Clean vs. Overlapped Speech-Music Detection System

Prabha Sharma - M22AIE224, Prateek Singhal - M22AIE215, Harsh Parashar - M22AIE210
Department of Computer Science, IIT-Jodhpur

1. Introduction

Speech and Music are often found as overlapping mixtures in most practical scenarios. It is very common to see in movies having dialogues with background music going on. These segments need to be identified beforehand and processed separately, these may disrupt the performance of high-level applications like automatic speech recognition and music information retrieval.

This project is using the paper titled "Clean vs. Overlapped Speech-Music Detection Using Harmonic-Percussive Features and Multi-Task Learning"[1] published by IEEE in 2022, link to paper

Github link to this project code is https://github.com/parasharharsh16/Clean-vs-Overlapped-Speech-Music-Detection

2. Problem Statement

Speech signals have wavy and continuous harmonics, while music signals exhibit horizontally linear and discontinuous harmonic patterns. Music signals also contain more percussive components than speech signals, manifested as vertical striations in the spectrograms. In case of speech music overlap, it might be challenging for automatic feature learning systems to extract class-specific horizontal and vertical striations from the combined spectrogram representation.

3. Literature Review

Initial studies in speech overlapped with music detection were performed using traditional feature engineering approaches and ML algorithms. There are studies on Non-Negative Matrix Factorization, Auto-correlation based features, or Principal Component Analysis to suppress or separate the presence of background music from the speech. Deep-learning-based algorithms have also been explored in the task of speech overlapped with music detection. To the best of the authors' knowledge, all previous works have used a combined harmonic and percussive representation. Speech and music signals have distinct harmonic and percussive characteristics. The harmonics in speech have a wavy structure, while music harmonics are relatively more stable (horizontally linear). This idea is the main motivation of using Harmonic-Percussive Source Separation (HPSS) to compute features previously unexplored in this task.

4. Proposed Solution

4.1. Classical Threshold Model

Proposed Solution for this problem works on the principal of HPSS decomposition where separability can be analyzed based on harmonics spans over rows and columns for speech and music. Separately presenting the harmonic and percussive information might help in better learning the discrimination. Class separability in HPSS with the fact that harmonics in music span only few adjacent rows whereas harmonics in speech may span over many rows in spectrogram because of wavy nature. The main task is to have 3 class classification of isolated music, speech, and speech+music (mixed)

- **HPSS (Harmonic Percussive Source Separation):** We will use HPSS decomposition method proposed by Fitzgerald et al. where harmonic enhanced spectrogram (H) and percussion enhanced spectrogram (P) is used to obtain respective decompositions (Hard and Soft mask). This work derives Soft mask using Spectrograms to obtain decomposition.

$$\mathbf{M}_H[i,j] = \frac{H^r[i,j]}{(H^r[i,j] + P^r[i,j])}$$
$$\mathbf{M}_P[i,j] = \frac{P^r[i,j]}{(H^r[i,j] + P^r[i,j])}$$

Figure 1: *Hard Mask and Soft Mask*

Where H and P are Harmonic enhanced spectrogram and percussion enhanced spectrogram.

Below are overall approach to solve the problem. We have used the threshold model where following conditions are used to classify the given classes (Speech, Music or Mixture):

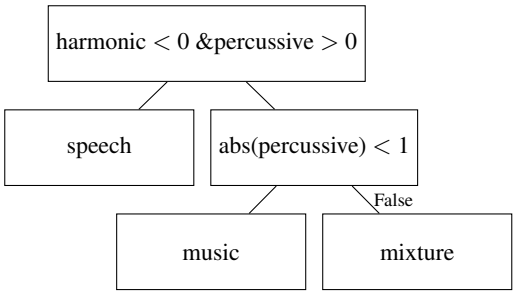

Figure 2: *Conditional view for prediction threshold, both persuasive and harmonic values are skewed*

4.2. Classical Machine Learning Model

To classsify the audios in classes given above, we used the audio data spectrogram to train the classical ML models and tried to achieve better accuracy in classification. spectrogram of Music and Speech varies at many levels, like music has

continuous distribution on the other hand, speech has gaps (pauses) which are clearly indicative in plots as well. We have used following two models to fit the spectrogram data and classify the audios.

