EURASIP Journal on Audio, Speech, and Music Processing

SOFTWARE

Open Access

Singer identification model using data augmentation and enhanced feature conversion with hybrid feature vector and machine learning

Serhat Hizlisoy[1*], Recep Sinan Arslan[1] and Emel Çolakoğlu[2]

Abstract

Analyzing songs is a problem that is being investigated to aid various operations on music access platforms. At the beginning of these problems is the identification of the person who sings the song. In this study, a singer identification application, which consists of Turkish singers and works for the Turkish language, is proposed in order to find a solution to this problem. Mel-spectrogram and octave-based spectral contrast values are extracted from the songs, and these values are combined into a hybrid feature vector. Thus, problem-specific situations such as determining the differences in the voices of the singers and reducing the effects of the year and album differences on the result are discussed. As a result of the tests and systematic evaluations, it has been shown that a certain level of success has been achieved in the determination of the singer who sings the song, and that the song is in a stable structure against the changes in the singing style and song structure. The results were analyzed in a database of 9 singers and 180 songs. An accuracy value of 89.4% was obtained using the reduction of the feature vector by PCA, the normalization of the data, and the Extra Trees classifier. Precision, recall and f-score values were 89.9%, 89.4% and 89.5%, respectively.

Keywords Singer identification, Octave-based spectral contrast, Mel-frequency cepstral coefficients, Extra Trees classifier

1 Introduction

Although music and song are often thought of as the same thing, they actually consist of different sounds. While the song expresses certain emotions with melodious human voices, music is an art that tries to express these emotions with instrumental sounds within certain rules. Music is one of the most widely used multimedia content on the Internet [1]. Music information retrieval (MIR) covers most audio analysis tasks such as music genre classification [2], song recognition [3], singer identification [4], instrument recognition [5], and emotion recognition from music [6]. Songs are defined by mixing human vocal frequencies with instrumental music. Because the song is a mixture of these frequencies, often the frequency of the singer's voice can be matched with the frequency of any instrumental music [7].

Major changes in technology have also affected the digital cloud, giving rise to millions of songs. Due to a large amount of data, some related songs of interest to the listener may disappear in the end. In such cases, singer knowledge can be accepted for recommendation systems as a primary feature. Most of the digital songs available in online music stores in recent years are classified and labeled according to album name, music genre, musical sentiment, and singer information. The singer identification and classification process is also highly studied by the research community, also called the MIREX group

*Correspondence:
Serhat Hizlisoy
serhathizlisoy@kayseri.edu.tr
[1] Department of Computer Engineering, Faculty of Engineering, Architecture and Design, Kayseri University, Kayseri, Turkey
[2] Calculated Sciences and Engineering, Graduate School of Education, Kayseri University, Kayseri, Turkey

[8]. While it is easy to find a song using the singer name, song title, or a few keywords within the song, it can be quite difficult to identify a singer using a music excerpts. In order to prevent dependency on access to keywords, singer identification systems have begun to be developed that find the name of the singer through a small piece of music. Today, these systems are used to classify millions of irregular music data [1]. In addition, these systems also help identify similarities between different singers and detect copyright infringement. As an example of this system, we can say the commercial application made by Shazam Entertainment [9]. These applications automate the task of identifying the given song and its singer through popularized search algorithms. With the song classification based on the singer, it recognizes the songs and recommends them to the listeners according to their favorite singer [10]. An example of these music recommendation systems is the Spotify [11] application. However, these applications still have not reached the desired level, because there is usually a limited number of training samples for each singer. Moreover, a singer's style changes over time, and samples change in style from one song to another [12]. As another reason for this, it can be said that singers use their voices differently in each song. This makes the modeling of the problem difficult and causes high accuracy in recognition not to be achieved. Apart from this, the use of instrumental music in the song in a way that suppresses the voice of the singer may also reduce the success of the system. In such a situation, it may be difficult for the singer's voice to separate from the song. In order to overcome such a problem, optimum features should be selected, and analysis should be made according to these features.

While the task can be manageable for people when given a small number of singers, it gets harder as the number of singers starts to increase, which requires knowing the songs of a large number of singers. Therefore, successfully modeling singer classification can have benefits for the music industry. In contrast, people can easily identify the singer from that singer's new song if they know the singer. Because people recognize a singer using a small part of the voice, thanks to their perceptual structure and neural training mechanisms. In addition, each singer's voice has certain acoustic features that represent it, such as fingerprints. These features, which facilitate identification, are formed by many factors such as the shape and size of the singer's vocal tract, chin movement, tongue, and teeth.

Singer identification is a challenging task compared to speaker identification. Since the speaker identification task is largely depends on the speaker's distinctive features, it can be easily accomplished with a comprehensive analysis of these features. Some distinctive patterns such as pauses, phonemes, and unit separation can be observed while processing the speech signal [13, 14]. On the other hand, singing is a continuous speech that is a conscious change in pitch and vocal tract behavior [15]. In particular, considering trained singers, different singing styles and differences in perceived intensity, pitch, and timbre from the singer's voice further diversify the acoustic differences between speaking and singing [16]. Also, the instrumental music behind the songs makes it difficult to identify the singer. This situation is the most important obstacle to the determination of the features required for a robust singer identification system.

The process of estimating a singer's perceptual characteristics for the researcher to design an efficient system for singer recognition is ambiguous. So, many features useful for speaker recognition can also properly classify singers. In our study, features that have been used in many fields, such as speech recognition, speaker recognition, emotion recognition, and singer recognition proven to be successful, have been tested using combinations of various feature vectors, both separately and together. Taking these features together powerfully explains the precise definition of the singer's voice. These features are mel-frequency cepstral coefficients (MFCCs), spectrogram, beats, contrast, cq, cents, stft, centroid, bandwidth, and rmse, which define the vocal production of the singer, i.e., the vocal system including vocal folds, vocal tract, and the respiratory parts [17].

The task of identifying singers has gained a lot of importance in recent years, and many researchers have done research on this subject in the last two decades [10, 18, 19]. However, when these studies are examined, it is seen that there is not much research on Turkish songs and singers. Turkish songs, which differ from today's songs performed in other languages with their unique melodies, genres, and makams, are worth investigating. Therefore, this study focused on Turkish songs and singers. In the first step of the study, an inhomogeneous dataset containing the songs of popular singers in Turkey was created and contributed to the literature. In this dataset, there are 20 songs per singer, and this situation can cause generalization failure, poor model performance, low accuracy, and reliability problems in the analysis. Thus, 10 data augmentation with resampling methods have been applied to the dataset to improve performance of machine learning models and expand the dataset size.

In this study, three different low-level features were extracted and combined for singer identification. Thanks to merging, feature vectors containing more expressive features of the song were produced. Thus, it was observed that higher accuracy was achieved than a single feature extraction method. In order to achieve the best results, tests were carried out with different machine learning

methods. It was aimed to achieve more successful results compared to previous studies. 30-second excerpts were used instead of longer songs to fulfill the task of identifying the singer, and logistic regression was used as a classifier in the proposed model.

