ACADEMY JOURNAL
OF
SCIENCE AND ENGINEERING
Available online @
www.academyjsekad.edu.ng
AJSE 19 (1) 2025
Original Research

CLASSIFICATION OF AUDIO SOURCE CHANNELS USING TIME-FREQUENCY ANALYSIS AND DECISION TREE CLASSIFIER

Sagir Lawan[1], Ashraf A. Ahmad[1]*, Amina Jibril[1], Idrisu Mohammed[2], Kulu Ahmad Amalo[3]

[1]Dept. of Electrical/Electronic Engineering, Faculty of Eng. and Tech., Nigerian Defence Academy, Kaduna, Nigeria
[2]Dept. of Electrical/Electronic Engineering, School of Engineering Technology, Federal Polytechnic, Bali, Nigeria
[3]Dept. of Electrical Technology Education, College of Technical Education, Federal Polytechnic, Kaduna, Nigeria

Abstract

This paper presents the development and classification of audio source separation algorithms utilizing tim frequency analysis techniques, specifically Wigner-Ville distributions (WVDs), compact kernel distributi (CKD) and multi-directional distribution (MDD). These methods are integrated with a decision tr classifier to enhance the separation of audio sources in noisy environments. The algorithms were tested the AI-generated audio signals across various signal-to-noise ratio (SNR) levels. The CKD meth demonstrated exceptional performance, achieving a classification accuracy of 100% at 0dB to 15dB SN for the multichannel AI-generated audio signals. This study highlights the effectiveness of the advanc time-frequency analysis techniques CKD and MDD in improving audio source separation and their potenti for real-time audio processing applications. The results indicate that these techniques can significant enhance the clarity and quality of separated audio signals, providing an effective solution for tasks such music source separation, speech enhancement, and environmental sound separation.

Keywords- Audio source separation, Time-frequency analysis, Wigner-Ville distribution (WVD), compact kernel distribution (CKD), multi-directional distribution (MDD), AI- generated audio signals, signal-to-noise ratio (SNR).

Corresponding author's e-mail: aaashraf@nda.edu.ng

1.0 INTRODUCTION

Audio source separation is a technology designed to isolate one or more specific source signals from an audio recording containing multiple sound sources (Li et. al., 2022). This technology is especially valuable in situations where audio quality is compromised by background noise or overlapping speakers. Its applications span various industries, enhancing audio processing, improving communication, and providing a richer audio experience in noisy environments. Notable applications include speech communication, speech enhancement, hearing aids, automatic speech recognition (ASR), music separation in recording and production, broadcasting and entertainment, surveillance systems, assistive listening devices, and virtual and augmented reality (Michelsanti et. al. 2021; Martinek et. al. 2021; Richard et. al. 2023; Zhu et. al. 2024).

However, traditional audio source separation methods face significant challenges in achieving accurate and efficient signal isolation. Conventional techniques, such as Independent Component Analysis (ICA) and Non-Negative Matrix Factorization (NMF), often rely on strong assumptions about the statistical independence of sources or the sparsity of signals in the time-frequency domain. These assumptions may not hold in real-world scenarios with overlapping speakers, complex noise profiles, or reverberant environments, leading to degraded performance (Jorgenson, 2022). Additionally, traditional methods struggle to handle signals with highly dynamic spectral structures or to adapt to variations in noise characteristics, limiting their effectiveness in practical applications

In many real-world scenarios, audio recordings are frequently contaminated by background noise, concurrent speakers, or reverberation, severely impacting the quality and intelligibility of the desired audio signals. This presents a significant challenge for applications such as speech communication, automatic speech recognition (ASR), speech enhancement, music production, and assistive listening devices. Therefore, it is crucial to develop audio source separation techniques capable of effectively extracting the desired audio signals from noisy environments. The primary motivation for developing audio source separation algorithms is to enhance audio processing in noisy environments, thereby improving the quality, intelligibility, and user experience across various audio applications. Background noise and interfering speakers can impede effective communication, reduce speech recognition accuracy, and detract from speech enhancement efforts. By isolating desired audio signals from unwanted noise and interference, overall audio quality can be significantly improved, leading to better speech intelligibility, more accurate speech recognition, and a more immersive audio experience.The aim of this research was achieved through the following objectives: first, to configure multichannel noisy audio signals using AI-generated audio, and second, to develop source separation algorithms based on time-frequency analysis and decision tree classifiers.

2 REVIEW OF RELATED WORK

In the realm of audio signal processing, several recent research endeavours have illuminated various aspects of speech enhancement, source separation, and classification, each contributing significantly to the field. This section reviews some of the most recent works related to the research.

A modified minimum mean square error (MMSE) method for enhancing speech signals in noisy mixtures was proposed (Tengtrairat et al., 2016). Their method outperformed standard MMSE and other existing methods in terms of signal quality and intelligibility. Specifically, the proposed enhancement method achieved an average Perceptual Evaluation of Speech Quality (PESQ) improvement of 27% and 19% over standard and modified MMSE methods, respectively, across a range of input Signal-to-Noise Ratios (SNRs). Additionally, the method demonstrated an average improvement of 3.0 dB (76%) for segmental SNR (SegSNR) and 0.4 (12%) for PESQ compared to noisy mixtures. The study hinted at future research prospects related to multi-channel source separation.

A blind source separation method suitable for real-time speech and noise separation on smartphones contributed to the field by developing a computationally efficient method (Bhat et al., 2019). Their approach integrated a neural network-based sound source localization method with Independent Vector Analysis (IVA) to enhance efficiency and accuracy. Remarkably, the proposed method achieved a 10-fold cross-validation accuracy of 86.2% with a standard deviation of 0.93%. The accuracy of Generalized Cross Correlation (GCC) is around 44.1%. Additionally, the Feed Forward Neural Network (FNN)-based Direction of Arrival (DOA) estimation (FNNDOA) method was evaluated across various Signal-to-Noise Ratios (SNRs) and noise types. The study highlighted the need for further research aimed at improving computational complexity and traditional IVA under realistic scenarios, such as situations where speakers change position while speaking, such as during a presentation on a podium.

In 2019, a cooperative system leveraging distributed microphone arrays and wearable devices to enhance audio source separation and improve listening performance was introduced (Corey et al., 2019). Their work involved deploying 160 microphones in a reverberant room, where the system exhibited significant improvements in source separation performance. Notably, there was an approximate 5 dB performance difference between the filters designed from unprocessed reference microphone signals and those designed from Independent Vector Analysis (IVA) estimates. Furthermore, the study revealed a direct correlation between separation Signal-to-Noise Ratio (SNR) and enhancement SNR, with every 1 dB improvement in the separation SNR providing about a 1 dB improvement in enhancement SNR. To advance this research, further investigations are warranted, especially concerning separation methods capable of effectively handling large numbers of sources in highly reverberant environments by harnessing spatial diversity.

