SINGING PITCH EXTRACTION AT MIREX 2009

Chao-Ling Hsu **Jyh-Shing Roger Jang** **Liang-Yu Chen**
Multimedia Information Retrieval Laboratory
Computer Science Department, National Tsing Hua University
Hsinchu, Taiwan
{leon, jang, davidson833} @mirlab.org

ABSTRACT

This extended abstract describes our submissions to the MIREX 2009 evaluation task on Audio Melody Extraction. The algorithms are designed for vocal F0 extraction from the music accompaniment. Although the low voicing recall decreases the overall accuracy of our algorithms, our algorithms still perform above average.

1. INTRODUCTION

Two algorithms were submitted to Audio Melody Extraction task of MIREX 2009. Both of them are designed for vocal (singing voice) F0 extraction. One of them enhances the harmonic structures of singing voices and extracts the strength of each pitch candidates. Dynamic Programming (DP) technique is then used to determine the vocal F0. The other one uses not only the enhanced singing harmonic but also considers the music contextual information. A two-stream Hidden Markov Model (HMM) is employed to determine the most likely singing pitches. Both the algorithms are described in detail in [1].

2. ALGORITHM 1

Fig. 1 shows the overview of the submitted algorithm 1. The strength of each pitch candidate is extracted from the normalized sub-harmonic summation (NSHS) map of the input polyphonic song. A DP technique is then employed to determine the unbroken pitch vectors. On the other hand, the MFCCs (Mel-frequency cepstral coefficients) are extracted to perform the voiced/non-voiced detection. Lastly, the singing pitch vectors are produced by integrating the results of these two processes. The following subsections explain these blocks in detail.

2.1 Features Extraction from a Spectrum

This block extracts 39-dimensional MFCCs (12 cepstral coefficients plus a log energy, together with their first and second derivatives) as the features for a 2-state HMM for voiced/non-voiced detection.

2.2 HMM-based Voiced/Non-voiced Detection

This block employs a continuous 2-state HMM to decode the mixture input into voiced and non-voiced segments, similar to the one proposed by Fujihara et al. [2]. Note that the "voiced" here indicates the voiced singing voice, and "non-voiced" indicates the unvoiced singing voice and music accompaniments. Given the MFCC feature

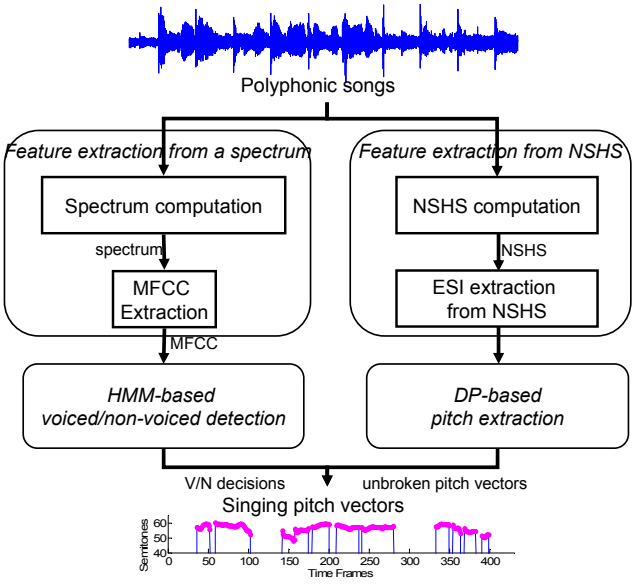

Figure 1. Algorithm 1 overview.

vectors $X = \{x_0, \cdots, x_t, \cdots\}$ of the input mixtures, the problem is to find the most probable sequence of voiced/non-voiced states, $\hat{S} = \{s_0, \cdots, s_t, \cdots\}$:

$$\hat{S} = \arg\max_{S} \left\{ p(s_0)p(x_0 \mid s_0) \prod_t \{ p(s_t)p(x_t \mid s_t)p(s_t \mid s_{t-1}) \} \right\}, \quad (1)$$

where $p(x \mid s)$ is the output likelihood of a state s, $p(s_t \mid s_{t-1})$ is the state transition probability from state s_{t-1} to s_t, and $p(s_t)$ is the prior of the state s_t. Note that $p(s_t \mid s_{t-1})$ and $p(s_t)$ can be obtained from the actual song data with manual annotations.

