Speech/Music Classification Using Features from Spectral Peaks

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

***Abstract*—Spectrograms of speech and music contain distinct striation patterns. Traditional features represent various properties of the audio signal but do not necessarily capture such patterns. This work proposes to model such spectrogram patterns using a novel Spectral Peak Tracking (SPT) approach. Two novel time-frequency features for speech vs. music classification are proposed. The proposed features are extracted in two stages. First, SPT is performed to track a preset number of highest amplitude spectral peaks in an audio interval. In the second stage, the location and amplitudes of these peak traces are used to compute the proposed feature sets. The first feature involves the computation of mean and standard deviation of peak traces. The second feature is obtained as averaged component posterior probability vectors of Gaussian mixture models learned on the peak traces. Speech vs. music classification is performed by training various binary classifiers on these proposed features. Three standard datasets are used to evaluate the efficiency of the proposed features for speech/music classification. The proposed features are benchmarked against five baseline approaches. Finally, the best-proposed feature is combined with two contemporary deep-learning based features to show that such combinations can lead to more robust speech vs. music classification systems.**

***Index Terms*—Spectral peak tracking, time-frequency audio features, speech music classification, spectrogram, SVM, CNN, GMM,**

I. Introduction

CONTENT-based audio indexing and retrieval applications often involve a critical preprocessing step of segmenting and classifying audio signals into distinct categories. Apart from general environmental sounds, speech and music are two basic audio categories. Preprocessing steps require classification algorithms that ensure the homogeneity of individual classes in audio segments [1]. This work focuses on proposing features for better discrimination of speech and music for such audio segmentation applications.

Researchers have exploited various acoustic differences between speech and music signals for classifying them [2], [3]. Saunders et al. [4] mention that pitch information usually exists for only three octaves in speech, whereas fundamental tones in music span up to six octaves. Sell et al. [5] state that unlike speech, music is expected to have strict structures in the frequency domain since specific tones play an essential part in its production. Panagiotakis et al. [6] show that the amount of silence present in the signal may also be a good discriminator between the two classes. Short silences usually punctuate speech sound units while music is generally continuous (Fig. 1).

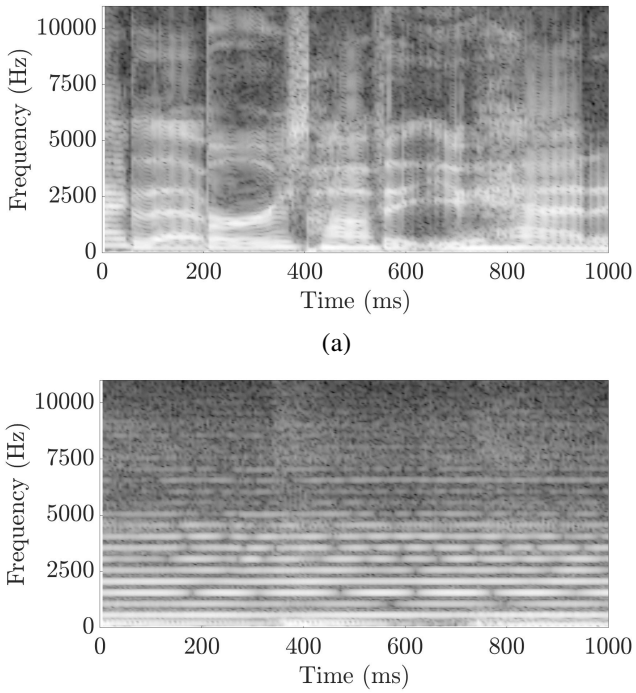

Fig. 1. Spectrograms of (a) Speech and (b) Music, computed using frame size of $10ms$ and frame shift of 5ms. Note the distinct striation patterns of speech and music. This observation motivated our proposal of time-frequency audio features for speech-music discrimination.

Many standard audio features have been used in literature to model the distinct behaviors of speech and music. The most widely used spectral features in this task are Zero-Crossing Rate [5], Spectral Centroid, Spectral Roll-off, and Spectral Flux [7]. Energy [8], Entropy [9] and Root Mean Square [5] values are the most popular temporal features used in speech vs. music classification (SMC, henceforth). Khonglah et al. [10] have proposed that features predominantly used in speech processing tasks (like the speech-specific modulation spectrum features) can be effective in the current task also. On the other hand, Sell et al. [5] suggest that chroma-based features are better in modeling the octave patterns in music and thus might

Mrinmoy Bhattacharjee, S.R.M. Prasanna and P. Guha are with the Dept. of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati-781039, India.

S.R.M. Prasanna is also with the Dept. of Electrical Engineering, Indian Institute of Technology Dharwad, Dharwad-580011, India.

Corresponding author: Mrinmoy Bhattacharjee (email: mrinmoy.bhattacharjee@iitg.ac.in).

The authors would like to thank the Visvesvaraya Ph.D. scheme of MeitY, GOI for supporting this work.

be useful in discriminating it from speech.

Existing works in SMC have mostly employed Gaussian Mixture Models (GMM) [5], [10], [11], Artificial Neural Networks (ANN) [9], k-Nearest Neighbors (kNN) [12], [13], [14] and Support Vector Machines (SVM) [11], [7], [8] as classifiers. Recently, authors have also used deep learning techniques for solving the SMC task [15], [16]. CNNs are very popular in image processing applications for feature learning and classification. This motivated researchers to use CNNs for learning features from the time-frequency representation of audio segments. Doukhan et al. [17] propose a semi-supervised training procedure for solving the SMC task. The first convolution layer in their architecture is pre-trained in an unsupervised manner using the spherical k-means algorithm and later kept constant throughout the model training phase. The input to their model is the stacked Mel-frequency coefficients of 50 frames. Papakostas et al. [18], on the other hand, use transfer learning to fine-tune an existing CNN model trained on image classification problems to learn the discrimination between spectrograms of speech and music.

There are works in literature that have explored features that can capture simultaneous variations in temporal and spectral domains for achieving better performance in SMC [5], [7], [9], [10], [19]. Spectrograms are a popular method of visualizing the tempo-spectral properties of an audio signal. Fig. 1 (a) and (b) show the spectrograms of speech and music respectively. Note that spectrograms can either have high time-resolution (wideband spectrograms) or high frequency-resolution (narrowband spectrograms), but not both at the same time [20]. Wideband spectrograms are generated with short temporal windows. They are characterized by vertical striations that represent pitch period and formant frequencies (in case of speech) in the form of prominent horizontal bands [20]. Narrowband spectrograms are generated using longer analysis windows and have horizontal striations that depict the fundamental frequency and its harmonics. Peaks in the spectra of audio frames may appear as striation patterns in spectrograms. Distinct class-specific properties can be captured by tracing trajectories of the highest spectral peaks in the spectrograms. Researchers have also used spectrograms for feature extraction. For example, Mesgarani et al. [21] were inspired by the auditory cortical processing methods to use Gabor-like spectro-temporal response fields for feature extraction from spectrograms. Whereas Neammalai et al. [8] performed thresholding and smoothing on standard spectrograms to form binary images and used them as features for classification.

Peak tracking has been a widely explored approach in the field of speech coding and synthesis. McAulay et al. [22] proposed that a speech segment can be represented as a combination of various sinusoids of specific frequency and definite lifetime, called partials. They generated high-quality artificial speech by adding together different partials with time weighing and amplitude modulation. Smith et al. [23] proposed an approach similar to [22], but for representing polyphonic music. Lagrange et al. [24] proposed an improved partial tracking algorithm based on the linear prediction algorithm that can better model the pseudo periodic part of polyphonic sounds. In other works, researchers have used a technique called Spectral Peak Tracking (SPT, henceforth) to trace the trajectory of fundamental frequency across consecutive frames in the spectrogram [25], [26]. Techniques similar to SPT have been used for feature generation in SMC literature as well. Seyerlehner et al. [27] proposed a feature called Continuous Frequency Activation (CFA, henceforth), which measures the steadiness of spectral components within a block of audio. Since music is considered to be relatively more stationary, this feature provided improved results in case of music detection. In works like [28], authors use a pre-determined threshold to binarize the magnitude spectrum of each audio frame in an interval to 1's and 0's. Subsequently, they count the number of 1's that appear for each frequency channel and use this measure as a feature for classification. Padmanabhan et al. [29] processed speech signals using band-pass filters and tracked spectral peaks in each band for speech recognition.