Machine learning algorithms for real-time blind audio source separation with natural language detection were also evaluated (Alghamdi et al., 2021). Conv-TasNet and Demucs algorithms were assessed for the quality and execution time of separation output signals, as well as the effectiveness of natural language detection. Both algorithms exhibited high accuracy and excellent results in the separation process. Conv-TasNet achieved the highest Signal-to-Distortion Ratio (SDR) score of 9.21 for music at the (music & female) experiment, and the highest SDR value for the child signal is 8.14. The SDR score for music at the (music & female) experiment is 7.8 during the Demucs algorithm, where the child output signal has the highest SDR score of 8.15 for the same experiment. Future research directions include diversifying training datasets, exploring alternative separation algorithms using deep learning approaches, and expanding the range of data categories.

Another paper titled "Informed Audio Source Separation with Deep Learning in Limited Data Settings" contributed to the field in multiple ways (Schulze-Forster, 2021). The study focused on three primary aspects: Supervised Setting with Limited Data, Text-Informed Singing Voice Separation, and an Unsupervised Deep Learning Approach. Baseline models achieved median Signal-to-Distortion Ratio (SDR) scores of 3.0 dB and 3.33 dB for BL1 and BL2, respectively, representing appropriate baselines given the simplicity and limited training data. The Percentage of Correctly Aligned Segments (PCAS) exceeded 80% for SNRs of 0 dB and above, making the proposed approach suitable for phoneme alignments in various datasets. Future research avenues include exploring training objectives involving the reconstruction of observed text and audio sequences, potentially combining attention mechanisms and the Connectionist Temporal Classification (CTC) loss.

In a parallel study, AVLIT, an innovative Audio-Visual Lightweight Iterative model designed for audio-visual speech separation in noisy environments, was introduced (Martel et al., 2023). AVLIT employs a combination of audio and video branches, both utilizing Asynchronous Fully Recurrent Convolutional Neural Network (A-FRCNN) blocks. The iterative design of AVLIT allows for efficient and lightweight processing while maintaining high separation quality. AVLIT-8 and AVLIT-4 demonstrated superior performance compared to Visual Voice and dual-path recurrent neural network (DPRNN), achieving roughly 1 dB and 0.5 dB substantial improvements in Scale-Invariant Signal-to-Distortion Ratio (SI-SDRi). These results highlight the potential of AVLIT as a practical solution for enhancing speech separation in challenging acoustic environments. Further investigation into AVLIT's behavior in reverberant environments is needed to fully assess its real-world applicability.

Recent advancements in audio signal processing have demonstrated significant improvements in speech enhancement, source separation, and classification techniques. For instance, a modified minimum mean square error (MMSE) method has shown superior performance in enhancing speech signals within noisy mixtures, achieving substantial improvements in signal quality and intelligibility compared to standard methods. Additionally, a blind source separation method designed for real-time applications on smartphones has integrated neural network-based sound source localization with Independent Vector Analysis (IVA), enhancing both efficiency and accuracy. Cooperative systems utilizing distributed microphone arrays have further improved source separation performance, particularly in reverberant environments. Machine learning algorithms such as Conv-TasNet and Demucs have also exhibited high accuracy in real-time blind audio source separation, demonstrating notable performance in various noise scenarios. Moreover, the study on "Informed Audio Source Separation with Deep Learning in Limited Data Settings" highlighted the efficacy of deep learning approaches even with limited training data, indicating potential for phoneme alignments and other applications. The innovative Audio-Visual Lightweight Iterative model (AVLIT) has shown superior performance in audio-visual speech separation under noisy conditions, underscoring its practicality for challenging acoustic environments.

Despite these advancements, there is a discernible gap in the literature regarding the classification of audio source channels using time-frequency analysis coupled with a Decision Tree Classifier (DTC). While existing research has predominantly focused on enhancing and separating audio sources, the specific challenge of accurately identifying the number of channels in an audio source remains underexplored. This paper aims to address this gap by leveraging time-frequency techniques alongside a Decision Tree Classifier to improve the classification of multi-channels audio source, thereby advancing the field of audio signal processing.

3 METHODOLOGY

In this section, the methodology for the classification of audio source channels using time-frequency analysis and a Decision Tree Classifier is meticulously detailed. The section outlines the systematic approach undertaken to achieve the research objectives, which include configuring multichannel noisy audio signals using AI-generated audio, developing source separation algorithms based on advanced time-frequency analysis techniques, and implementing a Decision Tree Classifier for accurate channel identification.

3.1 Audio Source Separation

Several methods have been proposed for audio source separation, including single-channel and multichannel separation methods, as well as methods for separating moving sound sources. Single-channel separation methods are designed to separate sources from a single audio channel, while Multichannel separation methods use multiple channels to separate sources (Gidlöf & Nyberg, 2023). Multichannel methods can be further classified into time-domain and frequency-domain methods. Time-domain methods use spatial information to separate sources, while frequency-domain methods use spectral information(Li et al., 2023).

The audio signals, for the research, employed a detail and comprehensive data collection strategy for speech enhancement for High-quality audio recordings was captured, meticulously categorized by speaking scenarios in noisy environments to construct the foundational dataset. To diversify this dataset,They AI generated audio signal base on the first objective, were generated using Python software. The single channel audios were then mixed together to form the multichannel audio by concatenating the individual audio signal using the MATLAB software.

Additive White Gaussian Noise (AWGN) is a type of noise commonly introduced into signals within communication systems. It is termed "additive" because it is added to the original signal, "white" due to its consistent power spectral density across all frequencies, and "Gaussian" because it follows a Gaussian (normal) probability distribution. In the context of audio signals, assessing performance necessitates accounting for such noise. During the signal pre-processing stage, a standard AWGN model is employed to generate and inject noise into all audio signals.The equation for the output signal (y) is expressed as the sum of the input signal (x) and the noise (n):

$$y = x + n \quad \text{(Cohen, 1995)} \qquad (1)$$

where (y) represents the output signal, (x) denotes the input signal, and (n) stands for the AWGN. This approach ensures that the developed audio source separation algorithm can effectively handle and mitigate the impact of noise, thereby enhancing audio processing in noisy environments.

