MUSICAL GENRE CLASSIFICATION USING MELODY FEATURES EXTRACTED FROM POLYPHONIC MUSIC SIGNALS

Justin Salamon[⋆] *Bruno Rocha*[†] *Emilia Gómez*[⋆]

Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
[⋆]{justin.salamon,emilia.gomez}@upf.edu
[†]brunomachadorocha@gmail.com

ABSTRACT

We present a new method for musical genre classification based on high-level melodic features that are extracted directly from the audio signal of polyphonic music. The features are obtained through the automatic characterisation of pitch contours describing the predominant melodic line, extracted using a state-of-the-art audio melody extraction algorithm. Using standard machine learning algorithms the melodic features are used to classify excerpts into five different musical genres. We obtain a classification accuracy above 90% for a collection of 500 excerpts, demonstrating that successful classification can be achieved using high-level melodic features that are more meaningful to humans compared to low-level features commonly used for this task. We also compare our method to a baseline approach using low-level timbre features, and study the effect of combining these low-level features with our high-level melodic features. The results demonstrate that complementing low-level features with high-level melodic features is a promising approach.

***Index Terms*—** Genre classification, melody extraction, pitch contour

1. INTRODUCTION

Genre classification involves the assigning of categorical labels to pieces of music to group them by common characteristics [1]. Genres are commonly used to organise large music collections both private and commercial, and the benefits of automating the classification process mean this topic has received much attention from the Music Information Retrieval (MIR) community in recent years.

Various approaches have been proposed, utilising features that describe different aspects of music such as pitch, timbre, rhythm and their combination [2]. Approaches using source separation [3] or models of auditory human perception [4] have also been proposed. However, one key aspect in music that has received little attention in the context of genre classification is the melody. Melodies in different genres can be expected to have different characteristics, especially in the case of sung melodies where it has been shown that the human voice is used in different ways depending on the musical genre [5]. Whilst there have been studies on genre classification using melody characteristics computed from symbolic data (MIDI files) [6], to the best of our knowledge there is no study on genre classification using high-level melodic characteristics extracted directly from the audio signal of polyphonic music.

In this paper we propose a method for genre classification based on melodic characteristics extracted from polyphonic music excerpts. To obtain the melody, we use a state-of-the-art melody extraction system which extracts and characterises pitch contours describing the predominant melodic line [7, 8]. The initial set of melody features is extended by computing characteristics derived from a musicological study of melody pitch contour. We use standard machine learning algorithms for the classification and compare our results to a baseline approach based on timbral features. An important aspect of our system is that the melodic features we use are what we consider high-level features. That is, unlike some of the features commonly used for genre classification, the features we use can be easily understood by humans. This means the classification results can be (in some cases) directly linked to aspects of the melody, for example that the vibrato rate applied in flamenco singing is on average lower than that applied in opera singing, or that the average pitch range used in vocal jazz is greater than that used in pop music.

In section 2 we describe the proposed method including the melodic features extracted and classification algorithms used. In section 3 we describe the data-sets used for evaluation and present the classification results obtained. Conclusions of the work are provided in section 4.

2. METHOD

2.1. Melody Extraction

To obtain a representation of the melody from polyphonic music, we use an automatic melody extraction system [7, 8]. Given the audio signal the system computes a salience function describing pitch salience over time as detailed in [9]. Peaks of the function are grouped over time using auditory streaming cues into *pitch contours*: time and frequency continuous sequences of salient pitches. A set of contour characteristics is computed for each contour and used to filter out non-melodic contours. The remaining contours are used to output the final melody as a sequence of fundamental frequencies (F0s) by selecting at each frame the pitch belonging to the most salient contour present.

For genre classification, rather than use the final sequence of F0s we use the set of pitch contours from which the sequence was selected. A pitch contour $c(n)$ is described as a discrete series of $n = 1 \ldots N$ frequency values (in cents). The time difference between each value (determined by the hop size H used in the melody extraction) is 2.9ms ($H = 128$ with sampling rate $f_S = 44100$Hz), and the frequency resolution is 10 cents. Note that the contours are

This research was funded by the Program de Formación del Profesorado Universitario of the Ministerio de Educación de España, COFLA (P09-TIC-4840-JA) and DRIMS (TIN2009-14247-C02-01-MICINN).

not segmented into notes nor quantized into semitones. This means a pitch contour may span a single note in the shortest case or a short phrase in the longest. It also means the contours allow us to capture aspects of the pitch evolution that are important for genre characterisation such as vibrato. An example of contours extracted from excerpts of different genres is provided in Figure 1. Melody contours are highlighted in bold.