It can be observed from Fig. 1 that speech and music signals produce quite distinct striation patterns in their respective spectrograms. Pitch and harmonics in speech slowly change from one sound unit to another [30]. These gradual transitions create arc-like striations in speech spectrograms. On the other hand, music spectrograms contain many horizontal line segments caused by relatively stationary pitch and harmonics, and broken by their sharp transitions [31], [32]. These spectro-temporal differences observed in the spectrograms of speech and music can be attributed to the following reasons. First, the speech production system possesses inertia [33], [34]. It requires a relatively large amount of time to change from one sound unit to another, leading to a smooth transition between sound units in speech spectrograms. On the other hand, individual notes of music have a specific onset instant, marked by a relatively large burst of energy that makes its striation patterns discontinuous [35]. Second, music tones decay slowly [36]. Comparatively, the speech production system is a damped system where sound units decay quite fast [37], which explains the presence of horizontal patterns in music but not in speech. Third, musical instruments produce only a fixed number of tones and their overtones [5]. On the other hand, the speech production system generates a large number of intermediate frequencies while transitioning from one sound unit to another [38], leading to the formation of arc-like patterns in speech spectrograms.

The observed differences in the spectrograms of speech and music motivated us to design features that can capture these distinct class-specific striation patterns for speech vs. music classification. However, hand-crafted features have a high dependence on problem-specific assumptions. On the other hand, automatic feature learning methods (like CNNs) can efficiently learn underlying patterns in the data. However, it is not very easy to interpret the information learned by such deep-learning methods due to their inherent stochastic nature. In applications where domain knowledge is available, it is also worthwhile to explore hand-crafted features that can show decent performance. Efficient hand-crafted features may be combined with deep-learning-based features to build more robust systems for SMC. These ideas form the basis of current proposal. This work has the following contributions:

- Proposal of a novel approach for SPT (subsection II-A)

which is capable of capturing prominent striation patterns present in spectrograms of speech and music signals.
- Proposal of two novel features sets – (a) MSD feature (subsection II-B) constructed using first and second order moments of location and amplitude values of peak traces obtained by SPT, (b) CBoW features (subsection II-C) constructed using averaged posterior probabilities obtained from Gaussian mixture models learned on peak traces obtained from entire training data.
- Proposal of a combination of the proposed features with deep-learning based features, that shows that such combinations can build more robust SMC systems.

The rest of this paper is organized as follows. The proposed scheme for SPT and subsequent feature extraction is described in detail in Section II. Three standard speech-music datasets are used to benchmark the proposal in this paper. We compare the performance of our proposal with five baseline approaches. We have performed extensive experiments and reported the results in Section III. Finally, we conclude in Section IV and sketch the possible future extensions of the present proposal.

II. Proposed Work

Speech and music are complex non-stationary signals. Spectra of speech consist of source harmonics superimposed by vocal tract formants. Energy concentrations in the spectrograms of speech signals are a manifestation of formants [39]. These energy concentrations are formed by high-amplitude peaks in speech spectra. It has been established that high-amplitude peaks provide information about dominant formants in the speech spectra [40], [29]. However, formants may be defined only for voiced segments, which constitute a majority of the speech content. On the other hand, music does not have any formant structure in its spectrum. It is composed of harmonics and resonances of the generating instruments. High amplitude spectral peaks in the music spectra will mostly correspond to resonant frequencies. The number of high amplitude spectral peaks to be considered for tracking must be a tunable parameter that depends upon the task at hand. For example, in polyphonic music or multi-speaker speech, where many fundamental frequencies might be present, considering more spectral peak tracks may help in capturing better discriminating information.

Spectral peak trajectories carry valuable information about the underlying sound segment. However, most speech processing systems use perceptually motivated cepstral features that do not explicitly model the peak trajectory information [29]. SPT in speech and music spectrograms might be a very effective way of extracting these discriminating features. There can be at least three strong reasons for claiming this. First, speech formants have a well-studied structure and show a predictable behavior [41]. However, resonances in music have a dynamic nature depending upon the composition and set of instruments used to produce the signal. Second, speech production uses only a single resonant cavity, i.e., the vocal tract, whereas the music signal is composed of multiple resonant devices depending upon the number of instruments used. Any deviation from the resonance patterns of speech may indicate the presence of music in the current two-class scenario. Third, as discussed in Section I, music signals tend to maintain some of their spectral properties for a considerable duration of time, whereas speech is highly non-stationary. We assume that trajectories of spectral peaks could capture such distinct behavior of speech and music. The proposed SPT technique and the subsequent feature extraction procedure are described in detail next.

TABLE I
Repeating procedure statistics for $p = 10$.

Dataset	Percentage of peak-repeated frames		
	Music	Speech	Overall
GTZAN	0%	0%	0%
Scheirer-slaney	0%	0%	0%
Musan	0.022%	0.020%	0.021%

A. Proposed SPT method

An audio segment $\mathbf{x}$ $(\mathbf{x}[n] \in \mathcal{R}; n = 0, \ldots N_s - 1)$ is divided into L overlapping frames $\mathbf{x}_l$ $(l = 0, \ldots L-1)$ of size $2N_f$. Let the kth DFT coefficient of $\mathbf{x_l}$ be,

$$\mathbf{X}_l[k] = \sum_{m=0}^{2N_f-1} \mathbf{x}_l[m] e^{-jk\frac{2\pi}{2N_f}m} \tag{1}$$

where, $k = 0 \ldots 2N_f - 1$. These frames $(\mathbf{x}_l)$ are sequences of real numbers. Hence, only the first N_f DFT coefficients (i.e. $\mathbf{X}_l[k]; k = 0, \ldots N_f - 1$) are considered for further processing. Next, we identify the frequency locations of all spectral peaks in lth frame to construct the following set $\mathbf{H}_l$.

$$\mathbf{H}_l = \{k : (|\mathbf{X}_l[k-1]| < |\mathbf{X}_l[k]|) \wedge (|\mathbf{X}_l[k]| > |\mathbf{X}_l[k+1]|)\} \tag{2}$$

where $0 \leq k < (N_f - 1)$ and $|\mathbf{X}_l[k]|$ indicates the magnitude of $\mathbf{X}_l[k]$. The number of spectral peaks in each frame varies. Not all spectral peaks are important for the task at hand. Thus, only a fixed number (atmost p, say) of highest amplitude peaks are identified from the spectrum of each frame. These highest spectral peaks from each frame are used to construct the truncated frequency location set $\tilde{\mathbf{H}}_l$:

$$\tilde{\mathbf{H}}_l = \left\{k_{(0)}, k_{(1)}, \ldots k_{(q)}\right\} \qquad \left(\tilde{\mathbf{H}}_l \subseteq \mathbf{H}_l\right) \tag{3}$$

such that $|\mathbf{X}_l[k_{(0)}]| \geq |\mathbf{X}_l[k_{(1)}]| \geq \ldots \geq |\mathbf{X}_l[k_{(q)}]|$ and $0 < q \leq (p-1)$. If for any lth frame, $q < (p-1)$, then the highest frequency location (i.e., $max(\tilde{\mathbf{H}}_l)$) in $\tilde{\mathbf{H}}_l$ is repeated $p - 1 - q$ times to maintain uniform cardinality of $\tilde{\mathbf{H}}_l$ (i.e., $|\ \tilde{\mathbf{H}}_l\ | = p$) for all frames. When just a small number of highest amplitude spectral peaks are considered ($p = 10$, say), this repeating procedure has negligible effect on the peak amplitude and location distributions. It is evident from Table I that the percentage of frames requiring peak repetition is *Nil* for two of the datasets used for evaluation, (GTZAN and Scheirer-Slaney, see Section III), while only a minuscule percentage of frames from the third dataset (MUSAN dataset, Section III) require peak repetition.

