Dastgàh Recognition in Iranian Music: Different Features and Optimized Parameters

Peyman Heydarian
Department of Computer Science
University of Waikato
Hamilton, New Zealand
pheydari@waikato.ac.nz

David Bainbridge
Department of Computer Science
University of Waikato
Hamilton, New Zealand
davidb@waikato.ac.nz

ABSTRACT

In this paper we report on the results of utilizing computational analysis to determine the dastgàh, the mode of music in the Iranian classical art music, using spectrogram and chroma features. We contrast the effectiveness of classifying music using the Manhattan distance and Gaussian Mixture Models (GMM). For our database of Iranian instrumental music played on a santur, using spectrogram and chroma features , we achieved accuracy rates of 90.11% and 80.2% when using Manhattan distance respectively. When using GMM with chroma, the accuracy rate was 89.0%. The effects of altering key parameters were also investigated, varying the amount of the training data and silence, as well as high frequency suppression on the results. The results from this phase of experimentation indicated that a 24 equal temperament was the best tone resolution. While experiments focused on dastgàh, with only minor adjustments the described techniques are applicable to traditional Persian, Kurdish, Turkish, Arabic and Greek music, and therefore suitable to use as a basis for a musicological tool that provides a broader form of cross-cultural audio search.

KEYWORDS

Dastgàh recognition, Persian mode, Maqàm, Chroma, DSP, Computational musicology

ACM Reference Format:
Peyman Heydarian and David Bainbridge. 2019. Dastgàh Recognition in Iranian Music: Different Features and Optimized Parameters. In *6th International Conference on Digital Libraries for Musicology (DLfM '19), November 9, 2019, The Hague, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3358664.3361873

1 INTRODUCTION

Although music search through audio content analysis has been explored extensively during the past two decades, the focus tends to be on mainstream Western genres, such as pop music. For other genres and musical styles, or when more specialist forms of study are required such as those undertaken by musicologists, there can be assumptions made in how the content analysis techniques have been devised that are not fully compatible with these other forms of music. This means that the techniques are not directly applicable. There is, therefore, value in turning attention to genres that sit outside of this mainstream, and to use the outcome of such investigations as a way of broadening the basis of melodic, modal, and rhythmic recognition analysis in tools that support musicological study.

In this paper we focus on techniques for detecting the mode in maqàmic musical traditions, and in particular the traditional Iranian dastgàh. The latter is the underlying modal system of Iranian classical music, and represents the scale and tonic, and to some extent the mood of a piece, with the dastgàh system itself closely related to the maqàm style in Turkish, Kurdish, and Arabic music.

With the emergence of websites, such as YouTube, as popular sources for all forms of music [3], knowledge of the mode of the music can potentially provide connections for users hailing from different countries who are otherwise separated by borders and language barriers, such as Greek and Iranian music listeners. However, this is currently difficult to achieve as people of one cultural group do not normally listen to the music of the other, and so such sites are lacking uniformly labeled user data to make such connections. Utilizing a computational approach to determine the mode gets past this impediment, and would pave the way for music search sites to enrich the features provided to users, driven by sound theoretical musical knowledge.

Such an enhanced system, could, for example, recommend Persian audio files in avàz-e esfehàn as well as Greek and Turkish pieces in maqàm nahàwand, a mode in Turkish/Greek music, with a similar scale. Relying on a manual approach, this could not have occurred unless a person who has the skill to recognise such modes, had manually labeled the files with the modal information in both language scripts. The algorithms presented in this paper aim to do this automatically and so provide a way to enhance how people can access music by introducing a new search criterion and, more generally, by increasing their knowledge about the music.

The general approach we take is to use the frequency spectrum and chromagrams computed from the audio as feature vectors to a classification algorithm that determines the mode of the music. A chromagram (also called chroma feature) is a simplified spectrum with logarithmically-spaced frequency domain components, similar to how a person perceives music and also the way the major musical scales are created. These features are then passed to a classifier which predicts the mode.

