Circuits, Systems, and Signal Processing (2023) 42:6929–6950
https://doi.org/10.1007/s00034-023-02440-0

Long-Term Multi-band Frequency-Domain Mean-Crossing Rate (FDMCR): A Novel Feature Extraction Algorithm for Speech/Music Discrimination

Mohammad Rasoul Kahrizi[1] · **Seyed Jahanshah Kabudian**[1]

Received: 19 April 2022 / Revised: 21 June 2023 / Accepted: 21 June 2023 /
Published online: 9 July 2023

Abstract

Multimedia data have increased dramatically today, making the distinction between desirable information and other types of information extremely important. Speech/music discrimination is a field of audio analytics that aims to detect and classify speech and music segments in an audio file. This paper proposes a novel feature extraction method called Long-Term Multi-band Frequency-Domain Mean-Crossing Rate (FDMCR). The proposed feature computes the average frequency-domain mean-crossing rate along the frequency axis for each of the perceptual Mel-scaled frequency bands of the signal power spectrum. In this paper, the class-separation capability of this feature is first measured by well-known divergence criteria such as Maximum Fisher Discriminant Ratio (MFDR), Bhattacharyya divergence, and Jeffreys/Symmetric Kullback–Leibler (SKL) divergence. The proposed feature is then applied to the speech/music discrimination (SMD) process on two well-known speech-music datasets—GTZAN and S&S (Scheirer and Slaney). The results obtained on the two datasets using conventional classifiers, including k-NN, GMM, and SVM, as well as deep learning-based classification methods, including CNN, LSTM, and BiLSTM, show that the proposed feature outperforms other features in speech/music discrimination.

Keywords FDMCR · Speech/music discrimination (SMD) · Speech detection · Audio signal processing · Speech processing · Spectral feature extraction

✉ Mohammad Rasoul Kahrizi
kahrizi.mr@hotmail.com

Seyed Jahanshah Kabudian
kabudian@razi.ac.ir

[1] Department of Electrical and Computer Engineering, Razi University, Kermanshah, Iran

1 Introduction

Every day, massive amounts of useful data are broadcast and propagated on TV, radio channels, and the Internet, a large percentage of which consists of audio signals. These signals might have been combined with other signals that are not useful or may not be desirable for our goals, like noise, all kinds of music, or anything other than speech. Accordingly, a mechanism is required to distinguish the undesirable and the valueless audio signals from the useful and the valuable audio signals.

A system is required to discriminate valueless data from intended data to reduce the storage volume of audio files. In mobile networks, mobile operators may need to eliminate or differentiate silence from speech to minimize the amount of data transferred. Discriminating systems can also be employed to switch radio channels during commercial breaks that typically involve music. Additionally, SMD (speech/music discrimination) systems can be utilized to detect and categorize music into different classes by distinguishing it from speech in audio signals, which differs from the aforementioned applications. SMD systems can also apply to speech enhancement, noise reduction, speaker identification, speech command recognition, emotion detection, and speech-to-text conversion.

In this paper, speech discrimination was used to detect and classify speech segments of a signal comprising different types of audio. Here, speech/music discrimination does not refer to separating the voice of the singer from a melody in a music track. Additionally, a feature extraction method called Long-Term Multi-band Frequency-Domain Mean-Crossing Rate (FDMCR) was proposed to discriminate speech from music in audio signals. The characteristics of the audio signal in the frequency domain were used to calculate the FDMCR. The proposed method is a long-term feature that is robust against sudden changes in audio signals. Below are the highlights of this paper:

- Introducing a new concept of mean-crossing rate in the frequency domain
- Proposing a novel feature extraction method based on the average frequency-domain mean-crossing rate along the frequency axis for each of the Mel-scaled perceptual frequency bands of a signal spectrum
- Demonstrating the class-separation ability of the proposed feature extraction method in discriminating between speech and music using well-known divergence criteria, such as MFDR, Bhattacharyya, and SKL
- Demonstrating higher accuracy of the proposed feature extraction method compared to other features in the speech/music discriminating process using conventional classifiers (SVM, GMM, and k-NN) and deep learning methods (CNN, LSTM, and Bi-LSTM) on two popular speech-music datasets GTZAN and S&S

The rest of the article is organized as follows. In Sect. 2, some of the most recent studies and notable works related to speech/music discrimination are briefly mentioned. The proposed method is described in Sect. 3, along with relevant mathematical equations. Section 4 deals with the comprehensive performance evaluation of the proposed method. Finally, the conclusions are presented in Sect. 5.

2 Literature Review

Discriminating speech from music has various applications. One of the earliest applications of SMD systems involved real-time discrimination of speech from other contents being transmitted on FM radio channels [37]. As advertisements begin on a radio channel, the system changes the channel automatically. It was mentioned in [37] that the performance of this system is high, and the ZCR feature is used for this purpose. In [20], the SMD system was employed to analyse radio channels. Authors in reference [44] used other features (i.e., spectral, rhythmic, and harmonic features) of audio signals, especially music signals, for better classification. In [38], the low-energy measure was introduced for the SMD system. Also, using peak energy in speech signals was discussed in [34]. In [10, 30], MFCC was employed to improve discrimination.

Authors in [14, 31] used the LPC feature for speech/music discrimination. As mentioned, the application of SMD systems is not limited to a few cases. In another case, the SMD system was used for Enhanced Voice Services (EVS) [29]. Authors in [25] used a series of features obtained from empirical mode decomposition, such as mean, absolute mean, variance, skewness, and kurtosis. The proposed method in [3] is based on the Single Frequency Filtering (SFF) approach so that in audio signals, the parts associated with speech have a higher signal-to-noise ratio (SNR) in all frequencies compared to other parts which include noise.

In [12], a set of filter-based features was combined to separate speech from background sounds (noises). In that study, the discrimination accuracy increased by about 24% in all environmental conditions. In [24], features specific to human speech, such as features that represent excitation source, vocal tract, and speech syllable rate, were used for speech/music classification. In [39], features obtained from chroma vectors and combinations with other features were used to discriminate speech from music accurately. The proposed features in that study are effective and efficient.

In [40], the proposed SMD method is founded on the distinction that music carries a melody, while speech does not. The author employs deviation distribution linked with fundamental frequencies of various audio signal components to suggest a technique that can differentiate between speech and music. Additionally, [11] utilizes the Stereo-Input Mix-to-Peripheral Level (SIMPL) feature for discrimination, which is commonly utilized to approximate the speech-to-music ratio. In [41], several methods have been proposed for automatically classifying a large number of audio signals. It has been mentioned that SMD accuracy using the proposed method would be more than 94%.

