STATISTICAL CHARACTERISATION OF MELODIC PITCH CONTOURS AND ITS APPLICATION FOR MELODY EXTRACTION

Justin Salamon
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
justin.salamon@upf.edu

Geoffroy Peeters, Axel Röbel
Sound Analysis-Synthesis Team
IRCAM - CNRS STMS, 75004 Paris, France
{geoffroy.peeters,axel.roebel}@ircam.fr

ABSTRACT

In this paper we present a method for the statistical characterisation of melodic pitch contours, and apply it to automatic melody extraction from polyphonic music signals. Within the context of melody extraction, pitch contours represent time and frequency continuous sequences of pitch candidates out of which the melody must be selected. In previous studies we presented a melody extraction algorithm in which contour features are used in a heuristic manner to filter out non-melodic contours. In our current work, we present a method for the statistical modelling of these features, and propose an algorithm for melody extraction based on the obtained model. The algorithm exploits the learned model to compute a "melodiness" index for each pitch contour, which is then used to select the melody out of all pitch contours generated for an excerpt of polyphonic music. The proposed approach has the advantage that new contour features can be easily incorporated into the model without the need to manually devise rules to address each feature individually. The method is evaluated in the context of melody extraction and obtains promising results, performing comparably to a state-of-the-art heuristic-based algorithm.

1. INTRODUCTION

Melody extraction algorithms can be divided into several categories, based on their underlying approach. Some systems extract the melody by first separating it from the rest of the audio signal using source separation techniques [6, 11]. Purely data-driven approaches have also been proposed, such as [14] in which the entire short-time magnitude spectrum is used as training data for a support vector machine classifier. Still, the largest set of methods to date are those that can be referred to as *salience-based* algorithms, which derive an estimation of pitch salience over time and then apply tracking or transition rules to extract the melody line without separating it from the rest of the audio [4, 8, 12, 16, 18]. A review of salience-based systems can be found in [15].

For salience-based methods, one of the most important steps is the tracking and selection of pitch candidates. That is, given a set of pitch candidates at each frame, the system must decide which candidate belongs to the melody. This is also one of the steps that varies most between systems: in [8], tracking agents compete for candidates on a per frame basis using a set of heuristics, and the most salient agent at the end of the tracking is selected as the final melody. Tracking agents are also used in [5], where pitch candidates are first grouped into tone objects which are added to the agents using rules based on auditory streaming. In [16] Hidden Markov Models (HMM) are used to model the pitch evolution of single notes, and then the models are combined into a single HMM with inter-note transition probabilities learned from a training data-set.

In [18], we proposed a method for melody extraction based on pitch contour characteristics. In our approach, pitch candidates are first grouped over time into *pitch contours* – time and frequency continuous sequences of pitch candidates, whose length may vary from a single note in the shortest case to a short phrase in the longest. Given all the pitch contours generated from the audio signal of a polyphonic piece of music, we compute a set of contour characteristics, or features, related to contour salience, length, height and pitch evolution (namely pitch deviation and the presence of vibrato). These contour features are then used to devise filtering rules for filtering out non-melodic contours. Given the final set of contours after filtering, the melody is selected as the pitch candidate belonging to the most salient contour present in each frame. In the most recent Music Information Retrieval Evaluation eXchange (MIREX 2011) [3], the algorithm was shown to outperform all alternative approaches, obtaining the highest mean overall accuracy achieved by a melody extraction algorithm for the current MIREX data-sets [17].

Similar to other extraction algorithms such as [5,8], one characteristic of our approach is that it heavily relies on heuristics for the candidate selection stage. Whilst this in itself is not a problem (some of the most successful algorithms also rely on heuristics [5]), it has the disadvantage that new heuristics must be devised whenever we want to incorporate new musical information into our algorithm (i.e. new contour features). This motivates us to explore the possibility of exploiting contour features in an automated manner, that is, creating a model based on contour features that can be easily updated whenever we want to incorporate new features.

In this paper, we present a method for the statistical characterisation of pitch contours using contour feature distributions. We do this by combining the distributions of different contour features into a single multivariate Gaussian distribution which embodies most of the features currently used by the algorithm. By learning separate feature distributions for melodic and non-melodic contours, we are able to create two different multivariate distributions, one for computing the likelihood that a contour is melodic, and the other for computing the likelihood that it's not melodic (i.e. accompaniment). The likelihoods are used to compute a single "melodiness" index, which is then used to select the final melody sequence. As can be inferred from the above description, the proposed method is flexible in that new features can be easily incorporated into the model without the need to manually devise rules to address them.

