MINISTERUL EDUCAŢIEI NAŢIONALE
ROMÂNIA
UNIVERSITATEA DE MEDICINĂ, FARMACIE, ŞTIINŢE ŞI TEHNOLOGIE „GEORGE EMIL PALADE" DIN TÂRGU MUREŞ

Acta Marisiensis. Seria Technologica
Vol. 21 (XXXVIII) no. 2, 2024
ISSN 2668-4217, ISSN-L 2668-4217

DEVELOPMENT OF AUDIO SOURCE SEPARATION ALGORITHM IN NOISY ENVIRONMENT USING COMPACT KERNEL TIME-FREQUENCY DISTRIBUTION

Amina JIBRIL, Ashraf Adam AHMAD, Sagir LAWAN, Farouk Muhammad ISAH
[1,2,3,4]Nigerian Defence Academy

Dept of Electrical/Electronic Engineering, Nigerian Defence Academy, Kaduna.

[1]amina.jibril2021@nda.edu.ng
[1]aaashraf@nda.edu.ng
[1]slawan@nda.edu.ng
[1]farouk.isah2021@nda.edu.ng

Abstract

This research focuses on the development of a source separation algorithm tailored to significantly enhance audio processing in noisy environments. By utilising advanced signal processing techniques and algorithms based on time-frequency analysis, the study explores the effectiveness of the Compact Kernel Distribution (CKD) for this purpose. Performance was evaluated using key metrics such as Signal-to-Interference Ratio (SIR), Source-to-Distortion Ratio (SDR), and Signal-to-Noise Ratio (SNR). Notable improvements were observed: SIR improved by 4.90% and decreased by 5.36%, while SDR improved by 66.47% and 58.08% at SNR of 5 dB SNR for the audio recordings signals compared to Max-Corr and Simplex-Corr, respectively. The results demonstrate that the developed algorithm significantly enhances SIR, SDR, and SNR metrics, with potential applications across various industries, including speech enhancement.

.

Key words: compact kernel distribution, wigner-ville distribution, digital signal processing, automatic speech recognition, additive white gaussian noise.

1. Introduction

Audio source separation refers to the process of separating individual sound sources from a mixture of audio signals. It has numerous applications in audio processing, particularly in noisy environments, where it can significantly improve the quality and intelligibility of audio signals [1].

Also, Audio source separation is a technology designed to isolate one or more specific source signals from an audio recording containing multiple sound sources [2]. This technology is particularly valuable in scenarios 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 [1][2].

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 [2]. 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 hinder 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.

This research aims to address these challenges by configuring multichannel noisy audio signals using recorded audio and developing source separation algorithms based on time-frequency analysis. The effectiveness of these algorithms will be analysed and validated to ensure significant improvements in audio source separation in noisy environments.

Recent advancements in audio source separation have significantly addressed challenges in both signal processing and practical applications. A comprehensive review in 2018 [3] highlighted the transition of multichannel audio source separation techniques from theoretical models to real-world scenarios, addressing issues like moving sources, varying sensor counts, and high reverberation levels. This review emphasized the need for future research to combine array processing, source separation, and machine learning to enhance robustness in complex conditions. Another key review [4] provided an extensive overview of machine listening research, covering advancements in sound source detection, classification, localization, enhancement, and separation. It introduced a workflow for selective hearing systems, emphasizing the importance of real-time processing and perceptual quality, and suggested further exploration of joint models for Active Noise Control (ANC) and machine listening. Complementing these reviews, a 2020 study [5] introduced blind audio source separation methods using expectation-maximization algorithms, showing improved speech quality and intelligibility. In 2021, an Enhanced Residual Filter (ERF) approach combined deep neural networks with traditional algorithms to improve localization performance [6], and research on real-time separation algorithms, such as ConvTasNet and Demucs, demonstrated significant performance improvements [7]. Additionally, various separation methods and optimization techniques, including the auxiliary function-based discriminative nonnegative matrix factorization, were validated for their effectiveness [8].

Despite these advancements, existing methods still face challenges, such as dealing with high reverberation, spatially diffuse sources, and synchronization delays, which limit their practical application in diverse and noisy environments. Current techniques often struggle with scalability and computational efficiency, particularly in multi-channel scenarios. To address these limitations, this research proposes using Compact Kernel Distribution (CKD) time-frequency distribution (TFD) methods, aimed at enhancing the resilience and reliability of audio source separation in noisy environments. This approach leverages advanced signal processing techniques to improve performance and adaptability, thereby contributing to the field by addressing key challenges in audio processing applications.