The structure of the paper is as follows. First we provide the background music theory that is pertinent to classifying the mode of dastgàh which focuses on intervals and tuning, as well mode,

DLfM '19, November 9, 2019, The Hague, Netherlands

ACM ISBN 978-1-4503-7239-8/19/11…$15.00
https://doi.org/10.1145/3358664.3361873

before going on to summarized related work. Next, in Section 3, we review the key techniques we build upon to develop the classifiers used in our experiments. We trialed two forms of classification scheme: the Manhattan distance, and Gaussian Mixture Models (GMM). In Sections 4 & 5 we respectively present details about the dataset used, which centres on a set of recordings played on a santur, and the results of our evaluation, before concluding with a summary of our findings.

2 BACKGROUND

2.1 The Persian intervals and modes

There are different views on Persian intervals [7, 20, 26]. Vaziri, for example, proposed a 24 equal temperament system (24-TET), by analogy with the Western equally tempered scale: Sori (𝄲) and Koron (𝄳) symbols show half-sharp and half-flat quartertones [26]. Farhat [7] suggests that in addition to the Western semitones, there are two intervals between a semitone and a whole tone, called the small and large neutral tones (n and N), and an interval greater than a whole tone, called the plus tone (p).

The implication of this from a signal processing standpoint is that, in addition to the Western intervals, there are flexible quartertones in Persian music that lie between two neighbouring semitones. Only a few of the quartertones are used in practice, with the Persian repertoire represented by 13 different notes: 7 diatonic notes, 3 semitones and 3 quartertones [20]: E F F𝄲 F# G G# A B𝄳 B C C𝄲 C# D. Persian music is based on a modal system, consisting of seven main modes and their five derivatives: Shur, Abu'Atà, Bayàt-é Tork, Afshàri, Dashti, Homàyun, Bayàt-é Esfehàn, Segàh, Chàhàrgàh, Màhur, RàstPanjgàh, and Navà. They are played in five different scales: Homàyun and Bayàt-é Esfehàn, Chàhàrgàh, Shur, Màhur, and Segàh. Figure 1 shows the tuning system for different modes. Moving accidentals are shown in parentheses. A mode implies the scale intervals and is to some extent an indication of the emotions of a piece. A person listening a piece of music recognises its mode by a combination of the following:

- Perceptually: based on the emotions of a piece
- Through melody/theme recognition: by matching a melody with known patterns
- By identifying the pitches: based on the intervals, frequency of occurrences and order of the notes

2.2 The santur

The database samples used in this paper are played on a Santur. The Santur is a trapezoidal string instrument, played by a pair of delicate hammer sticks, typically coated by a piece of cotton or leather (felt). The santur originated in Iran and is played in various countries including India, China, Thailand, Greece, Germany, where it is called by different names such as Yang-jin, Khim, Santouri, Hackbrett, Tsimble and the Hammered Dulcimer.

The santur found its way to Europe in the middle ages, where it has been known as the dulcimer in English literature since the fifteenth century. The instrument appears in The Oxford Companion to Music, with the first known reference being in 1660 AD [2]. In some ancient books its invention was associated with Farabi (870-950 AD). However, Masoudi includes the santur in a list of instruments of the Sassanid era (224-651 AD), indicating that it existed before Farabi [23]. Setayeshgar traces back the santur to Assyrian and Babylonian inscriptions (669 BC) [25].

