Transcription of vocal melodies using voice characteristics and algorithm fusion

Christopher Sutton, Emmanuel Vincent, Mark D. Plumbley
Centre for Digital Music
Queen Mary, University of London
Mile End Road, London E1 4NS, UK
christopher.sutton@cantab.net

Juan P. Bello
The Steinhardt School
New York University
35 W 4th Street, New York NY 10012, USA
jpbello@nyu.edu

Abstract

This paper deals with the transcription of vocal melodies in music recordings. The proposed system relies on two distinct pitch estimators which exploit characteristics of the human singing voice. A Hidden Markov Model (HMM) is used to fuse the pitch estimates and make voicing decisions. The resulting performance is evaluated on the MIREX 2006 Audio Melody Extraction data.

Keywords: Melody, singing voice, algorithm fusion.

1. Introduction

A key goal of digital music research is the automatic transcription of polyphonic music recordings. Systems seeking to perform full transcription have met with limited success so far. Higher transcription accuracy has been obtained by systems seeking to perform only a partial transcription consisting of the chord sequence, the drum track or the melody.

The melody of a piece of music is generally defined as the sequence of notes played by the lead instrument, but this leaves considerable ambiguity since the factors determining which instrument is the "lead" to a human listener are somewhat subjective and ill-defined. The fact that the raw pitch accuracy scores reported in the MIREX 2005 Audio Melody Extraction evaluation were considerably lower than for monophonic recordings suggests that the systems entered struggled to consistently identify the lead instrument.

In this paper, we aim to avoid this ambiguity by focusing on the case where melody is carried by the main vocal line, which is better defined objectively. Unlike standard transcription systems based on a single pitch estimator, the proposed system relies on two distinct pitch estimators which exploit characteristics of the human singing voice. A HMM is used to produce the final transcription by fusing the pitch estimates and making voicing decisions.

Useful voice characteristics are described in Section 2, followed in Section 3 by details of the system's design. The resulting performance is evaluated in Section 4 and conclusions are given in Section 5.

2. Characteristics of the human singing voice

To avoid transcribing non-vocal instruments, the proposed system exploits two salient characteristics of singing voice: *pitch instability* and *high-frequency dominance*.

2.1. Pitch instability

Pitch instability refers to the property of the singing voice that its pitch varies considerably over time compared with other pitched instruments. This is mostly due to the fact that *vibrato* typically exhibits an extent of $\pm$60–200 cents for singing voice and only $\pm$20–35 cents for other instruments [1]. Also, vocalists almost always sing *legato*, changing pitch smoothly during note attacks and transitions.

This characteristic has been exploited recently by a vocal detection system [2]. After identification of the musical key, the system filters the input audio by an inverse comb filter which attenuates all the harmonic partials of the seven notes in the key. Since vocal notes are rarely at exactly the intended pitch, their partials survive this process while other pitched instruments are attenuated.

2.2. High-frequency dominance

High-frequency dominance refers to the property of the singing voice that the power of its upper partials is larger than with other instruments. This has been observed in a study on vocal melody transcription [3], where the high frequency (over 800Hz) channels of a correlogram led to more accurate vocal pitch estimates than the low frequency channels.

We further investigated this effect in [4]. Figure 1 shows the minimum, mean and maximum reliability of correlogram channels for the estimation of vocal pitch over a range of recordings, where reliability is defined as the proportion of resulting pitch estimates within 50 cents of the ground truth. The recordings used were the nine training files for the MIREX 2005 Audio Melody Extraction evaluation featuring singing voice as lead instrument. The figure demonstrates that channels in the 3–15kHz range provide more reliable vocal pitch estimates than other channels.

