MELODY ESTIMATION FOR MIREX 2011

Wei-Hsiang Liao and Alvin W. Y. Su
National Cheng-Kung University
Tainan, Taiwan

Chunghsin Yeh and Axel Roebel
IRCAM/CNRS-STMS
Paris, France

ABSTRACT

The most dominant source in audio is usually perceived as melody. The estimation of it's fundamental frequency can be treated as melody estimation. The task become difficult when the polyphony of the audio is high. To determine the most dominant source in human perception, perceptual properties have to be accounted. In this algorithm, we applied psychoacoustic property to the data to estimate the melody.

1. INTRODUCTION

This algorithm is mainly derived from [2]. In this study, we propose to evaluate the following perceptual criteria: loudness, masking, and timbre similarity within the proposed melody estimation system. The auditory filters and other multi-resolution analysis methods are not explored here because we believe that the melody source stream is usually significantly present in the mid-frequency range and a fixed resolution of STFT (short-time Fourier transform) can thus be sufficiently adapted.

The proposed system consists mainly of two parts: candidate selection and tracking. As the salience of an F0 candidate is derived from the the dominant peaks that are harmonically matched, we propose to utilize perceptually-motivated criteria for dominant peak selection. Similarly, candidate scoring based on perceptual criteria is also evaluated to reveal how a correct candidate can be more favored than others. Based on the algorithm previously proposed in [1], a tracking algorithm dedicated to melody estimation is developed to determine the coherent source stream with an optimal trade-off among candidate score, smoothness of frequency trajectory and spectral envelope similarity.

The article is organized as follows: In Section 2, we present the methods for dominant peak selection and candidate scoring. In Section 3, the components of the tracking system is detailed.

2. CANDIDATE EXTRACTION

Extraction of compact F0 candidates from polyphonic signals is not an easy task because concurrent sources interfere with each other and spectral components from different sources may form reasonable F0 hypotheses [3]. Although a proper multiple-F0 estimation allows proper treatment of overlapping partials, a simpler scheme shall meet our needs for melody estimation.

Under the assumption that the melody stream is generated by the most dominant source, the interference from other sources has less impact on its spectral components. By means of selecting the dominant peaks, we can avoid excessive spurious candidates and efficiently establish a compact set of F0 hypotheses with reliable salience.

2.1 Peak Selection

The peak selection is mainly rely on masking effect. The masking effect depicts how a tone can mask its neighboring components across critical bands, which can be represented by the spreading function (on dB scale)

$$\begin{aligned} S_f(i,j) =& 15.81 + 7.5((i-j)+0.474) \\ &- 17.5(1+((i-j)+0.474)^2)^{0.5} \end{aligned} \tag{1}$$

where i is the bark frequency of the masking signal, and j is the bark frequency of the masked signal. The formula of converting frequency f_k from "kHz" to the bark scale is:

$$B(f_k) = 13 \cdot \arctan(0.76 \cdot f_k) + 3.5 \cdot \arctan(\frac{f_k}{7.5})^2 \tag{2}$$

The strength of masking of a peak is not only determined by the magnitude of the peak, but also related to its being *tonal* or *noisy*. We used a simple scheme to classify a peak : If a peak is 7dB higher than its neighboring component, it is considered tonal. Otherwise, it is considered noisy. Accordingly, the mask contributed by a peak is thus (on dB scale):

$$\begin{aligned} M(i,j) =& S_f(i,j) - (14.5+i)\cdot\alpha - 5.5\cdot(1-\alpha) \\ &(tonal : \alpha = 1, noisy : \alpha = 0) \end{aligned} \tag{3}$$

By means of selecting the maximal mask overlaying at each bin, the masking curve X_m is constructed:

$$20\log_{10} X_m(k) = \max\{M(i, B(f_k))\},\ \forall i \in I \tag{4}$$

where I is the set of all peaks. The peaks which are larger than the masking curve are selected.

2.2 Candidate Generation and Scoring

Similar to the *inter-peak beating* method proposed in [3], we present a method for generating F0 candidates from the selected dominant peaks. First, the F0 hypotheses are generated by collecting the spectral intervals between any pair of dominant peaks in the spectrum. Then, the spectral location principle is applied: If the generated hypothesis is not harmonically related to the peaks that support its spectral interval, it is not considered a reasonable candidate. Due to the overlapping partials, frequencies of the peaks are not sufficiently precise. Thus, a semitone tolerance is allowed for the harmonic matching.