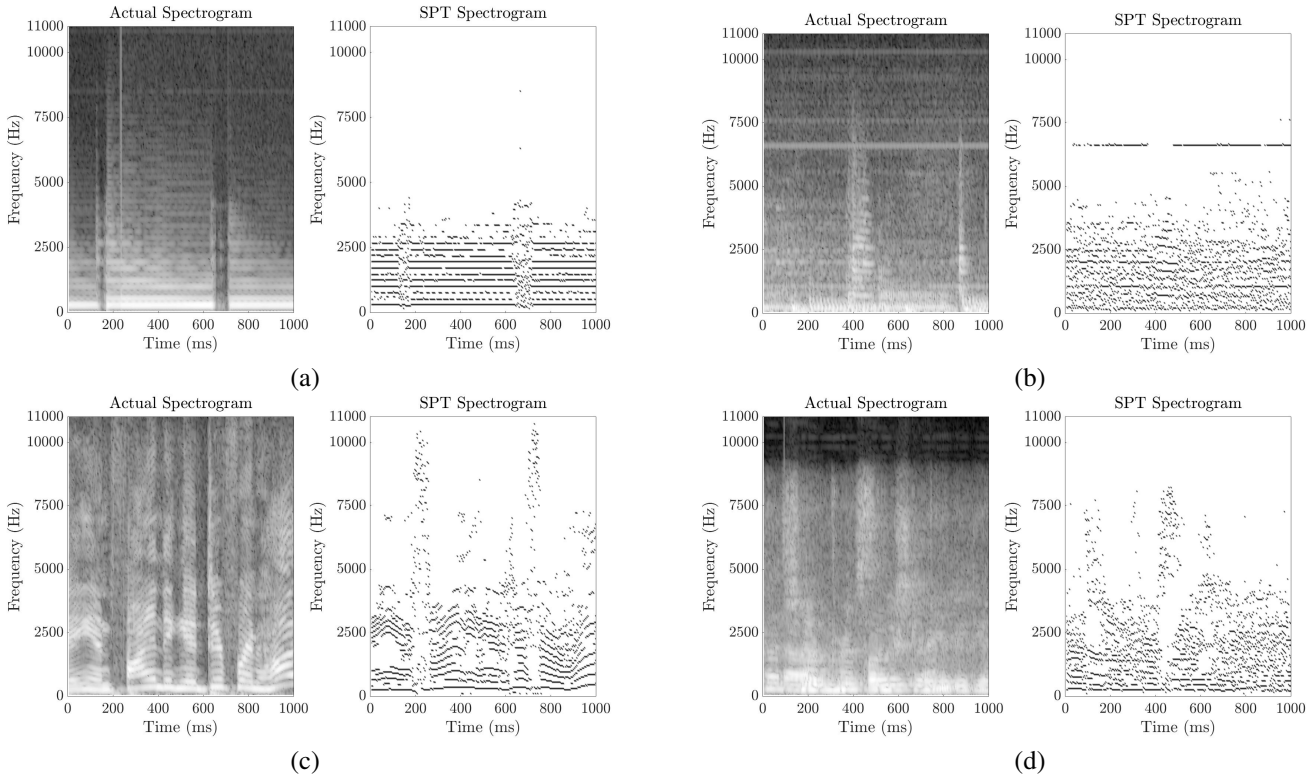

Fig. 2. Illustrating the effect of multiple fundamental frequencies on the spectrograms and subsequently on the proposed peak tracking algorithm. The figures shown are spectrograms computed from 1s intervals of (a) monophonic music, (b) polyphonic music, (c) single speaker speech, and (d) multi-speaker speech. In each subfigure, the original spectrogram is shown in the left, and the SPT spectrogram is shown in the right.

Next, the elements of $\tilde{\mathbf{H}}_l$ (frequency locations of the p highest peaks in the l^{th} frame spectra) are further sorted in descending order to construct the vector $\mathbf{fH}_l$ such that:

$$\mathbf{fH}_l[0] \geq \mathbf{fH}_l[1] \geq \ldots \geq \mathbf{fH}_l[p-1] \tag{4}$$

Here, $\mathbf{fH}_l[r] \in \tilde{\mathbf{H}}_l$ and $r = 0, \ldots p-1$. $\mathbf{fH}_l$ contains the sorted frequency locations of the spectral peaks in $\tilde{\mathbf{H}}_l$. The vectors $\mathbf{fH}_l$ $(l = 0, \ldots L-1)$ are used to construct a $p \times L$ Peak Location Matrix (PLM, henceforth) $\mathcal{L}$ for an audio interval. The lth column of $\mathcal{L}$ is defined as

$$\mathcal{L}_l = \mathbf{fH}_l^{\mathrm{T}} \tag{5}$$

Similarly, a Peak Amplitude Matrix (PAM, henceforth) $\mathcal{A}$ can be constructed. The elements of $\mathcal{A}$ are defined as

$$\mathcal{A}[r, l] = \mathbf{X}_l[h] \tag{6}$$

where $h = \mathcal{L}[r, l]$, $r = 0, \ldots (p-1)$ and $l = 0, \ldots (L-1)$. A flow-chart describing the procedure of computing the PLM (or PAM) matrix is provided in Fig. 4. Each row of $\mathcal{L}$ is defined as Location Sequence of Peak Traces (LSPT, henceforth). Similarly, each row of $\mathcal{A}$ is defined as Amplitude Sequence of Peak Traces (ASPT, henceforth). Note that the first row of $\mathcal{L}$ (or $\mathcal{A}$) corresponds to the peak traces of highest end of the spectrum. Similarly, the last row of $\mathcal{L}$ (or $\mathcal{A}$) corresponds to the peak traces of lowest end of the spectrum. LSPT and ASPT are sequences of peak location and amplitude values, respectively. This work views these peak traces as sub-channels of information extracted from the spectrogram of an audio interval.

In Fig. 2, the traces of identified spectral peaks are shown in the actual time-frequency scale as a separate representation, termed as an SPT-spectrogram in this work. An SPT-spectrogram is a matrix of the same size as the actual spectrogram, initialized with zeros. On this matrix, each spectral peak is located with its frequency bin and frame index and initialized with its amplitude. This SPT-spectrogram, when plotted as an image, shows the peak traces extracted from the corresponding spectrogram. We observe (Fig. 2) that peak traces capture unique striation patterns of speech and music spectrograms. Each LSPT (or ASPT) represents a part of this striation information. When multiple sources (hence multiple fundamental frequencies) are present in the audio segment, the harmonic patterns are disturbed, and the spectrogram becomes noisy. However, the signal retains a basic property of its audio class. As can be observed in Fig. 2, both monophonic (Fig. 2 (a)) and polyphonic (Fig. 2 (b)) music contain relatively linear striation patterns. Whereas, single-speaker (Fig. 2 (c)) and multi-speaker (Fig. 2 (d)) speech have curvy striations. Since the basic assumption of this work is preserved even in the case of multiple F_0 signals, the proposed approach is still able to capture the required discriminative information for classification. Efficacy of the proposed SPT approach can be confirmed by observing the SPT spectrograms shown in Fig. 2 (a)-(d) that contain all the prominent spectral striations for all the four cases. Note that these SPT-spectrograms are generated just for visualization purposes and are not used for feature computation. The proposed features are computed using the PLM (and PAM) matrix. Our proposal for modeling the