3. Proposed System

Experimentally, the voice characteristics described above are difficult to combine into a single standard pitch estimator. Therefore we adopt a novel approach for melody transcription, in which multiple transcriptions produced by parallel estimators are fused into a single transcription, hope-

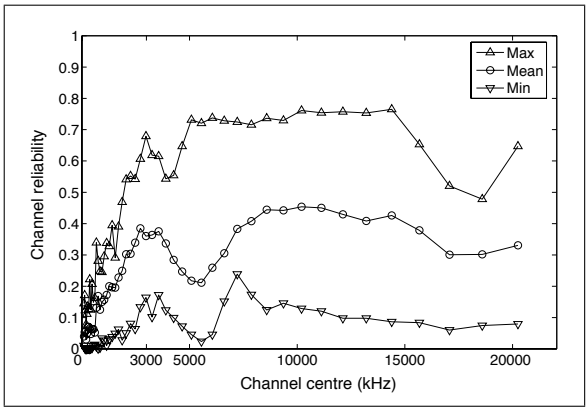

Figure 1. Reliability of correlogram frequency channels

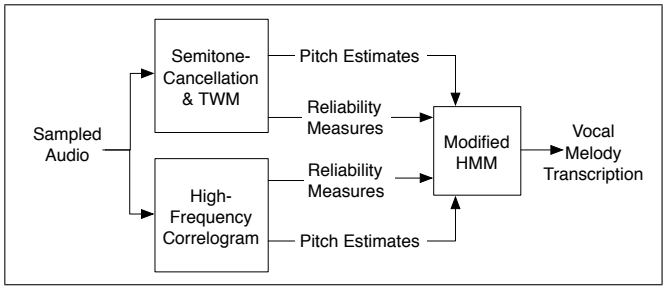

Figure 2. Diagram of the proposed system

fully more accurate than using any one of the estimators. In the following, two pitch estimators are used, but the system design and the fusion method generalise to a larger number of estimators.

The system diagram is shown in Figure 2. The input audio is processed by two pitch estimators, each producing a series of pitch estimates and associated *reliability measures* at 10 ms intervals. These values are then input to a HMM system to produce a single series of pitch estimates, with unvoiced segments represented by 0 Hz estimates.

3.1. Semitone-cancellation & TWM

The first vocal pitch estimator consists of a pre-processing stage in which a semitone-cancellation procedure emphasises the vocals, followed by the standard Two-Way Mismatch (TWM) [5] monophonic pitch transcription algorithm.

3.1.1. Semitone-cancellation procedure

Experimentally, we found that the non-vocal cancellation procedure proposed in [2] was too destructive of vocal pitch and did not allow accurate pitch estimation. Thus, instead of eliminating all the harmonic partials of interfering notes, we eliminate fundamental frequencies only. Since most music contains notes not in the musical key, the key detection stage is discarded and all semitone notes are eliminated. Based on the relative *vibrato* extent of vocals and other instruments (see Section 2.1), the bandwidth of the cancellation filters is set to ± 20 cents. This process is implemented in the frequency domain by zeroing suitable FFT bins [4]. Since most

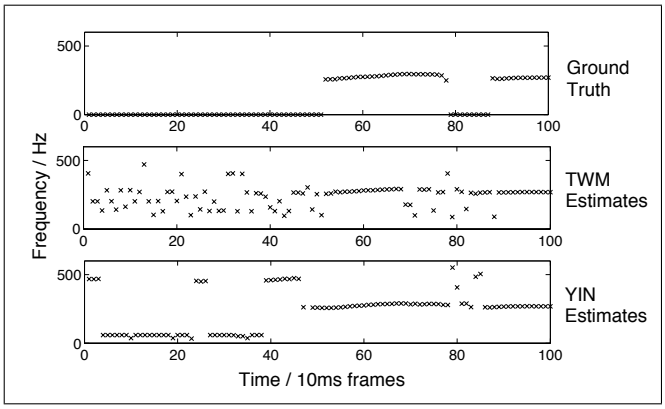

Figure 3. Pitch estimates after semitone-cancellation

partials of non-vocal notes survive this procedure, the output is finally bandpass-filtered to 300–2000Hz, which roughly corresponds to the pitch range of the human singing voice.

