On the Use of Perceptual Properties for Melody Estimation

Wei-Hsiang Liao, Alvin. Wen-Yu Su, Chunghsin Yeh, Axel Roebel

ON THE USE OF PERCEPTUAL PROPERTIES FOR MELODY ESTIMATION

Wei-Hsiang Liao and Alvin W. Y. Su
Dep. of Computer Science and Information Engineering
National Cheng-Kung University, Tainan, Taiwan
whsng.liao@gmail.com, alvinsu@mail.ncku.edu.tw

Chunghsin Yeh and Axel Roebel
Analysis/Synthesis team
IRCAM/CNRS-STMS Paris, France
cyeh@ircam.fr, roebel@ircam.fr

ABSTRACT

This paper is about the use of perceptual principles for melody estimation. The melody stream is understood as generated by the most dominant source. Since the source with the strongest energy may not be perceptually the most dominant one, it is proposed to study the perceptual properties for melody estimation: loudness, masking effect and timbre similarity. The related criteria are integrated into a melody estimation system and their respective contributions are evaluated. The effectiveness of these perceptual criteria is confirmed by the evaluation results using more than one hundred excerpts of music recordings.

1. INTRODUCTION

Auditory scene analysis of music signals is an ongoing active research in recent years as encouraging results continue to explore various applications in the field of digital audio effects (DAFx) and music information retrieval (MIR) [1]. Among the harmonic sources present in the music scene, the "melody" source usually forms perceptually and musically the most dominant stream [2] [3] [4]. The problem of melody estimation is difficult because it requires not only low-level information about sound signals but also high-level information about perception of music. In this article, we define the melody estimation problem as the estimation of the fundamental frequency (F0) of the most dominant source stream. Since the source with the strongest energy may not be perceptually the most dominant one, our study will make use of perceptual properties and evaluate their effectiveness.

In addition to the perceptual grouping cues of harmonic sounds in auditory scene analysis [5], many of the existing methods for melody estimation further make use of other perceptual properties such as loudness [6] [7], masking [8], timbre similarity [6] [9] [10] and auditory filters [3] [11] [12]. If one looks at the evaluation results of the MIREX (Music Information Retrieval Evaluation eXchange) campaign for the "Audio Melody Estimation" task, the systems that make use of these perceptual properties seem to show certain advantages in performance. In fact, the perceptually-motivated system proposed by Dressler [13, 14, 9, 15] always ranks the top [16]. Although important details of perceptual criteria are missing in her descriptions, it is nevertheless reasonable to assume that the key problem of melody estimation is related to perceptual criteria. 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 compare perceptually-motivated criteria with low-level signal features 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 [17], 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 paper 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. In Section 4, the effectiveness of the perceptual criteria are evaluated and the performance of the proposed system is compared to the state-of-the-art systems. Finally, conclusions are drawn and future works are proposed.

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 [18]. 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. The remaining problem is then to avoid extracting subharmonic F0 candidates that are supported by the combination of spectral components from different sources. They appear to be very competitive to the correct F0 and are very likely to cause octave errors. Since the target source is assumed to be dominant, its harmonic components should be present as dominant spectral peaks. 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

We propose four peak selection methods. The first two are based on loudness weighting and masking effects respectively to select *perceptually dominant* peaks, and the other two are based on cepstral envelope and noise envelope respectively to select *energy dominant* peaks.

Select by Loudness

It is known that the relative energy of the spectral components one measures is very different from the relative loudness one perceives [19]. Since calculating the loudness for complex sound is not straightforward, a common approach is to apply proper spectral weighting by a selected equal-loudness contour to imitate the perceptual dominance of spectral components. Accordingly, we weight the spectrum X with a frequency dependent equal-loudness curve L to obtain the **loudness spectrum** X_L:

$$X_L(k) = \frac{X(k)}{L(k)}; \tag{1}$$

where k is the frequency bin. We choose the equal-loudness curve proposed by Fletcher and Munson [20] measuring at 0dB SPL (sound pressure level) for L:

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

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 (see Fig. 1(a)).

