AUDIO TONALITY MODE CLASSIFICATION WITHOUT TONIC ANNOTATIONS

Zhiyao Duan[1,2*], *Lie Lu*[1], *and Changshui Zhang*[2]

[1]Microsoft Research Asia (MSRA), Sigma Center, Haidian District, Beijing 100080, China.
[2]State Key Laboratory on Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Automation, Tsinghua University, Beijing 100084, China.
duanzhiyao00@mails.tsinghua.edu.cn, llu@microsoft.com, zcs@mail.tsinghua.edu.cn

ABSTRACT

Traditional tonality mode (major or minor) classification or audio key finding algorithms often rely on tonic annotations (key names) of the training songs. However, unlike classical music whose keys are usually explicitly labeled in their titles, the keys of numerous popular music are hard to obtain. In contrast, it is much easier to only label the mode for each song. With only modes labeled, traditional approaches to key or mode classification cannot be directly applied, due to the lack of the reference point to transpose and align the chroma features with different keys. In this paper, we present an alignment approach to transpose chroma features within each mode to a reference (but unknown) tonic. Then several methods, including Single Profile Correlation, Multiple Profile Correlation and Support Vector Machine, are exploited to address mode learning and classification. Experimental results show the feasibility of the proposed approach.

Index Terms— Tonality classification, Audio key finding, Music information retrieval.

1. INTRODUCTION

Tonality or key is one of the most important aspects of music. It is highly related to chord, melody and emotion. Audio key finding is to estimate the mode (major or minor) and/or the tonic (the key name) of a piece of music from the audio input. It is an active topic in the music information retrieval area, since it is potentially helpful for many other topics such as chord recognition, music transcription, and recommendation.

A number of audio key finding algorithms have been proposed in the literature. Most of them employed a *profile correlation* method. These methods firstly extract the *chroma* feature [1] to represent a song, and then correlate it with existing chroma profiles corresponding to 24 major and minor keys. The key of the profile with the largest correlation value is assigned to the song. The chroma profiles can either be adapted from previously established ones [2, 3], such as Krumhansl's [4] and Temperley's [5] tone profiles proposed for MIDI key finding, or calculated using a data-driven method from the chroma features of a training data set [6, 7]. For instance, Gómex [2] and Peeters [3] generated their profiles by emphasizing the tonic, dominant and subdominant notes in Temperley's profiles, and Gómex [2] also deemphasized the harmonics of each note. As an alternative way, van de Par et al. [6] calculated the chroma profiles for major and minor by averaging the chroma vectors in a training data set after transposing them to C-major or c-minor according to their key labels. İzmirli [7] learned the chroma templates from monophonic sounds by averaging their chroma vectors which were weighted by adapted Temperley's profiles.

Other techniques were also employed in some work. For instance, Chuan and Chew [8] adapted the Spiral Array Center of Effect Generator (CEG) algorithm in symbolic key finding to polyphonic audio. İzmirli [7] used Principle Components Analysis (PCA) to reduce the dimensionality of the chroma features. Hidden Markov Models (HMM) were employed in [9, 10] to estimate key modulations.

While satisfying results achieved, a common requirement of the above approaches is that the training songs (if applicable) are labeled with specific key names, say C-major, a-minor, etc., since these labels are needed to either build key-dependent models, or in other cases, transpose chroma features to C-major or c-minor to build mode-dependent but key-independent models. This requirement can be easily satisfied in classical music pieces, since many of them are labeled with key names in their titles explicitly. However, for numerous popular music, key annotations is hard to obtain since it requires much expert knowledge and immense labor. In contrast, the mode (major or minor) of each song is relatively easy to label. In addition, in many applications, people care much more about modes than tonics in popular music. For example, in music mood detection, music recommendation, and music playlist generation, the mode label is sufficient to represent tonality information. Therefore, in this paper, we will address the problem of tonality mode classification when the data set is labeled with modes only.

*This wok was performed when the first author was a visiting student in Microsoft Research Asia.