2.3 Features Extraction from NSHS

This block extracts the feature vector which represents the energy distributions of the enhanced harmonic structures of singing voices. The harmonic structures can be enhanced by sub-harmonic summation (SHS) proposed by Hermes [3]:

$$H_t(f) = \sum_{n=1}^{N} h_n P_t(nf), \quad (2)$$

where $H_t(f)$ is the sub-harmonic summation value of the frequency f at time frame t, $P_t(*)$ is the power spectrum calculated from STFT, n is the index of harmonic components, N is the number of the harmonic components in consideration, and h_n is the weight indicating the contribution of the n th harmonic component. Usually we set $h_n = h^{n-1}$, where $h \le 1$. In order to further enhance the harmonics of singing voices, we propose the use of normalized SHS (NSHS) defined as follows:

$$\hat{H}_t(f) = \frac{\sum_{n=1}^{N_f} h_n P_t(nf)}{\sum_{n=1}^{N_f} h_n}, \quad (3)$$

where the number of harmonic components N_f depend on the frequency under consideration:

$$N_f = floor\left(\frac{0.5 f_s}{f}\right), \quad (4)$$

with f_s being the sampling rate. The reason of the modification is based on the observation that most of the energy in a song in located at the low frequency bins, and the energy of the harmonic structures of the singing voice seems to decay slower than that of instruments [4]. Therefore, when more harmonic components are considered, energy of the vocal sounds is further strengthened. Although some percussive instruments (e.g. cymbals) present high energy at higher frequency bins, their non-harmonic nature does not affect the NSHS much.

Based on the proposed NSHS, we can extract a feature vector of Energy at Semitones of Interests (ESI) for each given frame.

For each integer semitone of interests within the range $[40,75]$, we identify its maximum energy as an element of the feature vector. Take semitone 69 for example, the search range in semitone is $[68.5,69.5]$, corresponding to a frequency bin of $[427.47,452.89]$ in terms of Hertz. Then we find the maximum power spectrum within this range as the feature associated with semitone 69. Since there are 36 elements within semitone of interests, the length of the feature vector of ESI is also 36. Please refer to [1] for more details.

2.4 DP-based Pitch Extraction

The goal of using the DP technique is to find a path $f = [f_0, \cdots, f_i, \cdots, f_{n-1}]$ that maximizes the score function:

$$score(f, \theta) = \sum_{t=0}^{n-1} Y_t(f_t) - \theta \times \sum_{t=1}^{n-1} |f_t - f_{t-1}|, \quad (5)$$

where $Y_t(f_t)$ is a feature vector extracted from spectrum/NSHS at the frame t and frequency f_t. The first term in the score function is the sum of energy of the pitches along the path, while the second term controls the smoothness of the path with the use of a penalty term θ (which is set to 2 in our submission). If θ is larger, then the computed path are smoother.

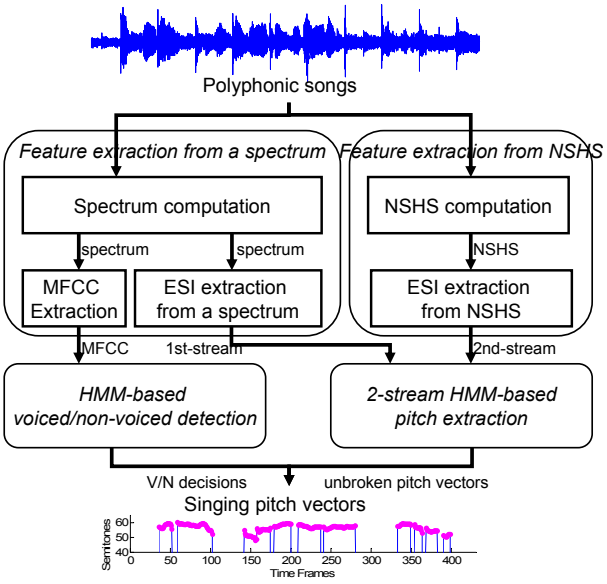

Figure 2. Algorithm 2 overview

The algorithm 1 employs dynamic programming to find the maximum of the score function, where the optimum-valued function $D(t,m)$ is defined as the maximum score starting from frame 1 to t, with $f_t = m$:

$$D(t,m) = Y_t(m) + \max_{k \in [0,35]} \{D(t-1,k) - \theta \times |k-m|\}, \quad (6)$$

where n is the number of frames, $t = [1, n-1]$, and $m = [0,35]$. The initial conditions are $D(0,m) = Y_0(m)$, and the optimum score is equal to $\max_{m \in [0,35]} D(n-1,m)$.

3. ALGORITHM 2

Fig. 2 shows the overview of the submitted algorithm 2. Two streams of features are extracted from the spectrogram and the NSHS map, respectively, of the input polyphonic song. A 2-stream HMM is then employed to decode the input songs into the most likely unbroken pitch vectors. On the other hand, the MFCCs (Mel-frequency cepstral coefficients) are extracted to perform the voiced/non-voiced detection which is same as the one used in algorithm 1. Lastly, the singing pitch vectors are produced by integrating the results of these two processes. The following subsections explain these blocks in detail.

3.1 Features Extraction from a Spectrum

This block extracts two types of features, including MFCCs and ESI (Energy at Semitones of Interests). MFCCs are the features for a 2-state HMM for voiced/non-voiced detection. ESI is the 1st-stream feature for a 2-stream HMM for pitch extraction. Different with algorithm 1, algorithm 2 extracts ESI from both spectrum and NSHS map.

Table 1. *MIREX 2009 Audio Melody Extraction Overall Summary results - Weighted (by Number of Files) Avg. of all Datasets*

Rank	Participant	Voicing Recall	Voicing False alarm	Raw Pitch Accuracy	Raw Chroma Accuracy	Overall Accuracy
1	Dressler	90.5445 %	47.2188 %	77.5947 %	79.5145 %	**66.7341 %**
2	Rao and Rao	89.2142 %	49.0418 %	67.4012 %	70.4479 %	60.618 %
3	Durrieu and Richard 1	91.7457 %	53.7476 %	68.6411 %	71.3884 %	60.0492 %
4	Durrieu and Richard 2	87.2511 %	45.2327 %	65.3401 %	69.8261 %	59.5668 %
5	Hsu, Jang and Chen 1	37.8763 %	**2.8033 %**	68.4293 %	72.0766 %	54.9707 %
6	Hsu, Jang and Chen 2	37.8763 %	**2.8033 %**	50.4366 %	66.9551 %	54.1855 %
7	Tachibana, Ono, Ono and Sagayama	99.8103 %	99.1475 %	**80.0558 %**	**83.7591 %**	52.711 %
8	Cancela	73.4387 %	42.4767 %	51.778 %	54.1777 %	52.5412 %
9	Cao and Li 2	77.2452 %	59.2516 %	58.9205 %	62.7661 %	49.5138 %
10	Joo, Jo and Yoo	38.0818 %	17.8244 %	73.0217 %	77.7374 %	48.6962 %
11	Cao and Li 1	92.3264 %	83.1113 %	58.9205 %	62.7661 %	44.3575 %
12	Wendelboe	**99.9896 %**	99.3837 %	66.4069 %	70.34 %	43.6651 %
Mean	**All Participants**	**76.2833 %**	**50.1702 %**	**65.5789 %**	**70.1463 %**	**53.9674 %**

3.2 HMM-based Voiced/Non-voiced Detection

This block is same as the one used in 2.2.

3.3 Features Extraction from NSHS

This block is same as the one used in 2.3. It extracts the 2nd-stream feature vector which represents the energy distributions of the enhanced harmonic structures of singing voices.