3.1.2. Pitch estimation by TWM

Informal listening tests show that the output of the semitone cancellation procedure is generally dominated by vocals, with components from other instruments being unpitched or much quieter. Thus it is feasible to transcribe vocal pitch by passing this output to a monophonic transcription algorithm. This algorithm should favour predominant partials on voiced frames to achieve high pitch accuracy. Since the fusion system (see Section 3.3) favours pitch continuity, it should also produce scattered pitch estimates on unvoiced frames to achieve high voicing detection accuracy. The TWM algorithm was chosen, as it offers a good compromise between these two objectives [1]. Other algorithms were found to generally transcribe weak instrumental notes on unvoiced frames [4], as illustrated in Figure 3.

3.1.3. Reliability measure

In order to assess which TWM pitch estimates are likely to be correct, each estimate is further associated with a reliability measure. This measure is obtained simply by mapping the TWM error [5] linearly to the interval $[0, 1]$.

3.2. High-frequency correlogram

The second vocal pitch estimator consists of a correlogram-based monophonic pitch transcription algorithm using only certain channels where the voice is likely to be predominant.

3.2.1. Correlogram design

The input audio is filtered by a 50-channel gammatone filterbank spanning the range 100Hz–22kHz [2]. The unbiased autocorrelation function (ACF) of each channel is computed in 50ms frames at 10ms intervals. The predominant pitch is then estimated in each channel and each frame by summing

[1] We used the implementation described in U. Zölzer, editor. *DAFX : Digital Audio Effects*. Wiley, 2002.

[2] This filterbank was implemented using the Auditory Toolbox available at http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/

the ACF value at the first three multiples of each integer lag in the singing voice range (1–12.5 ms) and picking the lag resulting in the largest sum. We found this method more reliable than full harmonic comb matching of the ACF.

3.2.2. High-frequency bias

Based on the channel reliability measures computed in Section 2.2, only 19 correlogram channels in the range 3–15kHz are used. The vocal pitch is then estimated for each time frame by clustering together channel-wise pitch estimates within 50 cents of each other and selecting the cluster with largest population. Experimentally, this approach provides the desired behaviour of accurate pitch estimates on voiced frames and scattered estimates on unvoiced frames. Other transcription algorithms applied to the input audio bandpass-filtered to 3–15kHz also produced scattered estimates on unvoiced frames, but achieved lower pitch accuracy [4].

3.2.3. Reliability measure

As above, each estimate is associated with a reliability measure. In this case, we wish to mark estimates as reliable when there is a strong consensus among correlogram channels. Thus reliability is defined as the proportion of channel-wise estimates belonging to the selected cluster.

3.3. Modified HMM

The fusion system is based on a HMM in which the hidden states represent the exact pitch sung, and the observed data are the pitch estimates and reliability measures from the two estimators described above. The Viterbi algorithm is used to produce the output transcription.

3.3.1. Dynamic state generation

Rather than defining an infinite number of hidden states to model continuous frequency, the states of the HMM are defined dynamically based on the input pitch estimates. With K pitch estimates $\{e_{k,t}\}_{1\le k\le K}$ at time t, the set of $(K+1)$ states is defined by $\Omega_t = \{\omega_{j,t}\}_{0\le j\le K}$ where

$$\omega_{j,t} = \begin{cases} \text{unvoiced} & \text{if } j = 0, \\ e_{j,t} & \text{if } 1 \le j \le K. \end{cases} \quad (1)$$

The notations $\omega_{j,t}$ and $e_{j,t}$ refer both to states and observations and to their assigned frequency values. The proposed system uses two pitch estimators, and so $K = 2$.

To avoid transcription errors when both estimators briefly fail, an additional *dummy state* is generated at time t for each state at $t-1$ for which there is no nearby estimate at t. More precisely, a state with frequency $\omega_{j,t-1}$ is added to Ω_t if there is no k for which $e_{k,t}$ is within 50 cents of $\omega_{j,t-1}$. A *pruning* process is introduced in the Viterbi algorithm to prevent such states persisting indefinitely [4].

