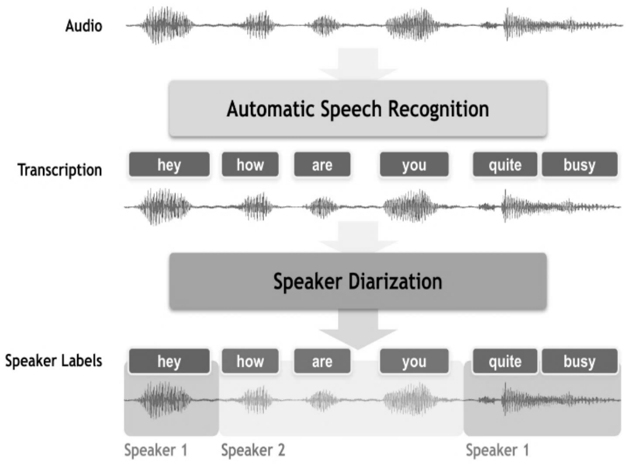

Figure 1 Speaker Diarization System

In domains like Interactive Voice Response (IVR) and Text-to-Speech (TTS) models, automated speech recognition is rapidly advancing within the realms of engineering and speech science [4]. The evolution of Automatic Speech Recognition (ASR) has found diverse applications in everyday life, impacting various aspects of computer science. It is applied in phone applications tailored for individuals with physical challenges or limited education. Speech recognition, functioning as both input and output in Human-Computer Interaction (HCI) [1], plays a vital role in tasks where understanding complex words or challenging written content is essential. To address these challenges, integrating an audio diarization model becomes indispensable. Speaker diarization [2], characterized as a method for interpreting input audio across chronological domains of signal energy and distinct sources, serves to segment audio signals into homogeneous pieces based on audio recognition. This methodology proves valuable in enhancing the comprehension and processing of speech signals, particularly in dealing with intricate words and diverse datasets.

1.1.Need for Speaker Diarization

The automated process of speaker diarization focuses on identifying speakers and determining their speaking intervals within a given audio recording. With the continuous increase in audio data volume, this technique has become essential, finding practical applications in diverse domains like information retrieval, conversation analysis, and meeting annotations. A recent improvement, known as Rich Transcription (RT), enhances speech-to-text conversions by incorporating speaker indexes. To fulfill these needs, various speaker clustering methods, including Agglomerative Hierarchical Clustering (AHC), top-down and bottom-up approaches, as well as global optimization strategies, have been developed [7]. Numerous methodologies are dedicated to the analysis of speech signals, encompassing vocal signals emitted by speakers. These vocal signals typically constitute single-channel inputs, encompassing a mix of diverse audios such as music and noises [8]. It's crucial to recognize that the format of these audio inputs is customized for particular uses rather than adhering to standardized sources [9]. Ongoing research has shaped speaker diarization, involving processes like indexing, clustering, and segmentation of audio [11]. This facilitates the systematic partitioning of audio into coherent segments by discerning speech signals from speakers and non-speech signals. Two pivotal techniques, namely Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) play a crucial role in enhancing the efficiency of audio diarization by simplifying the segmentation and clustering of audio [10-12]. Addressing the complexity of noise represents a significant challenge handled by various audio diarization methods [13]. In the realm of broadcasting applications, challenges such as determining the application model and ensuring audio signal quality prevail [14]. Other techniques underscore the imperative of efficient speaker diarization methods [15], making transcriptions more interpretable and effective. Consequently, optimization techniques [16] are applied to tackle classical problems and introduce intelligent technologies. This presents substantial opportunities for researchers to devise new models aimed at augmenting the efficacy of speaker diarization [5-6]. Voice disorder assessment deploys diverse techniques focusing on linear features of voice signals and amplitude perturbation [17]. Among these, wavelet decomposition emerges as a pivotal method for feature extraction. Various wavelet models, including wavelet packets and Discrete Wavelet Transforms (DWT), find application in this context [18].

1.2.Problems with Current Diarization Techniques

The issues faced by speaker diarization techniques are enlisted.

- The speaker diarization model faces several issues, like audio stream, which comprises overlapped speech and input audio heterogeneity [19].
- In [1], active learning technique supposed that the assistance of human for offering ideal answers to any query pair is complex task. In addition, the errors caused by human are expected.
- The DE-based K-means technique failed to offer improved outcomes all time. Hence, the issues rely on testing the hybridized method with other clustering validity criterion, like Clustering Separation criterion (CS) and Davies and Bouldin (DB) index [20].
- In [4], the GMM did not discover the speaker whenever various speakers speak concurrently. Hence, the issues rely on planning to combine overlapping speech discovery in video in which various speakers speak concurrently.

