TOWARDS THE CHARACTERIZATION OF SINGING STYLES IN WORLD MUSIC

Maria Panteli[1]*, Rachel Bittner*[2]*, Juan Pablo Bello*[2]*, Simon Dixon*[1]

[1]Centre for Digital Music, Queen Mary University of London, UK
[2]Music and Audio Research Laboratory, New York University, USA

ABSTRACT

In this paper we focus on the characterization of singing styles in world music. We develop a set of contour features capturing pitch structure and melodic embellishments. Using these features we train a binary classifier to distinguish vocal from non-vocal contours and learn a dictionary of singing style elements. Each contour is mapped to the dictionary elements and each recording is summarized as the histogram of its contour mappings. We use K-means clustering on the recording representations as a proxy for singing style similarity. We observe clusters distinguished by characteristic uses of singing techniques such as vibrato and melisma. Recordings that are clustered together are often from neighbouring countries or exhibit aspects of language and cultural proximity. Studying singing particularities in this comparative manner can contribute to understanding the interaction and exchange between world music styles.

Index Terms— singing, world music, pitch, features, unsupervised learning

1. INTRODUCTION

Singing is one of the primitive forms of musical expression. In comparative musicology the use of pitch by the singing voice or other instruments is recognized as a 'music universal', i.e., its concept is shared amongst all music of the world [1]. Singing has also played an important role in the transmission of oral music traditions, especially in folk and traditional music styles. We are interested in an across-culture comparison of singing styles using signal processing tools to extract pitch information from sound recordings.

In order to compare singing styles across several music cultures we require sound recordings to be systematically annotated. In the field of comparative musicology, annotation systems such as 'Cantometrics' [2] and 'Cantocore' [3] have been introduced. Pitch descriptors are well represented in such annotation systems. The most popular descriptors include the use of scales and intonation, the shape of the melodic contour, and the presence of melodic embellishments. For example, a study of 6251 European folk songs supports the hypothesis that musical phrases and melodies tend to exhibit an arch-shaped pitch contour [4].

In the field of Music Information Retrieval (MIR), research has focused on the extraction of audio features for the characterization of singing styles [5, 6, 7, 8]. For example, vibrato features extracted from the audio signal were able to distinguish between singing styles of, amongst others, opera and jazz [5]. Pitch class profiles together with timbre and dynamics were amongst the descriptors capturing particularities of a capella flamenco singing [6]. Pitch contours have also been used to model intonation and intonation drift in unaccompanied singing [7] and melodic motif discovery for the purpose of Indian raga identification in Carnatic music [8].

Singing style descriptors in the aforementioned MIR approaches are largely based on pre-computed pitch contours. Pitch contour extraction from polyphonic signals has been the topic of several studies [9, 10, 11, 12]. The most common approaches are based on melodic source separation [9, 10] or salience function computation [11, 12], combined with pitch tracking and voicing decisions. The latter two steps are usually based on heuristics often limited to Western music attributes, but data-driven approaches [13] were also proposed.

In this paper we focus on the characterization of singing styles in folk and traditional music from around the world. We develop a set of contour features capturing pitch structure and melodic embellishments. We train a classifier to identify pitch contours of the singing voice and separate these from non-vocal contours. Using features describing the vocal contours only we create a dictionary of singing style descriptors. The distribution of dictionary elements present in each recording is used for inter and intra singing style comparisons. We use unsupervised clustering to estimate singing style similarity between recordings and refer to culture-specific metadata and listening examples to verify our findings.

The contributions of this paper include a set of features for pitch contour description, a binary classifier for vocal contour detection, and a dictionary of singing style elements for world music. Our findings explore similarity within and between singing styles. Studying singing particularities in this comparative manner can contribute to understanding the interaction and exchange between world music styles.

This work was partially supported by the NYUAD Research Enhancement Grant # RE089 and the EPSRC-funded Platform Grant: Digital Music (EP/K009559/1).