Extra trees is an ensemble supervised learning approach that uses tree-like structures to decide the target class of generated feature vectors. After the features were extracted and selected from the data, the model was trained with the training data and then tested with the test data. With this process, an accuracy of 89.4% was achieved, which means the prediction is highly accurate.

The main contributions of this study to singer identification research can be summarized as follows:

- Generation of feature vectors containing more meaningful data of the song by using more than one feature extraction method.
- Creating a dataset of Turkish singers and songs to test the proposed model.
- Examining different feature extraction methods (mfcc, spectrogram, rmse, centroid, contrast, beats, stft, bandwidth, etc.)
- Comparative evaluation of the proposed approach with different dataset, parameters, methods, and feature vectors.

Experimental results show that the proposed system is robust and independent of song genre and singing style. The rest of this article is organized as follows. Section 2 explains previous studies on the subject. Section 3 describes the proposed model and classifier model for the singer identification method. Section 4 explains the experimental design and talks about the created Turkish Music Database, the extracted acoustic features, and their normalization process with experimental results. In Section 5, the effects of used features and classifiers on performance are discussed and compared.

2 Related works

Singer identification is one of the areas that has attracted attention for many years and has been studied by many researchers. These researchers aimed to find the singer who sang that song by analyzing a song with various methods. The basis of the research is based on audio signal processing and was initially tried by applying speech recognition methods. In recent years, with the rapid transition of the music industry to cloud-based systems, the importance of machine learning methods used to customize, automate, and categorize services has increased. In addition, interest in research areas such as singer identification [4], music emotion recognition [20], and music genre classification [2] used in music recommendation systems [11] has increased significantly.

In one of the first singer identification studies, databases with 10 singers and 21 singers, in which instrumental and singer voices are not separated, were used. In this study, Whitman et al. (2001) [21] obtained MFCC feature vectors from the data set and classified them with ANN and SVM. In another study, Kim and Whitman (2002) [22] used warped linear prediction coefficients (WPC) and linear prediction coefficients (LPC) as feature vectors and GMM and SVM as classifiers for singer identification on 200 songs of 17 different singers. Ellis (2007) [23] created the Artist20 dataset from 1413 songs by 20 pop music singers. In the identification task, a simple Gaussian classifier model using randomly sampled MFCC vectors and chroma vectors from each singer in this dataset achieved higher accuracy than can be achieved with MFCCs alone.

Patil et al. (2012) [24] applied the MFCC features together with the MFCC-based cepstral mean subtraction features on a 500-song dataset with a polynomial classifier to examine the singer identification problem. Tsai and Lee (2012) [25] proposed a very different approach to singer identification. In this approach, the attributes obtained from the singer's voices were combined with the lyrics of the song to find out who the singer was. Experimental results showed that the model trained using only lyrics performed poorer in recognizing singers compared to the system trained using acapella song recordings. D. Dharini and Revathy (2014) [26] used perceptual linear prediction (PLP) as features for singer identification and clustered with K-means. In terms of duration, they worked on 20-s song samples.

In this study by Saurabh and Bhirud (2014) [27], it is seen that noisy music, noise filtered music and finally North Indian classical music in which only the Tanpura instrument is used in the background. Three different systems, consisting of 10, 5, and 3 singers, in which each singer has 10 5-s songs, were tested. Later, features such as roll-off, brightness, roughness, and irregularity, which fall under the timbre category and are thought to play an important role in identifying a singer, were obtained from these data. The best accuracy results were obtained in the model in which three singers were tested. Masood et al. (2016) [28] identified Indian singers with good accuracy using neural network by extracting spectral features such as MFCC from clipped song at 2-s length. However, in the field of singer identification, deep learning-based research is still rare. Shen et al. (2019) [1] introduced a new deep learning approach based on multilayer LSTM and MFCC features, inspired by the success of recurrent neural networks in acoustic modeling, to recognize the singer of a song in big datasets.

Murthy et al. (2018) [8] proposed a technique for identifying singer from song by applying deep neural network (DNN) to linear predictive cepstral features and MFCCs. Random forest (RF) method with deep neural network is also used. The dataset contains 20 singers and 100 songs in 5 s in Hindi from each singer, which means 500 s of audio data for each singer.

Nasrullah and Zhao (2019) [12] obtained features from audio spectrograms containing frequency content over time and used them as input in convolutional recurrent neural network to form a basis for singer identification with deep learning. Representing the music data obtained using the Artist20 dataset as a spectrogram allows the convolutional layers to learn the spherical structure and the repetitive layers to learn the temporal structure.

Biswas and Solanki (2021) [7] thought that there could be different approaches that can be done, such as detecting the peak values of the sound in that music, in order to achieve the goal of separating the audio 36 frequency range from the music signal. For this purpose, spectral features such as spectral bandwith, rms, 37 spectral centroid, and MFCCs were extracted from the dataset containing 8 singers and 18 songs belonging to each singer and used for classification with important machine learning methods. The best results have been achieved with the multilayer perceptual neural network (MLP), where hyperparameter tuning can also be done.

Costa et al. (2017) [29] used the Latin Music Database, ISMIR 2004, and the African music collection as datasets in their study. The attributes are spectrograms, RLBP, acoustic attributes (rhythm patterns (RP), statistical spectrum descriptors (SSD), and rhythm histograms (RH)), and the classifiers are CNN and SVM. The outputs obtained with these two classifiers were combined with the late fusion strategy. In the African music database, CNN outperformed all results reported in the literature, while in the LMD database, the combination of a CNN with an RLBP-trained SVM achieved a recognition rate of 92%.

Li et al. (2021) [30] proposed a model called KNN-Net. It is basically a CRNN-based study with attention layers. Unlike other deep neural networks that use the softmax layer as the output layer here, they used KNN as a more interpretable layer to output target singer tags. Artist20, singer32, and singer60 are the datasets used in the study. Spectrograms, on the other hand, are attributes. With the attention-CRNN-KNN model, 85% accuracy value was obtained in the Artist20 dataset.

In their study, Sharma et al. (2019) [31] extracted singing vocals from polyphonic songs using the Wave-U-Net-based sound source separation approach. They also tried harmonic/percussive and CNN-based methods for this process. An i-vector-based system was used to model singer characteristics. The i-vector is a factor analysis approach that represents the dominant speaker information in terms of a low-dimensional vector. MFCC features have been extracted from the dataset. The features were then subjected to cepstral mean normalization of variance (CMVN) before the models were developed. A universal background model (UBM) with 1024 Gaussian components and a total variability matrix (T-matrix) with a factor of 400 were used for i-vector inference. An accuracy of 89.97% was achieved with six-fold cross-validation.

Zhang et al. (2022) [32] used the Artist20 dataset in their study. The best result obtained is 81% F1 value. Here, a model is proposed that uses convolutional recurrent neural network (CRNN) and uses frame-level sound properties (mel-spectrogram), intermediate features (melodiousness, articulation, rhythmic complexity, rhythmic stability, dissonance, tonal stability, and modality), and timbre properties (x-vector) of spectrograms to identify the singer.

