ON FINDING MELODIC LINES IN AUDIO RECORDINGS

Matija Marolt

Faculty of Computer and Information Science
University of Ljubljana, Slovenia
matija.marolt@fri.uni-lj.si

ABSTRACT

The paper presents our approach to the problem of finding melodic line(s) in polyphonic audio recordings. The approach is composed of two different stages, partially rooted in psychoacoustic theories of music perception: the first stage is dedicated to finding regions with strong and stable pitch (melodic fragments), while in the second stage, these fragments are grouped according to their properties (pitch, loudness...) into clusters which represent melodic lines of the piece. Expectation Maximization algorithm is used in both stages to find the dominant pitch in a region, and to train Gaussian Mixture Models that group fragments into melodies. The paper presents the entire process in more detail and provides some initial results.

1. INTRODUCTION

With the recent explosion of researches in computer music and especially in the field of music information retrieval, one of the problems that remain largely unsolved is the extraction of perceptually meaningful features from audio signals. By perceptually meaningful, we denote features that a typical listener can perceive while listening to a piece of music, and these include tempo and rhythm, melody, some form of harmonic structure, as well as the overall organisation of a piece.

It is clear that a set of tools that could handle these tasks well would be useful in a variety of applications that currently rely on symbolic (i.e. MIDI) as opposed to audio data. Such tools would bridge the gap between a large number of researches made on parametric (MIDI) data that amongst other include similarity measures, estimation of rhythm, GTTM decomposition and also query by example searching systems, where large musical databases could be made available, tagged with information extracted from audio. Audio analysis, learning and compositional systems could also make use of such information.

An overview of past researches shows that techniques for tempo tracking in audio signals are quite mature; several tools (i.e. [1]) are available for use, some of them work in real-time. Most have little problems with modern pop styles with small variations in tempo, while tracking an expressive piano performance usually still causes headaches to algorithms or their authors. Rhythmic organisation is already a harder problem, as it has more to do with higher level musical concepts, which are harder to represent [2]. A promising approach to finding harmonic structure in audio signals has been presented by Sheh and Ellis [3].

Our paper deals with extraction of melodic lines from audio recordings. The field has been extensively studied for monophonic signals, where many approaches exist (i.e. [4, 5]). For polyphonic signals, the work of several groups is dedicated to complete transcription of audio signals, with the final result being a score that represents the original audio ([6, 7, 8]). Algorithms for simplified transcriptions, like extraction of melody, have been studied by few, with the notable exception of the work done by Goto [9].

Our work builds on ideas proposed by Goto with the goal of producing a tool for extraction of melodic lines from audio recordings. The approach includes extraction of sinusoidal components from the original audio signal, EM estimation of predominant pitches, their grouping into melodic fragments and final clustering of melodic fragments into melodic lines. The paper briefly describes each of these stages and presents some preliminary results.

2. DISCOVERING MELODIC FRAGMENTS

Our approach to finding melodic lines begins with discovery of fragments that a melodic line is composed of – melodic fragments. Melodic fragments are defined as regions of the signal, that exhibit a strong and stable pitch. Pitch is the main attribute according to which fragments are discovered; other features, such as loudness or timbre, are not taken into consideration. They come into picture when fragments are merged into melodic lines according to their similarity.

2.1. SMS analysis

To locate melodic fragments, we initially need to estimate the predominant pitch(es) in the input signal. To achieve that, we first separate the slowly-varying sinusoidal components (partials) of the signal from the rest (transients and noise) by the well known spectral modelling synthesis approach (SMS, [10]). SMS analysis transforms the signal into a set of sinusoidal components with time-varying frequencies and amplitudes, and a residual signal, obtained by subtracting the sines from the original signal. We used the publicly available SMSTools software (http://www.iua.upf.es/mtg/clam) to analyse our songs with a 100 ms Blackman-Harris window, 10 ms hop size. Non-harmonic style of analysis was chosen, as our signals are generally polyphonic and not necessary harmonic (drums...).

2.2. Masking

The obtained sinusoidal components are subjected to a psychoacoustic masking model that eliminates the components masked by other, stronger ones. Only simultaneous masking within critical bands is taken into consideration – temporal masking is ignored. Tonal and noise maskers are calculated from the set of sinusoidal components and the residual signal, as described in [11], and components that fall below the global masking threshold removed. The masking procedure is mainly used to reduce the computational load

of predominant pitch estimation, as it on average halves the maximal number of sinusoidal components (to approx. 60 per frame).