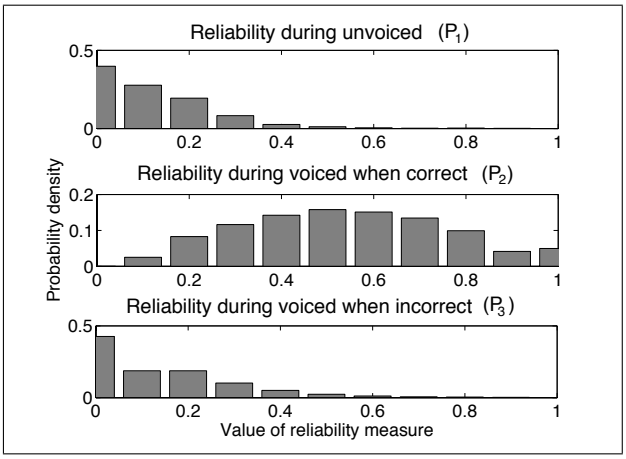

Figure 4. Distributions of High-Frequency Correlogram reliability measures

3.3.2. Observation probabilities

Unlike previous post-processing HMM systems [6], the proposed system considers all pitch estimates $\{e_{k,t}\}_{1\le k\le K}$ and reliability measures $\{r_{k,t}\}_{1\le k\le K}$ when calculating the observation probability of a given state $\omega_{j,t}$, defined by

$$P(\{e_{k,t}\}_{1\le k\le K}, \{r_{k,t}\}_{1\le k\le K} \mid \omega_{j,t}) = \prod_{k=1}^{K} P(e_{k,t}, r_{k,t} \mid \omega_{j,t}). \quad (2)$$

Each per-estimate observation probability $P(e_{k,t}, r_{k,t} \mid \omega_{j,t})$ is calculated using one of three probability distributions specific to the corresponding estimator k, depending on whether the state is voiced or unvoiced and whether the difference $d = 1200 \times |\log_2(\omega_{j,t}/e_{k,t})|$ between state and estimate frequencies is larger than 50 cents:

$$P(e_{k,t}, r_{k,t} \mid \omega_{j,t}) = \begin{cases} P_{1,k}(r_{k,t}) & \text{if } j = 0, \\ P_{2,k}(r_{k,t}) & \text{if } j \ne 0 \text{ and } d \le 50, \\ P_{3,k}(r_{k,t}) & \text{if } j \ne 0 \text{ and } d > 50. \end{cases} \quad (3)$$

These distributions were learnt by applying the two pitch estimators to the nine training recordings mentioned in Section 2.2 and forming histograms of the resulting reliability measures. The distributions for the high-frequency correlogram estimator are shown in Figure 4 and those for the semitone-cancellation-TWM estimator have similar shape. It can be seen that the reliability measures are generally low on unvoiced frames and for incorrect pitch estimates, but higher for correct pitch estimates.

3.3.3. Transition probabilities

Transition probabilities between voiced states are modelled using a combination of Gaussians with variances of 50 and 100 cents representing the variation in pitch during a note

Table 1. Summary evaluation for 19 30-second test recordings

System	Voicing Recall	False Alarm	d-prime Measure	Raw Pitch Accuracy	Raw Chroma Accuracy	Overall Accuracy
HF Corr.	58%	17%	1.20	59%	63%	63%
SC/TWM	68%	29%	1.00	56%	67%	58%
Proposed	71%	24%	1.25	71%	77%	67%

and between successive notes respectively. Other transition probabilities are estimated from the ground truth transcriptions for the training set. The transition probability from state $\omega_{i,t-1}$ to state $\omega_{j,t}$ is therefore defined as

$$P(\omega_{j,t} \mid \omega_{i,t-1}) = \begin{cases} 0.97 & i=0, j=0, \\ 0.03 \times \frac{1}{|\Omega_t|-1}, & i=0, j\neq 0, \\ 0.014 & i\neq 0, j=0, \\ c_{i,t} \times (0.936 \times \mathrm{e}^{\frac{-d^2}{100}} + 0.05 \times \mathrm{e}^{\frac{-d^2}{200}}), & i\neq 0, j\neq 0 \end{cases} \quad (4)$$

where $d = 1200 \times |\log_2(\omega_{j,t}/\omega_{i,t-1})|$ denotes the pitch difference in cents and $c_{i,t}$ is a normalisation factor chosen such that the transition probabilities sum to one.

