Features for comparing tune similarity of songs across different languages

Naveen Kumar, Andreas Tsiartas, Shrikanth Narayanan

Signal Analysis and Interpretation Lab
Department of Electrical Engineering, University of Southern California
Los Angeles, USA
komathnk@usc.edu, tsiartas@usc.edu, shri@sipi.usc.edu

***Abstract*—Finding tunes that are similar across languages and cultures offers new ways to study global musical influences and similarities. From a signal processing point of view, we find that the availability of vocal music tracks provides us a means for computing tune similarity even in the presence of language differences. While the different acoustic characteristics of each language add to the inherent ambiguity in these kind of problems, the guarantee that a vocal track exists can be a boon in disguise. For this purpose we use the Multi Band Autocorrelation Peak (MBAP) features, extracted in multiple bands providing complementary information which helps to improve the accuracy. Results obtained on a classification task suggest that these features can outperform traditional features like Chroma which capture information from the entire spectrum. Alignment cost using the dynamic time warping algorithm was used a classification metric on a dataset of songs obtained from Youtube.**

I. Introduction

Digitization of music was one of the most important changes that happened to the music industry in the recent decades. With the advent of technology, music today has crossed all possible barriers, including language and culture. As a result of the rapid growth of Internet, we now have worldwide access to a wide variety of music. With such a huge amount of music at our disposal, studies related to the analysis of music signals have gained interest in the community. Many of the problems related to music signals involve measuring music similarity, for instance, cover song identification, genre classification, tune similarity computation etc.

The basic notion of music similarity relies on finding patterns in a complex signal. Hence, similarity itself has been defined at different granularities for music signals. For example, one might be interested in a broader genre-level similarity [1], for use in recommendation systems. Alternatively, if the focus is more on applications such as song indexing, metrics based on individual song characteristics might be more helpful [2]. While the former problem requires a more subjective notion of music similarity, the latter defines music similarity in more concrete terms viz. similarity in tune. With such a variety of different approaches to specifying music similarity, it is not surprising that a plethora of features have been proposed for this purpose [3].

Some of these features try to make use of the structure in musical signals. The widely popular, chroma features [4], for example, try to specifically exploit the intrinsic structure involved in music composition. In addition other feature invariances like beat synchronization [5] have also been proposed to make this representation more robust to standard variabilities in songs. This treatment from first principles, however, works well only as long as a clean signal is available, where the tones are easy to detect. Hence, chroma features might tend to perform better on a *Music Information Retrieval* (MIR) task on a classical music corpus [6] as compared to a cover song identification on a more generic database comprising pop songs [7].

In cover song identification problems, the objective is to be able to recognize a song as an alternative or cover version of another previously recorded song. Depending on the artistic expressions of the cover artist, these versions may vary from the original version in tempo, key or arrangement [8]. Irrespective of this, the task is usually relatively easy for the human listener. However, the similarity in the signal is often not as easy to spot for a computer as it is for humans. Nevertheless, the notion of similarity in these problems is comparatively quite objective because of the existence of a "golden" or original version to which any of the cover versions can be compared.

In this work, we consider a sub-variant of the cover song identification problems viz. cover songs differing in the language of the vocals. Different acoustic-phonetic characteristics of each language add to the ambiguity in this case making the problem more challenging. On other hand, for a human listener, vocals in a song might provide additional cues for recognition due to the remarkable ability of the human auditory system in separating voice from background signals. As an application, the ability to detect cover songs can have interesting implications in cross-linguistic music indexing. In particular, this might be useful in searching for versions of a popular song in different languages. An interesting extension to this could be for plagiarism detection in songs with applications to digital media rights management.

Previously proposed features for this task attempt to extract melody from the vocals of a song [9]. This is supported

MMSP'12, September 17-19, 2012, Banff, Canada.
???-?-????-????-?/10/$??.??

by works in psychoacoustics [10] which suggest that certain idiosyncrasies of the singing voice might add to its saliency compared to other musical accompaniments. Thus, a vocal based melody estimate is robust, albeit constraining the choice of songs to only those with vocals. In this work, this is the domain of interest focused on cross-linguistic factors. The fact that these "melody" features indeed use information from the vocals is established by experiments performed on songs with the non-vocal regions removed. The authors in [9] propose to do this automatically by training models on labeled data.