2.3. Predominant pitch estimation

After the sinusoid components have been extracted, and masking applied, we estimate the predominant pitch(es) in short (50 ms) segments of the signal. Our pitch estimating procedure is based on the *PreFEst* approach introduced by Goto [9], with some modifications.

The method is based on the Expectation-Maximisation (EM) algorithm, which treats the set of sinusoidal components at each time instant as a probability density function (observed PDF), which is considered to be generated from a weighted mixture of tone models of all possible pitches at this time instant. A tone model is defined as a PDF, corresponding to a typical structure of a harmonic tone (fundamental frequency + overtones). The EM algorithm iteratively estimates the weights of all tone models, while searching for one that maximizes the observed PDF. Consequently, each tone model weight represents the dominance of the tone model and thereby the dominance of the tone model's pitch in the observed PDF.

Our modified iterative EM procedure is summarized as follows. At a given time instant t SMS provides us with a set of sinusoidal components with frequencies $F^{(t)}$ and amplitudes $A^{(t)}$.

Our observed state $O^{(t,n)}$ is represented by a set of sinusoids in the time interval $[t,t+n]$:

$$O^{(t,n)} = \left\{ F^{(t,n)}, A^{(t,n)} \right\}$$
$$F^{(t,n)} = \left[F^{(t)}, ..., F^{(t+n-1)} \right] \quad A^{(t,n)} = \left[A^{(t)}, ..., A^{(t+n-1)} \right] \tag{1}$$

The observed state $O^{(t,n)}$ is considered to be generated by a model $p^{(t)}$, which is a weighted sum of tone models M of all possible pitches $G^{(t)}$:

$$p^{(t)}(F^{(t,n)}) = \sum_{g \in G^{(t)}} w^{(t)}(g) M\left(F^{(t,n)}, g, C^{(t)}(g)\right) \tag{2}$$

The set of possible tone model pitches $G^{(t)}$ is derived from frequencies of sinusoidal components $F^{(t,n)}$, by encompassing all frequencies below 4200 Hz, and adding the frequencies of the first and second subharmonic components of each pitch, to account for missing fundamentals.

A tone model M with pitch f can be described as:

$$M\left(F^{(t,n)}, g, C(g)\right) = \sum_{h=1}^{H} m\left(F^{(t,n)}, g, h, C(g)\right)$$
$$m\left(F^{(t,n)}, g, h, C(g)\right) = \frac{c(h,g)\, G(F^{(t,n)}, hg, \sigma_h)}{norm\left(F^{(t,n)}, hg, \sigma_h\right)}$$
$$norm(F^{(t,n)}, f, \sigma) = \begin{cases} G(f,f,\sigma); \sum_{x \in F^{(t,n)}} G(x,f,\sigma) < nG(f,f,\sigma) \\ \sum_{x \in F^{(t,n)}} G(x,f,\sigma); \ otherwise \end{cases} \tag{3}$$
$$C(g) = \{\, c(h,g) \mid h = 1...H \,\}$$

$C(g)$ represents a set of relative amplitudes $c(h,g)$ of individual harmonics (1..h) in the tone model with frequency g and $G(x,\mu,\sigma)$ Gaussian distribution with mean μ and variance σ. The idea behind the normalization function *norm* lies in psychoacoustic models of loudness perception. The function serves as a limiter that limits the contribution of closely-spaced sinusoidal components, occurring when several strong components fall within the width of a Gaussian, representing a tone model component. In this case, the function limits the sum of contributions of all components, which in a simplified way mimics the effects that distance between frequency components plays in the perception of loudness [12].

The process is illustrated in Fig. 1, where a tone model with pitch 329 Hz is applied to a series of partials found by the SMS algorithm. The model acts as a sieve, picking and summing up contributions of individual partials that would fit into a tone with a pitch of 329 Hz. Only the first six tone model partials are shown.