In the study of Murthy et al. (2021) [8], two different datasets were used, artist20 and a database of Indian popular singers consisting of 20 singers. Genetic algorithm-based feature selection (GAFS) approach was preferred in feature selection. Two different classification techniques, artificial neural networks (ANNs) and random forest (RF), are discussed on features (MFCC, LPCC, SDC, chroma). In addition, the spectrograms and chromagrams of the audio clips are fed directly into the CNN for classification. In MFCCs + LPCCs + SDCs + chroma (with genetic algorithm), an accuracy of 61.69% (Artist20) was obtained with the random forest classifier. The CNN value of Artist20 is quite low. In the Indian dataset, 75.5% accuracy was obtained with CNN.

There are few studies on singer identification where maximum accuracy is achieved on less complex datasets with logistic regression. Therefore, these techniques are especially used in this study to determine the identity of the singer and will be detailed in the following sections.

3 Methodology

3.1 Proposed model

Within the scope of this study, a unique model for singer recognition was designed, and the training and testing processes of this model were shared in detail. The structure of this designed model is shown in Fig. 1. The model has three main and five sub-process steps, and each of them is explained in detail in this section.

3.2 Data augmentation

Data augmentation is creating copies of data using specific methods to prevent models' overfitting during machine learning model training. It is necessary to use

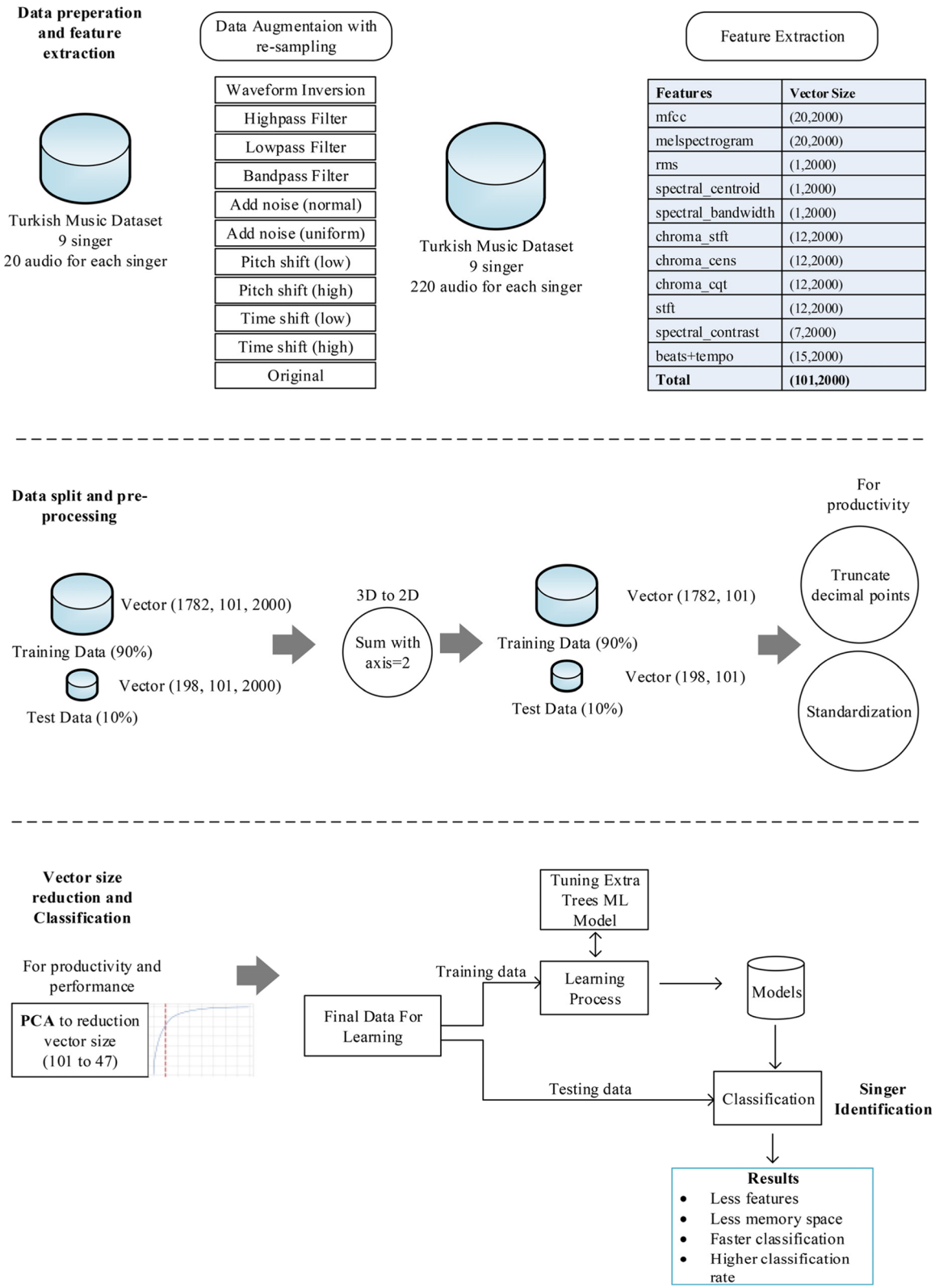

Fig. 1 A proposed model of singer identification

different data augmentation methods depending on the type of data such as image or audio. Some of the data augmentation methods used for the audio signal shown in Fig. 1 are as follows: waveform inversion, high-pass filter, low-pass filter, band-pass filter, time shift (low, high), pitch shift (low, high), and noise addition (normal, uniform).

In this study, a dataset consisting of 180 songs in total, belonging to 9 different Turkish singers, 20 songs of each singer, was prepared. Although 30-s recordings were sufficient for singer recognition, the number of data might not have been enough to achieve high performance. Therefore, the number of songs has been increased from 20 to 220 for each singer, including their original versions, using these 10 different data augmentation methods.

The waveform inversion filter is applied to invert the sound wave by 180°. At the end of this process, while the difference cannot be understood in a single waveform, it can be understood when combining waveforms [33]. The high-pass filter allows high frequencies to pass through in the audio signal. It prevents the transmission of low frequencies [34]. The signals cleaned with this filter are generally background noise or low-frequency unwanted signals. The low-pass filter allows the low frequency of the signal to pass. In this case, high-frequency values are ignored [34]. The aim is to attenuate signals above a certain frequency. Researchers working in the field of audio generally use the low-pass filter to eliminate high-frequency noise and unwanted high-pitched sounds in the sound. A band-pass filter allows the passage of components in a specific frequency band, called the passband, but blocks components with frequencies above or below this band [35]. In audio processing applications, it can be used to emphasize or process sounds in certain frequency ranges.

Pitch shift is a sound recording and data augmentation technique in which the original pitch of a sound is raised or lowered [36]. In summary, pitch shift (low) lowers the pitch of the sound, while pitch shift (high) raises the pitch of the sound.

Time shift refers to moving a signal forward or backward on the time axis [36]. It is the process of slowing down or speeding up audio or video recordings. While audio or video recording playback speed is slowed down with time shift (slow), playback speed is also increased with time shift (fast).

