CAN YOU HEAR ME CLEARLY?: AN EXPLORATORY STUDY ON ACOUSTIC MODELS IN SPEAKER SEPARATION

Kedar P Pandya
Department of Computer Science
Dalhousie University
Halifax, NS, Canada
kedar.pandya@dal.ca

ABSTRACT

This research project delves into speaker separation within mixed audio waveforms containing 2 and 3 speakers by employing advanced acoustic models -—WavLM [1] , Wav2Vec2 [2], and HuBERT [3] . The study assesses the efficacy of these state-of-the-art models in capturing nuanced details within audio waveforms to facilitate robust speaker separation. Processing the mixed and individual waveforms through WavLM [1], Wav2Vec2 [2], and HuBERT [3] yields embeddings, and clustering algorithms (k-means [4], agglomerative hierarchical clustering, DBSCAN [5],, and Gaussian Mixture Model [6]) are then applied to disentangle distinct sources within the mixed audio signals. The research is motivated by the evolving demands of practical applications, acknowledging the need for adaptable techniques that navigate complexities introduced by real-world noise scenarios, represented here by the incorporation of WHAM[7] dataset noise. In addition to providing insights into the comparative strengths of WavLM [1], Wav2Vec2 [2], and HuBERT [3] in the context of 2 and 3-speaker scenarios, the study contributes to a nuanced understanding of their interplay with clustering algorithms. Evaluation metrics such as Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Silhouette Score, Calinski Harabasz Index, and Davies-Bouldin Index are utilized comprehensively to assess the performance of the proposed approach. Beyond the immediate task of speaker separation, this research offers valuable insights into the adaptability and robustness of advanced acoustic models when faced with the challenges posed by real-world noise scenarios. The findings hold implications for diverse applications, including telecommunication, voice recognition, and audio transcription, where accurate speaker discrimination is essential for optimal performance.

1 Introduction

In the domain of audio signal processing, the endeavour to disentangle distinct voices within mixed audio waveforms is a significant challenge with broad applications. This study specifically focuses on the intricate task of speaker separation within mixtures comprising 2 and 3 voices, further complicated by including real-world noise from the WHAM[7] dataset. To address this challenge, we enlist the capabilities of advanced acoustic models, namely WavLM [1], Wav2Vec2 [2], and HuBERT [3], seeking to assess their efficacy in navigating the complexities inherent in real-world auditory environments.

The practical implications of accurate speaker separation resonate across applications such as telecommunication, voice recognition, and audio transcription. Our study introduces an additional layer of realism by incorporating noise from the WHAM[7] dataset, acknowledging the non-ideal conditions prevalent in actual audio scenarios.

To tackle this multifaceted challenge, we leverage the capabilities of WavLM [1], Wav2Vec2 [2], and HuBERT [3]. These acoustic models process mixed and individual waveforms, yielding embeddings that serve as distilled representations of the audio data. Moving beyond model-based processing, we employ clustering algorithms, including k-means [4] [4], agglomerative hierarchical clustering, DBSCAN [5], [5], and the Gaussian Mixture Model [6] [6], to discern and separate the distinct voices intermingled within the mixed audio signals.

The evaluation of our approach relies on a set of established metrics—Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Silhouette Score, Calinski Harabasz Index, and Davies-Bouldin Index. These metrics provide a quantitative foundation for assessing the performance of the proposed methodology in the context of speaker separation.

This paper navigates the complexities of speaker separation within two and 3-speaker scenarios enriched with real-world noise. The outcomes are anticipated to contribute to the refinement of speaker separation techniques and deepen our understanding of how advanced acoustic models handle the practical challenges inherent in audio processing.

2 Literature Review

The main focus of the paper is to use the acoustic models like HuBERT [3], Wav2Vec2 [2] and WavLM [1] to extract features from the input and use these features for further exploratory analysis. In the case of HuBERT [3], the raw waveform data is fed into the context network, which is essentially a transformer network. Through a process involving masked hidden unit prediction, HuBERT [3] learns to derive meaningful speech representations. Initially, this learned representation is compared with the clustered Mel-frequency cepstral coefficients (MFCC) representation of the same waveform in the first iteration of the model. Subsequently, it is compared to the intermediate layer clustered representation within the context network. This intricate process allows HuBERT [3] to excel in self-supervised speech representation learning, exhibiting a remarkable 13% reduction in Word Error Rate (WER) when applied to automatic speech recognition tasks.

In a fashion similar to the HuBERT [3] model, another noteworthy contender in the field is Wav2Vec2 [2], as introduced by Baveski et al. Wav2Vec2 [2], like HuBERT [3], takes the raw audio waveform as its input, forming the basis of its speech representation learning.