Certain studies have emphasized the classification stage of the discrimination process. These studies have introduced methods that enhance the discrimination process by suggesting a new machine-learning algorithm, rather than proposing a novel feature extraction method. Among these studies, [18, 26, 32] can be mentioned in which convolutional neural networks (CNN) were applied to music detection from broadcast contents and speech/music discrimination, respectively. Also, in [17], the method proposed for the classification stage is based on the recurrent neural network (RNN) and has higher efficiency and lower error. In [5, 8, 9, 15, 19, 22, 43, 45], other methods were proposed for speech detection and SMD. Moreover, in surveys regarding speech

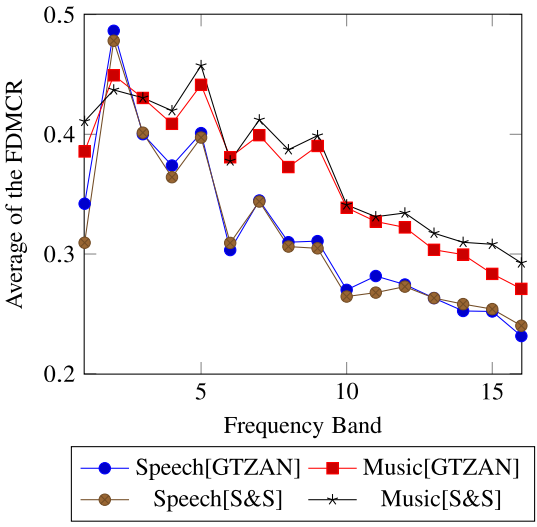

Fig. 1 The average of the FDMCR

discrimination, especially in [2, 4, 13, 33], more information can be obtained to discriminate speech from the audio signal, including history, previous studies, employed datasets, and methods associated with the discrimination process.

3 Long-Term Multi-band Frequency-Domain Mean-Crossing Rate (FDMCR)

In our extensive research, the behaviour of the frequency signal of speech and music was compared based on previous research data. The main hypothesis here is that the different behaviour of speech and music signals in the time domain (like ZCR) could have different manifestations in the frequency spectrum. Therefore, here, the frequency spectrum of speech and music signals was investigated extensively. As shown in Figs. 1 and 2, each same band of speech and music frequency spectra have different crossing rates of the mean frequency. The crossing rate of the mean frequency in the frequency spectrum of speech signals is lower than that of music. The following lines discuss the implementation steps of the proposed method.

As mentioned, the proposed method uses the frequency characteristics of audio signals. In summary, the FDMCR[1] is defined as the crossing rate of mean frequency for each pre-specified band on the frequency spectrum of an audio signal.

Here, the FDMCR was calculated by framing the primary signal to frames with 25 ms length and 50% overlap. Then as shown in Fig. 3, a short-time Fourier transform was applied to each frame (Eq. (1)). In the next step, the power spectrum for each frame was calculated (Eq. (2)). Then, it was smoothed in a 25-frame window (Eq. (3)).

[1] The source code of the FDMCR has been registered as "Long-Term Multi-band Frequency-Domain Mean-Crossing Rate (FDMCR) feature" on IEEE DataPort [21] with this DOI: "10.21227/H2NW6G".

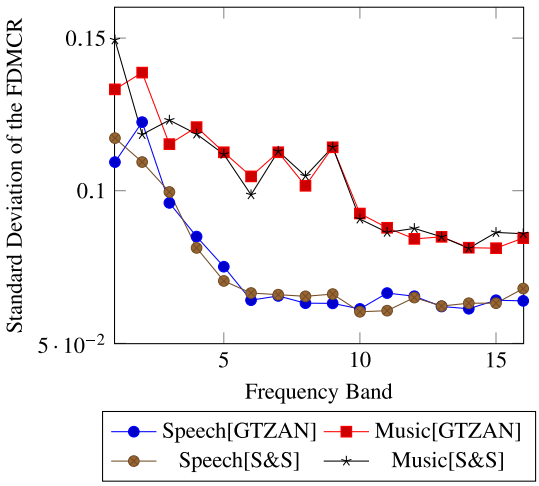

Fig. 2 The standard deviation of the FDMCR

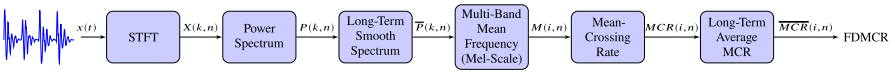

Fig. 3 The block diagram of the FDMCR

The smoothed power spectrum was subsequently partitioned into specific intervals (bands) along the frequency axis. Note that the width of these bands on the frequency axis is not uniform, and the bandwidth increases incrementally. The bands are narrower in lower frequencies and wider in higher frequencies. However, the lengths of these bands are uniform on the perceptual frequency axis and are established using Bilinear Frequency-Warping functions, as shown in Eqs. (4) or (5).

In this step, each band's mean value for the specific frame was calculated using Eq. (6). The mean-crossing rate for each band was calculated based on Eq. (7). In the final step, according to Eqs. (8, 9), mean-crossing rates obtained in a specific interval were averaged (an interval with 25 frames (312.5 ms)—12 frames before the current frame and 12 frames after it. The obtained result was the FDMCR. All steps of extracting the proposed feature from an audio signal are formulated as follows:

$$X(k,n) = \sum_{l=0}^{N_w-1} w(l)x(l+(n-1)N_{\text{sh}})e^{-j\frac{2\pi kl}{N_w}} \tag{1}$$

where X is the short-time Fourier transform (STFT) and N_w is the number of samples in the nth frame or window equal to the number of samples in a frame with a length of 25 ms. Moreover, N_{sh} is the number of signal samples that is equivalent to a frameshift half of N_w; this means that the overlap of the frames is fifty per cent here. $w(.)$ shows the window function, and k indicates the frequency-bin index for the desired frame.

$$P_x(k,n) = |X(k,n)|^2 \tag{2}$$

where P is the power spectrum for the n_{th} frame.