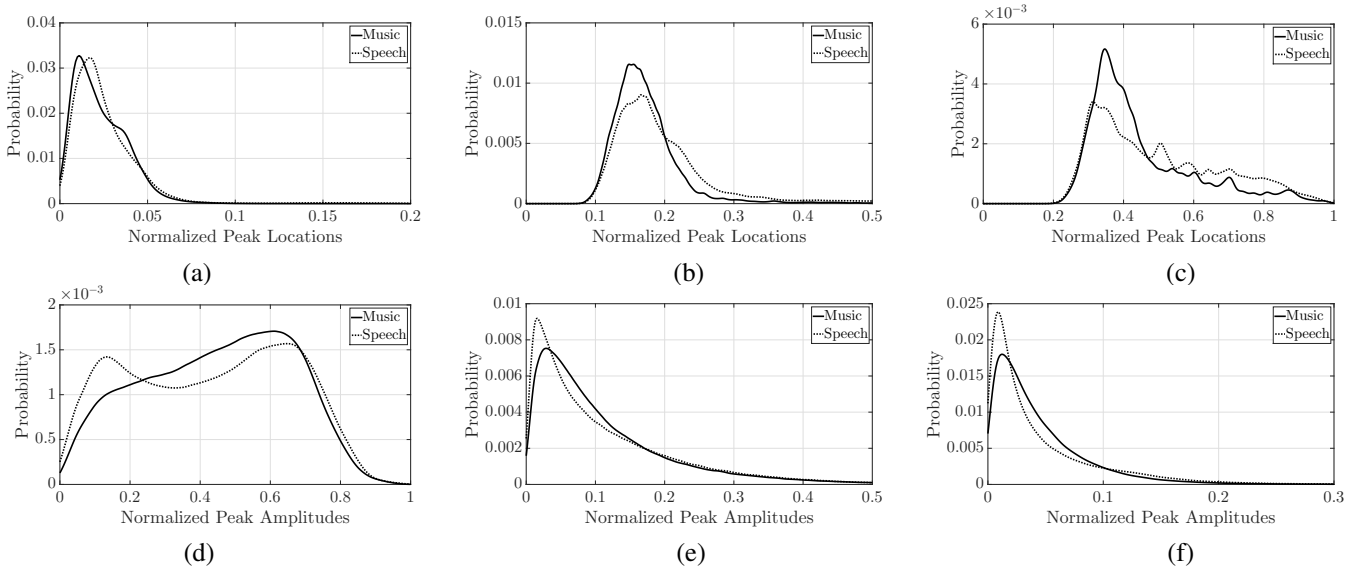

Fig. 3. Illustration of the peak location and amplitude distributions for 1^{st}, 5^{th}, and 10^{th} peak traces, generated using frame size of $10ms$ and frame shift of 5ms. Figures **(a)-(c)** show peak location distributions and Figures **(d)-(f)** show peak amplitude distributions. The data is drawn from the GTZAN dataset.

distributions of LSPTs (or ASPTs) as features are discussed next.

B. Statistical moments of peak traces as feature

Fig. 3 (a)-(c) respectively show distributions of first, fifth and tenth LSPT distributions of speech and music across all data in GTZAN dataset. Similarly, Fig. 3 (d)-(f) respectively show first, fifth and tenth ASPT distributions of speech and music. The LSPTs (and ASPTs) have been computed using short-term frames of size $10ms$ and a frame shift of 5ms. It can be seen that the corresponding distributions of LSPT and ASPT of speech and music are different. This difference might be useful in classifying these two classes if the distributions of these sequences are represented in suitable feature space. Our first proposal involves the use of Mean and Standard Deviation (MSD, henceforth) for modeling these distributions. Accordingly, the features extracted from PLM ($\mathcal{L}$, (5)) and PAM ($\mathcal{A}$ (6)) are named as MSD-LSPT and MSD-ASPT respectively. For notational convenience, the index r ($0 \leq r < p$) is used for referring to the rth row of $\mathcal{L}$ and $\mathcal{A}$. Attributes derived from the rth LSPT (or ASPT) will also be indexed by r. Mean $\mu_r^{\mathcal{L}}$ and standard deviation $\sigma_r^{\mathcal{L}}$ of the rth LSPT is computed as:

$$\mu_r^{\mathcal{L}} = \frac{1}{L}\sum_{l=0}^{L-1}\mathcal{L}[r][l] \quad \sigma_r^{\mathcal{L}} = \sqrt{\frac{1}{L}\sum_{l=0}^{L-1}\left(\mathcal{L}[r][l]-\mu_r^{\mathcal{L}}\right)^2} \tag{7}$$

The MSD feature computed from PLM is proposed as a $2p$-dimensional vector given by:

$$\text{MSD-LSPT} = \left[\mu_0^{\mathcal{L}}, \ldots \mu_{p-1}^{\mathcal{L}}, \sigma_0^{\mathcal{L}}, \ldots \sigma_{p-1}^{\mathcal{L}}\right] \tag{8}$$

Similarly, the mean ($\mu_r^{\mathcal{A}}$) and standard deviation ($\sigma_r^{\mathcal{A}}$) of ASPT can be computed from PAM, and used to construct the MSD feature as a $2p$-dimensional vector given by:

$$\text{MSD-ASPT} = \left[\mu_0^{\mathcal{A}}, \ldots \mu_{p-1}^{\mathcal{A}}, \sigma_0^{\mathcal{A}}, \ldots \sigma_{p-1}^{\mathcal{A}}\right] \tag{9}$$

Additionally a $4p$-dimensional feature vector MSD-ASPT-LSPT can be formed by concatenating MSD-ASPT and MSD-LSPT. The MSD features extracted from individual intervals of speech and music data are provided for training classifiers. Performance of MSD features in SMC on standard datasets are shown in Section III. This proposal uses only the first and second order statistics of LSPT (or ASPT). Our next proposal employs Gaussian mixture models for modelling peak location and amplitude distributions.

C. Component Bag-of-Words (CBoW) features from peak traces

Peak traces are temporally ordered sequences of prominent peaks occurring in successive frames of an audio interval. We believe that these sequences are capable of capturing the highest energy striation patterns observed in spectrograms. It can be observed from Fig. 3 that the peak traces exhibit multi-modal distributions. Thus, the use of only mean and standard deviation might be insufficient to model these distributions. Moreover, the MSD features are extracted from individual audio intervals and hence, are oblivious to their global distribution. This motivated us to propose another set of features that are capable of representing the inherent multi-modality of the underlying global distribution.

Gaussian mixture models (GMMs, henceforth) are widely used to characterize multi-modal data. A K-component GMM $\mathcal{G} = \left\{\mathcal{C}_0, \mathcal{C}_1, \ldots \mathcal{C}_{(K-1)}\right\}$ consists of the component Gaussians $\mathcal{C}_j = \{\pi_j, \mu_j, \nu_j\}$ $(j = 0, \ldots (K-1))$. Here, π_j is the mixing parameter, μ_j is the mean and ν_j is the variance of $\mathcal{C}_j$. Usually, a GMM is learned using the Expectation-Maximization algorithm. The number of GMM components (K) is selected empirically based on experimental results. In this work, single dimensional GMMs are trained with an optimal number of modes (K) to model the distribution of any rth peak trace across the whole training dataset. Let $\mathcal{G}^r$ be a GMM trained on any rth peak trace of either speech or music. Let u be the location (or amplitude) of a member peak of the

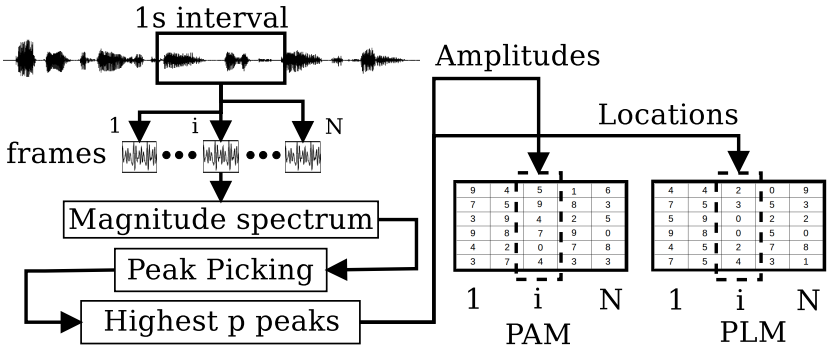

Fig. 4. Flow chart illustrating the process of computing the PAM and PLM matrices.

rth peak trace. The posterior probability of the jth component $\mathcal{C}_j^r$ of $\mathcal{G}^r$ with respect to u can be computed as

$$\mathcal{P}\left(\mathcal{C}_j^r \mid u\right) = \frac{P\left(u \mid \mathcal{C}_j^r\right)\pi_j^r}{\sum_{i=0}^{K-1} P\left(u \mid \mathcal{C}_i^r\right)\pi_i^r} \tag{10}$$

Here, the likelihood function $P\left(u \mid \mathcal{C}_j^r\right)$ is defined as

$$P\left(u \mid \mathcal{C}_j^r\right) = \frac{1}{\sqrt{2\pi\nu_j^r}}\, exp\left(-\frac{\left(u-\mu_j^r\right)^2}{2\nu_j^r}\right) \tag{11}$$

Let ${}_m\mathcal{G}^r = \left\{{}_m\mathcal{C}_j^r\right\}$ and ${}_s\mathcal{G}^r = \left\{{}_s\mathcal{C}_j^r\right\}$ $(j = 0, \ldots K-1)$ be two GMMs learned from the rth peak traces across the whole training set of music and speech respectively. The peak trace distributions of music and speech are observed to be distinctly different. This will lead to two GMMs with different component Gaussians. Thus, we assume that for most cases:

$$\frac{1}{L}\sum_{l=0}^{L-1}\mathcal{P}\left({}_m\mathcal{C}_j^r \mid u_l\right) \neq \frac{1}{L}\sum_{l=0}^{L-1}\mathcal{P}\left({}_s\mathcal{C}_j^r \mid u_l\right) \tag{12}$$

where u_l are location (or amplitude) of peak traces of an interval of L frames. This motivated us to propose the Component Bag-of-Words (CBoW, henceforth) features as averaged K posterior probabilities obtained from speech and music GMMs learned from p peak traces. These features have been named such because of their similarity to bag-of-words representation existing in the literature. CBoW feature extraction is a two stage process. The first stage involves estimation of separate GMMs from peak traces of all speech and music training data. This is described next.