2. Research Problem

The objective is to create a speaker diarization method utilising a collection of audio samples. Feature extraction, SAD (Speech Activity Detection), speaker segmentation, and clustering are the stages in the speaker diarization process that entail processing. The input audio samples are initially given to the feature extraction phase, which then mines pertinent features and extracts MFCC, LPCC, and LSF. Then, using the processes for removing music and silence, speech activity detection is done to locate speech activity. After locating the speech activity, the speakers are segmented based on the score. Finally, design and develop effective Speaker Diarization with effective Speaker Segmentation and clustering to overcome overlapping speach problem and huge amount of audio data [21-23].

3. Feature Extraction

Feature extraction in speaker diarization involves converting the raw audio signal into feature vectors, which are numerical representations of specific characteristics of the sound [24]. Typically, these features are based on acoustic properties, such as spectral characteristics, which capture information about the frequency content of the audio over time. Spectral characteristics include features like the distribution of energy across different frequency bands, the shape of spectral peaks, and the variation of spectral content over time. By extracting these features, the diarization system can capture distinctive patterns in the audio that are indicative of different speakers. This process is crucial for subsequent stages of speaker diarization, where algorithms analyze these feature vectors to determine speaker turns and segment the audio accordingly. Feature extraction serves as a crucial preprocessing step, facilitating the identification and discrimination of speakers within the audio recording [25].

3.1.Segmentation

In the segmentation stage of speaker diarization, the continuous audio stream is divided into shorter segments, typically spanning from milliseconds to seconds in duration. This partitioning allows for a more granular analysis of the audio data, making it easier to identify distinct speaker turns and transitions between speakers. The segmentation process involves breaking down the audio stream at points where there are pauses or changes in speaker activity, creating discrete segments for further analysis. These segments are often chosen to be short enough to capture rapid changes in speaker identity or speech content while still providing enough contextual information for accurate speaker clustering. By segmenting the audio stream, the diarization system can better isolate individual speech segments for subsequent speaker recognition and clustering. This step is crucial for accurately identifying and distinguishing between different speakers within the audio recording, laying the groundwork for subsequent stages of the diarization process [26].

3.2.Clustering

In the clustering stage of speaker diarization, the segments generated during the segmentation process are grouped together based on similarities in their acoustic features. Segments exhibiting similar acoustic properties, such as spectral characteristics and energy distribution, are presumed to originate from the same speaker and are clustered accordingly. This process involves analyzing the feature vectors extracted from each segment and comparing them to identify patterns and similarities. Segments with closer acoustic feature representations are grouped together into clusters, suggesting they likely belong to the same speaker. Clustering algorithms, such as k-means or Gaussian mixture models, are commonly employed to partition the segments into speaker-specific clusters [27-30]. By organizing segments into clusters, the diarization system can effectively group together speech segments from individual speakers, laying the foundation for speaker identification and labeling. This stage is critical for accurately delineating speaker boundaries and distinguishing between different speakers within the audio recording.

3.3.Speaker Labeling

In the speaker labeling stage of speaker diarization, each cluster generated during the clustering process is assigned a speaker label. Given that speaker identities are typically unknown at the outset, generic labels such as Speaker 1, Speaker 2, and so forth are assigned to the clusters. These labels serve as placeholders to represent each distinct speaker segment identified within the audio recording. The assignment of speaker labels is based on the assumption that each cluster corresponds to a unique speaker, as inferred from the similarities in acoustic features. While initially generic, these labels provide a means of distinguishing between different speaker segments within the audio stream. As speaker identities become more apparent through subsequent analysis or user input, these generic labels can be replaced with more specific speaker identities. Speaker labeling is a crucial step in the diarization process, as it facilitates the organization and identification of individual speakers within the audio recording, enabling downstream tasks such as speaker recognition and transcription.