Select by Masking Curve

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) [21]

$$\begin{aligned} S_f(i,j) =& 15.81 + 7.5((i-j)+0.474) \\ &- 17.5(1+((i-j)+0.474)^2)^{0.5} \end{aligned} \tag{3}$$

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 [22]:

$$B(f_k) = 13 \cdot \arctan(0.76 \cdot f_k) + 3.5 \cdot \arctan(\frac{f_k}{7.5})^2 \tag{4}$$

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 follow the MPEG's standard to classify a peak [23]: 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):

$$\begin{aligned} M(i,j) =& S_f(i,j) - (14.5+i)\cdot\alpha - 5.5\cdot(1-\alpha) \\ &(tonal: \alpha = 1, noisy: \alpha = 0) \end{aligned} \tag{5}$$

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

$$20\log_{10} X_m(k) = \max\{M(i, B(f_k))\},\ \forall i \in I \tag{6}$$

where I is the set of all peaks. The peaks which are larger than the masking curve are selected (see Fig. 1(b)).

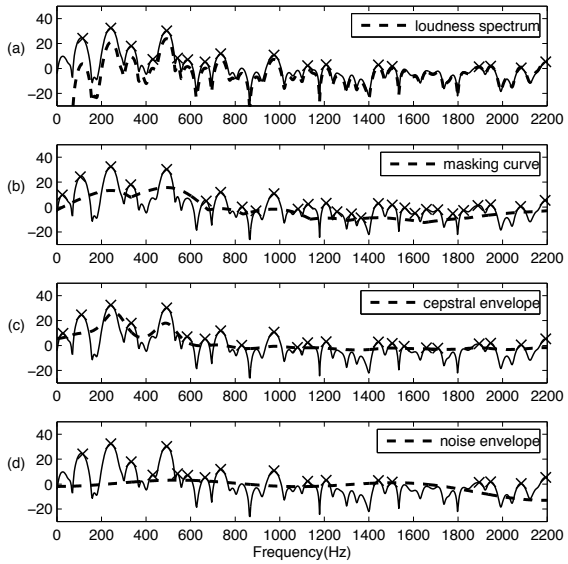

Figure 1: Dominant peak selection by (a) loudness spectrum, (b) masking curve, (c) cesptral envelope, and (d) noise envelope. The original spectrum is plotted as thin solid line and the selected peaks are marked by crosses. The y-axis is the log-amplitude in dB.

Select by Cepstral Envelope

The cepstral envelope is an approximation of the expected log-amplitude of the spectrum [24]. That is, it is a frequency-dependent curve that passes through the mean log-amplitudes at respective frequencies. Accordingly, it is reasonable to assume that the spectral peaks of the most dominant source lie above the cepstral envelope (see Fig. 1(c)). An optional raise of δ_C dB can be used to prevent selection of noise peaks.

Select by Noise Envelope

For the case of polyphonic signals, the cepstral envelope may not give reasonable estimation due to dense distribution of sinusoidal peaks. Besides, it allows some noise peaks to be selected because it passes through the mean of the noise peaks. A solution to these problems is the use of the noise envelope which is the raise of the mean noise level [18]. The proposed noise level estimation makes use of the Rayleigh distribution to model the spectral magnitude distribution of noise and is adaptive in frequency [25]. We raise the mean noise level by δ_NdB as the noise envelope to select dominant peaks (see Fig. 1(d)).

2.2. Candidate Generation and Scoring

Harris suggested locating all groups of pitch harmonics by means of identifying equally spaced spectral peaks on which the salience of a group is built [26]. This method belongs to the **spectral interval** type F0 estimators [27]. For polyphonic signals, however, partials belonging to different sources may form a group of harmonics which results in subharmonic F0s. One way to avoid generating subharmonic F0 candidates is to cast further constraints on the **spectral location** of each partial. Similar to the *inter-peak*

beating method proposed in [18], 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. 1): 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 [17] 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 [17], 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 (see Fig. 2). The trajectory smoothness is then modeled by

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

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).