Wav2Vec2 [2] employs a strategy involving masking of the quantized input waveform, along with the inclusion of a transformer network as its context network, mirroring the principles that have proven effective in models like HuBERT [3]. However, a pivotal difference arises in how Wav2Vec2 [2] handles its latent speech representations. Unlike the HuBERT [3] model, Wav2Vec2 [2] selects a subset of the latent speech representation it derives from the latent feature encoder, which is typically implemented as a Convolutional Neural Network (CNN) encoder. This selected subset is then used to compute the contrastive loss concerning the representation generated by the context network.

Notably, Wav2Vec2 [2] strikes a balance between model efficiency and performance on Automatic Speech Recognition (ASR) tasks. It achieves remarkable results, particularly when applied to the LibriSpeech[8] dataset with a full 960 hours of labeled data. In this context, it managed to achieve a Word Error Rate (WER) of 1.6/3.0 on the dev-clean/dev-other subsets, underscoring its prowess in ASR tasks while maintaining a relatively efficient parameter configuration.

However, relying solely on the representations generated by HuBERT [3] or Wav2Vec2 [2] for the task of speaker separation yielded suboptimal results. These models were primarily designed for Automatic Speech Recognition (ASR) tasks, which limited their efficacy in the context of speaker separation. In response to this limitation, Chen et al. [1], introduced WavLM [1], a model tailored to address the comprehensive spectrum of speech processing tasks.

WavLM [1] stands out for its ability to generate speech representations suitable for a wide range of speech processing tasks. This model achieved a state-of-the-art performance on the SUPERB tasks, which encompass a diverse array of challenges such as keyword spotting, intent classification, slot filling, speaker diarization, and, notably, speaker separation.

WavLM [1] shares some architectural similarities with HuBERT [3], as it draws inspiration from the self-attention mechanism within a transformer encoder. However, it differentiates itself by accepting a mixture of noise and speech utterances as its input. Notably, WavLM [1] incorporates a gated relative position bias in its transformer layers, a key innovation based on the offset between the "key" and "query" components within the self-attention mechanism. This innovation enhances its ability to handle complex speech processing tasks effectively.

3 Methodology

Speaker separation within mixed audio signals is complex, requiring a systematic and nuanced methodology. In this research, I integrate three advanced acoustic models—WavLM [1], Wav2Vec2 [2], and HuBERT [3]—pre-trained on the widely used LibriSpeech[8] dataset. I load and subsequently freeze these models, utilizing them as potent feature extractors for the task.

3.1 Getting Embedding

The initial step involves processing the mixed waveform through the frozen acoustic models to obtain embeddings. This waveform represents mixed audio signals, encompassing 2 and 3 speakers, further compounded by real-world noise sourced from the WHAM[7] dataset. The embeddings generated by these pre-trained models encapsulate intricate features within the audio data, establishing a robust foundation for subsequent analysis.

Simultaneously, the source waveforms, corresponding to individual speaker signals, undergo the same process. These individual signals are also passed through the frozen acoustic models, producing embeddings that capture the unique characteristics of each speaker. The frozen models serve as sophisticated encoders, distilling the complexity of audio signals into informative embeddings.

With the obtained embeddings in hand, the next crucial phase involves the application of clustering algorithms. Specifically, I employ four prominent clustering models—k-means [4], agglomerative hierarchical clustering, DBSCAN [5],, and Gaussian Mixture Model [6] (GMM). These models operate on the embeddings derived from the mixed waveform and source waveforms, aiming to segregate and categorize the speakers in the mixed audio signals.

3.2 Clustering Models

Using clustering models in the speaker separation task within mixed audio signals is crucial for several reasons. Clustering operates in an unsupervised learning paradigm, ideal for scenarios with limited labelled data. Its adaptability to variable speaker counts allows it to identify and differentiate speakers, accommodating dynamic scenarios easily. Clustering enhances the system's robustness by capturing speaker-specific characteristics and leveraging embeddings from pre-trained acoustic models, particularly in real-world noise. It facilitates a comprehensive exploration of the embedding space, revealing underlying patterns and relationships crucial for nuanced speaker separation.

3.2.1 k-means [4] Clustering

k-means [4] is a centroid-based clustering algorithm that partitions the data into K clusters, where K is a user-defined parameter. The algorithm iteratively assigns data points to the cluster with the nearest centroid and updates the centroids based on the mean of the assigned points. Partitioning the embedded space into K clusters excels in grouping similar embeddings, effectively isolating distinct speakers within mixed audio signals. Its straightforward approach makes it computationally efficient and a valuable tool for real-time speaker separation applications. k-means [4] provides a clear delineation of speaker clusters, contributing to the effectiveness of the overall methodology.