Lacking tonic annotations, previous data-driven methods [6, 7] cannot be directly used, since the chroma vectors cannot be straightforwardly transposed to C-major or c-minor to build mode-dependent (but key-independent) models. In this paper, an alignment scheme is first introduced to transpose the chroma features within each mode to some reference (but unknown) tonic to build mode models. Then three approaches, including Single Profile Correlation (SPC), Multiple Profile Correlation (MPC) and Support Vector Machine (SVM), are exploited to build the mode models and classify the songs. The flow-chart of the approach is shown in Fig. 1.

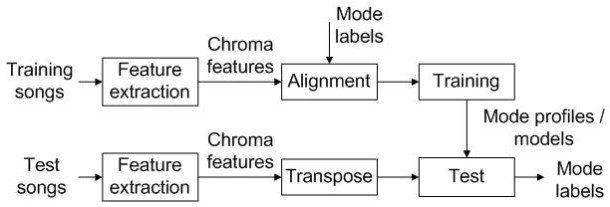


Fig. 1. Flow chart of the proposed approach.

2. FEATURE EXTRACTION AND ALIGNMENT

The chroma feature [1], which is commonly used for chord recognition and audio key finding, is employed in this paper. Due to the lack of tonic annotations, an alignment scheme is proposed here to transpose the chroma vectors within the same mode (but may be from different keys) to the same reference tonic, before building models of tonality modes.

2.1. Chroma feature extraction

Chroma feature can be either calculated by Constant Q Transform (CQT) [11], or by mapping the Fast Fourier Transform (FFT) spectra to the semitone scale. The former method is adopted in our approach. Each song is first divided into excerpts with equal lengths (15s, 30s or the whole song). In each excerpt, a 48-bins CQT in the frequency range from 130Hz (C3) to 1975Hz (B6) is calculated in each audio frame. The frame length is set to 130ms, which is decided by the finest frequency resolution of CQT: $130\text{Hz} \times (2^{\frac{\log_2(1975/130)}{48}} - 1) = 7.6\text{Hz}$; the frame hop is set to 10ms. Then for each excerpt, the average of the CQT vectors is calculated, and a 12-d chroma vector is calculated from the average CQT vector by summing the bins that have the same pitch class. Finally, the chroma vector is normalized so that its elements sum to 1.

2.2. Alignment

Given N chroma vectors $\mathbf{c_1}, \mathbf{c_2}, \cdots, \mathbf{c_N}$ within the same mode, we need to transpose them to the same tonic to get aligned vectors $\hat{\mathbf{c}}_\mathbf{1}, \hat{\mathbf{c}}_\mathbf{2}, \cdots, \hat{\mathbf{c}}_\mathbf{N}$. In general, two chroma vectors in the same mode correlate most if and only if they share the same tonic. Therefore, to align a set of chroma vectors, we select to maximize the overall correlation among them:

$$\{\hat{\mathbf{c}}_\mathbf{1}, \hat{\mathbf{c}}_\mathbf{2}, \cdots, \hat{\mathbf{c}}_\mathbf{N}\} = \arg \max_{\{\mathbf{c_i^{j_1}}, \cdots, \mathbf{c_i^{j_N}}\}} \sum_{i=1}^{N} \frac{\langle \mathbf{c_i^{j_i}}, \mathbf{q} \rangle}{\|\mathbf{c_i^{j_i}}\| \cdot \|\mathbf{q}\|} \quad (1)$$

$$\mathbf{c_i^j} = \text{Transpose}(\mathbf{c_i}, j) \quad (2)$$

$$\mathbf{q} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{c}}_\mathbf{i} \quad (3)$$

where $\langle \cdot, \cdot \rangle$ represents the inner product; $\| \cdot \|$ represents the norm; $\mathbf{c_i^j}$ is the transposed vector of $\mathbf{c_i}$, generated by circularly shifting the items j positions to the left; $\hat{\mathbf{c}}_\mathbf{i}$ is the ith aligned vector; and $\mathbf{q}$ is the average vector of them.

