

RESEARCH ARTICLE

Fusion of Multimodal Audio Data for Enhanced Speaker Identification Using Kolmogorov-Arnold Networks

ARYAMAN TAMOTIA¹, DHARUV SHANTU KARMOKAR¹, RETHI KOMAL¹,
K. KHADAR NAWAS¹, A. SHAHINA², AND A. NAYEEMULLA KHAN¹

¹School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India

²Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Tamil Nadu 603110, India

Corresponding author: K. Khadar Nawas (khadarnawas.k@vit.ac.in)

This work was supported in part by the Department of Science and Technology (DST), India, under the Fund for Improvement of Science and Technology Infrastructure in Universities and Higher Educational Institutions (FIST) Program under Grant SR/FST/ET-I/2022/1079; and in part by the Vellore Institute of Technology (VIT) University.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Speaker identification using audio data is quite challenging because of inherent differences between people, ambient noise, and variable recording conditions. Although the classical deep learning methods are effective, they have rather high computational cost, which leads to usually cumbersome parameter tuning processes and hence reduce their applicability to real-world deployments. In this paper, we discuss an efficient version of Kolmogorov-Arnold Network for audio-based speaker identification using three different types of audio inputs: bone conduction, air conduction, and throat microphone recordings, alongside evaluation on the publicly available LibriSpeech dataset to enhance reproducibility. To optimize the performance of these modalities, we investigate data-level, feature-level, and decision-level fusion techniques. An exhaustive comparison is performed against some of the prominent deep learning architectures, namely VGG-19, ResNet-152, EfficientNetV2, and DenseNet-121, both as standalone architectures and within multimodal combinations for the purpose of evaluating the proposed framework. We demonstrate that our model achieves a significant increase in accuracy up to 99.375% through the use of MFCC and GFCC feature inputs with data-level fusion through MFCC features and achieves this level of accuracy using substantially fewer parameters than those of conventional deep learning models. This makes the KAN framework a highly efficient choice for audio-based speaker identification. One of the critical innovations within this design is a class of trainable activation functions that eliminate the need for traditional weight parameters, thereby providing scalable and computationally efficient solutions. This work contributes to the field of biometrics by providing a refined, resource-efficient approach to speaker identification based on multimodal audio data.

INDEX TERMS Kolmogorov-Arnold networks, speaker identification, multimodal fusion, audio classification, MFCC, GFCC, trainable activation functions.

I. INTRODUCTION

Speaker identification using audio data is one of the essential concepts in modern biometric security and authentication

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy¹.

systems [1], [2]. In the context of this work, “speaker identification” refers specifically to the biometric task of recognizing a speaker’s identity from audio data, often interchangeably termed as “speaker recognition” in the literature. Increased demand for efficient, non-intrusive methods of identity verification arises in conjunction with the

overall growth of digital communication and the increasing demand for secure remote interactions. Audio-based speaker identification provides a very interesting solution because of the inherent convenience of the approach and the availability of audio capture devices as a part of modern technology. However, the process is fraught with problems due to variation in voice characteristics among different individuals and noise in the environment and inconsistencies in recording conditions [3], [16].

Traditionally, in audio-based identification, a lot of reliance is placed on manual features like Mel-frequency Cepstral Coefficients and Gammatone Frequency Cepstral Coefficients [4], [16]. For the capture of spectral envelopes of audio signals, MFCCs are useful as they model human auditory perception with a nonlinear Mel scale. However, GFCCs approximate cochlear filtering using gammatone filterbanks, hence achieving better robustness to noise as it more closely models human auditory system frequency selectivity [9], [16]. These features have proven successful in encoding vocal tract information and, thus, enabling speaker recognition tasks; however, they often tend to be limited in their robustness to various adverse acoustic environments, as reported in [1], [6]. Extraction and processing such features at scale is computationally expensive and can be prohibitive in resource-constrained settings.

Deep learning has revolutionized the field of audio processing and brought advanced methodologies for extracting features and classification [3], [5]. The architectures like VGG-19, ResNet-152, EfficientNetV2, and DenseNet-121 are regarded as highly developed deep architectures of neural networks that have been observed to possess great abilities in the direct learning of complex audio patterns from raw data [7], [8], [10], [14]. These models exploit the deep hierarchical structure to automatically extract salient features leading to significant enhancements in identification accuracies. In addition to MFCCs and GFCCs, other studies commonly leverage spectrograms and log-mel filterbank energies as feature representations in speaker recognition tasks, further enhancing the robustness of learned models. However, practical deployments of such a model are very much limited due to massive computational requirements, protracted training times, and dependency on large volumes of labeled data to achieve effective learning [5], [6], [14]. Additionally, architectures such as Time-Delay Neural Networks (TDNN) are widely adopted in speaker recognition tasks, particularly in x-vector-based systems, due to their effectiveness in capturing temporal speech dependencies [30]. These also usually operate on single-modal data acquired mainly from air-conduction microphones, thus less robust and sensitive to environmental noise and varying speaker characteristics.

In order to overcome these disadvantages, the use of multimodal inputs of different types of microphones, including bone conduction and throat microphones, is considered a promising approach [9], [10]. Audio signals captured using bone conduction microphones rely on vibrations

transmitted via the skull and are therefore particularly robust against ambient noise, moreover, and gain access to vocal qualities that could not easily be captured with conventional air-conduction microphones. Throat microphones capturing vibrations directly from the vocal cords provide another dimension of audio information less vulnerable to extrinsic noise and with a capacity for finer-grained phonation features [13]. Combination of these modalities can improve the robustness and accuracy of speaker recognition systems, especially in challenging acoustic environments [9], [17]. In this study, we used the Temco HG-17 bone conduction microphone, the Tactic Throat Microphone for throat input, and a Rode NT-USB Mini for conventional air conduction audio capture.

Despite its promise, there are still relatively few studies focusing on speaker re-identification using multimodal audio data [10], [17]. Moreover, exploration of different architectures than can be useful in processing multimodal inputs in an efficient manner is crucial towards realizing the potential of such a system. This is one of the reasons why the KAN architecture—invented based on the Kolmogorov superposition theorem—was assumed to be useful in addressing some of the aforementioned challenges [12], [13], [15]. KANs enable one to model multivariable functional relations as composition of several univariate functions with addition operations, thus lowering possible computational complexity and requirements on parameters.

This paper introduces development on robust and effective speaker identification based on multimodal audio data by utilizing the efficient KAN framework. In our work, we explore data-level, feature-level, and decision-level fusion techniques to enhance the fusion of bone conduction, air conduction, and throat microphone recordings for speaker identification. The proposed KAN model is tested with a rigorous comparison with established deep learning architectures in single-modal and multimodal settings, using MFCC and GFCC features to capture a comprehensive representation of the audio signals. Our contributions are valuable insights into the efficacy of KANs in speech and biometric applications and the benefits of multimodal fusion for audio-based speaker identification.

II. RELATED WORK

A. TRADITIONAL AUDIO-BASED SPEAKER IDENTIFICATION

Audio-based speaker identification has been one of the most prominent areas of research in the field of biometrics, where early systems relied heavily upon single-modal audio inputs and handcrafted techniques of feature extraction. MFCCs have gained immense popularity because they effectively capture the short-term power spectrum of audio signals on a Mel scale, more closely approximating the human auditory system's response than linear frequency scales [4], [16]. This feature extraction technique transforms an audio signal into the frequency domain; it also applies a Mel-scaled filterbank in order to calculate the discrete cosine transform, thus decorrelating coefficients.

GFCCs use gammatone filters instead in the cochlear filtering function of the human ear [9], [16]. GFCCs are very useful with high noise since they emphasize the spectral features that are more resistant to noise [6], [9]. The process of using GFCCs involves the application of a bank of gammatone filters to the audio signal, logarithmic compression, and analysis through cepstral analysis.

In addition to MFCCs and GFCCs, recent speaker recognition models also leverage spectrograms and log-Mel filterbank energies. Spectrograms offer a visual representation of the signal's frequency content over time, enabling CNNs to capture localized temporal and spectral patterns effectively. Log-Mel filterbank energies, derived directly from Mel-scaled filterbanks without the DCT transformation, preserve more raw spectral detail and have proven effective in deep learning-based pipelines, particularly for end-to-end learning systems in noisy or unconstrained environments.