Let ${}_s\mathcal{L}^t$ be the PLM matrix constructed from the tth interval $(t = 0, \ldots T_s - 1)$ of speech training data. Similarly, let ${}_m\mathcal{L}^\tau$ denote the PLM matrix constructed from the τth interval $(\tau = 0, \ldots T_m - 1)$ of music training data. Let the rth rows of ${}_s\mathcal{L}^t$ and ${}_m\mathcal{L}^\tau$ be denoted by ${}_{ls}R_t^r$ and ${}_{lm}R_\tau^r$ respectively $(r = 0, \ldots (p-1))$. Two different sets:

$${}_{ls}\mathbf{S}^r = \{{}_{ls}R_t^r[i]; t = 0, \ldots T_s - 1, i = 0, \ldots L-1\} \tag{13}$$

$${}_{lm}\mathbf{S}^r = \{{}_{lm}R_\tau^r[i]; \tau = 0, \ldots T_m - 1, i = 0, \ldots L-1\} \tag{14}$$

are constructed for accumulating the frequency locations of the rth peak traces of respective speech and music training data. Single dimensional K-component GMMs ${}_{ls}\mathcal{G}^r$ and ${}_{lm}\mathcal{G}^r$ are estimated from the elements of ${}_{ls}\mathbf{S}^r$ and ${}_{lm}\mathbf{S}^r$ respectively. Note that, two GMMs are learned for any rth peak trace. Thus, a total of $2p$ GMMs are estimated for p peak traces. We next describe the second stage of CBoW feature extraction that involves the computation of posterior probability vectors for any given audio interval.

Let $\mathcal{L}$ be the $p \times L$ PLM matrix of an audio interval containing L frames. Let ${}_lR^r$ be the rth row of $\mathcal{L}$. The learned GMMs ${}_{lm}\mathcal{G}^r$ and ${}_{ls}\mathcal{G}^r$ are used to obtain component-wise posterior probabilities for each element of ${}_lR^r$. The averaged posterior probability vector ${}_{lm}\mathcal{H}^r$ for ${}_lR^r$ is obtained by using ${}_{lm}\mathcal{G}^r$. This is computed as follows.

$${}_{lm}\mathcal{Z}^r(i) = \left[\mathcal{P}\left({}_{lm}\mathcal{C}_0^r \mid {}_lR^r[i]\right), \ldots \mathcal{P}\left({}_{lm}\mathcal{C}_{(K-1)}^r \mid {}_lR^r[i]\right)\right]$$
$${}_{lm}\mathcal{H}^r = \frac{1}{L}\sum_{i=0}^{L-1} {}_{lm}\mathcal{Z}^r(i) \tag{15}$$

Similarly, the averaged posterior probability vector ${}_{ls}\mathcal{H}^r$ for ${}_lR^r$ is computed (using ${}_{ls}\mathcal{G}^r$) in the following manner.

$${}_{ls}\mathcal{Z}^r(i) = \left[\mathcal{P}\left({}_{ls}\mathcal{C}_0^r \mid {}_lR^r[i]\right), \ldots \mathcal{P}\left({}_{ls}\mathcal{C}_{(K-1)}^r \mid {}_lR^r[i]\right)\right]$$
$${}_{ls}\mathcal{H}^r = \frac{1}{L}\sum_{i=0}^{L-1} {}_{ls}\mathcal{Z}^r(i) \tag{16}$$

Note that, both ${}_{ls}\mathcal{H}^r$ and ${}_{lm}\mathcal{H}^r$ are K length vectors. We construct the proposed CBoW-LSPT feature as a $2 \times K \times p$ dimensional vector and is given by

$$\text{CBoW-LSPT} = \left[{}_{lm}\mathcal{H}^0,\ {}_{ls}\mathcal{H}^0, \ldots,\ {}_{lm}\mathcal{H}^{p-1},\ {}_{ls}\mathcal{H}^{p-1}\right] \tag{17}$$

Similarly, PAM matrices computed from both speech and music intervals are denoted as ${}_m\mathcal{A}^\tau$ and ${}_s\mathcal{A}^t$ respectively. The rth rows ${}_{am}R_\tau^r$ and ${}_{as}R_t^r$ of ${}_m\mathcal{A}^\tau$ and ${}_s\mathcal{A}^t$ are used to form the sets ${}_{am}\mathbf{S}^r$ and ${}_{as}\mathbf{S}^r$. The respective GMMs ${}_{am}\mathcal{G}^r$ and ${}_{as}\mathcal{G}^r$ are estimated from ${}_{am}\mathbf{S}^r$ and ${}_{as}\mathbf{S}^r$. For any given audio interval with PAM matrix $\mathcal{A}$, the averaged posterior probability vectors ${}_{am}\mathcal{H}^r$ and ${}_{as}\mathcal{H}^r$ are computed in a similar manner as described in eqn 15 and eqn 16. The CBoW-ASPT feature is constructed as a $2 \times K \times p$ length vector and is given by

$$\text{CBoW-ASPT} = \left[{}_{am}\mathcal{H}^0, {}_{as}\mathcal{H}^0, \ldots, {}_{am}\mathcal{H}^{p-1}, {}_{as}\mathcal{H}^{p-1}\right] \tag{18}$$

Additionally, a $4 \times K \times p$-dimensional feature vector CBoW-ASPT-LSPT can be formed by concatenating CBoW-ASPT and CBoW-LSPT. Fig. 5 shows a functional block diagram

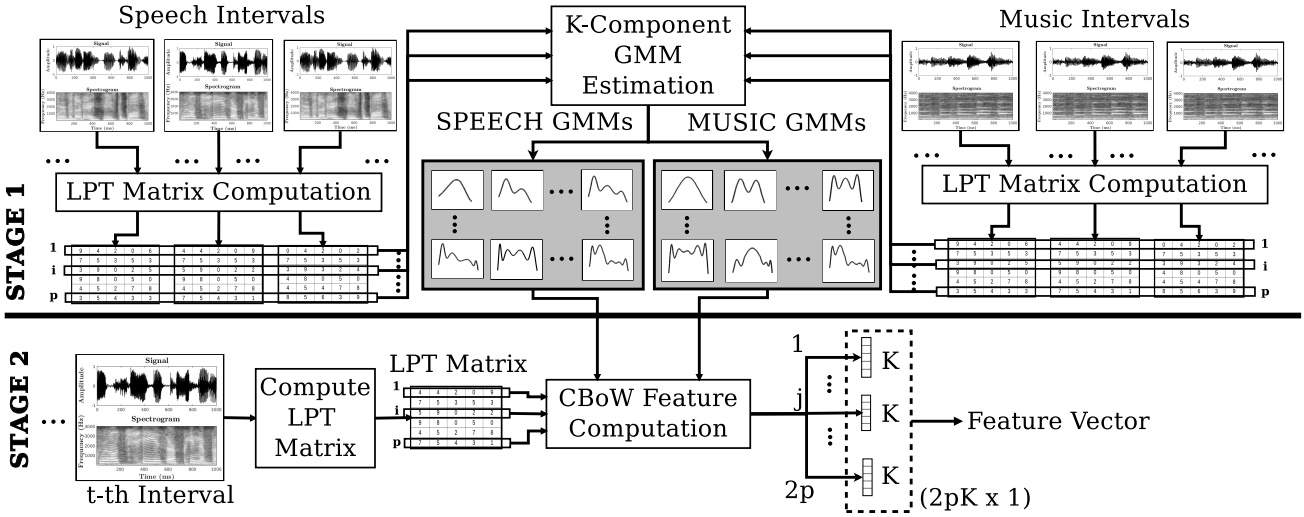

Fig. 5. Schematic diagram representing the procedure for computing the CBoW-LSPT feature. The feature computation is a two stage process. **STAGE 1** estimates separate GMMs for speech and music peak traces from entire training data. These learned GMMs are used in the **STAGE 2** to construct the CBoW-LSPT features. While computing CBoW-ASPT feature, the *LPT Matrix Computation* block is replaced by the *APT Matrix Computation* block.

for computing the CBoW features. Detailed experimentation with proposed features and the results of performance analysis are presented next.