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

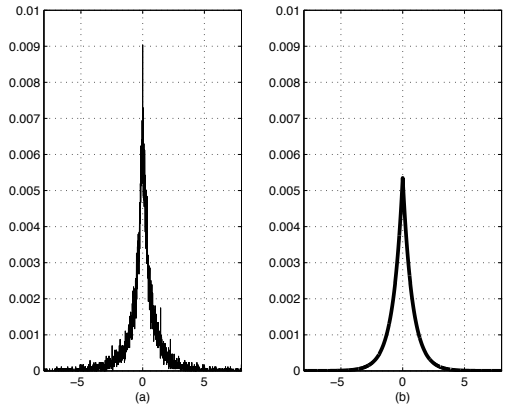

Figure 2: (a) The probability distribution of frequency deviation from ADC04 database (b) The probability density function modeled by the Laplace distribution. The x-axis is the frequency deviation in percentage.

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

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

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. EVALUATION

In this section, we present the evaluation of the effectiveness of the perceptual criteria. Firstly, the different peak selection methods are evaluated. Then, the system with/without perceptual criteria is evaluated. Finally, the performance is compared with that of MIREX participants. The databases used are listed below:

- ADC04: 20 excerpts of about 20s including MIDI, Jazz, Pop and Opera music as well as audio pieces with a synthesized voice. It is used for our training database [28].
- MIREX05: 25 excerpts of 10-40s from the following genres: Rock, R&B, Pop, Jazz, Solo classical piano [29]. Only 13 excerpts are made publicly available.
- RWC: 100 excerpts, 80 from Japanese hit charts in the 1990s and 20 from American hit charts in the 1980s [30]. This large database is rarely used in existing publications on melody estimation.

Peak selection

To evaluate the performance of different peak selection methods, we use two metrics: *recall rate* and *mean rank*. Recall rate is the

percentage of the correct melody F0 being extracted in the candidate set. A good peak selection method shall not exclude too many peaks that support the correct F0. Mean rank is the average score ranking of the correct melody F0 in the candidate set. As long as the dominant partials of the correct F0 are selected, the resulting score shall be high and the ranking of the correct F0 be on top. For the methods implying thresholds, several values are tested in search of the best configuration. The result is shown in Fig. 3. A good configuration shall result in a point located more to the top-right corner in the figure. The reasonable results obtained seem to locate in the region of which recall rate varies from 0.85 to 0.9 and mean rank varies from 2 to 1. In general, the perceptual criteria seem to be more effective than the spectral envelopes in favoring the correct F0s.

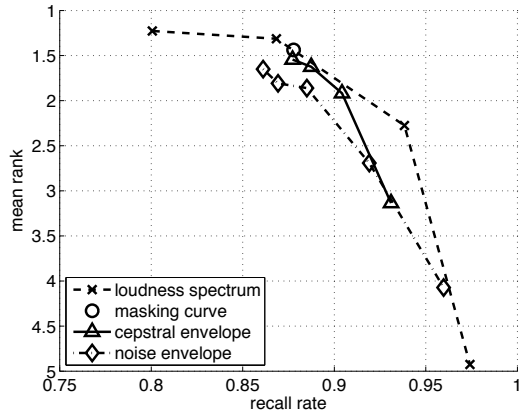

Figure 3: Evaluation results of different peak selection methods. The parameters tested are δ_L:(48,36,24,12), δ_C:(18,12,6,0) and δ_N:(12,9,6,3,0). The masking curve method does not involve any parameter and is shown as a single point.