Add noise is adding noise to the audio data. The purpose of doing this is to help the model learn these conditions in order to achieve successful results at different noise levels [37]. Additionally, this method is also used as a data augmentation process. Add noise (normal) is used to add normally distributed noise, and add noise (uniform) is used to add uniformly distributed noise.

3.3 Feature extraction

Feature extraction methods are used to convert audio signals into usable form. The goal is to keep the distinctiveness of the signal while reducing the input vector size [38]. For this step, there are many methods that can be used as feature extraction such as mfcc, mel-spectrogram, beats, rmse, centroid, bandwidth, cens, stft, cq, and contrast. In this study, many different techniques are used to extract features from speech signals, and the features of a total of 1980 songs belonging to 9 singers were extracted [39]. As a result of the tests, it was seen that the feature vector obtained by combining the features extracted with MFCC, mel-spectrogram, and octave-based spectral contrast contributed positively to the classification success.

MFCC is a feature extraction method used in singer recognition commonly. In MFCC, the pre-emphasis stage is used to increase the energy level of the high frequencies of the input signals. Thus, the data in these regions are better received, and these data are beneficial in the training phase. Windowing divides the input signal into discrete-time parts. This is accomplished using M millisecond windowsin and M millisecond offsets. DFT processing is applied to each signal segment. As a result of this process, the magnitude and phase representation of the signal are provided. This result contains information for each frequency range. However, humans are less sensitive to frequencies above 1 kHz. For this reason, the logarithmic Mel scale is applied to the outputs. Finally, the MFCC are obtained by performing the cosine transform to the Mel spectrum data [40, 41].. The Mel scale can be used to distinguish people from each other because the sounds that are equidistant from each other are close to each other in Mel scale values. Since spectral shape is a critical attribute for understanding the type of song, proposed model will generate the spectrogram showing the spectrographic representation of the signal [28].

The spectral contrast value represents the spectral peaks and minimums and the differences between them in the sub-bands. Sound harmonic features correspond to strong spectral peaks. Conversely, nonharmonic features and noises are concentrated in the minimum region. Therefore, the spectral contrast feature reflects the distribution of harmonic and nonharmonic components in the spectrum.

In order to obtain the spectrum from digital samples, firstly, FFT process is applied. Octave-scale filters are used to divide the frequency domain into sub-bands. The strength and differences of the spectral peaks and subregion are predicted for all sub-band. It is mapped to orthogonal space with the Karhunen-Loeve transform, and the dependency between different dimensions is removed, and after, the raw spectral contrast feature is

converted to log space. Thus, the spectral contrast value for each frame is estimated [42].

3.4 Data normalization and standardization

After the feature vector is obtained, it is necessary to separate the feature vector for use in the training and testing stages. The training and test rate are set at 90% and 10%. After that, the decimal values of the data were truncated and standardized to increase the productivity. Then, the data is normalized, and the redundancy is removed. Thus, it is possible to reveal more logical and relational data that serve to solve the problem. After normalization, data cleaning is done and NaN, and Inf value is corrected. Thus, suitable data were obtained for making analysis and prediction.

Since the range values of the MFCC coefficients vary greatly, their effects within the learning structure may not be homogeneous. In this case, the gradient descent can converge to zero very quickly. For this reason, MFCC values should be normalized before being used in model training [43]. The MFCC coefficients cannot be normalized separately for each 30 or 60 ms sound frame. This process is performed at once using the mean value and standard deviation with calculating over coefficients. The resulting mean value is subtracted from each MFCC value and divided by the standard deviation value.

3.5 Dimension reduction (PCA)

PCA is a technique that determines the direction of the greatest variance in the data and rotates the axes in this direction, thus finding the weighted linear combinations of the original variables [44]. This way, each component will not be related to the other. The first principal component is the component that takes into account the maximum variation in the original data. The second component represents the maximum change that the first component does not take into account. Similarly, the third component shows the maximum variation that the first two components do not take into account. In this way, all PCA components become a vector that is more easily classified and has border regions between classes, since they create a lower-dimensional vector instead of attributes on the main data. PCA is best applied to feature vectors to find a set of uncorrelated PCA components that can improve accuracy and increase feature size. The number of resulting PCA components is less than the number of original features.

3.6 Classification

In order to determine the model with high classification performance, which will be trained first and then used in the testing phase, tests were carried out with 11 different algorithms, and the results were compared. It was aimed to find the algorithm with the highest accuracy. In order to find the best model during the classification phase, hyper-parameter tuning was performed, and the parameters providing the highest accuracy were used. The resulting model was tested for the multi-classification problem, and its results were evaluated against different metrics.

In this study, extensive tests were performed with 11 different machine learning techniques for classification, and comparisons were made. Logistic regression, random forest, decision tree, gradient boosting, Extra Tree, Ada-Boosting, Gaussian Naive Bayes, KNN, linear discriminant analysis, SVM, and XGBoost algorithms were used as classifiers. For each algorithm, the hyperparameter values were determined to be the best with the GridSearchCV and RandomizedSearchCV algorithms. Thus, it was created to reveal the most successful model with the resulting concatenated feature vector.

4 Experimental design, analysis, and results

In this paper, it is aimed to find the model with the most successful classification performance. Experiments were carried out to predict the singers successfully. The steps taken in this regard are given in this section.

4.1 Datasets preparation

Creating a dataset is a difficult and time-consuming process. In recent years, a large number of dataset in many different languages have been created to automatically identify singers from a song. However, publishing these datasets are not very easy due to copyrights, so the number of quality databases still has not reached the desired level. Such situations compelled researchers to create their own databases to use in their studies. As a result of the researches, it has been seen that there is no dataset of Turkish song to be used in the task of identifying the singer. For this reason, a New Turkish Music Dataset has been created both to contribute to the literature and to use it in this study.

4.1.1 New Turkish Music Dataset

In the New Turkish Music Dataset, chosen 30-s parts of the songs represent the most important part of the music, especially where the voice of the singer and the music are intertwined and are difficult to distinguish from each other. Thus, it is desired that the model to be created is reliable in the face of the difficulties of the singer identification task. In this dataset, there are a total of 180 songs belonging to 9 popular singers, 20 of each singer of different genders, and the gender balance of the singers is distributed as 56% female and 44% male. After the songs were converted from stereo-channel format to mono-channel format using Audacity [45] program, they were divided into 30-s parts as shown in Fig. 2. Then, all songs

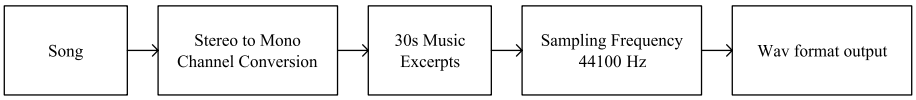

Fig. 2 Pre-processing schema of proposed model

were converted to wav format at a sampling frequency of 44,100, standardized, and recorded.