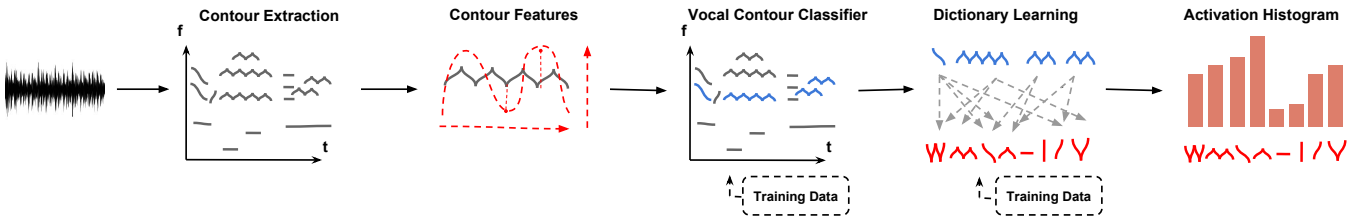

Fig. 1. Overview of the methodology (Section 3): Contours detected in a polyphonic signal, pitch feature extraction, classification of vocal/non-vocal contours and learning a dictionary of vocal features. Vocal contours are mapped to dictionary elements and the recording is summarized by the histogram of activations.

2. DATASET

Our dataset consists of 2808 recordings from the Smithsonian Folkways Recordings[1]. We use the publicly available 30-second audio previews and metadata and we choose information on the country, language, and culture of the recording as a proxy for similarity. In order to study singing style characteristics we select recordings that, according to the metadata, contain vocals as part of their instrumentation. We sample recordings from 50 different countries for geographical diversity and balance the dataset by selecting a minimum of 40 and maximum of 60 recordings per country (mean=56, standard deviation=6). Recordings span a minimum of 28 different languages and 60 cultures, but a large number of recordings lacks language or culture information. Additionally, a set of 62 tracks from the MedleyDB dataset [14] containing leading vocals was used as a train set for the vocal contour classifier (Section 3.3) and a set of 30 world music tracks containing vocal contours annotated using the Tony software [15] was used as a test set.

3. METHODOLOGY

We aim to compare pitch and singing style between recordings in a world music dataset. The methodology is summarized in Figure 1. We detect pitch contours for all sources of a polyphonic signal and characterize each contour by a set of pitch descriptors (Section 3.2). We use these features to train a binary classifier to distinguish between vocal and non-vocal contours (Section 3.3). Vocal contours as predicted by the classifier are further processed to create a dictionary of singing style elements. Each contour is mapped to the dictionary matrix and each recording is summarized by the histogram of its contour mappings (Section 3.4). Similarity between recordings is modeled via unsupervised clustering and intra- and inter-singing style connections are explained via references to the metadata and audio examples.

[1] http://www.folkways.si.edu

3.1. Contour extraction

We use the contour extraction method of Salamon et al. [12], which uses a "salience function", i.e. a time-frequency representation that emphasizes frequencies with harmonic support, and performs a greedy spectral magnitude tracking to form contours. Pitch contours detected in this way correspond to single notes rather than longer melodic phrases. The extracted contours covered an average of 71.3% (standard deviation of 24.4) of the annotated vocal contours across the test set (using a frequency tolerance of ± 50 cents). The coverage was computed using the multi-f_0 recall metric [16] as implemented in `mir_eval` [17]. Out of the 2808 recordings, the maximum number of extracted contours for a single track was 458, and the maximum number of extracted *vocal* contours was 85. On average, each track had 26 vocal contours (± 14), with an average duration of 0.6 seconds. The longest and shortest extracted vocal contours were 11.8 and 0.1 seconds respectively.

3.2. Contour features

Each contour is represented as a set of time, pitch and salience estimates. Using this information we extract pitch features inspired by related MIR, musicology, and time series analysis research. We make our implementations publicly available[2].

Let $c = (t, p, s)$ denote a pitch contour for time $t = (t_1, ..., t_N)$, pitch $p = (p_1, ..., p_N)$, salience $s = (s_1, ..., s_N)$, and N the length of the contour in samples. We compute a set of basic descriptors such as the standard deviation, range, and normalized total variation for pitch and salience estimates. Total variation TV summarizes the rate of change defined as

$$TV(x) = \sum_{i=1}^{N-1} |x_{i+1} - x_i|. \quad (1)$$

We compute $TV(p)$ and $TV(s)$ normalized by $\frac{1}{N}$. We also extract temporal information such as the time onset, offset and duration of the contour. These descriptors capture the structure of the contour at the global level but have little information at the local level such as the turning points of the contour or the use of pitch ornamentation.