System configurations

To understand the contribution of each component in the system, we propose to evaluate the system with different configurations. Since our current system does not detect if the melody is present (voiced) or not (unvoiced), we choose the following evaluation metrics [4]

$$\text{Raw Pitch Accuracy} = \frac{\text{number of correct estimates}}{\text{number of ground truth}} \qquad (10)$$

which is defined as the proportion of the voiced frames in which the estimated F0 is within one semitone of the ground truth.

The *baseline configuration* does not take into account any perceptual properties. The peak selection simply picks the first 20 largest peaks and the tracking does not use the envelope similarity measure ($\gamma = 0$). The *perceptual configuration* uses the loudness spectrum for peak selection, the envelope similarity compression factor $\gamma = 2.4$ and the emission probability compression factor $\beta = 0.1$. These parameters are trained from the data set ADC04. For each configuration, we further evaluate how the tracking mechanism improves the average raw pitch accuracy. The results without tracking simply reports the best candidate at each frame. The comparison is shown in Table 1. It is found that the perceptual configuration performs better than the baseline configuration by about 3 to 4%. The tracking mechanism slightly improve about 1 to 2%. Further investigation is ongoing to improve the tracking algorithm.

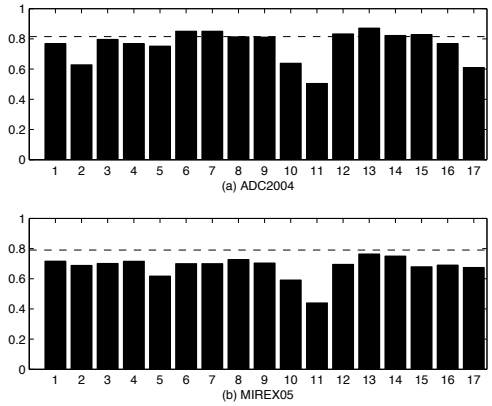

Figure 4: Raw pitch accuracy comparisons: (a) The MIREX participant results for ADC04 database (b) The MIREX participant results for MIREX05 database. The indices corresponding to MIREX participant IDs are: the first five for MIREX 2010 (HJ1, TOOS1, JJY2, JJY1, SG1) and the remaining twelve for MIREX 2009 (CL1, CL2, DR1, DR2, HJC1, HJC2, JJY, KD, MW, PC, RR, TOOS). Please refer to MIREX website for the respective systems [16]. The horizontal line shows the results of the proposed system.

	best candidate	candidates + tracking
Baseline config.	73.16%	74.03%
Perceptual config.	76.92%	78.10%

Table 1: Average raw pitch accuracy for baseline configuration(without perceptual properties) and perceptual configuration. For each configuration, the frame-based estimation (reporting the best candidate) is evaluated against the tracking system.

Comparison with the state-of-the-art system

Thanks to the MIREX campaign, the performance of the start-of-the-art systems are publicly evaluated (see Fig. 4). Although the MIREX database is only partially available for our evaluation, the results (see Table 2) still demonstrate its competitive performance among the top-ranked systems.

ADC04	MIREX05	RWC
81.53%	79.00%	74.49%

Table 2: Average raw pitch accuracy of proposed system evaluated on three databases.

5. CONCLUSION

The effectiveness of perceptual properties in the context of melody estimation has been studied. For the proposed melody estimation system, the accuracy is improved by more than 3% while taking into account perceptual properties. The use of either loudness or masking curve demonstrates advantages over the proposed spectral envelope features. The envelope similarity is found to slightly improve the accuracy, too. The proposed system is evaluated on more than one hundred excerpts of music recordings and demonstrates its competitive performance to the state-of-the-art systems. Future work will be the improvement of the tracking algorithm and the development of the voicing detection algorithm.

6. REFERENCES

[1] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, New York, 2006.

[2] 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 (ISCA Journal)*, vol. 43, no. 4, 2004.

[3] R. P. Paiva, T. Mendes, and A. Cardoso, "Melody detection in polyphonic musical signals: exploiting perceptual rules, note salience, and melodic smoothness," *Computer Music Journal*, vol. 30, no. 4, pp. 80–98, 2006.