3.2 Time-Frequency Analysis/Distribution

Time-frequency analysis/distribution (TFD) is a technique used in signal processing to analyse and represent signals in both the time and frequency domains. It provides a view of a signal represented over both time and frequency, allowing for the analysis of signals containing multiple time-varying frequencies. (Ahmad et al., 2024) Time-frequency analysis/distribution techniques find applications in various areas, including multichannel audio source separation, music source separation, and speech separation. The TFD used in this research are explain as follows:

3.2.1 Wigner-Ville Distribution of a Signal (WVD)

The Wigner-Ville Distribution (WVD) is a powerful method for estimating the power spectral function of a nonstationary signal through a time-frequency energy distribution approach. Initially introduced by Wigner and later adapted by Ville for signal processing applications (Boashash, 2016), the WVD is denoted as $P_{z,}WVD(t, f)$and is defined as:

$$P_{z,}\mathrm{WVD}(t,f) = \int_{-\infty}^{\infty} z\left(t + \frac{\tau}{2}\right) z^{*}\left(t - \frac{\tau}{2}\right) e^{-j2\pi f\tau} d\tau \qquad (2)$$

where $P_{z,}WVD(t, f)$,represents the Wigner-Ville distribution of a signal at time (t) and frequency (f), $z\left(t + \frac{\tau}{2}\right)$is a complex-valued function typically indicating the analysing signal or window function, and∗is the complex conjugate. While the theoretical evaluation from minus infinity to plus infinity is impractical, a pseudo-Wigner-Ville distribution (PWVD) mitigates this by using a running window (Boashash, 2016):

$$P_{z,}PWVD(t,f) = \int_{-\infty}^{\infty} h(\tau)z\left(t + \frac{\tau}{2}\right)z^{*}(t - \frac{\tau}{2})e^{-j2\pi f\tau}d\tau \qquad (3)$$

where $h(\tau)$is the window function. The WVD has unique kernel functions from the bilinear generalized class of time-frequency distribution, exhibiting excellent time-frequency aggregation, particularly in low Signal-to-Noise Ratio (SNR) conditions (Boashash, 2016).

The WVD process involves obtaining the instantaneous autocorrelation function (IAF) and converting it to a time-frequency distribution using the Fourier transform(Boashash, 2016):

$$K_z\,(t,\tau) \;=\; z\,(t + \frac{\tau}{2})\,z^{*}\,(t - \frac{\tau}{2}) \qquad (4)$$

$$P_{z,}WVD(t,f) = \int_{-\infty}^{\infty} K_z\,(t,\tau)\,e^{-j2\pi f\tau}d\tau \quad (5)$$

However, the WVD suffers from limitations such as inner artifacts and cross terms. To address these, the windowed WVD (WWVD) is used, incorporating a window function after obtaining the IAF. The Hamming window is selected for its ability to provide better frequency resolution and side lobe suppression(Boashash, 2016):

$$K_{z,w}(t,\tau) \;=\; g_w(\tau)\,K_z\,(t,\tau) \qquad (6)$$

$$P_{z,WWVD}(t,f) \;=\; \int_{-\infty}^{\infty} g_w(\tau)\,z\,(t + \frac{\tau}{2})\,z^{*}\,(t - \frac{\tau}{2})\,e^{-j2\pi f\tau}d\tau \qquad (7)$$

$$P_{z,WWVD}(t,f) \;=\; \int_{-\infty}^{\infty} 0.54 - 0.46\cos(\frac{2\pi\tau}{T})\,z\,(t + \frac{\tau}{2})\,z^{*}\,(t - \frac{\tau}{2})\,e^{-j2\pi f\tau}d\tau \qquad (8)$$

3.2.2 Compact Kernel Distribution (CKD) and Multidirectional Kernel Distribution (MDD)

Compact Kernel Distribution (CKD): The Compact Kernel Distribution (CKD) method is an advanced version of the pseudo-Wigner-Ville Distribution (WVD), designed to provide a compact support kernel window function that effectively vanishes outside a specified range in the ambiguity domain(Boashash, 2016). Unlike Gaussian windows with infinite lengths, CKD does not require truncation using rectangular windows, preventing the loss of valuable information. CKD is known for its superior performance in suppressing cross-terms while maintaining auto-term resolution, achieved by combining compact support with flexible adjustments to the kernel's shape and size independently (Boashash, 2016):

$$g(v,\tau) = G_1(v)g_2(\tau) = \begin{cases} e^{2c\frac{cD^2}{eV^2-D^2}+\frac{cE^2}{\tau^2-E^2}} & |V|<D, |\tau| < E, \\ 0 & \text{otherwise} \end{cases} \qquad (9)$$

where ν and τ are the Doppler and lag windows determined by parameters D and E, and Ccontrols the shape. The kernel width in the ambiguity domain is determined by prior knowledge of the signal components. The Instantaneous Autocorrelation Function (IAF) of WVDs, represented as $K_z\,(t,\tau)$, is central to CKD, and its Time-Frequency Distribution (TFD) is given by (Boashash, 2016):

$$K_z\,(t,\tau) \;=\; z\,(t + \frac{\tau}{2})\,z^{*}\,(t - \frac{\tau}{2}) \qquad (10)$$

$$P_{z,CKD}(t,f) = \int_{-\infty}^{\infty} g(t,\tau) * K_z\,(t,\tau)\,e^{-j2\pi f\tau}d\tau \qquad (11)$$

Where $g(t,\tau)$is obtained through:

$$g(t,\tau) = \int_{-\infty}^{\infty} g(v,\tau)e^{-j2\pi v\tau}dv \qquad (12)$$

From this, the CKD TFD is:

$$P_{z,CKD}(t,f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(v,\tau)\,z\left(u + \frac{\tau}{2}\right)z^{*}\left(u - \frac{\tau}{2}\right)e^{-j2\pi f\tau}.\,e^{-j2\pi v\tau}dudvd\tau \qquad (13)$$

The CKD performance can degrade for signals with auto-terms oriented away from the time or frequency axis in the (t,f) domain. To address this, the Multidirectional Kernel Distribution (MDD) can be used.

Multidirectional Kernel Distribution (MDD): The Multidirectional Kernel Distribution (MDD) is particularly suited for signals with energy concentrated along multiple directions in the (t,f) domain, such as multicomponent linear frequency modulated (LFM) signals with different nonzero chirp rates(Boashash, 2016). To achieve high-resolution TFDs for such signals, a rotation parameter is included in the formulation of smoothing kernels (Boashash, 2016):

$$g_\theta(v,\tau) = \begin{cases} e^{\frac{c}{\left(\frac{v\cos(\theta)-\tau\sin(\theta)}{D}\right)-1}} e^{\frac{c}{\left(\frac{\sin(\theta)v+\cos(\theta)\tau}{E}\right)}} & \text{for} |\cos(\theta)v-\sin(\theta)\tau|<D \\ & \text{for} |\sin(\theta)v+\cos(\theta)\tau|<E \\ 0 & \text{otherwise,} \end{cases} \qquad (14)$$

where θ is the angle of the kernel with the Doppler axis in the ambiguity domain, D is the half-support of $g_\theta(v,\tau)$, and E is the half-length along its principal direction. For signals with multiple directions of energy concentration, the smoothing is performed along multiple directions, resulting in a summation of a predetermined numberN_D of directional kernels (Boashash, 2016):

$$g(v,\tau) = \frac{e^c}{N_D} \sum_{i=1}^{N_D} g\theta_i(v,\tau), \qquad (15)$$

The MDD TFD can be formulated using the IAF:

$$\mathrm{K_z}\ (\mathrm{t},\tau) = \mathrm{z}\,(\mathrm{t} + \frac{\tau}{2})\, \mathrm{z}^*\,(\mathrm{t} - \frac{\tau}{2}) \qquad (16)$$

$$\mathrm{P_{z,MDD}}(\mathrm{t},\mathrm{f}) = \int_{-\infty}^{\infty} \mathrm{g}(\mathrm{t},\tau) * \mathrm{K_z}\ (\mathrm{t},\tau)\ \mathrm{e}^{-\mathrm{j}2\pi \mathrm{f}\tau} \mathrm{d}\tau \qquad (17)$$

where $\mathrm{g}(\mathrm{t},\tau)$ is derived as:

$$\mathrm{g}(\mathrm{t},\tau) = \int_{-\infty}^{\infty} g_\theta(v,\tau) \mathrm{e}^{-\mathrm{j}2\pi v\tau} \mathrm{d}v \qquad (18)$$

