YOU CALL *THAT* SINGING? ENSEMBLE CLASSIFICATION FOR MULTI-CULTURAL COLLECTIONS OF MUSIC RECORDINGS

Polina Proutskova
Department of Computing
Goldsmiths, London, UK
p.proutskova@gold.ac.uk

Michael Casey
Bregman Music Research Laboratory
Dartmouth College, USA
mcasey@Dartmouth.edu

ABSTRACT

The wide range of vocal styles, musical textures and recording techniques found in ethnomusicological field recordings leads us to consider the problem of automatically labeling the content to know whether a recording is a song or instrumental work. Furthermore, if it is a song, we are interested in labeling aspects of the vocal texture: e.g. solo, choral, acapella or singing with instruments. We present evidence to suggest that automatic annotation is feasible for recorded collections exhibiting a wide range of recording techniques and representing musical cultures from around the world. Our experiments used the Alan Lomax *Cantometrics* training tapes data set, to encourage future comparative evaluations. Experiments were conducted with a labeled subset consisting of several hundred tracks, annotated at the track and frame levels, as acapella singing, singing plus instruments or instruments only. We trained frame-by-frame SVM classifiers using MFCC features on positive and negative exemplars for two tasks: per-frame labeling of singing and acapella singing. In a further experiment, the frame-by-frame classifier outputs were integrated to estimate the predominant content of whole tracks. Our results show that frame-by-frame classifiers achieved 71% frame accuracy and whole track classifier integration achieved 88% accuracy. We conclude with an analysis of classifier errors suggesting avenues for developing more robust features and classifier strategies for large ethnographically diverse collections.

1. INTRODUCTION

We explore approaches for MIR and ethnomusicology to support each other in the area of cross-cultural research and to contribute new tasks and observations to both fields. Ethnomusicological recordings constitute a major challenge to MIR tools, due to their musical, acoustic and technical diversity, so they can help improve our understanding of machine-music interaction. MIR methods are also driving interest in larger-scale, data-intensive cross-cultural studies in music.

Ethnomusicological recordings document musical repertories outside of Western classical and popular music, often those that are endangered or extinct today. These recordings are used for education or research on these repertories. Some collections have been commercially released by record labels or cultural organizations [1]. Now, due to easily accessible recording equipment, the volume of recordings is growing exponentially. This poses new challenges in managing ethnomusicological collections which can hold up to hundreds of thousands recorded items with tens to thousands of hours of recordings, though only a fraction of these are currently digitized [2].

Recording quality within collections varies greatly and there is often little or no information about the technical and acoustic context for the recording. Sometimes the singer or the leading instrument are not the most dominant part of the recording; social contexts vary greatly from the concert or album settings common in Western musical culture; field recordings often contain sounds of social and natural environments as well as other noise conditions. However, the greatest challenge is the variance in musical material: even if a given collection is homogeneous, the content will differ greatly from Western music so requiring new MIR approaches.

This paper concerns automatic annotation, both at the frame level and track level, of ethnomusicological field recordings. The qualitative nature of their research requires us to approach ethnomusicologists carefully when proposing new technology. For example, a single classification error can have a far reaching impact on interpretation, so we must consider classification errors and their causes and document these for users of automatically labeled archives. To this end, we give an overview of previous work in Section 2, describe our classification experiments in Section 3, present our results and offer detailed observations on classifier errors in Section 4 and conclude in Section 5.

2. PREVIOUS WORK

Downie [3] and Tzanetakis et al. [4] have stressed the need for research on ethnomusicological collections. But publications in this area are still rare in large part due to there being few recorded collections available with annotations to use as ground truth data. Tzanetakis et al. [4]

provide an overview of MIR work related to non-Western musical content and suggests basic guidelines for this kind of study, but this work does not consider cross-cultural research, involving heterogeneous collections, which we are predominantly interested in.