2. METHODOLOGY

The methodology employed in this study focuses on developing and validating an audio source separation algorithm designed specifically for noisy environments. Utilising compact kernel time-frequency distribution. This section outlines the systematic application of these methodologies, detailing their theoretical foundations, implementation strategies, and expected contributions to improving signal clarity and intelligibility in audio recordings.

A. BLOCK DIAGRAM OF AUDIO SOURCE SEPARATION SYSTEM

The working of the system is best summarized with a block diagram-which represented with blocks shown in Figure 1.

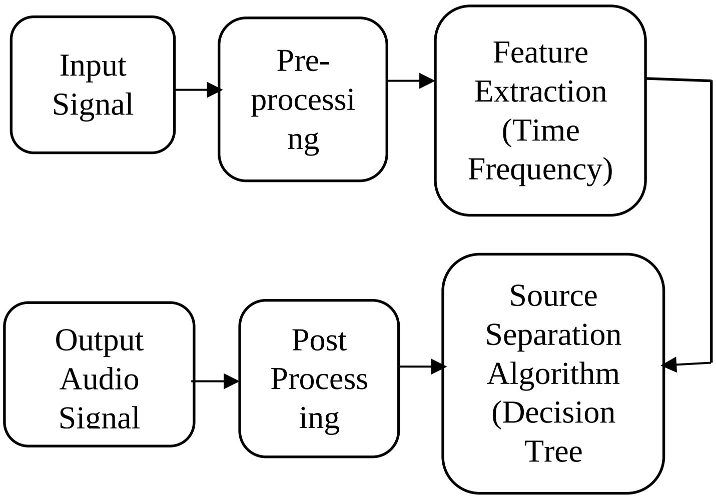
Input Signal
Pre-processing
Feature Extraction (Time Frequency)
Source Separation Algorithm (Decision Tree
Post Processing
Output Audio Signal

Fig. 1 Block diagram of the Audio Source Separation system

From Fig. 1, the process begins with the input audio signal, containing a mixture of desired sources and unwanted noise or interference. Following this, preprocessing steps are applied to remove artifacts and unwanted components, including filtering and noise reduction. Feature extraction is then performed using this techniques Wigner- vile distribution (WVD), Window Wigner- vile distribution (WWVD), Compact kernel distribution (CKD) TFD, capturing spectral and temporal characteristics. These features serve as input to the source separation algorithm, which utilises a decision tree classifier to separate individual audio sources from the mixture. Post-processing techniques, such as denoising, is then applied to enhance the quality of the separated sources. Finally, the output audio signal comprises the isolated desired sources, effectively separated from the unwanted noise or interference.

B. AUDIO RECORDING SIGNALS

The audio signals, for the research, employed a data collection strategy for speech enhancement for High-quality audio recordings was captured, and categorized by speaking scenarios in noisy environments to construct the foundational dataset. The audio recorded signal, is derived from Waveform Audio File Format (WAV) files sourced from the Kaggle Dataset known as LJ Speech (https://www.kaggle.com/dataset/mathurinache/the-lj-speech-dataset) [9]. LJ Speech is a publicly available speech dataset which comprises of 13,100 brief audio clips featuring a solitary speaker reciting passages from seven non-fiction books. Each clip is accompanied by a corresponding transcription. The clips exhibit varying durations, spanning from 1 to 10 seconds, with a cumulative duration of approximately 24 hours. The texts, published between 1884 and 1964, fall within the public domain. Librivox project recorded the audio in 2016-2017, and the recordings are also part of the public domain. The single channel audio recording was then mixed together to form the multichannel audio using the algorithm developed. The mixing equation used in audio signal processing, represents the output signal $y_c(n)$ at discrete time (n) as the sum of the products of input signals $x_k\ (n)$with corresponding coefficients a_k, where (K) is the total number of input signals [10] is show in Equation (1):

$$y_c(n) = \sum_{k=1}^{k} a_k\ .x_k\ (n) \qquad (1)$$

Where: $y_c(n)$is the mixed output signal at time (n), $x_k\ (n)$is the k^{th}input signal at time (n) and

$a_k$is the coefficient or gain applied to the k^{th}input signal.