4. Evaluation

The proposed system was first tested on 19 30-second extracts covering a wide range of genres and instrumentations, and evaluated according to the criteria used in the MIREX 2005 Audio Melody Extraction task. For comparison, the two pitch estimators were tested individually by running the HMM with a single set of pitch estimates. The results for the three systems are shown in Table 1. It can be seen that the proposed system considerably outperforms either single-estimate system, with a better d-prime value for voicing detection and substantially higher pitch accuracy. This demonstrates that there is a benefit to using multiple pitch estimators in parallel, and that the modified HMM system is a suitable fusion method.

The system was also entered for the MIREX 2006 Melody Extraction Task, with results being compiled for vocal melodies, non-vocal melodies, and all melodies. In the case of vocal melodies (see Table 2), the system achieved a similar transcription accuracy as above, ranking it third out of five in both categories. The same test set was used in 2005 and when vocal melody results are compiled for the 2005 systems also, the proposed system ranks 4/15 for raw pitch accuracy and 5/15 for overall accuracy.

The voicing performance was better than during previous testing, achieving a d-prime measure of 1.74, compared with the top-scoring system's d-prime measure of 1.75. The system's specialisation for vocal melodies is demonstrated well by the results for non-vocal melodies, where both voicing and pitch estimation performance fall considerably, and overall accuracy drops to around 30%.

Table 2. MIREX 2006 results - Vocal Melodies

System	Voicing Recall	False Alarm	d-prime Measure	Raw Pitch Accuracy	Raw Chroma Accuracy	Overall Accuracy
Dressler	85.5%	28.7%	1.62	78.5%	81.6%	73.7%
Ryynänen	77.0%	15.6%	1.75	75.7%	76.9%	72.5%
Poliner	93.7%	44.3%	1.68	69.1%	70.6%	65.0%
Sutton	71.8%	12.3%	1.74	70.7%	71.6%	67.3%
Brossier	99.6%	97.9%	0.63	42.7%	53.5%	30.7%

5. Conclusion

It was hoped that by narrowing the melody transcription task to vocal melodies only, a higher accuracy of transcription would be achievable. Though results do not yet show this, the pitch accuracy obtained is promising for a system which has not yet been extensively developed. The pitch estimation results also demonstrate the potential of using multiple pitch estimators in parallel. The benefit of specialising in vocal melodies is shown by the strong voicing performance, where a relatively simple method achieves voicing performance similar to the top-scoring system. More information about the proposed system and a discussion of potential improvements are available in [4].

6. Acknowledgments

We would like to thank the staff of UIUC for all their hard work organising and running the 2006 MIREX competition. We are especially grateful for the extra work involved in producing separate results for vocal melodies.

E. Vincent is funded by EPSRC grant GR/S75802/01.

References

[1] R. Timmers and P. W. M. Desain. Vibrato: The questions and answers from musicians and science. In *Proc. Int. Conf. on Music Perception and Cognition (ICMPC)*, 2000.

[2] A. Shenoy, Y. Wu, and Y. Wang. Singing voice detection for karaoke application. In *Proc. Int. Symp. on Visual Communications and Image Processing (VCIP)*, 2005.

[3] Y. Li and D. L. Wang. Detecting pitch of singing voice in polyphonic audio. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages III–17–20, 2005.

[4] C. Sutton. Transcription of vocal melodies in popular music. Master's thesis, Dept. of Electronic Engineering, Queen Mary, University of London, 2006.

[5] R. C. Maher and J. W. Beauchamp. Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of the Acoustical Society of America*, 95(4):2254–2263, 1994.

[6] M. P. Ryynänen and A. P. Klapuri. Polyphonic music transcription using note event modeling. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 319 – 322, 2005.