From this, the MDD TFD is:

$$\mathrm{P_{z,MDD}}(\mathrm{t},\mathrm{f}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_\theta(v,\tau)\, \mathrm{z}\left(\mathrm{u} + \frac{\tau}{2}\right) \mathrm{z}^*\left(\mathrm{u} - \frac{\tau}{2}\right) \mathrm{e}^{-\mathrm{j}2\pi \mathrm{f}\tau}.\ \mathrm{e}^{-\mathrm{j}2\pi v\tau}\ \mathrm{dudvd}\tau \qquad (19)$$

The parameters for CKD and MDD, such as the kernel shape C, Doppler cut-off D, and lag cut-off E, are chosen based on prior knowledge of the signal components. Table 1 and Table 2 show the ranges for these parameters. The CKD and MDD approaches are implemented in MATLAB functions like `tf_kernel_ckd` and `tf_kernel_mdd`, which generate the respective kernels based on the specified parameters. The IAF functions `IAF_CKD` and `IAF_MDD` smooth the IAF in the lag domain using these kernels, followed by applying CKD and MDD to obtain the TFD

Table 1: C, D, E Range of Values for CKD Kernel (Al-Sa'd et al., 2021)

S/N	PARAMETERS	RANGE OF VALUE
1	C	[0, 3]
2	D	[0, 1]
3	E	[0, 1]

Table 2: C, D, E Range of Values for MDD Kernel (Al-Sa'd et al., 2021)

S/N	PARAMETERS	RANGE OF VALUE
1	C	[0, 3]
2	D	[0, 1]
3	E	[0, 1]

By incorporating these methods, the CKD and MDD approaches enhance time-frequency analysis, particularly for signals with complex energy distributions in the (t,f) domain.

3.2.3 TFR Feature Extraction

Extracting features from Time-Frequency Representations (TFR) allows for the characterization of different aspects of the signal, aiding in tasks like audio source separation and classification (Sharma et al., 2020). TFR extraction features like Spectral Centroid and Spectral Bandwidth, selected for this research, are commonly used to describe the distribution of energy in the frequency domain over time.

Spectral Centroid: The Spectral Centroid is a feature that represents the centre of mass of the spectrum of a signal, providing information about where the "centre" of the signal's frequency content lies (Sharma et al., 2020). It is calculated as the weighted mean of the frequencies present in the signal, with higher values indicating a higher concentration of energy towards higher frequencies, and vice versa. Spectral Centroid is a useful feature for audio analysis as it can help differentiate between sounds with different spectral characteristics, aiding in tasks like instrument recognition and audio source separation. In the context of developing an audio source separation algorithm, the Spectral Centroid helps in distinguishing between different sources in a mixture based on their frequency content over time. Mathematically, it is represented as (Sharma et al., 2020):

$$Centroid(t) = \frac{\sum_{\mathrm{f}} \mathrm{f.}|\mathrm{X(t,f)}|}{\sum_{\mathrm{f}} \mathrm{f.}|\mathrm{X(t,f)}|} \qquad (20)$$

Spectral Bandwidth: Spectral Bandwidth is a feature that describes the width of the frequency range occupied by a signal and provides information about the spread of energy across the frequency spectrum (Sharma et al., 2020). It is calculated as the standard deviation of the frequencies around the Spectral Centroid, reflecting how dispersed the frequencies are around the centre of mass. Spectral Bandwidth is useful for characterizing the timbral qualities of audio signals, as signals with broader bandwidths tend to sound brighter or noisier compared to signals with narrower bandwidths. In the context of audio source separation, Spectral Bandwidth helps in distinguishing between sources with different spectral shapes and can aid in separating sources with overlapping frequency content. Mathematically, it is represented as (Sharma et al., 2020):

$$\mathrm{Bandwidth(t)} = \sqrt{\frac{\sum_{\mathrm{f}}\left(\mathrm{f}-\mathrm{Centroid(t)}\right)^2 |\mathrm{X(t,f)}|}{\sum_{\mathrm{f}} \mathrm{f.}|\mathrm{X(t,f)}|}} \qquad (21)$$

By using Spectral Centroid and Spectral Bandwidth features, this research enhances the ability to analyse and separate audio sources based on their distinct frequency characteristics over time.

3.3 Classifiers in Audio Source Separation

Classifiers are machine learning algorithms used for various applications (Ahmed et al, 2024), including playing a crucial role in developing real-time audio source separation algorithms for enhanced audio processing in noisy environments. They can be trained on labelled or unlabelled data and can be supervised, unsupervised, or semi-supervised. Various types of classifiers, including linear regression, decision trees, random forests, support vector machines, clustering algorithms, Principal Component Analysis (PCA), Independent Component Analysis (ICA), self-training, co-training, and multi-view learning, are utilized in audio processing to achieve tasks like music source separation, speech enhancement, and environmental sound separation (Li et al., 2023).

Decision Trees in Audio Source Separation: Decision trees, a type of supervised learning algorithm, are used for both classification and regression problems. They function by recursively splitting the data into smaller subsets based on the most significant features. This process, known as bagging, involves using decision trees as parallel estimators. In classification problems, the final result is determined by the majority vote from each decision tree, while in regression problems, the prediction is the mean value of the target values in the leaf node (Josephine et. al,, 2021; Costa & Pedreira, 2023).

In this research, a binary decision tree was employed, where each internal node has exactly two outgoing edges, representing Yes/No questions. The training set is split into two disjoint subsets, D = D_{Yes} + D_{No}. The subset D_{Yes} is associated with the left branch of the split and D_{No} to the right branch. This splitting criterion is applied recursively on each branch using only the samples that reach that node until a stopping criterion is met.

Decision trees are integral to the methodology, offering an effective means of classification by incorporating prior knowledge and handling features of various scales. Their adaptability to different data types and optimization of decision-making by minimizing impurity through appropriate questioning strategies make them highly effective in this context. The system flow chart illustrating this process is shown in Figure 1.

From Figure 1, the System Flow Chart above, the program starts by adding a noise signal to theAI-generated audio. Next, the Time-Frequency Distribution (TFD) is performed. Further processing involves feature extraction using the spectral centroid and spectral bandwidth to estimate their values. These values are then used to set the upper and lower limits of the classifier (Decision Tree Classifier). If the sum of the squared spectral centroid and the spectral bandwidth is greater than the lower limit, the signal is classified as multi-channel. If not, the program checks if this sum is less than the upper limit. If it is, the signal is classified as double-channel; otherwise, it is classified as single-channel, and the process ends.