The structure of the remainder of the paper is as follows. In Section 2 we describe the proposed approach, including the creation of pitch contours, computation of contour features and their distributions, and the statistical modelling of these distributions. In Section 3, we describe the evaluation of the proposed approach, including evaluation material, measures and results. Finally, in Section 4 we conclude the paper with discussion of the results and some propositions for future work.

2. METHOD

In this section we describe the steps performed to obtain our contour feature model. These include the creation of pitch contours, computation of contour features and feature distributions, and finally the modelling of these distributions as a multivariate normal distribution.

2.1 Creating pitch contours

A summary of the contour creation process is provided here. For further details, the reader is referred to [18]. A block diagram of the process is provided in Figure 1.

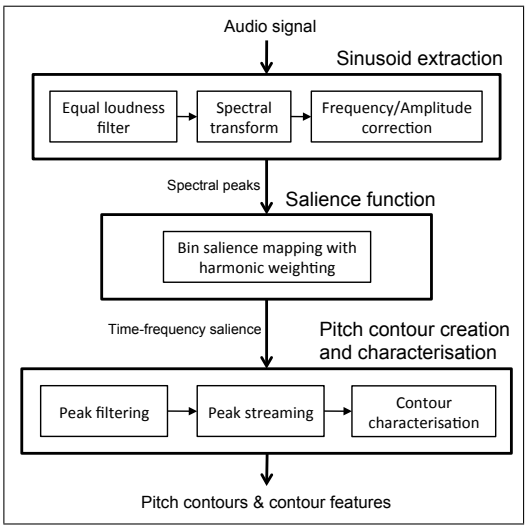

Figure 1. Block diagram of the steps involved in the creation of pitch contours.

In the first stage, sinusoids (spectral peaks) are extracted from the signal at each frame. We start by applying an equal loudness filter which attenuates frequencies where the melody is usually not present [19]. Next we compute the Short-Time Fourier Transform, and take the peaks of the spectrum at each frame. Peak frequencies and amplitudes are re-estimated by computing each peak's instantaneous frequency using the phase vocoder method [7]. In the next stage, the spectral peaks are used to create a salience function based on weighted harmonic summation [19]. The salience function is quantised into 600 bins covering a range of nearly five octaves from 55Hz to 1760Hz. The peaks of the salience function at each frame are considered as pitch candidates for the melody. In the next stage, the pitch candidates are grouped over time and frequency using rules based on auditory streaming [2] to create pitch contours. In Figure 2 we provide examples of contours generated from excerpts of different musical styles. Contours belonging to the melody are highlighted in bold.

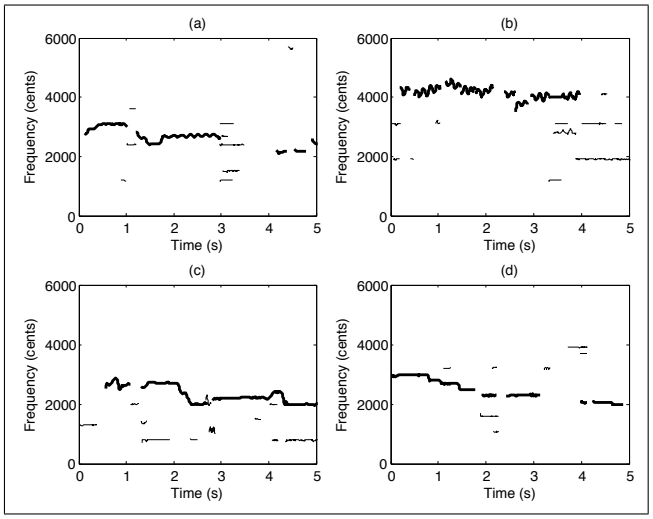

Figure 2. Pitch contours generated from excerpts of (a) vocal jazz, (b) opera, (c) pop and (d) instrumental jazz.