3.2.2 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering operates under a bottom-up approach, starting with each data point as a singleton cluster and iteratively merging clusters based on proximity. The resulting hierarchy forms a tree-like structure known as a dendrogram. This approach provides insights into the relationships between data points, enabling the identification of clusters at various levels of granularity. This approach benefits speaker separation as it captures fine-grained and high-level structures, allowing for a nuanced understanding of speaker dynamics. The dendrogram generated by hierarchical clustering reveals the hierarchy of clusters, aiding in the identification and isolation of individual speakers within mixed audio signals.

3.2.3 DBSCAN [5], (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN [5], is a density-based clustering algorithm that groups data points based on their density within the embedded space. It identifies core points with sufficient neighbours within a specified radius and expands clusters by connecting density-reachable points. Unlike centroid-based methods, DBSCAN [5], can discover clusters of varying shapes and sizes, making it robust in scenarios where speakers may exhibit diverse patterns. In speaker separation, DBSCAN [5], excels at discovering clusters of different shapes and sizes, enabling effective separation of speakers even in irregular patterns. Its adaptability to the density of embeddings contributes to the accuracy of the speaker separation methodology.

3.2.4 Gaussian Mixture Model [6] (GMM)

Gaussian Mixture Model [6] (GMM) is a probabilistic clustering algorithm that models the data as a mixture of several Gaussian distributions. Each cluster is associated with a Gaussian distribution, and the model parameters are estimated through the Expectation-Maximization (EM) algorithm. GMM [6] is particularly useful when the underlying data distribution is not well-defined and exhibits probabilistic characteristics. In the context of speaker separation, GMM [6]

captures the probabilistic relationships between embeddings, providing a sophisticated representation of the complex structure within the embedded space. It excels in scenarios where the boundaries between clusters are not well-defined.

The frozen acoustic models, acting as feature extractors, provide a universal representation for both mixed waveform and source waveforms. This uniformity facilitates a seamless integration of clustering algorithms, allowing for a comprehensive exploration of the embedded space.

The evaluation of the speaker separation process is a critical aspect of my methodology. I employ a set of well-established metrics to quantify the clustering algorithms' performance.

The metrics collectively provide a comprehensive understanding of the efficacy of my speaker separation methodology. The quantitative assessment allows for comparisons between different clustering algorithms. It provides insights into their suitability for the unique challenges posed by mixed audio signals containing 2 and 3 speakers and real-world noise.

In summary, this methodology encompasses the use of pre-trained acoustic models frozen after exposure to the diverse LibriSpeech[8] dataset. The freezing ensures that the models act as fixed feature extractors, capturing the essence of mixed and source waveforms. Subsequent application of diverse clustering algorithms aims to unravel the intricate tapestry of speakers within the mixed audio signals. The rigorous evaluation using a diverse set of metrics validates my approach's effectiveness, providing valuable insights for advancing speaker separation techniques in real-world audio scenarios.

4 Experimental Details

4.1 Dataset

The dataset employed in this research is derived from the comprehensive LibriSpeech[8] data corpus and specifically utilizes the LibriMix[9] dataset. To construct LibriMix[9], samples are selectively drawn from the train-clean-100 subset, although variations such as dev-clean, test-clean, or train-clean-360 may be chosen based on specific input requirements. The dataset creation process involves the generation of a metadata sheet, pairing two distinct audio file names for scenarios with two speakers and three instances with three speakers. Subsequently, a script orchestrates the combination of these audio files by overlapping them at random starting times, mimicking the natural variability encountered in real-world audio mixtures. WHAM[7] noise is integrated into the mixed audio signals to enhance the dataset's realism further. This meticulous approach to dataset creation ensures that the experimental setup closely mirrors actual scenarios, providing a robust foundation for assessing the effectiveness of the proposed speaker separation methodology.