Al-Karawi et al. have proved that if MFCC features are used together with traditional neural networks, the accuracy rate is around 85% in the case of speaker recognition applications [6]. In most cases, such systems fail to maintain their high performance when they are exposed to environmental noise, emotional variations of speakers, and vocal symptoms associated with health problems [1], [16]. In addition, the handcrafted features rely on models to capture complex, non-linear patterns inherent in audio signals and potentially reduce their effectiveness in more challenging identification scenarios.

B. DEEP LEARNING IN AUDIO PROCESSING

Deep learning methodologies have greatly improved the capabilities in audio processing by automatically taking out features and allowing a model to learn hierarchical representations of data directly from raw inputs [3], [5]. CNNs, in particular, have been instrumental in capturing spatial and temporal dependencies of audio signals when represented in spectrograms or other time-frequency representations [5], [14].

VGG-19 is a CNN architecture that contains 19 weight layers and has been used to perform speaker recognition tasks with great accuracy [32]. VGG-19 is able to learn complex features at different abstraction levels since it has a high depth, but its substantial number of parameters necessitates the use of substantial computational resources and memory, thus testing the limits for the case of real-time applications and deployment on devices with limited capabilities [7], [21].

Residual learning through skip connections in ResNet-152 makes it possible to train very deep networks that do not suffer from vanishing gradient issues, therefore, this architecture is relatively deep and efficient in computation making it adaptive for tough classification tasks, such as audio-based identification of a speaker [8].

More efficiently, EfficientNetV2 uses compound scaling to uniformly scale the network dimensions concerning depth, width, and resolution by a compound coefficient [7].

This has ensured that EfficientNetV2 has notably high accuracy with fewer parameters and lower computations as compared to traditional scaling methods [7].

DenseNet-121 applies densely connected convolutional layers that enhance information flow and gradient propagation throughout the network [10]. It also avoids the vanishing gradient problem by connecting each layer to every other successive layer and encourages feature reuse, which enables the model to learn better with fewer parameters [10]. However, deep learning models require an overwhelming amount of labelled data and remarkable computational resources to train most of the time, making them non-progressive in some applications [3], [5].

Time-Delayed Neural Networks (TDNNs) have also been important in the field of speaker recognition [30]. In contrast to traditional CNNs, TDNNs are designed to learn temporal dependencies in sequential data, like speech signals, through context windows over time-delayed inputs. This design allows the network to learn temporal relationships between phonetic units, and it is very effective for speaker verification and identification tasks. TDNNs have been broadly used in state-of-the-art systems such as x-vectors for extracting speaker embeddings, particularly in challenging conditions such as noisy environments or cross-channel speech. Their capacity for handling variable-length input sequences while preserving temporal resolution has made TDNNs a solid baseline for contemporary speaker recognition systems.

C. MULTIMODAL FUSION TECHNIQUES

Multimodal data fusion has emerged as a powerful strategy to enhance speaker identification systems by leveraging complementary information from diverse sources [9], [11], [17]. Fusion strategies are classified mainly into three levels:

1) DATA-LEVEL FUSION

This fusion technique merges raw data from multiple modalities at an early stage; therefore, the combined data is treated as a single input for feature extraction [11]. Data-level fusion preserves the integrity of the original data, while allowing the model to learn inherent correlations between modalities. However, it requires high accuracy in synchronizing data streams and may prove difficult if the modalities have different data formats or different sampling rates.

2) FEATURE-LEVEL FUSION

In feature-level fusion, features independently obtained from one modality are concatenated into a single feature vector [11]. This method balances between the richness of data-level fusion and simplicity of decision-level fusion. Furthermore, dimension reduction techniques, such as PCA, can be used to handle the size of computational complexity. Therefore, feature-level fusion can express the complementary information from each modality while keeping model sizes reasonable.

3) DECISION-LEVEL FUSION

This is the combination of the outputs of individual classifiers that have learned separately for other different modalities [11]. The approaches include majority voting, weighted averaging, and more advanced ensemble techniques to fuse the decisions. A benefit of decision-level fusion is robustness due to combining independent predictions; however, it may increase the complexity in the process of decision making and fail to exploit intermodal correlations.

In the context of audio-based speaker identification, there is a potential to exploit multimodal data coming from bone conduction, air conduction, and throat microphones with a possible effect on significantly improving system robustness and accuracy. Despite the potential, the applications of multimodal fusion techniques in this field are also mostly unexplored [10], [17].

D. KOLMOGOROV-ARNOLD NETWORKS (KAN) AND THEIR APPLICATIONS

The Kolmogorov-Arnold Network (KAN) relies on the Kolmogorov superposition theorem, which asserts that any continuous multivariate function could be expressed as a finite sum of continuous univariate functions and addition operations [12], [13], [15]. This theorem can be applied by KAN to represent complex functions using fewer parameters than the traditional neural network, which may reduce computations and enhance potential scalability [12], [13]. The benefits of KANs in modeling complex systems:

1) REDUCED COMPUTATIONAL COMPLEXITY

Decomposing multivariate functions into sums of univariate functions by KANs enables it to model complex relationships with very few parameters [12], [15]. That is to say, the size of such models may be drastically reduced, which in turn facilitates easier or quicker deployment and improved real-time performance.

2) TRAINABLE ACTIVATION FUNCTIONS

Unlike traditional neural networks, KANs utilize trainable activation functions [12], [13]. With this degree of freedom, an activation function that is best suited for the task at hand can be learned by the network, which would be more efficient for the accuracy and adaptability of the model.

3) SCALABILITY AND EFFICIENCY

The inherent superposition structure of KANs, since they are constructed from functions of univariate variables, are computationally efficient and scalable [12], [15], especially in high-dimensional inputs or when dealing with multivariate integration.

It has already been proved by former studies that KANs may be applied for function approximation and signal processing tasks, where the possibility for using effective models in complicated systems has been demonstrated [12], [13], [15]. Its application in audio-based speaker

identification is limited, yet it opens possibilities to research further effectiveness of KANs in this specific domain.

E. RESEARCH GAPS AND OPPORTUNITIES

From the review of previous literature, gaps and opportunities exist in the research that combines audio-based speaker identification:

1) UNDERUTILIZATION OF KAN IN AUDIO PROCESSING

The theoretical merits of KANs in computation complexities in modeling complex functions, though excellent in theory, have not thoroughly been researched and tested for application in audio-based speaker identification [12], [13], [15]. Research in this area will likely allow for more efficient models with great accuracy but fewer resources.

2) LIMITED EXPLORATION OF MULTIMODAL FUSION STRATEGIES

Much of the existing work tends to focus on data that are single-modal or use very simplistic fusion strategies [10], [11], [17]. It is quite necessary to have a deeper evaluation of the existing fusion strategies for multimodal fusion, especially when one is interested in multimodal audio inputs like bone conduction and throat microphones. It may help increase the robustness and accuracy of speaker identification systems.

3) HIGH COMPUTATIONAL COSTS OF DEEP LEARNING MODELS

Most traditional deep learning models require complicated time-consuming parameter tuning with massive computational resources, which are not practicable for all applications, especially in real-time or resource-limited environments [5], [6]. New architectures that are highly performant but have lower computational cost need to be designed.

This paper attempts to bridge such gaps in speaker identification, developing a strong and effective system with the help of an efficient KAN framework and multimodal audio data. The proposed approach is to exploit the effectiveness of the KAN structure and multimodal fusion to overcome these limitations imposed by the traditional methods that will contribute to improving the level of biometric security.

III. METHODOLOGY

This section reviews the methodological framework shown in (Figure 1), it is developed to enhance the Kolmogorov-Arnold Network (KAN) for speaker identification with multimodal audio data. The approach includes collection of datasets, feature extraction and processing, as well as pre-trained deep learning models built upon comparative analysis, which leads to optimization in identification accuracy.