[2] https://github.com/rabitt/icassp-2017-world-music

The second set of features focuses on local pitch structure modeled via curve fitting. We fit a polynomial y of degree d to pitch and salience estimates,

$$y[n] = \sum_{i=0}^{d} \alpha_i t_n^i \tag{2}$$

for polynomial coefficients α_i and sample $n = 1, ..., N$. We denote $y_p[n]$ and $y_s[n]$ as the polynomials fit to the pitch and salience features respectively. We store the coefficients α_i and the $L2$-norm of the residuals $r_p[n] = y_p[n] - p_n$ and $r_s[n] = y_s[n] - s_n$. The degree of polynomial is set to $d = 5$. These descriptors summarize the local direction of the pitch and salience sequences.

The third set of features models vibrato characteristics. Vibrato is an important feature of the singing voice and the characteristic use of vibrato can distinguish between different singing styles [5]. We model vibrato from the residual signal between the pitch contour and the fitted polynomial. The residual signal defines fluctuations of the pitch contour not captured via the smoothed fitted polynomial and is thus assumed to carry content of vibrato and other pitch embellishments. From the residual signal we extract descriptors of vibrato rate, extent, and coverage.

We approximate the residual $r_p[n]$ by a sinusoid $v[n]$ and amplitude envelope $A[n]$,

$$r_p[n] \approx A[n] * v[n] = A[n] cos(\bar{\omega} t_n + \bar{\phi}) \tag{3}$$

where $\bar{\omega}$ and $\bar{\phi}$ denote the frequency and phase of the best sinusoidal fit. The residual $r_p[n]$ is correlated against ideal complex sinusoidal templates along a fixed grid of frequencies, and $\bar{\omega}$ and $\bar{\phi}$ are the frequency and phase of the template with highest correlation. The amplitude envelope $A[n]$ is derived from the analytic signal of the Hilbert transform of the residual. The frequency $\bar{\omega}$ denotes the rate of vibrato and is constrained by the vibrato range of the singing voice as well as assumptions of fluctuation continuity in time. The latter is modeled via the vibrato coverage descriptor C which evaluates the goodness of sinusoidal fit in short consecutive time frames. This is modeled as

$$C = \frac{1}{N} \sum_{i=1}^{N} u_i \tag{4}$$

where

$$u_i = \begin{cases} 1, & \text{if } \frac{1}{w} \sum_{k=i-\frac{w}{2}}^{i+\frac{w}{2}-1} |r_p[k] - v[k]| < \tau \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

for some threshold τ, time frame of length w centered at sample i, and $r_p[k], v[k]$ the value of the residual and sinusoid, respectively, at sample k. The frame size w is set to the length of half a cycle of the estimated vibrato frequency $\bar{\omega}$.

Vibrato extent E is derived from the average amplitude of the residual signal, $E = \frac{1}{\hat{N}} \sum_{i=1}^{N} u_i A_i$ for $\hat{N}$ the total number of samples where vibrato was active. The pitch contour p is reconstructed by the sum of the fitted polynomial, the fitted sinusoidal (vibrato) signal, and some error, $p[n] = y_p[n] + E * u[n] * v[n] + \epsilon$. The reconstruction error ϵ is also included in our set of pitch contour features.

We extract in total 30 descriptors summarizing pitch content for each contour. These features are used as input to the vocal contour classifier (Section 3.3) and subsequently to learning a dictionary of singing elements (Section 3.4).

3.3. Vocal contour classifier

We trained a Random Forest Classifier to distinguish vocal contours from non-vocal contours using the features described in Section 3.2. Training labels were created by computing the percentage a given contour overlapped with the annotated vocal pitch, and labeling contours with more than 50% overlap as "vocal" (for more details, see [13]). The classifier was trained on 62 tracks from the MedleyDB dataset [14] containing leading vocals. The resulting training set contained a total of $\approx 60,000$ extracted contours, ≈ 7400 of which were labeled "vocal". Hyperparameters of the classifier were set using a randomized search [18], and training weights were adjusted to be inversely proportional to the class frequency to account for the unbalanced training set.

3.4. Dictionary learning

Given a selection of vocal contours and their associated features we learn a dictionary of the most representative pitch characteristics. Dictionary learning denotes an unsupervised feature learning process which iteratively estimates a set of basis functions (the dictionary elements) and defines a mapping between the input vector and the learned features. In particular, K-means is a common learning approach in image and music feature extraction [19, 20].