$$\overline{P_x(k,n)} = \frac{1}{M+1} \sum_{m=n-\frac{M}{2}}^{n+\frac{M}{2}} P_x(k,m) \tag{3}$$

where $\overline{P_x}$ is the smoothed power spectrum in an interval with $M+1$ frames, and M is an even and positive integer equal to 24. Frequency warping can be performed according to Eqs. (4) or (5), which are different in shape but equal to each other. The bilinear frequency warping is used as follows [28]:

$$\tilde{\omega} = 2 \arctan\left(\frac{1+\alpha}{1-\alpha} \tan\frac{\omega}{2}\right) \tag{4}$$

$$\tilde{\omega} = \omega + 2 \arctan\left(\frac{\alpha \sin(\omega)}{1-\alpha\cos(\omega)}\right) \tag{5}$$

in which $\tilde{\omega}$ is the warped frequency (estimate of human's perceptual frequency), which is divided into bands with equal width proportional to the number of bands: if the number of bands is eight and the sampling frequency is 8 kHz, there will be eight bands with equal width in the interval [0, Mel (4000)] on the warped frequency axis. It should be noted that bandwidths are not the same after mapping to unwarped frequency (Hz scale, calculated by the inverse of Eqs. (4) or (5)). And ω is the normal frequency in the interval $[0, \pi]$. α is the warping factor between $[-1, 1]$ and determines the degree of nonlinearity in frequency warping, which is equal to 0.3 here(approximately equals Mel scale).

$$M(i,n) = \frac{\sum_{f\in F_i} \overline{P_x}(f,n)}{|F_i|} \tag{6}$$

where $M(i,n)$ is the mean smoothed power spectrum of the nth frame and ith band, and F_i is the set of frequency bins in the ith band and $|F_i|$ is the number of frequency bins in the ith band (cardinality).

$$\mathrm{MCR}(i,n) = \frac{1}{|F_i|-1} \sum_{f\in F_i} \frac{\left|\mathrm{sign}(P_x(f,n)-M(i,n)) - \mathrm{sign}(P_x(f-1,n)-M(i,n))\right|}{2}$$

$$\text{where } \mathrm{sign}(x) = \begin{cases} +1, & \text{if } x \geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

where MCR(i,n) is the mean-crossing rate for the nth frame and ith band. $P_x(f, n)$ is the nth frame and fth frequency bin's power spectrum.

$$\overline{\mathrm{MCR}}(i, n) = \frac{1}{N+1} \sum_{m=n-\frac{N}{2}}^{n+\frac{N}{2}} \mathrm{MCR}(i, m) \tag{8}$$

$$\mathrm{FDMCR}(n) = \left[\overline{\mathrm{MCR}}(1, n), \overline{\mathrm{MCR}}(2, n), ..., \overline{\mathrm{MCR}}(B, n)\right]_{1\times B} \tag{9}$$

where $(\overline{\mathrm{MCR}})(i, n)$ is the smoothed mean-crossing rate in an interval with $N + 1$ frames. N is an even and positive integer equal to 24, and n and i are indices for frames and bands, respectively. B is the number of bands, which is equal to 16. The value of all parameters is quantified experimentally to achieve the proposed feature's best performance. The optimization algorithms like [23] can also be used to initialize parameters to obtain an optimized form.

Finally, the B-dimensional FDMCR(n) feature vector is obtained for each frame. This feature vector, corresponding to the nth frame, is given to a classifier for decision-making, and the classifier labels the frame with speech or music.

4 Experimental Protocols and Results

To measure the proposed method's efficiency in discriminating speech from music, the FDMCR is evaluated and compared with features mentioned in Sect. 2 and many other well-known features in the speech processing context. First, the separability criteria and then machine learning accuracy metrics were used to that end. Generally, several conventional types of classical and deep techniques are used for machine learning. It should be noted that the proposed method was compared with other methods under the same conditions described below. All results were achieved by running methods in MATLAB (R2020b). In addition, all features were used in the long-term form to further the authenticity of the comparison under identical conditions, meaning that if a feature is not inherently long-term, after computing the feature for each frame, the final value of the feature for the frame would be obtained by averaging among a certain number of frames. This is while in this paper, twelve frames before and after along with the current frame were used. Accordingly, the long-term features were calculated within a 25-frame time window.

For the evaluation, two well-known datasets were used for comparison: GTZAN[2] [44] with 128 30-second audio files, 64 of which are related to speech class, and the rest are related to music and S&S[3] (LabROSA) [6, 38, 46] with contains 244 15-second audio files, 140 of which are related to speech, and the rest are related to music and noise.

[2] Available: "http://opihi.cs.uvic.ca/sound/music_speech.tar.gz". Accessed: 3/13/2021.

[3] Accessible from this address: "http://www.ee.columbia.edu/~dpwe /sounds/musp/".

4.1 Comparison Based on Separability

In the first experiment, the class-separability of the proposed feature was compared with other features. Among well-known criteria for measuring the discriminability or distance between two probability distributions used here are D_{MFDR} (Maximum Fisher Discriminant Ratio) [27], D^{Bhat} (Bhattacharyya Distance), and D^{SKL} (Symmetric Kullback–Leibler Distance) [1]. The mathematical formulae of the mentioned measures are as follows:

$$x_{\text{speech}} \sim N(x; \mu_1; \Sigma_1) \tag{10}$$

$$x_{\text{music}} \sim N(x; \mu_2; \Sigma_2) \tag{11}$$

$$\Gamma = \frac{\Sigma_1 + \Sigma_2}{2} \tag{12}$$

$$\Psi = \Sigma_1^{-1} + \Sigma_2^{-1} \tag{13}$$

$$D_{\text{MFDR}} = \frac{1}{2}(\mu_1 - \mu_2)^T \Gamma^{-1}(\mu_1 - \mu_2) \tag{14}$$

$$D_{\text{Bhat}} = \frac{1}{8}(\mu_1 - \mu_2)^T \Gamma^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln\left(\frac{\det(\Gamma)}{\sqrt{\det(\Sigma_1)\det(\Sigma_2)}}\right) \tag{15}$$

$$D_{\text{SKL}} = \frac{1}{2}(\mu_1 - \mu_2)^T \Psi(\mu_1 - \mu_2) + \frac{1}{2}\text{tr}\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\right) \tag{16}$$

where I is the identity matrix. μ_1 and μ_2 are the samples mean of all feature vectors, which belong to speech and music classes, respectively, Σ_1 and Σ_2 are the sample covariance matrices. If covariance matrices of speech and music classes are assumed to be the same, all the above measures would be the same, and they would only differ in scale. However, this assumption is ignored in our computations.

The separability metrics presented in Eqs. (10–16) were derived through statistical calculations. In simpler terms, this comparison method primarily relies on the difference in the mean value of a feature between speech and music classes. A larger difference value for a feature implies that it is more effective in distinguishing speech from music.

Table 1 shows the features' class separability for the two speech and music classes on the GTZAN and S&S datasets. The results indicate that the proposed method exhibits greater discriminability across all between-class distance criteria, suggesting its superior ability to accurately classify speeches and music pieces in audio signals.

