AN AUDITORY STREAMING APPROACH FOR MELODY EXTRACTION FROM POLYPHONIC MUSIC

Karin Dressler
Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany
kadressler@gmail.com

ABSTRACT

This paper proposes an efficient approach for the identification of the predominant voice from polyphonic musical audio. The algorithm implements an auditory streaming model which builds upon tone objects and salient pitches. The formation of voices is based on the regular update of the frequency and the magnitude of so called streaming agents, which aim at salient tones or pitches close to their preferred frequency range. Streaming agents which succeed to assemble a big magnitude start new voice objects, which in turn add adequate tones. The algorithm was evaluated as part of a melody extraction system during the MIREX audio melody extraction evaluation, where it gained very good results in the voicing detection and overall accuracy.

1. INTRODUCTION

Melody is defined as a linear succession of tones which is perceived as a single entity. One important characteristic of the tone sequence is the smoothness of the melody pitch contour. There are different techniques to avoid large frequency intervals in the tone sequence – at present two main algorithm types can be distinguished:

On the one hand, there are probabilistic frameworks that combine pitch salience values and smoothness constraints in a cost function that is evaluated by optimal path finding methods like the hidden Markov Model (HMM), the Viterbi algorithm or dynamic programming (DP). On the other hand, there are rule based approaches that trace multiple F0 contours over time using criteria like magnitude and pitch proximity in order to link salient pitch candidates of adjacent analysis frames. Subsequently, a melody line is formed from these tone-like pitch trajectories, using rules that take the necessary precautions to assure a smooth melodic contour. Of course such a division is rather artificial. It is easy to imagine a system that uses tone trajectories as input for a probabilistic framework. And vice versa a statistical approach can be used to model tones. In fact, Ryynänen and Klapuri have implemented a method for the automatic detection of singing melodies in polyphonic music, where they derive a HMM for note events from fundamental frequencies, their saliences and an accent signal [8].

There are many stable probabilistic relationships that can be observed in melody tone sequences [6]. This fact makes the application of a statistical model so useful, because such characteristics can easily be expressed mathematically in order to find the optimal succession of tones. Hence, most approaches to voice processing are statistical methods that accomplish the tone trajectory forming and the identification of the melody voice simultaneously [4, 5, 7]. Rao and Rao advocate DP over variants of partial and tone tracking, but also clearly state the problems of most statistical methods [7].

While for rule-based approaches alternative melody lines can be recovered quite easily, there is no effective possibility to retrieve alternative paths for DP approaches, because the mathematical optimization of the methods depends on the elimination of concurrent paths. Hence, it is not easy to state whether the most likely choice stands out from all other paths. This problem is most evident if two or more voices of comparable strength occur simultaneously within a musical piece. Work towards a solution to this problem was presented in [7], giving an example for DP with dual fundamental frequency tracking. The system tracks an ordered pair of two pitches, but it cannot ensure that the two contours will remain faithful to their respective sound sources.

Another challenging problem is the identification of non-voiced portions, e.g. frames where no melody voice occurs. The simultaneous identification of the optimal path together with the identification of melody frames is not easy to accomplish within one statistical model, so often the voicing detection is performed by a separate processing step. Nonetheless, optimal path finding algorithms may be confused by breaks in the tone sequence, especially because the usual transition probabilities do not apply in between melodic phrases.

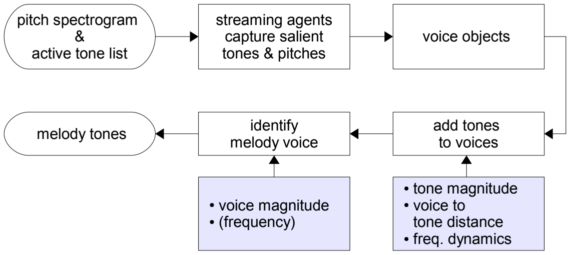

Figure 1. Overview of the voice estimation algorithm

In this paper, we present an algorithm for the identification of predominant voices in music that addresses some of the above-mentioned problems. Although no statistical model is implemented, probabilistic relationships that can be observed in melody tone sequences are exploited.