In this work,, we propose features that estimate the fundamental frequency in multiple bands, using auto correlation. These multi band auto correlation peak (MBAP) features, contain a more complete description of the song using complementary information in different bands. Unlike features directly computed from the spectrum, MBAP feature estimates the the information in a way that gets increasingly sparse, in the higher bands. We compare the features using a dynamic time warping algorithm [11] which computes a cost for aligning a sequence of feature vectors with respect to another. A high alignment cost means that a significant warping was required to time-align the two sequences which will typically be the case for dissimilar signals. This allows us to use the alignment cost as a simple similarity metric. Results on a dataset of songs from Youtube suggest that these feature representations are promising.

II. Data Collection

Unlike cover songs in the same language for which the identity of the original song is typically well established, it is often not easy to obtain a solid ground truth for the reference song in the cross-lingual case. This makes it challenging to get a suitable labeled dataset to evaluate the performance of our classification experiment (Section V). To develop and test the ideas of tune similarity in this paper we focus on a subset of these songs which were either explicitly covered or dubbed in other languages. A song dubbed into another language shares the same tune as that of the original song. However, unlike ordinary cover songs, both versions are created by the same composer. This ensures that for each dubbed song in the database the identity of the reference song is known by definition. A dubbed song is usually sung by another artist in the new language, with slight modifications to the music. Dubbings are common for popular Indian songs (popular music, from films/musicals) when the composer wants to cater to a multilingual audience. We use parallel songs in the following Indian languages: Hindi, Tamil and Bengali.

In certain cases covers of popular songs are often as good as dubbings, because of the artist's attempt to maintain its original form. To create a more diverse database, we additionally selected such cover songs in other languages. About a third of our dataset comprises original or covered songs in Japanese, Russian, Korean, Mandarin, English and German. To obtain parallel versions of all these songs, we turn to clips posted on [1]Youtube; this adds complexity to the problem because of the differences in encoding, quality and source of each clip. This however makes the problem closer to a real world scenario where the songs will probably differ in a lot more than just vocals.

A total of 48 songs were collected in the above mentioned languages, of an average duration of 5 minutes (Table I). This gave us 24 pairs of files which we shall refer to as $\texttt{orig}_n$ for the original song and $\texttt{alt}_n$ for the alternative song in another language. Given an `orig` song, the task is to identify the corresponding `alt` song. For ease of representation these clips are so arranged such that $\texttt{alt}_m$ is the matching song corresponding to $\texttt{orig}_n$ for $n = m$. Hence, each song matching is a 24-way classification problem in our experiment.

TABLE I
DATASET SPLIT BETWEEN DIFFERENT LANGUAGES

Language	#
Hindi	17
Tamil	11
Bengali	6
Korean	6
Japanese	3
English	2
Others	3
Total	48

III. Baseline Features

Our baseline classification uses chroma features popularly used in music similarity literature [4][7]. We additionally compare against melody features proposed in [9] for a similar task of cross-lingual query by example. Although melody features compute information similar to chroma features, they are better suited to track the melody of the vocals, which adds to their robustness.

Unlike many generic audio features used in music similarity, chroma features were specifically designed to capture musical information. They intend to measure musical similarity by not focusing on the exact frequency, but rather on the chromatic scale where frequencies in a higher scale are also mapped to a standard scale. Psychoacoustically speaking, this is inspired by the way we usually perceive music in terms of relative and not absolute frequency. Chroma features are 12 dimensional corresponding to the twelve semitones in the western chromatic scale (A-G#).

Motivated by the success of Query By Humming systems[12] which use a hummed query as an exemplar, it has been suggested that extracting melody from only the vocals might add to the robustness of the feature in cover song identification problems. In [13], the authors propose to do this using an initial segmentation of non-vocal regions which are excluded from any further analysis. *Melody extraction* is then performed on only the remaining segments, yielding a 12 dimensional description similar to the chroma feature.

