Clean vs. Overlapped Speech-Music Detection Using Harmonic-Percussive Features and Multi-Task Learning

Mrinmoy Bhattacharjee, *Graduate Student Member, IEEE*, S. R. M. Prasanna, *Senior Member, IEEE*, and Prithwijit Guha, *Member, IEEE*

***Abstract*—Detection of speech and music signals in isolated and overlapped conditions is an essential preprocessing step for many audio applications. 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. A pre-processing step of separating the harmonic and percussive components before training might aid the classifier. Thus, this work proposes the use of harmonic-percussive source separation method to generate features for better detection of speech and music signals. Additionally, this work also explores the traditional and cascaded-information multi-task learning (MTL) frameworks to design better classifiers. MTL framework aids the training of the main task by employing simultaneous learning of several related auxiliary tasks. Results have been reported both on synthetically generated speech music overlapped signals and real recordings. Four state-of-the-art approaches are used for performance comparison. Experiments show that harmonic and percussive decomposition of spectrograms perform better as features. Moreover, the MTL-framework based classifiers further improve performances.**

***Index Terms*—Speech music overlap detection, harmonic percussive source separation, multi-task learning, radio broadcast audio classification.**

I. Introduction

SPEECH and music are the most frequently encountered audio categories in movies, TV shows, web series, and radio broadcasts. Researchers have been tackling the problem of speech vs. music classification for a long time now. State-of-the-art methods [1]–[4] can identify isolated speech and music segments with impressive accuracy. However, speech and music are often found as overlapping mixtures in most practical scenarios. For example, sentimental scenes in movies and TV shows frequently have speech with background music to highlight the scene's mood. If such segments are not identified beforehand and processed separately, these may disrupt the performance of high-level applications like automatic speech recognition and music information retrieval. Hence, this work focuses on discriminating isolated speech and music segments from their overlapping mixtures.

Initial studies in speech overlapped with music detection were performed using traditional feature engineering approaches and machine learning algorithms. Some authors dealt with the presence of background music by compensating for it [5], using Non-Negative Matrix factorization to suppress it [6] or separate it using Independent Component Analysis [7]. Others detected the presence of background music using autocorrelation-based features [8] or Principal Component Analysis [9]. Some works attempted to segment overlapping soundtracks by using Singular Spectrum Analysis [10] or Self-Similarity Matrix-based approach [11]. Most works used classifiers like Gaussian Mixture Models (GMM, henceforth) [5], [8], Support Vector Machines (SVM, henceforth) [5], [8], [11], Random Forests [11] and Logistic Regression [11].

Deep-learning-based algorithms have also been explored in the task of speech overlapped with music detection. Jia et al. [12] detected the presence of music using a novel Hierarchical Regulated Iterative Network, while Gimeno et al. [13] used Recurrent Neural Network trained on limited data for the task. Venkatesh et al. [14] used Convolutional Recurrent Neural Network with artificially synthesized radio-broadcast like speech and music data. In a recent work of this paper's author's [15], an enhanced version of spectrograms called pyknograms were explored in the task of speech overlapped with low-energy music detection with a fully convolutional network.

In this context, it is relevant to review the popular Albayzín campaigns, a set of audio processing challenges open for public participation. Audio segmentation evaluation (ASE, henceforth) was one of the tasks in their 2010, 2012 [16] and 2014 [17] editions. In the Albayzín-2014 ASE, participating systems were required to identify the presence of speech, music, or noise,

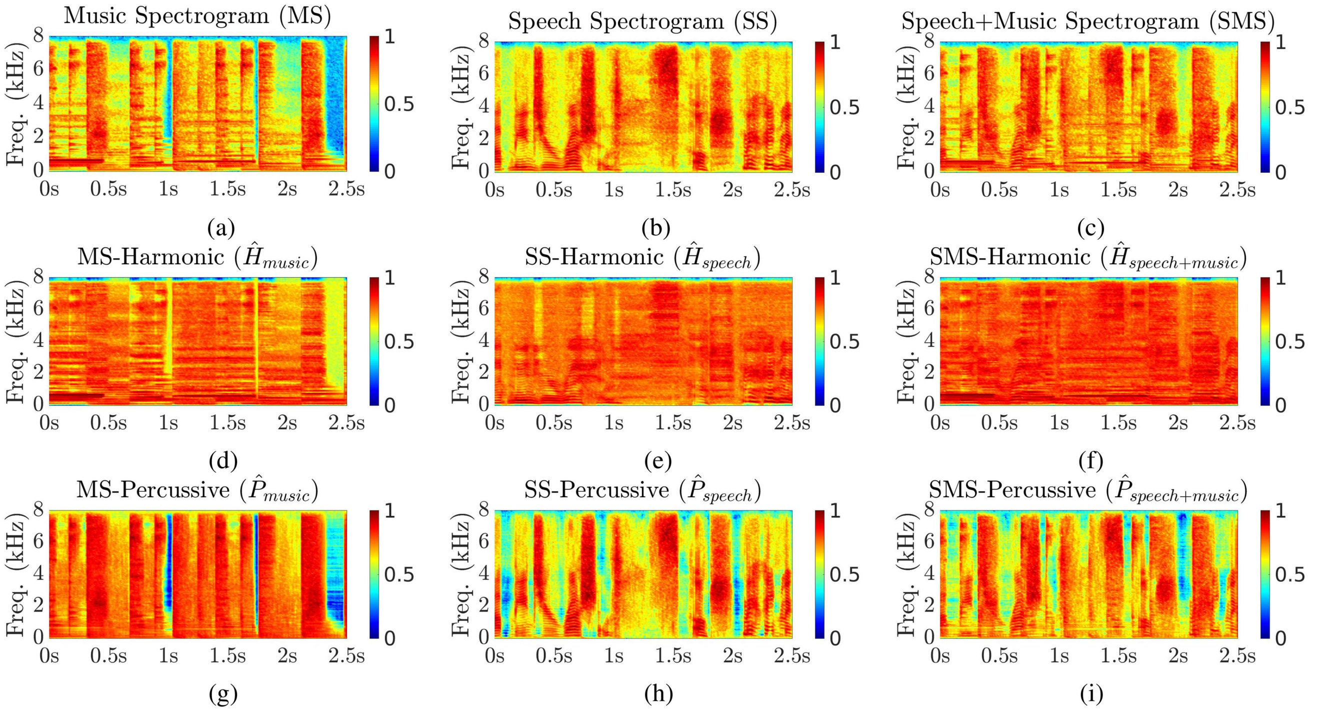
Manuscript received 5 November 2021; accepted 15 March 2022. Date of publication 1 April 2022; date of current version 2 December 2022. This work was supported in part by Visvesvaraya Ph.D. Scheme, MeitY, Govt. of India – MEITY-PHD-1230, and in part by the Department of Biotechnology, Government of India under Project BT/COE/34/SP28408/2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tom Backstrom. *(Corresponding author: Mrinmoy Bhattacharjee.)*

Mrinmoy Bhattacharjee and Prithwijit Guha are with the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India (e-mail: mrinmoy.bhattacharjee@iitg.ac.in; pguha@iitg.ernet.in).

S. R. M. Prasanna is with the Department of Electrical Engineering, Indian Institute of Technology Dharwad, Dharwad 580011, India (e-mail: prasanna@iitdh.ac.in).