4.2 The Evaluation Based on Machine Learning Accuracy

In this step, the Equal-Error-Rate (EER) metric is used for comparing methods. This error metric equals the arithmetic mean of the False Positive Rate (FPR) and False Negative Rate (FNR). Also, we use the accuracy measure for collation methods. This metric is equal to the precision criterion. Furthermore, the F-measure or F-score is used for comparing the correctness of methods. This metric is equal to the harmonic mean of precision and recall criterion. Moreover, for a more accurate comparison of

Table 1 Comparing separability of the proposed method with other methods on the GTZAN and S&S datasets

Method	Corpus	Between-class distance criterion		
		D^{SKL}	D^{Bhat}	D^{MFDR}
AutoCorrelation	S&S	0.17742	0.02204	0.08162
	GTZAN	0.18250	0.02222	0.05669
Centroid	S&S	0.03245	0.00403	0.00390
	GTZAN	0.20214	0.02431	0.04457
Chroma HighFreq [39]	S&S	0.42438	0.04811	0.00095
	GTZAN	0.29236	0.03411	0.00006
ChromaDiff [39]	S&S	0.46550	0.05231	0.00151
	GTZAN	0.31322	0.03637	0.00001
Drugman [12]	S&S	1.45549	0.17217	0.34920
	GTZAN	1.13883	0.13438	0.22108
Energy	S&S	0.14288	0.01725	0.00433
	GTZAN	0.00582	0.00073	0.00290
Energy ratio	S&S	0.02196	0.00274	0.00923
	GTZAN	0.08546	0.01048	0.01221
Entropy	S&S	0.27536	0.03334	0.09652
	GTZAN	0.12874	0.01560	0.00080
Proposed FDMCR	S&S	**9.67128**	**1.02872**	**1.44,986**
	GTZAN	**9.68412**	**0.98527**	**1.11791**
Flatness	S&S	0.21552	0.02599	0.05774
	GTZAN	0.16433	0.01974	0.00525
LTSD [35]	S&S	2.85802	0.34194	1.32325
	GTZAN	1.92524	0.24048	0.96120
LTSV (GammaTone)	S&S	1.54539	0.15956	0.40769
	GTZAN	2.19741	0.21122	0.52657
LTSV [16]	S&S	2.03353	0.19391	0.43608
	GTZAN	2.25962	0.21154	0.48664
MBLTSV [42]	S&S	3.20428	0.34772	0.34758
	GTZAN	3.33674	0.36800	0.37335
MFCC	S&S	3.85658	0.42784	0.68933
	GTZAN	4.04613	0.41768	0.46964
Peakiness	S&S	0.28927	0.03592	0.13683
	GTZAN	0.13711	0.01658	0.00257
Pitch	S&S	0.63164	0.07876	0.31256
	GTZAN	0.23494	0.02895	0.10028
Sadjadi [36]	S&S	1.16890	0.13303	0.18281

Table 1 continued

Method	Corpus	Between-class distance criterion		
		D^{SKL}	D^{Bhat}	D^{MFDR}
	GTZAN	0.70738	0.08420	0.14021
SNR (dB)	S&S	7.62710	0.39399	0.09793
	GTZAN	0.06099	0.00755	0.01711
Spread	S&S	0.09646	0.01186	0.02683
	GTZAN	0.27416	0.03354	0.11045
ZC	S&S	0.01518	0.00189	0.00338
	GTZAN	0.27047	0.03173	0.01290

The best results are written in bold

methods, another metric called Area-Under-Curve (AUC) is used, which is equal to the AUC of the Receiver Operating Characteristics (ROC) curve. It should be noted that the rates of all comparison metrics for each method are computed in average mode and using the 10-fold way here.

4.2.1 The Comparison Based on Classical Machine Learning Methods

In this part of the evaluation, the k-NN, GMM, and SVM classifiers were used for machine learning. Generally, in cases with the k-NN classifier, 128-NN ($k = 128$) is selected to ensure better accuracy, lower error, and the practical efficiency of the k-NN with $k = 128$ compared to other types of k-NN. One feature vector was produced every 12.5-ms (frameshift size). Therefore, the number of generated feature vectors and training samples was remarkably high. Evidently, more training samples improve the efficiency of the k-NN classifier. Also, the Mahalanobis distance is used in the k-NN classifier here. In this study, GMM models with eight Gaussian components for each music and speech class were used. Moreover, the RBF kernel was selected when SVM was used.

Tables 2, 3 and Figs. 4 and 5 compare the various methods' errors and accuracy using different classifiers according to EER, F-score, and AUC criteria. Here, the AUC metric, which represents the area under the ROC curve, was calculated for each method on average.

First, the features of the GTZAN corpus were compared. As shown in Table 2 and Fig. 4, proposed method produced the best result in this section and among all evaluations using different classifiers on the GTZAN dataset. It was found that the FDMCR with the k-NN classifier has the best performance among the FDMCR results (in terms of EER, F-score, and accuracy measures), followed by the evaluations of FDMCR with SVM and GMM, respectively.

The proposed method ranked first on the GTZAN dataset according to the AUC criterion The AUC shows the average performance for all possible values of decision thresholds. Some of these decision threshold values lead to high levels of False Positive Rates (FPR) or low levels of True Positive Rates (TPR) in the ROC curve, which are not

Table 2 The results of different methods on the GTZAN dataset using k-NN, GMM, and SVM classifiers (%)

Method	Classifier	AUC	EER	F-score	Accuracy
AutoCorrelation	GMM	61.51	42.32	57.65	57.66
	KNN	61.44	42.66	57.30	57.32
	SVM	61.82	41.66	58.31	58.32
Centroid	GMM	60.22	41.79	58.18	58.19
	KNN	57.71	44.11	55.85	55.88
	SVM	61.15	42.08	57.88	57.90
ChromaHighFreq	GMM	57.29	43.04	56.92	56.94
	KNN	56.50	45.03	54.93	54.97
	SVM	57.81	43.60	56.36	56.39
ChromaDiff	GMM	57.39	43.65	56.32	56.37
	KNN	56.36	45.51	54.44	54.48
	SVM	55.52	45.91	54.06	54.08
Drugman	GMM	77.06	30.03	69.94	69.95
	KNN	80.22	26.55	73.41	73.41
	SVM	75.51	30.18	69.79	69.79
Energy	GMM	49.32	51.29	48.68	48.69
	KNN	52.10	48.19	51.76	51.82
	SVM	50.97	50.29	49.67	49.71
Energy Ratio	GMM	53.64	47.54	52.39	52.46
	KNN	51.72	48.51	51.44	51.49
	SVM	53.34	47.42	52.55	52.58
Entropy	GMM	42.04	55.82	44.16	44.18
	KNN	54.85	46.37	53.57	53.63
	SVM	46.54	53.28	46.70	46.73
Proposed FDMCR	GMM	88.14	18.56	81.42	81.41
	KNN	**89.23**	**17.01**	**82.95**	**82.99**
	SVM	89.11	17.99	81.98	81.97
Flatness	GMM	48.85	52.70	47.26	47.29
	KNN	56.93	44.40	55.56	55.60
	SVM	49.87	51.43	48.53	48.58
LTSD	GMM	83.49	24.21	75.78	75.77
	KNN	82.70	24.07	75.89	75.88
	SVM	78.19	24.50	75.49	75.52
LTSV (GammaTone)	GMM	77.52	29.47	70.51	70.53
	KNN	76.74	30.02	69.95	69.95