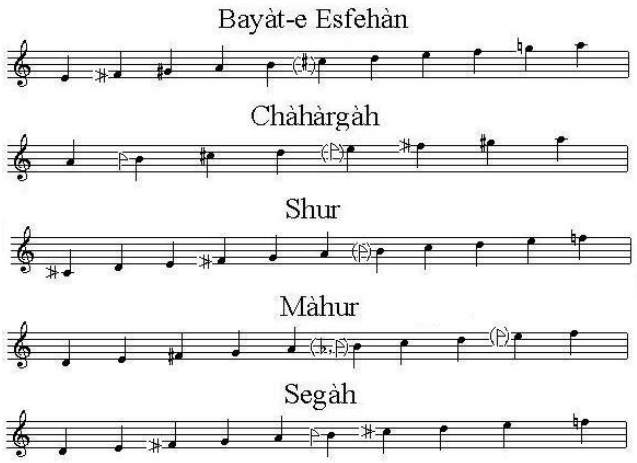

Figure 1: The tuning system

2.3 Related work

Gedik and Bozkurt [8] provided a method for makam (aka maqàm) recognition in Turkish music. They construct a high-dimensional pitch histogram of 53 commas per octave. Tonality is estimated first, by shifting and comparing the normalised pitch histograms with templates. The mode is then recognized by comparing the pitch histograms with the templates of each makam. Şentürk et al. presented a score informed tonic identification system for Turkish makam music [6] and Karakurt et al. developed MORTY: a toolbox for mode recognition and tonic identification which is developed in Python and is publicly accessible [21]. They consider an n-dimensional space, compared a single distribution for each mode in [8]. Chordia and Rae present a raag recognition system based on pitch class and pitch class dyad distributions [5]. And Gulati et al. [9] propose a melody representation for Indian raga recognition. Abdoli [1] presented a method for dastgàh recognition, based on fuzzy logic that takes into account the flexibility of Persian intervals and the dastgàh is recognised based on the similarity between these fuzzy sets and theoretical data. Heydarian et al. proposed a set of algorithms for pitch, and tonic and mode recognition in Persian music, assuming a 24-TET as the optimum tone resolution for Persian music [11–13, 15–20]. They used a constant Q transform for pitch tracking [4, 11, 19, 20], and the spectrum, the chroma and pitch histograms for the classification of the Persian audio musical files, based on their dastgàh /maqàm. This paper builds on this research.

3 MODE IDENTIFICATION

We now review the key techniques utilized in our work, and explain how they are combined to perform the task of mode identification. Frequency spectrum and chroma features constitute the input features we work with; the geometric classifier Manhattan distance, and the generative method Gaussian Mixture Model the two classifiers used to gauge the similarity between an unknown test samples

and the set of templates which were made during the training phase for each mode. A relative pitch basis is assumed, where transposed versions of a melody are considered the same. A 24-TET is found to be the optimum resolution and templates are constructed for each mode based on the chroma of the training samples. Spectrogram and spectral average can be calculated through the FFT of the signal. Minkowski's distance of order one, also called the Taxi Cab or Manhattan distance is used as the distance measure. Initially, a shift and multiplication process is performed, the shifted Harmonic Pitch Class Profile (HPCP) or THPCP (Transposed HPCP) is obtained, and the test samples are aligned with the training templates. Subsequently, the minimum distance between a test sample and the templates for each class determines the mode.

The Gaussian Mixture Model (GMM) is also used in this research as another classification scheme. A Mixture is a probabilistic model that describes a sub-population. The Gaussian Mixture Model has been used in classification tasks such as image recognition, text-independent speaker verification and MIR (Music Information Retrieval) [10, 22, 24]. In this work, the method is used for mode recognition, where the distribution of the chroma features is modeled by a Gaussian mixture density.

The mathematical formulas for making spectrogram and chroma, and the classifiers are provided in [17].

4 THE DATABASE AND FREQUENCY RESOLUTION

A database of 5706 seconds of music (91 pieces) in five Persian modes is used. It includes scale notes, opening sections, melodies and random sequences of notes, played on a santur by the first author. Santur notes are sustained and in some cases two or more notes are heard at the same time. The recordings include fast sequences of notes too. The samples are mono, 16 bit wave files at a sampling rate of Fs=44100 Hz. For processing, the signal is down sampled by a factor of 8 and the constant Q transform is calculated with minimum and maximum frequencies of 130 Hz and 1400 Hz, 72 bins and a window length of 8192. The 72 bins per octave provide enough resolution to distinguish between neighbouring quartertones regardless of the tuning. Having 72 frequency bins makes the smallest interval step of 0.00967. Thus the frequency resolution is 0.00967*130=1.257 Hz, and as the signal is down sampled to 5512, the required frame size, corresponding to this frequency resolution is N=5512 / 1.257 = 4385 samples. The resulting frequency resolution is 5512/8192= 0.67 Hz, which is more accurate than the required 2.56 Hz resolution (F3-Fs3). Every three bins are merged subsequently and a quartertone frequency resolution of 0.0293 times 130 Hz = 3.81 Hz is resulted. The chroma feature vectors of the database are made accessible publicly [14].