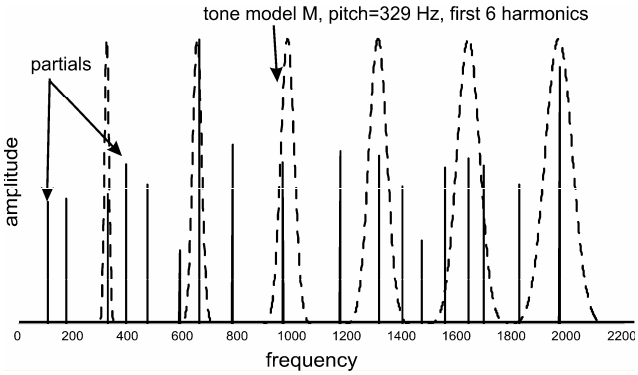

Figure 1: *Applying a tone model on a set of partials*

The weights w of all possible tone models (eq. 2) and amplitudes of their harmonics (c), are iteratively calculated by the EM algorithm:

$$\overline{w^{(t)}(g)} = \sum_{\{f,a\} \in O^{(t+n)}} \frac{w^{(t)}(g)\, a M\left(f, g, C^{(t)}(g)\right)}{\sum_{\bar{g} \in G^{(t)}} M\left(f, \bar{g}, C^{(t)}(g)\right)} \tag{4}$$

$$\overline{c^{(t)}(h,g)} = \sum_{\{f,a\} \in O^{(t,n)}} \frac{w^{(t)}(g)\, a m\left(f, g, h, C^{(t)}(g)\right)}{\sum_{\bar{g} \in G^{(t)}} M\left(f, \bar{g}, C^{(t)}(g)\right)} \tag{5}$$

When the iterative algorithm converges, the pitch of the tone model with the highest weight w is taken to be the predominant pitch. We use early stopping to stop the convergence prematurely and take the first few highest weights to represent the predominant pitches in the time window under consideration. These are later tracked and grouped into melodic fragments.

In the beginning, all tone model weights and amplitudes are initialized to the same value. Tone models contain a maximum of 20 harmonics, values of σ_h range between 50 cents (1st harmonic) to 100 cents (20th harmonic). After some experiments, the value of n, representing the width of the analysis window, was set to 5, thereby encompassing a time interval of 50 ms. This significantly reduced the effects of "noisy" partials, found by SMS analysis, on estimation of predominant pitch.

The effect can be seen in Fig. 2, representing the outcome of the EM algorithm on a short fragment from Aretha Franklin's interpretation of song Respect. Both figures show the distribution of tone model weights (predominant pitches) through time. The left side of

the figure shows results obtained by using individual time frames produced by the SMS analysis (10 ms) to calculate tone model weights, while in the figure on the right, 5 frames of SMS output (50 ms) were taken to calculate the weights. It is clear that by using a larger window, melodic fragments in the noisier sections stand out much clearer.

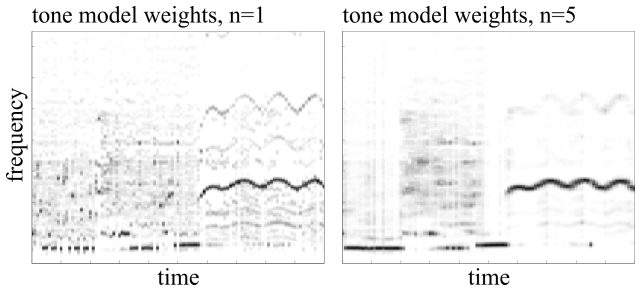

Figure 2: *Effect of window size n on the EM algorithm for predominant pitch estimation*

2.4. Forming melodic fragments

Weights produced by the EM algorithm indicate the pitches that are dominant at each time instance. Melodic fragments are formed by tracking dominant pitches through time and thereby forming fragments that have continuous pitch contours. The first part of the procedure is similar to pitch salience calculation as described by Goto [13]. For each pitch with weight greater than a dynamically adjusted threshold, salience is calculated according to its dominance in a 50 ms look-ahead window. The procedure tolerates pitch deviations of up to 100 cents per 10 ms window and also tolerates individual noisy frames that might corrupt pitch tracks by looking at the contents of the entire 50 ms window.

After saliences are calculated, grouping into melodic fragments is performed by continuously tracking the top three salient peaks and producing fragments along the way as follows:

- the procedure ignores all time instances, where total loudness of the signal, calculated according to Zwicker's loudness model [12] falls below a set threshold;
- the initial set of melodic fragments F is empty; the initial set of candidate melodic fragments C is empty;
- the following operations are repeated:
  - in each time instance t, select the top three salient peaks that differ from each other by more than 200 cents and find their exact frequencies f_i, according to the largest weight w_i in the neighbourhood:
  - in the set of candidate fragments C, find a fragment c with average frequency closest to f_i
    - if the difference in frequencies between c and f_i is smaller than 200 cents, add f_i to the current candidate fragment;
    - otherwise, start a new candidate fragment
  - after the top three pitches at time t have been processed, find all candidate fragments, that have not been extended during the last 50 ms. If their length exceeds 50 ms, add them to the set of melodic fragments F and remove them from the set of candidates C. If their length is shorter than 50 ms, remove them from C.
- after the signal has been processed, merge harmonically related melodic fragments, appearing at the same time (only 1st and 2nd overtones are taken into consideration) and join continuous fragments (in time and frequency).