2. METHOD

2.1 Overview

Figure 2.1 shows an overview of the algorithm. The input to the proposed algorithm are the tone objects and/or salient pitches of the current frame. The formation of musical voices is a continuous process destined by the frame-wise evolution of so-called streaming agents, which are distributed along the frequency spectrum. A streaming agent gains power by the capturing of salient tones or pitches. Moreover, it changes its position in the frequency spectrum in order to move towards salient sounds. Voice objects can be derived from the streaming agents. Then, adequate tone objects are assigned to the respective voices. Finally, the melody voice is chosen from the set of voices. The main criterion for the selection is the magnitude of the voice. Only tone objects of the melody voice qualify as melody tones.

2.2 Formation of Streaming Agents

The voice detection is based on 18 streaming agents (SA). Each streaming agent denotes a very simple voice formation unit, which independently selects a succession of strong tones or pitches. It is mainly characterized by its magnitude $\bar{A}_{\text{sa}}$, and two frequency based measures: a variable position $\bar{f}_{\text{sa}}$ and a fixed home position $f_{\text{sa\_home}}$, which are both given in cent. The home positions of the streaming agents are distributed evenly with a distance of 300 cent over the allowed melody frequency range.

Over time, each streaming agents gradually moves towards the selected sound sources and assembles a magnitude corresponding to the rating magnitude of the captured tone objects.

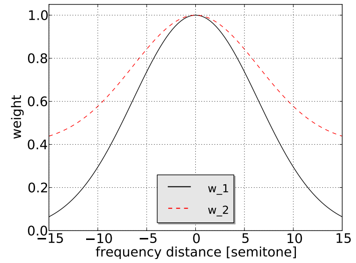

Figure 2. Gaussian Weighting Functions

2.2.1 Selection of Tones

In each analysis frame, the streaming agents searches for strong tones and pitches. In the further description, we refer only to tone objects, although the method can be also used for frame-wise estimated pitch magnitudes as described for example in [3]. For the identification of the best matching tone a rating is calculated from four criteria:

- *magnitude:* The tone magnitude A_{tone} is a good indicator for the perceptual importance of a tone.
- *frequency distance weight:* It is due to the fixed home position that each SA may pick different notes in a polyphonic signal. While at the one hand a strong selection criterion is the magnitude of the tone object, at the other hand the agent's choice is strongly biased towards its own home position. The frequency distance Δf in cent between the tone's pitch f_{tone} and the streaming agent's home position $f_{\text{sa\_home}}$ enters into the rating as a weighting factor that is calculated using a Gaussian function $w_1(\Delta f)$:

$$w_1(\Delta f) = e^{-0.5\frac{(\Delta f)^2}{640^2}} \qquad (1)$$

  Figure 2.2.1 shows the weighting function, which reaches half the maximum value at a frequency difference of approximately 750 cent.
- *frequency deviation:* Human listeners draw particular attention to all sounds with changing attributes. If a tone has a varying frequency deviation (persistently more than 20 cent frequency difference in between analysis frames) the rating is doubled. Accordingly, the deviation factor D is set to one or two in the final rating.
- *capture mode:* There should be a tendency of the SA to continuously track an already captured tone object. If a tone object has already been captured by a SA, the rating for the tone object is boosted by the factor $C = 1.5$. Otherwise, the factor is set to one. (See section 2.2.2 for a detailed explanation of the capture mode.)

The rating is estimated for all streaming agents and finally, each streaming agent "picks" only one tone – the object with the maximum rating magnitude A_{rating}:

$$A_{\text{rating}} = D \cdot C \cdot A_{\text{tone}} \cdot w_1(f_{\text{tone}} - f_{\text{sa\_home}}) \qquad (2)$$

For the rating of pitches the boost factors D and C are omitted – the rating is simply the product of pitch magnitude and the frequency distance weight.

2.2.2 *Modes of Tone Capturing*

As the streaming agent approaches salient sound sources, two different modes are distinguished within the tone capturing process: *aim* and *captured*. In the *aim* mode the streaming agent aims at a distinct tone object and moves slowly towards the selected pitch or tone.

In order to capture a tone, the SA must aim at the distinct tone for a specific time span. The demanded time depends on the difference between the variable position of the SA $\bar{f}_{\text{sa}}$ and the tone's frequency f_{tone} [1] . As long as the SA aims at the same tone object, a capture counter n is incremented in each analysis frame. The tone is captured if [2] :

$$n > \frac{1}{30} \left| \bar{f}_{\text{sa}} - f_{\text{tone}} \right| . \qquad (3)$$

As the SA moves towards the selected pitch, the frequency difference between tone and streaming agent becomes smaller during the capturing process. Since the adaptation speed of the variable position $\bar{f}_{\text{sa}}$ depends on many parameters, the duration needed to capture a tone cannot be immediately assessed from the frequency difference between successive notes. As soon as the SA aims at a sound object in a different frequency region, the capture counter is set to zero.

