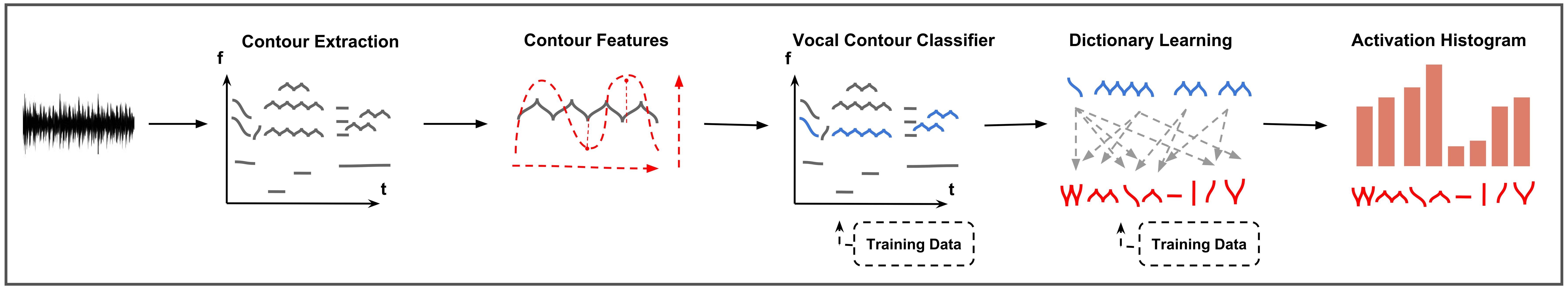
Towards the Characterization of Singing Styles in World Music

Maria Panteli[1], Rachel Bittner[2], Juan Pablo Bello[2], Simon Dixon[1]
m.panteli@qmul.ac.uk
[1] Centre for Digital Music, Queen Mary University of London, UK
[2] Music and Audio Research Laboratory, New York University, USA

Motivation

Singing has played an important role in the transmission of oral music traditions, especially in folk and traditional music styles [1]. By processing melodic contours from vocal recordings and applying unsupervised clustering we explore the space of singing style similarity in world music.



Dataset

- 2808 recordings from 50 countries
- mean=56, std=6 recordings per country
- 28 languages and 60 cultures

Smithsonian Folkways Recordings

Contour Extraction

- method: salience function [2]
- mean=26, std=14 vocal contours per track
- duration mean=0.6 seconds
- contour $c = (t, p, s)$, time $t = (t_1, ..., t_N)$, pitch $p = (p_1, ..., p_N)$, salience $s = (s_1, ..., s_N)$

Vocal Contour Classifier

- train set: 60,000 contours from MedleyDB tracks [3]
- test set: 4,000 contours from world tracks annotated with Tony [4]
- random forest classifier, classes: vocal, non-vocal
- class-weighted accuracy: **0.74**
- vocal contour recall: **0.64**

Contour Features

bit.ly/contours_code

30 features capturing:

1) global structure of the contour:
   pitch range, duration, total variation of pitch/salience estimates
2) local pitch structure modeled via curve fitting:
   polynomial coefficients α_i, residual from fitted curve (L2-norm)

$$y_p[n] = \sum_{i=0}^{d} \alpha_i p_n^i \qquad r_p[n] = y_p[n] - p_n$$

3) vibrato characteristics modeled from residual of fitted curve

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

- vibrato rate $\bar{\omega}$: frequency of best sinusoidal fit
- vibrato extent, average $A[n]$: analytic signal of Hilbert transform
- vibrato coverage : goodness of sinusoidal fit over time

Dictionary Learning

- dictionary learning via spherical K-means [5], K=100
- linear encoding: project contour features onto cluster centroids
- activation histogram for each recording: sum of contour mappings
- singing styles: via K-means clustering of activation histograms, K=9 based on silhouette score