Noise Signals

The noise signal used is a database of 16-channel environmental noise recording name DEMAND (Diverse Environments Multichannel Acoustic Noise Database) by Emmanuel Vincent *et'al*, consisting of 16 single channel WAV files in one directory at both 48KHz and 16KHz sampling rate and all files were compressed into 'zip' files [11]. Two zip files were used for this research work which are OHALLWAY_16k.zip and OFFICE_16k.zip to denote common interference of hallway and office respectively. The single channel WAV file noise signal downloaded was then added to the multichannel audio signal.

Additive White Gaussian Noise (AWGN) is a type of interference commonly inserted into signals within communication systems. It is referred to as "additive" because it is added to the original signal, "white" because it maintains consistent power spectral density across all frequencies, and "Gaussian" because it adheres to a Gaussian (normal) probability distribution. When evaluating the performance of audio signals, consideration of such noise is essential. In the signal pre-processing phase, a standard AWGN model is utilized to generate and introduce noise into all audio signals. Equation (2) describe the resulting signal (y) is formulated as the sum of the input signal (x) and the noise component (n) [12].

$$y = x + n \qquad (2)$$

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.

C Time-Frequency Analysis/Distribution

Time-frequency analysis/distribution (TFD) is a signal processing technique utilized to examine and depict signals across both time and frequency domains. It offers a comprehensive representation of signals that vary over time and frequency, enabling the analysis of signals with multiple time-varying frequencies [13]. Applications of time-frequency analysis/distribution techniques span diverse fields such as multichannel audio source separation, music source separation, and speech separation. The specific TFD methods employed in this study are detailed as follows:

I. Wigner-Ville Distribution of a Signal (WVD)

The Wigner-Ville Distribution (WVD) is a method used to estimate the power spectral function of a nonstationary signal by analysing its time-frequency energy distribution. Initially introduced by Wigner and later adapted by Ville for signal processing applications [14], the WVD is represented as $P_{z,}WVD(t,f)$ and is represented mathematically in Equation (3):

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

Here, $P_{z,}WVD(t,f)$denotes the Wigner-Ville distribution of a signal at time (t) and frequency (f), where $z\left(t + \frac{\tau}{2}\right)$ represents a complex-valued function indicating the analysing signal or window function, and * denotes complex conjugation. Due to the impracticality of theoretically evaluating over infinite limits, a pseudo-Wigner-Ville distribution (PWVD) addresses this by employing a running window [14]:

$$P_{z,}\mathrm{PWVD}(t,f) = \int_{-\infty}^{\infty} h(\tau) z\left(t + \frac{\tau}{2}\right) z^*(t - \frac{\tau}{2}) e^{-j2\pi f\tau} d\tau \qquad (4)$$

where $h(\tau)$ is the window function. The WVD utilizes specific kernel functions from the bilinear generalized class of time-frequency distributions, demonstrating effective time-frequency resolution, particularly in low Signal-to-Noise Ratio (SNR) conditions [14]. However, inherent artifacts and cross terms limit its performance. To mitigate these issues, the windowed WVD (WWVD) incorporates a window function post obtaining the instantaneous autocorrelation function (IAF). The Hamming window is preferred for its enhanced frequency resolution and suppression of side lobes [15]:

$$P_{z,\mathrm{WWVD}}(t,f) = \int_{-\infty}^{\infty} g_w(\tau)\, z\,(t + \frac{\tau}{2})\, z^*\,(t - \frac{\tau}{2})\, e^{-j2\pi f\tau} d\tau \qquad (5)$$

$$P_{z,\mathrm{WWVD}}(t,f) = \int_{-\infty}^{\infty} 0.54 - 0.46\cos\left(\frac{2\pi\tau}{T}\right) z\,(t + \frac{\tau}{2})\, z^*\,(t - \frac{\tau}{2})\, e^{-j2\pi f\tau} d\tau \qquad (6)$$

II Compact Kernel Distribution (CKD)