A. DATASET COLLECTION

A comprehensive multimodal audio dataset was curated to capture a great deal of variability in voice characteristics using three modalities for microphones that include

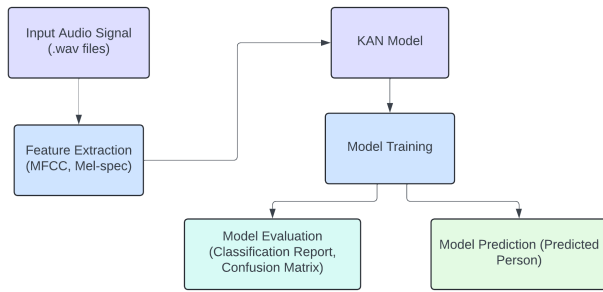


FIGURE 1. KAN model workflow for audio classification.

bone conduction, air conduction, and throat microphones. This will provide such variability that will allow robust and accurate models in speaker identification.

1) PARTICIPANTS AND RECORDING CONDITIONS

- **Participants:** The recordings contain contributions from 15 participants, all of whom have contributed several recordings under different environmental conditions. Among these, all participants are Indian, with a gender distribution of 8 male and 7 female speakers. The spoken language is English, although the speakers exhibit varying regional accents due to differences in their state of origin. The age range of the participants spans from 19 to 40 years, providing a diverse demographic profile. Overall, there are 30 recordings per speaker, which ensures enough data to capture distinctive voice characteristics and intra-speaker variability.
- **Controlled Environment:** The recordings were made in a soundproof room, so as to avoid ambient noise and provide a controlled baseline for model performance under ideal conditions. In such a controlled environment, it can be ensured that the fundamental vocal characteristics of each participant are captured without interference from any other sources of noise.
- **Variable Conditions:** To simulate the scenario better and strengthen the dataset, the recordings were done outdoor and in the regular office settings. These would introduce environmental noise, reverberations, and other acoustic variations in the settings, making it more challenging for models to adapt to multiple conditions. This will then increase the models' generalizing capabilities.

The dataset was split into 70% training, 15% validation, and 15% test sets using stratified sampling to ensure balanced representation of all speakers across the splits.

2) MICROPHONE MODALITIES

- **Bone Conduction Microphone:** This captures audio signals through vibrations transmitted via the skull bones. Bone conduction microphones are less prone to ambient noise and will capture unique vocal features associated with bone vibration patterns. This modality is particularly useful in noisy environments where air-conducted audio might be compromised [13].

- **Air Conduction Microphone:** This is the classical method of audio recording and captures sound waves in the air. Since most people use this method, it is more vulnerable to environmental noises, reverberations, and echoes that degrade the quality of captured sound and may reduce the identification accuracy [1], [16].
- **Throat Microphone:** This receives vibrations directly from the vocal cords, capturing phonation characteristics that might not be present in air-conducted recordings. This modality is valuable in picking speech under very noisy acoustic conditions because it is less impacted by ambient sound and could pick subtle nuances in voice production [13].

The microphones used include the Temco HG-17 Bone Conduction Microphone, the Rode NT-USB Mini for air conduction recordings, and the Tactic Throat Microphone for throat conduction. The combination of the above modalities gives a large and heterogeneous dataset to test multimodal fusion approaches and develop an efficient speaker identification system that could operate well in different acoustic conditions.

3) PUBLIC BENCHMARK DATASET: LIBRISPEECH

In addition to our custom multimodal recordings, we also evaluate the proposed models on the publicly available LibriSpeech corpus [31], a standard benchmark dataset widely used for speaker and speech recognition tasks. We utilize the 'train-clean-100' and 'train-clean-360' subsets, which contain approximately 100 and 360 hours of clean speech respectively, from 251 speakers in the case of train-clean-100 and 921 speakers for train-clean-360, all recorded under controlled conditions.

This dataset serves two purposes:

- **Benchmark Comparison:** It enables direct comparison with existing literature using standard benchmarks.
- **Generalizability Assessment:** It helps test whether the proposed models generalize well to clean, real-world recordings beyond our custom acquisition environment.

The LibriSpeech data were preprocessed using MFCC extraction pipeline for consistency with our in-house dataset.

B. FEATURE EXTRACTION AND PROCESSING

Effective feature extraction is crucial for accurate speaker identification. The study employs two primary acoustic features—Mel-frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC)—to capture comprehensive spectral and temporal characteristics of the audio signals. Combining these features leverages their complementary strengths to enhance model performance.

1) MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

MFCCs are widely used in speech and audio processing due to their ability to represent the spectral envelope of sounds in a manner that closely aligns with human

auditory perception [4], [16]. The extraction process involves several steps:

Extraction Process:

- **Frame Segmentation:** The continuous audio signal is segmented into overlapping frames of 25 milliseconds with a 10-millisecond overlap. This segmentation allows for the analysis of short-term stationary properties of the audio signal.
- **Windowing:** A Hamming window is applied to each frame to reduce spectral leakage and improve frequency resolution.
- **Pre-emphasis:** A pre-emphasis filter is applied to amplify the high-frequency components of the signal, compensating for the natural roll-off in the speech spectrum and enhancing discriminative features.

Feature Computation:

- **Fast Fourier Transform (FFT):** The windowed frames are transformed from the time domain to the frequency domain using FFT, providing a spectrum of frequencies present in each frame.
- **Mel Filterbank Application:** The magnitude spectrum is passed through a filterbank consisting of 40 triangular filters spaced according to the Mel scale. This scale reflects the human ear's perception of sound frequencies, with finer resolution at lower frequencies and coarser resolution at higher frequencies.
- **Logarithmic Compression:** The filtered energies are converted to the logarithmic scale to mimic the human auditory system's logarithmic perception of loudness.
- **Discrete Cosine Transform (DCT):** DCT is applied to the log filterbank energies to decorrelate the coefficients and reduce dimensionality. The first 12-13 coefficients are typically retained, as they contain the most relevant information for speaker recognition.

MFCCs effectively represent the vocal tract configurations and are robust to variations in speech content and environmental noise, making them suitable for speaker identification tasks [4], [16].

2) GAMMATONE FREQUENCY CEPSTRAL COEFFICIENTS (GFCC)

GFCCs are designed to model the human auditory system's frequency selectivity more accurately than MFCCs by using gammatone filters [6], [9]. The extraction process includes:

Extraction Process:

- **Gammatone Filterbank Application:** The audio signal is filtered through a bank of 64 gammatone filters, covering frequencies from 50 Hz to 8000 Hz. The filters are spaced according to the Equivalent Rectangular Bandwidth (ERB) scale, providing finer resolution in frequency bands critical for speech perception.
- **Temporal Integration:** The energy outputs of the gammatone filters are integrated over each frame, similar to the MFCC process, capturing the temporal dynamics of the signal.

- **Logarithmic Compression:** Logarithmic scaling is applied to the integrated energies to simulate the non-linear loudness perception of the human ear.
- **Cepstral Coefficient Computation:** A DCT is performed on the log energies to obtain the GFCCs, which provide a compact representation of the spectral information.

GFCCs are particularly effective in noisy environments due to their emphasis on auditory filter shapes that are robust to noise. They complement MFCCs by capturing additional auditory information, enhancing the model's ability to distinguish between speakers [6], [9].

3) FEATURE COMBINATION

- **Feature Fusion:** MFCC and GFCC features are concatenated to form a unified feature vector for each audio frame. This combination enriches the feature representation by leveraging the strengths of both feature sets.
- **Dimensionality Reduction:** Due to the increased dimensionality resulting from feature concatenation, Principal Component Analysis (PCA) is applied to reduce the feature vector's dimensionality while retaining the most significant variance components. This step helps manage computational complexity and mitigates the risk of overfitting.
- **Normalization:** The resulting feature vectors are standardized to have zero mean and unit variance. Normalization ensures that all features contribute equally to the model training process and improves convergence during optimization.

These steps help in improving generalization and reducing sensitivity to variations across recording conditions.

C. IMPLEMENTATION OF PRE-TRAINED DEEP LEARNING MODELS

To benchmark the performance of the efficient KAN framework, four pre-trained deep learning models are evaluated: VGG-19, ResNet-152, EfficientNetV2, and DenseNet-121. As seen in Figure 2 and 3, each model's unique architecture and feature extraction capabilities make it suitable for comparative analysis in speaker identification through audio data. The pre-trained models were directly downloaded and adapted for our dataset without any additional pretraining on external datasets. These models were originally pre-trained on the ImageNet dataset, which provides general-purpose feature representations that can be adapted for audio spectrogram inputs through architectural adjustments.