This article has supplementary downloadable material available at https://doi.org/10.1109/TASLP.2022.3164199, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2022.3164199


Music Spectrogram (MS)

(a)

Speech Spectrogram (SS)

(b)

Speech+Music Spectrogram (SMS)

(c)

MS-Harmonic ($\hat{H}_{music}$)

(d)

SS-Harmonic ($\hat{H}_{speech}$)

(e)

SMS-Harmonic ($\hat{H}_{speech+music}$)

(f)

MS-Percussive ($\hat{P}_{music}$)

(g)

SS-Percussive ($\hat{P}_{speech}$)

(h)

SMS-Percussive ($\hat{P}_{speech+music}$)

(i)

Fig. 1. This figure illustrates the spectrograms of (a) music, (b) speech, and (c) speech+music mixed at 0 dB, along with their harmonic (2^{nd} row), and percussive (3^{rd} row) decompositions. It may be noted that speech+music spectrograms carry the combined striations of both the component signals.

either isolated or overlapped. Albayzín-2014 ASE provided a more general and realistic database than those used in the Albayzín-2010 and 2012 ASE. The submitted systems used two distinct approaches for the task [17]. About half of the submissions followed the segmentation-and-classification strategy using techniques like Bayesian Information Criterion. The remaining systems employed a segmentation-by-classification strategy whereby models of individual classes are used to generate predictions that are smoothed in a post-processing step. GMMs followed by Hidden Markov Models (HMM, henceforth) were a popular choice for this strategy. The most common feature choices were Mel Frequency Cepstral Coefficients (MFCC, henceforth), Perceptual Linear Prediction coefficients, short-term energy, and other standard spectral features. GMMs, i-vectors, HMM, Logistic Regression, and SVM were popularly used for classification. The participating authors observed that the presence of noise class increased the difficulty of the task. Overlapping segments of speech, music, and noise were mostly confused with the pair-wise overlaps. Many recent works have also tried to solve the Albayzín-2014 ASE task using deep-learning-based approaches. Gimeno et al. [18] employed a sequence of Bidirectional Long Short Term Memory units to segment audio sequences followed by classification.

Another popular challenge, known as the Music Information Retrieval Evaluation eXchange (MIREX, henceforth), tasked the detection of speech and music in its 2015 and 2018 editions. Classification-and-segmentation and classification-by-segmentation were the major approaches used in MIREX challenges as well. Frame energy, zero-crossing rates, spectral features, MFCCs, Chroma-based features were popular features used in MIREX 2015. Most of the authors adopted Mel-spectrograms as the input feature in MIREX 2018. Classification systems based on expert systems, SVM, Restricted Boltzmann Machines, Logistic Regression, Random Forests, and shallow Neural Networks were popular in MIREX 2015. However, most systems switched to some form of deep-learning-based classifier in MIREX 2018.

Previous works in speech and music detection have used Mel-scaled spectrograms (MS, henceforth) or its derivatives as the principal feature [14], [18], [19]. Few submissions to the MIREX challenges have explored Constant-Q Transform spectrograms and Periodograms. Other features used were self-similarity matrix [11], spectral tracking [20], Continuous Frequency Activations [21], Pyknograms [15], Mel Frequency Cepstral Coefficients [22] and standard tempo-spectral features. 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. Fig. 1(a)-(c) show the spectrograms of music, speech, and speech overlapped with music (speech+music, henceforth) at 0 dB Speech-to-Music Ratio (SMR, henceforth), respectively. The harmonics in speech have a wavy structure, while music harmonics are relatively more stable (horizontally linear). Percussive components characterized by an impulse like vertical striations are present more in music [23] than speech. It might be challenging for an automatic feature learning system to isolate and learn the class-specific patterns from a combined representation like spectrogram. Separately presenting the harmonic and percussive information might help in better learning the discrimination. This idea is the main motivation of using

Harmonic-Percussive Source Separation (HPSS, henceforth) to compute features previously unexplored in this task. HPSS based features have been successfully used previously in Jazz solo instrument classification [24], time-scale modifications of music signals [25], and music genre classification [26]. We believe that HPSS representations might perform better in the current task as well.

Another contribution of this work is exploring the Multi-Task Learning (MTL, henceforth) framework in the current task. The motivation of using MTL in the current task can be justified using the following reasons. First, MTL has been successfully explored previously in different speech and audio processing applications with considerable success. Notable examples include speech recognition [27], speaker verification [28], harmony recognition of symbolic music [29], analysis of acoustic scenes and events [30], Speech synthesis [31], End-to-end speech translation [32], Neural machine translation [33] and several others. Second, the current task has some auxiliary information like mixing SMR ratio that can be used for additional supervision in training the detection model. Third, a model trained using this framework learns to perform multiple tasks for any given input. For memory-constrained systems [34], such a model will be extremely beneficial. Fourth, training networks with highly related auxiliary tasks and sufficient noise levels inherent in the data can improve generalization capabilities [35]. In addition to the traditional MTL framework, this work also explores cascaded information in the MTL framework that is found to be beneficial in literature [36]–[38].

This work has four principal contributions. First, HPSS based features are explored in the task of detecting speech overlapped with music. Second, traditional and cascaded-information MTL frameworks are explored to enhance classifier performances. Third, the effect of challenging mixing SMRs (say -5dB and 20 dB) [5], [6], [9], [10] on the speech+music detection performance is analysed. Fourth, the effectiveness of the proposed system when employed to real signals containing isolated or overlapped speech and music is discussed.

The rest of the paper is organized in the following manner. Section II discusses the proposed approach for the detection of speech overlapped with music. A brief description of the procedure for HPSS is provided in Section II-A. An analysis of class separability provided by the proposed features is provided in Section II-B. The proposed MTL design is explained in Section II-C. The experiments performed and results obtained are discussed in Section III. Finally, conclusions and possible future directions of extending this work are discussed in Section IV.

II. Proposed Feature and Network Architectures

This work explores representations obtained from HPSS as features to detect speech+music signals. Moreover, a classifier designed in the MTL framework might leverage additional implicit information associated with the underlying task and perform better than a traditional classifier. The following subsections describe the methodology for HPSS decomposition and the design of proposed models in the MTL framework.

A. Harmonic Percussive Source Separation

This work uses the HPSS decomposition method proposed by Fitzgerald et al. [39]. Let, S be a complex-valued Discrete Fourier Transform (DFT, henceforth) based spectrogram, and $\|S\|$ be the magnitude spectrogram derived from S. The spectrogram $\|S\|$ can be further decomposed into separate harmonic and percussive components. A harmonic enhanced spectrogram (H) is computed by median filtering the rows of $\|S\|$ with a window size of l_{harm}. Similarly, a percussion enhanced spectrogram (P) is computed by median filtering the columns of $\|S\|$ with a window size of l_{perc}. The equations for computing H and P are as follows.

$$H[i, 1:n_t] = \text{median}\left(\|S\|\,[i, 1:n_t], l_{\text{harm}}\right)$$
$$P[1:n_f, j] = \text{median}\left(\|S\|\,[1:n_f, j], l_{\text{perc}}\right)$$

where, $i = 1, \ldots n_f$ are the indices of n_f frequency bins in S, $j = 1, \ldots n_t$ are the indices of n_t frames in S, and median($\bullet$) signifies the median filter. Masks are generated using these enhanced spectrograms that are multiplied with the original spectrogram S to obtain the respective decompositions. Two variants of these masks can be computed, hard masks and soft masks. This work uses soft masks for decomposition. The soft masks are based on Wiener filtering and computed as follows.