The mode *captured* might not be reached by every tone in very complex or noisy music signals. Yet, it is not necessary that a tone is captured by a streaming agent to qualify as a melody tone. The *aim* mode is generally sufficient to ensure the propagation of the streaming agents towards the most significant sound sources. Still, the additional mode enhances the movement of the streaming agent towards the selected tone objects.

2.2.3 *Magnitude Update*

The streaming agent is able to increase its magnitude $\bar{A}_{\text{sa}}$ whenever it reaches the capture mode *captured*. The magnitude it assembles depends on the current rating magnitude of the selected tone as given in equation 2, but without taking into account the boosting factor C. The slightly altered rating magnitude is labeled A^*_{rating}. The use of this rating magnitude implies that a streaming agent which captures a tone far away from the home positions will not build up a high magnitude.

However, for the computation of the magnitude, the initial rating A^*_{rating} is weighted by a second frequency distance weighting which exploits the distance between the variable position of the streaming agent $\bar{f}_{\text{sa}}$ and the tone's frequency f_{tone}. The weighting function remains the same: the Gaussian function w_1 given in equation 1. The additional weighting assures that the streaming agent profits more from tone magnitudes which are close to its current position $\bar{f}_{\text{sa}}$.

In order to update the magnitude values we use the exponential moving average (EMA) [3] :

$$\bar{A}_{\text{sa}} \rightarrow \alpha_{\text{x}} \cdot \bar{A}_{\text{sa}} + (1 - \alpha_{\text{x}}) \cdot A^*_{\text{rating}} \cdot w_1(f_{\text{tone}} - \bar{f}_{\text{sa}}). \qquad (4)$$

The start value for the iterative calculation of the EMA is zero. The smoothing factor α_{x} depends on the current weighted rating $A^*_{\text{rating}} \cdot w_1(f_{\text{tone}}, \bar{f}_{\text{sa}})$. If the current value is higher than the actual EMA value, α_{x} corresponds to a half life period of 1 second, otherwise the half life period is set to 500 ms. If the streaming agent is only in *aim* capture mode, the magnitude of the streaming agent is damped with a half life period of 500 ms.

2.2.4 *Position Update*

The streaming agent changes its variable position $\bar{f}_{\text{sa}}$ towards salient tones or pitches. The speed of the position adaptation is mainly determined by three factors:

- *the tone's magnitude:* the bigger the tone magnitude in comparison to the long term average weightings, the faster the SA changes its position.
- *the distance between captured tone and the streaming agent's home position:* the SA tends to move faster towards its own home position. This behavior ensures the stream segregation for a cycle of quickly alternating high and low tones as described in [1, chapter 2].
- *the frequency deviation:* the SA moves faster towards frequency modulated tones.
- *the capture mode*: the SA moves faster towards captured tones.

From this it follows that the basic weighing for the position update is similar to the rating magnitude A_{rating} for the tone selection process as given in equation 2. In order to

[1] All frequencies are measured in cent.

[2] The condition assumes a hop-size of 5.8 ms between two analysis frames.

[3] The EMA applies weighting factors to all previous data points which decrease exponentially, giving more importance to recent observations while still not discarding older observations entirely. The smoothing factor α determines the impact of past events on the actual EMA. It is a number between 0 and 1. A lower smoothing factor discards older results faster. A more intuitive measure than the smoothing factor is the so called half-life period. It denotes the time span over which the initial impact of an observation decreases by a factor of two. Taking into account the desired half-life t_h and the time period between two EMA calculations $\Delta t \approx 5.8$ms, the corresponding smoothing factor is calculated as follows: $\alpha = 0.5^{\Delta t / t_h}$.

estimate the significance of the current rating magnitude, it has to be set into relation with the ratings of previous analysis frames. That's why we introduce a position magnitude, which is the exponential moving average of previous ratings:

$$\bar{A}_{\text{pos}} \rightarrow \alpha_{1.5\text{s}} \cdot \bar{A}_{\text{pos}} + (1 - \alpha_{1.5\text{s}}) \cdot A_{\text{rating}}. \qquad (5)$$

In order to adapt the variable position $\bar{f}_{\text{sa}}$ of the streaming agent, the current rating is set into relation with the EMA of previous ratings:

$$\bar{f}_{\text{sa}} \rightarrow \frac{\bar{A}_{\text{pos}}\bar{f}_{\text{sa}} + (1 - \alpha_{500\text{ms}}) \cdot A_{\text{rating}} \cdot f_{\text{tone}}}{\bar{A}_{\text{pos}} + (1 - \alpha_{500\text{ms}}) \cdot A_{\text{rating}}}. \qquad (6)$$

The initial value for the iterative calculation is the home position $f_{\text{sa\_home}}$. Parameter $\alpha_{500\text{ms}}$ is a smoothing factor, which corresponds to a half life time of 500 ms [4] .