1) MODEL SELECTION AND FINE-TUNING

- **VGG-19:** VGG-19 is known for its deep architecture with 19 weight layers, utilizing small convolutional filters (3×3) and a uniform architecture throughout the network [7], [21]. The model is fine-tuned by replacing the final classification layers to match the number

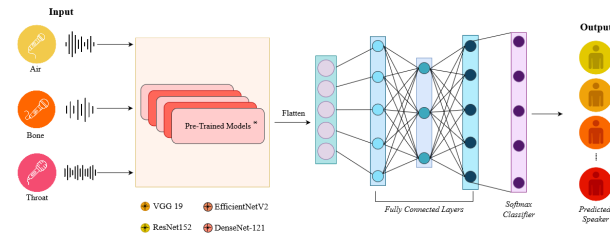


FIGURE 2. Unimodal pre-trained architecture.

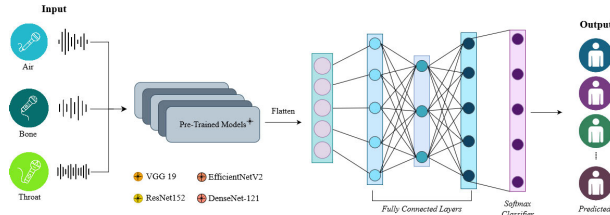


FIGURE 3. Trimodal pre-trained architecture.

of classes in our dataset and adjusting the learning rates of the earlier layers to prevent overfitting. The input convolutional layer was modified to accept single-channel MFCC/GFCC input features in place of RGB images, this was done for all the pre-trained models.

- **ResNet-152:** ResNet-152 employs residual learning with skip connections, enabling the training of very deep networks [8], [14]. The model is fine-tuned by unfreezing the later layers to learn domain-specific features while keeping earlier layers frozen to retain the pre-trained weights' general feature extraction capabilities.
- **EfficientNetV2:** EfficientNetV2 utilizes a compound scaling method to optimize the network's depth, width, and resolution, achieving high performance with fewer parameters [7]. Fine-tuning involves adjusting the scaling coefficients to balance model complexity and accuracy for our specific dataset.
- **DenseNet-121:** DenseNet-121 features densely connected layers, where each layer receives inputs from all preceding layers, promoting feature reuse and efficient gradient flow [10]. The model is adapted by modifying the classification layer and fine-tuning the network on our dataset.

2) TRAINING AND EVALUATION

Datasets: The models are trained and evaluated on three types of datasets:

- **Unimodal Datasets:** Contain audio data from a single microphone modality (air, bone, or throat).
- **Bimodal Datasets:** Combine data from two modalities (e.g., air-bone, bone-throat, throat-air) to assess the benefits of dual-modal fusion.

- **Trimodal Dataset:** Includes data from all three modalities, enabling the evaluation of comprehensive multimodal fusion.

Training Parameters:

- **Optimizer:** The Adam optimizer is utilized for its adaptive learning rate capabilities and efficient convergence.
- **Learning Rate:** An initial learning rate of 0.001 is set, with a decay schedule to reduce the rate as training progresses.
- **Batch Size:** A batch size of 32 is selected to balance memory constraints and training efficiency.
- **Epochs:** Models are trained for up to 50 epochs, with early stopping implemented based on validation loss to prevent overfitting.

Data Augmentation: To enhance generalization and prevent overfitting, data augmentation techniques are applied, including:

- **Time Shifting:** Randomly shifting the audio signal in time to simulate slight variations in speech timing.
- **Pitch Alteration:** Modifying the pitch of the audio signal within a small range to account for natural variations in voice pitch.
- **Background Noise Addition:** Adding ambient noise samples at various signal-to-noise ratios to simulate different environmental conditions.

Performance Metrics: Model performance is assessed using accuracy, precision, recall, and F1-score on both validation and test sets. Confusion matrices are also analyzed to understand class-specific performance.

3) RESULTS SUMMARY

a: UNIMODAL PERFORMANCE

ResNet-152 achieved the highest accuracy on the bone and throat modalities, indicating its effectiveness in capturing complex features inherent in these inputs. EfficientNetV2 also demonstrated robustness, particularly on the bone and throat datasets, suggesting good generalization capabilities.

TABLE 1. Unimodal pretrained models (MFCC).

Model	Air (%)		Bone (%)		Throat (%)	
	Val	Test	Val	Test	Val	Test
VGG-19	93.62	94.43	95.73	93.37	93.71	91.83
ResNet-152	93.86	92.74	98.89	98.31	98.73	98.57
EfficientNetV2	91.72	92.67	98.16	98.43	98.13	97.83
DenseNet-121	93.34	93.62	97.83	97.94	97.61	97.73

b: BIMODAL PERFORMANCE

DenseNet-121 showed strong performance in the air-bone combination, demonstrating effective feature integration. ResNet-152 maintained high accuracy across all bimodal combinations, showcasing its adaptability to diverse audio inputs.

TABLE 2. Bimodal pretrained models (MFCC).

Model	Air-Bone (%)		Bone-Throat (%)		Throat-Air (%)	
	Val	Test	Val	Test	Val	Test
VGG-19	98.18	97.13	93.70	91.82	91.36	89.27
ResNet-152	98.56	95.45	97.67	95.45	97.73	95.57
EfficientNetV2	97.63	94.64	97.71	95.52	97.76	93.91
DenseNet-121	99.47	98.38	97.63	97.73	95.65	91.64

c: TRIMODAL PERFORMANCE

In the trimodal setup, ResNet-152 achieved the highest accuracy, affirming its strength in multimodal scenarios and its ability to leverage comprehensive feature sets for improved identification.

TABLE 3. Trimodal pretrained models (MFCC).

Model	Air-Bone-Throat (%)	
	Val	Test
VGG-19	95.24	97.76
ResNet-152	99.25	98.75
EfficientNetV2	95.78	95.28
DenseNet-121	97.27	97.79

D. KOLMOGOROV-ARNOLD NETWORKS (KAN) IMPLEMENTATION

The Kolmogorov—Arnold Network (KAN) shown in Figure 4, is a computationally efficient framework with which to deal with high-dimensional representations of data and complex classification tasks. This method relies on the Kolmogorov-Arnold representation theorem-whereby complex multivariate functions are broken down into compositions of operations on univariate functions and additions. It is this feature that enables the KAN approach to model very complex functions without excessive computational overhead. In this work, KAN was implemented using a custom efficient implementation, not relying on the PyKAN library, to better optimize memory and inference efficiency on audio tasks. This makes KAN well-suited for high-dimensional analysis and real-time audio processing applications.

The core formula of KAN, enhanced by B-splines, allows us to decompose a multivariable function as follows:

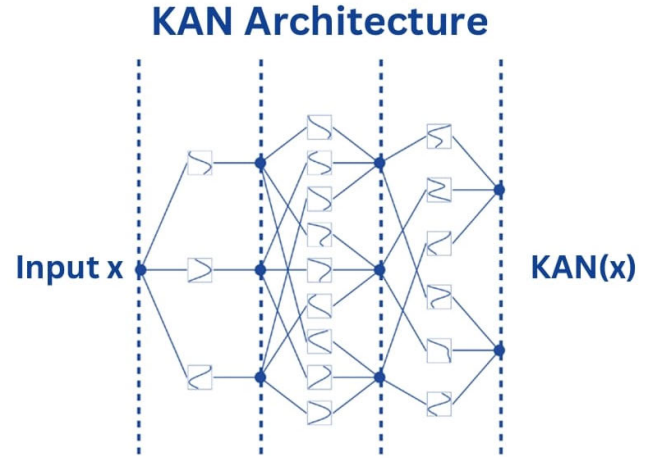
$$f(x_1, x_2, \dots, x_n) = \sum_{i=1}^{2n+1} \phi_i \left(\sum_{j=1}^n \psi_{ij}(x_j) \right) \quad (1)$$

In this decomposition:

- ϕ_i functions can be represented by B-spline basis functions to ensure smoothness in the approximation,
- ψ_{ij} functions may also be B-spline transformations on individual variables x_j , allowing each input dimension to be approximated using splines.