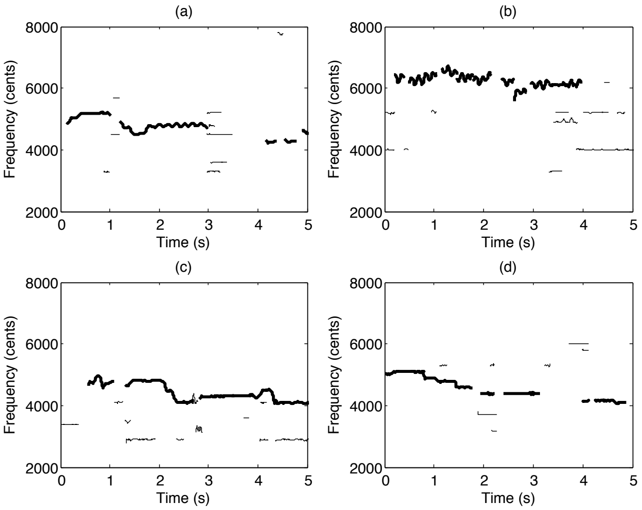

Fig. 1. Pitch contours extracted from excerpts of different genres: vocal jazz (a), opera (b), pop (c) and instrumental jazz (d).

2.2. Melody Features

For each contour, a set of melodic features is automatically computed. We divide the features into three categories, detailed in sections 2.2.1, 2.2.2 and 2.2.3. Then in section 2.2.4 we explain how the contour features are used to compute global per-excerpt features for use in the classification.

2.2.1. Pitch and duration features

The following features are related directly to contour pitch or length:

- **Duration** $t = N \cdot \frac{H}{f_S}$ (in seconds). (1)
- **Mean pitch height** $\mu_p = \frac{1}{N}\sum_{n=1}^{N} c(n)$. (2)
- **Pitch deviation** $\sigma_p = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(c(n)-\mu_p)^2}$. (3)
- **Pitch range** $r_p = \max(c(n)) - \min(c(n))$. (4)

2.2.2. Vibrato features

Vibrato is a periodic variation of pitch that is characterised by its rate and extent (depth) [10]. Apart from being a distinctive element of the singing voice, the way in which it is applied varies between different singing styles [5], and thus we expect features related to vibrato to be important for genre classification. As a first step the system detects whether a contour has vibrato or not. This is done by applying the STFT to the pitch contour $c(n)$ as in [11] and checking for a prominent peak in the magnitude spectrum $|C(k)|$ at the expected range for vibrato in human voice (5-8Hz). If vibrato is detected, the rate and extent can be computed from the peak's frequency and magnitude respectively. We use a frame size of 120 samples (350ms) to ensure we capture at least 2 cycles of the lowest period expected for vibrato, and a hop size of 1 sample.

In addition to these features, we wanted to capture the amount of vibrato applied throughout a contour. That is, the proportion of the contour in which vibrato is applied. We refer to this as vibrato coverage, and we expect it to vary between genres where vibrato is used a lot (e.g. opera) and genres where it might be applied just at the end of a phrase (e.g. vocal jazz). A summary of the vibrato features is:

- **Vibrato rate** v_r: frequency of prominent peak of $|C(k)|$ in expected vibrato range (in Hz).
- **Vibrato extent** v_e: magnitude of said peak (in cents).
- **Vibrato coverage** v_c: ratio of samples with vibrato to total number of samples in the contour (value between 0-1).

2.2.3. Contour typology

In [12] Adams proposes to categorise melodic segments based on the "distinctive relationship among their minimal boundaries". By categorising the possible relationship between a segment's initial (I), final (F), highest (H) and lowest (L) pitch, 15 "contour types" are defined. An example of three different contour types is provided in Figure 2.