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

where, power r can be either 1 or 2. In this work, $r = 2$ is considered. These masks are multiplied element-wise ($\bigotimes$) to the original complex spectrogram (S) to generate the harmonic decomposition ($\hat{H} = \mathbf{M}_H \bigotimes S$) and percussive decompositions ($\hat{P} = \mathbf{M}_P \bigotimes S$). For a detailed treatment of the method, the reader is encouraged to refer [39]. Fig. 1(d)-(f) show $\hat{H}$ and Fig. 1(g)-(i) show $\hat{P}$ of the spectrograms in Fig. 1(a)-(c). It can be observed from the figures that $\hat{H}$ can clearly capture harmonic striations of the signal, while $\hat{P}$ contains the signal's percussion patterns.

B. Class-Separability Provided by HPSS

This subsection describes a method employed to gauge the enhancement in class separability induced by HPSS. The linear harmonics in music might span only a few adjacent rows (along the frequency dimension) in the spectrogram. In contrast, harmonics in speech may span over many rows in the spectrogram because of their wavy nature. Therefore, the harmonics' rows might have a localized energy distribution over successive audio frames in music but not in speech. Note that the rows without any harmonics in either signal's spectrogram would mainly contain background information and not provide much separability. Alternatively, in the case of spectrogram columns containing percussive striations, the frame's energy is almost evenly distributed across all frequency bins. For non-percussive frames, the frame energy is contained only in a small number of frequency bins. Thus, energy distribution might be localized in spectrogram

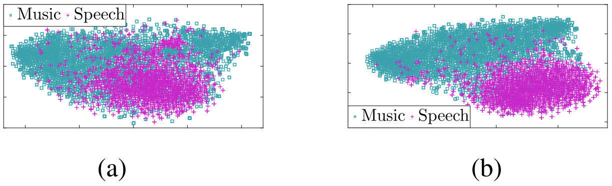


Fig. 2. The t-SNE plots illustrate the distribution of skewness vectors. The subfigures are generated by concatenating $\mathbf{r}_{skew}$ and $\mathbf{c}_{skew}$ vectors computed from $\|S\|$ (shown in (a)), and $\hat{H}$ and $\hat{P}$ (shown in (b)). It can be observed that harmonic and percussive decompositions can improve the class separability of speech and music. For more details, refer Section II-B.

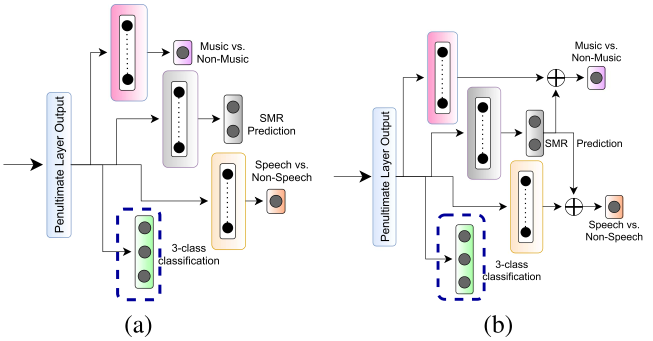


Fig. 3. Illustrating the proposed design of (a) Traditional MTL and (b) Cascaded-information MTL frameworks.

columns containing percussion and widely distributed otherwise. Hence, it may be expected that the respective row-energy and column-energy distributions of speech and music signals would be different.

The nature of row-energy and column-energy distributions are studied in this work by computing their skewness across rows ($\mathbf{r}_{skew}$, henceforth) and columns ($\mathbf{c}_{skew}$, henceforth) using the *Scipy* [40] python library. The motivation and methodology for computing these measures are described in Section I of the supplementary material (Supp. Mat., henceforth). The spectrograms are Mel-scaled with 21 filters for this experiment to smooth along the frequency axis. The distributions of row and column energies using skewness measure are visualized in Fig. 2 using 2-dimensional embeddings generated using the t-SNE [41] algorithm. Fig. 2(a) shows t-SNE visualizations generated with the concatenated row and column skewness vectors computed from the Mel-spectrogram. Similarly, Fig. 2(b) shows the visualization generated by concatenating the row and column skewness vectors computed from Mel-harmonic and Mel-percussive spectrograms, respectively. The representation for Mel-spectrogram has much overlap between the classes, even though it inherently contains both harmonic and percussive information. However, separately computing row and column skewness vectors from the Mel-harmonic and Mel-percussive decompositions enhances the class separability, as can be observed in Fig. 2(b). Since the HPSS based decomposition enhances the class separability of speech and music signals, it might also be useful in detecting speech+music signals mixed at various SMR levels.

C. Multi-Task Learning Framework

Individual models are trained for different problems in the single-task learning (STL, henceforth) framework. The STL models with sufficient parameters to approximate the underlying distribution and learned from large enough datasets are known to perform reasonably well. For example, STL architectures proposed in [1]–[4] have been successfully used for speech and music detection. However, in most practical cases, the performance of these models is constrained by the complexity of the underlying task and generalizability to unseen data. A Multi-Task Learning (MTL) framework attempts to overcome these problems by learning multiple closely related subproblems using a single model. Such a technique aids the learning of the main task by joint supervision of related auxiliary targets. This work explores the MTL framework for improving the speech+music detection performance.

This work's main task is the 3-class classification of isolated speech, music, and overlapped speech+music. This work explores both the traditional MTL framework and a cascaded-information variant of MTL. The proposed models designed in the traditional MTL framework involve simultaneous training of three auxiliary tasks (AT, henceforth) to aid learning of the main task. First, a speech vs. non-speech classifier (AT_S, henceforth) learns to differentiate between speech and non-speech. Music and speech+music are considered as non-speech for AT_S. Second, the music vs. non-music classifier (AT_M, henceforth) learns to discriminate between music and non-music. Here, speech and speech+music are considered non-music. The AT_S and AT_M are learned using a binary cross-entropy loss function. Third, a regression-based task (AT_R, henceforth) tries to estimate the SMR of a given audio signal. The AT_R task is trained using a $l2$ loss-based optimization scheme. The target output of AT_R task, $\mathbf{t} = [\mathbf{t}_M, \mathbf{t}_S]$, is a 2-dimensional vector that indicates the scaling factor of music ($\mathbf{t}_M$) with respect to speech ($\mathbf{t}_S$) in the input signal. Let the set of speech, music, and speech+music signals be denoted by Γ_S, Γ_M, and Γ_{SM}, respectively. For a given input signal $x[n]$ and an SMR of v dB, the AT_R task target $\mathbf{t}$ is computed as follows.

$$\mathbf{t} = \begin{cases} [0,1]^T, & \text{if, } x[n] \in \Gamma_S \\ [1,0]^T, & \text{if, } x[n] \in \Gamma_M \\ [10^{-\frac{v}{10}}, 1]^T, & \text{if, } \{x[n] \in \Gamma_{SM}\} \wedge \{v \geq 0\text{dB}\} \\ [1, 10^{\frac{v}{10}}]^T, & \text{if, } \{x[n] \in \Gamma_{SM}\} \wedge \{v < 0\text{dB}\} \end{cases}$$

The proposed traditional MTL architecture is shown in Fig. 3(a). The hyper-parameters of each auxiliary sub-network have been tuned over a subset of the training data. The number of hidden layers were varied over $[1, 2, 3]$, while the number of hidden neurons were varied over $[16, 32, 64, 128]$. For the AT_S and AT_M tasks, both *hinge* loss and *binary-crossentropy* loss were tested. The final tuned sub-networks of all the auxiliary tasks consist of a single fully connected hidden layer of 16 nodes with *ReLU* activation. For regularization, the hidden layer is equipped with *BatchNormalization* and a *Dropout* fraction of 0.4. The AT_S and AT_M have a single neuron in the output layer that is *Sigmoid* activated with *binary-crossentropy* loss function.