1) B-SPLINE BASIS FUNCTIONS

A B-spline basis function $B_{k,d}(x)$ of degree d is defined recursively. B-splines are defined over a set of control

**FIGURE 4.** KAN Architecture.

points (knots) t , and for each point x , the B-spline is calculated as:

$$B_{i,d}(x) = \frac{x - t_i}{t_{i+d} - t_i} B_{i,d-1}(x) + \frac{t_{i+d+1} - x}{t_{i+d+1} - t_{i+1}} B_{i+1,d-1}(x) \quad (2)$$

where

$$B_{i,0}(x) = \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

By using B-spline basis functions in the hidden layer nodes of KAN, each component function ψ_{ij} or ϕ_i can be made more flexible and smoother.

2) HIDDEN LAYER TRANSFORMATION WITH B-SPLINES

In KAN with B-splines, each hidden neuron transformation is represented by:

$$h_i = \phi_i \left(\sum_{j=1}^n B_{ij,d}(x_j) \right) \quad (3)$$

where $B_{ij,d}$ is a B-spline basis function of degree d , applied to each input variable x_j . This transformation allows for a smooth, flexible representation of each hidden layer's contribution to the overall network.

3) OUTPUT LAYER WITH SUMMATION OF B-SPLINE TRANSFORMATIONS

Finally, the output of the KAN is given by summing the B-spline transformations from each hidden layer node:

$$y = \sum_{i=1}^{2n+1} h_i = \sum_{i=1}^{2n+1} \phi_i \left(\sum_{j=1}^n B_{ij,d}(x_j) \right) \quad (4)$$

This output layer formula allows for the flexible combination of all single-variable approximations, ensuring that KAN can approximate the original function f with high accuracy.

Many data augmentation approaches are used to enhance the robustness and adaptability of the KAN model for different types of input data. These include noise injection, which involves the introduction of random noise into data for generalization; time stretching, modification of the time characteristics of signals without changing pitch and is often applied to audio signals; pitch shifting, alteration in the pitch of input data without time period change; and random cropping, variability through selecting portions of the data. These augmentation techniques strengthen KAN's generalization capability across different input variations, hence making it much more robust and adaptable for general applications in real life. Batch normalization and dropout improve training stability and prevent overfitting by normalizing activations and deactivating random neurons during training.

KAN's trainable activation functions learn optimal non-linear transformations to capture complex data patterns dynamically. This flexibility ensures that the model adapts without requiring predefined mappings. Feature extraction is performed with Mel-frequency cepstral coefficients (MFCC), Gammatone frequency cepstral coefficients (GFCC), and a combined MFCC+GFCC approach, enhancing the model's ability to leverage both unimodal and bimodal datasets efficiently.

KAN employs the Adam optimizer with an initial learning rate of 0.001 for efficient convergence, paired with cross-entropy loss for stable classification task performance. Evaluation metrics like accuracy, precision, recall, and F1-score demonstrate KAN's effectiveness. With MFCC-based data-level fusion, KAN achieves an accuracy of around 99.375%, surpassing traditional deep learning architectures. Lower parameter requirements further improve KAN's computational efficiency.

Notably, MFCC features consistently yielded higher test accuracy than MFCC+GFCC combinations, likely due to better generalization and reduced feature redundancy in high-dimensional space. Tables 4 and 5 present KAN's unimodal and bimodal feature extraction performance.

TABLE 4. Unimodal feature extraction (KAN).

Feature	Air (%)		Bone (%)		Throat (%)	
	Val	Test	Val	Test	Val	Test
MFCC	93.49	95.74	98.89	98.45	96.18	98.64
GFCC	89.51	93.62	98.56	95.18	92.09	93.91
MFCC+GFCC	87.72	93.62	98.17	96.57	93.73	94.18

E. MULTIMODAL DATA FUSION MECHANISMS

1) OVERVIEW OF FUSION MECHANISMS

Data fusion at the data, feature, and decision levels facilitates the effective integration of information from multiple audio modalities. Each approach operates at a distinct stage of the pipeline, allowing models to leverage unique advantages and address challenges specific to speaker identification.

TABLE 5. Bimodal feature extraction (KAN).

Feature	Air-Bone (%)		Bone-Throat (%)		Throat-Air (%)	
	Val	Test	Val	Test	Val	Test
MFCC	98.89	99.03	98.46	99.11	98.64	98.84
GFCC	97.76	97.64	96.39	97.73	91.36	95.45
MFCC+GFCC	98.67	96.18	95.35	95.45	93.64	92.36

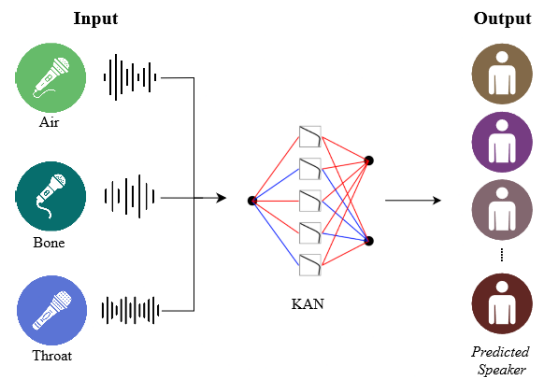


FIGURE 5. Data level fusion model.

2) DATA-LEVEL FUSION

As shown in Figure 5, combining modalities at the early stages, or at the data level, involves directly merging raw audio data, requires the model to learn an integrated representation. By preserving temporal correspondence across modalities—such as audio streams from Bone, Air, and Throat microphones—synchronized audio signals are concatenated to form a single input sequence for generating MFCC and GFCC features.

a: ADVANTAGES

Data-level fusion captures complementary information from the start, avoiding the need for separate feature extraction for each modality.

b: CHALLENGES:

It requires maintaining accurate temporal alignment and handling the high dimensionality of combined raw data, which can be computationally intensive.

3) FEATURE-LEVEL FUSION

Feature-level fusion shown in Figure 6, combines features extracted from each modality into a single vector, enabling mid-level integration. After individual feature extraction, the features are combined, and PCA is applied for dimensionality reduction and normalization.

a: ADVANTAGES

This fusion method selectively integrates meaningful features, improving computational efficiency.

b: CHALLENGES

Effective feature engineering is crucial to ensure proper integration across modalities.

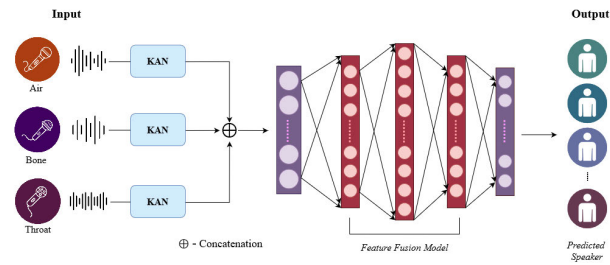


FIGURE 6. Feature level fusion model.

4) DECISION-LEVEL FUSION

Decision-level fusion shown in Figure 7, aggregates outputs from individual classifiers developed for each modality (Bone, Air, and Throat) as a late-stage integration method. By applying majority voting or weighted averaging, classifier outputs are combined to produce the final decision.

a: ADVANTAGES

This modular design enhances robustness, as classifiers can be trained independently and fine-tuned for resilience against specific modality failures.

b: CHALLENGES

Managing multiple classifiers increases complexity and may duplicate computations, potentially increasing processing time.

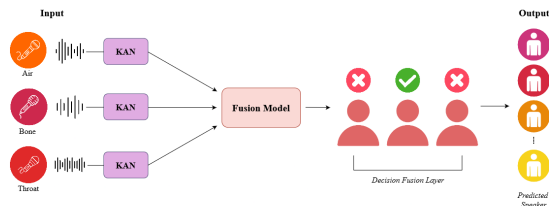


FIGURE 7. Decision Level Fusion Model.

5) EXPERIMENTAL RESULTS

Each fusion strategy yields distinct results. The following summary outlines the strengths and weaknesses of each fusion technique in processing multimodal audio data for speaker identification:

- **Data-Level Fusion:** This strategy achieved the highest accuracy by aggregating Bone, Air, and Throat audio signals prior to feature extraction, allowing it to capture complementary information that enhances speaker identification performance.
- **Feature-Level Fusion:** Although this approach demonstrated moderate performance, there is potential for optimization. The selective integration of MFCC and GFCC features is promising, though fine-tuning feature selection and fusion policies could further

improve results. Properly designed feature combinations are essential for maximizing benefits.