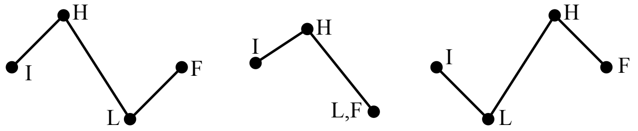

Fig. 2. Different types of melodic contour.

We adopt Adam's melodic contour typology and compute the type of each contour. Before the type is computed the contour pitch is quantized into a quarter-tone resolution, to avoid smaller pitch variations affecting the contour type. To summarise:

- **Contour type** ζ_i: one of 15 melodic contour types ($i = 1 \ldots 15$).

2.2.4. Global features

The contour features are used to compute global excerpt features which are used for the classification. For the pitch, duration and vibrato features we compute the mean, standard deviation, skewness and kurtosis of each feature over all contours. The contour typology is used to compute a type distribution describing the proportion of each contour type out of all the pitch contours forming the melody. In addition to these features several global features are added:

- **Global highest pitch** p_h: The highest pitch in the melody.
- **Global lowest pitch** p_l: The lowest pitch in the melody.
- **Global pitch range** $r_g = p_h - p_l$. (5)
- **Global vibrato presence**: the ratio between contours with vibrato to all contours in the melody (between 0-1).
- **Interval features**: we compute the interval between each pair of consecutive contours as the difference between their mean pitch height. We then compute the mean, standard deviation, skewness and kurtosis of all intervals in the melody.

This gives us a total of 51 features. Initial experiments revealed that some features resulted in better classification if they were computed using only the longer contours in the melody. This is probably because long contours are less likely to be an error of the melody extraction algorithm, and also there is a greater chance to detect vibrato features in longer contours. For this reason we computed for each feature (except for the interval features) a second value computed using only the top third of the melody contours when ordered by duration. This gives us a total of 98 features for the next stage.

2.3. Classification

To classify the excerpts we compare several classification algorithms from the Weka data mining software [13]. We start by performing attribute selection using the *CfsSubsetEval* attribute evaluator and *BestFirst* search method [14] with a 10-fold cross validation, only keeping features that were used in all folds. Each attribute is normalised feature-wise between 0 and 1. For each classification algorithm we use a 10-fold cross validation and repeat the experiment 10 times, reporting the average accuracy. The algorithms compared are Support Vector Machines (SMO; radial basis function kernel), Random Forest (RF), K-Nearest Neighbours (K*) and Bayesian Network (BNet).

3. EVALUATION

3.1. Data-sets

For evaluation we constructed a data-set of five musical genres in which the melody plays an important role: opera, pop, flamenco, vocal jazz and instrumental jazz (where the melody is played by a saxophone or trumpet rather than sung). For initial experiments the data-set consisted of fifty 30-second excerpts per genre (250 excerpts in total). The set was later expanded to include 100 excerpts per genre (500 excerpts in total). To cover variations within a genre the the excerpts for each genre were selected from a wide set of artists. All excerpts were taken from a section of the song where the melodic line is clearly present.

As a final experiment we evaluated our method on the GTZAN [1] collection, consisting of 10 genres with one hundred 30-second excerpts per genre (1000 excerpts in total). Note that in this collection some excerpts might not have a melody at all, and for some genres (e.g. metal) the melody extraction may not perform very well. Still, we wanted to see what could be achieved for this collection without any modification to the method or excerpts.

3.2. Baseline and combined feature sets

To compare our results we computed a baseline set of low-level timbral features which are commonly used in genre classification. For each excerpt we computed the first 20 Mel-frequency cepstral coefficients (MFCCs) as in [15], using a 23ms window size with 50% overlap, taking 40 mel-frequency bands up to 16kHz. We compute the mean and variance of each coefficient, resulting in a total of 40 descriptors. We also wanted to see whether results could be improved by combining low-level and high-level information. To do this we created a third feature set which combines our melodic features with the MFCC features, giving a total of 138 descriptors.