By leveraging these classifiers, particularly decision trees, the research aims to enhance audio source separation capabilities, contributing to more efficient audio processing in diverse and noisy environments.

Table 3: Simulation Set up Values for Lower and Upper Limits

SIGNAL	TFD	LOWER LIMIT (LL)	UPPER LIMIT (UL)
AI Generated	WVD	0.9460	0.9550
	WWVD	0.9080	0.9150
	CKD	0.9250	0.9280
	MDD	0.8500	0.9100

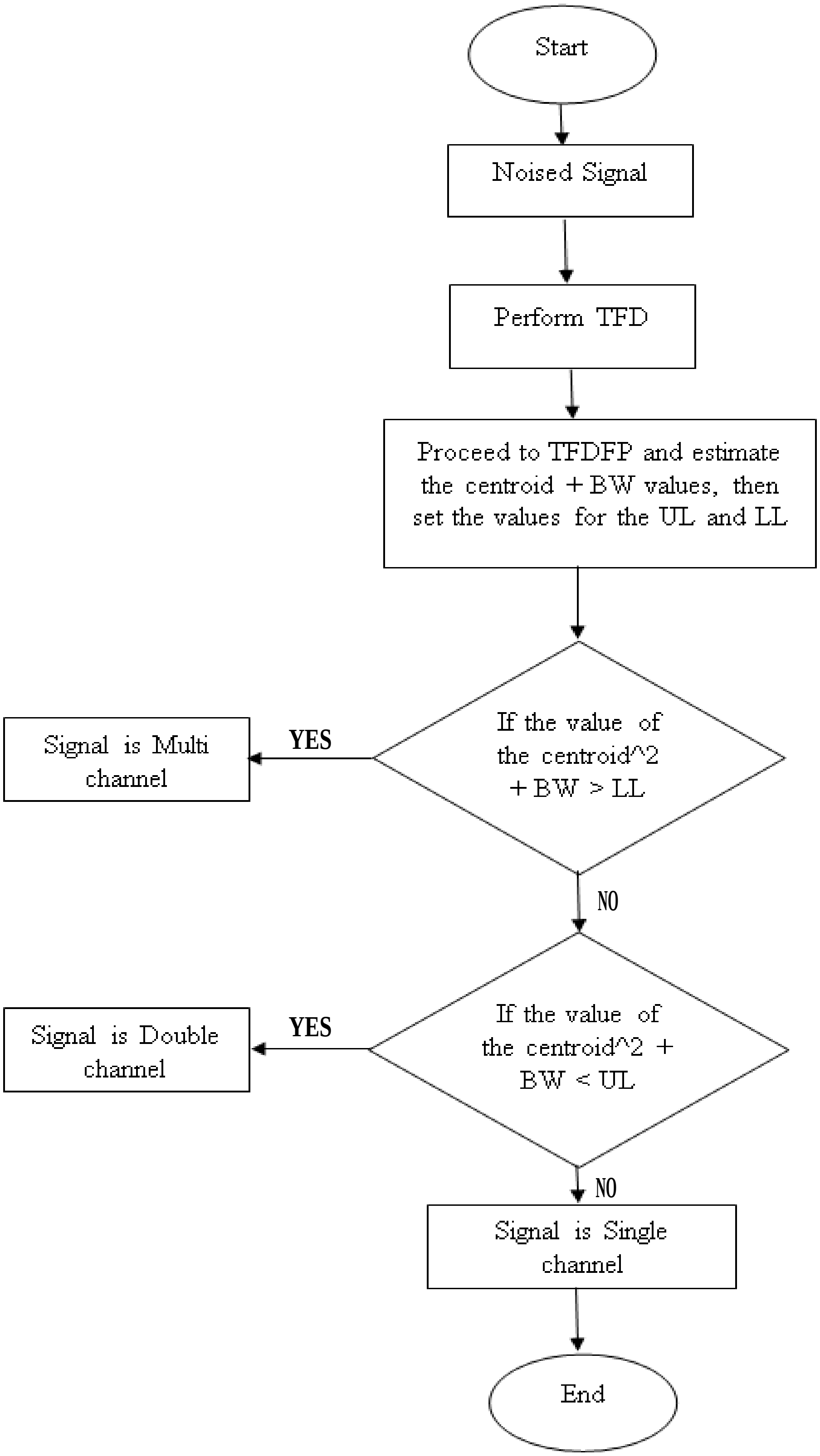
Start
Noised Signal
Perform TFD
Proceed to TFDFP and estimate the centroid + BW values, then set the values for the UL and LL
If the value of the centroid^2 + BW > LL
YES
Signal is Multi channel
NO
If the value of the centroid^2 + BW < UL
YES
Signal is Double channel
NO
Signal is Single channel
End

Figure 1: Audio Source Separation System Flow Chart

3.4 Performance Analysis/ Simulation Set-Up

The algorithms developed to test the accuracy of channel separation using various time-frequency domain processing methods for the AI-generated audio signals setups incorporate all the previously discussed steps, such as loading audio segments, adding noise, applying a Hilbert transform, and conducting processing and classification tasks using different signal processing methods including WVD, WWVD, CKD, and MDD.

Finally, each TFD (WVD, WWVD, CKD, and MDD) is analysed and visualized, assessing the probabilities of detecting 1, 2, or 3 channels in the audio signal across various Signal-to-Noise Ratio (SNR) levels. Table 3 below shows the Simulation Set-up Values for Lower and Upper Limits obtained.

Table 3 above shows the lower and upper limit values for multichannel audio sources used for both AI-generated and audio recording signals. These values are based on the TFR Feature Extraction detailed discussed earlier, which utilizes spectral centroid and spectral bandwidth values. The upper and lower limit values were set accordingly before classification was conducted.

4 RESULTS AND DISCUSSION

In this section, we present the results and discuss the performance of Wigner-Ville Distribution (WVD), Windowed Wigner-Ville Distribution (WWVD), Compact Kernel Distribution (CKD), and Multidirectional Kernel Distribution (MDD) in the context of multichannel audio source separation. Analysing their effectiveness under varying Signal-to-Noise Ratio (SNR) conditions, and evaluating their ability to detect and classify 1, 2, or 3 channels for the AI-generated audio signals. Results are visualized through plots, focusing on the impact of noise and the influence of spectral centroid and spectral bandwidth on classification accuracy. The discussion highlights the strengths, limitations, and practical implications of each method in noisy environments.

4.1 AI Generated Audio Signals

Figure 2 below shows the time representation plot for AI generated audio signals.