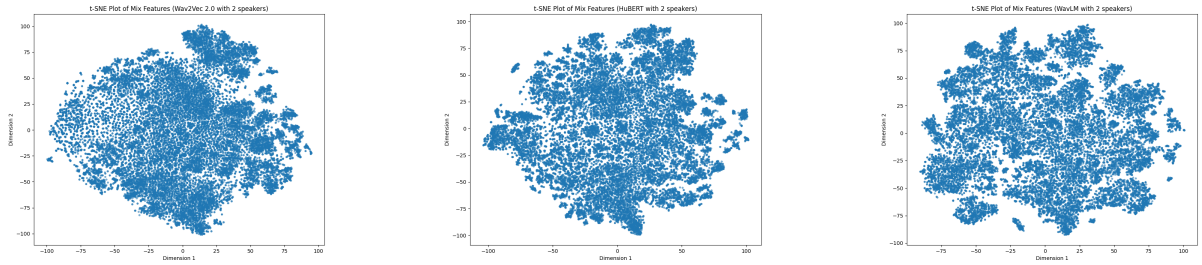


Figure 1: This is a visualization of how Wav2Vec2 [2], HuBERT [3] and WavLM [1] model extract the features from the mixed waveforms for two speakers. It is a visible difference how these model are extracting the features from the input from this t-SNE plots.

4.2 Dataset Preprocessing

The dataset utilized in this study is sourced from the LibriMix[9] dataset, and its integration into the research framework is facilitated by the PyTorch LibriMix[9] class. This class is instrumental in loading the dataset and presenting the data in a structured Tuple format—specifically, Tuple[int, Tensor, List[Tensor]]. Here, the 'int' component represents the sample rate, the 'Tensor' encapsulates the mixture waveform, and the 'List[Tensor]' corresponds to a list of source waveforms. Considering these distinct components enables a comprehensive representation of the audio data.

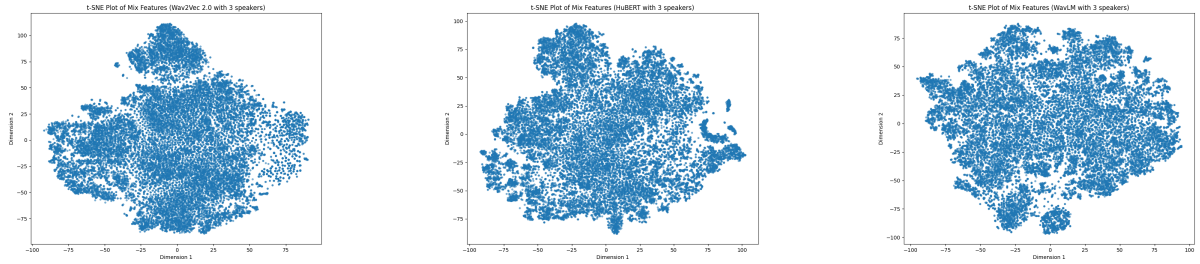


Figure 2: This is a visualization of how Wav2Vec2 [2], HuBERT [3] and WavLM [1] model extract the features from the mixed waveforms for three speakers. It is a visible difference how these model are extracting the features from the input from this t-SNE plots.

A meticulous preprocessing strategy is employed during batch formation to address the inherent variability in feature vector sizes across the dataset. After the formation of batches, the waveforms undergo a padding procedure, ensuring uniformity in the size of feature vectors within each batch. This crucial step is executed by the collate function, which pads the waveforms to the highest value of the feature vector within the given batch. The rationale behind this approach is to homogenize the input dimensions before presenting the batch to the acoustic models, thereby facilitating coherent and consistent processing by the models during the training and evaluation phases.

In adopting such a formalized approach to dataset loading and preprocessing, the research ensures a standardized and reproducible foundation for subsequent experimentation. This systematic treatment of the dataset aligns with best practices in deep learning research, contributing to the robustness and reliability of the overall research methodology.

4.3 Acoustic Models

The acoustic models deployed in this research—namely Wav2Vec2 [2], WavLM [1], and HuBERT [3]—were sourced from the HuggingFace model repository, a widely recognized and reputable resource for pre-trained models in natural language processing and audio-related domains. Leveraging the convenience and reliability of HuggingFace, the pre-trained weights for these models were obtained from their official repository. Specifically, the Wav2Vec2 [2] model utilized in this study derived its weights from the base model trained on an extensive corpus of 960 hours of data, ensuring a robust representation of audio features. In parallel, the WavLM [1] base model, trained on a concise yet diverse dataset spanning 100 hours, and the HuBERT [3] large model, trained on an extensive 960 hours of data, were selected for their distinct characteristics and capabilities.

