
NNDL Spring 2025 Project Proposal

Chaudhary, Abhishek
ac5003@columbia.edu

Park, Sungjoon
sp4050@columbia.edu

1 Introduction

The Cocktail Party Problem in Machine Learning refers to the task of separating individual source signals from a mixture of signals. The central challenge of this task involves focusing on and isolating a single voice (output) from the mixed signal of all the voices (input), when the mixing process – how each voice contributes to the overall mixture – is unknown. Mathematically, we can express this problem as:

$$\mathbf{x}(t) = A\mathbf{s}(t)$$

Where:

- $\mathbf{x}(t) \in \mathbb{R}^n$: (observed) mixture signal at time t
- $\mathbf{s}(t) \in \mathbb{R}^m$: (unknown) vectorized source signals at time t
- $A \in \mathbb{R}^{n \times m}$: (unknown) mixing matrix

The cocktail party problem has direct, real-world applications in areas such as speech recognition, automatic transcription, and music source separation. More broadly, the ability to isolate meaningful signals from background noise is valuable in fields such as climate science, telecommunications, and financial markets. As a result, there has been continued progress toward developing more effective models to address this challenge.

2 Related Work

1. Previous attempts to solve the Cocktail Party Problem

A 2018 review by Wang and Chen on supervised speech separation highlights MLPs as the most widely used early neural network model for the task, but notes that these approaches remained largely ineffective, even with deeper architectures. Subsequent improvements in model architecture, such as CNNs, RNNs, LSTMs, and GANs offered better performance.

RNNs with LSTM offer increased performance at capturing temporal dynamics and generalize well across speakers. Weninger et al. 2015, for instance, showed that RNNs with LSTM outperform feedforward DNNs in modeling long-term context in speech, especially in noisy conditions.

On the other hand, Hershey et al. 2016 introduced Deep Clustering, a novel framework for speaker-independent speech separation. In their approach, each time-frequency unit is mapped by a deep neural network to a high-dimensional embedding space, where units associated with the same speaker are positioned closer together. This method enables clustering-based separation without requiring speaker identity during training or inference, making it scalable and generalizable to unseen speakers.

2. The Kolmogorov-Arnold network (KAN)

The theoretical basis of the KAN is the Kolmogorov-Arnold representation theorem, which states that **any multivariate continuous function** can be represented as a **finite sum of compositions of univariate continuous functions**, where addition is the only multivariate operation allowed. (SOURCE 1)

KAN was first proposed by Liu et al. 2024 as an alternative to the multilayer perceptron (MLP), offering increased interpretability, accuracy, and parameter efficiency. The method trades linear weights for learnable univariate splines, opting for simple sum inputs rather than nonlinear activation functions such as ReLU. A visual example of KANs in action can be seen in Figure 1:

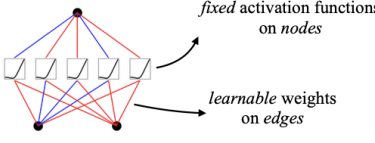
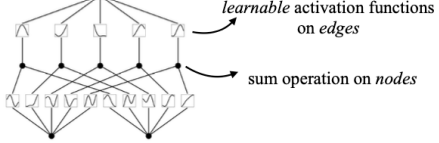
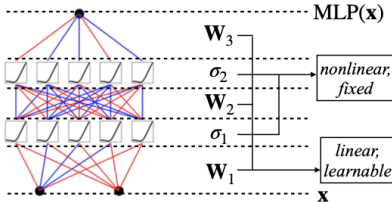
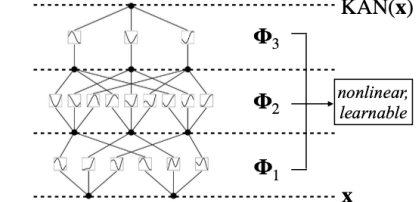
Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(\epsilon)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$
Model (Shallow)	(a) 	(b) 
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$
Model (Deep)	(c) 	(d) 

Figure 1: Multi-Layer Perceptrons (MLPs) vs. Kolmogorov-Arnold Networks (KANs)

Why it is relevant:

- The K-A representation theorem allows complex multivariate functions to be represented as sums of univariate functions. This compositional view aligns with the task at hand, where audio signals are mixed together through weighted addition.
- KANs allow inspection and even symbolic extraction of activation functions, which improves interpretability. In our task, we hope to be able to symbolify the internal structures of the voice recordings, which may correspond to distinct audio source components.

3 Method/Algorithm

Our research will apply Kolmogorov-Arnold network (KANs) to the blind source separation, specifically targeting two-speaker environments. KANs offer a promising alternative to traditional deep learning approaches due to their ability to leverage compositional functions for complex signal processing tasks.

3.1 Proposed Model Architecture

Our proposed model architecture is an Encoder-Latent-Decoder where each component is constituted via a KAN. This architecture leverages existing PyTorch implementations of KAN blocks. The encoder transforms the mixed audio signal into latent representations. The latent space is structured to effectively disentangle speaker-specific features. The decoder will then reconstruct the individual speech signals from the separated latent space representations.

3.2 Training Methodology

Our approach will employ supervised training with audio tracks serving as both inputs and outputs. To ensure robustness and expand our training dataset, we implement two complementary training procedures:

3.2.1 Single-Speaker Training

We will utilize isolated audio tracks containing single speakers as clean signals. The model is trained to reproduce these clean signals as outputs. While this approach is common in encoder-decoder architectures for data compression, we recognize a significant challenge: the model could potentially learn to pass inputs directly to outputs without meaningful processing or separation capability.

3.2.2 Mixed-Signal Training

To address the limitation above, we incorporate a second training procedure that directly addresses the cocktail party problem. Here, inputs consist of overlaid clean signals from two speakers, with individual speaker tracks serving as ground truth. This forces the model to develop genuine separation capabilities.

3.2.3 Loss Function Design

A fundamental challenge in speaker separation tasks is the ambiguity in speaker ordering. The model may correctly separate speakers but assign them differently than our arbitrary ground truth designation. To address this, we implement a specialized loss function that calculates the minimum loss between

- Output 1 compared to Clean Signal A and Output 2 compared to Clean Signal B
- Output 1 compared to Clean Signal B and Output 2 compared to Clean Signal A

This approach prevents penalizing the model when it correctly separates speakers but assigns them in the reverse order of our designated ground truth, ensuring the model focuses on separation quality rather than arbitrary speaker ordering.

4 References

1. Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10), 1702–1726. <https://doi.org/10.1109/TASLP.2018.2842159>
2. Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 31–35). <https://doi.org/10.1109/ICASSP.2016.7471631>
3. Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., et al. (2015). Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Liberec, Czech Republic. [Online]. Available: <https://hal.science/hal-01163493>
4. Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *Doklady Akademii Nauk*, 114, 953–956.
5. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov–Arnold Networks. *arXiv preprint arXiv:2404.17307*. <https://arxiv.org/abs/2404.17307v5>