2.2 Contour features and distributions

Once the contours are created, we compute a set of contour characteristics, or features, for each contour. Similarly to some melody extraction systems, we define features based on contour pitch, length and salience [12]. However, by avoiding the quantisation of contours into notes we are able to extend this set by introducing features extracted from the pitch trajectory of the contour, namely its pitch deviation and the presence of vibrato. Every pitch contour is represented by two discrete series $c(n)$ and $s(n)$, $n = 1 \dots N$. The former contains the frequency values (in cents) of every pitch candidate in the contour, and the latter its corresponding salience value. Thus, for every pitch contour we compute the following characteristics:

- **Pitch mean** $C_{\bar{p}} = \frac{1}{N} \sum_{n=1}^{N} c(n)$.
- **Pitch deviation** $C_{\sigma_p} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (c(n) - C_{\bar{p}})^2}$.

- **Total salience** $C_{\Sigma s} = \sum_{n=1}^{N} s(n)$.
- **Mean salience** $C_{\bar{s}} = \frac{1}{N} C_{\Sigma s}$.
- **Salience deviation** $C_{\sigma_s} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (s(n) - C_{\bar{s}})^2}$.
- **Length** $C_l = N \cdot \frac{H}{f_S}$ (in seconds, where H and f_S are the hop size (128) and sampling frequency (44100) used by algorithm respectively).
- **Vibrato presence** C_v**:** whether the contour has vibrato or not (true/false). Vibrato is automatically detected by the system using a method based on [9].

In [18] these features were used to filter out non-melody pitch contours. To do this, we computed the distribution of each feature [1] for melody and non-melody contours using a representative data-set (c.f. Section 3.1), reproduced in Figure 3. Each plot includes the feature distribution for melody contours (solid red line) and non-melody contours (dashed blue line). In plots (c), (d) and (e) the feature values are normalised by the mean feature value for each excerpt. Observing these graphs we see how melodic contour characteristics differ from non-melodic contours: a mid-frequency pitch range, greater pitch variance, greater salience (both mean and total) and salience variance, and longer contours. These observations concur with voice leading rules derived from perceptual principles [10]. Note that in most (but not all) of the excerpts in this data-set the melody is sung by a human voice. Additionally, for vibrato presence we found that 95% of all contours in which vibrato was detected were melody contours.

In [18], these observations were exploited by devising a set of heuristic filtering rules to remove non-melodic contours. As mentioned in the introduction, whilst this approach was shown to be very successful for filtering non-melodic contours, in our current work we raise the question of whether the contour feature distributions can be exploited in a more general way using statistical modelling.

2.3 Statistical Modelling

Our goal is to define a statistical model that encompasses all of the contour feature distributions provided in Figure 3. To do so, we represent all feature distributions as two multivariate normal distributions, one for melodic contour features and one for non-melodic contour features. In [13] a multivariate Gaussian was shown to obtain comparable classification performance to GMMs when the amount of training data is relatively small.

As seen in the plots, though some distributions (in particular the distribution of pitch height for melodic contours) appear normal, this is not the case for all distributions. Thus, in the first step of the modelling we apply a power transform to obtain a normal-like distribution for each contour feature. Specifically, we apply the Box-Cox transform [1], which for a variable Y with data samples $y_i > 0$ is defined as:

[1] With the exception of vibrato presence which is a binary value (true/false).

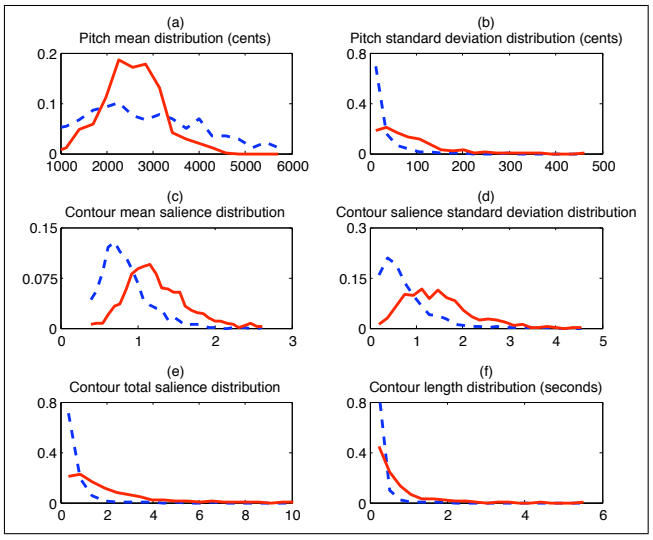

Figure 3. Pitch contour feature distributions (relative frequency vs. feature value): (a) Pitch mean, (b) Pitch std. dev., (c) Mean salience, (d) Salience std. dev., (e) Total salience, (f) Length. The red solid line represents the distribution of melody contour features, the blue dashed line represents the distribution of non-melody contour features.