[1]http://www.youtube.com

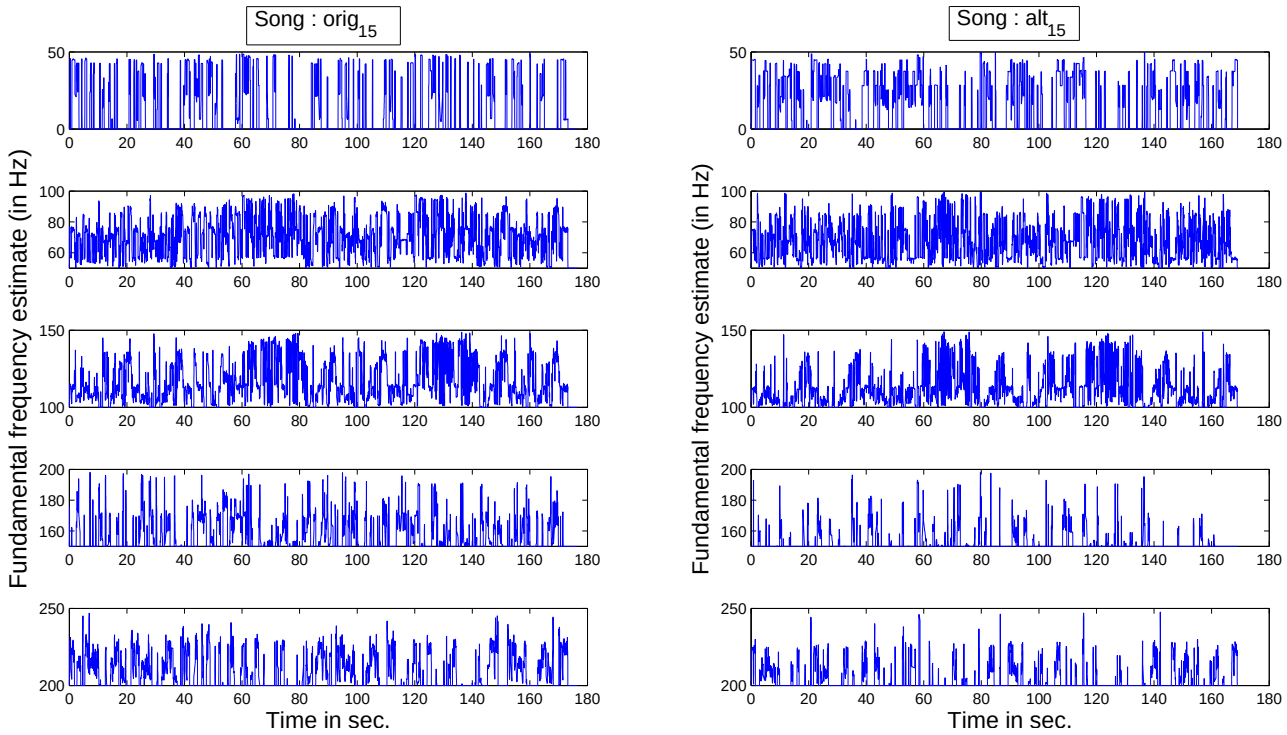

Fig. 1. MBAP features computed in the first five bands for the songs $\texttt{orig}_{15}$ (left) and $\texttt{alt}_{15}$ (right). Note the similarity in trend for features in the same band.

However, this method requires a large amount of training data for automatic segmentation of non vocal regions, and might be limited by the ambiguity in cases when there is a significant overlap between vocals and accompaniments. Instead, we seek to have feature representations that can retain the salient information irrespective of the segment of the song.

For a fair comparison, in this paper, we manually segment the non vocal regions in the songs, since the the best accuracy of the system reported in [13] was also corresponding to manual segmentation. All experiments are performed on both segmented and unsegmented versions of the songs.

