Singer Identification Using Time-Series Auto-Regressive Moving Average

Ananya Bonjyotsna and Manabendra Bhuyan

Abstract Singer Identification (SID) is one of the major interests in the field of Music Information Retrieval (MIR). The researches in SID in the last decade have been primarily focused in improving the identification accuracy by using better features in addition to Mel-frequency Cepstral Coefficients (MFCC). This work primarily attempts to explore a time-domain feature from the model parameters of the time-series Auto-regressive-Moving Average (ARMA) model to be used as one of the features for SID. The ARMA features are also combined along with MFCC to compare the results and observe its performance in SID. The MFCC and ARMA features are trained and classified using the Gaussian Mixture Model (GMM). Most of the literature deals in the spectral domain for feature extraction. Therefore, this paper mainly seeks to find the scope of using time-domain model parameters as one of the features in decision-making problems in the field of MIR.

Keywords Singer identification · MFCC · ARMA · GMM · MIR

1 Introduction

Singer Identification (SID) is the process of retrieving the identification of the singer in a song through extraction of the most efficient and robust features from the singers voice and their processing. This interest in identifying the singer is motivated by the growing amount of music exchange using the internet and the need to categorize the songs based on the singer. One of the method of characterization of most audio systems, music servers and online music stores is by the name of the singers. For content based MIR, it is necessary to represent the singing voice by its characteristics. Generally the process of SID can be divided into three steps—(i) Locating vocal/non-vocal

A. Bonjyotsna (✉) · M. Bhuyan
Tezpur University, Tezpur, India
e-mail: ananyab@tezu.ernet.in

M. Bhuyan
e-mail: manab@tezu.ernet.in

G. R. Kadambi et al. (eds.), *Emerging Trends in Photonics, Signal Processing and Communication Engineering*, Lecture Notes in Electrical Engineering 649,
https://doi.org/10.1007/978-981-15-3477-5_27

segments in the song and extracting the vocal segments, (ii) Feature extraction of the vocal segments and (iii) Statistical Classification. If the data used are monophonic, the first step may not be taken into consideration. Some of the commonly used features in SID are Mel-frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPCs), Perceptual Linear Prediction Coefficients (PLPs) and the Harmonic Coefficients. Similarly, most common classifiers that are seen to be used in SID are Gaussian Mixture Models (GMM), Hidden Markov Model (HMM), Support Vector Machines (SVMs) and Multilayer Perceptrons (MLPs). Kim and Whitman [6] used traditional LPC features for speech coding to extract singing voice features. They used two different classifiers GMM and SVM to model their feature vectors using established pattern recognition algorithms. Zhang [10] computed the LPCC to extract the audio features of each audio frame which form the feature vectors of the audio frame and used a GMM classifier classify singers. Tsai and Wang [9] and Nwe and Li [7] have also employed MFCC and another feature called Octave Frequency Cepstral Coefficient (OFCC), respectively. Bartsch and Wakefield [1] proposed an algorithm for identifying the singing voice in a song with no accompaniment using the spectral envelope of the signal to build a composite transfer function as the feature and have used a standard quadratic classifier for classification. Cai et al. [2] proposed a new auditory feature called Gammatone Cepstral Coefficient combined with MFCC and LPCC to represent different singing voice features. Patil et al. [8] used Cepstral Mean Subtraction (CMS) based MFCC (CMSMFCC) feature vectors for SID and compared the result with MFCC using second-order polynomial classifier. Johnson et al. in [5] analysed the Long-term Average spectrum as an objective measurement for singing voice classification. Deshmukh et al. [3] worked only on timbre features in SID applied to North Indian classical singers. Devaney in [4] has also used similar features like pitch, timing, dynamics and timbre descriptors and implemented them on western classical trained singers for intra and inter-singer similarity.

The existing research works mentioned above have mostly considered the frequency-domain auditory features for characterizing the singing voice signal. And also it is seen that time-domain stochastic methods were only used in speech signals and not in singing voice yet. In this work, time-domain approach is directly applied to singing voice and explored to find a new feature of the singing voice signal. The model builds a transfer function model of the framed input signal and the coefficients of the polynomial transfer function are used to form the feature vector of the signal. This work finally compares the performance of SID using MFCC and ARMA model coefficients as the primary features.

2 Methodology

The SID technique is performed using the conventional system as in [9]. The singing voices of the singers are the signals that are fed to the system. There are two phases in the system, i.e. Training phase and Testing phase. In the training phase, the Feature Vectors (FV) formed from the feature extraction are fed to the training block to

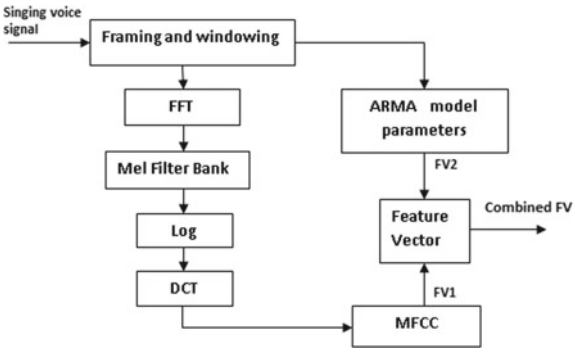


Fig. 1 Feature extraction for SID

build GMM model for each singer. In the testing phase, the FVs extracted from the unknown sample are then subjected to the classifier block to classify the given test signal to one of the trained singer models.