Previous systems for detection of singing employ frame-by-frame classification on data consisting primarily of Western popular music [5,6,7]. These studies employ MFCC features, sometimes with derivatives and other spectral features, combined with statistical models using combinations of Neural Networks, HMMs, GMMs or SVMs to classify into two categories: frames containing singing and non-sung frames. Temporal smoothing is often applied to reduce labeled region fragmentation [6]. Our work differs in the variance of both acoustic and musical conditions of the training and testing data and in the detailed consideration of classifier errors with respect to this variation.

Wembu and Baumann [8] suggested that SVM classifiers yield slightly better results than HMMs and GMMs. Only a few studies used frame-by-frame labeling of the ground truth [9] which we consider to be essential to accurate evaluation. Two MIR studies are related to singing in non-Western cultures: singer identification in Greek Rembetiko [9] and in South Indian Carnatic music [10]. Both of these employed MFCCs. The latter team used signal separation for distinguishing between vocal and instrumental frames.

Other studies involving non-Western music recordings consider single musical cultures or repertories, each of which provides some homogeneity of the musical material. Several researchers performed rhythmic analysis and classification based on beat features for such repertories like Malay, Greek, Central African traditional music as well as Afro-Cuban music [11,12,13]. Chordia et al. [14] found a statistical measure based on pitch classes that distinguishes between different ragas. Sridhar and Geetha [10] developed Carnatic interval cepstral coefficients (CICC), based on the division of the octave in 22 Sruti, to better suit the tonality structure of the Indian Carnatic music.

Holzapfel et al. [9] compiled a database of Rembetiko singers for their artist recognition experiment. They deliberately chose recordings that are very similar in style to avoid identification due to style differences. They labelled training data frame-by-frame with one second window hop. They used aggregate "world model" GMMs to distinguish vocal from instrumental frames using intersection of the maximum-likelihood and minimum likelihood frames of opposing classifiers. This technique resulted in a classification accuracy of 99%. The data set included historical grammophone recordings, but the study was for a homogenous musical style over 21 Rembetiko singers.

Cross-cultural MIR studies include discriminating mood taxonomy of Chinese traditional music and Western classical music [15], retrieval through metric similarity for Greek and Central African music [12] and a study on metrical ambiguity in Bossa Nova, Gahu, Rumba, Soukous, Son, and Shiko [16].

We take a more general approach. To establish a model for cross-cultural research we require a sufficiency of training data to account for variance caused by difference in cultural origins, in recording techniques and in musical textures.

3. VOCAL/INSTRUMENT CLASSIFICATION

The purpose of our study is the evaluation of the baseline performance of widely-used MIR methods on an ethnographically diverse data set and to gain insight into future research potential by performing a detailed analysis of any misclassifications. We prepared a data set consisting of excerpts taken from the Lomax Cantometrics training tapes collection [17,18] which contains a high degree of cultural, technical and textural variance since the data was originally collected to find correlations between musical style and cultural traits such as social organization of a society.

3.1 Data

The Lomax data set consisted of 1000 tracks from all over the globe including recordings sung in different languages, music played on "exotic" instruments, singing, polyrhythmic as well as rhythmically free melodies, non-diatonic, non-tempered scales, a great diversity of voice timbres, rhythms, harmonies and textures from heterophonic to uncoordinated, with considerable variation in the social organization of the performing group.

For our experiments we used 355 of approximately 1000 sound samples. Of these 355 tracks 297 contain singing and 58 are purely instrumental. Of the singing tracks 185 are a'capella singing and 112 contain accompanying instruments; 110 are sung solo, 130 are choral and 57 contain both solo and group singing; 166 tracks contain primarily male singing (solo or group), 60 female singing and 51 mixed male and female singing, 10 are sung by children. More than 50 cultures are represented in the database from 5 continents as well as from large and small islands. Instruments include all kinds of idiophones (rattles, drums, frame drums, sticks, gamelans, xylophones), aerophones (flutes, clarinets, trumpets, tuba, didgeridoo), chordophones (all kinds of lutes, zithers, bow chordophones like fiddles and classical violins).