- **Support Vector Machine**
- **Random Forest Classifier**

4.3. MTLF (Multi-Task Learning Framework)

Keeping the consideration of complexity of underlying task and generalizability to unseen data, a MTL framework been used. A Multi-Task Learning (MTL) framework attempts to overcome the problems of performance by learning multiple closely related subproblems using a single model.

Model Architecture: Temporal Convolutional Network (TCN): The model utilizes a TCN as its main feature extraction component. We are using one TCN layer followed by convolutional layers and multi dense layers.

The architecture is given below:

Table 1: *Model Architecture of MtlCascadeModel*

Layers	Description
TCN	– num_channels: 1 – kernel_size: 3. – dropout: Randomly chosen between 0.05 and 0.5. – dilations: Exponentially growing with the base of 2.
Dense	– dense_sp, dense_mu, dense_smr: Separate dense layers for different outputs, with each having its own specified number of hidden nodes and layers (20 nodes, 1 layer).
Output	– Each dense layer's output is then passed to a final output layer (output_sp, output_mu, output_smr), each concluding with a sigmoid activation function for binary classification or probability prediction.

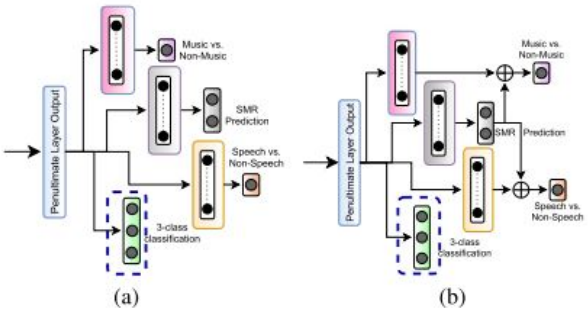

Figure 3: *Traditional MTL and Cascaded Information MTL*

4.4. Dataset

Dataset used for this solution is publicly available speech and music dataset MUSAN which is a corpus of music, speech, and noise recordings. Dataset link

Speech (Hrs)	Music (Hrs
60 hr	42 hr

Table 2: *Audio Files*

4.5. Data Prepration

To prepare the dataset we followed a systematic approach consisting of several steps. Firstly, we loaded the Musan metadata file, which provided us with paths of both types of files (speech or music). Subsequently, we prepared a combination of music and speech to create mixed signals. This involved randomly selecting windows within the signals(speech and music), ensuring that the data's uniqueness remained high and avoiding monotony.

1. **Data for threshold model** The Short-Time Fourier Transform (STFT) is applied to segment the input signal into time-frequency components. Using the HPSS algorithm, the STFT representation is split into harmonic and percussive components, discerning tonal and transient elements. Skewness, a measure of asymmetry, is computed for both harmonic spectrogram rows and percussive spectrogram columns, providing insights into their distributions.

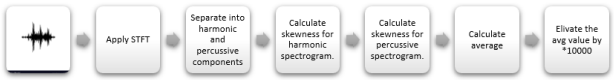

Figure 4: *Data Processing for Classical Model*

2. **Data for ML and MTL Model** After generating the mixed signals, we modified their decibels within a specified range. In our experiments, we adjusted the decibels to fall within the range of [-5, 5] to ensure consistent level of audibility Finally, we converted the speech, music, and mixed signals into spectrograms. This conversion allowed us to analyze the frequency distribution and patterns within the signals, enabling deeper exploration of harmonic/non-harmonic patterns.

Figure 5: *Data Processing for ML/DL Models*

5. Model Evaluation

The procedure entails conducting evaluation for each category: speech, music, and mixed. It assesses the prediction and target values of each category using test data, which are then utilized to determine precision, F1 scores, and accuracy. Subsequently,

this assessment data is visualized through ROC and AUC graphs for each category.The evaluation is conducted for all models, including classical, threshold, and DL, across various fractions of data.