In order to reflect the perceptual dominance of a candidate, we propose to score F0 candidates based on the loudness spectrum X_L (eq. 5):

$$X_L(k) = \frac{X(k)}{L(k)}; \tag{5}$$

where k is the frequency bin. We choose the equal-loudness curve proposed by Fletcher and Munson measuring at 0dB SPL (sound pressure level) for L:

$$\begin{aligned} 20\log_{10} L(k) =& 3.64 \cdot f_k^{-0.8} - 6.5 \cdot e^{-0.6\cdot(f_k-3.3)^2} \\ &+ (10^{-3}) \cdot f_k^4 \end{aligned} \tag{6}$$

where the frequency f_k in “kHz” is converted from the respective frequency bin k. Then, we select the peaks that are not smaller than δ_LdB of the maximum of X_L.

The score of a candidate is the summation of the first $H = 10$ partials in the loudness spectrum. The contribution of a partial is determined by the harmonically matched peak with the largest loudness nearby. The partials not selected as dominant peaks will not contribute to the score.

3. TRACKING BY DYNAMIC PROGRAMMING

Given a sequence of candidates extracted from the spectrogram, we adapt the tracking algorithm proposed in [1] to decode the melody stream. Since the melody stream may not be always the most dominant source at each short-time instant, decoding with the maximal score will not yield the optimal result. Therefore, we propose to integrate an additional criterion, spectral envelope similarity, into the dynamic programming scheme. Following [1], we describe the problem using the hidden Markov model (HMM):

- Hidden state: true melody F0
- Observation: loudness spectrogram
- Emission probability: normalized candidate score
- Transition probability
  - trajectory smoothness: the frequency difference between two connected F0 candidates
  - spectral envelope similarity: the spectral envelope difference between two connected candidates

Compared with the previous method, two novelties are introduced in the transition probability. One is the probability distribution of the melody F0 difference between frames for evaluating the trajectory smoothness. Learned from the ADC04 training database, the distribution is approximated by the Laplace distribution. The trajectory smoothness is then modeled by

$$F(c_n, c_m) = \frac{1}{2b}\exp(-\frac{|f_{c_n} - f_{c_m}|}{b \cdot f_{c_m}}), b = 0.0077889 \tag{7}$$

where c_n, c_m represent the two candidates with frequencies f_{c_n}, f_{c_m}. Notice that c_n, c_m may be located at different analysis frames and the distance allowed for connection is three frames.

The other novelty is the integration of the spectral envelope similarity in the transition probability. This is intended to favor candidate connection with similar timbre such that the decoded stream is locked to the same source even when it becomes less dominant (smaller score).

$$A(c_n, c_m) = 1 - \frac{\sum_{h=0}^{H} |X_L(t_n, hf_{c_n}) - X_L(t_m, hf_{c_m})|^2}{\sum_{h=0}^{H} X_L(t_m, hf_{c_m})^2} \tag{8}$$

where t_n, t_m denotes the frames where c_n, c_m are extracted. The transition probability is thus given by

$$T(c_n, c_m) = F(c_n, c_m)A(c_n, c_m)^{\gamma} \tag{9}$$

where γ is a compression parameter which should reflect the importance of the envelope similarity measure. In order to obtain the optimal trade-off between the emission probability (score) and the transition probability, we further apply a compression factor β on the emission probability.

The connection weight between two nodes is defined by the product of the emission probability and the transition probability, from which the forward propagated weights can be accumulated. The optimal path (melody stream) is then decoded by backward tracking through the nodes of locally maximal weights.

4. REFERENCES

[1] W.-C. Chang, W.-Y. Su, C. Yeh, A. Roebel, and X. Rodet. Multiple-f0 tracking based on a high-order HMM model. In *Proc. of the 11th Intl. Conf. on Digital Audio Effects (DAFx-08), Espoo, Finland*, 2008.

[2] Wei-Hsiang Liao, Alvin W. Y. Su, Chunghsin Yeh, and Axel Roebel. On the use of perceptual properties for melody estimation. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-11)*, Paris, France, Sept. 19–23, 2011.

[3] C. Yeh. *Multiple fundamental frequency estimation of polyphonic recordings*. PhD thesis, Université Paris 6, 2008.