We received the audio in MP3 128kbt/sec, 44,1kHz. Files had durations of between 10 and 150 seconds. For each audio file we extracted 20-band Mel Frequency Cepstral Coefficients (MFCC). These were extracted using a short-time Fourier transform with hop size 100ms (2205 samples), window length 185.76ms (8192 samples), FFT length 16384 samples (2.69Hz frequency bins). For classification and evaluation we developed tools in Matlab using the libsvm package [19].

Audio files were annotated at the whole track level as being predominantly sung acapella (sa), instrumental (i) or singing plus instruments (si). For a subset we performed frame-by-frame labelling at 50ms increments: 111 for singing (sa + si) vs purely instrumental (i) and 77 for acapella singing (sa) vs instrumental or accompanied singing (i + si).

3.2 Frame-level and track-level classification

The first experiment was to determine the performance of frame-by-frame two-class SVM classifiers consisting of a) frames with singing, consisting of (sa) acapella and singing with instruments (si), verses instruments only (io) and b) acapella frames (sa) versus non-acapella frames; i.e. frames containing singing with instruments (si) and frames containing instruments only (io). The positive and negative training labels used for the binary SVM classifiers are summarized in Table 1.

Building on these two classifiers, we defined a third task to classify whole tracks as predominantly acapella singing (SA), instruments only (IO) or singing plus instruments (SI). The method used for the third task was whole-track integration of the frame-by-frame two-class SVM classifier outputs from the first two tasks.

Classifier	Labels +	Labels -
Sung Frame (s)	(sa) (si)	(io)
Acapella Frame (sa)	(sa)	(si) (io)

Table 1: training labels for binary classifiers used in the experiments.

For the first round of our experiment we conducted leave-one-out cross-validation on 36 songs labelled frame-by-frame for singing (s) vs. pure instrumental (si) and 30 tracks for acapella singing (sa) vs instrumental (si). We tested on whole songs so that the test set did not consist of frames drawn from any training track.

In the second round we used 111 labelled tracks for cross-validation (77 tracks for pure singing vs instrumental classifier). We applied different ways of preprocessing features to obtain better results: i.e. removing the first MFCC band, unit-norming feature vectors, detecting and removing quiet frames. We also incorporated temporal aspects of features in two ways: derivatives of the feature vectors and concatenation of sequences of three to five feature vectors. These modifications did not influence our results significantly.

A third experiment was conducted for the sung (s) vs pure instrumental (io) classifier for which we trained an SVM model on all 111 frame-by-frame labelled songs and predicted labels using this model for a test set of 244 new tracks, of which 237 contained singing and 7 were purely instrumental. We integrated the classifier output labels to construct whole song predictions for the test set using 30% sung frames as a threshold for a sung track. We compared these predictions with our manual annotation of the test songs to evaluate accuracy.

4. RESULT

The results are summarized in Tables 2-4. The first experiment yielded a mean accuracy of 74% for the sung frame classifier and 77% for the acapella frame classifier. There were a number of problem cases which had a major impact on prediction accuracy, these are discussed in the next section. When recordings from these problem groups were removed from the data set, the mean accuracy of the cross-validation on remaining 27 songs was 87.9% with 18% standard deviation for the sung frame classifier, a substantive improvement. For the next round of the experiment we paid special attention to these problem cases and included additional tracks with these characteristics into the training set.

For the third experiment, the accuracy was 83.6% for sung track classification and 62.2% of acapella tracks in the collection were correctly identified. However, 97% of tracks labeled as sung contained singing which means that the false positive rate for instrumental tracks was impacting performance. We discuss the false positives in the next section as well as what might be done to improve classifier performance.

27 tracks (problem cases excluded)	Mean accuracy	87.9%
	Std. deviation	18.0%
36 tracks	Mean accuracy	73.6%
	Std. deviation	26.0%
111 tracks	Mean accuracy	71.5%
	Std. deviation	22.4%
	Mean recall singing	83.9%
	Mean precision singing	52.6%
	Mean recall pure instrumental	76.1%
	Mean precision pure instrumental	60.5%

Table 2: singing vs. pure instrumental classifier: leave-one-out cross-validation results.