The optimal alignment above needs exhaustive search of all kinds of combinations of the vector shifts. Here we address this problem by an iterative method. The average vector is initialized with the first chroma vector in the data set, i.e. $\mathbf{q_0} = \mathbf{c_1}$. Then chroma vectors are aligned to it one by one, and the average is updated at the same time:

$$\hat{\mathbf{c}}_\mathbf{i+1} = \arg \max_{\mathbf{c_{i+1}^j}} \frac{\langle \mathbf{c_{i+1}^j}, \mathbf{q_i} \rangle}{\|\mathbf{c_{i+1}^j}\| \cdot \|\mathbf{q_i}\|} \quad (4)$$

$$\mathbf{q_{i+1}} = \frac{i \cdot \mathbf{q_i} + \hat{\mathbf{c}}_\mathbf{i+1}}{i+1} \quad (5)$$

After N times updates, $\mathbf{q_N}$ is used to initialize $\mathbf{q}$ again, and the above steps are performed once more to obtain the final aligned vectors. Although this is a "greedy" algorithm, it is found that the calculated average vector is stable when we randomly change the sequence of the training chroma vectors.

3. LEARNING AND CLASSIFICATION

For mode profiles/models learning and classification, the traditional Single Profile Correlation (SPC) method is employed. Besides, Multiple Profile Correlation (MPC) and Support Vector Machine (SVM) methods are also proposed.

3.1. Single Profile Correlation

Traditional profile correlation method [2, 3, 6, 7] can be employed for mode modeling and classification, where each mode is represented by *one* chroma profile. We call this method SPC in contrast to MPC proposed in Section 3.2. In SPC, the chroma profile is calculated by averaging the transposed training chroma vectors, where the transposition is performed by Eq. (1), since the key labels are unknown.

The obtained chroma profile is a 12-d vector, representing the distribution of each pitch class. İzmirli [7] argued that diatonic notes are most important to represent a key or a mode, and work better for key finding. Therefore, we also generate

the 7-d profile, where each element corresponds to the diatonic note of the 12-d chroma profile.

When classifying the chroma vector $\mathbf{c}$ of a test excerpt, it is circularly shifted 12 times to generate 12 vectors $\mathbf{c^1}, \mathbf{c^2}, \cdots, \mathbf{c^{12}}$ with different keys. These shifted vectors are correlated with a mode profile $\mathbf{p}$ to get correlation scores, from which the highest is taken as the confidence score indicating that $\mathbf{c}$ is in this mode (see Eq. (6) and (7)). The mode having the highest confidence score is assigned to the excerpt.

$$\text{Score}(\mathbf{c}, Mode) = \max_i \text{Score}(\mathbf{c^i}, Mode) \tag{6}$$

$$\text{Score}(\mathbf{c^i}, Mode) = \frac{\langle \mathbf{c^i}, \mathbf{p} \rangle}{\|\mathbf{c^i}\| \cdot \|\mathbf{p}\|} \tag{7}$$

Finally, given the mode labels of all its excerpts, the mode of a song is decided by majority voting.

3.2. Multiple Profile Correlation

Compared with single chroma profile, using multiple profiles may improve the representation of a mode, since the chroma vectors transposed from different keys have bigger variance than those of the same key. In our approach, K profiles are built using a K-kernel Gaussian Mixture Model. The centers and weights of the mixtures compose the profiles $\mathbf{p_1}, \cdots, \mathbf{p_K}$ and their weights $w_1, \cdots, w_K$. Correspondingly, the profiles with 7 diatonic elements are also generated.

To be consistent with SPC, the confidence score of a test chroma vector $\mathbf{c}$ being in a mode is also calculated based on Eq. (6), and the mode with the highest score is assigned to $\mathbf{c}$. Moreover, $\text{Score}(\mathbf{c^i}, Mode)$ can be defined in two methods: the maximum or the weighted summation of the correlations between $\mathbf{c^i}$ and the multiple profiles of the mode:

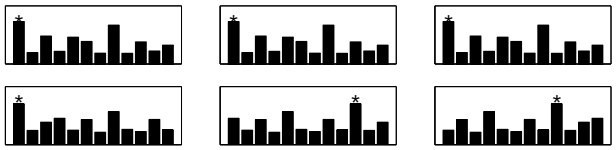
$$\text{Score}(\mathbf{c^i}, Mode) = \max_k \frac{\langle \mathbf{c^i}, \mathbf{p_k} \rangle}{\|\mathbf{c^i}\| \cdot \|\mathbf{p_k}\|} \tag{8}$$

$$\text{Score}(\mathbf{c^i}, Mode) = \sum_k \frac{w_k \cdot \langle \mathbf{c^i}, \mathbf{p_k} \rangle}{\|\mathbf{c^i}\| \cdot \|\mathbf{p_k}\|} \tag{9}$$