Table 2 continued

Method	Classifier	AUC	EER	F-score	Accuracy
	SVM	76.34	29.69	70.27	70.27
LTSV	GMM	79.90	28.06	71.88	71.96
	KNN	79.38	28.79	71.17	71.17
	SVM	79.65	28.42	71.55	71.54
MBLTSV	GMM	76.21	31.15	68.81	68.83
	KNN	74.79	33.29	66.68	66.68
	SVM	77.95	28.90	71.06	71.06
MFCC	GMM	79.46	27.63	72.35	72.35
	KNN	79.74	28.17	71.80	71.79
	SVM	79.34	26.48	73.48	73.47
Peakiness	GMM	55.01	46.70	53.27	53.29
	KNN	58.60	44.98	54.98	55.01
	SVM	56.47	45.63	54.33	54.35
Pitch	GMM	64.38	40.58	59.40	59.41
	KNN	63.66	40.64	59.31	59.35
	SVM	64.10	40.40	59.57	59.59
Sadjadi	GMM	71.45	34.41	65.56	65.57
	KNN	72.57	32.07	67.90	67.90
	SVM	63.43	38.86	61.11	61.12
SNR	GMM	59.41	44.65	54.64	55.69
	KNN	56.91	45.52	54.45	54.48
	SVM	59.67	43.46	56.50	56.52
SNR (dB)	GMM	59.68	42.69	57.28	57.31
	KNN	56.69	45.31	54.65	54.69
	SVM	59.37	43.07	56.79	56.88
Spread	GMM	62.99	41.08	58.89	58.91
	KNN	60.33	43.20	56.76	56.78
	SVM	61.81	42.18	57.78	57.79
ZC	GMM	60.70	41.85	58.13	58.14
	KNN	61.11	42.36	57.58	57.61
	SVM	61.19	42.15	57.82	57.83

The best results are written in bold

reliable working points for real-world applications. Hence, EER, F-score, or accuracy could be more realistic criteria for ranking methods.

Now we compare features on the S&S dataset. As shown in Fig. 5 and Table 3, the FDMCR produced the best results when using k-NN (in terms of EER, F-Score, accuracy, and AUC). Furthermore, the FDMCR demonstrated a better performance in both SVM and GMM results in this section of the experiments. Overall, the proposed method outperforms other methods on the S&S dataset, yielding the best results.

Table 3 The results of different methods on the S&S dataset using k-NN, GMM, and SVM classifiers (%)

Method	Classifier	AUC	EER	F-score	Accuracy
AutoCorrelation	GMM	63.85	39.40	60.21	60.36
	KNN	64.11	40.47	59.18	59.31
	SVM	64.36	39.50	60.12	60.26
Centroid	GMM	50.97	49.02	50.64	50.97
	KNN	53.14	47.70	51.97	52.24
	SVM	54.49	46.02	53.58	53.87
ChromaHighFreq	GMM	63.53	40.03	59.46	59.71
	KNN	65.45	39.19	60.41	60.54
	SVM	64.81	38.06	61.57	61.65
ChromaDiff	GMM	54.93	44.59	55.01	55.18
	KNN	65.39	39.22	60.41	60.52
	SVM	50.92	48.48	51.09	52.88
Drugman	GMM	74.98	31.85	67.73	67.74
	KNN	80.05	27.84	71.78	71.74
	SVM	74.55	31.73	67.88	67.90
Energy	GMM	49.22	50.60	49.11	49.41
	KNN	50.11	50.16	49.44	49.82
	SVM	49.86	49.32	49.69	51.11
Energy Ratio	GMM	56.99	44.55	55.10	55.36
	KNN	56.71	44.34	55.35	55.57
	SVM	57.49	44.65	55.03	55.27
Entropy	GMM	65.18	39.22	60.40	60.55
	KNN	63.59	40.69	58.94	59.10
	SVM	64.55	39.84	59.79	60.96
Proposed FDMCR	GMM	90.05	16.38	83.32	83.17
	KNN	**91.46**	**15.18**	**84.42**	**84.82**
	SVM	91.25	15.58	84.12	83.97
Flatness	GMM	63.38	40.82	58.81	59.00
	KNN	62.33	41.13	58.47	58.67
	SVM	63.04	41.35	58.29	58.49
LTSD	GMM	86.13	20.19	79.47	79.34
	KNN	85.99	20.06	79.62	79.49
	SVM	81.24	20.17	79.77	79.89
LTSV (GammaTone)	GMM	74.61	31.59	69.97	68.41
	KNN	74.06	31.95	67.65	67.64
	SVM	73.10	31.72	67.88	67.87
LTSV	GMM	78.27	29.77	69.83	69.87

Table 3 continued

Method	Classifier	AUC	EER	*F*-score	Accuracy
	KNN	77.82	29.89	69.72	69.67
	SVM	77.96	29.88	69.72	69.67
MBLTSV	GMM	73.37	33.29	66.27	66.29
	KNN	71.61	35.13	64.47	64.51
	SVM	74.51	31.82	67.76	67.75
MFCC	GMM	81.88	25.69	73.92	73.86
	KNN	82.60	25.44	74.19	74.09
	SVM	79.65	26.22	73.40	73.31
Peakiness	GMM	67.82	36.74	62.85	62.94
	KNN	67.25	37.69	61.95	62.04
	SVM	67.92	36.94	62.67	63.76
Pitch	GMM	72.24	33.88	65.70	65.71
	KNN	71.69	34.14	65.42	65.47
	SVM	71.62	33.81	65.78	66.79
Sadjadi	GMM	76.10	30.43	69.17	69.16
	KNN	77.25	29.75	69.87	69.86
	SVM	69.90	34.70	64.93	64.98
SNR	GMM	61.36	43.14	51.48	62.38
	KNN	63.85	39.46	60.15	60.31
	SVM	53.59	45.68	53.44	54.52
SNR (dB)	GMM	61.25	41.21	56.95	59.29
	KNN	63.70	39.10	60.52	60.67
	SVM	52.17	48.82	49.76	51.60
Spread	GMM	51.76	49.73	49.92	50.22
	KNN	53.00	48.38	51.26	51.56
	SVM	52.12	49.36	50.12	50.64
ZC	GMM	49.28	51.20	48.43	48.84
	KNN	55.32	46.03	53.55	53.87
	SVM	49.60	51.18	48.50	48.84

The best results are written in bold

4.2.2 The Comparison Based on Deep Learning

As mentioned in [7, 32], deep learning methods effectively identify and separate speech from music. Therefore, we intend to use these methods to compare and evaluate various methods' performance, including our proposed method.