2.3 Formation of Voices

The positions and magnitudes of the 18 streaming agents are the foundation for the voice estimation. Figure 3 shows how the progress of the multiple streaming agents is influenced by salient tone objects. It can be noted that the approximate progression of musical voices is already suggested by the distribution of the streaming agents.

Each streaming agent which poses a local magnitude maximum is a candidate for the formation of a voice object. This means that each voice object is in general linked to a streaming agent with the peak magnitude compared to the magnitude of the neighboring agents. Of course, the local maximum may shift from one streaming agent to another. In this case, the voice may gradually change the assigned link to a neighboring streaming agent within the duration of approximately 20 analysis frames. The position of a voice f_{voice} is defined by the position $\bar{f}_{\text{sa}}$ of the linked streaming agent. The magnitude of the voice A_{voice} is defined by the streamer magnitude $\bar{A}_{\text{sa}}$. If the voice is currently adapting to a new streaming agent, weighted average values of the concerning two streaming agents are used.

If a streaming agent with a local maximum magnitude is not assigned to a voice object, it may start a new one. However, a new voice is created only if the streaming agent is more than 4 streaming agents away from a any streaming agent linked to another voice, or if the frequency difference between the streaming agent and all other existing voices is greater than 600 cent. A voice object is eliminated if the voice magnitude is smaller than 5 percent of the global maximum voice magnitude or if two voices aim at the same streaming agent. In the latter case the voice with the smaller magnitude is eliminated.

2.4 Adding Tones to Voices

Now that voice objects have been defined, adequate tone objects must be added. The only voice tone candidate is actually the currently selected tone of the corresponding streaming agent. If the voice is adapting to a new streaming agent, the closest streaming agent is used as a reference. Several measures are taken to ensure a reliable voicing detection. This means even if the corresponding streaming agent has selected a tone, the tone candidate has to be validated in order to qualify as a voice tone.

2.4.1 Distance Threshold

Although the proposed algorithm does not apply a common statistical model, it takes advantage of the most eminent probabilistic relationships in melodic tone sequences [6]: 1) Melodies consist typically of tones that are close to one another in pitch. 2) There is a strong tendency for a regression to the mean pitch.

The frequency of the voice represents the weighted average frequency of the recently selected tone objects, so in a way the voice position can be seen as the adaptive computation of the mean pitch. Consequently, the best voice tone candidates are close to the actual voice position. Adequate voice tones have to be within an octave range of the actual voice position. Another obvious thing to do would be the adjustment of the magnitude thresholds according to the frequency distance. This idea is implemented in the short term magnitude threshold described in section 2.4.3.

2.4.2 Global Long Term Magnitude Threshold

The global long term magnitude threshold is implemented as an adaptive threshold that is valid for all voices. It decays with a half life period of 5 seconds. If a tone magnitude appears which is larger than the current long term magnitude value, the magnitude threshold is updated to the new maximum.

The magnitude of the candidate voice tone is compared to the long term maximum value. In order to pass the global threshold, tones should not be more than 8 dB below the decaying maximum value. Still, other criteria may alter the effective threshold value – in the best case the allowed dynamic range is increased from 8 dB to 20 dB:

- The capture level of the assigned streaming agent and its two neighbors are evaluated. Depending on how many streaming agents are in capture mode *captured* concerning the candidate voice tone, the effective threshold may decreased to 14 dB below the decaying maximum. On the other hand, the threshold is increased for all tones that are not selected (aimed) by at least 5 streaming agents in the long term average.
- A variation of the fundamental frequency (vibrato or glides) increases the noticeability of tones. In this

[4] Since the position weight depends on many factors, parameter α does not exactly set any half life period for the position update. Yet the corresponding time span gives a reference point for the approximate adaptation speed.