3.4.Refinement

In the refinement stage of speaker diarization, the system enhances its initial clustering and labeling by incorporating additional information. This can include factors such as the evolution of speaker characteristics over time or the utilization of language models to improve accuracy. Through iterative processes, the system may reassess the clustering of segments, considering factors such as speaker pitch, speaking rate, or other temporal patterns indicative of speaker identity. Additionally, language models can aid in identifying linguistic patterns or vocabulary preferences unique to individual speakers, further refining the speaker labels. By leveraging such information, the system iteratively updates its clustering and labeling decisions, aiming to achieve a more accurate representation of the speakers present in the audio recording. This refinement process enhances the overall performance of the speaker diarization system, improving its ability to accurately segment and label speaker turns within the audio data. The development of diarization methods started more than a decade ago, intensive research continues to improve the accuracy and computational efficiency of diarization algorithms. Diarization systems mainly utilize unsupervised machine learning algorithms when utterances are shared between speakers, but it is not known which diarization label applies to a particular speaker. This approach is known as unsupervised diarization. However, in some applications, it is necessary to identify multiple speakers that interact freely in an audio recording. This task can be completed using a supervised diarization approach, which combines diarization and identification methods. The main distinction between unsupervised and supervised diarization involves the different ways of segment indexation. The former relies on grouping of similar segments into separate categories (clustering), while the latter requires matching a segment with a certain speaker on the basis of voice samples (classification). Thus, the last transformation step for supervised diarization is performed with the help of an additional output module the speaker‘s classifier. However, the identification of the speaker using supervised diarization has a significant peculiarity, as it is necessary, unlike in the common biometric voice recognition systems, to analyze the utterances of more than one subjects. This is a challenge, especially under low-data conditions, as new speakers are represented only by one short audio recording lasting dozens of seconds at most. In the developing modern digital world, application scenarios of diarization systems are constantly expanding, which often appear as components of more complex information systems in human–machine communication. Therefore, research aimed at the development and implementation of a flexible and open-source system capable of solving the main problems of diarization, both in a supervised and unsupervised manner, is relevant. Recent advancements in speaker recognition have been driven by the need to more effectively handle variability between train and test speech samples. Research efforts have focused predominantly on sources of variability external to the speaker, such as channel, language, and duration. Within-speaker sources of variability, such as vocal effort, emotion, and aging, have received less research attention. Whisper is a common mode of low vocal effort speech, generally produced with the aim of maintaining intelligibility for an intended listener, while restricting intelligibility for others. It is therefore utilized in scenarios where a speaker wishes to conceal or disguise their identity, or communicate information discretely. Such scenarios are commonly encountered in forensic applications of speaker recognition, and therefore it is important to establish the effect of whisper in this context. Whisper differs significantly from neutral speech (defined here as modal speech produced with normal vocal effort). Whisper is produced without vibration of the vocal folds, and therefore the signal contains no periodic excitation. Additionally, the center frequencies and bandwidths of formants generally increase in whisper, while the overall signal energy decreases. Is it really necessary to detect whisper for speech recognition? Yes, the main objective is to improve the current speaker diarization accuracy by investigating appropriate approaches and calibrate the whisper presence in speech.