Compact Kernel Distribution (CKD): The Compact Kernel Distribution (CKD) technique represents an enhanced iteration of the pseudo-Wigner-Ville Distribution (WVD), designed to feature a kernel window function with compact support that effectively diminishes outside a specified range in the ambiguity domain [14]. Unlike Gaussian windows of infinite length, CKD avoids the need for truncation using rectangular windows, thereby preserving information. CKD is distinguished for its capability to suppress cross-terms while maintaining high resolution of auto-terms, achieved through a combination of compact support and adaptable adjustments to the kernel's shape and size independently:

$$g(v,\tau) = G_1(v) g_2(\tau) = \begin{cases} e^{2c\,\frac{cD^2}{eV^2 - D^2} + \frac{cE^2}{\tau^2 - E^2}} & |V| < D, |\tau| < E, \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

where ν and τ represent Doppler and lag windows determined by parameters D and E, and C controls the shape. The width of the kernel in the ambiguity domain is determined based on prior knowledge of the signal components. The Instantaneous Autocorrelation Function (IAF) of WVDs, forms the basis of CKD, and its Time-Frequency Distribution (TFD) is given in Equation (8):

$$P_{z,\mathrm{CKD}}(t,f) = \int_{-\infty}^{\infty} g(t,\tau) * z\,(t + \frac{\tau}{2})\, z^*\,(t - \frac{\tau}{2})\, e^{-j2\pi f\tau} d\tau \qquad (8)$$

Where $g(t,\tau)$ is derived from:

$$g(t,\tau) = \int_{-\infty}^{\infty} g(v,\tau) e^{-j2\pi v\tau} dv \qquad (9)$$

From this, the CKD TFD is expressed as:

$$P_{z,\mathrm{CKD}}(t,f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(v,\tau)\, z\left(u + \frac{\tau}{2}\right) z^*\left(u - \frac{\tau}{2}\right) e^{-j2\pi f\tau}. e^{-j2\pi v\tau} du dv d\tau \qquad (10)$$

Parameters such as the kernel shape C, Doppler cut-off D, and lag cut-off E are selected based on prior knowledge of the signal components. Table 1 outlines the ranges for these parameters.

The implementation of CKD is facilitated through the MATLAB function `tf_kernel_ckd`, which generates respective kernels based on specified parameters. The IAF function `IAF_CKD` smoothes the IAF in the lag domain using these kernels, followed by applying CKD to obtain the TFD. By employing these methods, the CKD approach enhances time-frequency analysis, particularly suitable for signals exhibiting complex energy distributions in the (t, f) domain.

Table 1: C, D, E Range of Values for CKD Kernel [16]

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

By implementing these techniques, the CKD method improves time-frequency analysis, particularly for signals exhibiting complex energy distributions in the (t, f) domain.

III Instantaneous Power Formulations From TFD:

Following the completion of TFD analysis, the subsequent step involves examining the results obtained from both WVD and CKD analyses. Typically, TFD involves several parameters that influence its characteristics and applicability. These parameters allow analysts to customize the TFD method to suit specific applications and signal attributes. This research specifically focuses on time-domain power analysis. More precisely, the analysis in this study concentrates on instantaneous power (IP), calculated as the integral of TFD over frequency [14]. The mathematical representation of instantaneous power at time t is given by Equation (11):

$$P(t) = \int_{-\infty}^{\infty} P_z(t,f) df \qquad (11)$$

D. Performance analysis/ simulation set-up

Performance indicators for audio processing are measures used to evaluate the effectiveness and quality of audio processing algorithms. These indicators provide objective assessments of the performance of the algorithms in various aspects, such as noise reduction, source separation, and audio quality[2].

Objective Performance Indicators:

Energy Ratio (ER), Signal-to-Distortion Ratio (SDR), Signal-to-Noise Ratio (SNR), Signal-to-Interference

Ratio (SIR): These indicators focus on quantifying the quality of the processed audio signal objectively. They provide numerical measures that can be used to compare different algorithms and assess their performance.

Subjective Performance Indicators:

Perceptual Evaluation of Audio Quality (PEAQ), Mean Opinion Score (MOS), and Objective Difference Grade (ODG): These indicators consider

human perception and provide subjective evaluations

of the audio quality. They are obtained through listening tests and subjective assessments by human listeners.

Signal-to-Distortion Ratio (SDR): SDR is a commonly used objective performance indicator for Audio source separation. It measures the ratio of the power of the desired source signal to the power of the distortion introduced during the separation process. Higher SDR values indicate better separation performance [17]. Mathematically it represented in Equation (12):

$$\mathrm{SDR} = 10 * \log 10 \frac{(\text{Power of Source Signal})}{(\text{Power of Distortion})} \qquad (12)$$

Signal-to-Interference Ratio (SIR): SIR measures the ratio of the power of the desired source signal to the power of interfering sources or artifacts. It assesses the ability of the algorithm to suppress unwanted sources or artifacts in the separated signal. Higher SIR values indicate better separation performance [18]. Mathematically it represented in Equation (13):

$$\mathrm{SIR} = 10 * \log 10 \frac{(\text{Power of Source Signal})}{(\text{Power of Interfering Sources})} \quad (13)$$

Signal-to-Noise Ratio (SNR)

SNR is a fundamental objective metric used to quantify the ratio of the desired signal (source) to the unwanted noise in an audio signal. It provides a numerical value that measures the extent to which the algorithm successfully separates the target audio source from the surrounding noise.