IV. Proposed Features

A very useful feature in speech processing, pitch is widely used in speech prosody and expressive speaking style modeling. Pitch has also been used in music similarity tasks like Query by Humming (QBH) systems [12], where they are employed to extract the melody of the tune from the hummed voice. Typically defined as the period of the glottal pulses in speech, the concept of pitch in musical signals is usually not so well defined. This is mainly due to the existence of multiple pitch trajectories in music. The pitch, could for instance, correspond to a note played by some instrument, or alternatively to a singer's voice. There is however a difference in these two pitches, and neither are they perceived the same. Related studies in psychoacoustics [10] have shown that the singing voice in fact acquires some typical characteristics in its attempt to stand out from the rest of the accompaniments. This notion is used in fields like *predominant pitch detection* [14] to define a fundamental frequency for the signal.

While this ambiguity might limit the performance of overall pitch as a feature, it also provides an insight into the design of our proposed feature. Since defining the notion of a fundamental frequency for music is tricky, we first proceed to formally define it. For this purpose, we borrow from the auto correlation based pitch estimation algorithm commonly used in speech signals [15].

A. Fundamental Period Estimation

This estimation method assumes that the signal is quasi-periodic, and tries to use short time auto correlation to search for a fundamental period or frequency. Since this method is based on time domain processing, the local range that we search in determines the periods that we might expect to find. Specifically for a music signal $x[n]$ we define the auto correlation function as follows

$$R[k] = \sum_{m=0}^{L-1-k} x[m]x[m+k]; \forall k \in \left(\frac{F_s}{f_{max}}, \frac{F_s}{f_{min}}\right), k \in N$$

where L is the length of the signal, F_s the sampling frequency of the signal $x[n]$, while f_{min} and f_{max} set the upper and lower range for the fundamental period that we are trying to estimate.

Then, the largest peak of the auto correlation can be compared against some fixed threshold (fraction of $R[0]$) to check if a periodic component is present. If the the peak value is above the threshold, the period is defined to be the position

of the largest peak. This gives a fundamental frequency $f \in (f_{min}, f_{max})$. In practice, direct auto correlation is rarely used. Typically a method like 3 level center clipping is first used to preprocess the noisy signal. In this work, we used an implementation of this algorithm found in *Praat* [16].

B. Multi Band Auto correlation Peaks

Unlike [14] which attempts to model and track the "predominant pitch" in a song, we hypothesize that there exist multiple such pitch trajectories of interest in different bands, that might provide discriminative information. An attempt to extract a single pitch from such a signal would lead to a noisy pitch estimate, possibly jumping between these multiple pitch trajectories. Under the assumption that these trajectories do not overlap, we try to measure this information by computing the fundamental frequency in different bands by modifying the ranges f_{min} and f_{max} above. Specifically we use the following bands (in Hz).

TABLE II
BANDS IN WHICH MBAP FEATURES ARE EXTRACTED.

i	f_{min}	f_{max}
1	5	50
2	50	100
3	100	150
4	150	200
5	200	250
6	250	350
7	350	450
8	450	550

The MBAP features f_i are thus computed for the 8 bands in Table II. We do not consider any bands beyond these because any higher fundamental frequencies were found to be rare for the songs in our database.

Typically, feature estimates obtained via this method are noisy because of the interleaving non-periodic regions which have missing feature values. The usual convention is to indicate the absence of periodicity using zero values. Hence, post-processing of these features hence includes median filtering, interpolation and normalization of the ranges. All zero valued regions of length less than 0.5 seconds are replaced with linearly interpolated values, while the ones with a larger duration are clipped to a constant value of f_{min}.

Figure 1 shows sample MBAP features for the first 5 bands extracted for an `orig` song and its corresponding `alt` sample. The minimum value in each band is f_{min}. Also note the sparsity of MBAP features in higher bands because of absence of higher fundamental frequencies. In spite of differences in some regions, the feature trajectories have similar trends. Additionally, the figure adds to the intuition of using a multiple band description to help in cases when the features differ locally in one of the bands.