4. Literature Survey

The first automatic speaker recognition (ASR) system came into existence in 1962 through an article by Lawrence G. Kersta, a Bell Laboratories physicist designated, “Voiceprint Identification”. In 1960, Gunnar Fant developed a physiological model of humans voice production system, which sets a speech analysis base. The speaker recognition system’s evolution from the late 1900s to the early 2000s. "Speaker Diarization: A Review of Recent Research" by Fred Richardson and Douglas Reynolds (2018): This comprehensive review covers various aspects of speaker diarization, including techniques, challenges, and applications. While it primarily focuses on general speaker diarization methods, it provides valuable insights into the state-of-the-art techniques applicable to whispered speech. "Whispered Speech Processing: A Review" by Bhiksha Raj and Rita Singh (2014): This paper discusses the challenges and techniques specific to processing whispered speech. It covers aspects such as feature extraction, modeling, and recognition in whispered speech scenarios, offering valuable background information for research in whispered speech diarization. "Supervised Learning for Speaker Diarization" by Hervé Bredin et al. (2017): This paper explores supervised learning approaches for speaker diarization, focusing on techniques such as deep neural networks (DNNs) and convolutional neural networks (CNNs). While it doesn't specifically address whispered speech, the principles and methodologies discussed are relevant for adapting supervised techniques to whispered speech diarization. "Deep Neural Network Embeddings for Text-Independent Speaker Verification" by Li Wan et al. (2018): Although focused on speaker verification, this paper introduces deep neural network embeddings for speaker representation. The techniques presented could be adapted for speaker diarization, including whispered speech, by leveraging the discriminative power of deep embeddings. "A Comparative Study of Recent Speaker Diarization Systems on Real and Simulated Data"by R. Gangashetty et al. (2019): This study evaluates the performance of various speaker diarization systems on both real and simulated data. While not specifically addressing whispered speech, the findings offer insights into the strengths and limitations of different diarization approaches, which can inform research in whispered speech diarization. "Deep Neural Network Embeddings for Text-Independent Speaker Verification" by Li Wan et al. (2018): Although focused on speaker verification, this paper introduces deep neural network embeddings for speaker representation. The techniques presented could be adapted for speaker diarization, including whispered speech, by leveraging the discriminative power of deep embeddings. "A Comparative Study of Recent Speaker Diarization Systems on Real and Simulated Data" by R. Gangashetty et al. (2019): This study evaluates the performance of various speaker diarization systems on both real and simulated data. While not specifically addressing whispered speech, the findings offer insights into the strengths and limitations of different diarization approaches, which can inform research in whispered speech diarization. "Whispered Speech Recognition Using Deep Recurrent Neural Networks with Attention" by Wei Xiong et al. (2017): Although focusing on whispered speech recognition, this paper introduces deep recurrent neural networks with attention mechanisms. These techniques could be adapted for speaker diarization by incorporating attention mechanisms to focus on speaker-specific features in whispered speech. "Supervised and Unsupervised Speaker Diarization of Overlapping Speech in Meetings" by M. J. F. Gales et al. (2010): This study addresses the challenges of speaker diarization in meetings, where overlapping speech is common. While not specific to whispered speech, the methods proposed for handling overlapping speech could be relevant for whispered speech diarization, given the similarities in acoustic challenges. Previous research has shown that, in comparison to neutral-neutral comparisons, mismatched whisper-neutral comparisons severely impair the performance of traditional speaker recognition frameworks. To mitigate this problem, a number of front-end feature adjustments have been suggested. These objectives have already brought up a number of significant research questions, such as the ability of a certain representation to discriminate and the underlying model of human speech when employing that representation. They also oversee studies on sound interpretation and human speech comprehension. The intention of this work was to use the framework to develop a general-purpose diarization system that can function in supervised mode. To increase the system's capability and enhance the diarization outcomes, additional speaker identification modules must be added along with stage-by-stage parameter adjustments. The research's long-term objective is to investigate speaker diarization literature using a vector-based method for whisper identification that can function accurately across datasets even for brief periods of time. Create and implement a vector Probabilistic Linear Discriminant Analysis system with a traditional frontend to do a comparative speaker recognition experiment across two datasets of whispered speech.

5. Research Gap Analysis

Till now, speaker diarization and whisper detection technology development have been thoroughly covered in two comprehensive overview articles that have different focus. Various speaker diarization systems and their subtasks are reviewed through 2018 in the context of broadcast news. Thus, the evolution of speaker diarization technology during the 1990s and early 2000s is discussed historically. On the other hand, speaker diarization for meeting speech and its associated issues was given more attention in whisper detection. One of the primary challenges in developing supervised speaker diarization systems for whispered speech is the scarcity of annotated datasets specifically tailored to this domain. Existing research often relies on conventional feature representations designed for normal speech, which may not effectively capture the unique spectral characteristics of whispered speech. While supervised learning approaches show promise for speaker diarization, there is a lack of comprehensive exploration of different architectures and methodologies specifically tailored to whispered speech. Overlapping speech is common in whispered speech recordings, posing challenges for accurate speaker segmentation and identification. Existing research often focuses on benchmark datasets and controlled experimental settings, with limited evaluation on real-world applications and practical deployment scenarios. Many existing diarization systems may lack scalability and efficiency, particularly when applied to large-scale datasets or real-time applications. Research efforts in whispered speech diarization may sometimes overlook user needs and application requirements, resulting in solutions that may not align with real-world use cases or user preferences. I believe that this survey work is a valuable contribution to the community to consolidate the recent developments with neural methods and thus facilitate further progress toward a more efficient diarization. A number of recent works also explored how to improve clustering-based methods with deep learning based techniques in order to let them deal also with overlapped speech.