4.1.2 Artist20 dataset

Artist20 is one of the most commonly used datasets in the literature on singer identification. It was created by Labrosa [23]. It consists of 1413 songs covering different musical genres from 6 albums of each of 20 singers. Since the songs are very long and this may have a negative impact on the identification of the singers, the data was increased by regularly dividing each of the 1413 songs into 10-s segments, and the number of data was increased to 23,885. Also, these songs were converted into wav format at a sampling frequency of 44,100 and standardized. In this dataset, training and testing were set to 90% and 10%, respectively. In order to determine the model with high classification performance, tests were carried out with the same algorithms used in the previous dataset, and the best result was obtained with KNN. Hyper-parameter tuning was settled to find the best model, and the parameters providing the highest performance were used. As a parameter to KNN, leaf size = 20, distance metric = "minkowski," and number of neighbors = 1 were selected.

4.2 Experimental results

In this study, it is aimed to find the singer to which song belongs with the model developed using a dataset consisting of a total of 180 songs taken from 9 different Turkish singers. The suggested model was explained in Section 3.1 in detail. However, a series of experiments were carried out in order to decide on the methods and classifiers to be used in the proposed model. In this section, the intermediate result values obtained during the emergence of the best model are shown sequentially.

According to Table 1, the classification results were at the highest level of 50%. In this situation, it has been understood that the used of a single feature extraction method is not sufficient for Turkish singer identification. For this reason, it is thought that training and testing with a vector containing more six features for each song, which will emerge as a result of combining more than one feature extraction method, can contribute positively to the results. The tests were repeated by making models in which the MFCC, spectral constrast, chroma-CENS and root-mean-square error, and algorithms, which showed the highest classification success, were used together. Accordingly, it was understood that the most successful results were obtained with MFCC + spectral contrast.

The test results obtained with 11 different classifiers are shown in detail in Table 2. Accuracy, precision, F-score, and recall were used as classification performance metrics. As a result of the tests performed, the highest classification rate achieved with Extra Tree with 89.4%. Precision, sensitivity, and F-score were 89.9%, 89.4%, and 89.5%, respectively. The results for this problem, which was trained with 198 songs for each class and tested with 22 songs, were quite successful. The confusion matrix for the most success model obtained was as shown in Fig. 3.

As seen in Figs. 3 and 4, three singers had varying degrees of recognition success, with the highest

Table 1 Best results for each feature extractor with selected classifiers

Feature extraction method	Classifier	Accuracy	Precision	Recall	F-score
MFCC	Logistic regression	0.500	0.429	0.500	0.424
Spectogram	Extra tree	0.333	0.304	0.333	0.290
Beats	KNN	0.310	0.330	0.310	0.305
Spectral contrast	Logistic regression	0.500	0.526	0.500	0.505
Constant-q	SVM	0.357	0.334	0.357	0.321
Chroma-CENS	Random forest	0.405	0.406	0.405	0.379
Short-time Fourier transform (STFT)	Logistic regression	0.333	0.405	0.333	0.321
Spectral centroid	Decision tree	0.262	0.255	0.262	0.234
Bandwith	SVM	0.262	0.239	0.262	0.241
Root-mean-square error	SVM	0.381	0.278	0.381	0.313

Table 2 MFCC + spectral contrast classification results for New Turkish Music Dataset

Classifier	Accuracy	Precision	Recall	F-score
Extra tree	**0.894**	**0.899**	**0.894**	**0.895**
Logistic regression	0.763	0.772	0.763	0.762
Decision tree	0.525	0.532	0.525	0.523
Gaussian NB	0.561	0.624	0.561	0.560
Linear discriminant analysis	0.727	0.737	0.727	0.728
AdaBoost	0.429	0.441	0.429	0.431
GradientBoost	0.798	0.804	0.798	0.798
KNN	0.813	0.830	0.813	0.814
XGBoost	0.833	0.835	0.833	0.831
Random forest	0.818	0.827	0.818	0.818
SVM	0.768	0.769	0.768	0.767

achieving a perfect score. This shows that 100% recognition success can be achieved for singers with the proposed model. The remaining six singers achieved 86% recognition success. It has been understood that higher values can be achieved in classification success if the necessary increase is made in training and test data.

Tests were also carried out for the Artist20 dataset, which is frequently used in the literature, on the model with the best results. In this way, it was aimed to compare the performance of the model with the studies in the literature. In this context, the results of 11 machine learning algorithms were obtained and evaluated.

As a result of the tests performed, the average classification rate achieved with KNN with 85.4%. Precision, sensitivity and F-score were 86.6%, 85.4%, and 85.4%, respectively. In addition, accuracy values of over 70% were achieved with Extra Tree and XGBoost, which are other classifiers. These classifiers also had high accuracy values in the other dataset. The results of KNN for this problem, which was trained with approximately 1094 songs for each class and tested with 100 songs, were quite successful. The confusion matrix for the most success model obtained was as shown in Fig. 5. As seen in Fig. 5, the recognition success rate was over 80% for all but three singers and over 85% for all but seven singers. The band with the highest accuracy value is Metallica (97%); unlike other singers, this band is a heavy metal band. The singer with the lowest accuracy value is garth (63%), an artist who operates in the country music genre. The other singers in the study generally produced works in the genre of rock music.

4.3 Discussion and comparison of similar works

There have been many studies from past to present regarding singer identification. These studies are not directly equivalent to each other but show experimental results in different languages with different datasets. Some examples on this subject are given in Table 3. While different algorithms such as MFCC, LPCC, PLP, RMS, and spectral contrast are uses for feature extraction, KNN, GMM, LSTM, SVM, MLP, CNN, and logistic

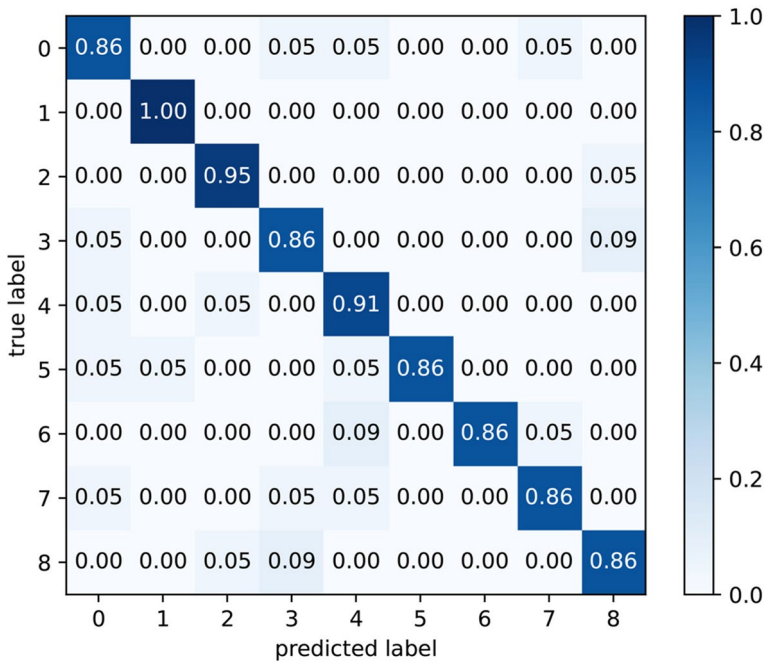

Fig. 3 Confusion matrix of proposed model with Extra Tree for New Turkish Music Dataset