The AT_R task has two nodes in its output layer with *linear* activation and $l2$ loss function. In a similar manner, the proposal for cascaded-information MTL variant is shown in Fig. 3(b). The 2-dimensional output from AT_R is concatenated with hidden layer outputs of the AT_S and AT_M tasks. This way, the cascading of predicted SMR values might aid the speech vs. non-speech and music vs. non-music auxiliary tasks.

The proposed models use four separate loss functions. Let $\mathcal{L}_s$ be the loss function of AT_S, while y_s and $\hat{y}_s$ be its respective ground-truth and predicted outputs. Let, $\mathcal{L}_m$ be the loss function for the AT_M task with respective true and predicted outputs as y_m and $\hat{y}_m$. The AT_R branch estimates the SMR proportion of speech and music in an input signal. The AT_R regression task is learned using a $l2$ loss $\mathcal{L}_{smr}$. These losses are defined as follows:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{k=1}^{N}\left(y_s[k]\log\left(\hat{y}_s[k]\right)+(1\text{-}y_s[k])\log\left(1-\hat{y}_s[k]\right)\right)$$

$$\mathcal{L}_m = -\frac{1}{N}\sum_{k=1}^{N}\left(y_m[k]\log\left(\hat{y}_m[k]\right)+(1\text{-}y_m[k])\right.$$
$$\left.\log\left(1-\hat{y}_m[k]\right)\right)$$

$$\mathcal{L}_{smr} = \frac{1}{N}\sum_{k=1}^{N}\left(y_{smr}[k]\text{-}\hat{y}_{smr}[k]\right)^2$$

Here, $k = 1, \ldots N$ are the samples in a training batch of size N. Also, y_{smr} and $\hat{y}_{smr}$ are the respective ground-truth and predicted values of the SMR proportion. The final 3-class classification loss function $\mathcal{L}_c$ for the main task is learned using a *categorical-crossentropy* loss function and is defined as

$$\mathcal{L}_c = -\frac{1}{N}\sum_{k=1}^{N}\sum_{\vartheta=1}^{3}\left(y_c^{(\vartheta)}[k]log\left(\hat{y}_c^{(\vartheta)}[k]\right)\right)$$

Here, $y_c^{(\vartheta)}$ are the one-hot encoded ground truth and $\hat{y}_c^{(\vartheta)}$ are the predicted outputs of the ϑ^{th} output neuron of the main task network. The total loss $\mathcal{L}_{Total}$ can be defined as the weighted sum of these four losses mentioned above. The MTL-based model is learned by minimizing $\mathcal{L}_{Total}$.

$$\mathcal{L}_{Total} = w_s.\mathcal{L}_s + w_m.\mathcal{L}_m + w_{smr}.\mathcal{L}_{smr} + w_c.\mathcal{L}_c$$

The loss weights w_s, w_m, w_{smr} and w_c can be varied to obtain optimal loss minimization for a given task. Setting equal weights for all losses was found to be best for this work (see Section II, Supp. Mat.).

III. Experiments and Results

The proposed approach is validated using various experiments as described in this section. The music and speech data from the MUSAN - A Music, Speech, and Noise corpus [42] (≈ 102 hours) are used as experimental data. The MUSAN dataset is popularly used in a variety of speech and audio processing tasks, like speech music detection [14], general-purpose audio representation learning [43], music relative loudness estimation [12], sound source separation works [44], speech enhancement [45] voice activity detection in the wild [46], etc. For the initial experiments, data for the speech+music class is generated synthetically. However, in a later subsection (see Section III-H), results on real mixed speech and music signals are also reported to establish the efficacy of the proposed approach. This work uses three-fold cross-validation. Each experiment is run for three iterations, considering one of the folds as a test set and the remaining folds as the training set. Results are reported in the form of mean (μ) ± standard deviation (σ) computed over the three test runs. The performance metrics used in this work are accuracy (Acc, henceforth), precision (Prec, henceforth), recall (Rec, henceforth), and $F1$-score ($F1$, henceforth) (see Section III, Supp. Mat.). The process of generating the mixed signals is described next.

A. Synthetic Speech+Music Signal Generation

The MUSAN dataset contains ≈ 42 hours of music and ≈ 60 hours of speech. The available music and speech files are divided into three (almost equal) folds. Music files in the MUSAN dataset are provided with genre annotations, while many speech files have gender information. Such available information was considered while grouping the files so that similar distribution of music and speech could be maintained across the folds. The speech+music data for each fold was created by mixing random pairs of music and speech files from the same fold. Files for mixing were chosen so that files from speech class (more in number) were sampled without replacement, while some files from the music class (less in number) were sampled at most twice. For simulating real mixed signals, all integer SMR levels in the range -5dB to 20 dB with a step of 1 dB were considered. Here, SMR is defined by considering speech as the reference signal. It was ensured that for the speech+music data in each fold, an almost equal number of file pairs were mixed at each SMR level. The division of files from the MUSAN dataset into folds and speech+music file pairs with SMR annotations used in this work have been shared publicly (along with the codes[1]).

B. Baseline Methods for Comparison

The proposed approach is validated using four state-of-the-art speech music detection methods from literature. First, the method proposed by Doukhan et al. [1] ($B1$, henceforth) uses 21-Mel spectrogram input (MS, henceforth) with a CNN to classify speech and music signals. The $B1$ classifier has four convolutional layers followed by four fully-connected layers (512 neurons each) leading to ≈1.4 million parameters. Second, the proposal of Papakostas et al. [2] ($B2$, henceforth) uses a CNN classifier with grayscale spectrogram input (S, henceforth) to classify speech and music. The $B2$ classifier consists of three convolutional layers and two fully-connected layers (4096 neurons each), leading to ≈44 million parameters. Third, Lemaire et al. [3] ($B3$, henceforth) proposed a non-causal Temporal Convolution Network (TCN, henceforth) architecture with log-scaled 80-Mel spectrogram input (LMS, henceforth) to detect speech and music in radio broadcasts. The $B3$ classifier consists of one TCN unit of three residual block stacks that

[1] [Online]. Available: https://github.com/mrinmoy-iitg/SM_HPSS_MTL

TABLE I
ILLUSTRATING THE PERFORMANCES OF BASELINE METHODS USED FOR COMPARISONS IN THIS WORK.

			Music			Speech			Speech+Music			Avg.
Feature	**Classifier**	**Acc**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**F1**
MS	$B1$	71.68 ±2.28	49.5 ±3.37	77.64 ±11.28	60.19 ±4.29	81.03 ±1.35	87.83 ±6.03	84.19 ±2.21	75.18 ±4.66	51.86 ±4.18	61.32 ±3.9	68.67 ±3.06
S	$B2$	67.39 ±2.02	40.26 ±3.08	61.4 ±8.82	48.59 ±5.04	78.31 ±4.4	93.99 ±4.3	85.38 ±3.52	74.85 ±1.18	47.36 ±5.25	57.88 ±3.86	64 ±2.65
LMS	$B3$	79.23 ±1.37	62.66 ±2.72	84.4 ±0.85	71.91 ±2	80.41 ±2.81	88.86 ±3.07	84.35 ±0.12	83.57 ±0.7	64.65 ±5.86	72.8 ±3.78	76.33 ±2.08
LS	$B4$	56.86 ±10.41	67.45 ±12.02	46.22 ±29.72	49.16 ±18.32	90.92 ±11.82	40.98 ±33.76	48.58 ±36.4	59.19 ±7.47	84.29 ±9.61	69.04 ±4.81	55.67 ±13.32
LS	$B4$ (NoFC)	69.95 ±16.97	48.59 ±13.58	88.06 ±11	61.09 ±9.6	89.02 ±8.22	63.05 ±46.68	63.93 ±39.83	81.08 ±18.57	59.67 ±15.59	68.01 ±14.05	64.35 ±20.52