Results

bit.ly/TSNE_demo

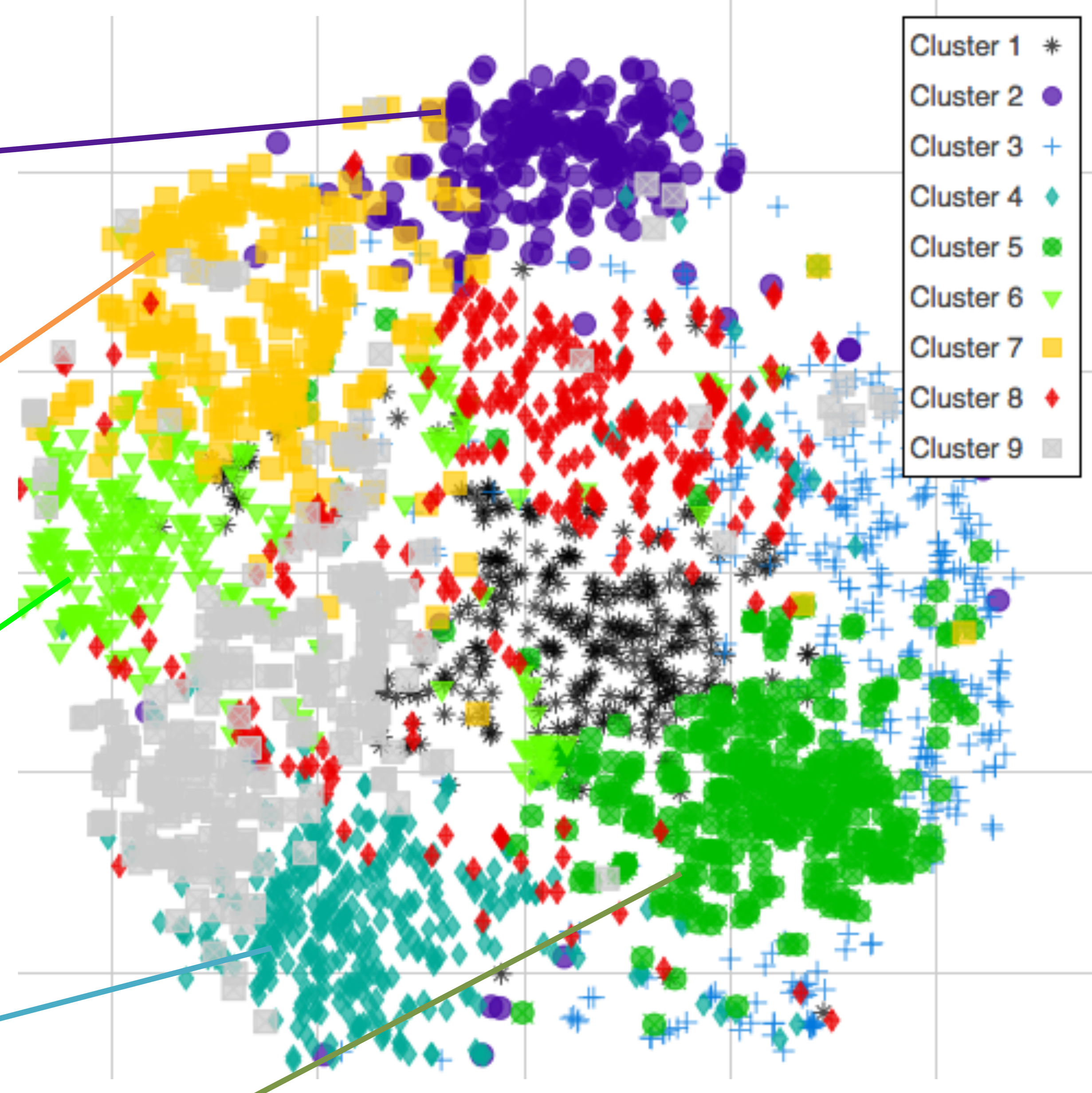
cluster 2: speech, recitation of poems and sacred text

cluster 7: extensive vibrato, opera and throat singing, northern European

cluster 6: prominent melisma, Eastern Mediterranean

cluster 4: medium-slow syllabic singing, choir, roughness

cluster 5: fast syllabic, African and Caribbean



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] J. Salamon, E. Gomez, and J. Bonada, “Sinusoid extraction and salience function design for predominant melody estimation,” in *DAFx*, 2011, pp. 73–80.

[3] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, “MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research,” in *ISMIR*, 2014, pp. 155–160.

[4] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Sala- mon, J. Dai, J. Bello, and S. Dixon, “Computer-aided melody note transcription using the tony software: Accuracy and efficiency,” in *Technologies for Music Notation and Representation*, 2015.

[5] S. Dieleman and B. Schrauwen, “Multiscale Approaches To Music Audio Feature Learning,” in *ISMIR*, 2013, pp. 116–121.