$$y_i^{(\lambda)} = \begin{cases} \frac{(y_i^{\lambda} - 1)}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y_i), & \text{if } \lambda = 0, \end{cases} \tag{1}$$

where the power parameter λ is selected such that it maximises the log-likelihood of λ given the transformed data, which is assumed to be normally distributed. An example of the distributions for the contour total salience feature $C_{\Sigma s}$ before an after transformation is provided in Figure 4. In plots (a) and (b) we show the feature distribution for melodic contours before and after transformation respectively, and in plots (c) and (d) we plot the corresponding distributions for non-melodic contours. In plots (b) and (d) we also display the normal distribution that best fits the transformed data.

The mean vectors μ and covariance matrices Σ (of size $N \times N$ where N is the number of features used) of the transformed distributions are obtained using the standard maximum likelihood estimators, allowing us to construct a multivariate normal distribution with parameters $\theta = (\mu, \Sigma)$ of the form:

$$f_\theta(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right). \tag{2}$$

This procedure is repeated twice, once for the melodic contour feature distributions, and once for the non-melodic (i.e. background) contour feature distributions, resulting in two multivariate normal distributions which we denote $f_{\theta m}$ and $f_{\theta bg}$ respectively. Given the feature vector $\mathbf{x}$ of a pitch contour, we can now use $f_{\theta m}$ and $f_{\theta bg}$ to compute the likelihood of the contour being a melodic contour and the likelihood of it being a non-melodic contour (equations 3 and 4 respectively):

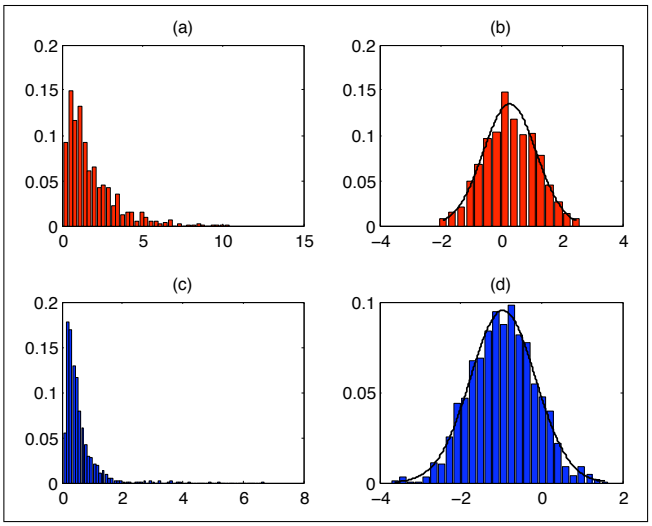

Figure 4. Contour total salience distributions. For melodic contours: (a) raw data, (b) after Box-Cox. For non-melodic contours: (c) raw data, (d) after Box-Cox.

$$\mathcal{L}(\theta_m|\mathbf{x}) = f_{\theta m}(\mathbf{x}) \tag{3}$$

$$\mathcal{L}(\theta_{bg}|\mathbf{x}) = f_{\theta bg}(\mathbf{x}) \tag{4}$$

Given the two likelihoods, we define the "melodiness" index $\mathcal{M}(\mathbf{x})$ of a pitch contour with feature vector $\mathbf{x}$ as the likelihood ratio of the melodic and non-melodic likelihoods:

$$\mathcal{M}(\mathbf{x}) = \frac{\mathcal{L}(\theta_m|\mathbf{x})}{\mathcal{L}(\theta_{bg}|\mathbf{x})} \tag{5}$$

3. EVALUATION

We evaluate the proposed contour characterisation approach in the context of melody extraction. To do so, we define a straight-forward rule for melody selection based on our proposed melodiness index $\mathcal{M}(\mathbf{x})$. Given the contours generated for a musical excerpt, at each frame we check to see which contours are present in the frame and select as melody the pitch candidate belonging to the contour whose features $\mathbf{x}$ result in the highest melodiness index $\mathcal{M}(\mathbf{x})$. The resulting melody sequence is evaluated using the standard measures employed for melody extraction evaluation. In the following sections we describe the music collection and measures used for evaluation, and compare the results obtained using the proposed method to those obtained by a state-of-the-art melody extraction algorithm.