5.1. Evaluation Results

Model	Data	Speech				Music				Mixed			
		Precn	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc
Threshold	1%	1.00	1.00	1.00	8.4%	1.00	1.00	1.00	47.8%	0.00	0.00	0.00	46.2%
SVM	1%	1.00	0.27	0.42	7.2%	1.00	1.00	1.00	23.9%	1.00	1.00	1.00	14.5%
Random Forest	1%	1.00	0.42	0.59	15.5%	1.00	1.00	1.00	39.6%	1.00	1.00	1.00	33.2%
MTL	**1%**	0.94	0.84	0.89	**92.7%**	0.66	0.79	0.72	**79.3%**	0.85	0.23	0.36	**73.6%**
Threshold	5%	1.00	1.00	1.00	8.3%	1.00	1.00	1.00	47.9%	0.00	0.00	0.00	46.6%
SVM	5%	1.00	0.22	0.37	7.5%	1.00	1.00	1.00	28.6%	1.00	1.00	1.00	17.6%
Random Forest	5%	1.00	0.27	0.43	11.0%	1.00	1.00	1.00	54.8%	1.00	1.00	1.00	27.7%
MTL	**5%**	0.98	0.89	0.93	**95.7%**	0.74	0.85	0.79	**85.1%**	0.93	0.43	0.59	**79.7%**
Threshold	10%	1.00	1.00	1.00	8.1%	1.00	1.00	1.00	47.7%	0.00	0.00	0.00	46.1%
SVM	10%	1.00	0.45	0.62	19.0%	1.00	1.00	1.00	13.8%	1.00	1.00	1.00	17.4%
Random Forest	10%	1.00	0.68	0.81	34.0%	1.00	1.00	1.00	35.8%	1.00	1.00	1.00	34.9%
MTL	**10%**	0.98	0.84	0.90	**94.0%**	0.83	0.73	0.78	**86.3%**	0.86	0.63	0.73	**84.3%**
Threshold	20%	1.00	1.00	1.00	8.2%	1.00	1.00	1.00	47.9%	0.00	0.00	0.00	47.0%
SVM	20%	1.00	0.24	0.38	6.7%	1.00	1.00	1.00	25.5%	1.00	1.00	1.00	14.9%
Random Forest	20%	1.00	0.34	0.51	13.5%	1.00	1.00	1.00	43.2%	1.00	1.00	1.00	33.1%
MTL	**20%**	0.98	0.96	0.97	**98.0%**	0.78	0.88	0.83	**88.0%**	0.95	0.53	0.68	**83.3%**

Table 3: *Evaluation table for all the models*

5.2. AUC ROC Curve

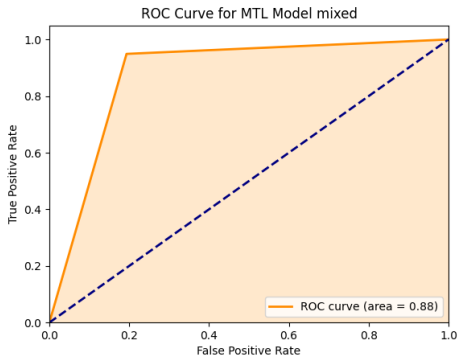

Figure 6: *ROC curve illustrating the performance of the MTL model on Speech+Music*

6. Result

MTL has emerged as the most effective approach across different tasks and performance thresholds. It consistently outperforms other models like Support Vector Machines (SVM), Random Forest, and the threshold model in terms of accuracy. MTL achieves notably high accuracy scores ranging from 73% to 98%, indicating its capability to handle multiple tasks simultaneously and utilize shared information among tasks to enhance overall performance.

Although the threshold model demonstrates exceptional precision, recall, and F1 scores in distinguishing between pure speech and music, it encounters difficulties in scenarios involving mixed data, resulting in consistently low accuracy metrics. In contrast, MTL exhibits a more balanced performance across all tasks and thresholds, highlighting its adaptability and efficiency in managing various data types and situations.

This suggests that while the threshold model may excel in specific contexts, MTL offers a more dependable solution for real-world applications where accuracy across diverse tasks is essential.

7. References

1. Mrinmoy Bhattacharjee, S. R. M. Prasanna, Prithwijit Guha, "Clean vs. Overlapped Speech-Music Detection Using Harmonic-Percussive Features and Multi-Task Learning" *IEEE*, 2022.