2.1 Feature Extraction

Prior to training the singer models, the singing voice waveforms are converted to MFCC frame by frame and ARMA model functions are computed frame by frame. The computation of ARMA model functions is described in Sect. 2.2. The block diagram for the feature extraction process is shown in Fig. 1. Let $FV1 = \{X_1, X_2, \ldots, X_L\}^T$ and $FV2 = \{Y_1, Y_2, \ldots, Y_L\}^T$ be the M-dimensional MFCCs and N-dimensional ARMA coefficients computed, respectively, for each signal where 'L' is the total number of frames. The combined FV is then given as $FV = \{FV1, FV2\}$. The experiment is conducted at first by taking MFCC features only, then the combined FV and lastly by taking ARMA features alone. And the results are compared thereafter.

2.2 Time-Series ARMA Modeling

Generally in ARMAX (Auto-regressive-Moving Average with eXogenous input), the present output has a relationship with the previous values of the inputs and outputs. Moving average denotes the noise model used in the system. The system also depends not only on the present inputs but also the previous inputs, i.e. the exogenous inputs. Considering a linear time-invariant system with an input signal u(n), output y(n) and disturbance e(n), first we define the univariate time-series 'y' into delay coordinates Y(n) = y(n−1), y(n−2), …, y(n−na); 'u' into U(n) = u(n−1), u(n−2), …, u(n−na)

and 'e' into E(n) = e(n−1), e(n−2), …, e(n−nc) which is given by a difference equation,

$$y(n) + a_1 y(n-1) + a_2 y(n-2) + \cdots + a_{na} y(n-n_a) = b_1 u(n-1) + \cdots + b_{nb} u(n-n_b) + e(n) + c_1 e(n-1) + \cdots + c_{nc} e(n-n_c) \quad (1)$$

Equation 1 can be written as

$$y(n) + \sum_{k=1}^{na} a_{na} y(n-k) = \sum_{k=1}^{nb} b_{nb} u(n-k) + e(n) + \sum_{k=1}^{nc} c_k e(n-k) \quad (2)$$

Expressing Eq. 2 in discrete domain, we get

$$y(n)\left[1 + \sum_{k=1}^{na} a_k q^{-k}\right] = u(n) \sum_{k=1}^{nb} b_k q^{-k} + e(n)\left[1 + \sum_{k=1}^{nc} c_k q^{-k}\right] \quad (3)$$

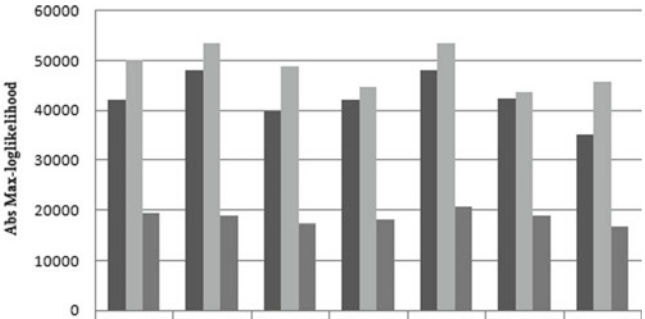
where 'n_a' is the order of the polynomial A(q), 'n_b' is the order of the polynomial B(q) and 'n_c' is the order of the polynomial C(q).

Now, the singing voice as a time-series model without any input to the system is expressed as

$$y(n)\left[1 + \sum_{k=1}^{n_a} a_k q^{-k}\right] = e(n)\left[1 + \sum_{k=1}^{n_c} c_k q^{-k}\right] \quad (4)$$

3 Experiments and Results

The SID accuracy is evaluated first by using MFCC alone, then by using both MFCC and ARMA features and lastly by using only ARMA features. The method is implemented in MATLAB platform. 10 monophonic songs of duration 30–50 s were considered for each of the 7 singers. The recording is done in the recording setup by using condenser microphone under a noise-proof environment. The sampling frequency of all the audio segments is 44.1 KHz of 16-bit mono tracks. A total of 35 songs with 5 songs from each singer were used for training and the rest 35 songs are used for testing. The extracted features from the audio segments are subjected to the training phase to build GMM model for the singers. 12 MFCC coefficients are considered for the experiment and the FV1 is a 12-dimensional matrix $L \times 12$ where *'L'* is the total number of frames computed from the segment. The frame size in samples is considered to be 1024 and the moving window size is taken to be half of the frame size, i.e 512 samples. In order to find the ARMA model parameters, the AIC values were calculated for different orders and the minimum AIC value was found for the order(4, 2). So the fourth-order transfer function was computed for each frame of