add up to ≈0.11 million parameters. The optimal $B3$ classifier was obtained by tuning the hyperparameters mentioned by the authors [3] on a subset of this work's training data. Fourth, a CNN with 64 trainable Mel-scale convolutional filters with log-spectrogram input (LS, henceforth) was proposed by Jang et al. [4] ($B4$, henceforth) for music detection. With three convolution layers in addition to the Mel-scale one and two fully connected layers (2048 and 1024 neurons, respectively), the number of trainable parameters in the $B4$ classifier is ≈21 million. Among all the baselines considered, $B1$ uses the smallest and most challenging context window size of 695 ms. Hence, all results reported in this work are computed for 695 ms windows. However, the effect of varying context window size is also analyzed (see Section IV, Supp. Mat.). All the baselines were proposed as binary classification tasks. However, this work is designed as a 3-category classification task where there is considerable overlap between the classes, thereby increasing the overall complexity. Hence, the performances reported for the baselines are lesser than their original binary classification performances. However, comparable results have been obtained for binary speech vs. music classification performance of all baseline methods computed using the setup of this work (see Section V, Supp. Mat.).

C. Feature Computation Details

The spectrograms in this work are computed using a short-term frame size of 25 ms and a frameshift of 10 ms. A heuristic-based energy threshold is used to remove silences in audio files. The HPSS decomposition of spectrograms is performed using the *Librosa* [47] python library. Classifiers are designed using the Keras [48] and Tensorflow [49] libraries. The median-filtering window sizes used for HPSS decomposition in this work are l_{harm}=21 and l_{perc}=11 (tuned experimentally, see Section VI, Supp. Mat.). The spectrogram and its harmonic-percussive decompositions have been Mel-scaled using n_{mels}=120 filters (tuned experimentally, see Section VI, Supp. Mat.). The classifiers are trained using feature patches with $n_t = 68$ frames (695 ms) as the temporal context. Patches are extracted with a shift of 68 frames. The models are trained with a batch size of 48 and a maximum epoch of 50. An early-stopping criterion has been used while training to avoid overfitting. Early-stopping is a regularization approach that monitors the validation loss and terminates the model training if there is no improvement for consecutive 5 epochs. The best model with the lowest validation loss obtained in the process is retained. All codes used for performing experiments in this work have been shared publicly (see Section III-A). The results are discussed in the following subsections.

D. Performance of Harmonic-Percussive Features

The 3-class classification performance of $B1$, $B2$, $B3$ and $B4$ are tabulated in Table I. The best average $F1$-score of 76.33 ±2.08 is obtained for the $B3$ baseline with the LMS feature. The $B4$ baseline with LS input seems to perform the poorest. However, it was observed that the $B4$ model overfits the training data. The reason seemed to be too many training parameters. Reducing the number of parameters by removing the fully-connected (FC, henceforth) layers in the $B4$ architecture greatly improved its performance (see $B4$ (NoFC) in Table I). The best performing $B3$ model is used in further experiments for developing the best feature and classifier combination. Later, performance improvements obtained with all baselines using the proposed methods are described in Section III-F.

Performance of the proposed HPSS decomposed features in the current 3-class classification task with the best baseline $B3$ is tabulated in Table II. The performance of $B3$ was further improved by setting n_{mels}=120 (first row in Table II). The performance of $B3$ with log-scaled 120-Mel harmonic spectrogram input (LMHS, henceforth) computed with tuned l_{harm}=21 is listed in the second row. In the third row, the performance of $B3$ with log-scaled 120-Mel percussive spectrogram input (LMPS, henceforth) computed with tuned l_{perc}=11 is provided. In the last row, the performance of $B3$ with the early-fusion (EF, henceforth) of the LMHS and LMPS features concatenated along the feature dimension (LMHPS-EF, henceforth) is listed. The EF strategy performed the best among intermediate-fusion and late-fusion strategies explored in this work (see Section VII, Supp. Mat.). It can be observed that the LMHS feature does not perform better than LMS. A possible reason might be that Mel-scaling reduces the resolution of high-frequency harmonics that hampers discrimination. LMPS provides almost similar results as that of LMS, although with a lower standard deviation. However, the LMHPS-EF significantly improves upon the baseline LMS performance (average $F1$-score) by around 7%. The LMHPS-EF feature also performs better for each of the individual classes. Such performances support the proposal of this work that HPSS decomposition of the spectrogram feature

TABLE II
ILLUSTRATING THE EFFECT OF USING OPTIMIZED NUMBER OF MEL-FILTERS ALONG WITH HARMONIC AND PERCUSSIVE FEATURES WITH THE BEST PERFORMING BASELINE ($B3$). HERE, n_{mels}=120, l_{HARM}=21, AND l_{PERC}=11 ARE USED AS THE TUNED PARAMETERS.

			Music			Speech			Speech+Music			Avg.
Feature	**Classifier**	**Acc**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**F1**
LMS	$B3$	81.75 ±1.01	63.92 ±3.26	87.86 ±0.54	73.97 ±2.14	80.91 ±4.65	90.95 ±1.84	85.54 ±1.87	86.88 ±0.57	64.83 ±8.06	74.04 ±5.3	77.67 ±3.21
LMHS	$B3$	79.02 ±5.02	52.61 ±8.02	87.14 ±1.77	65.43 ±6.63	83.59 ±4.59	97.11 ±1.25	89.78 ±2.1	88.91 ±2.74	52.24 ±15.48	65.12 ±12.55	73.33 ±6.81
LMPS	$B3$	81.77 ±1.31	56.48 ±3.49	86.29 ±3.2	68.17 ±1.66	89.95 ±5.88	95.26 ±4.03	92.36 ±1.87	89.06 ±1.26	64.16 ±4.11	74.52 ±2.67	78.33 ±1.53
LMHPS-EF	$B3$	86.87 ±1.38	70.67 ±5.06	86.85 ±4.38	77.74 ±1.18	90.47 ±3.1	97.1 ±1.13	93.64 ±1.41	92.06 ±2.13	77.82 ±6.06	84.2 ±2.67	85.19 ±1.75

TABLE III
ILLUSTRATING THE EFFECT OF MODIFYING $B3$ WITH TRADITIONAL MTL-BASED AND CASCADED-INFORMATION MTL-BASED (C-MTL) FRAMEWORKS, AS SHOWN IN FIG. 3.

			F1			
	Feature	**Acc**	**music**	**speech**	**speech+music**	**Avg.**
$B3$-MTL	LMS	81.79 ±2.59	75.63 ±3.28	85.98 ±2.82	77.22 ±2.74	79.33 ±2.52
	LMHS	83.27 ±1.08	71.28 ±4.39	90.93 ±0.77	76.82 ±3.86	79.67 ±2.52
	LMPS	85.44 ±0.08	73.42 ±1.29	94.39 ±1.38	81.79 ±0.37	83.20 ±0.24
	LMHPS-EF	89.12 ±1.67	82.74 ±0.95	93.76 ±1.92	87.71 ±1.92	88.07 ±1.59
$B3$-C-MTL	LMS	82.72 ±1.72	75.80 ±4.01	86.25 ±0.94	77.46 ±3.48	79.67 ±2.52
	LMHS	84.62 ±1.97	74.45 ±3.68	91.61 ±1.55	80.59 ±3.94	82.33 ±3.06
	LMPS	85.11 ±1.99	72.74 ±2.64	93.91 ±1.84	81.06 ±3.65	82.67 ±2.52
	LMHPS-EF	90.09 ±0.66	81.54 ±0.75	94.49 ±0.79	86.95 ±0.51	87.33 ±0.58

helps in better detection of speech, music, and speech+music signals.