III. Experiments and Results

The proposed features are validated on three datasets. These are (a) GTZAN Music/Speech collection [42], (b) Scheirer-Slaney Music-Speech Corpus [43], (c) MUSAN - A Music, Speech and Noise corpus [44]. Both GTZAN and Scheirer-Slaney contain 1 hour of data. On the other hand, MUSAN is a much larger dataset containing around 102.5 hours of speech and music data.

Five baseline approaches have been used to benchmark our proposal. The first baseline uses speech specific features set (FS) proposed by Khonglah et al. [10] (Khonglah-FS, henceforth). The second baseline (Sell-FS, henceforth) uses chroma based features to represent music tonality for enhanced speech music classification [5]. The Mel Frequency Cepstral Coefficients (MFCCs) are widely used in most speech processing applications, including SMC [7]. This work uses the 39-dimensional MFCCs along with their Δ and $\Delta\Delta$ coefficients as the third baseline (MFCC-39, henceforth) for performance comparison. Keum et al. in [45] proposed features derived from a variant of spectral peak tracking for speech and music discrimination. This approach is adopted as our fourth baseline (Keum-FS, henceforth). Finally, to contemporize the proposal and for benchmarking against current trends of deep-learning based methods, this work uses the CNN architecture proposed in [18] as the fifth baseline (Papakostas-CNN, henceforth). Spectrogram images of $1s$ intervals are used to train the CNN and are used to generate results for comparison with the proposed approach.

The experiments are performed using standard python packages [1]. This work uses a train-test split ratio of 80 : 20. The examples in each of the two sets are sampled randomly without replacement to ensure that there is no overlap between the two sets. Classifier hyperparameters have been tuned over a validation set extracted from the training set, keeping the testing set untouched until final evaluation. Classification performance is reported using the mean and standard deviation of F1-scores [46] obtained from 10 independent trials. The number of peak-traces (p) and the number of GMM mixtures (K) for computing the proposed features are set to 10 and 5 respectively, based on experimental results. For GTZAN and Scheirer-Slaney datasets, classification results for baseline and proposed features are generated using SVM (RBF kernel). The cost and gamma parameters of SVM are tuned using a grid search. For MUSAN dataset, results for baseline and proposed features are computed using Bagged RBF-SVM and Deep neural network (DNN) based classifier. The Bagged SVM classifier ensemble has 10 base SVM classifiers with 20% bootstrap in each bag. The cost and gamma parameters of all base SVMs are optimized using a grid search. Table II shows the DNN architecture used in this work. The DNN model is trained for 100 epochs with a batch size of 64.

TABLE II
Architechture of DNN used in Table VI, VII and VIII.

Input Layer: Size $L1$ = Feature Dimension
Layer 2: Size $L2 = 2 \times L1$, Activation = ReLU
Layer 3: Size $L3 = \frac{2}{3} \times L2$, Activation = ReLU
Layer 4: Size $L4 = \frac{1}{2} \times L3$, Activation = ReLU
Layer 5: Size $L5 = \frac{1}{3} \times L4$, Activation = ReLU
Output Layer: Size $L6 = 2$, Activation = SoftMax

A. Effect of varying frame and interval size

Table III presents the effect of changing short-term audio frame size from 10 ms to 30 ms (frame shifts are taken as half of the frame sizes). A smaller short-term frame gives a smoother spectrum that resembles the formant structure in speech. The presence of formants in speech discriminates it from music. Thus, the performance of proposed features is

[1]Codes used in this work can be found at https://github.com/mrinmoy-iitg/Speech-Music-Classification-Using-SPT

TABLE III
PERFORMANCE OF PROPOSED FEATURES FOR *different audio frame sizes* OVER **GTZAN** DATASET USING 10-COMPONENT GMM CLASSIFIER. INTERVAL SIZE IS FIXED AT 1S. FRAME SHIFTS ARE TAKEN AS 5MS, 10MS AND 15MS, CORRESPONDING TO FRAME SIZES OF 10MS, 20MS AND 30MS, RESPECTIVELY.

Features	Frame Size (in miliseconds)		
	10	20	30
MSD-ASPT	86.96 ±1.65	86.42 ±1.38	86.73 ±1.91
MSD-LSPT	87.51 ±0.89	88.68 ±0.85	86.60 ±1.39
MSD-ASPT-LSPT	90.92 ±1.24	91.21 ±0.92	89.41 ±1.38
CBoW-ASPT	86.56 ±1.28	86.98 ±1.74	88.53 ±1.29
CBoW-LSPT	87.51 ±2.16	87.40 ±1.35	85.79 ±2.16
CBoW-ASPT-LSPT	92.67 ±0.84	93.10 ±0.93	91.79 ±1.06

TABLE IV
PERFORMANCE OF PROPOSED FEATURES FOR *different interval sizes* OVER **GTZAN** DATASET USING 10-COMPONENT GMM CLASSIFIER.

Features	Classification Interval Size (in seconds)		
	0.50	1.00	2.00
MSD-ASPT	84.34 ±1.37	86.96 ±1.65	86.83 ±2.08
MSD-LSPT	82.77 ±1.52	87.51 ±0.89	90.70 ±1.28
MSD-ASPT-LSPT	88.19 ±1.55	90.92 ±1.24	92.81 ±1.90
CBoW-ASPT	83.48 ±1.03	86.56 ±1.28	89.55 ±2.11
CBoW-LSPT	82.63 ±1.61	87.51 ±2.16	90.57 ±1.37
CBoW-ASPT-LSPT	89.36 ±1.01	92.67 ±0.84	94.16 ±1.68

expected to be better for smaller frame sizes. The performance of the proposed features drops for a frame size of 30 ms. On the other hand, almost similar performances are noted for frame sizes of 10 ms and 20 ms. For all further experiments, this work uses a 10 ms frame size (and 5 ms frame shift).

Table IV presents the performance of proposed features computed for three different audio interval sizes – 0.5s, 1s and 2s. We observe an improvement in classification performance for an increase in interval size. This result indicates that using larger interval sizes lead to better modeling of the spectral peak traces. However, in real scenarios, larger interval sizes will lead to poor (time) resolution of classifier decisions. On the other hand, a smaller interval size will result in poor performance. Hence, this work considers 1s audio intervals as units for classification decision as a compromise between lower resolution and better performance.

B. Performance analysis

Table V presents performance of baseline and proposed features using SVM (RBF kernel) over the GTZAN and Scheirer-Slaney datasets. The CBoW-ASPT-LSPT and MSD-ASPT-LSPT features provide best and second-best performance over these two datasets. Table VI shows the classification performance of baseline and proposed features on MUSAN dataset using bagged SVM (RBF kernel) and DNN. All baseline and proposed features show better performance over this dataset due to the availability of a large amount of training data. Even the standard deviations of F1-scores are observed to reduce significantly. The MFCC-39 turns out to be the best baseline feature. The CBoW features individually perform better than individual MSD features. MSD-ASPT-LSPT feature substantially improves upon the MSD features taken separately. However, CBoW-ASPT-LSPT stands out as the overall best performer with the DNN classifier.

Table VII presents the performance comparison of Papakostas-CNN, the best of the other four baselines, and the best of the proposed features for all three datasets. For GTZAN and Scheirer-Slaney, the performance of Papakostas-CNN is significantly lower than the best baseline and proposed features. This can be attributed to the lack of sufficient training data available in smaller datasets. However, for the larger MUSAN dataset, Papakostas-CNN outperforms all other methods. The proposed CBoW-ASPT-LSPT feature provides comparable performance on MUSAN dataset. This indicates the efficiency of proposed CBoW features in SMC.