	Singer 1	Singer 2	Singer 3	Singer 4	Singer 5	Singer 6	Singer 7
Only MFCC	42065.8	48115.3	39915.0	42210.6	47986.6	42450.6	35287.2
MFCC+ARMA	50174.2	53634.8	48938.7	44861.1	53648.8	43699.0	45672.1
Only ARMA	19454.4	18824.7	17402.7	18127.1	20703.5	18813.5	16886.0

Fig. 2 Scores of absolute Max log-likelihood of the three cases

the audio segment. For each frame, the fourth-order polynomial transfer function is expressed as

$$G(q) = \frac{1 + c_1 q^{-1} + c_2 q^{-2}}{1 + a_1 q^{-1} + a_2 q^{-2} + a_3 q^{-3} + a_4 q^{-4}} \tag{5}$$

The coefficients of the polynomial transfer function a_1, a_2, a_3, a_4, c_1 and c_2 are used to form the feature vector. The 12 MFCC coefficients and 6 coefficients of the ARMA model are combined to form an 18-dimensional Feature Vector FV for modeling the GMM. For each class, 8 mixture Gaussian model is built. The feature vectors of the testing class is then matched to the already built GMM singer model.

Considering only MFCC and both MFCC and ARMA, the testing accuracy came out to be 100% for both the cases which means that all the audio segments were correctly identified to the respective singers. In case of only ARMA, FV2 is trained instead of FV. It has been observed that testing resulted is 94.28% accuracy, i.e. 33 audio segments out of 35 have been identified correctly. According to the results, the percentage accuracy of ARMA is less than MFCC. Now, in order to compare which method performs better, the maximum log-likelihood score is used as the metric. The higher the value of this metric, the closer the test signal is to the target singer. In Fig. 2, the absolute values are taken which means the smaller values denote the higher probabilities of matching. It could be seen that the combined Features showed increase in the absolute maximum log-likelihood score compared to only MFCC. This shows that adding the ARMA parameters as the feature decreases the performance of SID compared to only MFCC. However, another observation that is worth analysing is that the maximum log-likelihood scores in case of ARMA shows much smaller values than the other methods. It has been observed that in case of ARMA, the values are considerably less than the other two cases.

4 Conclusion

The objective of this work was to explore the possibility of using time-domain ARMA model parameters as a characteristic feature for classification of singers in the application of SID. The research works of the existing literature mentioned in Sect. 1 have been trying to improve the percentage accuracy of SID by adding new features in addition to MFCC. However, it is seen that no work has been done to inspect the applicability of time-domain features in SID. From this experiment, it is found that ARMA alone or when combined with MFCC could not perform better than MFCC in SID accuracy. But we cannot nullify the time-domain approach right away based on the accuracy results alone. The scores of the max log-likelihood of the ARMA case contradicts the accuracy performance. The reason for this should be because of considering the time-domain signal directly and applying to modeling without any transformation. Since we know that voice signals are quasi-stationary in nature, therefore such short-time modeling may further be investigated to improve the accuracy of SID. Therefore, the work in this paper could lead the research toward analysing the applicability of ARMA model coefficients as a feature in applications of SID or other identification applications of MIR.

References

1. Bartsch MA, Wakefield GH (2004) Singing voice identification using spectral envelope estimation. IEEE Trans Speech Audio Process 12(2):100–109. https://doi.org/10.1109/TSA.2003.822637
2. Cai W, Li Q, Guan X (2011) Automatic singer identification based on auditory features. In: 2011 seventh international conference on natural computation, vol 3, pp 1624–1628. https://doi.org/10.1109/ICNC.2011.6022500
3. Deshmukh S (2014) North indian classical musics singer identification by timbre recognition using mir toolbox. Int J Comput Appl 91(4):1–4
4. Devaney J (2016) Inter-versus intra-singer similarity and variation in vocal performances. Int J Comput Appl 45(3):252–264. https://doi.org/10.1080/09298215.2016.1205631
5. Johnson AM, Kempster GB (2011) Classification of the classical male singing voice using long-term average spectrum. J Voice 25(5):538–543. https://doi.org/10.1016/j.jvoice.2010.05.009, http://www.sciencedirect.com/science/article/pii/S089219971000130X
6. Kim YE, Whitman, B (2002) Singer identification in popular music recordings using voice coding features. In: Proceedings of the 3rd international conference on music information retrieval, pp 164–169
7. Nwe TL, Li H (2007) Exploring vibrato-motivated acoustic features for singer identification. IEEE Trans Audio Speech Lang Process 15(2):519–530. https://doi.org/10.1109/TASL.2006.876756
8. Patil HA, Radadia PG, Basu TK (2012) Combining evidences from Mel cepstral features and cepstral mean subtracted features for singer identification. In: 2012 international conference on Asian language processing, pp 145–148. https://doi.org/10.1109/IALP.2012.33

9. Tsai WH, Wang HM (2006) Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. IEEE Trans Audio Speech Lang Process 14(1):330–341. https://doi.org/10.1109/TSA.2005.854091
10. Zhang T (2003) Automatic singer identification. In: Proceedings of the 2003 international conference on multimedia and expo, ICME '03, vol 2, pp 33–36. IEEE Computer Society, Washington, DC. http://dl.acm.org/citation.cfm?id=1170745.1171614