All three acoustic models—Wav2Vec2 [2], WavLM [1], and HuBERT [3]—share a common origin, having undergone training on the LibriSpeech[8] data corpus. This choice of training data ensures a consistent foundation across the models, aligning with the overarching goal of maintaining a standardized experimental setup. By drawing from pre-trained models, the research capitalizes on the wealth of knowledge encoded in these models' weights, acquired during their training on diverse and extensive datasets. This strategic selection of acoustic models leverages the expertise embedded in state-of-the-art pre-trained models and positions the research within the broader context of advancements in deep learning for audio processing. The visualization of the features vectors after being passed through the acoustic models is shown in 1, 2

4.4 Clustering Models

In the speaker separation pipeline, the clustering models play a pivotal role by operating on the embeddings extracted from the acoustic models. Building on the advantages highlighted earlier, these clustering models—k-means [4], Agglomerative Clustering, and Gaussian Mixture Model [6] (GMM)—are strategically applied to discern and isolate individual speakers within the mixed audio signals. The hyperparameter tuning for these clustering models is executed meticulously to optimize their performance in the speaker separation task.

Following a hypothesis grounded in the notion that the number of clusters should align with the number of speakers in the audio mixture, the k-value, representing the number of clusters, is set to 2 for scenarios with two speakers and 3 for scenarios with three speakers. This careful selection is driven by the expectation that each cluster would correspond to an individual speaker, facilitating the subsequent extraction of distinct speaker sources. However, acknowledging the inherent challenges posed by real-world noise in the input audio, alternative configurations have been explored.

In particular, recognizing that noise may introduce complexities in the clustering process, experiments were conducted with modified k-values. For two-speaker scenarios, a k-value of 3 was tested in k-means [4] clustering, and for three-speaker scenarios, a k-value of 4 was employed in k-means [4] clustering. This exploration aims to assess the clustering models' robustness in scenarios where noise may influence the actual number of clusters. By systematically varying the k-values, the research captures the intricate dynamics of mixed audio signals, further refining the clustering models' adaptability to real-world acoustic environments.

4.5 Evaluation Metrics

A suite of standard clustering evaluation metrics was utilized to provide a comprehensive quantitative evaluation in assessing the efficacy of the clustering models employed for speaker separation. The commonly employed metrics include the Silhouette score, Calinski-Harabasz index, and Davies-Bouldin index, each offering distinct insights into the quality and characteristics of the generated clusters.

The **Silhouette score** (S) is a metric that quantifies how well-separated the clusters are. For each sample i, it is computed as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where $a(i)$ is the average distance from the i-th sample to other samples in the same cluster, and $b(i)$ is the average distance from the i-th sample to samples in the nearest cluster (excluding the cluster to which i belongs). The overall Silhouette score is the mean of $S(i)$ across all samples. The typical range of the silhoutte score ranges from -1 to 1 where a score of 1 indicates well-defined clusters and -1 indicates instances being assigned in a wrong clusters.

The **Calinski-Harabasz index** is a ratio of the between-cluster variance to the within-cluster variance, providing a measure of cluster compactness and separation. A higher value of CHI indicates a better, well-defined and compact clusters. It is calculated as:

$$Calinski - HarabaszIndex = \frac{Between - ClusterVariance}{Within - ClusterVariance}$$

The **Davies-Bouldin index** assesses the compactness and separation of clusters, with lower values indicating better clustering. It is computed as the average similarity ratio of each cluster with its most similar cluster:

$$Davies - BouldinIndex = \frac{1}{N}\sum_{i=1}^{N}\max_{j\neq i}\left(\frac{S_i + S_j}{d(C_i, C_j)}\right)$$

Here, N is the number of clusters, S_i and S_j are the average distances from points in clusters C_i and C_j to their respective cluster centers, and $d(C_i, C_j)$ is the distance between the cluster centers.

Additionally, given the availability of true labels in the form of source waveforms, further evaluation metrics include the **Adjusted Rand Index** (ARI) and **Normalized Mutual Information** (NMI).

The **Adjusted Rand Index** measures the similarity between the true and predicted clustering, adjusting for chance. It is computed as:

$$ARI = \frac{RI - \text{Expected\_RI}}{\max(RI_{\max} - \text{Expected\_RI}, 0)}$$

Where RI is the Rand Index, $Expected\_RI$ is the expected value of the Rand Index under a null hypothesis, and RI_{max} is the maximum possible value of the Rand Index. An ARI value typically ranges between -1 and 1 with 1 indicating perfect clustering, 0 indicating random clustering and a value less than 0 indicating worse than random.

The **Normalized Mutual Information** measures the mutual dependence between the true and predicted clusterings, normalized to fall between 0 (no mutual information) and 1 (perfect agreement). It is computed as:

$$NMI = \frac{I(X;Y)}{\sqrt{H(X) \cdot H(Y)}}$$

where $I(X;Y)$ is the mutual information between the true (X) and predicted (Y) clusterings, and $H(X)$ and $H(Y)$ are the entropies of X and Y respectively.