- **Decision-Level Fusion:** While robust, decision-level fusion showed a slight drop in accuracy compared to data-level fusion. Classifier output aggregation improved resilience to noise and variability but resulted in a minor precision trade-off.

These findings indicate that each fusion mechanism offers specific benefits for certain performance metrics, underscoring the importance of a balanced multimodal audio processing system design for speaker identification. In this study, the final speaker identification performance of the KAN model was evaluated primarily using data-level fusion, as it consistently outperformed the other strategies across all feature settings.

TABLE 6. Trimodal feature extraction with different fusion techniques(KAN).

Fusion Mechanism	Test Accuracy (%)		
	MFCC	GFCC	MFCC+GFCC
Data Level	99.375	98.458	98.57
Feature Level	95.88	84.27	95.31
Decision Level	96.31	93.19	94.27

IV. EXPERIMENTAL SETUP

A. HARDWARE CONFIGURATION

A robust hardware configuration has been selected to afford efficient model training and processing. The multimodal dataset experiments were conducted on Kaggle, while the LibriSpeech dataset experiments were performed on a specialized system. The experimental setup included an Intel Xeon Platinum multi-core processor (48 Cores), equipped with 2x NVIDIA H100 Tensor Core GPUs with 80 GB HBM2e memory each. The system featured 2 TB of system memory (32 DIMMs) for handling memory-intensive operations, and storage consisting of 2 × 1.92 TB NVMe M.2 SSDs in RAID 1 for the operating system. The environment ran on Ubuntu Linux 20.04 LTS with NVIDIA Driver 535.230.02 and CUDA 12.2, selected for their reliability and compatibility with deep learning frameworks.

B. SOFTWARE ENVIRONMENT

The development environment includes Python 3.8, a flexible language with extensive libraries for data processing and machine learning. The main framework adopted is PyTorch 1.10, due to its flexibility in neural network research. Supporting libraries include NumPy 1.21 for numerical computations, SciPy 1.7 for scientific operations, and scikit-learn 0.24 for machine learning functions.

C. AUDIO PROCESSING AND DATA HANDLING

Audio processing was performed with librosa 0.8, which enabled feature extraction, while data manipulation was handled by pandas 1.3, facilitating structured data handling. Data visualization and analysis were conducted using matplotlib

and seaborn, providing insights into model performance and data characteristics. LibriSpeech audio files were processed using the same feature extraction pipeline, and incorporated into the evaluation framework using consistent sampling rates and feature normalization techniques. This ensures fair benchmarking and highlights the adaptability of the models across datasets.

D. EVALUATION METRICS

Model performance was evaluated using accuracy, precision, recall, and F1-score to comprehensively assess classification performance. These metrics ensure fair evaluation across imbalanced class distributions and highlight the model's ability to generalize across diverse speakers.

1) CONFUSION MATRIX

The confusion matrix provided a clear visualization of true versus predicted classifications, offering an overview of performance across all classes. This metric highlights areas where the model excels or requires improvement, assisting in accuracy refinement.

2) COMPUTATIONAL EFFICIENCY METRICS

The computational efficiency metrics included total trainable parameters, time to convergence, and average prediction time per sample. These metrics helped balance accuracy with computational resource usage, supporting effective model deployment.

3) K-FOLD CROSS-VALIDATION

To address overfitting, a 5-fold cross-validation strategy was applied, splitting the data into five subsets for model training across different subset permutations. This approach provides a more robust generalization ability estimate and reduces dependency on a single train-test split.

V. RESULTS AND DISCUSSION

A. PERFORMANCE COMPARISON

The models were rated on accuracy, precision, recall, and F1-score metrics, and then categorized into unimodal, bimodal, and trimodal analyses for an overall rating. The metrics were computed on the validation and test splits, using stratified sampling to ensure class distribution consistency. This ensured comprehensive evaluation across class boundaries and helped detect imbalances and misclassifications. In the unimodal analysis, the performance of the models was tested over the different audio modalities: Air, Bone, and Throat. The accuracy results are given in Table 1. The Efficient KAN model surpasses all of the pre-trained deep learning models in each of the unimodal datasets and achieved the best accuracy on all of the three modalities: Air, Bone, and Throat. ResNet-152 demonstrated strong performance, especially in the Bone and Throat modalities, outperforming VGG-19, EfficientNetV2, and DenseNet-121 in these cases; however, DenseNet-121 produced the poorest

pre-training results, suggesting that it may not acquire speaker-specific representations as effectively.

TABLE 7. Unimodal pre-trained models performance (MFCC).

Model	Air (%)	Bone (%)	Throat (%)
VGG-19	94.43	93.37	91.83
ResNet-152	92.74	98.31	98.57
EfficientNetV2	92.67	98.43	97.83
DenseNet-121	93.62	97.94	97.73
Efficient KAN	95.74	98.45	98.64

In the Bimodal test, two audio modalities were considered to evaluate the ability of the models to learn from complementary information. The accuracy of these bimodal configurations is presented in Table 2. Here, Efficient KAN presented outstanding performance with near-perfect accuracy in all bimodal combinations. ResNet-152 also produced good performance in most of the bimodal combinations especially Air-Bone and Bone-Throat combinations. On the other hand, DenseNet-121 showed relatively low accuracy in the combination of Bone-Throat and Throat-Air. This indicates that it lacks to merge the data of different modalities effectively. In these instances, it presents the possibility of Efficient KAN models for complementary exploitation of information from different modalities.

TABLE 8. Bimodal pre-trained models performance (MFCC).

Model	Air-Bone (%)	Bone-Throat (%)	Throat-Air (%)
VGG-19	97.13	91.82	89.27
ResNet-152	95.45	95.45	95.57
EfficientNetV2	94.64	95.52	93.91
DenseNet-121	98.38	97.73	91.64
Efficient KAN	99.03	99.11	98.84

Under trimodal analysis, the models were tested based on their ability to integrate all three audio modalities at the same time. Table 3 summarizes the accuracy results for that particular trimodal setup. In the trimodal configuration, the Efficient KAN model achieved an impressive accuracy of 99.375%, which performs better than any of the pre-trained models. ResNet-152 turned out to be the best-performing pre-trained model of the trimodal setup, but still lagged behind Efficient KAN by about 10 %. Everywhere the pre-trained models demonstrated systematically improved performance when scaled from unimodal to trimodal configurations, but no one was able to reach the high level of performance of Efficient KAN. These results, as presented in Tables 1, 2, and 3, clearly reflect the outstanding capability of the Efficient KAN model to exploit multimodal data for better performance in classification tasks.

To assess generalization, the models were also tested on the LibriSpeech “train-clean-100” and “train-clean-360” subsets. The Efficient KAN model showed great results on these benchmarks.

These results shown in Table 10 confirm that the KAN framework generalizes well even on external datasets not

TABLE 9. Trimodal pre-trained models performance (MFCC).

Model	Trimodal Accuracy (%)
VGG-19	97.76
ResNet-152	98.75
EfficientNetV2	95.28
DenseNet-121	97.79
Efficient KAN	99.375

TABLE 10. LibriSpeech dataset on KAN.

Dataset	Val Accuracy (%)	Test Accuracy (%)
train-clean-100	96.57	95.87
train-clean-360	95.71	95.28

included in its training phase, such as LibriSpeech, and reinforces its robustness in speaker identification tasks.

B. COMPUTATIONAL COMPLEXITY

Computational complexity was assessed in addition to the classification accuracy to measure the efficiency of each model: number of trainable parameters, training time, and inference time. As shown in Table 4, the Efficient KAN model required several orders of magnitude fewer parameters, at 0.9 million, compared to the pre-trained deep learning models that have between 8.0 million (DenseNet-121) and 144 million (VGG-19) parameters. This decrease in the number of parameters improved the training and inference times immensely. For example, Efficient KAN trained in 2 hours while all the models trained the pre-learned models took 4 to 6 hours. At the inference level, the Efficient KAN model processed samples at a maximum rate of 1.5 ms, while DenseNet-121 at 2.1 ms and VGG-19 at 5.2 ms.

TABLE 11. Computational complexity of models.