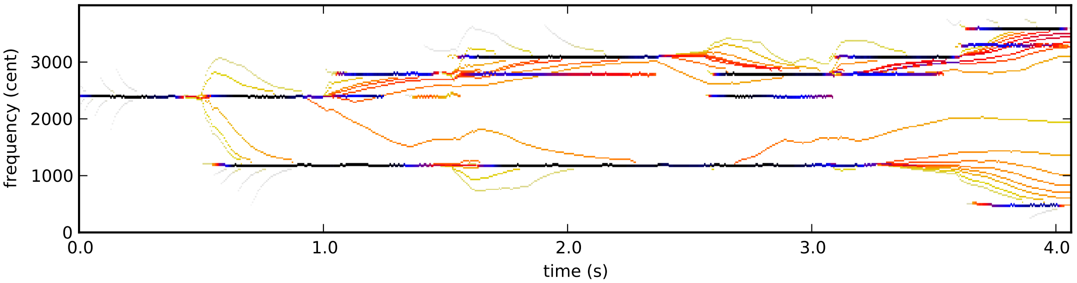

Figure 3. Streaming Agents: It can be seen how the streaming agents (thin lines) move towards salient tones and pitches. To maintain clarity salient pitches are not shown. The identified tone objects are indicated by dark bold lines. When the bass voice comes in, some streaming agents turn to the bass voice as it is closer to their preferred home position.

case the threshold is lowered by 6 dB.

2.4.3 Short Term Magnitude Threshold

The short term magnitude threshold is estimated separately for each voice. It secures that shortly after a strong tone is finished no weaker tone is included as a voice tone, so it is especially useful to bridge small time gaps between strong tones of a voice. Furthermore, the threshold delays the inclusion of tones that are far away from the current voice position. To achieve this the tone magnitude is again weighted with a frequency distance weight, evaluating the frequency offset between tone and voice:

$$w_2 = r + (1 - r) \cdot w_1(f_{\text{tone}} - f_{\text{voice}}). \qquad (7)$$

Figure 2.2.1 shows that the weighting function w_2 is asymmetric. Tones in the lower frequency range of an instrument or the voice are often softer. Hence, parameter r is set to 0.4 for tones with a lower frequency than the current voice position, otherwise $r = 0.2$.

The short term threshold is adaptive and decays with a half life time of 100 ms. If a weighted tone magnitude $w_2 \cdot A_{\text{tone}}$ appears which is larger than the current short term magnitude threshold, the threshold is updated to the new maximum. The tone passes the threshold if it is no more than 6 dB below the current threshold value.

2.5 The Identification of the Melody Voice

The most promising feature to distinguish melody tones from all other sounds is the magnitude. The magnitude of the tones is of course reflected by the voice magnitude. Hence, in general the voice with the highest magnitude is selected as the melody voice. It may happen that two ore more voices have about the same magnitude and thus no clear decision can be taken. In this case, the voices are weighted according to their frequency: voices in very low frequency regions receive a lower weight.

3. EVALUATION

3.1 Qualitative Evaluation

A striking advantage of the proposed method is its computational efficiency and the continuously updated voice information in real time. Moreover, the algorithm is flexible enough to track a variable number of concurrent voices. This is the main reason for the good melody detection accuracy for instrumental music excerpts with two or more strong voices like the one shown in figure 4.

The segregation of notes into different auditory streams depends on many aspects – like for example the magnitude, frequency and timbre of tones. Psychoacoustic experiments have shown that the grouping of tones also depends on the rate [1]. Due to the delayed capturing of tone objects, the presented method is able to take into account temporal aspects of the evolving signal. For example a series of alternating high and low tones will be integrated into one auditory stream at a low playback speed. Yet, with increasing rate high and low tones are grouped into individual voices.

Nonetheless, it must be noted that many aspects of human perception cannot be covered. Although the algorithm allows a broad dynamic range for melody tones, in some interpretations an even greater dynamic range can be found, especially if the melody is sung by a human. Still, by lowering the magnitude thresholds many tones from the accompaniment will be selected by mistake. A simple magnitude threshold cannot avoid all errors.