Experiments were also performed to show the effectiveness of combining CBoW-ASPT-LSPT feature with two contemporary deep-learning based features. First is the deep bottleneck feature (DBF, henceforth), which has gained popularity in many speech processing applications in recent times [47]. DBFs are generated using a deep neural network. One of the hidden layers in this network, called the bottleneck layer, has significantly less number of nodes compared to other layers. Embeddings generated from this layer are referred to as DBF. The DBF network considered in this work has 5 hidden layers, and the middle one is the bottleneck layer. The bottleneck layer has a size of 50, the input and other hidden layers have 1313 nodes each, and the output layer has 2 nodes. MFCC (13-dimensional) features for every frame in a 1s interval are concatenated and passed as input to the DBF network. Second, feature embeddings generated from Papakostas-CNN network (Papakostas-CNN-Embed, henceforth) are used as the other deep-learning based feature. Papakostas-CNN-Embed feature is extracted from the penultimate layer of the CNN network proposed in [18] and has a dimension of 4096. The DNN architecture described in Table II is used as the classifier in this experiment.

The results obtained from these experiments are listed in Table VIII. It can be observed from the Table that Papakostas-CNN-Embed is the best performer individually. However, when the DBF and Papakostas-CNN-Embed features are separately combined with CBoW-ASPT-LSPT, the results improve. Although both early and late fusion strategies show an improvement in the performance, it can be observed that late fusion provides better results in both cases. As such, it can be said that the proposed CBoW-ASPT-LSPT features can capture some additional information that is missed by the deep learning methods, which leads to improvement in performance upon combination.

C. Discussions

The following intuitive reasoning can be proposed to explain the efficiency of the CBoW features. First, each one-dimensional element of LSPT (or ASPT) is projected to a K-dimensional posterior probability vector. Such a non-linear transformation to a higher dimensional space might induce separability of features. Second, the averaged posterior probability vectors of all peak traces are concatenated together. This concatenation leads to enhanced chances of separability

TABLE V
PERFORMANCE OF SMC USING **SVM** (RBF KERNEL) CLASSIFIER ON **GTZAN** AND **SCHEIRER-SLANEY** DATASETS. PERFORMANCE IS REPORTED AS: AVERAGE F1-SCORE ± STANDARD DEVIATION. THE TOP THREE PERFORMANCES ARE INDICATED BY: ★ (BEST), ♡ (2^{nd} BEST) AND ♣ (3^{rd} BEST).

Dataset	Baseline				Proposed					
	Khonglah-FS	Sell-FS	MFCC-39	Keum-FS	MSD-ASPT	MSD-LSPT	MSD-ASPT-LSPT	CBoW-ASPT	CBoW-LSPT	CBoW-ASPT-LSPT
GTZAN	91.02 ±1.53	**94.00 ±0.86** ♣	93.48 ±0.94	90.75 ±1.34	92.00 ±2.35	89.05 ±2.47	**94.17 ±2.19** ♡	92.58 ±2.24	92.46 ±2.90	**95.25 ±2.24** ★
Scheirer Slaney	**95.22 ±0.86** ♣	94.99 ±0.63	93.86 ±1.20	88.35 ±1.54	94.33 ±1.87	92.75 ±0.85	**96.12 ±1.78** ♡	95.03 ±1.15	93.89 ±1.40	**96.15 ±1.09** ★

TABLE VI
PERFORMANCE OF SMC USING **BAGGED-SVM** (RBF KERNEL) AND **DNN** CLASSIFIERS ON **MUSAN** DATASET. PERFORMANCE IS REPORTED AS: AVERAGE F1-SCORE ± STANDARD DEVIATION. THE TOP THREE PERFORMANCES ARE INDICATED BY: ★ (BEST), ♡ (2^{nd} BEST) AND ♣ (3^{rd} BEST).

Classifier	Baseline				Proposed					
	Khonglah-FS	Sell-FS	MFCC-39	Keum-FS	MSD-ASPT	MSD-LSPT	MSD-ASPT-LSPT	CBoW-ASPT	CBoW-LSPT	CBoW-ASPT-LSPT
Bagged-SVM	91.09 ±0.12	97.32 ±0.05	**98.29 ±0.05** ♡	95.37 ±0.06	94.28 ±0.08	93.25 ±0.07	**98.10 ±0.05** ♣	95.32 ±0.08	96.75 ±0.09	**98.99 ±0.04** ★
DNN	92.49 ±0.12	97.62 ±0.09	**98.56 ±0.09** ♡	95.40 ±0.52	94.52 ±0.59	91.88 ±1.56	**97.51 ±0.55** ♣	95.35 ±0.67	97.14 ±0.63	**98.87 ±0.25** ★

TABLE VII
PEFORMANCE COMPARISON OF BEST BASELINE AND PROPOSED FEATURES WITH **2D CNN** BASED BASELINE (PAPAKOSTAS-CNN).

Dataset	Papakostas-CNN	Best baseline	Best Proposed (CBoW-ASPT-LSPT)
GTZAN	89.76 ±3.16	94.00 ±0.86 (Sell-FS)	95.25 ±2.24
Scheirer-slaney	90.85 ±4.29	95.22 ±0.86 (Khonglah-FS)	96.15 ±1.09
Musan	99.36 ±0.76	98.56 ±0.09 (MFCC-39)	98.87 ±0.25

TABLE VIII
RESULT OF COMBINING THE PROPOSED FEATURES WITH CONTEMPORARY DEEP NETWORK BASED TECHNIQUES.

Feature	**F1-score**
CBoW-ASPT-LSPT (CAL)	98.87 ±0.25
DBF [47]	98.87 ±0.41
Papakostas-CNN-Embed [18]	99.17 ±0.2
CAL + DBF (Early fusion)	99.50 ±0.17
CAL + DBF (Late fusion)	99.61 ±0.15
CAL + Papakostas-CNN-Embed (Early Fusion)	99.66 ±0.06
CAL + Papakostas-CNN-Embed (Late Fusion)	99.80 ±0.05

in a higher-dimensional space. Third, the process of averaging posterior probability vectors over an interval possibly emphasizes the importance of class-specific components.

The results reported in Table VII indicate that the proposed feature is unable to improve upon the baseline Papakostas-CNN approach over the larger MUSAN dataset. Even though the proposed feature does provide comparable performance to Papakostas-CNN, yet it seems to fail in taking full advantage of more training data. Thus, for larger datasets, Papakostas-CNN, in the combination of the proposed feature (see Table VIII) may be a better choice for this task. However, Papakostas-CNN is unable to learn correctly in low data cases (GTZAN and Scheirer-Slaney datasets). As such, the proposed feature may be the better choice for this task in the case of smaller datasets.

In practical scenarios, classification models trained on clean speech and music data may be tested with data that have a mixture of both the classes. To gauge the effectiveness of proposed features in a mixed data scenario, a set of mixed-class (MC, henceforth) data experiments have been performed. In one set, test data contains pure speech vs. MC (PS-MC, henceforth), while in the other set, test data contains pure music vs. MC (PM-MC, henceforth). In the case of PS-MC, speech is mixed with music in the specified ratios (-20dB to $+20$dB) to generate MC data. Similarly, for the PM-MC experiment, music is mixed with speech in the specified ratios (-20dB to $+20$dB) to generate MC data. It can be observed from Fig 6 that as the amount of mixture in MC data increases, the performance gradually drops. With an increasing amount of mixture, MC data becomes increasingly similar to the pure class data considered in the experiment. For e.g., in the PS-MC case, with increasing amount of mixture, MC data becomes increasingly similar to speech, and the vice verse for PM-MC case. As such, all testing samples get recognized as any one of the two classes. The lowest performance reaches close to 50% and thus justifies this reasoning. Thus, it can be observed from Fig. 6 that the proposed feature shows stable performance with graceful degradation as long as either speech or music is the dominating content over the added noise in MC data. Thus, it can be said that the proposed feature is robust to a tolerable extent.