6. Results using the Recorded Database

Table 1 Comparative results using MFCC, LFCC, MFCC-LPC

Database	No of Speakers	Train-test mode	Classifier	Features	% Accuracy
Recorded	24	Whisper-Whisper	K-means	MFCC	62.5
				LFCC	65.2
				MFCC+ LPC	45.8

However, prior to the development of timbre features, this investigation also looks into more conventional features like MFCC, LFCC, and MFCC-LPC fusion. In comparison to MFCC, the results were better when the linear filter (i.e., LFCC) was used in place of the Mel-filter. The International Journal of Pure and Applied Mathematics has published the result demonstrating this improvement. The result depicted in Table 1 are the additional support to prove the superior performance of the features compared to traditional MFCC, LFCC, and MFCC-LPC fusion. These results are also accepted by International journal of Computer Applications, Taylor and Francis.

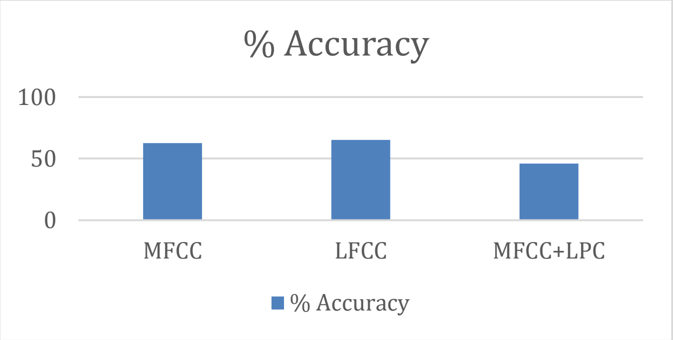

Figure 2 Comparison- MFCC, LFCC, MFCC-LPC Fusion

Figure 2 proves that the identification accuracy is highest with whispered speaker database using LFCC features among all the well-known audio features.

Conclusion

This article highlights a number of issues that need to be resolved in the coming years and gives an overview of the state-of-the-art in speaker diarization systems. For instance, speaker diarization is not developed enough yet so these techniques are easily transferable between domains, as demonstrated in Section V, wherein slight alterations in meeting data (collected at the same locations) result in significant performance disparities. To make results more meaningful and systems more resilient to unknown fluctuations, additional datasets must be assembled in the interim. Naturally, systems will need to become more effective in order to process such data in an acceptable amount of time as dataset sizes increase. Nevertheless, managing overlapping speech—which must be ascribed to several speakers—is probably the largest obstacle of all. There are thus tremendous prospects for major advancements and considerable improvements to the somewhat ad hoc and heuristic approaches that currently dominate the field of speech and speaker recognition, given that the community is still very young, at least in comparison to the more established fields.

References

[1]. Shrawankar, U., &Mahajan, A. (2013). Speech: A Challenge to Digital Signal Processing Technology for Human-Computer Interaction [arxiv:1305.1925]. arXiv.org.

[2]. May, T, Van de Par, S, &Kohlrausch, A. (2011). Noise-robust speaker recognition combining missing data techniques and universal background modelling. IEEE transactions on audio, speech, and language processing, 20(1), 108-121. doi: 10.1109/TASLP.2010.2095124

[3]. Abrol V and MalhotraJ.Data dashboard-integrating data mining with data deduplication. International Journal of Computer Applications.2013;71(22).

[4]. Richardson, F, Reynolds, D &Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. IEEE signal processing

letters, 22(10), 1671-1675. doi: 10.1109/LSP.2015.2420092

[5]. Stafylakis T, Kenny P, Alam MJ and Kockmann M (2015). Speaker and channel factors in text-dependent speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(1), 10.1109/TASLP.2014.2382203

[6]. Sethuram, V., Prasad, A., &Rao, R. R. (2020). Optimal trained artificial neural network for Telugu speaker diarization. Evolutionary Intelligence, 13(4), 631-648. doi: 10.1007/s12065-020-00378-9

[7]. Yu, C., & Hansen, J. H. L. (2017). Active learning based constrained clustering for speaker diarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(11), 2188-2198. doi: 10.1109/TASLP.2017.2747097

[8]. Karim, D., Salah, H., &Adnen, C. (2019). Hybridization DE with K-means for speaker clustering in speaker diarization of broadcasts news. International Journal of Speech Technology, 22(4), 893-909. doi: 10.1109/TASLP.2017.2747097.

[9]. Darekar RV and DhandeAP.Emotion Recognition from Speech Signals Using DCNN with Hybrid GA-GWO Algorithm.Multimedia Research.2019; 2(4): 12-22.

[10]. Srinivas V and SanthiraniCh.Hybrid Particle Swarm Optimization-Deep Neural Network Model for Speaker Recognition. Multimedia Research.2020; 3(1): 1-10.