3.3. Results

We start by presenting the results for the initial 250 excerpt data-set. A total of 10 attributes were selected out of the initial 98 (a * indicates the feature was computed from long contours only): r_p:mean, μ_p:mean, v_r:mean*, v_r:skewness*, v_e:mean*, v_c:mean*, v_c:stddev*, ζ_9*, ζ_{10}*, ζ_{14}*. We see that most descriptors are computed from the longer contours of the melody. We also note a strong presence of vibrato related features. In Figure 3 we present the classification results comparing the melodic, MFCC and combined feature sets. The number of features selected for each set is indicated in brackets.

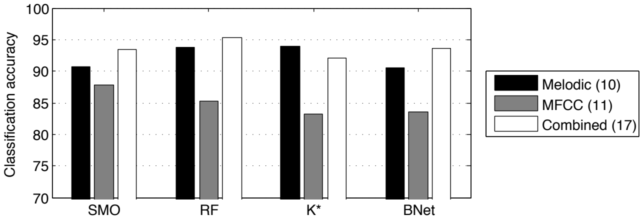

Fig. 3. Classification results for the initial 250 excerpt data-set.

We see that with all classifiers we obtain a classification accuracy of over 90% using the melodic features. In all cases the melodic feature set outperforms the baseline approach. Next, we note that for most classifiers we can increase the classification accuracy by combining the MFCC features with our high-level melodic features.

To see whether any descriptors were especially discriminative we also classified the data using a decision tree. It turned out that two important features are the mean vibrato coverage and mean vibrato rate. In Figure 4 we see that the genres can be fairly well separated using just these two descriptors. Furthermore, both descriptors are musically meaningful (the former expressing the degree to which vibrato is applied and the latter the average rate of the vibrato).

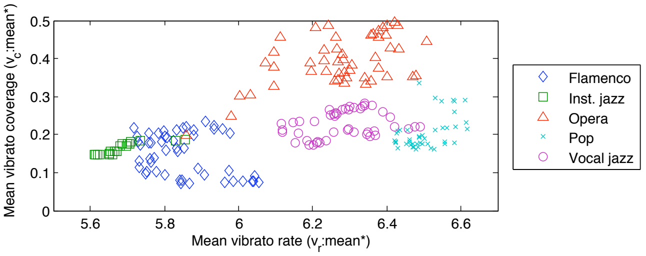

Fig. 4. Mean vibrato coverage vs mean vibrato rate.

Next we examine the results for the extended data-set (500 excerpts), provided in Figure 5. Note that this time only 7 descriptors were selected for the melodic feature set.

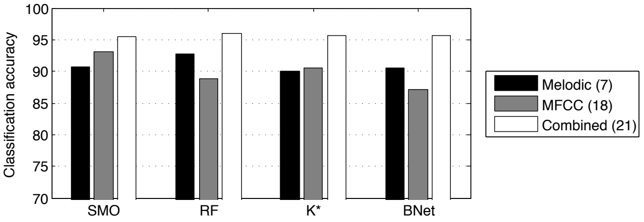

Fig. 5. Classification results for the extended 500 excerpt data-set.

We see that for all classifiers the melodic feature set maintains classification accuracies above 90%. We also note that for RF and

BNet the melodic set still outperforms the baseline approach even though it uses less than half the amount of descriptors. This time results for all classifiers are improved when combining the two different sets of descriptors. To ensure the results were not biased by the different size of each feature set, we ran two further experiments imposing a fixed number of descriptors for all three sets (21 and 10). In both cases the results were consistent with those of Figure 5, with the combined set outperforming the other two. Examining the confusion matrices of the classification results, we found that for the melodic feature set the confusion occurs primarily between pop and vocal jazz. This is understandable as these singing styles have common characteristics, making them hard to distinguish even for humans [5]. Combining the melodic features with the MFCC features reduces this confusion, leading to an overall increase in accuracy.

Finally, we examine the results obtained for the GTZAN collection, provided in Figure 6. As expected, the classification results are not as high as those obtained for the collections where we ensured that there is a melody in each excerpt. Still, with the SMO classifier and the combined feature set we obtain an accuracy of 82%, improving significantly on both the melodic and MFCC feature sets. Whilst this does not surpass the highest accuracy reported for this collection to date [4], the results provide an important proof of concept – that combining low-level features with high-level melodic features is a promising approach for improving genre classification.