3.1 Music Collection

The collection used for evaluation is comprised of musical excerpts with per-frame annotations of the melody F0 which are freely available for research purposes (cf. collection 3 in [18]). Our collection includes 65 audio excerpts from a variety of musical genres including rock, pop, R&B, jazz and opera singing. Excerpt durations range from 5 to 35 seconds. For each excerpt, the annotation is comprised of two columns, one containing the time-stamp for the frame, and the other containing the F0 of the melody in that frame. If there is no melody present in the frame (i.e. the frame is 'unvoiced'), a value of 0Hz is placed in the annotation.

3.2 Evaluation Measures

The extracted melody sequence is evaluated using the standard measures employed for melody extraction evaluation in the MIREX campaign. The measures are designed to evaluate the performance of the algorithm in several aspects: voicing detection (determining when the melody is present and when it is not), pitch accuracy (estimating the correct F0 when the melody is present), and overall accuracy (the combination of voicing and pitch accuracy). It should be noted that our proposed approach, as presented above, does not include a method for voicing detection. That is, at each frame a non-zero frequency value is returned by choosing a pitch candidate from one of the contours present in the frame. The only exception are frames in which no contours are present, in which case the algorithm outputs 0Hz. For this reason, in the first part of the evaluation the proposed approach is evaluated only in terms of its pitch accuracy. In the second stage of the evaluation, we combine the proposed approach with the voicing detection method proposed in [18], and evaluate the results obtained using this combined approach. When voicing detection is included, the algorithm indicates whether a frame is voiced or unvoiced by returning either a positive or negative frequency value respectively (e.g. 300Hz or -300Hz). The negative values represent the pitch estimate of the algorithm for frames it has detected as unvoiced. When evaluating the algorithm's pitch accuracy the sign is ignored, meaning incorrect voicing detection will not affect the pitch (and chroma) accuracy. The overall accuracy (see below) serves as a global measure which considers both pitch and voicing detection accuracy. A summary of the evaluation measures, which are detailed in [15], is provided in Table 1.

Voicing Recall Rate: the proportion of frames labeled as voiced in the ground truth that are estimated as voiced by the algorithm.
Voicing False Alarm Rate: the proportion of unvoiced frames in the ground truth that are estimated as voiced by the algorithm.
Raw Pitch Accuracy: the proportion of voiced frames in the ground truth for which the F0 estimated by the algorithm is within $\pm\frac{1}{4}$ tone (50 cents) of the ground truth annotation.
Raw Chroma Accuracy: same as the raw pitch accuracy, except that both the estimated and ground truth F0 sequences are mapped into a single octave, in this way ignoring octave errors in the estimation.
Overall Accuracy: combines the performance of the pitch estimation and voicing detection to give an overall performance score. Defined as the proportion of frames (out of the entire piece) correctly estimated by the algorithm, where for non-voiced frames this means the algorithm labeled them as non-voiced, and for voiced frames it means the algorithm both labeled them as voiced and provided a correct F0 estimate for the melody (i.e. within $\pm\frac{1}{4}$ tone of the ground truth).

Table 1. Evaluation measures for melody extraction.

3.3 Results

To avoid any bias in the results, we separate the training and evaluation material by conducting a 3-fold cross validation, and report the results averaged across all folds. In Table 2 we provide the results obtained by our proposed approach (without any voicing detection method). For completeness we calculate all evaluation measures, though as explained above, since we do not attempt to perform any voicing detection, only the raw pitch and raw chroma measures (highlighted in bold) should be taken into consideration at this point. For comparison, we include the results obtained by the algorithm presented in [18] (which includes voicing detection), which obtained the highest mean overall accuracy results in MIREX 2011 (denoted SG) [17].

Alg.	Voicing Recall	Voicing False Alarm	**Raw Pitch**	**Raw Chroma**	Overall Accuracy
Prop.	0.95	0.60	**0.77**	**0.83**	0.65
SG	0.86	0.19	**0.81**	**0.83**	0.77

Table 2. Results obtained using the proposed method without voicing detection, compared to those obtained by SG.

We see that the proposed approach obtains the same chroma accuracy as the state-of-the-art algorithm. The lower raw pitch accuracy indicates that the proposed approach makes slightly more octave errors. Nonetheless, the results are definitely promising, and their comparability to SG suggests that the proposed approach is also comparable with other state-of-the-art melody extraction algorithms evaluated in MIREX [2] .