27 tracks (problem cases excluded)	Mean accuracy	85.3%
	Std. deviation	19.5%
30 tracks	Mean accuracy	77.4%
	Std. deviation	26.6%
77 tracks	Mean accuracy	71.1%
	Std. deviation	22.4%
	Mean recall singing	85.2%
	Mean precision singing	51.9%
	Mean recall pure instrumental	76.7%
	Mean precision pure instrumental	58.9%

Table 3: acapella singing vs. instrumental classifier: leave-one-out cross-validation results.

	Singing vs pure instrumental classifier	Acapella singing vs instrumental classifier
Nr training tracks	111	77
Nr test tracks	244	278
Accuracy	83.61%	62.23%
Recall positive	85.65%	42.28%
Precision positive	97.13%	76.83%
Recall negative	14.29%	85.27%
Precision negative	2.86%	56.12%

Table 4: whole track tests results.

5. PROBLEM CASES ANALYSIS

5.1 False positives / false negatives analysis

In general, the reasonable accuracy of our frame-by-frame classifiers, which is further improved in the whole track predictions, suggests this kind of classification can be achieved independently of the origin of musical material and the variance in stylistic parameters.

Similar results of the first and the second round show that we were able to include some of the variance of the data which caused problems in the first round into the training set of the second round. The following sections outline our analysis of the errors of classification and what might have caused them.

The following instrumental sounds were predicted badly during the first experiment and also negatively influenced the ability of the classifiers to predict singing when they were present in the training set:

1. woodwind instruments with a lot of "air" in the sound, like pan pipes also organs, mouth organs
2. shortly plucked or hammered instruments, like mbira, xylophone or marimba, some lutes
3. the dominant instrument playing the melody was misclassified as singing
4. tracks containing both singing and instruments were misclassified as purely instrumental
5. well blended choral singing with a wide pitch range was misclassified as instrumental
6. yodeling was misclassified as instrumental

Cases 3 and 4 could be practically eliminated in the second round of cross-validation with more training data. The performance of the classifiers with the other problem cases has improved with additional training data, but examples of these cases were still present among false positives/false negatives in the second round.

In the second experiment there were some tracks that were misclassified: mbira and mouth organ, gamelans and xylophones, flutes and clarinets (especially when played while "singing" into them) could "cheat" the classifier yielding frame-by-frame accuracies of 30% and lower in these cases. False instrumenttal positives were caused by the Russian state choir with very well blended, wide ranging vocals; the raspy, narrow, heterophonic choral singing of Marajin from Australian Arnhemland; a gospel choir with the roaring sound of Louis Armstrong and an extremely high soprano voice.

Another group of tracks with had classification results close to chance. These also included the examples from above plus:

i. choral singing of complex interlocked motivic structures; this is what Victor Grauer [20] calls pygmy/bushmen style. Also interlocked pan pipes playing, which he considers to be evolutionary related to the pygmy/bushmen style
ii. discoordinated singing in a big group
iii. fiddles, whistles, country/blues guitar and mouth har
iv. wind section of a classical orchestra

The whole track tests problem cases analysis exposed similar problem cases (mbira, xylophones, flutes, discoordinated singing, narrow, low pitched voice, complex polyphonic choral performance) but also introduced new cases which apparently were not included into the training data, such as singing with strong accents, like e.g. Native Americans from the Iroquois Confederacy; voice imitating instruments, e.g. percussion; sprechgesang (very fast spoken/sung text).