[4] G. E. Poliner, D. P.W. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: approaches and evaluation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.

[5] A. S. Bregman, *Auditory Scene Analysis*, The MIT Press, Cambridge, Massachusetts, 1990.

[6] M. Marolt, "Audio melody extraction based on timbral similarity of melodic fragments," in *Proc. of Eurocon 2005*, 2005.

[7] J. Salamon and E. Gómez, "Melody extraction from polyphonic music audio," Music Information Retrieval Evaluation eXchange (MIREX) 2010.

[8] M. Marolt, "On finding melodic lines in audio recordings," in *Proc. of the Intl. Conf. on Digital Audio Effects (DAFx-04)*, 2004, pp. 217–221.

[9] K. Dressler, "Audio melody extraction for MIREX 2009," in *5th Music Information Retrieval Evaluation eXchange (MIREX'09)*, 2009.

[10] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[11] M. Ryynänen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. of the 7th Intl. Conf. on Music Information Retrieval (ISMIR'06)*, 2006.

[12] Y. Li and D.L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1475–1487, 2007.

[13] K. Dressler, "Extraction of the melody pitch contour from polyphonic audio," in *1st Music Information Retrieval Evaluation eXchange (MIREX'05)*, 2005.

[14] K. Dressler, "An auditory streaming approach on melody extraction," in *2nd Music Information Retrieval Evaluation eXchange (MIREX'06)*, 2006.

[15] K. Dressler, "Audio melody extraction - late breaking at ISMIR 2010," in *11th Intl. Conf. on Music Information Retrieval (ISMIR'10)*, 2010.

[16] "Music Information Retrieval Evaluation eXchange (MIREX) homepage," http://www.music-ir.org/mirex/wiki/.

[17] 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.

[18] C. Yeh, *Multiple fundamental frequency estimation of polyphonic recordings*, Ph.D. thesis, Université Paris 6, 2008.

[19] B. Bauer and E. Torick, "Researches in loudness measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 14, no. 3, pp. 141 – 151, 1966.

[20] H. Fletcher and W.A. Munson, "Loudness, its definition, measurement and calculation.," *Journal of the Acoustic Society of America*, vol. 5, pp. 82–108, 1933.

[21] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 314–323, 1988.

[22] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *Journal of the Acoustic Society of America*, vol. 33, no. 2, pp. 248–248, 1933.

[23] ISO/IEC 13818-3, "Information technology – generic coding of moving pictures and associated audio information – part 3: Audio," Tech. Rep., ISO/IEC JTC1/SC29 WG11, 1998.

[24] D. Schwarz and X. Rodet, *Analysis, Synthesis, and Perception of Musical Sounds*, chapter Spectral envelopes and additive + residual analysis/synthesis, pp. 175–227, Springer Science+Business Media, LLC, NY, USA, 2007.

[25] C. Yeh and A. Roebel, "Multipl-f0 estimation for MIREX 2010," Music Information Retrieval Evaluation eXchange (MIREX) 2010.

[26] C. M. Harris, "Pitch extraction by computer processing of high-resolution Fourier analysis data," *Journal of the Acoustical Society of America*, vol. 35, pp. 339–343, March 1963.

[27] A. Klapuri, *Signal Processing Methods For the Automatic Transcription of Music*, Ph.D. thesis, Tampere University of Technology, 2004.

[28] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack, "ISMIR 2004 audio description contest," Tech. Rep., UPF MTG, 2004.

[29] G. Poliner and D. Ellis, "A classification approach to melody transcription," in *11th Intl. Conf. on Music Information Retrieval (ISMIR'05)*, 2005.

[30] M. Goto, "AIST annotation for the RWC Music Database," in *Proc. of the 7th Intl. Conf. on Music Information Retrieval (ISMIR'06)*, 2006, pp. 359–360.