Model	Parameters (Mil)	Training (hrs)	Inference (ms)
VGG-19	144	6	5.2
ResNet-152	60	5.8	4.5
EfficientNetV2	66	5	3.2
DenseNet-121	8.0	4.5	2.1
Efficient KAN	0.9	2	1.5

C. ADVANTAGES OF KAN

The efficient Kolmogorov-Arnold Network, KAN, shows quite an obvious supremacy from the performance and efficiency of the traditional deep learning models. From Tables 1, 2, and 3, one can see how KAN performs better than pre-trained models in unimodal, bimodal, and trimodal, with up to a 99.375% accuracy in the trimodal configuration. Such a high accuracy rate suggests the capability of KAN to successfully catch very complex audio features, which proves highly useful for speaker identification purposes, among other applications.

The architecture of KAN consists of less than 0.9 million parameters, making it highly efficient in terms of computational resources. Tables 4, 5, 6 and 10 depict the

data for KAN. Its compact size takes less time for both training and inference in contrast to the larger models such as VGG-19. This enables direct real-time application that insists on rapid processing. Its efficient design also allows for better scalability toward handling larger datasets and other modalities without having a significant increase in resource requirements.

KAN also addresses overfitting by using dropout and batch normalization techniques, which together enhance its generalization performance on test data. KAN's trainable activation functions, learned from data rather than being fixed like ReLU or tanh, adapt more effectively to varying signal characteristics in speaker recognition. Its architecture is grounded in well-established mathematical principles, making it more straightforward to implement and maintain compared to complex models like ResNet or DenseNet. It retains high performance and adaptability while being compatible with complex, nonlinear data relationships.

D. LIMITATIONS AND CHALLENGES

However, despite several challenges and limitations which emerged during the study, there are strengths of the Efficient KAN model. Data alignment for the fusion mechanism still calls for the actual alignment of multiple audio streams to be effective for Data Level Fusion. Misalignment may lead to suboptimal outcomes, especially under dynamic recording settings, thereby demanding proper synchronization methods. KAN's performance is heavily dependent on the quality of the features that have been extracted, such as MFCC and GFCC; therefore, weak feature engineering can be detrimental to the model, making it require more accurate feature extraction.

KAN uses flexible and trainable activation functions; however, one might think that having more choices might make its performance even better, which would point to future research that follows other functions to enhance the potential of KAN. Other problems, such as issues of generalizability, are also encountered through the use of audio-based speaker identification: for wide applicability across biometric modalities like facial or fingerprint recognition, testing beyond that was conducted and possibly structural changes to KAN may be in order. While KANs have demonstrated scalability and efficiency in handling large and high-dimensional datasets, their performance on extremely large or complex datasets may necessitate careful tuning and optimization to maintain efficiency and accuracy.

KANs offer improved interpretability compared to most traditional neural network models, though challenges remain. Unlike conventional "black box" neural networks, KANs leverage interpretable univariate spline functions and a structure based on the Kolmogorov-Arnold representation theorem, allowing for better visualization and analysis of computational components. However, as KANs scale to handle complex tasks, fully understanding their decision-making processes can still be difficult. In subsequent research, further enhancing the interpretability of KANs to make

the decision-making process even more transparent and explainable remains important for their broader adoption and trustworthiness.

VI. CONCLUSION

This study introduces an efficient Kolmogorov-Arnold Network (KAN) framework designed for speaker identification utilizing multimodal audio data. By integrating inputs from bone conduction, air conduction, and throat microphones, the proposed KAN model effectively captures a wide range of speech characteristics, resulting in high identification accuracy. The experimental results indicate that the efficient KAN significantly surpasses leading deep learning models, including VGG-19, ResNet-152, EfficientNetV2, and DenseNet-121, achieving an outstanding accuracy rate of 99.375% in trimodal configurations. To demonstrate the scalability and generalizability of the proposed framework, we further evaluated the KAN model on the publicly available LibriSpeech dataset, achieving accuracy exceeding 95%, which confirms its robustness even on large and diverse datasets.

In addition, the KAN model demonstrates impressive computational efficiency, utilizing only 0.9 million parameters in contrast to the numerous parameters required by traditional models, and achieving quicker training and inference times. The benefits of the KAN framework extend beyond accuracy and efficiency; its reduced parameter complexity and trainable activation functions enhance scalability and adaptability, making it well-suited for deployment in resource-constrained environments. Nonetheless, this research also highlights several limitations, including challenges in data synchronization during fusion, reliance on high-quality feature extraction, and limited exploration of activation functions. Furthermore, while the KAN model performs exceptionally in audio-based identification, its applicability to other biometric modalities warrants further investigation.

Future work should address these limitations by improving synchronization techniques for multimodal data, exploring a wider range of activation functions to further enhance model performance, and extending the KAN framework to other biometric applications such as facial and fingerprint recognition. Additionally, testing the model on larger and more diverse datasets will help validate its scalability and generalization capabilities. Enhancing the interpretability of the KAN model is also essential for gaining deeper insights into its decision-making processes.

In conclusion, the efficient KAN framework offers a promising alternative to traditional deep learning models for audio-based speaker identification, combining high accuracy with computational efficiency. The superior performance and computational efficiency of the KAN framework stem from its foundational ability to learn activation functions as splines, allowing it to more accurately model the complex, non-linear relationships in multimodal audio data compared to traditional deep learning architectures like CNNs that rely on fixed activation functions and extensive parameterization.

This research contributes to the advancement of biometric security systems by offering a more scalable, interpretable, and computationally efficient solution for reliable speaker identification.

Data Availability: Datasets used in the current study are available from the corresponding author on request.

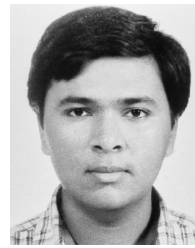
ACKNOWLEDGMENT

The authors gratefully acknowledge the support of Vellore Institute of Technology and DST-FIST for providing the resources necessary to conduct this research.

REFERENCES

- [1] K. Chavan and U. Gawande, "Speech recognition in noisy environment, issues and challenges: A review," in *Proc. Int. Conf. Soft-Comput. Netw. Secur. (ICSNS)*, Feb. 2015, pp. 1–5, doi: [10.1109/ICSNS.2015.7292420](https://doi.org/10.1109/ICSNS.2015.7292420).
- [2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, Jan. 2010, doi: [10.1016/j.specom.2009.08.009](https://doi.org/10.1016/j.specom.2009.08.009).
- [3] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, Aug. 2021, doi: [10.1016/j.neunet.2021.03.004](https://doi.org/10.1016/j.neunet.2021.03.004).
- [4] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012, doi: [10.1109/TASL.2011.2172422](https://doi.org/10.1109/TASL.2011.2172422).
- [5] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. Felipe Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," 2017, *arXiv:1705.09792*.
- [6] K. A. Al-Karawi, "Mitigate the reverberation effect on the speaker verification performance using different methods," *Int. J. Speech Technol.*, vol. 24, no. 1, pp. 143–153, Mar. 2021, doi: [10.1007/s10772-020-09780-1](https://doi.org/10.1007/s10772-020-09780-1).
- [7] A. Koumparoulis and G. Potamianos, "Accurate and resource-efficient lipreading with Efficientnetv2 and transformers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 8467–8471, doi: [10.1109/ICASSP43922.2022.9747729](https://doi.org/10.1109/ICASSP43922.2022.9747729).
- [8] C. Liang, Y. Chen, J. Yao, and X.-L. Zhang, "Multi-channel far-field speaker verification with large-scale ad-hoc microphone arrays," 2021, *arXiv:2110.05975*.
- [9] X. Song, H. Chen, Q. Wang, Y. Chen, M. Tian, and H. Tang, "A review of audio-visual fusion with machine learning," *J. Phys., Conf. Ser.*, vol. 1237, no. 2, Jun. 2019, Art. no. 022144, doi: [10.1088/1742-6596/1237/2/022144](https://doi.org/10.1088/1742-6596/1237/2/022144).
- [10] N. Bala, R. Gupta, and A. Kumar, "Multimodal biometric system based on fusion techniques: A review," *Inf. Secur. J., Global Perspective*, vol. 31, no. 3, pp. 289–337, 2022, doi: [10.1080/19393555.2021.1974130](https://doi.org/10.1080/19393555.2021.1974130).
- [11] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Inf. Fusion*, vol. 52, pp. 187–205, Dec. 2019, doi: [10.1016/j.inffus.2018.12.003](https://doi.org/10.1016/j.inffus.2018.12.003).
- [12] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "KAN: Kolmogorov-Arnold networks," 2024, *arXiv:2404.19756*.
- [13] D. A. Sprecher, "A universal mapping for Kolmogorov's superposition theorem," *Neural Netw.*, vol. 6, no. 8, pp. 1089–1094, Jan. 1993, doi: [10.1016/s0893-6080\(09\)80020-8](https://doi.org/10.1016/s0893-6080(09)80020-8).
- [14] A.-A. Tulbure, A.-A. Tulbure, and E.-H. Dulf, "A review on modern defect detection models using DCNNs – deep convolutional neural networks," *J. Adv. Res.*, vol. 35, pp. 33–48, Jan. 2022, doi: [10.1016/j.jare.2021.03.015](https://doi.org/10.1016/j.jare.2021.03.015).
- [15] J. Schmidt-Hieber, "The Kolmogorov-Arnold representation theorem revisited," *Neural Netw.*, vol. 137, pp. 119–126, May 2021, doi: [10.1016/j.neunet.2021.01.020](https://doi.org/10.1016/j.neunet.2021.01.020).
- [16] M. Al-Qaderi, E. Lahamer, and A. Rad, "A two-level speaker identification system via fusion of heterogeneous classifiers and complementary feature cooperation," *Sensors*, vol. 21, no. 15, p. 5097, Jul. 2021, doi: [10.3390/s21155097](https://doi.org/10.3390/s21155097).
- [17] T. Song, L. T. H. Nguyen, and T. V. Ta, "MPSA-DenseNet: A novel deep learning model for English accent classification," *Comput. Speech Lang.*, vol. 89, Jan. 2025, Art. no. 101676, doi: [10.1016/j.csl.2024.101676](https://doi.org/10.1016/j.csl.2024.101676).
- [18] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79236–79263, 2021, doi: [10.1109/ACCESS.2021.3084299](https://doi.org/10.1109/ACCESS.2021.3084299).