[11]. Al-Nasheri A, Muhammad G, Alsulaiman M, Ali Z, Malki KH, Mesallam TA and Ibrahim MF. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. Ieee Access. 2017;6:6961-6974.

[12]. Fang, S. H., Tsao, Y., Hsiao, M. J., Chen, J. Y., Lai, Y. H., Lin, F. C., & Wang, C. T. (2019). Detection of pathological voice using cepstrum vectors: A deep learning approach. Journal of Voice, 33(5), 634-641.

[13]. Bebortta, Sujit, Manoranjan Panda, and Shradhanjali Panda. "Classification of pathological disorders in children using random forest algorithm." 2020 international conference on emerging trends in information technology and engineering (ic-ETITE). IEEE, 2020.

[14]. Kręcisz, Krzysztof, and DawidBączkowicz. "Analysis and multiclass classification of pathological knee joints using vibroarthrographic signals." Computer methods and programs in biomedicine 154 (2018): 37-44.

[15]. Anita, J. S., and J. S. Abinaya. "Impact of supervised classifier on speech emotion recognition." Multimedia Res 2.1 (2019): 9-16.

[16]. Rashmi, N., and MrinalSarvagya. "Self-improved grey wolf optimization for estimating carrier frequency offset in SCM-OFDM systems." International Journal of Pervasive Computing and Communications 16.1 (2020): 53-73.

[17]. Rashmi, N., and MrinalSarvagya. "A new optimised interleaver design for high-dimensional data transmission in SCM-OFDM system." International Journal of Wireless and Mobile Computing 18.2 (2020): 153-166.

[18]. Saidi, Pouria, and FarshadAlmasganj. "Voice disorder signal classification using m-band wavelets and support vector machine." Circuits, Systems, and Signal Processing 34 (2015): 2727-2738.

[19]. Remmiya, R., and C. Abisha. "Artifacts removal in EEG signal using a NARX model based CS learning algorithm." Multimedia Research 1.1 (2018): 1-8.

[20]. Gregory Sell and Daniel Garcia-Romero, Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration, Spoken Language Technology Workshop (SLT), 2014 IEEE, 7-10 Dec. 2014.

[21]. S. E. Tranter, D. A. Reynolds, An overview of automatic speaker diarization systems, IEEE Transactions on Audio, Speech, and Language Processing 14 (2006) 1557–1565.

[22]. H. W. Kuhn, The hungarian method for the assignment problem, Naval research logistics quarterly 2 (1955) 83–97.

[23]. J. G. Fiscus, J. Ajot, M. Michel, J. S. Garofolo, The rich transcription 2006 spring meeting recognition evaluation, in: Proceedings of International Workshop on Machine Learning and Multimodal Interaction, NIST, 2006, pp. 309–322

[24]. Tanveer, M.I.; Casabuena, D.; Karlgren, J.; Jones, R. Unsupervised Speaker Diarization that is Agnostic to Language, Overlap-Aware, and Tuning Free. arXiv 2022, arXiv:2207.12504. [Google Scholar]

[25]. Dawalatabad, N.; Madikeri, S.; Sekhar, C.C.; Murthy, H.A. Novel Architectures for Unsupervised Information Bottleneck Based Speaker Diarization of Meetings. Ieee/Acm Trans. Audio Speech Lang. Process. 2021, 29, 14–27. [Google Scholar] [CrossRef]

[26]. Bredin, H.; Yin, R.; Coria, J.M.; Gelly, G.; Korshunov, P.; Lavechin, M.; Fustes, D.; Titeux, H.; Bouaziz, W.; Gill, M.P. Pyannote audio: Neural Building Blocks for Speaker Diarization. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7124–7128. [Google Scholar] [CrossRef]

[27]. Bai, Z.; Zhang, X.-L. Speaker recognition based on deep learning: An overview. Neural Netw. 2021, 140, 65–99, ISSN 0893-6080. [Google Scholar] [CrossRef] [PubMed]

[28]. Dimitriadis, D.; Fousek, P. Developing On-Line Speaker Diarization System. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 2739–2743. [Google Scholar] [CrossRef]

[29]. D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in INTERSPEECH, Florence, Italy, 2014, pp. 249–252.

[30]. M. Mandasari, R. Saeidi, M. McLaren, and D. van Leeuwen, "Quality measure functions for calibration of speaker recognition system in various duration conditions," IEEE Trans, Audio, Speech and Language Proc., vol. 21, no. 11, pp. 2425–2438, 2013.