3.3. Support Vector Machine

SVM is successfully applied to many classification problems. We exploit it to mode classification in our scenario. A SVM with a radial basis function kernel is trained using the aligned training chroma vectors. The same as the correlation method, the chroma vector of a test excerpt is transposed 12 times and each one is classified. The mode label of the one with the highest classification confidence is assigned to the excerpt.

Here, one important issue in employing SVM is: Although the training chroma vectors are aligned within each mode (as in Section 2.2), the alignment (or arrangement) between the vectors in the two modes need to be considered carefully to improve the classification accuracy. One possibility is to make the major profile (the average of major chroma vectors) and the minor profile have the same tonic (see Fig. 2(a)), supposing the pitch class with the maximum value in the mode profile is the tonic of the mode. Another possibility is to align the tonic of minor to the 6th note (i.e. the 9th pitch class) of major, to make them "relative" (see Fig. 2(b)), such as for C-major and a-minor. However, it is found that both alignments cannot produce good results.



(a) Same tonic (b) Relative (c) Correlate least

Fig. 2. Three alignment methods between the major and minor profiles of the training set. The upper and lower panels of each sub figure is the major profile, and the lower is the minor profile. The asterisk indicates the tonic.

This can be explained as follows: For both the above-mentioned alignments, the distance between the major profile and the minor profile in the training set is small. That is to say, the training samples of major and minor mode are close to each other in the feature space. Therefore, it is hard for SVM to find a good classification surface between two modes. Moreover, it is noticed that the decisive chroma vector among the 12 transposed vectors of each test excerpt is the furthest one from the classification surface. This makes the distribution of the decisive test chroma vectors different from that of the training vectors, so that the classifier cannot generalize well to test excerpts.

To solve this problem, we use an alignment way that makes the profiles of major and minor correlate least or apart as far as possible, as in Fig. 2(c). From another point of view, this alignment together with the alignment in Section 2.2, can be seen analog to minimize the intra-class distance while maximize the inter-class distance. Experiments show that it works better than the above two.

4. EXPERIMENTS

The experimental materials were about 5,000 popular songs of various genres including soft rock, hard rock, electronica, folk, country, jazz, etc., with modes annotated. Songs having ambiguous modes or major-minor modulations were discarded. Finally 4,528 songs (2,786 major and 1,742 minor) composed our data set. 25% of them were randomly selected as the training set and the left as the test set.

We first evaluate the performances of profile correlation methods, comparing various profiles, including Krumhansl's profiles [4], learned single profile with / without feature alignment, and learned multiple profiles with the MPC approach. We also evaluate the performances when using different

Accuracy (%)	15s	30s	song
SPC (Krumhansl)	71.8 / 76.8	72.4 / 77.0	70.8 / 76.2
SPC (not aligned)	61.2 / 63.8	61.2 / 63.5	61.1 / 61.8
SPC (Aligned)	74.4 / 76.8	75.5 / 77.0	75.5 / **76.3**
MPC4 (Max)	76.0 / 76.8	75.8 / 76.8	73.7 / 76.0
MPC4 (Sum)	76.3 / 77.2	75.6 / 77.5	74.9 / 76.1
MPC8 (Max)	75.6 / **77.6**	75.3 / **77.4**	73.1 / 75.5
MPC8 (Sum)	76.6 / 76.4	76.2 / 77.3	75.4 / 75.0
MPC12 (Max)	76.2 / 77.1	76.0 / 77.3	71.6 / 75.1
MPC12 (Sum)	77.1 / 77.0	76.0 / 77.2	75.6 / 75.9

Table 1. Classification results using correlation methods with various profiles. In SPC, "not aligned" and "Aligned" stand for the profiles which are calculated from not-aligned and aligned chroma vectors, respectively. In MPC, different number of profiles (4, 8 and 12) for each mode are compared; and "Max" and "Sum" represent the correlation calculation with Eq. (8 and 9). Each setting is evaluated in correlation with the 12 elements profiles (the first number) and the 7 diatonic elements profiles (the second number).

lengthes of excerpts, and profiles having all 12 elements or only 7 diatonic elements.