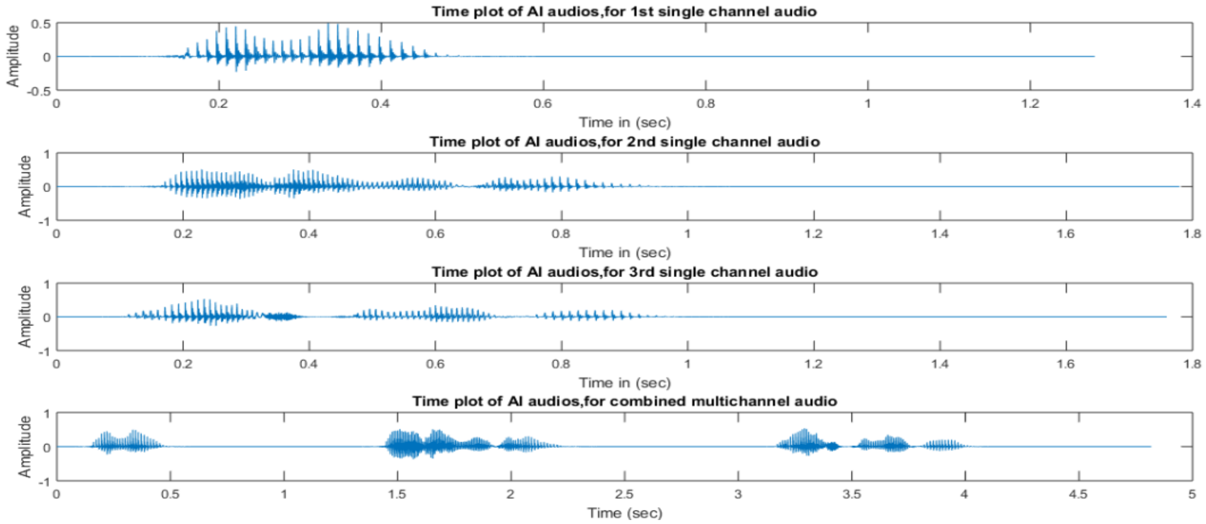

Figure 2: The time plot of AI Generated Audio signal

Figure 2 shows the time plot of the AI Generated audio signal for the individual single channel audio and the combined multichannel audio. Examining the combine audio plot of Figure 2 shows that all single audio channels have been appropriately captured.

4.2 Time-Frequency Representations (TFR) Of The TFDs

Figure 3(a), which is the 3D plot of multichannel AI generated audio signal illustrates a three-dimensional (3D) representation, specifically a waterfall plot, depicting the correlation between power, time, and frequency in the typical audio signal utilizing WVD. The signal runs for duration of 4.5secs, a sampling frequency of 22KHz and SNR of 10dB as depicted in Figure 4.5. The spike is power indicated presence of high speech sound which aids the identification, classification and performance indication measurement of number of channels present in the signal.

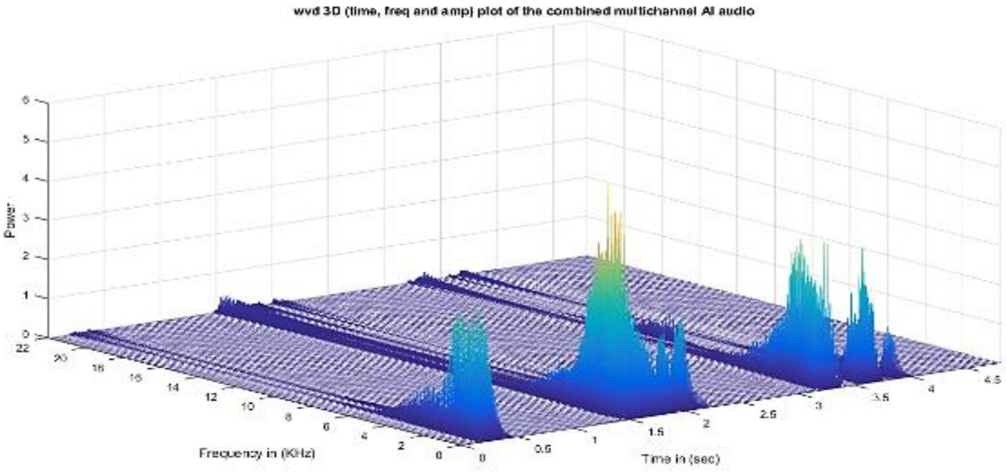

Figure 3 (a) 3D plot of multichannel AI generated audio signal

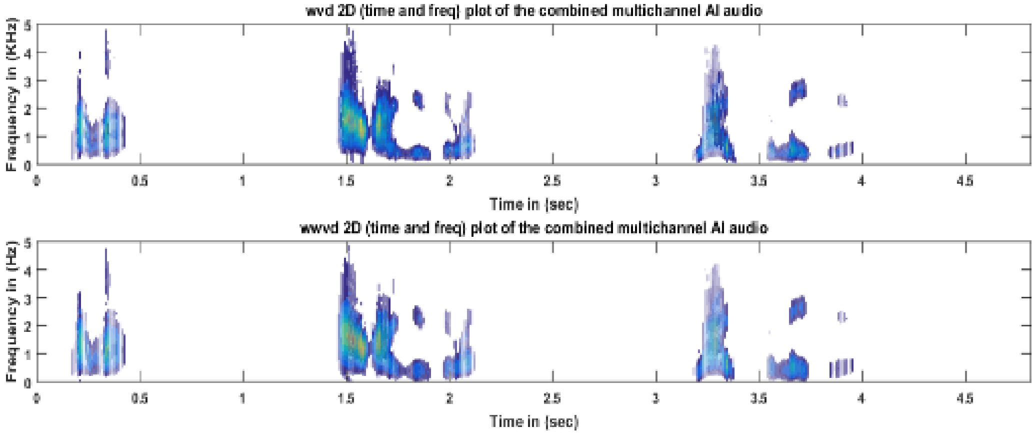

Figure 3 (b): 2D contour plot of multichannel AI generated audio signal

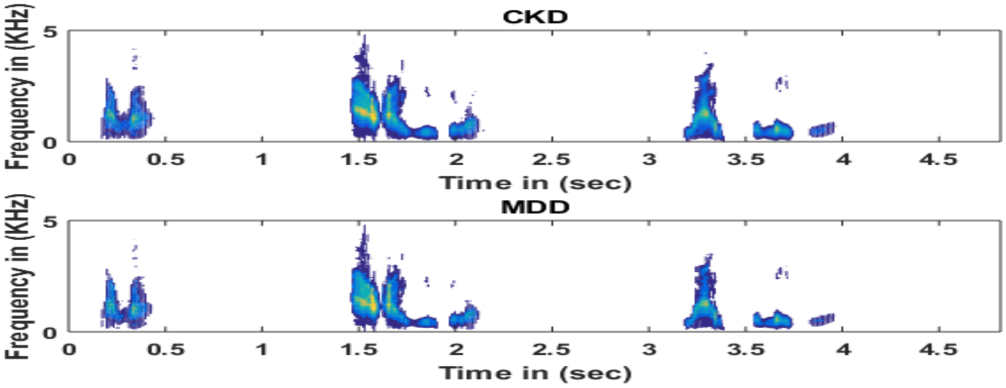

Figure 3 (c): 2D contour plot of CKD, and MDD of multichannel AI generated audio signal