V. CLASSIFICATION SCHEME

Before diving into the details of classification, it is necessary to realize that the parallel clips in our dataset might not exactly be time aligned. This might be the case when the songs do not have the same tempo. Alternatively they might be arranged differently or have parts in one that are missing in the other. $\texttt{orig}_1$ for example might be from a movie clip where the song was used as a background music, whereas $\texttt{alt}_1$ might be from the original soundtrack or a concert recording. This can cause variations between the two versions, necessitating the use of temporal modeling techniques to find an optimal alignment along time. This problem being similar to edit distance approach between two time series, our natural choice was to use Dynamic Time Warping (DTW). Thus, we use DTW to match two songs, using the alignment cost as a classification metric for tune similarity. The rationale is that dissimilar songs will require heavy warping to align, thereby incurring a large alignment cost.

A. Alignment

Dynamic Time Warping has been used to find the similarity between time series, in fields like speech recognition [17][18], where the signals are not expected to be exactly aligned in time. This helps in cases where the signals may vary in time or speed (e.g. varying speaking rates), which can easily be the case for the problem at hand. In general, if we can assume that one of the series is the result of a non-linear warping of another, then DTW tends to find the best alignment. Given that certain restrictions like monotonicity hold, DTW exploits a dynamic programming framework to compute the best path in polynomial time. DTW is especially well suited for matching sequences with missing information which makes it a good match for our problem.

Given two time series $\texttt{orig}_i[m]$ and $\texttt{alt}_j[n]$, DTW returns a sequence of index tuples : $\{(m_1, n_1), (m_2, n_2), \ldots\}$ corresponding to the two series defining the best alignment. Then, the best alignment cost $\mathcal{C}$ can be computed for the time-aligned signals, using a predefined distance function. By substituting an appropriate distance function this algorithm can be extended to multidimensional signals. For our work, we use Mahalanobis distance [19] which normalizes for unequal variances along different feature dimensions. This is useful when combining features with different ranges or units. For matching two sequences of length M and N each, we precompute this distance matrix of size $M \times N$.

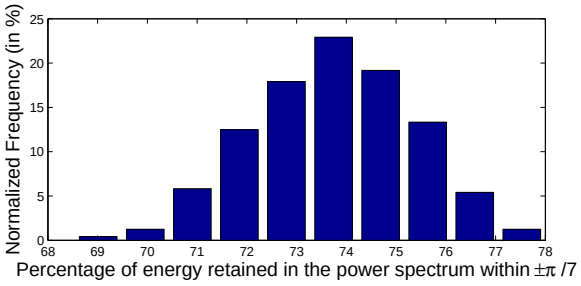

Fig. 2. Histogram showing effects of feature subsampling on the frequency content of the energy

B. Computational Feasibility

The use of dynamic programming for searching the best path through this matrix, reduces the complexity to polynomial time. However, the entire classification task still needs to compute L^2 DTW alignments on a database of $2L$ songs, for each choice of feature, which is significant computation. To reduce the complexity further, we subsample the time series of features by a factor of 7. The subsampling factor was chosen purely due to reasons related to computational capacity. We show that this subsampling doesn't alter the features significantly, because of the preceding smoothing operations (Section IV-B). In spite of subsampling the AC component of the features retain about 74% of the energy on an average (Figure 2).

After all the DTW alignments have been computed (Figure 3) classification comprises simply comparing the song alignment costs for all the song options. Let $\mathcal{C}_{ij}$ be the cost of aligning $\texttt{orig}_i$ with $\texttt{alt}_j$; $i, j = 1 \cdots 24$. We classify $\texttt{orig}_i$ as being similar to $\texttt{alt}_k$ for

$$k = \arg\min_j \mathcal{C}_{ij}$$

The classification accuracy is compared both ways considering `orig` and `alt` songs as the reference song by turns. The reported accuracy is the average of these two classification accuracies.

VI. Experiments and Results

We run classification experiments by computing DTW alignments for all pairs of files for all features. Then, this is used to compute the average classification accuracy as discussed above.