We learn a dictionary of contour features using spherical K-means, a variant of K-means found to perform better in prior work [21]. As a preprocessing step, we standardize the data and whiten via Principal Component Analysis (PCA). We use a linear encoding scheme to map contour features to cluster centroids, obtained by the dot product of the point with the dictionary matrix. We set $K = 100$ considering the diversity of countries, languages, and cultures in our dataset.

3.5. Singing style similarity

To characterize the singing style of a recording we sum the dictionary activations of its contours and standardize the result. We apply this to all recordings in our dataset which results in a total of 2808 histograms with 100 bins each. Using these histograms we apply K-means clustering to model similarity. The silhouette score is used to decide the number K of clusters that gives the best partition. Each cluster is considered a proxy of a singing style in our music collection.

4. RESULTS

4.1. Vocal Contour Classification

We tested the performance of the classifier on the 30 world music tracks (Section 2). The (class-weighted) accuracy on this set was 0.74 (compared with 0.95 on the training set), with a vocal contour recall of 0.64. This difference in performance can be attributed to differing musical styles in the training and test set - the training set contained primarily pop and Western classical vocals, while the test set contained vocal styles from across the world.

On the full dataset of 2808 recordings, extracted contours for which the probability of belonging to the vocal class was above 0.5 were considered vocal contours. False negatives (i.e., vocal contours undetected by the classifier) are of little consequence for subsequent analysis, as long as there are a sufficient number of vocal contours to describe the track. False positives, on the other hand, do affect our analysis, and we discuss an example of this in Section 4.2

4.2. Intra- and inter-style similarity

Using vocal contour features we learned a dictionary of singing elements (Section 3.4) and computed a histogram of activations for each recording. Similarity was estimated via K-means with $K = 9$ according to the silhouette score (Section 4.2). Figure 2 shows a visualization of the feature space of the recordings using a $2D$ TSNE embedding [22] and coloured by the cluster predictions[3]. Referring to the metadata we note that the majority of clusters represent recordings from neighbouring countries or similar culture or language. For example, cluster 6 groups mostly Eastern Mediterranean cultures, cluster 7 groups northern European cultures, clusters 3 and 5 group African and Caribbean cultures, and clusters $1, 8$ group mostly Latin American cultures.

Listening to some examples we observe that clusters can be distinguished by characteristic uses of vibrato, melisma, and slow versus fast syllabic singing. We note that vibrato denotes small fluctuations in pitch and melisma is the method of singing multiple notes to a single syllable. We observe that cluster 7 consists of slow syllabic singing examples with limited melisma but extensive use of vibrato. In this cluster we find examples of opera and throat singing techniques. Clusters $6, 8, 9$ consist of medium-slow syllabic singing with some use of vibrato but more prominent melisma. These clusters capture also instrumental examples of string instruments and aerophones. Clusters $3, 5$ consist of rather fast syllabic singing whereas cluster 1 consists of medium-fast syllabic singing with some use of melisma. Cluster 4 consists of medium-slow syllabic singing and some choir singing examples with voices overlapping in frequency range creating sometimes roughness or vibrato effects. Cluster 2, the points

[3] An interactive demo of Figure 2 can be found at `eecs.qmul.ac.uk/~mp305/TSNE.html`

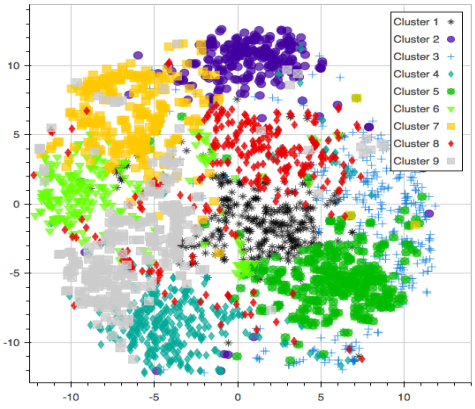

Fig. 2. A $2D$ TSNE embedding of the histogram activations of the recordings coloured by the cluster predictions.

of which seem to be slightly disconnected from the other clusters, denotes spoken language examples such as recitation of poems or sacred text.