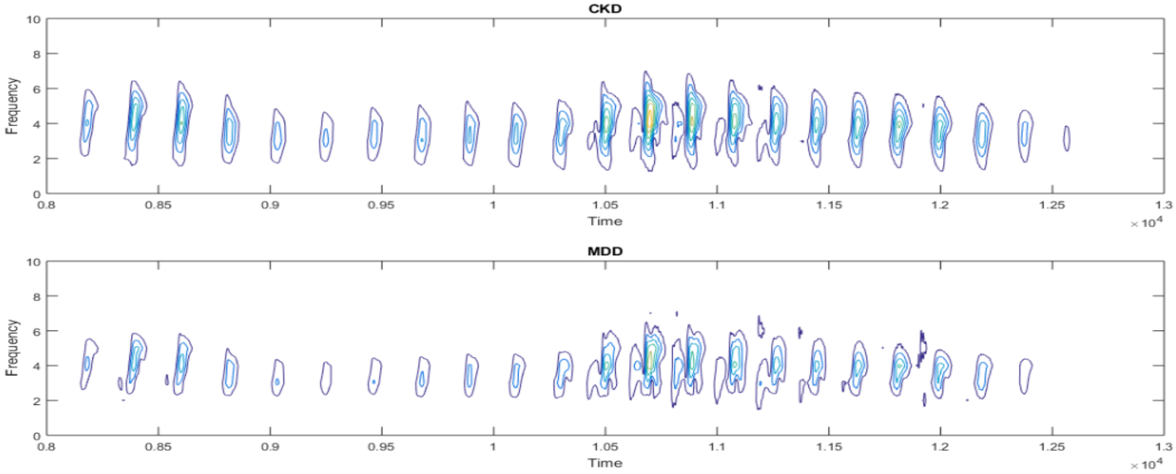

Figure 3 (d): Special Zoom of the 2D contour plot of CKD, and MDD of multichannel AI generated audio signal

Figure 3 (b), is a 2D contour plot illustrating time and frequency characteristics of the multichannel AI generated audio signal, as depicted in Figure 3 (a), is presented using the Wigner-Ville Distribution (WVD) and the Window Wigner-Ville Distribution (WWVD). The visual representation highlights the presence of cross terms in the WVD plot, indicating interference resulting from the interaction between the primary signal and the accompanying noise. At such, it shows the importance of mitigating cross term effects for accurate signal analysis. The WWVD plot, however, reveals a reduction in cross term effects, suggesting improved signal clarity and facilitating more precise feature extraction.

For Figure 3 (c), it presents a 2D contour plot illustrating the time and frequency characteristics of the CKD and MDD analyses applied to the multichannel AI-generated audio signal depicted in Figure 4.5, utilizing CKD and MDD methodologies. From the visual representation, the CKD parameters C, D, and E are set at specific values: C at 1.5, D at 0.1, and E at 0.1. Analysis of various tests involving these parameters reveals that while parameter C can vary between low and high values within its range, parameters D and E perform optimally at lower values, effectively reducing artifact presence. while, increasing their values increases the presence of artifact. For MDD, parameter C is ideally maintained at a low value, with a value of 0.1 employed in this instance. At higher values, it leads to increased artifact presence. Similarly, the threshold value can vary between low and high ranges, but a low value of 0.1 is preferred due to the nonlinearity of the audio signal. When the threshold is high, the direction of angle of the MDD becomes excessively raised, complicating audio signal tracking. Notably, both CKD and MDD analyses demonstrate eradication of cross terms compared to Figure 3 (b), undo, with a slight presence of internal artifacts.

Figure 3 (d) shows a Special Zoom of the 2D contour plot of CKD, and MDD time and frequency plot of the same multichannel AI generated audio signal of Figure 3 (c). The figure shows what the signal consists and the more circles inside each one indicates more frequency at different level and power that have been captured.

4.3 Classification Results of Audio Signals

The Audio signals are classified using the TFDs considered for this research, the plot of the classification accuracy and the discussion are as follows.

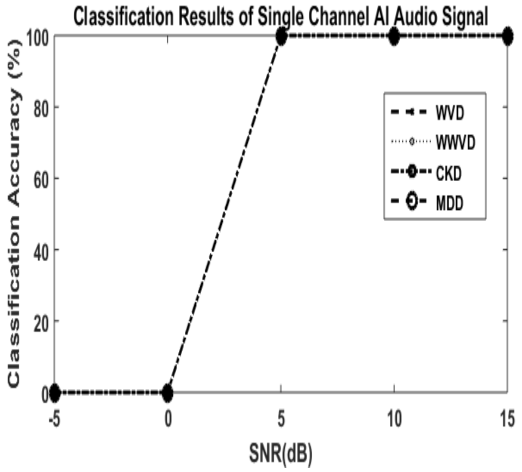

Figure 4 (a)

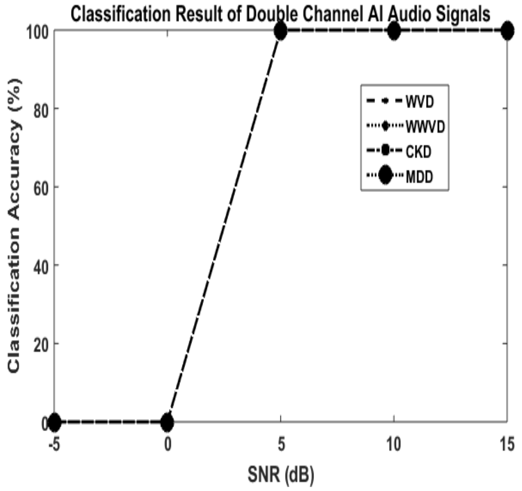

Figure 4 (b)

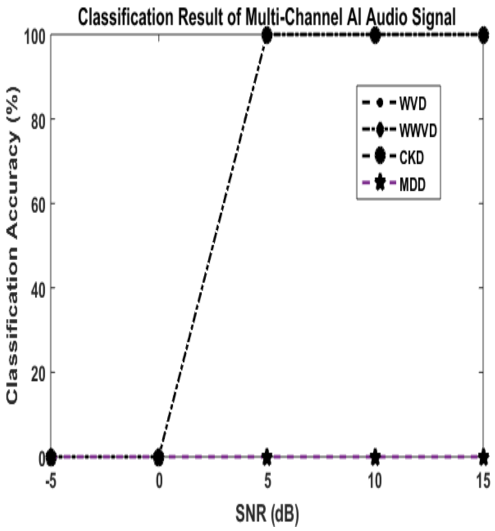

Figure 4 (c)

Figure 4 (a), (b), and (c): Classification Accuracy Results of AI Generated Audio Signals

The classification accuracy results depicted in Figure 4 (a), (b), and (c) reveal interesting variations across different scenarios. Firstly, the consistency of 100% classification accuracy across various Signal-to-Noise Ratio (SNR) levels (0, 5, 10, 15 dB) for single and double-channel configurations suggests robustness in the performance of the TFD methods utilized. However, the stark contrast observed in the multi-channel scenario (c), where the MDD method consistently yields a 0% classification accuracy across all SNR levels, prompts further investigation. The absence of differences in the results of the TFDs, despite variations in SNR levels, could stem from several factors. One possibility is that the TFD methods employed may have inherent limitations in effectively distinguishing between signal and noise components, resulting in consistent accuracy regardless of SNR. Alternatively, it could indicate that the features extracted by the TFD methods do not significantly contribute to classification accuracy in the context of the dataset or classification task at hand. While Regarding the MDD method yielding 0% classification accuracy in the multi-channel scenario, several hypotheses could be considered. One explanation may be that the multi-channel setup introduces complexities or interferences that render the MDD method ineffective in accurately capturing signal characteristics. Additionally, limitations in the MDD algorithm's ability to handle multi-channel data or challenges in parameter tuning for the multi-channel scenario could contribute to the observed results. Further analysis and experimentation are warranted to elucidate the underlying reasons for these observations and to explore potential improvements or alternative approaches to address the identified limitations.