This section compares the three features that had the best results in previous comparisons using deep learning methods. As mentioned in [7, 32], using deep learning methods with image-based features to discriminate speech from music has shown

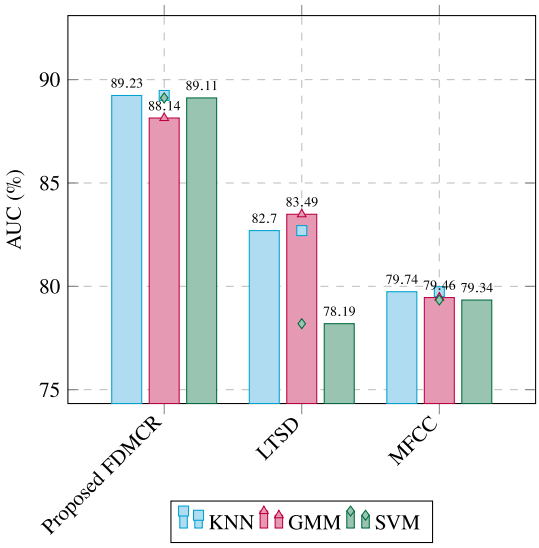

Fig. 4 Comparing the AUC of the top three superior methods using different classifiers on the GTZAN dataset

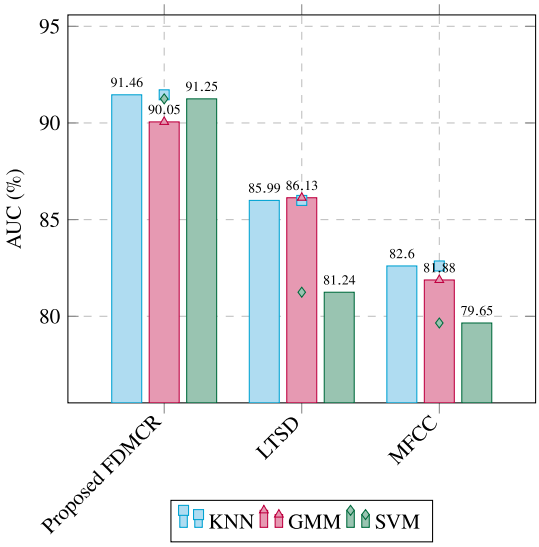

Fig. 5 Comparing the AUC of the top three superior methods with different classifiers on the S&S dataset

promising results. However, we compared our proposed feature with two other features using deep learning methods while all features are audio-based (not image-based) using deep learning methods, including CNNs and recurrent neural networks (RNNs).

Table 4specifies the characteristics of the deep learning methods used in this study. The network architecture of deep learning methods has a significant impact on the performance and results of learning methods. Here, the input of learning networks was in the form of a window of the desired feature for a 10-frame neighbourhood (Five frames before and after the desired frame).

Table 4 The architecture and characteristics of the used deep learning methods

Parameters	Methods			
	BiLSTM	CNN	CNNP [32]	LSTM
Layers	InputLayer BilstmLayer BilstmLayer FullyConnectedLayer SoftmaxLayer ClassificationLayer	InputLayer Convolution2dLayer BatchNormalizationLayer ReluLayer MaxPooling2dLayer Convolution2dLayer BatchNormalizationLayer ReluLayer MaxPooling2dLayer Convolution2dLayer BatchNormalizationLayer ReluLayer Convolution2dLayer BatchNormalizationLayer ReluLayer MaxPooling2dLayer DropoutLayer FullyConnectedLayer SoftmaxLayer WeightedClassificationLayer	InputLayer Convolution2dLayer ReluLayer MaxPooling2dLayer BatchNormalizationLayer Convolution2dLayer ReluLayer MaxPooling2dLayer BatchNormalizationLayer Convolution2dLayer ReluLayer MaxPooling2dLayer BatchNormalizationLayer FullyConnectedLayer ReluLayer DropoutLayer FullyConnectedLayer ReluLayer DropoutLayer FullyConnectedLayer SoftmaxLayer ClassificationLayer	InputLayer LstmLayer FullyConnectedLayer SoftmaxLayer ClassificationLayer

Table 4 continued

Parameters	Methods			
	BiLSTM	CNN	CNNP [32]	LSTM
MaxEpochs	20	25	16	100
MiniBatchSize	64	128	128	27
Solver	Adaptive moment estimation (ADAM)	Adaptive moment estimation (ADAM)	Stochastic gradient descent with momentum (SGDM)	Adaptive moment estimation (ADAM)
Other	InitialLearnRate=1e−3 LearnRateDropFactor=0.1 LearnRateSchedule=piecewise LearnRateDropPeriod=5 NumHiddenUnits=200	InitialLearnRate=3e−4 LearnRateDropFactor=0.1 LearnRateSchedule=piecewise LearnRateDropPeriod=20	InitialLearnRate=1e−3 LearnRateDropFactor=0.1 Momentum=0.9	InitialLearnRate=1e−3 LearnRateDropFactor=0.1 NumHiddenUnits=100 GradientThreshold=1

The best results are written in bold

Table 5 The results of the top three features on the GTZAN dataset using deep learning methods (%)

Method	Classifier	EER	F-score	Accuracy
Proposed FDMCR	BiLSTM	21.09	78.88	78.88
	CNN	17.04	82.96	82.95
	CNNP	**16.41**	**83.57**	**83.55**
	LSTM	21.74	78.22	78.21
LTSD	BiLSTM	20.48	79.49	79.48
	CNN	20.10	79.87	79.84
	CNNP	19.69	80.28	80.26
	LSTM	19.67	80.30	80.28
MFCC	BiLSTM	26.71	73.25	73.24
	CNN	21.94	78.02	78.01
	CNNP	20.28	79.69	79.68
	LSTM	29.78	70.13	70.28