TABLE III
CLASSIFICATION ACCURACIES ON 24 SONGS USING DIFFERENT FEATURES (IN %)

Feature	Seg	Full
By Chance	2.1	2.1
Chroma	12.5	14.5
Melody	56.2	50.0
Pitch (250Hz)	41.6	52.0
MBAP (5 bands)	58.3	58.3
MBAP + Melody	52.1	58.3
MBAP (8 bands)	58.3	64.5
Pitch (550Hz)	64.6	72.9

The results illustrate that feature representations like melody and chroma computed on the entire spectrum suffer from lack of robustness when a clean music signal is not available. Owing to this, chroma features perform the worst in the classification task (Figure 3). Melody features, which aim to extract melody from the vocals in a song add to the robustness leading to a higher performance on manually segmented songs (Table III). MBAP features perform better using information from multiple bands including information that may not be captured by melody features.

To compare against pitch features, we use the same frequency content as MBAP features. Pitches upto 250Hz and 550Hz are extracted corresponding to 5 and 8 bands MBAP features. The hypothesis is that the additional complementary information in the bands leads to a higher accuracy on the classification task. While this clearly holds for the lower 5 bands, the results are slightly counter-intuitive for the higher bands. In spite of the sparsity in the higher bands, the performance for MBAP features still continues to improve because of the additional information. However, pitch extracted upto 550 Hz now outperforms the 8 band MBAP features.

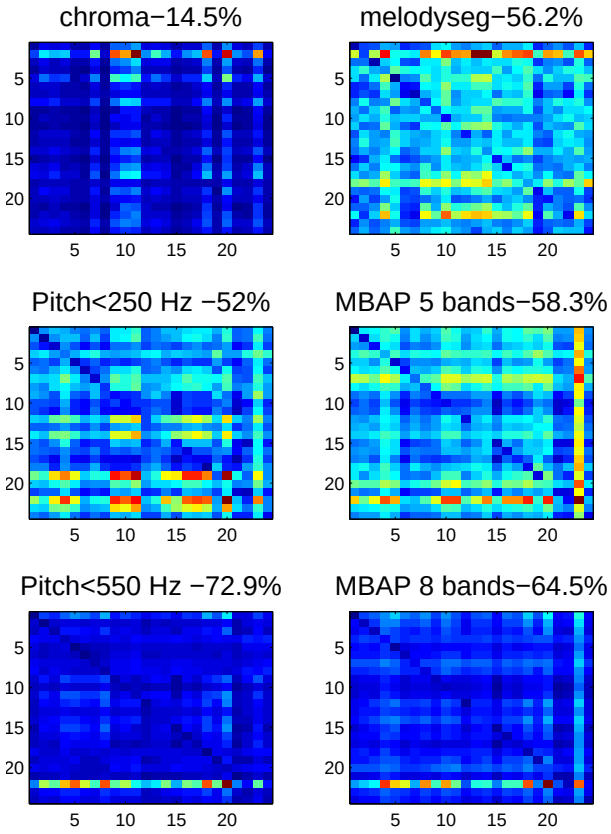

Fig. 3. 24×24 DTW alignment cost matrices for the features: Chroma, Melody, Pitch and MBAP. *A strong diagonal suggests a high classification accuracy (in title).*

This inconsistency in results can be explained by understanding the properties of MBAP features. Since these features are nothing but fundamental frequency estimates in different bands, they are sensitive to changes like a key/scale transpose. Moreover, since a transpose in a song shifts the frequencies [2]exponentially by a constant factor, the error is expected to be larger for the higher bands. In fact, such errors might cause pitch trajectories in one band to jump to another bands after transpose. Hence, classification accuracy might be penalized due to this modeling error. We verify this empirically, by performing the experiment only for the songs with no transposition. Not handicapped by the transpose variation, MBAP features perform better than pitch features(Table IV).

[2] $f = 2^{t/12} f_0$ where t is the transpose in semitones

TABLE IV
CLASSIFICATION ACCURACIES ON 13 SONGS IN THE SAME KEY (IN %)

Feature	Seg	Full
Pitch (250Hz)	61.5	61.5
MBAP (5 bands)	69.2	73.1
Pitch (550Hz)	65.4	76.9
MBAP (8 bands)	73.1	80.8

Since manual segmentation of non-vocal regions in the songs forms an important part of our experiments, it is also interesting to note its effect on different feature representations and their corresponding classification accuracies. We note that melody features are the only feature representations that benefit from removal of non-vocal regions in songs. This is supported by the intuition that these features are optimized to track melodies from vocals. For any other feature, removing non-vocal segments would mean throwing away discriminative information, which correlates with the decrease in performance. An exact binomial hypothesis test performed to compare our approach against baseline features, shows that the results are significant at the 10% significance level.