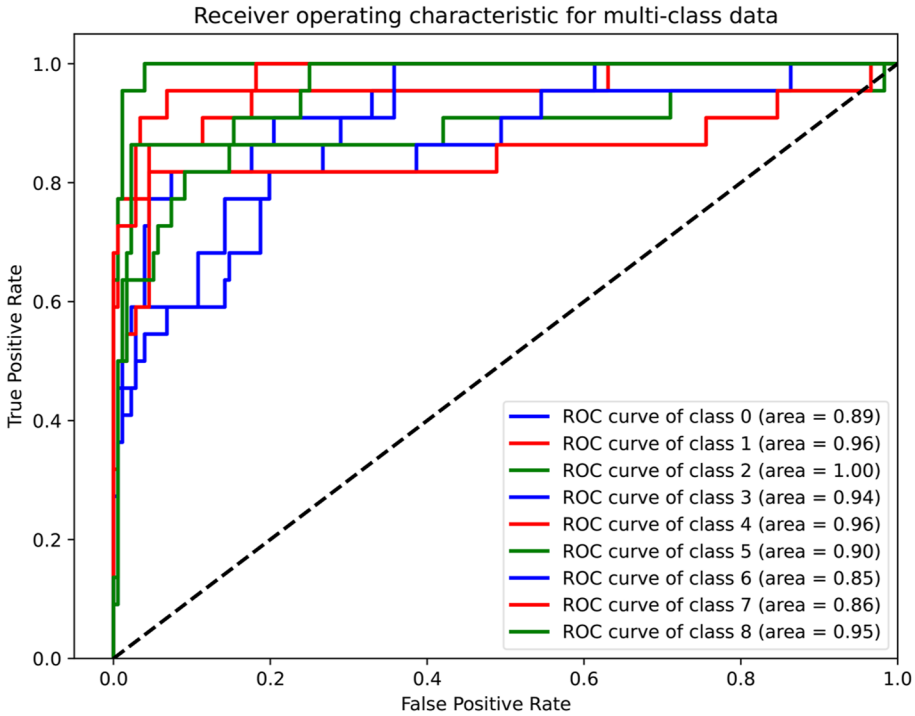

Fig. 4 ROC curve of proposed model for multi-class data for New Turkish Music Dataset

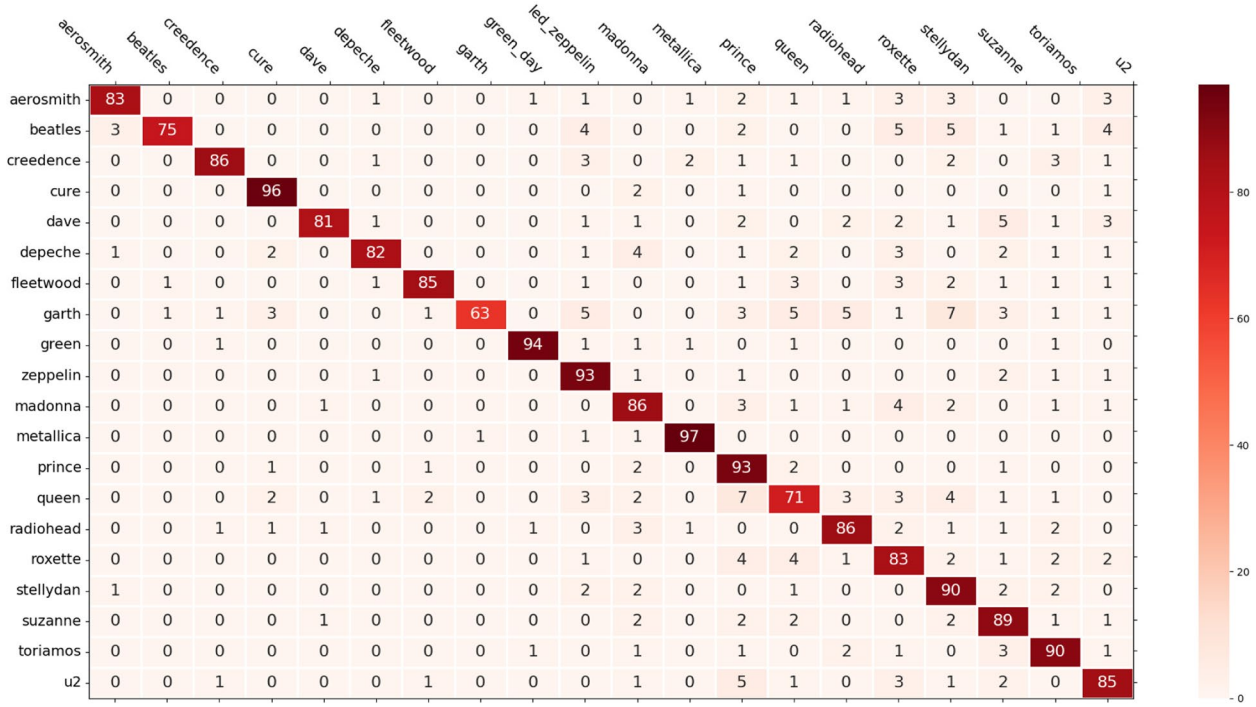

Fig. 5 Confusion matrix of KNN with mfcc + spectral contrast for Artist20 dataset

Table 3 MFCC + spectral contrast classification results

Article	Dataset	Features	Classifiers	Rate(%)
Liu and Huang (2002) [46]	Ten different singers, 30 different music for each singer (Chinese)	FMCV, PMCV	KNN	80.0
Tsai and Wang (2006) [47]	Twenty-three different singers, 10 different music for each singer	MFCC	GMM	87.8
Dharini and Revathy (2014) [26]	Ten different singers, 20 different soundtracks for each singer (Indian, Bengali)	PLP	K-means	55.56
Eghbal-Zadeh et al. (2015) [18]	Artist20 (20 singers, 1413 songs)	MFCCs	KNN	84.31
Xing (2017) [48]	Ten different singers, 10 different music for each singer	LPC	GMM	81.8
Shen et al. (2019) [1]	MIR-1K dataset	MFCCs	LSTM	88.4
Loni and Subbaraman (2019) [17]	Twenty-six different singers, 550 different songs (Indian)	Formants, vibrato, timbre, and harmonic spectral envelope	SVM	86
Murthy et al. (2021) [8]	Indian popular singers' database (IPSD), Artist20	MFCCs, LPCCs, SDCs, chroma, spectogram	YSA-RF-CNN	61.69–75.50
Noyum et al. (2021) [49]	Four different singers, 50 different songs for each singer	DWT	Linear SVM	83.96
Costa et al. (2017) [29]	Latin Music Database, ISMIR 2004, and African music collection dataset	Spectrogram, RLBP, rhythm patterns (RP), statistical spectrum descriptors (SSD), and rhythm histograms (RH)	CNN SVM	92
Li et al. (2021) [30]	Artist20, singer32 vs singer60	Spectrogram	CRNN	99.0–85.0
Nasrullah et al. (2019) [12]	Artist20	Spectrogram	CRNN	93.7 (F1)
Sharma et al. (2019) [31]	Artist20	MFCC	UBM T-matrix	89.97
Zhang et al. (2022) [32]	Artist20	Mel-spectrogram, articulation, rhythmic complexity, rhythmic stability, dissonance, tonal stability, modality, x-vector	CRNN	81 (F1)
Proposed model	Nine different singers, 20 different songs for each singer	MFCC, octave-based spectral contrast	Extra Tree	89.4
Proposed model	Artist20	MFCC, octave-based spectral contrast	KNN	85.4