In other word SNR is a measure of the ratio of the power of a signal to the power of the noise that interferes with the signal. It is a widely used objective measure of audio quality and is used to evaluate the effectiveness of noise reduction algorithms [19]. Mathematically, SNR is defined in Equation (14):

$$\mathrm{SNR} = 10.\log_{10}\left(\frac{\text{Power of signal(source)}}{\text{Power of noise}}\right) \qquad (14)$$

S/N	**Signal**	**Sampling frequency (fs)**	**Duration in seconds (sec)**
2	Audio Recording	22.050KHz	6.5 sec
3	Hallway Interference	16KHz	4.5 and 6.5 sec for AI and Audio Recording Respectively
4	Office Interference	16KHz	4.5 and 6.5 sec for AI and Audio Recording Respectively

Table 2 Simulation Set-up Values

From Table 2 Simulation Set-up Values above shows the sampling frequency and the length (duration in time) of the multichannel audio sources used for the Audio Recording signals. Also, the values of the sampling frequency and the length (duration in time) for Noise signal (hallway and office interference) downloaded.

Table 3 Simulation Set-up Parameters Values

S/ N	**SIGNAL**	**PARAMETERS C, D, and E for CKD**		
		C	D	E
2	Audio Recording	1.5	0.1	0.1

While, Table 3 Simulation Set-up Parameters Values shows the range of parameters (C, D, and E) used for the TFD algorithm developed for CKD kernel respectively for the multichannel audio source used for the Audio Recording signals based on their range of values as mention in Table 1.

The 'SIR_SDR' algorithm developed calculates the Signal-to-Interference Ratio (SIR) or Signal-to-Distortion Ratio (SDR) of a given signal series. It takes the input series 'IP' and a threshold 'thr' as its inputs. The function operates on the input series in blocks of

100 samples, determining whether each block represents a signal or noise based on whether its amplitude exceeds a threshold percentage of 1% selected of the maximum amplitude. It then computes the total signal and noise powers and calculates the Signal Ratio (SR), which may represent SIR or SDR, based on specific applications.

Furthermore, 'SR_Setup' and 'SR_Setup1', intended for both AI-generated and audio-recorded signals, evaluate the accuracy of the Signal-to-Interference Ratio (SIR) and Signal-to-Distortion Ratio (SDR) under various Signal-to-Noise Ratio (SNR) conditions. These setups generate noisy versions of a signal and compute the WWVD and CKD for each noisy version. Subsequently, they estimate the Instantaneous Power (IP) based on the WWVD and CKD and calculate the SIR or SDR for each IP. Ultimately, the setups average the SIR or SDR over Monte Carlo loops and plot the results for different SNR values.

3. RESULTS AND DISCUSSION

Results and discussions are presented in this section.

Audio Recorded Signals

Figure 2 shows the time representation plot for Audio Recording signals while Figure 3 shows the frequency representation plot for the multichannel Audio Recording signals

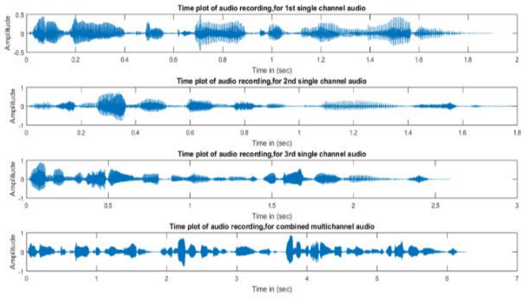

Fig 2: The time plot of Audio Recording signal

Figure 2 shows the time plot of the Audio Recording signals 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.