Table 1 summarizes the comparison results. First, for SPC, by using the profiles calculated from aligned chroma vectors, the accuracy is improved 13% compared with those without alignment. It is also seen that the profiles with feature alignment work better than Krumhansl's profiles, which indicates the feasibility of the data-driven profile learning method. Second, with MPC approach, the accuracy can be further improved around 1-2%; and Eq. (9) works slightly better than Eq. (8). Third, the results with 15s- and 30s- excerpts are generally better than those taking the whole song as an excerpt. This may be because of modulations, which are common in some popular music. Fourth, correlation with the profiles having only 7 diatonic elements outperform with those having all 12 elements. This is in accordance with the results in [7] and indicates that the diatonic items contain the most useful information for mode detection and key finding.

We then evaluate the performance of the SVM-based approach, and compare different alignment schemes between major and minor classes. Table 2 presents the comparison results. As discussed in Section 3.3, arranging chroma features of major and minor with least-correlation (or maximum apart) criteria works best. Moreover, similar to the results obtained with the correlation approach, results with 15s- and 30s- excerpts are slightly better than the song-level results. The best result using SVM is up to 78.2%, which is better than that of the profile correlation method. This also shows the effectivity of discriminative methods for this task.

We also performed other experiments, such as using logarithmic scale chroma features, using 24-d chroma features, and imposing different weights on features (e.g. to emphasize the beginning and the end of each song), etc. However, the obtained results did not show considerable improvements.

Accuracy (%)	15s	30s	song
Not aligned	61.0	61.0	61.0
Aligned, same tonic	71.0	69.7	70.9
Aligned, relative	59.0	69.7	72.6
Aligned, correlate least	**78.2**	**77.1**	**76.3**

Table 2. Classification results using SVM with different alignment methods between major and minor on training chroma vectors.

5. CONCLUSIONS

In this paper, we propose an approach to address mode classification in the scenario that the data set is labeled with modes (major or minor) only, but without tonic. In this scenario, traditional approaches to modes modeling cannot be directly applied, due to the lack of reference points to align the chroma features from different keys. Correspondingly, this paper proposes an alignment approach to align chroma features within each mode to the same (but unknown) reference tonic. Then, SPC, MPC and SVM are exploited to learn mode models. While experimental results show the effectivity of the proposed approach, there is still room for future work. For example, we may pursue key-independent features and exploit temporal information to improve the mode model building.

6. REFERENCES

[1] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. International Computer Music Conference (ICMC)*, pp. 464-467, 1999.

[2] E. Gómez, "Tonal description of polyphonic audio for music content processing," *INFORMS Journal on Computing, Special Cluster on Computation in Music*, vol. 18, no. 3, 2006.

[3] G. Peeters, "Chroma-based estimation of musical key from audio-signal analysis," in *International Conference on Music Information Retrieval (ISMIR)*, 2006.

[4] C. Krumhansl, *Cognitive Foundations of Musical Pitch*, Oxford University Press, New York, 1990.

[5] D. Temperley, *The Cognition of Basic Musical Structures*, Cambridge, MA: MIT Press, 2001.

[6] S. van de Par, M. McKinney and A. Redert, "Musical key extraction from audio using profile training," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, 2006.

[7] Ö. İzmirli, "Audio key finding using low-dimensional spaces," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, 2006.

[8] C. Chuan and E. Chew, "Polyphonic audio key finding using the spiral array CEG algorithm," in *International Computer Music Conference (ICMC)*, 2005.

[9] K. Noland and M. Sandler, "Key estimation using a hidden markov model," in *International Conference on Music Information Retrieval (ISMIR)*, 2006.

[10] W. Chai and B. Vercoe, "Detection of key change in classical piano music," in *Proc. 6th International Conference on Music Information Retrieval (ISMIR)*, London, 2005.

[11] J. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425-434, 1991.