E. Performance of MTL Framework

The second contribution of this work is an exploration of the popular MTL framework in the current task. This work explores traditional MTL architecture [28] and a cascaded-information MTL variant [38] (see Section II-C). Table III lists the performances of best baseline classifier $B3$ whose architecture is modified according to traditional MTL framework ($B3$-MTL, henceforth) and the cascaded-information MTL framework ($B3$-C-MTL, henceforth), as shown in Fig. 3. Only class-wise $F1$-score is listed here (detailed results are provided in Section VIII, Supp. Mat.). With the $B3$-MTL architecture, the performance of the baseline LMS feature improves by around 2%. Similarly, the performances of proposed features LMHS, LMPS, and LMHPS-EF improve by around 6%, 5%, and 3%, respectively. The $B3$-C-MTL architecture also improves the performances of LMS, LMHS, LMPS, and LMHPS-EF features by around 2%, 9%, 4%, and 2%, respectively. However, the overall best average $F1$-score of 88.07 ±1.59 for the 3-class classification task is obtained with the $B3$-MTL architecture with the LMHPS-EF feature. Performances of all three classes have also improved significantly with the LMHPS-EF feature and $B3$-MTL classifier. Hence, it can be inferred that the use of the MTL framework in the current task helps in learning more generalizable representations for distinguishing the different audio classes, thereby significantly improving the overall performance.

F. HPSS Features and MTL Framework With Baselines

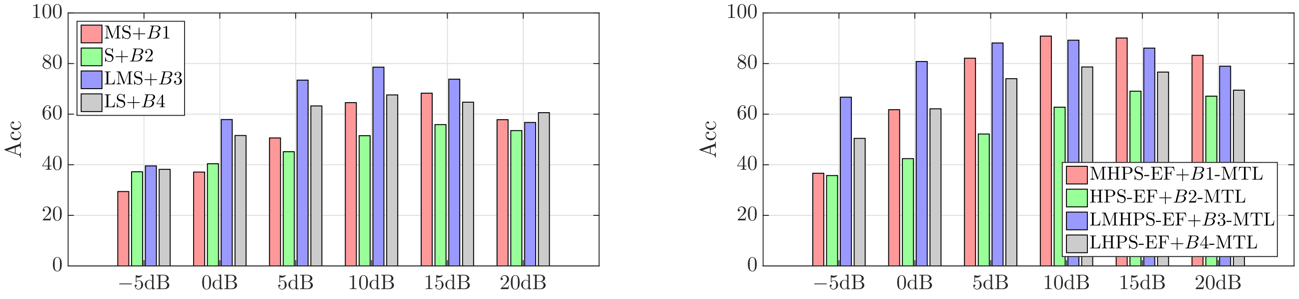
The improvements obtained for $B3$ with the usage of harmonic-percussive features and MTL-based classifier modification motivate the application of these changes to other baselines. Each baseline is fed with the EF of harmonic and percussive features with the respective preprocessing of each baseline. Moreover, all the baselines are equipped with the MTL modification that provided the best performance previously. Thus, $B1$ is modified to $B1$-MTL and provided the EF of 120-Mel harmonic and percussive spectrograms (MHPS-EF, henceforth) as input. $B2$ is modified to $B2$-MTL and given the EF of harmonic and percussive spectrograms (HPS-EF, henceforth) as input. Lastly, $B4$ is converted to $B4$-MTL and trained with EF of log-scaled harmonic and percussive spectrograms (LHPS-EF, henceforth) as input. In Table IV, the improved performances of all the baselines used in this work are provided. It can be observed that there are significant improvements to the performances of all the baselines. $B3$-MTL classifier with LMHPS-EF feature performs the best, while the $B2$-MTL provides a minor improvement. The relatively poor performance of $B2$-MTL may be improved by tuning its architecture for the current task and dataset.

G. Performance at Challenging SMR Levels

An important goal of this work is the detection of speech+music signals in challenging SMR scenarios. All performances reported till now are computed for speech+music signals mixed at different SMR levels in the aforementioned range $[-5, \ldots 20]$ dB. It has been encouraging to observe that the current proposal performs quite well in the presence of mixed signals at various SMR levels. However, it is also important to assess the capability of the proposed approach in detecting speech+music signals at specific challenging SMR levels. In this context, results are computed at -5dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. Here, -5dB and 20 dB are the most challenging cases since the music component is 5 dB louder than the speech component in the former, while speech is 20 dB louder in the later. In such cases, the speech+music signal is likely to be confused with the louder component. The performance in such cases will highlight the robustness of the proposed approach. Just for this experiment, the SMR annotations that were fixed for every

TABLE IV
ILLUSTRATING THE IMPROVEMENT IN PERFORMANCES OF ALL 4 BASELINES WITH THE BETTER HARMONIC-PERCUSSIVE FEATURE AND MTL MODIFICATION OF THEIR RESPECTIVE CLASSIFIER ARCHITECTURES.

			Music			Speech			Speech+Music			Avg.
Feature	**Classifier**	**Acc**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**F1**
MHPS-EF	$B1$-MTL	89.67 ±2.92	67.03 ±6.92	96.27 ±2.43	78.84 ±4.5	94.36 ±2.76	99.25 ±0.98	96.72 ±1.28	97.4 ±2.06	74.54 ±8.05	84.23 ±4.81	86.6 ±3.51
HPS-EF	$B2$-MTL	69.60 ±1.89	45.79 ±0.39	63.77 ±9.23	53.14 ±3.24	79.47 ±5.7	90.27 ±3.98	84.35 ±1.65	74.71 ±2.05	54.87 ±0.48	63.27 ±1.04	66.92 ±1.71
LMHPS-EF	$B3$-MTL	89.12 ±1.67	78.31 ±1.42	87.8 ±3.35	82.74 ±0.95	90.08 ±4.63	97.91 ±1.61	93.76 ±1.92	93.26 ±0.72	82.81 ±3.03	87.71 ±1.92	88.07 ±1.59
LHPS-EF	$B4$-MTL (NoFC)	73.8 ±6.93	59.03 ±7.29	63.3 ±22.54	58.89 ±6.18	83.88 ±4.45	87.57 ±12.61	85.31 ±6.24	77.27 ±10.43	69.68 ±10.35	72.3 ±2.18	72.17 ±4.04



(a) (b)

Fig. 4. This figure illustrates the performance of (a) baseline models, and (b) their modified versions, at varying SMR levels. Accuracy values are reported in percentage.

speech+music files (see Section III-A) were substituted with the particular SMR level being tested (among $[-5, 0, 5, 10, 15, 20]$ dB). The mean recall over 3-folds for detecting speech+music signals is reported.

In Fig. 4, the performance of all the baseline classifiers have been compared with their proposed modifications. The obtained results indicate that the performance of all baselines has improved with the proposed modifications at all six chosen SMR levels. Most significant improvements can be observed for the $B1$ and $B3$ baselines. The performances of all systems peak around 10 dB and expectedly drop towards the challenging SMR cases. The $B3$ model with the proposed modifications performs much better than the others at challenging SMRs. Also, the performance at -5dB is poorer than that at 20 dB for all systems. This result indicates that the models are confused more in the presence of loud music than speech. This observation might be attributed to the fact that speech is a relatively low-frequency signal [50] when compared to music. Thus, only a limited range of frequencies might be dominated by loud speech in a speech+music signal, enabling better detection. In comparison, loud music might be dominating a more extensive spectral range which creates more confusion. Nonetheless, the overall performances obtained at challenging SMR levels establish the efficacy of the current proposal.