5 Results and Discussion

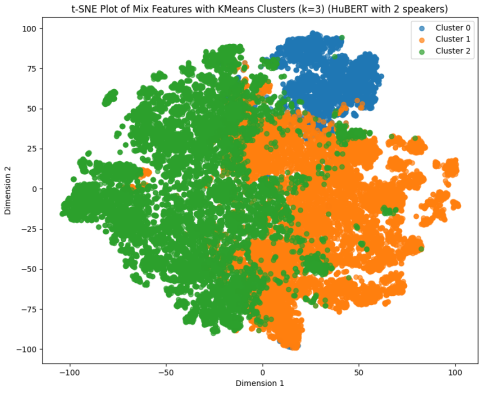
The unanticipated outcomes observed in the evaluation metrics, particularly the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), necessitate a nuanced examination to discern factors contributing to these unexpected results. In scenarios where ideal agreement between true labels, represented by embeddings from acoustic models, and predicted cluster assignments should approach a value of 1, the observed metrics consistently approximating 0 suggest a clustering pattern akin to randomness.

A plausible explanation for this phenomenon resides in the intricate nature of the audio data and the challenges introduced by real-world noise. The presence of noise, both in the original audio signals and during the clustering process, introduces complexities that impede the accurate discernment and segregation of individual speakers. The inherent variability in speaker characteristics, speech patterns, and background noise prevalent in real-world acoustic environments may pose challenges that the existing methodology may not fully address.

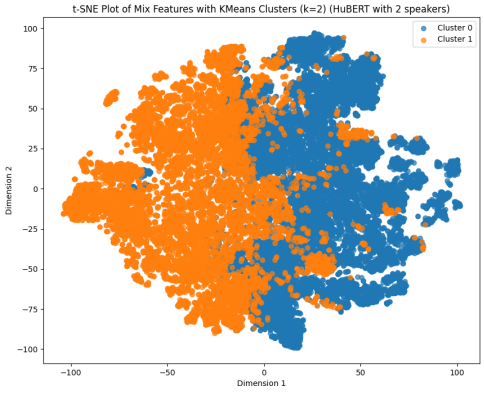
	Model	ARI	NMI	Silhouette Score	CHI	DBI
Two Speakers	k-means [4] (K = 3)	2.27e-8	0.18	0.18	5178.85	1.85
	k-means [4] (K = 2)	1.06e-8	0.13	0.16	4868.75	2.28
	Agglomerative Clustering (K = 2)	-2.57e-9	0.09	0.1	3859.17	1.58
	DBSCAN [5],	4.26e-9	0.04	-0.2	2.39	1.11
	Gaussian Mixture Model [6] (K = 2)	-2.34e-8	0.12	0.1	2871.09	2.95
Three Speakers	k-means [4] (K = 3)	-2.68e-9	0.18	0.19	5020.50	1.65
	k-means [4] (K = 4)	-2.09e-8	0.24	0.18	4476.77	1.90
	Agglomerative Clustering (K = 3)	1.21e-8	0.17	0.15	4064.24	1.77
	DBSCAN [5],	4.09e-10	0.01	-0.06	2.79	0.925
	Gaussian Mixture Model [6] (K = 3)	-1.56e-8	0.19	0.11	2698.74	3.27

Table 1: The performance of the selected clustering models on the audio features extracted from the WavLM [1] model.

Furthermore, metrics assessing cluster purity, such as the Silhouette Score, Calinski-Harabasz Index (CHI), and NMI, yielded results consistent with the ARI and NMI metrics, reinforcing the observed clustering pattern resembling randomness. A noteworthy observation emerged from the application of DBSCAN [5], clustering to all three acoustic embeddings. Despite suboptimal performance across all models, DBSCAN [5], exhibited notably poor results in the CHI metric. Visualization of the formed clusters revealed that the model tended to create clusters for individual feature vectors, predominantly designating most feature vectors as outliers. This behavior contributed to the subpar performance of DBSCAN [5],, indicating challenges in effectively discerning meaningful clusters within the data. The visualization of the cluster for one of the embeddings is shown in the Figure 5

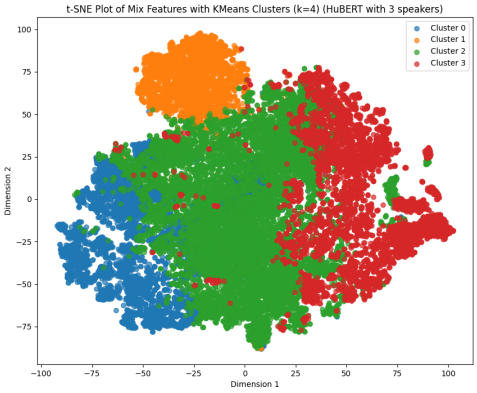


(a) t-SNE graph for the clusters formed by HuBERT [3] with 2 speakers and k=3 in k-means [4]