The best results are written in bold

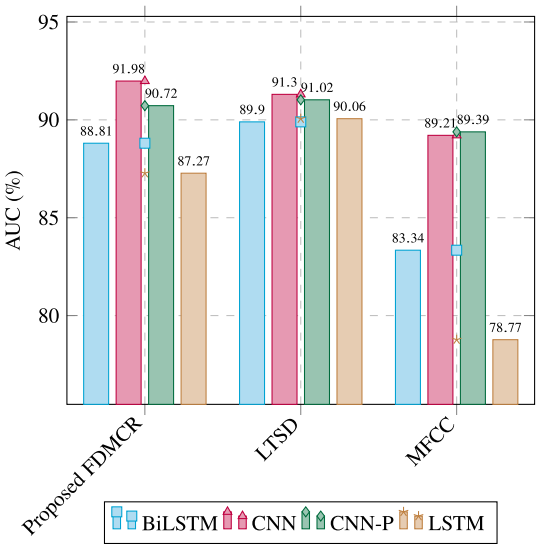

Fig. 6 Comparing the AUC of methods with different deep learning techniques on the GTZAN dataset

As in the previous section, we examined the desired methods on the GTZAN and S&S datasets. As shown in Table 5 and Fig. 6, the proposed method performed best when the GTZAN dataset is used. Also, the FDMCR produced the best results compared to other features when the S&S dataset is used for comparison, as shown in Table 6 and Fig. 7.

Compared to the results of other classical methods, like the results of comparisons in [7, 32], our results indicated that deep methods often outperform classical methods. Nonetheless, it was found that the results are not desirable in some cases because deep network architectures are very influential in this type of learning performance. In addition, deep learning methods are more compatible with image-based methods

Table 6 The results of top three features on the S&S dataset using deep learning methods (%)

Method	Classifier	EER	*F*-score	Accuracy
Proposed FDMCR	BiLSTM	18.23	81.41	81.23
	CNN	**14.98**	**84.80**	**84.67**
	CNNP	16.06	83.65	83.51
	LSTM	19.42	80.28	80.13
LTSD	BiLSTM	16.42	83.28	83.14
	CNN	14.99	84.79	**84.67**
	CNNP	15.63	84.10	83.97
	LSTM	16.37	83.34	83.20
MFCC	BiLSTM	24.23	75.48	75.39
	CNN	17.66	82.00	81.83
	CNNP	17.33	82.37	82.24
	LSTM	27.43	72.19	72.11

The best results are written in bold

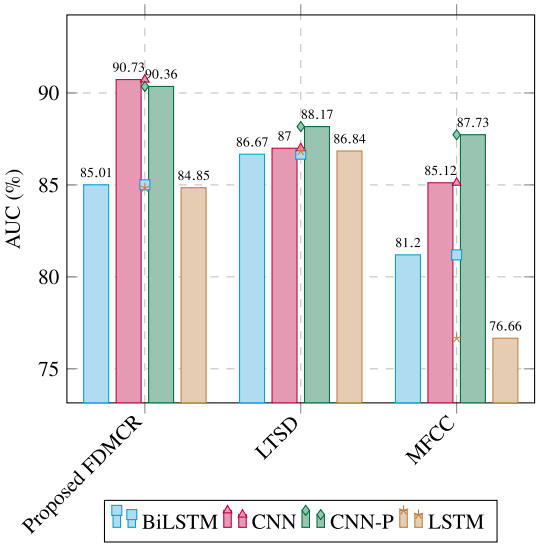

Fig. 7 Comparing the AUC of methods with different deep learning techniques on the S&S dataset

structurally and functionally. Therefore, these learning methods should not be expected to perform much better when audio-based (non-image-based) features are used.

5 Conclusion and Future Suggestions

This paper proposed a new feature extraction method called Long-Term Multi-band Frequency-Domain Mean-Crossing Rate (FDMCR) based on the new concept of mean-crossing rate in the frequency domain.

To prove the efficiency of the proposed method, first, the capability of the proposed feature for class discrimination was measured using famous divergence criteria such as Maximum Fisher Discriminant Ratio (MFDR), Bhattacharyya divergence, and Jeffreys/Symmetric Kullback–Leibler (SKL) divergence. This feature was then applied to the speech/music discrimination problem using conventional and deep learning-based classifiers on two popular speech-music datasets, GTZAN and S&S.

It was shown that the proposed feature in this paper leads to more separability between speech and music classes and performs better in the evaluations than other features. The proposed system's high computational complexity and memory consumption in the deep learning stage pose a limitation, given that a speech/music discrimination system should typically be fast and have low computational overhead. To address this issue, deep neural networks with fixed-point weights or approximate/stochastic computations can be utilized. Additionally, training deep learning systems on large speech-music datasets presents another challenge. One approach to tackle this problem is to divide the dataset into smaller parts and train a deep classifier on each part. The output of these classifiers can then be optimally combined using various ensemble learning methods.

To enhance the system's efficiency, one potential future approach is to combine the proposed feature with feature vectors from other algorithms. In addition, different algorithms for dimensionality reduction or feature selection must be used to reduce redundancy in the combined vector after combining vectors.

References

1. K.T. Abou-Moustafa, F.P. Ferrie, A note on metric properties for some divergence measures: The gaussian case. in Asian Conference on Machine Learning, pp. 1–15 (2012)
2. F. Alías, J. Socoró, X. Sevillano, A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Appl. Sci. **6**(5), 143 (2016)
3. G. Aneeja, B. Yegnanarayana, Single frequency filtering approach for discriminating speech and nonspeech. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(4), 705–717 (2015)
4. M. Anusuya, S. Katti, Front end analysis of speech recognition: a review. Int. J. Speech Technol. **14**(2), 99–145 (2011)
5. R.G. Balamurali, C. Rajagopal, Speech/music discrimination (2017). US Patent 9,613,640
6. A.L. Berenzweig, D.P. Ellis, Locating singing voice segments within music signals. in Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575), pp. 119–122 (2001)
7. M. Bhattacharjee, S.M. Prasanna, P. Guha, Speech/music classification using features from spectral peaks. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1549–1559 (2020)
8. G.K. Birajdar, M.D. Patil, Speech and music classification using spectrogram based statistical descriptors and extreme learning machine. Multimed. Tools Appl. **78**(11), 15141–15168 (2019)
9. G.K. Birajdar, M.D. Patil, Speech/music classification using visual and spectral chromagram features. J. Ambient Intell. Hum. Comput. **11**, 1–19 (2019)
10. M.J. Carey, E.S. Parris, H. Lloyd-Thomas, A comparison of features for speech, music discrimination. in 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258), vol. 1, pp. 149–152 (1999)
11. A. Chen, M.A. Hasegawa-Johnson, Mixed stereo audio classification using a stereo-input mixed-to-panned level feature. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 2025–2033 (2014)
12. T. Drugman, Y. Stylianou, Y. Kida, M. Akamine, Voice activity detection: merging source and filter-based information. IEEE Signal Process. Lett. **23**(2), 252–256 (2015)