VII. CONCLUSION

In this paper, we discuss methods for tune similarity of songs in different languages. An unsupervised method was developed to find a matching song, from a data set of candidate songs, similar in tune corresponding to a given test song. Classification results on a dataset of 48 songs from Youtube suggest that the Multi Band Auto Correlation Peak (MBAP) features show improved performance compared to traditional features which either describe the entire spectrum, or focus specifically on a single aspect [9][6]. Further investigation is needed to make MBAP features more robust to shifts in key.

For future work we would like to verify our proposed feature representations on a larger and more diverse database of songs. Hence, our future efforts would focus on improving the classification scheme used in this paper, both in terms of efficiency and also adaptability to other more supervised schemes for wider applications. In addition we would like to explore other distance metrics that are robust to missing values and can deal with the sparsity of MBAP features in the higher bands.

As we see above, music similarity tasks are quite sensitive to variabilities in songs. One approach to make the MBAP features invariant to these, might be to estimate the transpose and compensate accordingly. Alternatively, it might also be helpful to try with band intervals other than the uniform ones used in this paper. For example, search and use of bands in a logarithmic scale might make the MBAP features more robust due to the exponential nature of the key transpose shift. In addition, an adaptive choice of the bands catering to specific properties of a song might be useful.

REFERENCES

[1] A. Uitdenbogerd and J. Zobel, "Matching techniques for large music databases," in *Proceedings of the 7th ACM International Multimedia Conference*. Citeseer, 1999, pp. 57–66.

[2] H. Shih, S. Narayanan, and C. Kuo, "An hmm-based approach to humming transcription," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 337–340.

[3] R. Typke, F. Wiering, and R. Veltkamp, "A survey of music information retrieval systems," 2005.

[4] S. Kim and S. Narayanan, "Dynamic chroma feature vectors with applications to cover song identification," in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*. IEEE, 2008, pp. 984–987.

[5] D. Ellis, C. Cotton, and M. Mandel, "Cross-correlation of beat-synchronous representations for music similarity," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 57–60.

[6] S. Kim, E. Unal, and S. Narayanan, "Music fingerprint extraction for classical music cover song identification," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 1261–1264.

[7] D. Ellis and G. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–1429.

[8] J. Serra, E. Gómez, and P. Herrera, "Audio cover song identification and similarity: background, approaches, evaluation, and beyond," *Advances in Music Information Retrieval*, pp. 307–332, 2010.

[9] W. Tsai, H. Yu, and H. Wang, "A query-by-example technique for retrieving cover versions of popular songs with similar melodies," in *Int. Symp. on Music Information Retrieval (ISMIR)*. Citeseer, 2005, pp. 183–190.

[10] J. Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, no. 3, pp. 82–91, 1977.

[11] E. Gómez and P. Herrera, "The song remains the same: Identifying versions of the same piece using tonal descriptors," in *Proc. ISMIR*, 2006, pp. 180–185.

[12] A. Ghias, J. Logan, D. Chamberlin, and B. Smith, "Query by humming: musical information retrieval in an audio database," in *Proceedings of the third ACM international conference on Multimedia*. ACM, 1995, pp. 231–236.

[13] W. Tsai, H. Yu, and H. Wang, "Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval," *Journal of Information Science and Engineering*, vol. 24, no. 6, pp. 1669–1687, 2008.

[14] Y. Li and D. Wang, "Detecting pitch of singing voice in polyphonic audio," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 3. IEEE, 2005, pp. iii–17.

[15] L. Rabiner and B. Juang, "Fundamentals of speech recognition," 1993.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," *Version*, vol. 5, p. 21, 2005.

[17] C. Myers and L. Rabiner, "Connected digit recognition using a level-building dtw algorithm," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 3, pp. 351–363, 1981.

[18] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[19] P. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Science, Calcutta*, vol. 12, 1936, p. 49.