3.2 MIREX Audio Melody Extraction Task

The presented method for the detection of predominant voices has been implemented as part of a melody extraction algorithm which was evaluated at the Music Information Retrieval Evaluation eXchange (MIREX) [2]. Algorithm parameters regarding the width and the shape of the weighting functions as well as the timing constants of the adaptive thresholds have been adjusted using the melody extraction training data of ISMIR 2004 and MIREX 2005. Although the presented parameter sets maximize accuracy in the two

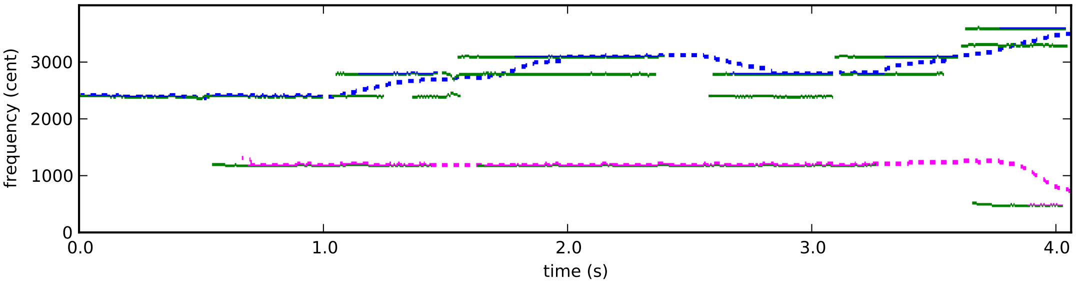

Figure 4. Voices: When the bass voice comes in, a second voice object is created. Two predominant voices are recognized: the melody voice (blue) and the bass voice (pink).

Algorithm	Voicing Recall (%)	Voicing False Alarm (%)	Raw Pitch (%)	Overall Accuracy (%)	Runtime (min)
proposed	90.9	41.0	80.6	73.4	24
dr1	92.4	51.7	74.4	66.9	23040
dr2	87.7	41.2	72.1	66.2	524
rr	91.3	51.1	72.2	65.2	26
pc	79.3	40.3	64.1	62.9	4677
jjy	61.0	29.4	73.3	56.6	3726
cl2	80.3	57.4	63.5	55.2	33
cl1	93.0	80.7	63.5	52.2	28
hjc1	43.6	9.7	66.1	50.5	344
hjc2	43.6	9.7	51.1	49.0	584

Figure 5. Melody Extraction Results of MIREX 2009

data sets, acceptable results are achieved on a wide parameter range. Moreover, the MIREX results show that the given settings generalize well on different kinds of data.

Table 5 shows the analysis results for systems that perform voicing detection. The melody extraction algorithm achieved the best overall accuracy and at the same time stands out due to very short run-times. The Raw Pitch measure represents the estimation performance for all voiced frames. For this measure the evaluation is constrained to time instants where the melody voice is present. The measure Overall Accuracy requires a voicing detection – the algorithm has to indicate whether the melody voice is present in the current frame or not. The MIREX results show that the implemented method allows a high Voicing Recall and at the same time a low Voicing False Alarm.

4. CONCLUSION

In this paper we presented an efficient approach to auditory stream segregation in polyphonic music. The MIREX results show that the proposed method allows a reliable identification of the predominant voice in different kinds of polyphonic music. The qualitative evaluation shows that the algorithm mimics some characteristics of stream segregation in the human auditory system, taking into account the magnitude of tones, note intervals and playback speed. However, timbral features are not exploited to group tones. In order to reach a higher accuracy an instrument/singing voice recognition is required.

5. REFERENCES

[1] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*, volume 1 MIT Press paperback. MIT Press, Cambridge, Mass., Sept. 1994.

[2] K. Dressler Audio Melody Extraction for MIREX 2009. In *5th Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

[3] K. Dressler Pitch estimation by the pair-wise evaluation of spectral peaks. In *AES 42nd Conference*, Ilmenau, Germany, July 2011.

[4] J.-L. Durrieu, G. Richard, B. David and C.Fvotte Source/Filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):564–575, Mar. 2010.

[5] C.-L. Hsu, L.-Y. Chen, J.-S. R. Jang, and H.-J. Li. Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement. In *Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, Oct. 2009.

[6] D. Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. The MIT Press, Cambridge, Massachusetts, 2006.

[7] V. Rao and P. Rao. Improving polyphonic melody extraction by dynamic programming based dual f0 tracking. In *Proc. of the 12th International Conference on Digital Audio Effects (DAFx)*, Como, Italy, Sept. 2009.

[8] M. Ryynänen and A. Klapuri. Transcription of the singing melody in polyphonic music. In *Proc. of the 7th International Society for Music Information Retrieval Conference (ISMIR)*, Victoria, Canada, Oct. 2006.