- [19] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection. Fundamentals and speech recognition system robustness," *Robust Speech Recognit. Understand.*, vol. 6, no. 9, pp. 1–22, 2007, doi: [10.5772/4740](#).
- [20] Z. Wu and Z. Cao, "Improved MFCC-based feature for robust speaker identification," *Tsinghua Sci. Technol.*, vol. 10, no. 2, pp. 158–161, Apr. 2005, doi: [10.1016/S1007-0214\(05\)70048-1](#).
- [21] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: [10.1109/ACCESS.2019.2936124](#).
- [22] C. Medjahed, A. Rahmoun, C. Charrier, and F. Mezzoudj, "A deep learning-based multimodal biometric system using score fusion," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 11, no. 1, p. 65, Mar. 2022, doi: [10.11591/ijai.v11.i1.pp65-80](#).
- [23] H. E. Çetingül, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker/speech recognition using lip motion, lip texture and audio," *Signal Process.*, vol. 86, no. 12, pp. 3549–3558, Dec. 2006, doi: [10.1016/j.sigpro.2006.02.045](#).
- [24] S. Meudt, M. Glodek, M. Schels, and F. Schwenker, "Multi-view video based tracking and audio-visual identification of persons in a human-computer-interaction scenario," in *Proc. IEEE Int. Conf. Cybern. (CYBCO)*, Jun. 2013, pp. 116–121, doi: [10.1109/CYBCONF.2013.6617454](#).
- [25] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on deep neural networks in speech and vision systems," *Neurocomputing*, vol. 417, pp. 302–321, Dec. 2020, doi: [10.1016/j.neucom.2020.07.053](#).
- [26] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7204–7208, doi: [10.1109/ICASSP.2013.6639061](#).
- [27] A. Mittal and M. Dua, "Automatic speaker verification systems and spoof detection techniques: Review and analysis," *Int. J. Speech Technol.*, vol. 25, no. 1, pp. 105–134, Mar. 2022, doi: [10.1007/s10772-021-09876-2](#).
- [28] M. Dua and S. Dua, "Noise robust automatic speech recognition: Review and analysis," *Int. J. Speech Technol.*, vol. 26, no. 2, pp. 475–519, Jul. 2023, doi: [10.1007/s10772-023-10033-0](#).
- [29] D. Liu, Z. Wang, L. Wang, and L. Chen, "Multi-modal fusion emotion recognition method of speech expression based on deep learning," *Frontiers Neurobotics*, vol. 15, Jul. 2021, Art. no. 697634, doi: [10.3389/fnbot.2021.697634](#).
- [30] X. A. Ji, T. G. Molnár, S. S. Avedisov, and G. Orosz, "Learning the dynamics of time delay systems with trainable delays," in *Proc. 3rd Conf. Learn. Dyn. Control*, vol. 144, 2021, pp. 930–942.
- [31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210, doi: [10.1109/ICASSP.2015.7178964](#).
- [32] M. Sharif-Noughabi, S. M. Razavi, and S. Mohamadzadeh, "Improving the performance of speaker recognition system using optimized VGG convolutional neural network and data augmentation," *Int. J. Eng.*, vol. 38, no. 10, pp. 2414–2425, 2025, doi: [10.5829/ije.2025.38.10a.17](#).



DHRUV SHANTU KARMOKAR is currently pursuing the bachelor's degree in artificial intelligence and machine learning with Vellore Institute of Technology (VIT), Chennai.

He has a strong interest in the diverse fields of deep learning, data science, generative AI, computer vision, speech recognition, and cybersecurity. He aspires to leverage his technical expertise to solve complex challenges and foster innovation in the technology sector. He is particularly motivated to contribute to organizations at the forefront of deploying intelligent systems to enhance processes and generate practical value.



RETHI KOMAL is currently pursuing the bachelor's degree in computer science with a focus on artificial intelligence and robotics with Vellore Institute of Technology.

He has contributed to research in audio-based speaker identification and deep learning architectures. He previously worked as a Data Science Intern with Tata Consultancy Services (TCS) and continues to pursue his interests in data science, machine learning, and multimodal signal processing through academic projects.



K. KHADAR NAWAS received the B.Tech. degree from Anna University, in 2006, and the M.Tech. degree in computer science from B. S. Abdur Rahman University, in 2011. He is currently pursuing the Ph.D. degree in speaker recognition with Vellore Institute of Technology (VIT), Chennai, India.

He has been an Assistant Professor with the School of Computing Science and Engineering, VIT, since 2012. He has 12 years of experience in academics and research. Previously, he was a Junior Research Fellow with the Christian Medical College Vellore. His research interests include speaker recognition, deep learning, and multi-modal recognition.



A. SHAHINA received the M.Tech. degree from IIT Madras, in 1998, and the Ph.D. degree from the Department of Computer Science and Engineering, IIT Madras, India, in 2008.

She is currently a Professor with the Department of Information Technology, SSN College of Engineering, Chennai, India, with 23 years of teaching and research experience. She has authored more than 55 research publications, including refereed international journals and conferences. Her

research interests include machine learning, deep learning, and speech processing.



ARYAMAN TAMOTIA is currently pursuing the bachelor's degree in artificial intelligence and machine learning with Vellore Institute of Technology (VIT), India.

His work bridges predictive modeling, deep learning, and AI-driven innovation, with research experience at IIT Kharagpur, on surface roughness prediction, and at Samsung, where he developed advanced models for sound wave direction-of-arrival. At Deloitte, he focuses on cutting-edge applications of generative AI and agentic AI, leveraging these technologies to design transformative solutions for complex real-world challenges. His expertise spans deep learning, speech and audio processing, computer vision, and the integration of generative and agentic AI for impactful, forward-looking applications.



A. NAYEEMULLA KHAN received the Ph.D. degree in computer science and engineering from IIT Madras, in 2006.

He is currently a Professor with the School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India. He has 17 years of industry experience and 11 years in teaching. He previously served as the Senior Manager (ATC) of the Airports Authority of India and later a Research Scientist with Acusis, India, an MNC where he led speech recognition efforts. His research interests include speech recognition, machine learning, deep learning, and brain-computer interfaces.

...