5 EXPERIMENTS AND RESULTS

The first five recordings in each dastgàh of the database are recordings of the scale notes, opening sections (daràmad), and simple metric and non-metric pieces, being played in a way to describe the five different scales of the dastgàhs (Figure 1). These form the training files used for the classifiers, with the remainder the test-set. In some cases, the training samples were used in the tests too, to compare the results with the cases where the training samples were not used. In the specific condition for the Manhattan distance classifier where a systematic way of generating the templates for the scales was used (i.e., artificially produced) to find out the effect of the amount of training data, these five recordings have not been used during the training phase, and so were added to the set of testing files. Using the Manhattan distance, the optimum recognition rates for spectral average and chroma average features achieved were 90.1% and 80.2% respectively (the training data was used in the tests and frames were non-overlapped). The estimation became 85.7%, using the chroma features and a machine learning method (GMM).

Looking at the confusion matrices, provided in [12], it was observed that the nature of the errors were different. Thus, these methods can be combined in order to improve the results. If the training templates are shifted 23 times, to reduce the effect of different tonalities between training and test templates, the recognition rate using chroma and Manhattan distance increases to 86.8%. This recognition rate was achieved using santur samples as the data-driven training templates.

5.1 Performance versus the amount of training data

Figure 2 compares performance versus the amount of training data, which is a smooth curve, obtained by averaging the following three cases using 0–151 s of files 1–6, 6–10, 8–15. Spectral average was used as the feature vector. The maximum recognition rates over the average curve of the three experiments are 89.4% without silence suppression and high-energy suppression and 83.9% with silence suppression and high-energy suppression. The average shows that over 65s of training data was needed to produce a recognition rate of around 83.2%. Thus performance is dependent on the duration of the training samples used and on whether silence suppression and high-energy suppression are implemented.

5.2 Comparison of the features

The highest recognition rate is achieved by using the spectrum, where the system depends on harmonic content, instrumentation, octave register, and tonality of the training samples, and the dimensionality of the feature space is high.

Chroma, a substantially smaller version of the spectrum inasmuch as frequency bins have merged the components, is less dependent on instrumentation than spectrum, and has a lower calculation cost, although the results are still affected by harmonic content, and are still dependent on tonality. With a tonic detection stage or a shift-and-compare process, the tonic can be shifted to the desired tonality.

Both features can either be used on a frame-by-frame basis or in average form. The summing-up process (averaging of all frames) has two desirable consequences: (1) reducing the effect of instrumentation and timbre in general; and (2) reducing the effects of noise as the fundamental frequencies and their harmonics are intensified. However, when averaged, temporal information, including the note sequence is lost. If a sample contains modulations to different modes or modal tonics, the points of modulation need to be found prior to averaging, where each segment is treated separately.

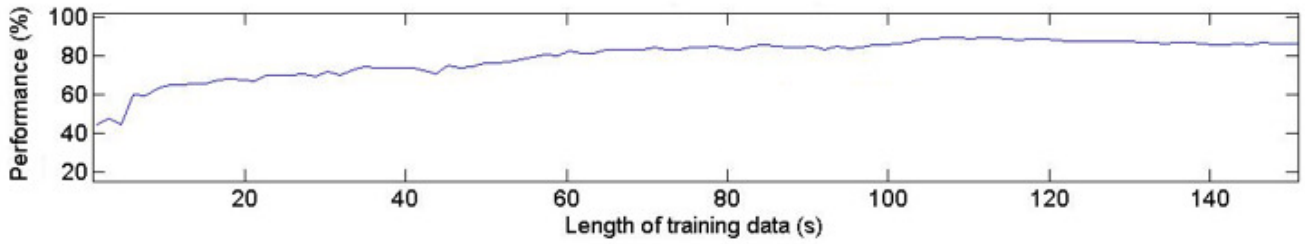

Figure 2: Performance versus amount of training data

5.3 Parameter optimisation

The parameters of the mode recognition algorithms developed in this research have been optimized, based on the training samples. The optimized parameters depend on aspects of the samples and the processes, including the onsets, silence and high-energy suppression, tone resolution, frequency range, and the amount of training data. The following parameter optimisations are made by straight comparison, with no shift and compare in templates.