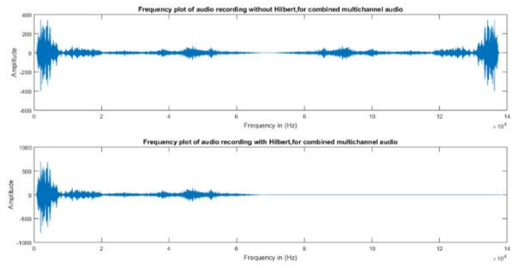

Fig 3: The frequency plot of Multichannel Audio Recording signal

Figure 3 shows the frequency plot of the multichannel Audio Recording signal without Hilbert and with Hilbert. Where the plot without Hilbert represents the original audio signal or the normal version and the plot with Hilbert represent the analytic audio signal. The analytic audio signal was obtained by applying the Hilbert transform to the original audio signal. Also, it is seen in the second aspect of Figure 3 that the non-required mirrored version has been eliminated.

Time-frequency representations (TFRs) of the TFDs considered in the research

This section presents the multi-channel Audio recording signals Time Frequency Representations of the TFDs considered in this research.

3D TFR of WVD for audio recording signal

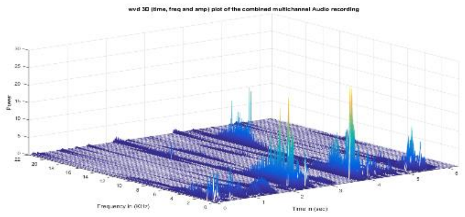

Fig. 4 3D plot of multichannel Audio Recording signal.

Figure 4 shows the correlation between power, time, and frequency in the typical audio signal utilising WVD. The signal, has a center frequency of 2 KHz, a sampling frequency of 22 KHz and SNR of 10 dB.

2D TFR of WVD, AND WWVD for audio recording signal

Figure 5 shows the two-dimension (2D) contour plot of multichannel Audio recording signal.

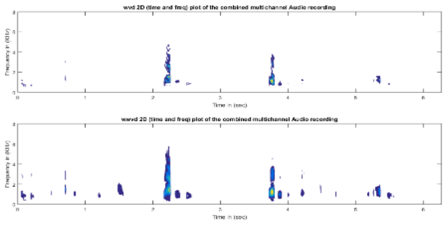

Fig. 5 2D contour plot of multichannel Audio Recording signal.

Figure 5 shows the 2D contour plot of the multichannel Audio Recording signal, as depicted in Figure 4, 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. This comparative analysis validates the effectiveness of employing the WWVD methodology in reducing interference and enhancing the clarity of underlying signal components, thereby justifying selection for further algorithmic design and analysis in subsequent stages of the research.

2D TFR of CKD for audio recording signal;

Figure 6 shows the two-dimension (2D) contour plot of CKD multichannel Audio Recording signal

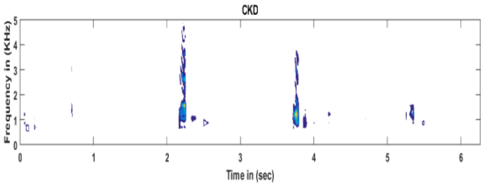

Fig. 6 2D contour plot of CKD multichannel Audio Recording signal.

Figure 6 presents utilizing 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. Notably, CKD analyses demonstrate eradication of cross terms compared to Figure 5, undo, with a slight presence of internal artifacts.

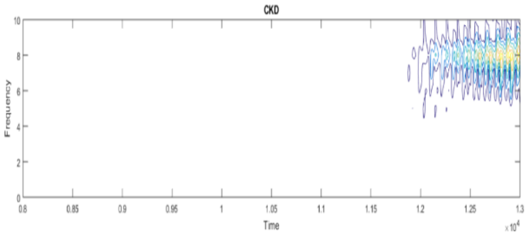

Fig. 7: Special Zoom of the 2D contour plot of CKD multichannel Audio Recording signal.

From Fig. 7, shows a Special Zoom of the 2D contour plot of CKD time and frequency plot of the multichannel Audio Recording signal of Figure 6. 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.

performance analysis results of audio signals based on SIR, SDR and SNR

The plots of the performance analysis considered for this research is presented and discussed in this section. The Audio signal performance analysis result for Audio Recording is given in Figure 8.