H. Performance With Real Mixed Signals

The performances reported so far were computed over synthetically generated speech+music signals. However, the efficacy of the proposed approach can be gauged when tested with real speech and music signals present as isolated and overlapping mixtures. Schlüter et al. [51] created a dataset (*DAFx12-dataset*, henceforth) of recorded Swiss and Austrian radio broadcasts. The authors manually annotated the recordings into speech/non-speech and music/non-music segments. The *DAFx12-dataset* consists of around 28 hours of music, 8 hours of speech, and 5 hours of speech+music segments. The dataset is divided into a training set of around 15 hours, a validation set of 6 hours, two test sets of 9 hours (Swiss recordings), and 12 hours (Austrian recordings). The reader is encouraged to refer [51] for more details about the *DAFx12-dataset*.

Schlüter et al. [51] trained two separate classifiers to detect speech and music separately. Following their approach for a fair comparison, two separate classifiers have been trained in this work to evaluate the proposed approach on the *DAFx12-dataset*. For generalizaton purposes, silence removal was not performed for the *DAFx12-dataset*. The $B3$-MTL model trained on the LMHPS-EF feature over 695 ms context was used in this experiment in a transfer learning mode. For the music detection classifier ($B3$-MTL-Mu, henceforth), except for the music/non-music output, others were stripped off from the $B3$-MTL model. The remaining weights in the $B3$-MTL-Mu model were initialized with those from the trained $B3$-MTL model. The weights were subsequently tuned over the training set of the *DAFx12-dataset*. Similarly, all but the speech/non-speech output were stripped off from the $B3$-MTL model to create the speech detection classifier ($B3$-MTL-Sp, henceforth). The weights of $B3$-MTL-Sp were initialized and tuned similarly as $B3$-MTL-Mu. Both the models were tuned with a Nadam optimizer [52] with an initial learning rate of $2e^{-3}$. The previously mentioned early-stopping criterion was also used.

TABLE V
PERFORMANCE OF THE PROPOSED APPROACH ON REAL SIGNALS FROM THE *DAFX12-DATASET* [51] ARE TABULATED HERE. BASELINE RESULTS ARE QUOTED DIRECTLY FROM THE REFERENCE.

		Music/Non-music				Speech/Non-speech				Speech+Music/Non-(speech+music)			
Method	**Test set**	**Acc**	**Prec**	**Rec**	**$F1$**	**Acc**	**Prec**	**Rec**	**$F1$**	**Acc**	**Prec**	**Rec**	**$F1$**
Schlüter et al. [51]	Swiss	97.30	98.80	98.00	98.40	98.40	96.40	96.50	96.40	–	–	–	–
	Austrian	95.60	95.30	97.40	97.30	97.00	95.90	95.10	95.50	–	–	–	–
Proposed	Swiss	96.37	97.53	**98.20**	97.86	97.58	95.08	94.37	94.72	95.04	71.39	64.86	67.97
	Austrian	93.68	94.96	**97.48**	96.21	95.91	94.55	93.36	93.95	91.19	74.12	73.03	73.57

Table V lists the results of the proposed method over *DAFx12-dataset*. The baseline results [51] are directly quoted from the paper. It can be observed that the proposed approach provides performances comparable with the baseline for both the test sets. For the music/non-music detection, the proposed method provides a slightly better recall as well. The results for speech+music detection in Table V are generated by combining the predictions from both the $B3$-MTL-Mu and $B3$-MTL-Sp classifiers. The proposed approach provides a decent $F1$-score for detecting speech+music as well. However, small differences in the music and speech detection performances are observed that can be reasoned as follows. First, the baseline result was computed using $\approx$46 ms frame-size, $\approx$23 ms frame-shift, and context window of $\approx$923 ms. The proposed approach uses a frame-size, frame-shift, and context window of 25 ms, 10 ms, and 695 ms, respectively. Second, this work employs transfer learning to tune the model trained on the MUSAN dataset to *DAFx12-dataset*. Whereas, the model of Schlüter et al. [51] is trained from scratch on the *DAFx12-dataset*. Despite the slight differences, the results obtained are encouraging. It can be said that the proposed method of using harmonic-percussive spectrogram decomposition with the MTL framework can be an effective method of detecting not only isolated speech and music signals but also their mixtures.

IV. CONCLUSION

This work proposes the use of harmonic-percussive source separation (HPSS) to generate features that are shown to be better suited for detecting speech+music signals mixed at varying SMR levels. Baseline classifiers were modified in the traditional and cascaded-information multi-task learning (MTL) framework to improve the classification performance further. The HPSS features are found to outperform state-of-the-art features. The use of the MTL framework also aided in further improvement of the performances. Results have been reported over both synthetic speech+music data generated using the MUSAN dataset and real mixed data from the *DAFx12-dataset* [51].

This work can be extended in the following directions. First, the harmonic-percussive decomposition algorithm may be automated by learning the decomposition from data, possibly leading to better extraction of signal-specific harmonic and percussive components. Second, a soft parameter sharing MTL framework [53] may be explored in this task, which might allow the auxiliary tasks to learn better goal-specific feature representations, further improving the main task's performance. Third, the system may be extended to detect speech overlapped with other kinds of background sounds, in addition to music.

ACKNOWLEDGMENT

The authors would like to thank Dr. Jan Schlüter of Institute of Computational Perception, Johannes Kepler University Linz, for sharing spectrograms of the audio files of *DAFx12-dataset*.

REFERENCES

[1] D. Doukhan, E. Lechapt, M. Evrard, and J. Carrive, "INA's MIREX 2018 music and speech detection system," in *Proc. Music Inf. Retrieval Eval. eXchange (MIREX)*, 2018, pp. 1–2.

[2] M. Papakostas and T. Giannakopoulos, "Speech-music discrimination using deep visual feature extractors," *Expert Syst. Appl.*, vol. 114, pp. 334–344, 2018.

[3] Q. Lemaire and A. Holzapfel, "Temporal convolutional networks for speech and music detection in radio broadcast," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Nov. 2019, pp. 229–236.

[4] B.-Y. Jang, W.-H. Heo, J.-H. Kim, and O.-W. Kwon, "Music detection from broadcast contents using convolutional neural networks with a mel-scale kernel," *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, no. 1, pp. 1–12, 2019.

[5] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, vol. 2, 1997, pp. 851–854.

[6] F. Weninger, J. Feliu, and B. Schuller, "Supervised and semi-supervised suppression of background music in monaural speech recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 61–64.

[7] P. Vanroose, "Blind source separation of speech and background music for improved speech recognition," in *Proc. 24th Symp. Inf. Theory*, 2003, pp. 103–108.

[8] K. Lee and D. P. Ellis, "Detecting music in ambient audio by long-window autocorrelation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 9–12.

[9] T. Izumitani, R. Mukai, and K. Kashino, "A background music detection method based on robust feature extraction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 13–16.

[10] D. Y. Mohammed and F. F. Li, "Overlapped soundtracks segmentation using singular spectrum analysis and random forests," in *Proc. IEEE 2nd Int. Conf. Knowl. Eng. Appl.*, 2017, pp. 49–54.

[11] N. Tsipas, L. Vrysis, C. Dimoulas, and G. Papanikolaou, "Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination," *Multimedia Tools Appl.*, vol. 76, no. 24, pp. 25 603–25 621, 2017.

[12] B. Jia, J. Lv, X. Peng, Y. Chen, and S. Yang, "Hierarchical regulated iterative network for joint task of music detection and music relative loudness estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, no. 29, pp. 1–13, Oct. 2021.