13. S. Duan, J. Zhang, P. Roe, M. Towsey, A survey of tagging techniques for music, speech and environmental sound. Artif. Intell. Rev. **42**(4), 637–661 (2014)
14. K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, Speech/music discrimination for multimedia applications. In: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), vol. 4, pp. 2445–2448 (2000)
15. G. Fuchs, A robust speech/music discriminator for switched audio coding. in 2015 23rd European Signal Processing Conference (EUSIPCO), pp. 569–573 (2015)
16. P.K. Ghosh, A. Tsiartas, S. Narayanan, Robust voice activity detection using long-term signal variability. IEEE Trans. Audio Speech Lang. Process. **19**(3), 600–613 (2010)
17. P. Gimeno, I. Viñals, A. Ortega, A. Miguel, E. Lleida, Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. EURASIP J. Audio Speech Music Process. **2020**(1), 1–19 (2020)
18. B.Y. Jang, W.H. Heo, J.H. Kim, O.W. Kwon, Music detection from broadcast contents using convolutional neural networks with a mel-scale kernel. EURASIP J. Audio Speech Music Process. **2019**(1), 11 (2019)
19. M. Joshi, S. Nadgir, Extraction of feature vectors for analysis of musical instruments. in 2014 International Conference on Advances in Electronics Computers and Communications, pp. 1–6 (2014)
20. S. Kacprzak, B. Chwiećko, B. Ziółko, Speech/music discrimination for analysis of radio stations. in 2017 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–4 (2017)
21. M.R. Kahrizi, Long-term multi-band frequency-domain mean-crossing rate (fdmcr) feature. https://doi.org/10.21227/H2NW6G
22. M.R. Kahrizi, S.J. Kabudian, Long-term spectral pseudo-entropy (ltspe): a new robust feature for speech activity detection. J. Inf. Syst. Telecommun. (JIST) **6**(4), 204–208 (2018). https://doi.org/10.7508/jist.2018.04.003
23. M.R. Kahrizi, S.J. Kabudian, Projectiles optimization: A novel metaheuristic algorithm for global optimiaztion. Int. J. Eng. (IJE) IJE Trans. A Basics **33**(10), 1924–1938 (2020). https://doi.org/10.5829/ije.2020.33.10a.11
24. B.K. Khonglah, S.M. Prasanna, Speech/music classification using speech-specific features. Digit. Signal Process. **48**, 71–83 (2016)
25. B.K. Khonglah, R. Sharma, S.M. Prasanna, Speech vs music discrimination using empirical mode decomposition. in 2015 Twenty First National Conference on Communications (NCC), pp. 1–6 (2015)
26. A.A. Khudavand, S. Chikkamath, S. Nirmala, N. Iyer, Music/non-music discrimination using convolutional neural networks, in *Soft Computing and Signal Processing*. ed. by V.S. Reddy, V.K. Prasad, J. Wang, K.T.V. Reddy (Springer Singapore, Singapore, 2021), pp.17–28
27. S.J. Kim, A. Magnani, S. Boyd, Robust fisher discriminant analysis. in: Advances in neural information processing systems, pp. 659–666 (2006)
28. A. Makur, S.K. Mitra, Warped discrete-fourier transform: Theory and applications. IEEE Trans. Circ .Syst. I Fundam. Theory Appl. **48**(9), 1086–1093 (2001)
29. V. Malenovsky, T. Vaillancourt, W. Zhe, K. Choo, V. Atti, Two-stage speech/music classifier with decision smoothing and sharpening in the evs codec. in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5718–5722 (2015)
30. O.M. Mubarak, E. Ambikairajah, J. Epps, Novel features for effective speech and music discrimination. in 2006 IEEE International Conference on Engineering of Intelligent Systems, pp. 1–5 (2006)
31. J.E. Muñoz-Exposito, S. Garcia-Galan, N. Ruiz-Reyes, P. Vera-Candeas, F. Rivas-Peña, Speech music discrimination using a single warped lpc-based feature. in Proc. ISMIR, vol. 5, pp. 16–25 (2005)
32. M. Papakostas, T. Giannakopoulos, Speech-music discrimination using deep visual feature extractors. Expert Syst. Appl. **114**, 334–344 (2018)
33. G. Peeters, A large set of audio features for sound description (similarity and classification) in the cuidado project (2004)
34. J. Pinquier, J.L. Rouas, R. André-Obrecht, A fusion study in speech/music classification. in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., vol. 2, pp. II–17 (2003)
35. J. Ramırez, J.C. Segura, C. Benıtez, A. De La Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3–4), 271–287 (2004)
36. S.O. Sadjadi, J.H. Hansen, Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process. Lett. **20**(3), 197–200 (2013)

37. J. Saunders, Real-time discrimination of broadcast speech/music. in 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 2, pp. 993–996 (1996)
38. E. Scheirer, M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator. in 1997 IEEE international conference on acoustics, speech, and signal processing, vol. 2, pp. 1331–1334 (1997)
39. G. Sell, P. Clark, Music tonality features for speech/music discrimination. in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2489–2493 (2014)
40. B. Thompson, Discrimination between singing and speech in real-world audio. in 2014 IEEE Spoken Language Technology Workshop (SLT), pp. 407–412 (2014)
41. W.H. Tsai, C.H. Ma, Automatic speech and singing discrimination for audio data indexing. in Big Data Applications and Use Cases, pp. 33–47. (Springer, 2016)
42. A. Tsiartas, T. Chaspari, N. Katsamanis, P.K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos, S. Narayanan, Multi-band long-term signal variability features for robust voice activity detection. in Interspeech, pp. 718–722 (2013)
43. N. Tsipas, L. Vrysis, C. Dimoulas, G. Papanikolaou, Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination. Multimed. Tools Appl. **76**(24), 25603–25621 (2017)
44. G. Tzanetakis, P. Cook, Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002)
45. E. Wieser, M. Husinsky, M. Seidl, Speech/music discrimination in a large database of radio broadcasts from the wild. in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2134–2138 (2014)
46. G. Williams, D.P. Ellis, Speech/music discrimination based on posterior probability features. in Sixth European Conference on Speech Communication and Technology (1999)