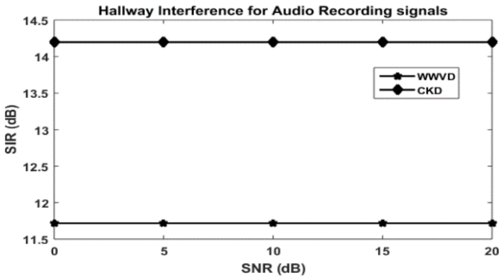

Fig. 8 (a)

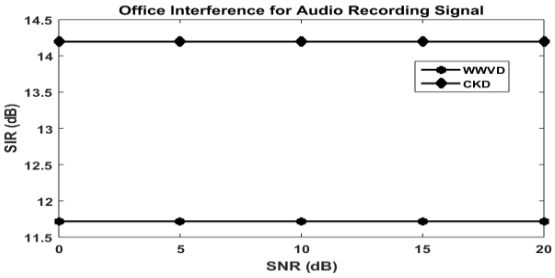

Fig. 8 (b)

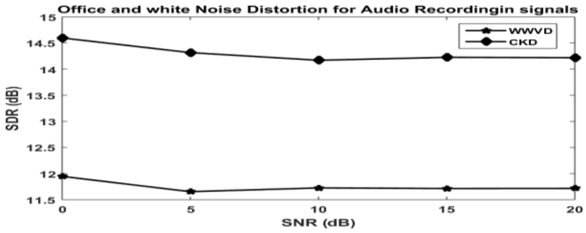

Fig. 8 (c)

Fig. 8 (a), (b), and (c): Performance Analysis Results of Audio Recording Signals

Figure 8 (a), (b), and (c) shows Performance Analysis Results of multi-channel Audio Recording Signals for 10 Iteration for each of the method with SNR from 0 to 20 dB at interval of 5dB, where (a) shows the result Hallway Interference, (b) Office Interference and (c) SDR. So also, from the result obtained, CKD outperformed WWVD with higher SIR in figure 8 (a) with 14.19, (b) with 14.19 and (c) with SDR of 14.16. As stated in section 3.8, ‘Higher SIR or SDR values indicate better separation performance’ at such, CKD has a better separation than WWVD

Performance validation analysis

The research paper used for validation of this thesis used two novel methods for separating speakers in multi-speaker audio recordings, especially when speakers have unbalanced or infrequent activity levels [20]. The first method, called the maximum correlation (Max-Corr) method, uses linear programming to maximize correlation between time frames to identify single-speaker frames. The second method, called the simplex correlation (Simplex-Corr) method, utilizes convex geometry tools on correlation vectors to detect vertices corresponding to single-speaker frames. Both methods estimate the activity probabilities of each speaker from the detected single-speaker frames. These estimated probabilities are then used to compute spectral masks for separating and enhancing the individual speaker signals via spatial and spectral processing.

Result Validation for Audio Recorded Signal

Figure 9 (a), and (b) below shows the SIR and SDR Results Validation for the multichannel Audio Recording Signals respectively.

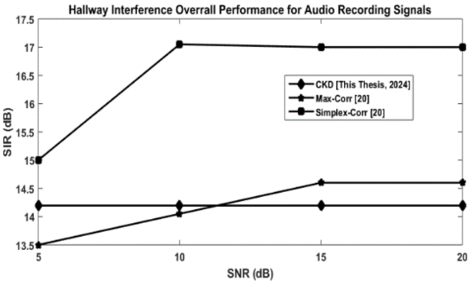

Fig. 9 (a)

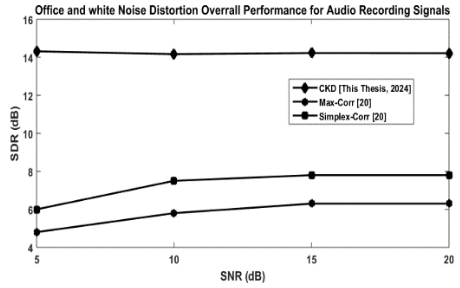

Fig. 9 (b)

Fig. 9 (a), and (b): Results Validation for Audio Recording Signals

Figure 9 (a), and (b) are the Results of Validation for Audio Recording Signals, based on SIR and SDR performance validation considered for this research. The SIR of the CKD for this thesis recorded an improvement of 4.90% compare to the Max-Correlation with a decrease of 5.36% for the Simplex-Correlation of [20] compare to CKD of this thesis at SNR of 5 dB respectively was achieved. While SDR also recorded percentage improvement of 66.47% and 58.08% for both Max-Correlation and Simplex-Correlation respectively compared to [20].

