BAYESIAN FRAMEWORK-BASED VOCAL MELODY EXTRACTION FOR MIREX 2014

Liming Song, Ming Li
Key Laboratory of Speech Acoustics and Content Understanding,
Chinese Academy of Sciences, Institute of Acoustics, CAS
{lmsong86, liming.ioa}@gmail.com

ABSTRACT

We present an effective approach for automatic extraction of the main melody from polyphonic music, especially vocal melody songs. The approach is based on a Bayesian framework by calculating the probability of each pitch candidate using a set of characteristics, and searching for the best pitch sequence by viterbi algorithm. We submitted our algorithm for the audio melody extraction task of the Music Information Retrieval Evaluation eXchange (MIREX) 2014.

1. INTRODUCTION

Melody is the a concise and representative description of polyphonic music defined as a succession of the predominant fundamental frequency of the musical source. It can be used in numerous applications such as automatic music transcription, music structure analysis, "Query-by-humming" system and music information retrieval.

Our algorithm is based on a Bayesian framework. Figure 1 is the block diagram of our system, which is consist of two module: a pitch evolution model and a acoustic model. Pitch evolution model describes how pitch contours change and contains two sub-models that are pitch transition model and harmonics variation model. The acoustic model represents what the acoustic characteristics would have when the pitch is the hypothesis one. The acoustic model includes three sub-models that are pitch periodicity model, harmonic shape model and vocal/non-vocal model. the output melody contour is obtained using Viterbi algorithm by maximize the weighted summation of the scores of each sub-models.

2. METHOD DESCRIPTION

We assume $\mathbf{F}_0$ is the pitch sequence of a polyphonic music song, $\mathbf{O}$ is the characteristics observed from the input musical signal, and $\mathbf{F}_0^c$ is the pitch candidate series. The optimal pitch sequence will satisfy the following equation

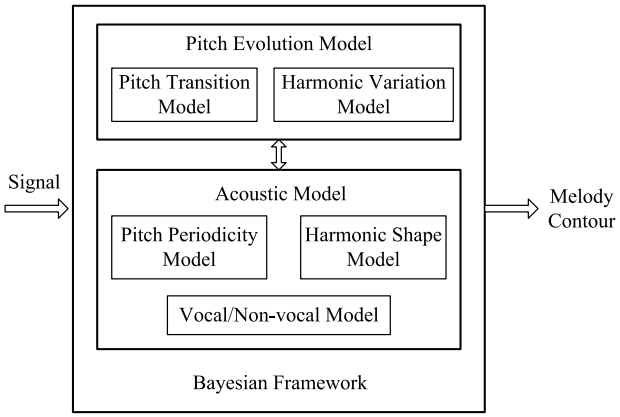

Figure 1. The block diagram of the system.

(1), where Bayes theorem is applied.

$$\begin{aligned}\mathbf{F}_0 &= \arg\max_{\mathbf{F}_0^c} p(\mathbf{F}_0^c|\mathbf{O}) \\ &= \arg\max_{\mathbf{F}_0^c} p(\mathbf{F}_0^c)p(\mathbf{O}|\mathbf{F}_0^c)\end{aligned} \tag{1}$$

Where

$$\begin{aligned}p(\mathbf{F}_0^c) &= p(f_{0,1}^c, f_{0,2}^c, ...f_{0,N}^c) \\ &= p(f_{0,1}^c)\prod_{t=2}^{N} p(f_{0,t}^c|f_{0,t-1}^c)\end{aligned} \tag{2}$$

and

$$p(\mathbf{O}) = \prod_{t=1}^{N} p(O_t^c|f_{0,t}^c) \tag{3}$$

2.1 Pitch Evolution Model

Considering that the melody contour is continuous in time and changes gradually, as well as the magnitude of the multiples of F0 and the pitch is only associated with the previous situation, the pitch evolution model is defined as

$$\begin{aligned}p(f_{0,t}^c|f_{0,t-1}^c) &= p(\Delta f, V_t|f_{0,t-1}^c) \\ &= p(\Delta f|f_{0,t-1}^c)p(V_t|f_{0,t-1}^c)\end{aligned} \tag{4}$$