3.4 2-Stream HMM-based Pitch Extraction

We employ a 2-stream HMM to model the relationship between the adjacent melody pitches and their corresponding audio context. Given the 1st-stream ESI feature vectors $V = \{v_0, \cdots, v_t, \cdots\}$ from spectrogram and the 2nd-stream ESI feature vectors $C = \{c_0, \cdots, c_t, \cdots\}$ from NSHS map, our goal is to find the most likely sequence of pitch states, $\hat{R} = \{r_0, \cdots, r_t, \cdots\}$:

$$\hat{R} = \arg\max_{R} \left\{ p(r_0)\, p(v_0, c_0 \mid r_0) \prod_t \left\{ p(r_t)\, p(v_t, c_t \mid r_t)\, p(r_t \mid r_{t-1}) \right\} \right\}, \tag{7}$$

where $p(r_t \mid r_{t-1})$ is the state transition probability from pitch state r_{t-1} to r_t, $p(r_t)$ is the prior of the pitch state r_t, and $p(v, c \mid r)$ is the joint output likelihood of the pitch state r defined as:

$$p(v, c \mid r) = p_v(v \mid r)\, p_c(c \mid r), \tag{8}$$

where $p_v(v \mid r)$ and $p_c(c \mid r)$ are the state likelihoods of feature vectors v and c, respectively, given the state r. This is a typical multi-stream HMM which is broadly used in speech processing [5]. The state likelihoods (or conditional observation likelihoods), transition probabilities, and priors of eq. (7) and (8) can all be obtained from the actual song data with manually annotated pitch contours.

4. RESULTS

Table 1 shows the weighted overall summary results. Although the dataset contains non-vocal songs (our algorithms are mainly designed for vocal songs), the results are still worthy of being analyzed because the number of the non-vocal songs is relatively small.

The results show that our algorithm 1 performs better than algorithm 2, and it seems contrary because algorithm 2 uses more information than algorithm 1. It may be resulted from the limited variety of HMM training dataset.

Comparing to other approaches, the raw pitch accuracy and raw chroma accuracy of our algorithms are roughly comparable as the algorithms ranked 2nd to 4th. However, the low voicing recall degenerates the overall accuracy. This can be improved by increasing the size and variety of the dataset to train the voiced/non-voiced detection HMM models. Moreover, some of the other submissions perform significantly better in raw pitch accuracy such as Dressler and Tachibana et al. The results show that there is still a large room for us to improve the performance.

5. CONCLUSIONS AND FUTURE WORK

This extended abstract describes our submissions to the MIREX 2009 evaluation task on Audio Melody Extraction. Our algorithms perform above average.

To improve our algorithm, we plan to develop a more robust voiced/non-voiced detector which can be used in

both vocal and non-vocal songs. On the other hand, we will try to find out the reason that causes the lower performance, compared to Tachibana et al., in raw pitch accuracy.

6. REFERENCES

[1] Chao-Ling Hsu, Liang-Yu Chen, Jyh-Shing Roger Jang, and Hsing-Ji Li, "Singing Pitch Extraction From Monaural Polyphonic Songs By Contextual Audio Modeling and Singing Harmonic Enhancement", International Society for Music Information Retrieval, Kobe, Japan, Oct. 2009.

[2] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic Synchronization between Lyrics and Music CD Recordings Based on Viterbi Alignment of Segregated Vocal Signals," *ISM*, pp. 257–264, 2006.

[3] D. J. Hermes, "Measurement of Pitch by Subharmonic Summation," *Journal of Acoustic Society of America*, vol.83, pp. 257-264, 1988.

[4] Y. Li and D. L. Wang, "Detecting Pitch of Singing Voice in Polyphonic Audio," *IEEE ICASSP*, pp. 17–20, 2005.

[5] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. J. Odell, D. Ollason, D. Povery, V. Valtchev, and P. C. Woodland: *The HTK Book (for HTK version 3.4)*, Cambridge University, 2006.