5. DISCUSSION

Results showed that some recordings contained instrumental (non-vocal) or speech contours. The vocal contour classification task can be improved with more training examples from world music, and enhanced classes to cover cases of speech. We also observed sub-groups within clusters, for example clusters $6, 8, 9$, which indicates that clustering partitions can be further investigated. We based our observations on qualitative measures via listening to some examples and visualizing the clustered data. Future work aims to evaluate further the singing style clusters via a quantitative comparison with the metadata and using feedback from musicology experts.

6. CONCLUSION

In this paper we focused on the extraction of pitch contour features for the characterization of singing styles in world music. We developed a set of pitch features and used this to train a vocal classifier as well as to learn a dictionary of singing style elements. We investigated similarity in singing styles as predicted by an unsupervised K-means clustering method. Preliminary results indicate that singing style clusters often group recordings from neighbouring countries or with similar languages and cultures. Clusters are distinguished by singing attributes such as slow/fast syllabic singing and the characteristic use of vibrato and melisma. The investigation of singing styles as proposed in this study can provide evidence of interaction and exchange between world music styles.

7. REFERENCES

[1] S. Brown and J. Jordania, "Universals in the world's musics," *Psychology of Music*, vol. 41, no. 2, pp. 229–248, 2011.

[2] A. Lomax, *Cantometrics: An Approach to the Anthropology of Music*, Uneversity of California Extension Media Center, Berkeley, 1976.

[3] P. E. Savage, E. Merritt, T. Rzeszutek, and S. Brown, "CantoCore: A new cross-cultural song classification scheme," *Analytical Approaches to World Music*, vol. 2, no. 1, pp. 87–137, 2012.

[4] D. Huron, "The melodic arch in Western folksongs," *Computing in Musicology*, vol. 10, pp. 3–23, 1996.

[5] J. Salamon, B. Rocha, and E. Gomez, "Musical genre classification using melody features extracted from polyphonic music signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 81–84.

[6] N. Kroher, E. Gómez, C. Guastavino, F. Gómez, and J. Bonada, "Computational Models for Perceived Melodic SImilarity in A Capella Flamenco Singing," in *International Society for Music Information Retrieval Conference*, 2014, pp. 65–70.

[7] M Mauch, K Frieler, and S Dixon, "Intonation in Unaccompanied Singing : Accuracy , Drift and a Model of Reference Pitch Memory," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 1–11, 2014.

[8] V. Ishwar, S. Dutta, A. Bellur, and H. A. Murthy, "Motif Spotting in an Alapana in Carnatic Music," in *International Society for Music Information Retrieval Conference*, 2013, pp. 499–504.

[9] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564– 1578, 2007.

[10] J. Durrieu, G. Richard, B. David, and C. Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[11] K. Dressler, "An Auditory Streaming Approach for Melody Extraction from Polyphonic Music," in *International Society for Music Information Retrieval Conference*, 2011, pp. 19–24.

[12] J. Salamon, E. Gomez, and J. Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," in *International Conference on Digital Audio Effects*, 2011, pp. 73–80.

[13] R. M. Bittner, J. Salamon, S. Essid, and J. P. Bello, "Melody Extraction by Contour Classification," in *International Society for Music Information Retrieval Conference*, 2015, pp. 500–506.

[14] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," in *International Society for Music Information Retrieval Conference*, 2014, pp. 155–160.

[15] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the tony software: Accuracy and efficiency," in *International Conference on Technologies for Music Notation and Representation*, 2015.

[16] M. Bay, A. F. Ehmann, and J. S. Downie, "Evaluation of multiple-f0 estimation and tracking sys- tems," in *International Society for Music Information Retrieval Conference*, 2009, pp. 315–320.

[17] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir eval: A transparent imple- mentation of common mir metrics," in *International Society for Music Information Retrieval Conference*, 2014, pp. 367–372.

[18] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[19] A. Coates and A. Y. Ng, "Learning feature representations with K-means," in *Neural Networks: Tricks of the Trade*, pp. 561–580. Springer Berlin Heidelberg, 2012.

[20] J. Nam, J. Herrera, M. Slaney, and J. Smith, "Learning Sparse Feature Representations for Music Annotation and Retrieval," in *International Society for Music Information Retrieval Conference*, 2012, pp. 565–570.

[21] S. Dieleman and B. Schrauwen, "Multiscale Approaches To Music Audio Feature Learning," in *International Society for Music Information Retrieval Conference*, 2013, pp. 116–121.

[22] L.J.P van der Maaten and G.E. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.