In the second stage of the evaluation, we combine the proposed approach with the voicing detection method proposed in [18]. The approach is based on filtering out contours by setting a salience threshold determined from the distribution of contour mean salience $C_{\bar{s}}$ in a given excerpt. The reader is referred to the article cited above for further details. Thus, the combined approach consists of first filtering out non-voiced contours using the voicing filter, and then selecting the melody out of the remaining contours based on their melodiness index $\mathcal{M}(\mathbf{x})$ as before. The F0 estimate for non-voiced frames (recall that algorithms can return F0 estimates for non-voiced frames so that pitch and voicing accuracies can be evaluated independently) is produced by selecting the pitch candidate belonging to the non-voiced contour (i.e. a contour that was removed by the voicing filter) with the highest $\mathcal{M}(\mathbf{x})$ out of all non-voiced contours present in the frame. The results are presented in Table 3, once again alongside the results obtained by SG for comparison. This time we focus on the voicing evaluation measures and the overall accuracy.

As expected, by combining our proposed approach with a voicing detection method we are able to considerably reduce the voicing false-alarm rate (from 60% to 25%) whilst maintaining a relatively high voicing recall rate (87%). As a result, the overall accuracy of the proposed approach goes up from 65% without voicing detection to 74% with voicing detection. Though the same voicing detection approach is applied in both cases, we note the voicing false alarm is not the same. This is because some steps in SG, though not designed to address voicing detection, have been shown to have a positive effect on it [18]. Whilst the combined approach does not outperform SG (no other system has, to date), the results serve as a promising proof-of-concept, with an overall accuracy which is comparable to other state-of-the-art melody extraction algorithms.

[2] Music Information Retrieval Evaluation eXchange [Online]: http://www.music-ir.org/mirex/wiki/Audio_Melody_Extraction (Apr. 12).

Alg.	**Voicing Recall**	**Voicing False Alarm**	Raw Pitch	Raw Chroma	**Overall Accuracy**
Prop.	**0.87**	**0.25**	0.78	0.83	**0.74**
SG	**0.86**	**0.19**	0.81	0.83	**0.77**

Table 3. Results obtained using the proposed method with voicing detection, compared to those obtained by SG.

As a final step, we inspect the values of our melodiness index $\mathcal{M}(x)$ for melody and non-melody contours. In Figure 5, we plot the values of $\mathcal{M}(x)$ for all pitch contours of all excerpts (on a log scale). For each excerpt we first normalise all $\mathcal{M}(x)$ values by the highest value in the excerpt, so that we can plot all values from all excerpts in a single graph. Values for melody contours are represented by a red circle, and values for non-melody contours by a blue x.

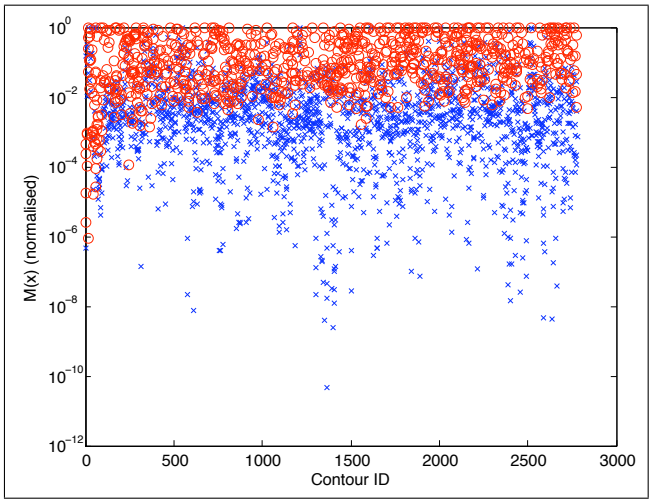

Figure 5. Normalised $\mathcal{M}(x)$ values for melody contours (red circle) and non-melody contours (blue x).

We see that the $\mathcal{M}(x)$ values for the two classes are fairly distinguishable, with the vast majority of melody contours having higher $\mathcal{M}(x)$ values than non-melody contours. This, apart from suggesting that $\mathcal{M}(x)$ is indeed a good indicator for melody contours, means we might be able to refine our melody selection algorithm by studying the distributions of $\mathcal{M}(x)$ for melody and non-melody contours. We intend to explore this possibility in future work.