regression are used as classifiers. As a result, recognition rates range from 55.67 to 92%. In the proposed model, the accuracy value of 89.4% was achieved with a New Turkish Music Dataset. The highest level of success was achieved for the Turkish language with the model in which the MFCC and octave-based spectral contrast features are used as a hybrid. On the other hand, when the same feature set was used with KNN on the artist20 dataset, a recognition success of 85.4% was achieved. It has been observed that even by just increasing the data for the artist20 dataset, better results are obtained than many studies in the literature. It was seen in the results obtained from both datasets that in addition to using different features and classifiers, various data augmentation methods also had a positive effect on recognition success.

Within the scope of the articles we examined, accuracy values between 80 and 99% were observed with machine learning algorithms in the studies using original datasets. In one of the studies [30], it was claimed that a very high classification performance of 99% was achieved. However, it was seen that not enough evidentiary arguments were presented. Dharini and Revathy [26] used the K-means clustering algorithm in their study and found a lower recognition rate. In other studies, modeling was done with classification algorithms. A nine-class structure is used in our model. In the publications examined, the number of classes varies between 4 and 26. The widespread use of acoustic attributes is also seen. In our model, acoustic features were extracted from the original dataset. A dataset consisting of the songs of Turkish singers, data attribute selection with PCA, and pre-processing process with normalization were included in the study. The use of the data augmentation method is another difference of the study. Data augmentation is a process used in machine learning and deep learning problems that increases the generalization ability of the model, reduces overfitting, increases the number of data, and provides better performance.

In addition, the results of the studies conducted on the Artist20 dataset, which is the frequently encountered dataset in the literature review, are also listed in the table. Since there is a dataset prepared for music genre classification

and our model has been tested with this dataset in order to compare the success of our own model with other inputs. In studies using this dataset, it has been observed that MFCC and spectrograms are more frequently preferred as attributes, and both machine learning and deep learning algorithms are used as classifiers. This dataset consists of 20 classes, and studies with accuracy values of 80% in machine learning algorithms and 90% in deep learning algorithms have been observed. In our study, after preprocessing, data augmentation, and attribute selection, high accuracy value was achieved with the machine learning algorithm.

As can be seen in the examples examined, higher accuracy values are observed in studies using deep learning algorithms. However, with the high accuracy value obtained in our model, which should be evaluated on the machine learning side, it shows that the model can generalize successfully. Of course, high accuracy value alone is not enough to interpret the model, but other metrics are already mentioned in the content of the article.

5 Conclusions

In this study, a model proposal is given that enables the identification of singers by using the acoustic features obtained from 30-s song recordings found in the new dataset. For this proposed, a database of 180 songs consisting of Turkish singers was created. By performing tests with the created dataset, the performance of the proposed hybrid feature vector in singer identification was measured. In our experiments, the effectiveness of mel-spectrogram and octave-based spectral contrast features in classifying singers has been demonstrated. The contribution of PCA and normalization processes to classification success is given comparatively. It was understood that the model showed stable results in the data consisting of the recordings of the singers from different albums and years. During this experiment, it was observed that many features obtained from the audio signal had a disruptive effect on singer identification. It has been seen that the logistic regression classifier provides the highest classification success with 89.4%. In addition, the artist20 dataset, which is frequently encountered in the literature, was used to compare the success of the model with other studies. It was observed that 85.4% success was achieved with the KNN classifier on the artist20 dataset, which was created using different data augmentation methods and tested with the proposed model. Therefore, it was seen that the model created was quite successful not only in recognizing Turkish singers but also in recognizing singers singing in different languages. In the future, studies will be conducted on the effectiveness of the proposed model for data consisting of songs in many different languages. In addition, the model proposed in this study will be expanded in order to detect both singers in songs performed by more than one singer.

Acknowledgements
Not applicable.

Authors' contributions
We provide methods for recognizing singers from Turkish songs.

Funding
Not applicable.

Availability of data and materials
The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 17 October 2023 Accepted: 20 February 2024
Published online: 26 February 2024

References

1. Z. Shen, B. Yong, G. Zhang, R. Zhou, Q. Zhou, A deep learning method for Chinese singer identification. Tsinghua Sci. Technol. **24**(4), 371–378 (2019)
2. S. Hizlisoy, Z. Tüfekci, Derin Öğrenme ile Türkce Müziklerden Müzik Türü Sınıflandırması. Avrupa Bilim ve Teknoloji Dergisi **24**, 176–183 (2020)
3. M.D. Ferreira, R.F. de Mello, Time complexity evaluation of cover song identification algorithms. Appl. Acoust. **175**, 107777 (2021)
4. Maddage, N. C., Xu, C., and Wang, Y. (2004, August), Singer identification based on vocal and instrumental models, In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 2, pp. 375-378). IEEE.
5. Shreevathsa, P.K., Harshith, M., Rao M, A., and Ashwini. (2020), Music instrument recognition using machine learning algorithms, 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM), 161-166.
6. S. Hizlisoy, S. Yildirim, Z. Tufekci, Music emotion recognition using convolutional long short term memory deep neural networks. Engineering Science and Technology, an International Journal **24**(3), 760–767 (2021)
7. S. Biswas, S.S. Solanki, Speaker recognition: an enhanced approach to identify singer voice using neural network. International Journal of Speech Technology **24**(1), 9–21 (2021)
8. Y.V. Murthy, T.K.R. Jeshventh, M. Zoeb, M. Saumyadip, G.K. Shashidhar, in *2018 eleventh international conference on contemporary computing (IC3)*. Singer identification from smaller snippets of audio clips using acoustic features and DNNs (IEEE, 2018), pp. 1–6
9. A. Wang, The Shazam music recognition service. Commun. ACM **49**(8), 44–48 (2006)
10. Y.S. Murthy, S.G. Koolagudi, T.J. Raja, Singer identification for Indian singers using convolutional neural networks. International Journal of Speech Technology, 1–16 (2021)
11. J. Haupt, Spotify (review). Notes. **69**, 132–138 (2012). https://doi.org/10.1353/not.2012.0115
12. Z. Nasrullah, Y. Zhao, in *2019 International Joint Conference on Neural Networks (IJCNN)*. Music artist classification with convolutional recurrent neural networks (IEEE, 2019), pp. 1–8
13. R.S. Arslan, N. Barışçı, Development of output correction methodology for long short term memory-based speech recognition. Sustainability **11**(15), 4250 (2019). https://doi.org/10.3390/su11154250
14. R.S. Arslan, N. Barışçı, N. Arıcı, S. Koçer, Detecting and correcting automatic speech recognition errors with a new model. Turk. J. Electr. Eng. Comput. Sci. **29**(5), 2298–2311 (2021). https://doi.org/10.3906/elk-2010-117