To summarize, following classes of sounds are likely to cause misclassification:

1. Instrumental:
   - All kinds of idiophones: drums, percussions, rattles; xylophones; lamellohpones like mbira; gongs like gamelans
   - Aerophones: flutes, clarinets, whistles, pan pipes (but not bagpipes), mouth harp, mouth organ
   - Fiddle and guitar, all kinds of lutes
2. Vocal (solo and homophonic):
   - Singing voice with extreme characteristics: very low or very high pitched; very narrow, nasal or raspy; voices with significant non-harmonic components in the spectrum; brilliance (strong higher frequency components) in the voice; yodeling.
   - Voice imitating instruments or singing with very strong accents
   - sprechgesang, very fast spoken/sung text
3. Polyphonic textures
   - contrapuntal, heterophonic, interlocked as well as discoordinated performances in a wide range, by an orchestra and/or a choir

Recording quality is an important factor for classification accuracy. Though the quality of audio on Cantometrics training tapes is much more varied then of any modern collection of classical or popular music, it is considerably better than many ethnomusicological datasets. We observed misclassification of recordings with extreme sound distortion, but in general the classifiers were able to cope with significant variation in recording quality.

Temporal changes presumably play an important role in distinguishing singing. As is known from speech signal processing, human speech as well as singing contains speech formants which are specific for each vocal ('a', 'e', 'i', 'o', 'u') and change from syllable to syllable. This kind of change in the formants is absent in the spectrum of practically all musical instruments.

We tried incorporating temporal changes, using MFCC derivatives, into the features but that didn't show any significant difference in the result. This suggests using a shorter hop size and window size for our features. The shorter hop size will increase the number of features so this will likely push the running time for experiments above acceptable durations.

5.2 Future Research

A MIREX-like comparison of performance on the Cantometric training tapes dataset would determine the best and cheapest approach and would uncover models impli-

citly relying on musical features of a specific culture like Western popular musics. Also systematic research into the feature selection for this type of classification is needed.

The next step in approaching the question of generalization with respect to cultural origin and musical style would be to test the SVM model we have trained on other musical collections. Will we be able to detect singing within collections of classical music, popular music, folk songs, non-Western recordings? If not, how much training data is missing, what kind of variance is not covered by our training set? Is the goal of having a single model to detect singing in all music achievable?

The statistical framework using SVMs of our experiments is scalable for use with tens of thousands of tracks. The training on 111 tracks takes a few minutes and prediction takes about 1 sec per song on a current high-end laptop. Prediction runs sequentially on every song, thus the testing is O(n) of the number of tracks to be predicted. With the current model we expect the approximate running time of prediction for 10 000 tracks to be less than 3 hours on our system. Assuming the generality of the model, it allows our software to be applied to much bigger collections of any musical style and origin.

It is also apparent that the same statistical infrastructure can be used to automatically classify other frame-level musical features. This will need a new round of frame-by-frame as well as whole track labeling. We plan to use this approach to classify tracks with singing into solo and choral singing, male, female and mixed singing, to detect specific style patterns like yodeling and drones. It also suggests that we should segment audio according to the prediction of more general classifiers and design hierarchical classification, for instance male/female on segments with singing; or otherwise to use this segmentation as a preliminary step for other techniques, such as pitch extraction for solo singing.

Combining these features with musical parameters obtained by other techniques (such as the amount of percussivity) or of a larger scope (such as average pitch) one would get a multi-facet description of the musical style of a track. Such a representation of a musical style applied to the Cantometrics training tapes opens up various possibilities: to study geographic distribution of a musical parameter, a combination of parameters or style patterns (e.g. choral vs. solo singing or yodel); to revise delineation of music cultures; to study the dynamics of musical style spread and influence. Since this style description is compact and can be extracted automatically from audio, it is easy to add further tracks to the data set and continually refine this research.

6. CONCLUSION

In this paper we presented our work on manual annotation of a diverse ethnomusicological collection for the purposes of testing MIR tools for automatic annotation. We conducted three experiments that served to demonstrate good performance on the data set for the task of labeling regions of different types of sung tracks versus non-sung tracks. We also presented an analysis of errors that suggests strategies for improving the overall accuracy of such classifiers.