Onset detection. Different parts of a signal carry different information that can affect the analysis results [12]. At an onset (when a note is struck), in addition to the fundamental frequency of the note and its partials, several other strings are also excited to vibrate, so several extra fundamental frequencies, their partials, and transients are briefly present in the spectrum of the signal. It was found that onset detection function affects the recognition rate. By starting the analysis frame an optimum distance from the onset, notes of short duration and rapid ornaments (e.g., consecutive quartertones and other ornaments involving intervals not usually allowed in a particular mode) can be excluded. Furthermore, some onsets can be skipped: for instance, if the analysis data is taken from the third onset instead of the start of the file, the recognition rate slightly increases, from 80.22% to 83.3% with no-overlap and slightly increases from 74.3% to 74.5% with 1/8-overlapped frames.

Silence and high-energy frame suppression. Silence suppression and high-energy frame suppression affect the recognition rate and change the nature of errors (silence suppression perhaps only marginally, as the santur has a relatively sustained sound). Our experiments show that silence suppression does not affect mode recognition results, whereas high-energy suppression does affect the results.

Frame size and the frequency range. In a previous research it was observed that the optimised frame size diminishes from Nf=131072 samples (with no suppression) to Nf=32768 when silence suppression and high-energy frame suppression are performed [12]. In both cases (without and with suppression), the maximum recognition rate was 86.36% when the training samples were not used in tests, or 90.11% if the training samples were used. We did the test including the training data to see the effect out of curiosity. As the mean of the training samples is used in training, the audio contents of the individual files are not reflected in the tests directly. Two alternative types of frames, Hamming and Hann, were compared in the pitch tracking task. Hamming worked better.

Frequency range is another parameter that affects the recognition rate. There are unwanted elements in very low and very high frequency contents of the dataset, and the signal could be high-pass filtered to amplify the effect of the low-amplitude, high-frequency components, and low-pass filtered to avoid the effect of the unnecessary components. Most of the notes were played in the middle register of the santur (277.2 Hz–698.5 Hz) and most of the energy as a whole is concentrated below 5 kHz. The effect of frequency range using spectral averages was examined: at a lower frequency bound of 458.9 Hz, a maximum performance of 89.4% is achieved, which is higher than was the case with the full frequency range. To reduce the amount of calculations and to include only the necessary harmonics, it would be advantageous to limit the frequency range to 458.9 Hz–6085 Hz.

Tone resolution. Different tone resolutions (12-TET, 24-TET, 48-TET, 53-TET) were compared. 24-TET produced the best results and is used thereafter. Sparse chroma average was made, where the least 5, 11, 22 and 22 components respectively were removed respectively. The recognition rates became 70.33%, 80.22%, 72.53% and 72.53% respectively.

The number of mixtures in GMM method. In the case of the Gaussian Mixture Models (GMM) method, further to the parameters considered above for chroma, the number of mixtures was varied between 1 and 10 to determine the optimum value. The highest estimation (85.7%) occurs when 1/8-overlapped frames with non-sparse chroma and 6 mixtures are used [12].

6 CONCLUSION

In this article we discussed the problem of mode identification in Persian music, and algorithms for automatic mode recognition have been presented. The frequency spectrum and chromagrams are used as the feature set, with Manhattan distance yielding recognition rates of 90.11%, 80.2% respectively for the dataset used in our experiments. The accuracy of GMM with chroma features was 89.0%. The effects of varying the different parameters such as the amount of training data, frame size, frequency and tone range were also investigated. With minor modifications, the approach is transferable to several other musical traditions in the Mediterranean and Near East region. For future work, tonic detection is recommended prior to scale recognition; the database can be extended to include vocal and multi-instrumental samples and musical samples of other traditions; finally, musical knowledge for example note successions and the intervals that may or may not occur in a mode performance are worthy of consideration.