5 CONCLUSION

In this research, the source separation algorithm was developed using time-frequency analysis techniques, specifically WVD, WWVD, CKD and MDD, alongside a decision tree classifier. The performance of these algorithms was evaluated on the AI-generated multichannel audiosignals under various Signal-to-Noise Ratio (SNR) conditions. The results demonstrated that the CKD-based approach achieved a remarkable classification accuracy of 100% at 0dB to 15dB SNR for the multichannel AI-generated audio signals. This high accuracy underscores the efficacy of CKD in enhancing audio processing in noisy environments. MDD also showed significant promise in handling signals with complex spectral structures. The integration of these time-frequency distribution methods with a decision tree classifier proved effective in accurately identifying and separating audio sources, highlighting their potential for real-time applications in audio processing and source separation tasks.

Despite these promising results, several limitations were identified. First, the study relied solely on AI-generated audio signals, which may not fully reflect the complexities of real-world audio data. Additionally, the performance of the algorithms at higher noise levels (above 15dB SNR) was not explored, leaving room for further evaluation in extreme noise conditions. Future research would try to address these limitations by testing the algorithms on real-world audio datasets with diverse noise characteristics and by exploring their robustness under higher SNR conditions. Furthermore, the integration of advanced machine learning techniques, such as deep learning-based classifiers, could be investigated to further enhance the accuracy and generalizability of the proposed methods. This would contribute to advancing the field of audio source separation and its applications in real-world scenarios.

REFERENCES

Ahmad, A. A., Aji, M.M., Abdulkadir, M., Adunola, F.O. & Lawan, S. (2024). Analysis of Normal Radar Signal based onn Different Time-Frequency Distribution Configuration. *Arid Zone Journal of Engineering, Technology and Environment*, 20(4), 699-712.

Ahmed, A., Ahmad, A. A. & F. O. Adunola (2024). Classification of vessel types using the visual geometry group based convolutional neural network. *Global Journal of Engineering and Technology Advances*, 21(02), 88-92. doi: 10.30574/gjeta.2024.21.2.0207

Al-Sa'd, M., Boashash, B., & Gabbouj, M. (2021). Design of an optimal piece-wise spline Wigner-Ville distribution for TFD performance evaluation and comparison. *IEEE Transactions on Signal Processing*, *69*, 3963–3976.

Alghamdi, A., Healy, G., & Abdelhafez, H. (2021). *Machine Learning Algorithms for Real Time Blind Audio Source Separation with Natural Language Detection*. *6*(5), 125–140. https://doi.org/10.25046/aj060515

Bhat, G. S., Reddy, C. K. A., Shankar, N., & Panahi, I. (2019). A computationally efficient blind source separation for hearing aid applications and its real-time implementation on smartphone. *Proceedings of Meetings on Acoustics*, *39*(1), 1–14.

Boashash, B. (2016), *Time-Frequency Signal Analysis and Processing* (2nd Edition). Elsevier Ltd.

Cohen, L. (1995). *Time-frequency analysis* (Vol. 778). Prentice Hall PTR Englewood Cliffs.

Corey, R. M., Skarha, M. D., & Singer, A. C. (2019). Cooperative audio source separation and enhancement using distributed microphone arrays and wearable devices. *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 296–300.

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, *56*(5), 4765-4800.

Gidlöf, A., & Nyberg, D. (2023). *Evaluation of Methods for Sound Source Separation in Audio Recordings Using Machine Learning*. MSc Thesis in Electrical Engineering, Linkoping University, Sweden.

Jorgensen, E. J. (2022). *Acoustic complexity in real-world noise and effects on speech perception for listeners with normal hearing and hearing loss* (Doctoral dissertation, The University of Iowa).

Josephine, P. K., Prakash, V. S., & Divya, K. S. (2021). Supervised Learning Algorithms: A Comparison. *Kristu Jayanti Journal of Computational Sciences (KJCS)*, 01-12.

Li, H., Chen, K., Wang, L., Liu, J., Wan, B., & Zhou, B. (2022). Sound source separation mechanisms of different deep networks explained from the perspective of auditory perception. *Applied Sciences*, *12*(2), 832.

Li, Y., & Ramli, D. A. (2023). Advances in Time-Frequency Analysis for Blind Source Separation: Challenges, Contributions, and Emerging Trends. *IEEE Access*, *11*, 137450-137474.

Martel, H., Richter, J., Li, K., Hu, X., & Gerkmann, T. (2023). Audio-Visual Speech Separation in Noisy Environments with a Lightweight Iterative Model. *ArXiv Preprint ArXiv:2306.00160*, 1–5.

Martinek, R., Jaros, R., Baros, J., Danys, L., Kawala-Sterniuk, A., Nedoma, J., Machacek, Z. & Koziorek, J., 2021. Noise Reduction in Industry Based on Virtual Instrumentation. *Computers, Materials & Continua*, *69*(1).

Michelsanti, D., Tan, Z. H., Zhang, S. X., Xu, Y., Yu, M., Yu, D., & Jensen, J. (2021). An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on*

Audio, Speech, and Language Processing, *29*, 1368-1396.

Richard, G., Smaragdis, P., Gannot, S., Naylor, P. A., Makino, S., Kellermann, W., & Sugiyama, A. (2023). Audio signal processing in the 21st century: The important outcomes of the past 25 years. *IEEE Signal Processing Magazine*, *40*(5), 12-26.

Schulze-Forster, K. (2021). *Informed audio source separation with deep learning in limited data settings*. Ph.D.dissertation Institut Polytechnique De Paris.

Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, *158*, 107020.

Tengtrairat, N., Woo, W. L., Dlay, S. S., & Gao, B. (2016). Online noisy single-channel source separation using adaptive spectrum amplitude estimator and masking. *IEEE Transactions on Signal Processing*, *64*(7), 1881–1895.

Zhu, W. (2024). *Quiet Interaction: Designing an Accessible Home Environment for Deaf and Hard of Hearing (DHH) Individuals through AR, AI, and IoT Technologies* (Doctoral dissertation, OCAD University).