We would like to see these methods applied to larger real world collections in ethnomusicological archives enhancing access to our cultural heritage. Practically every public ethnomusicological archive has poorly annotated holdings. Also small private archives are growing fast and could benefit from this kind of automatic annotation. Today, more than ever, technological infrastructure is needed for these types of recordings: archives are challenged to open up their collections, to make them "user-friendly", to provide not only content, but added value like expertise, easy access and fun. Having a whole collection annotated in a consistent way would allow the design of new user interfaces that are able to graphically represent style patterns and regions. Social tagging could be used to counter the errors generated in automatic annotation. Combined with archivists' expertise and moderation, this approach will enable archives to close gaps in annotation and offer hands-on activities to their user communities.

7. REFERENCES

[1] R. Reigle, "Humanistic motivations in ethnomusicological recordings," in *Recorded Music - Philosophical and Critical Reflections* (M. Dogantan-Dack, ed.), Middlesex University Press, 2009. with CD

[2] P. Proutskova, "Musical memory of the world - data infrastructure in ethnomusicological archives," *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[3] J. S. Downie, "Music information retrieval," *Annual Review of Information Science and Technology*, vol. 37, pp. 295–340, 2003.

[4] G. Tzanetakis, A. Kapur, W. A. Schloss, and M. Wright, "Computational ethnomusicology," *journal of interdisciplinary music studies*, vol. 1, no. 2, pp. 1–24, 2007.

[5] G. Tzanetakis, "Song-specific bootstrapping of singing voice structure," *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, vol. 3, pp. 2027– 2030, June 2004.

[6] A. Berenzweig, D. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," *AES 22nd International Conference*, 2002.

[7] A. Berenzweig and D. Ellis, "Locating singing voice segments within music signals," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[8] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," *Proceedings of the International Symposium on Music Information Retrieval*, 2005.

[9] A. Holzapfel and Y. Stylianou, "Singer identification in rembetiko music," *Sound and Music Computing*, 2007.

[10] R. Sridhar and T. Geetha, "Music information retrieval of carnatic songs based on carnatic music singer identification," *Computer and Electrical Engineering, 2008. ICCEE 2008. International Conference on*, pp. 407 – 411, Dec 2008.

[11] S. Doraisamy, S. Golzari, N. M. Norowi, N. Sulaiman, and N. I. Udzir, "A study on feature selection and classification techniques for automatic genre classification of traditional malay music," *Pro-*

ceedings of the International Symposium on Music Information Retrieval, 2008.

[12] I. Antonopoulos, A. Pikrakis, S. T. O. Cornelis, D. Moelants, and M. Leman, "Music retrieval by rhythmic similarity applied on greek and african traditional music," *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[13] M. Wright, G. Tzanetakis, and A. Schloss, "Analyzing afro-cuban rhythm using rotation-aware clave template matching with dynamic programming," *Proceedings of the International Symposium on Music Information Retrieval*, 2008.

[14] P. Chordia and A. Rae, "Raag recognition using pitch-class and pitch-class dyad distributions," *Proceedings of the International Symposium on Music Information Retrieval*, 2007.

[15] W. Wu and L. Xie, "Discriminating mood taxonomy of chinese traditional music and western classical music with content feature sets," *Image and Signal Processing, 2008. CISP '08. Congress on*, vol. 5, pp. 148 – 152, May 2008.

[16] G. Toussaint, "A mathematical analysis of african, brazilian, and cuban clave rhythms," *Proceedings of BRIDGES: Mathematical Connections in Art, Music and Science*, pp. 157–168, 2002.

[17] A. Lomax, *Folk Song Style and Culture*. New Brunswick, New Jersey: Transaction Books, 1968.

[18] A. Lomax, *Cantometrics: An Approach To The Anthropology Of Music*. The University of California, 1976. accompanied by 7 cassettes.

[19] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines,* 2001. Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

[20] V. Grauer, "Echoes of our forgotten ancestors," *The World Of Music*, vol. 2, 2006.s