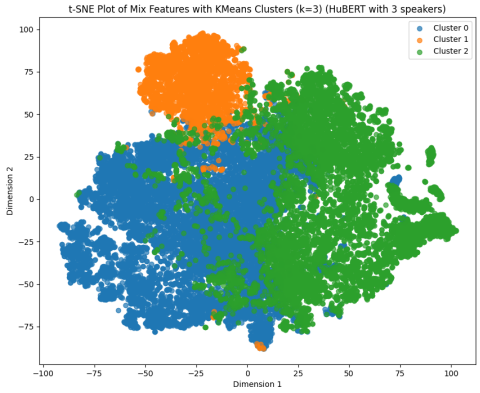


(b) t-SNE graph for the clusters formed by HuBERT [3] with 2 speakers and k=2 in k-means [4]

In contrast, for scenarios involving two and three speakers, k-means [4] clustering applied to HuBERT [3] embeddings demonstrated superior performance compared to other clustering algorithms. This observation is further supported by the t-SNE plot, which visually represents the clusters formed by the clustering algorithm. The discernible efficacy of k-means [4] clustering with HuBERT [3] embeddings suggests a potential alignment of this particular configuration with the intricate dynamics of speaker separation within mixed audio signals. The exploration of clustering algorithm



(a) t-SNE graph for the clusters formed by HuBERT [3] with 3 speakers and k=4 in k-means [4]



(b) t-SNE graph for the clusters formed by HuBERT [3] with 3 speakers and k=3 in k-means [4]

	Model	ARI	NMI	Silhouette Score	CHI	DBI
Two Speakers	k-means [4] (K = 3)	1.26e-8	0.17	0.16	5029.32	1.83
	k-means [4] (K = 2)	-1.6e-11	0.12	0.16	6388.83	1.99
	Agglomerative Clustering (K = 2)	-1.67e-8	0.11	0.14	3900.05	2.4
	DBSCAN [5],	4.88e-9	0.07	-0.32	2.49	1.61
	Gaussian Mixture Model [6] (K = 2)	5.13e-9	0.1	-0.02	602.17	4.96
Three Speakers	k-means [4] (K = 3)	-1.77e-8	0.17	0.16	4252.68	1.86
	k-means [4] (K = 4)	-2.46e-8	0.22	0.12	3604.98	2.19
	Agglomerative Clustering (K = 3)	-1.85e-8	0.17	0.12	3431.42	2.1
	DBSCAN [5],	5.46e-9	0.06	-0.31	2.02	1.80
	Gaussian Mixture Model [6] (K = 3)	-8.15e-9	0.16	0.03	1902.41	3.95

Table 2: The performance of the selected clustering models on the audio features extracted from the Wav2Vec2 [2] model.

behavior and performance nuances sheds light on the complexity of the task, motivating further refinement and exploration of methodologies to enhance speaker separation capabilities in real-world acoustic environments. The visualization of the clusters formed by k-means [4] can be seen in the Figure 3a 3b

	Model	ARI	NMI	Silhouette Score	CHI	DBI
Two Speakers	k-means [4] (K = 3)	1.31e-8	0.17	0.20	6660.79	1.57
	k-means [4] (K = 2)	5.22e-9	0.13	0.18	7459.92	1.80
	Agglomerative Clustering (K = 2)	1.39e-8	0.12	0.14	5255.93	2.06
	DBSCAN [5],	1.20e-11	0.01	-0.25	0.82	1.65
	Gaussian Mixture Model [6] (K = 2)	-5.47e-9	0.13	0.05	875.01	5.19
Three Speakers	k-means [4] (K = 3)	-2.44e-8	0.18	0.20	5823.02	1.62
	k-means [4] (K = 4)	2.53e-9	0.22	0.17	5128.35	1.80
	Agglomerative Clustering (K = 3)	-1.49e-8	0.15	0.16	4260.02	1.76
	DBSCAN [5],	3.49e-11	0.01	-0.26	1.02	1.67
	Gaussian Mixture Model [6] (K = 3)	-2.31e-8	0.19	0.07	1365.97	4.13

Table 3: The performance of the selected clustering models on the audio features extracted from the HuBERT [3] model.