4. Conclusion

This research successfully developed an audio source separation algorithm tailored for noisy environments using Compact Kernel Distribution (CKD) and time-frequency analysis techniques. The algorithm demonstrated significant improvements in Signal-to-Interference Ratio (SIR), Source-to-Distortion Ratio (SDR), and Signal-to-Noise Ratio (SNR), achieving notable performance improvement: SIR improved by 4.90% and decreased by 5.36%, while SDR improved by 66.47% and 58.08% at 5 dB SNR for the audio recordings compared to Max-Corr and Simplex-Corr, respectively. These advancements highlight the potential of the developed algorithm in applications such as speech enhancement and other audio processing tasks. For future research, integrating the developed algorithms with emerging technologies such as deep learning or reinforcement learning techniques could be explored. This integration may further enhance the algorithms' capabilities and adaptability to dynamic audio environments, paving the way for even more improved and versatile audio processing solutions.

References

[1] J. Chien, *Source Separation and Machine Learning*. Taiwan: Academic Press, 2019.

[2] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. West Sussex, UK: John Wiley & Sons Ltd, 2018.

[3] L. Girin, S. Gannot, and X. Li, “Audio Source Separation into the Wild,” in *Multimodal behavior analysis in the wild: Advances and Challenges*, N. S. X. Alameda-Pineda, E. Ricci, Ed., Academic Press, 2018, pp. 53–78.

[4] C. Estefan and L. Hanna, “Selective Hearing: A Machine Listening Perspective,” in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, Germany, 2019, pp. 1–6.

[5] A. Eisenberg, B. Schwartz, and S. Gannot, “Blind audio source separation using two expectation-maximization algorithms,” in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, Israel, 2020, pp. 1–6.

[6] L. Cheng, X. Sun, D. Yao, J. Li, and Y. Yan, “Estimation reliability function assisted sound source localization with enhanced steering vector phase difference,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 421–435, 2020.

[7] A. Alghamdi, G. Healy, and H. Abdelhafez, “Real Time Blind Audio Source Separation Based on Machine Learning Algorithms,” *2020 2nd Nov. Intell. Lead. Emerg. Sci. Conf.*, pp. 35–40, 2020.

[8] Li Li, “Study on audio source separation algorithms under various conditions, ranging from determined to more realistic conditions,” 2021.

[9] “The LJ Speech Dataset.” Accessed: Jan. 11, 2024. [Online]. Available: https://www.kaggle.com/datasets/mathurinache/the-lj-speech-dataset

[10] S. Marchand and P. Mahe, “Informed Source Separation for Stereo Unmixing--An Open Source Implementation,” in *International Conference on Digital Audio Effects (DAFx)*, 2023, pp. 102–109.

[11] “DEMAND_ a collection of multi-channel recordings of acoustic noise in diverse environments.” Accessed: Nov. 18, 2023. [Online]. Available: https://zenodo.org/records/1227121#.YlbIxWj0mUk

[12] L. Cohen, *Time-frequency analysis*, vol. 778. Prentice Hall PTR Englewood Cliffs, 1995.

[13] “Time–frequency analysis - Wikipedia.” Accessed: Aug. 30, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Time–frequency_analysis

[14] B. Boashash, “Time-Frequency Signal Analysis and Processing,” in *Time-Frequency Signal Analysis and Processing*, 2nd Editio., B. Boashash, Ed., 125 London Wall: Elsevier Ltd, 2016, pp. 31–63.

[15] A. A. Ahmad, A. Daniyan, and D. O. Gabriel, “Selection of window for inter-pulse analysis of simple pulsed radar signal using the short time Fourier transform,” 2015.

[16] M. Al-Sa’d, B. Boashash, and M. Gabbouj, “Design of an optimal piece-wise spline Wigner-Ville distribution for TFD performance evaluation and comparison,” *IEEE Trans. Signal Process.*, vol. 69, pp. 3963–3976, 2021.

[17] J. Abel, M. Kaniewska, C. Guillaumé, W. Tirry, and T. Fingscheidt, “An instrumental quality measure for artificially bandwidth-extended speech signals,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 384–396, 2016.

[18] D. Barry, “Real-time Sound Source Separation For Music Applications Real-time Sound Source Separation For Music Applications,” Doctoral Thesis, Technological University Dublin, 2019. doi: 10.21427/rn03-2738.

[19] M. Torcoli, T. Kastner, and J. Herre, “Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1530–1541, 2021.

[20] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, “Audio source separation by activity probability detection with maximum correlation and simplex geometry,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, pp. 1–16, 2021.

[1]