IV. CONCLUSION

This work proposes a novel two-stage feature extraction scheme for representing the time-frequency characteristics of an audio interval. The first stage uses a detects p prominent spectral peak traces in an interval of audio. Two sets of proposed features (MSD and CBoW) are computed from the locations (or amplitudes) of the detected peak traces. The

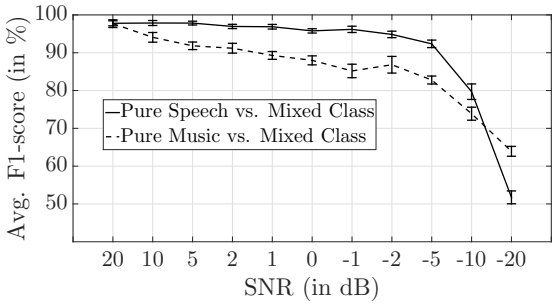

Fig. 6. Figure illustrating the performance of CBoW-ASPT-LSPT with mixed class (MC) data. The performance drops drastically for both the cases when SNR increases beyond 2 dB

performance of our proposal is validated on three standard datasets and compared with five baseline approaches. It is shown that the fusion of either MSD (i.e., MSD-ASPT-LSPT) or CBoW (i.e., CBoW-ASPT-LSPT) features provide better performance than the individual ones. Experiments show that CBoW-ASPT-LSPT stands out as the overall best feature. Further, a combination of the proposed CBoW-ASPT-LSPT feature with contemporary deep bottleneck features and deep CNN embeddings were shown to improve the classification performance, indicating that such a combination can form highly robust SMC systems.

The present proposal can be extended in the following directions. First, optimizing the number of GMM components (rather than using a fixed K) for different peak traces of speech and music might reduce the feature dimension and provide better classification performance. Second, the proposed features can be extended to increase their robustness towards mixed class data. Third, the proposed features may be employed in the task of identifying the dominant content in a mixed speech and music signal. Fourth, the present proposal focusses on the binary problem of SMC. We believe that the proposed features can be applied in other audio classification problems involving multi-category environmental sounds.

REFERENCES

[1] F. Kurth and M. Muller, "Efficient index-based audio matching," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 382–395, Feb 2008.

[2] V. A. Masoumeh and M. B. Mohammad, "A review on speech-music discrimination methods," *International Journal of Computer Science and Network Solutions*, vol. 2, Feb 2014.

[3] Y. Lavner and D. Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, no. 1, p. 239892, Jun 2009.

[4] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2, May 1996, pp. 993–996.

[5] G. Sell and P. Clark, "Music tonality features for speech/music discrimination," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2489–2493.

[6] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155–166, Feb 2005.

[7] E. Mezghani, M. Charfeddine, C. B. Amar, and H. Nicolas, "Multi-feature speech/music discrimination based on mid-term level statistics and supervised classifiers," in *Proc. of the IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, Nov 2016, pp. 1–8.

[8] P. Neammalai, S. Phimoltares, and C. Lursinsap, "Speech and music classification using hybrid form of spectrogram and fourier transformation," in *Proc. of the Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec 2014, pp. 1–6.

[9] M. Srinivas, D. Roy, and C. K. Mohan, "Learning sparse dictionaries for music and speech classification," in *Proc. of the 19th International Conference on Digital Signal Processing*, Aug 2014, pp. 673–675.

[10] B. K. Khonglah and S. R. Mahadeva Prasanna, "Speech / music classification using speech-specific features," *Digital Signal Processing*, vol. 48, no. C, pp. 71–83, Jan 2016.

[11] H. Zhang, X.-K. Yang, W.-Q. Zhang, W.-L. Zhang, and J. Liu, "Application of i-vector in speech and music classification," in *Proc. of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec 2016, pp. 1–5.

[12] J. G. A. Barbedo and A. Lopes, "A robust and computationally efficient speech/music discriminator," *J. Audio Eng. Soc.*, vol. 54, no. 7/8, pp. 571–588, 2006.

[13] E. Alexandre-Cortizo, M. Rosa-Zurera, and F. Lopez-Ferreras, "Application of fisher linear discriminant analysis to speech/music classification," in *Proc. of the EUROCON 2005 - The International Conference on "Computer as a Tool"*, vol. 2, Nov 2005, pp. 1666–1669.

[14] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification," *J. Audio Eng. Soc*, vol. 52, no. 7/8, pp. 724–739, 2004.

[15] A. Kruspe, D. Zapf, and H. Lukashevich, "Automatic speech/music discrimination for broadcast signals," in *Proc. of the INFORMATIK 2017*. Gesellschaft für Informatik, Bonn, 2017, pp. 151–162.

[16] A. Pikrakis and S. Theodoridis, "Speech-music discrimination: A deep learning perspective," in *Proc. of the 22nd European Signal Processing Conference (EUSIPCO)*, Sept 2014, pp. 616–620.

[17] D. Doukhan and J. Carrive, "Investigating the use of semi-supervised convolutional neural network models for speech/music classification and segmentation," in *Proc. of the The Ninth International Conferences on Advances in Multimedia (MMEDIA 2017)*, Venise, Italy, Apr 2017.

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

[19] C. Lim and J. h. Chang, "Enhancing support vector machine-based speech/music classification using conditional maximum a posteriori criterion," *IET Signal Processing*, vol. 6, no. 4, pp. 335–340, June 2012.

[20] S. Cheung and J. S. Lim, "Combined multi-resolution (wide-band/narrowband) spectrogram," in *Proc. of the 1991 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, April 1991, pp. 457–460.

[21] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, May 2006.

[22] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.

[23] J. O. Smith and X. Serra, "Parshl: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. of the 1987 International Computer Music Conference, ICMC*, International Computer Music Conference. Champaign/Urbana, Illinois [Michigan]: Michigan Publishing, Aug 23-26 1987, pp. 290–7.

[24] M. Lagrange, S. Marchand, and J. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1625–1634, July 2007.

[25] N. K. L. Murthy, P. C. Madhusudana, P. Suresha, V. Periyasamy, and P. K. Ghosh, "Multiple spectral peak tracking for heart rate monitoring from photoplethysmography signal during intensive physical exercise," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2391–2395, Dec 2015.

[26] Z. Zhang, Z. Pi, and B. Liu, "Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 522–531, Feb 2015.

[27] K. Seyerlehner, T. Pohle, M. Schedl, and G. Widmer, "Automatic music detection in television productions," in *Proc. of the 10th International Conference on Digital Audio Effects*, September 2007.

[28] J. Keum and H. Lee, "Speech/music discrimination using spectral peak feature for speaker indexing," in *2006 International Symposium on Intelligent Signal Processing and Communications*, Dec 2006, pp. 323–326.

[29] M. Padmanabhan, "Spectral peak tracking and its use in speech recognition," in *Proc. of the* 6^{th} *International Conference on Spoken Language Processing (ICSLP)*, October 2000, pp. 604–607.
[30] X. Xu, Yiand Sun, "Maximum speed of pitch change and how it may relate to speech," *The Journal of the Acoustical Society of America*, vol. 111, no. 3, pp. 1399–1413, 2002.
[31] J. F. Alm and J. S. Walker, "Time-frequency analysis of musical instruments," *Society for Industrial and Applied Mathematics Review*, vol. 44, no. 3, pp. 457–476, August 2002.
[32] M. J. Hawley, "Structure out of sound," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 1993, uMI Order No. not available.
[33] Z. Zhang, "Mechanics of human voice production and control," *The Journal of the Acoustical Society of America*, vol. 140(4), pp. 2614–2635, 2016.
[34] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, Nov 2008.
[35] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, Sept 2005.
[36] J. Meyer, *Structure of Musical Sound*. New York: Springer New York, 2009, pp. 23–44.
[37] L. L. Oller, *Analysis of Voice Signals for the Harmonics-to-noise Crossover Frequency*. UPC, Barcelona, Spain: KTH Royal Institute of Technology, 2008.
[38] B. K. Khonglah and S. R. M. Prasanna, "Low frequency region of vocal tract information for speech / music classification," in *Proc. of the IEEE Region 10 Conference (TENCON)*, Nov 2016, pp. 2593–2597.
[39] C. Glaser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and bayesian estimation for robust formant tracking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 224–236, Feb 2010.
[40] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010, pp. 312–314.
[41] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.
[42] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul 2002.
[43] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Apr 1997, pp. 1331–1334.
[44] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.
[45] J. Keum and H. Lee, "Speech/music discrimination based on spectral peak analysis and multi-layer perceptron," in *2006 International Conference on Hybrid Information Technology*, vol. 2, Nov 2006, pp. 56–61.
[46] E. Zhang and Y. Zhang, *F-Measure*. Boston, MA: Springer US, 2009, pp. 1147–1147.
[47] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on dnn-decision tree svm model," *Speech Communication*, vol. 115, pp. 29 – 37, 2019.