The observed results evoke a robust discussion on the challenges and insights garnered from the application of clustering models to speaker separation tasks within mixed audio signals. The notable discrepancy between true labels and predicted cluster assignments, as reflected in metrics like the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), underscores the intricate nature of the audio data and the formidable impact of real-world noise. The detailed results in numerical form is seen in Table 1, 2, 3 The presence of noise, both inherent in the original audio signals and introduced during the clustering process, presents complexities that challenge the models' capacity to accurately discern and isolate individual speakers. The consistent convergence of various metrics, including the

Silhouette Score, Calinski-Harabasz Index (CHI), and NMI, toward patterns resembling randomness accentuates the formidable nature of the task.

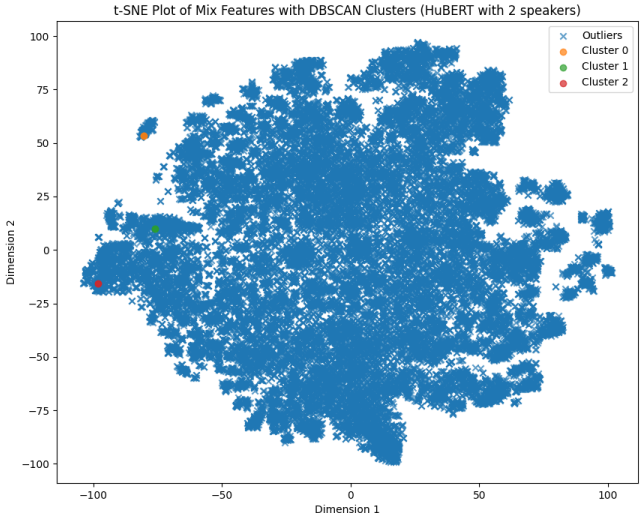


Figure 5: This is one of the clustering plot from the DBSCAN [5], model for 2 speakers HuBERT [3] features vectors.

An intriguing observation arises from the suboptimal performance of DBSCAN [5], clustering, particularly evidenced by its poor performance on the CHI metric. The visual inspection of clusters formed by DBSCAN [5],, with an inclination to designate individual feature vectors as outliers, provides valuable insights into the limitations of this algorithm in effectively capturing meaningful clusters within the data. This observation prompts a critical reevaluation of the suitability of DBSCAN [5], for the nuances of speaker separation tasks within mixed audio signals.

Contrastingly, the superior performance of k-means [4] clustering, specifically when applied to HuBERT [3] embeddings in scenarios involving two and three speakers, introduces a promising avenue for further exploration. The alignment of k-means [4] clustering with the HuBERT [3] embeddings, as substantiated by the t-SNE plot, suggests a potential synergy that merits deeper investigation. This nuanced understanding of clustering algorithm behavior and performance nuances not only highlights the complexities of speaker separation tasks in real-world acoustic environments but also motivates ongoing efforts to refine methodologies and enhance the efficacy of clustering models in addressing these challenges. The discussion encapsulates the multifaceted nature of the results, offering insights that pave the way for continued exploration and advancement in the domain of speaker separation within diverse and dynamic audio contexts. However, it is also notable that the performance of the clustering models on WavLM [1] embeddings was overall better as compared to other embeddings.

6 Conclusion

In conclusion, this study has illuminated the intricate challenges and potential breakthroughs associated with employing clustering models for speaker separation within mixed audio signals. The substantial discrepancy between true labels and predicted cluster assignments, as indicated by metrics like the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), underscores the complexity of audio data and the impact of real-world noise. The presence of noise, both in original audio signals and during the clustering process, introduces complexities that challenge the models' ability to discern and isolate individual speakers accurately. The convergence of various metrics, including the Silhouette Score, Calinski-Harabasz Index (CHI), and NMI, towards randomness accentuates the formidable nature of the task.

An intriguing finding pertains to the performance of clustering algorithms, notably DBSCAN [5], and k-means [4]. While DBSCAN [5], exhibited poor results, despite being a highly regarded clustering algorithm. k-means [4] on the other hand, was able to achieve better results than others due to its simplicity. Moreover, due to it's complex architecture involving gated relative position bias, WavLM [1] was able to produce embeddings that could produce better clusters than other embeddings. This nuanced understanding of clustering algorithm behavior underscores the complexities of

speaker separation tasks in real-world acoustic environments and motivates ongoing efforts to refine methodologies and enhance clustering model efficacy. The conclusion encapsulates the multifaceted nature of the results, offering insights for continued exploration and advancement in speaker separation within diverse audio contexts.

References

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, oct 2022.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[4] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Knowledge Discovery and Data Mining*, 1996.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[7] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," 2019.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

[9] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020.