4. CONCLUSION

In this paper we presented an approach for the statistical characterisation of melodic pitch contours. We explained how pitch contours can be generated from an audio excerpt

and how to calculate contour features. We then showed how these features can be used to build a model to describe melodic and non-melodic contours, leading to the computation of a melodiness index $\mathcal{M}(\mathbf{x})$. The proposed approach was evaluated in the context of melody extraction by using the melodiness index to select the melody out of the generated contours. The results of the evaluation showed that the approach achieves pitch and chroma accuracies comparable to a state-of-the-art melody extraction algorithm. By combining the proposed approach with a voicing detection method, we were able to obtained satisfying overall accuracy values as well.

When considering the caveats of the proposed approach compared to the state-of-the-art algorithm (SG), one clear difference is that whilst in SG temporal information is also taken into account, in the proposed approach the melody selection at each frame is performed using the melodiness index $\mathcal{M}(\mathbf{x})$ only, and no temporal continuity is taken into account. This means the pitch trajectory of the melody is allowed to contain large jumps which are not common in melodies, which tend to have a relatively smooth pitch trajectory in accordance with voice leading principles [10]. Thus, a possible direction for improving the performance of the proposed approach is to combine the melodiness index with some type of temporal evolution constraint. For instance, we could use the melodiness index in combination with one of the tracking techniques mentioned in the introduction of the paper, such as HMMs [16] or tracking agents [5, 8]. Another possibility for improvement is to consider more contour features. For instance, earlier in the paper it was explained that in 95% of the cases where the system detected vibrato in a contour, that contour belonged to the main melody. This information is not exploited in the current model (with the exception of the voicing detection method). An additional important research direction would be the gathering of more data to enable the use of more sophisticated statistical models (e.g. GMMs). Finally, another interesting research direction would be to learn genre specific feature distributions, and depending on the genre of the excerpt use a different model to compute $\mathcal{M}(\mathbf{x})$. This could be done by creating a two stage classification/melody extraction system, where the contour features could also be used for the classification stage as in [20].

5. ACKNOWLEDGMENTS

This work was supported by the Programa de Formación del Profesorado Universitario (FPU) of the Ministerio de Educación de España and the Quaero Program funded by Oseo French State agency for innovation.

6. REFERENCES

[1] G. E. P. Box and D. R. Cox. An analysis of transformations. *J. of the Royal Statistical Soc. Series B (Methodological)*, 26(2):211–252, 1964.

[2] A. Bregman. *Auditory scene analysis*. MIT Press, Cambridge, Massachussetts, 1990.

[3] J. S. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[4] K. Dressler. Audio melody extraction for mirex 2009. In *5th Music Inform. Retrieval Evaluation eXchange (MIREX)*, 2009.

[5] K. Dressler. An auditory streamin approach for melody extraction from polyphonic music. In *12th Int. Soc. for Music Inform. Retrieval Conference*, pages 19–24, Miami, USA, Oct. 2011.

[6] J.-L. Durrieu. *Automatic Transcription and Separation of the Main Melody in Polyphonic Music Signals*. PhD thesis, Télécom ParisTech, 2010.

[7] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell Systems Technical J.*, 45:1493–1509, 1966.

[8] M. Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43:311–329, 2004.

[9] P. Herrera and J. Bonada. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proc. Workshop on Digital Audio Effects (DAFx-98)*, pages 107–110, 1998.

[10] D. Huron. Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, 19(1):1–64, 2001.

[11] 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 Trans. on Audio, Speech, and Language Process.*, 15(5):1564–1578, Jul. 2007.

[12] R. P. Paiva, T. Mendes, and A. Cardoso. Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness. *Comput. Music J.*, 30:80–98, Dec. 2006.

[13] G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Audio Engineering Society Convention 115*, Oct. 2003.

[14] G. Poliner and D. Ellis. A classification approach to melody transcription. In *Proc. 6th Int. Conf. on Music Inform. Retrieval*, pages 161–166, London, Sep. 2005.

[15] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. Gómez, S. Steich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Trans. on Audio, Speech and Language Process.*, 15(4):1247–1256, 2007.

[16] M. Ryynänen and A. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Comput. Music J.*, 32(3):72–86, 2008.

[17] J. Salamon and E. Gómez. Melody extraction from polyphonic music: Mirex 2011. In *5th Music Inform. Retrieval Evaluation eXchange (MIREX)*, Miami, USA, Oct. 2011.

[18] J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, Aug. 2012.

[19] 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)*, pages 73–80, Paris, France, Sep. 2011.

[20] J. Salamon, B. Rocha, and E. Gómez. Musical genre classification using melody features extracted from polyphonic music signals. In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012.