[13] P. Gimeno, V. Mingote, A. Ortega, A. Miguel, and E. Lleida, "Partial auc optimisation using recurrent neural networks for music detection with limited training data," in *Proc. Interspeech*, 2020, pp. 3067–3071.

[14] S. Venkatesh *et al.*, "Artificially synthesising data for audio classification and segmentation to improve speech and music detection in radio broadcast," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 636–640.

[15] M. Bhattacharjee, S. R. M. Prasanna, and P. Guha, "Detection of speech overlapped with low-energy music using pyknograms," in *Proc. Nat. Conf. Commun.*, 2021, pp. 1–6.

[16] A. Ortega, D. Castan, A. Miguel, and E. Lleida, "The albayzin 2012 audio segmentation evaluation," in *Proc. IberSpeech*, Madrid, Spain, 2012, pp. 21–23.

[17] D. Castán *et al.*, "Albayzín-2014 evaluation: Audio segmentation and classification in broadcast news domains," *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 1, pp. 1–9, 2015.

[18] P. Gimeno, I. Viñals, A. Ortega, A. Miguel, and E. Lleida, "Multiclass audio segmentation based on recurrent neural networks for broadcast domain data," *EURASIP J. Audio, Speech, Music Process.*, vol. 2020, no. 1, pp. 1–19, 2020.

[19] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 131–135.

[20] T. Taniguchi, M. Tohyama, and K. Shirai, "Detection of speech and music based on spectral tracking," *Speech Commun.*, vol. 50, no. 7, pp. 547–563, 2008.

[21] K. Seyerlehner, T. Pohle, M. Schedl, and G. Widmer, "Automatic music detection in television productions," in *Proc. 10th Int. Conf. Digit. Audio Effects (DAFx'07)*, Bordeaux, France, vol. 10, 2007, pp. 1–8.

[22] D. Castán, A. Ortega, A. Miguel, and E. Lleida, "Audio segmentation-by-classification approach based on factor analysis in broadcast news domain," *EURASIP J. Audio, Speech, Music Process.*, vol. 2014, no. 1, pp. 1–13, 2014.

[23] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[24] J. S. Gómez, J. Abeßer, and E. Cano, "Jazz solo instrument classification with convolutional neural networks, source separation, and transfer learning," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 577–584.

[25] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 105–109, Jan. 2014.

[26] E. Tsunoo, G. Tzanetakis, N. Ono, and S. Sagayama, "Beyond timbral statistics: Improving music classification using percussive patterns and bass lines," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 1003–1014, May 2011.

[27] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 4835–4839.

[28] Z. Tan, M. Mak, and B. K. Mak, "DNN-Based score calibration with multitask learning for noise robust speaker verification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 700–712, Apr. 2018.

[29] T.-P. Chen *et al.*, "Functional harmony recognition of symbolic music data with multi-task recurrent neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 90–97.

[30] N. Tonami, K. Imoto, M. Niitsuma, R. Yamanishi, and Y. Yamashita, "Joint analysis of acoustic events and scenes based on multitask learning," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 338–342.

[31] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4460–4464.

[32] T. Kano, S. Sakti, and S. Nakamura, "End-to-end speech translation with transcoding by multi-task learning for distant language pairs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, no. 28, pp. 1342–1355, Apr. 2020.

[33] S. Zhou, X. Zeng, Y. Zhou, A. Anastasopoulos, and G. Neubig, "Improving robustness of neural machine translation with multi-task learning," in *Proc. 4th Conf. Mach. Transl.*, 2019, pp. 565–571.

[34] N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, C. Forlivesi, and F. Kawsar, "Squeezing deep learning into mobile and embedded devices," *IEEE Pervasive Comput.*, vol. 16, no. 3, pp. 82–88, 2017.

[35] T. Lee and A. Ndirango, "Generalization in multitask deep neural classifiers: A statistical physics approach," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 15862–15871.

[36] N. Zhuang, Y. Yan, S. Chen, and H. Wang, "Multi-task learning of cascaded cnn for facial attribute classification," in *Proc. 24th Int. Conf. Pattern Recognit.*, 2018, pp. 2069–2074.

[37] Y. Gong *et al.*, "Deep cascade multi-task learning for slot filling in online shopping assistant," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 6465–6472, 2019.

[38] D. Zhou and Q. He, "Cascaded multi-task learning of head segmentation and density regression for RGBD crowd counting," *IEEE Access*, vol. 8, pp. 101 616–101 627, 2020.

[39] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. 13th Int. Conf. Digit. Audio Effects (DAFx'10)*, Graz, Austria, vol. 13, 2010, pp. 1–4.

[40] P. Virtanen *et al.*, "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, pp. 261–272, 2020.

[41] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[42] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," pp. 1–4, Oct. 2015. [Online]. Available: https://arxiv.org/abs/1510.08484

[43] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 3875–3879.

[44] H. Li, K. Chen, and B. U. Seeber, "Auditory filterbanks benefit universal sound source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 181–185.

[45] B. J. Borgström and M. S. Brandstein, "Speech enhancement via attention masking network (seamnet): An end-to-end system for joint suppression of noise and reverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, no. 29, pp. 515–526, Dec. 2021.

[46] H. Dinkel, S. Wang, X. Xu, M. Wu, and K. Yu, "Voice activity detection in the wild: A data-driven approach using teacher-student training," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, no. 29, pp. 1542–1555, Apr. 2021.

[47] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," in *Proc. 14th Python Sci. Conf.*, vol. 8, 2015, pp. 18–25.

[48] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://keras.io

[49] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016. [Online]. Available: https://arxiv.org/abs/1603.04467

[50] K. N. Stevens, *Acoustic Phonetics*, vol. 30. Cambridge, MA, USA: MIT Press, 2000.

[51] J. Schlüter and R. Sonnleitner, "Unsupervised feature learning for speech and music detection in radio broadcasts," in *Proc. 15th Int. Conf. Digit. Audio Effects*, vol. 15, 2012, pp. 1–8.

[52] T. Dozat, "Incorporating Nesterov momentum into ADAM," in *Proc. 4th Int. Conf. Learn. Representations*, 2016, pp. 1–4.

[53] S. Ruder, J. Bingel, I. Augenstein, and A. Sogaard, "Latent multi-task architecture learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 4822–4829, 2019.

Mrinmoy Bhattacharjee (Graduate Student Member, IEEE) received the Bachelor of Technology degree with a specialization in information technology from Assam University Silchar, Silchar, India, in 2011, the Master of Technology degree with a specialization in computer technology from the National Institute of Technology, Raipur, India, in 2014. He is currently working toward the Ph.D. degree in signal processing with the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, India. His research interests include speech, music, and general audio processing.

S. R. M. Prasanna (Senior Member, IEEE) is currently a Professor with the Department of Electrical Engineering, Indian Institute Of Technology (IIT) Dharwad, Dharwad, India. He was earlier a Professor, an Associate Professor, and Assistant Professor with the Department of Electronics and Electrical Engineering, IIT Guwahati, Guwahati, India. He has supervised several Ph.D. theses in his research areas and published in reputed national and international journals and conferences, which include speech and handwriting processing.

Prithwijit Guha (Member, IEEE) received the bachelor's degree in electrical engineering from Jadavpur University, Kolkata, India, and the M.Tech. and Ph.D. degrees in electrical engineering from the Indian Institute of Technology Kanpur, Kanpur, India. He is currently an Associate Professor with the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, India. His research interests include computer vision, machine learning and signal processing.