REFERENCES

[1] Sajjad Abdoli. 2011. Iranian Traditional Music Dastgàh Classification. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*. 422–431.

[2] Denis Arnold (Ed.). 1984. *The New Oxford Companion to Music.* Oxford University Press, Oxford.

[3] Gianmario Borio. 2016. *Musical Listening in the Age of Technological Reproduction.* Routledge publishing.

[4] Judith C. Brown. 1991. Calculation of a Constant Q Spectral Transform. *Journal of the Acoustical Society of America* (1991).

[5] Parag Chordia and Alex Rae. 2007. Raag Recognition using Pitch-Class and Pitch-Class Dyad Distributions. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR).*

[6] Sertan Şentürk, Sankalp Gulati, and Xavier Serra. 2013. Score Informed Tonic Identification for Makam Music of Turkey. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR).*

[7] Hormoz Farhat. 1990. *The Dastgàh Concept in Persian Music.* Cambridge University Press.

[8] Ali Cenk Gedik and Baris Bozkurt. 2010. Pitch-frequency Histogram Based Music Information Retrieval for Turkish Music. *Signal Processing* 90, 4 (2010), 1049–1063.

[9] Sankalp Gulati, Joan Serrà, Kaustuv K Ganguli, Sertan Şentürk, and Xavier Serra. 2016. Time-delayed Melody Surfaces for Raga Recognition. In *Proceedings of the 17th International Society for Music (ISMIR).*

[10] Toni Heittola and Anssi Klapuri. 2002. Locating Segments with Drums in Music Signals. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR).*

[11] Peyman Heydarian. 2000. *Music Note Recognition for Santur.* Master's thesis. Tarbiat Modarres University.

[12] Peyman Heydarian. 2016. *Automatic Recognition of Persian Musical Modes in Audio Musical Signals.* Ph.D. Dissertation. London Metropolitan University.

[13] Peyman Heydarian. 2018. Tonic and Dastgàh Recognition in Iranian music. In *Proceedings of the 3rd Digital Music Research Network Workshop (DMRN).*

[14] Peyman Heydarian. 2019. The Santur Chroma Features. Retrieved August 31, 2019 from http://www.thesantur.com

[15] Peyman Heydarian and Lewis Jones. 2008. Measurement and Calculation of the Parameters of Santur. In *Proceedings of the Annual Conference of the Canadian Acoustical Association (CAA).*

[16] Peyman Heydarian and Lewis Jones. 2014. Tonic and Scale Recognition in Persian Audio Musical Signals. In *Proceedings of the 12th IEEE International Conference on Signal Processing (ICSP2014).*

[17] Peyman Heydarian, Lewis Jones, and Allan Seago. 2007. The Analysis and Determination of the Tuning System in Audio Musical Signals. In *Audio Engineering Society Convention 123.*

[18] Peyman Heydarian, Lewis Jones, and Allan Seago. 2012. Automatic Mode Estimation of Persian Musical Signals. In *Audio Engineering Society Convention 133.*

[19] Peyman Heydarian, Ehsanollah Kabir, and Mojtaba Lotfizad. 2001. Music Note Recognition for Santur. In *Proceedings of the 7th Annual Conference of the Computer Society of Iran.*

[20] Peyman Heydarian and Joshua D. Reiss. 2005. The Persian Music and the Santur Instrument. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR).*

[21] Altuğ Karakurt, Sertan Şentürk, and Xavier Serra. 2016. MORTY: A Toolbox for Mode Recognition and Tonic Identification. In *Proceedings of the 3rd International Digital Libraries for Musicology workshop (DLfM).*

[22] Matija Marolt. 2004. Gaussian Mixture Models for Extraction of Melodic Lines from Audio Recordings. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR).*

[23] A. Masudi. 1989. *The Meadows of Gold: The Abbasids.* Keegan Paul International.

[24] Douglas A. Reynolds. 1995. Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communication* 17 (1995), 91–108.

[25] Mehdi Setayeshgar. 1986. *Lexicon of Iranian music.* Ettela'at.

[26] Ali Naqi Vaziri. 1913. *Dastur-e Tàr.*