15. T. Ratanpara, N. Patel, Singer identification using perceptual features and cepstral coefficients of an audio signal from Indian video songs. Eurasip J. Audio Speech Music Process. **2015**(1), 1–12 (2015)
16. Chowdhury, A., Cozzo, A., and Ross, A. (2020), JukeBox: a multilingual singer recognition dataset, arXiv preprint arXiv:2008.03507.
17. D.Y. Loni, S. Subbaraman, Robust singer identification of Indian playback singers. Eurasip J. Audio Speech Music Process. **2019**(1), 1–14 (2019)
18. H. Eghbal-Zadeh, M. Schedl, G. Widmer, in *2015 23rd European Signal Processing Conference (EUSIPCO)*. Timbral modeling for music artist recognition using i-vectors (IEEE, 2015), pp. 1286–1290
19. S. Kooshan, H. Fard, R.M. Toroghi, in *2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA)*. Singer identification by vocal parts detection and singer classification using lstm neural networks (IEEE, 2019), pp. 246–250
20. S. Hizlisoy, Z. Tufekci, Farklı Kültürlere Ait Farklı Türdeki Müziklerden Duygu Tanıma. Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi **35**(3), 687–698 (2020)
21. Whitman, B., Flake, G., and Lawrence, S. (2001), Artist detection in music with minnowmatch, In Neural Networks for Signal Processing XI: Proceedings of the 2001 IEEE Signal Processing Society Workshop (IEEE Cat. No. 01TH8584) (pp. 559-568). IEEE.
22. Kim, Y. E., and Whitman, B. (2002), Singer identification in popular music recordings using voice coding features, In Proceedings of the 3rd international conference on music information retrieval (Vol. 13, p. 17).
23. D.P. Ellis, *Classifying music audio with timbral and chroma features* (2007)
24. H.A. Patil, P.G. Radadia, T.K. Basu, in *2012 international conference on Asian language processing*. Combining evidences from mel cepstral features and cepstral mean subtracted features for singer identification (IEEE, 2012), pp. 145–148
25. W.H. Tsai, H.C. Lee, Singer identification based on spoken data in voice characterization. IEEE Trans. Audio Speech Lang. Process. **20**(8), 2291–2300 (2012)
26. D. Dharini, A. Revathy, in *2014 International Conference on Communication and Signal Processing*. Singer identification using clustering algorithm (IEEE, 2014), pp. 1927–1931
27. S.H. Deshmukh, S.G. Bhirud, North Indian classical music's singer identification by timbre recognition using MIR toolbox. Int. J. Comput. Appl. **91**(4) (2014)
28. S. Masood, J.S. Nayal, R.K. Jain, in *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*. Singer identification in Indian Hindi songs using MFCC and spectral features (IEEE, 2016), pp. 1–5
29. Y.M.G. Costa, L.S. Oliveira, C.N. Silla Jr., An evaluation of convolutional neural networks for music classification using spectrograms. Appl. Soft Comput. **52**, 28–38 (2017)
30. Li, W., Zhang, X., Qian, J., Yu, Y., Sun, Y. (2021), Singer identification using deep timbre feature learning with KNN-Net, ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
31. B. Sharma, R.K. Das, H. Li, On the importance of audio-source separation for singer identification in polyphonic music. Proc. Interspeech **2019** (2019). https://doi.org/10.21437/Interspeech.2019-1925
32. X. Zhang, J. Wang, N. Cheng, J. Xiao, *Singer identification for metaverse with timbral and middle-level perceptual features, 2022 International Joint Conference on Neural Networks (IJCNN)* (2022). https://doi.org/10.1109/IJCNN55064.2022.9892657
33. Website: https://helpx.adobe.com/audition/using/inverting-reversing-silencing-audio.html
34. Çeven, S., and Bayır, R. (2020), Ortam Sesinden İnsan Sesinin Ayrıştırılması için Filtre Geliştirilmesi, Avrupa Bilim ve Teknoloji Dergisi, (Special Issue), 331-337.
35. L.J. Christiano, T.J. Fitzgerald, The band pass filter. Int. Econ. Rev. **44**(2), 435–465 (2003) http://www.jstor.org/stable/3663474
36. L. Ferreira-Paiva, E. Alfaro-Espinoza, V.M. Almeida, L.B. Felix, R.V. Neves, *A survey of data augmentation for audio classification, In XXIV Brazilian Congress of Automatics (CBA)* (2022)
37. Wei, S., Zou, S., and Liao, F. (2020), A comparison on data augmentation methods based on deep learning for audio classification. In Journal of physics: Conference series (Vol. 1453, No. 1, p. 012085). IOP Publishing.
38. R.S. Arslan, N. Barışçı, A detail survey of Turkish automatic speech recognition. Turk. J. Electr. Eng. Comput. Sci. **28**, 3253–3269 (2020)
39. M. Yadav, M.A. Alam, (2018), Speech recognition: a review. International Journal of Research in Electronics and Computer Engineering (IJRECE) **6**, 1–9 (2018)
40. S. Karpagavalli, E.A. Chandra, Review on automatic speech recognition architecture and approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition 2016 **9**(4), 393–404 (2016)
41. Arslan, R. S., and Barissi, N. (2018), The effect of different optimization techniques on end-to-end Turkish speech recognition systems that use connectionist temporal classification, 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). https://doi.org/10.1109/ismsit.2018.8567240
42. D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, L.-H. Cai, *Music type classification by spectral contrast feature* (Microsoft Research area, 2002)
43. E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, A. Othmani, MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomedical signal processig and control **71**(1-11), 2022 (2022)
44. Arslan, R. S. (2020), Development of output correction methodology for Turkish speech recognition and design of a recurrent neural network (Phd Thesis). Council of Higher Education Thesis Center. (626161).
45. Audacity(R) software is copyright (c) 1999-2021 Audacity Team. [Web site: http://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License.] The name Audacity(R) is a registered trademark."
46. Liu, C., and Huang, C. (2002), A singer identification technique for content-based classification of MP3 music objects. CIKM'02.
47. W. Tsai, H. Wang, Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. IEEE Trans. Audio Speech Lang. Process. **14**, 330–341 (2006)
48. L. Xing, *Singer Identification of Pop Music with Singing-voice Separation by RPCA* (2017). https://doi.org/10.15002/00021518
49. Noyum, V. D., Mofenjou, Y. P., Feudjio, C., Göktuğ, A., and Fokoue, E. (2021), Boosting the predictive accurary of singer identification using discrete wavelet transform for feature extraction, arXiv preprint arXiv:2102.00550.

Publisher's Note