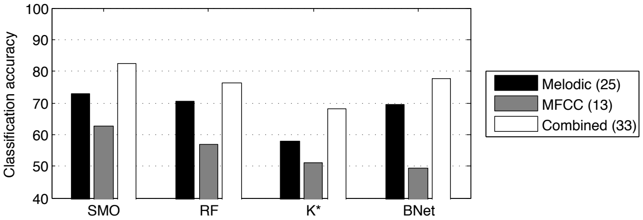

Fig. 6. Classification results for the GTZAN data-set.

4. CONCLUSIONS

We presented a genre classification method based on a novel set of melodic features. By using an automatic melody extraction system we were able to compute these features directly from the audio signal of polyphonic music, without the need to obtain the monophonic melody track beforehand. A set of melodic features was proposed, based on pitch, duration and vibrato characteristics, and on contour typology. The melodic feature set was evaluated on three different data-sets and was shown to outperform a baseline low-level timbral feature set based on MFCCs. Most importantly, we demonstrated that the classification accuracy can be improved by combining the two feature sets. This suggests that adding high-level melodic features to traditional low-level feature sets is a promising approach for genre classification. It is worth noting that the current performance of state-of-the-art melody extraction systems, including the one used in this paper, is around 75%[1]. The positive results obtained in this study demonstrate that an automatically extracted mid-level representation of the melody, though not 100% accurate, can still be used successfully to address related MIR challenges. Finally, another important aspect of the approach presented in this paper is the fact that most of the melodic features proposed can be easily understood by humans. This means that the classification results can be interpreted more easily, allowing us to make straight forward links between musical genres and melodic characteristics.

5. REFERENCES

[1] G. Tzanetakis and P. Cook, "Automatic musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, July 2002.

[2] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, March 2006.

[3] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagayama, "Autoregressive mfcc models for genre classification improved by harmonic-percussion separation," in *Proc. of the 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, Aug. 2010, pp. 87–92.

[4] Y. Panagakis, C. Kotropoulos, and G. Arce, "Music genre classification using locality preserving non-negative tensor factorization and sparse representations," in *Proc. of the 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, 2009, pp. 249–254.

[5] J. Sundberg and M. Thalén, "Describing different styles of singing: A comparison of a female singer's voice source in 'classical', 'pop', 'jazz' and 'blues'," *Logopedics, Phoniatrics Vocology*, vol. 26, no. 2, pp. 82–93, 2001.

[6] C. McKay and I. Fujinaga, "Automatic genre classification using large high-level musical feature sets," in *Proc. 5th International Conference on Music Information Retrieval*, Barcelona, Spain, October 2004.

[7] J. Salamon and E. Gómez, "Melody extraction from polyphonic music: Mirex 2011," in *5th Music Information Retrieval Evaluation eXchange (MIREX)*, extended abstract, Miami, USA, October 2011.

[8] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech and Language Processing*, In Press.

[9] J. Salamon, E. Gómez, and J. Bonada, "Sinusoid extraction and salience function design for predominant melody estimation," in *Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11)*, Paris, France, September 2011, pp. 73–80.

[10] J. Sundberg, "Acoustic and psychoacoustic aspects of vocal vibrato," *Vibrato*, pp. 35–62, 1995.

[11] P. Herrera and J. Bonada, "Vibrato extraction and parameterization in the spectral modeling synthesis framework," in *Proc. of the Digital Audio Effects Workshop (DAFx-98)*, 1998.

[12] C. Adams, "Melodic contour typology," *Ethnomusicology*, vol. 20, pp. 179–215, 1976.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, November 2009.

[14] M. Hall, *Correlation-based Feature Selection for Machine Learning*, Ph.D. thesis, University of Waikato, Hamilton, New Zealand, 1999.

[15] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classificaton," in *Proc. of the 6th International Society for Music Information Retrieval Conference*, London, UK, 2005, pp. 628–633.

[1] Music Information Retrieval Evaluation eXchange [Online]. Available: http://www.music-ir.org/mirex/wiki/Audio_Melody_Extraction (Jan. 2012).