Where $\Delta f = f_{0,t-1}^c - f_{0,t}^c$ represents the change of the pitch from one frame to the next and V_t is the normalized cross-correlation of the energy vectors $\mathbf{E}^h$ of the harmonic

of the neighboring pitch, representing the harmonic variation.

$$V_t = \frac{\mathbf{E}_t^h \cdot \mathbf{E}_{t-1}^h}{|\mathbf{E}_t^h| \cdot |\mathbf{E}_{t-1}^h|} \tag{5}$$

As suggested in [1], the distribution of Δf can be modeled by a Laplacian. The score of normalized cross-correlation is considered as a exponent function. In this submission, we limit $|\Delta f| \leq 30$ to further reduce search space.

2.2 Acoustic Model

The acoustic model given a pitch is defined as follow

$$\begin{aligned} p(O_t^c|f_{0,t}^c) &= p(E_t^s, M_t, S_t|f_{0,t}^c) \\ &= p(E_t^s|f_{0,t}^c)p(M_t|f_{0,t}^c)p(S_t|f_{0,t}^c) \end{aligned} \tag{6}$$

Where E_t^s represents the pitch periodicity. Here we use the sub-harmonic summation of the pitch candidate reported in [2], concluded as:

$$E_t^s = \sum_{m=1}^{M} h^{m-1} S(m f_0^c, t) \tag{7}$$

Where $S(f,t)$ is the power spectrum value of the frequency f at t^{th} frame, and h is a compression factor, $0 < h < 1$, implying that higher harmonics contribute less to the pitch than lower harmonics do.

The score of sub-harmonic summation is defined as a exponent function, noting that it is not a probability in a strict sense.

$$p(E_t^s|f_{0,t}^c) = exp[-\alpha_s(\frac{E_t^s}{\max_t E_t} - \beta_s)] \tag{8}$$

M_t relates to the harmonic shape, representing the likelihood whether the pitch candidate is generated by the singer or other instrument. We extract the inhibited magnitude of the harmonic of pitch candidates as a feature vector which further scored by a Gaussian Mixture Model(GMM). To decide whether the vocal exists at a certain time is a key task in melody extraction, named as Vocal/Non-vocal (VNV) decision (S_t). The human voice as a special instrument, has its own unique timbre characteristic compared to other instruments. We use a combination of Shifted Delta Cepstra for Mel-Frequency Cepstrum Coefficients (MFCC-SDC) [3], Spectral Contrast Feature (SCF) [5] and Harmonic Features (HF) [4] as the features to train two models via GMM: Vocal model λ_V and Non-vocal model λ_{NV}. Then the VNV likelihood can be measured as

$$S_t = \log p(\mathbf{x}|\lambda_V) - \log p(\mathbf{x}|\lambda_{NV}) \tag{9}$$

Then we perform a Viterbi search to obtain the best output pitch sequence $\mathbf{F}_0$. The above algorithm based on Bayes framework makes reasonable balanced decisions among different pitch hypotheses.

3. REFERENCES

[1] Z. Jin, D.L. Wang: "HMM-based multipitch tracking for noisy and reverberant speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 19, No. 5, pp. 1091–1102, 2011

[2] C. Cao, M. Li, J. Liu, Y.H. Yan: "Singing melody extraction in polyphonic music by harmonic tracking," *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, pp. 373–374, 2007.

[3] M.A. Kohler, M. Kennedy: "Language identification using shifted delta cepstra," *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, Vol. 3, pp. III–69, 2002.

[4] P. Cook: "Identification of Control Parameters inan Articulatory Vocal Tract Model, with Applicationsto the Synthesis of Singing," *Ph. D. diss., StanfordUniversity Electrical Engineering Department*, 1990.

[5] D. Jiang, L. Lu, H. Zhang, J. Tao, L. Cai: "Music type classification by spectral contrast feature," *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, Vol. 1, pp. 113–116, 2002.