The final result of this simple procedure is a set of melodic fragments, which may overlap in time, are at least 50 ms long and may have a slowly changing pitch. Parameters of each fragment are its start and end time, its time-varying pitch and its time-varying loudness. The fragments obtained provide a reasonable segmentation of the input signal into regions with stable dominant pitch. An example is given in Fig. 3, which shows segmentation obtained on a 5.5 seconds excerpt from Aretha Franklin's interpretation of the song Respect. 25 fragments were obtained; six belong to the melody sung by the singer, while the majority of others belong to different parts of the arrangement, which become dominant when lead vocals are out of the picture. Additionally, three noisy fragments were found, which were either due to consonants or drum parts. These can usually be dealt with in the last part of the procedure, where fragments are merged into melodic lines.

We performed informal subjective listening tests by resynthesizing the fragments (on the basis of their pitch and amplitude) and comparing these resynthesized versions with the original signal covering the same time spans. Most of the fragments perfectly captured the dominant pitch in the areas, even if, while listening to the entire original signal, some of the fragments found were not immediately obvious to the listener (i.e. organ parts in the given example). We carried out such tests on a set of excerpts from 10 different songs, covering a variety of styles, from jazz, pop/rock to dance, and the overall performance of the algorithm for finding melodic fragments was found to be satisfying; it discovered a large majority of fragments belonging to the lead melody, which is the main point of interest in this study.

3. FORMING MELODIC LINES

The goal of our project is to extract one or more melodic lines from an audio recording. How is a melodic line, or melody, defined? There are many definitions; Levitin describes melody as an auditory object that maintains its identity under certain transformations along

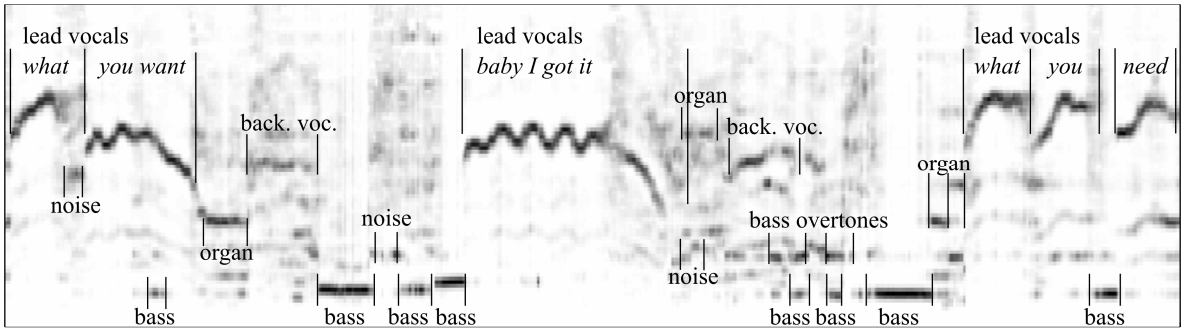

Figure 3: *Segmentation into melodic fragments of an excerpt from Otis Redding's song Respect sung by Aretha Franklin*

the six dimensions of pitch, tempo, timbre, loudness, spatial location, and reverberant environment; sometimes with changes in rhythm; but rarely with changes in contour [14]. Not only that melodies maintain their identity under such transformations, or rather because of that, melodies themselves are usually (at least locally in time) composed of events that themselves are similar in pitch, tempo, timbre, loudness, etc.

The fact becomes useful when we need to group melodic fragments, like the ones obtained by the procedure described before, into melodic lines. In fact, the process of discovering melodic lines becomes one of grouping melodic fragments through time into melodies. Fragments are grouped according to their properties. Ideally, one would make use of properties, which accurately describe the six dimensions mentioned before, especially pitch, timbre, loudness and tempo. Out of these, timbre is the most difficult to model; we are not aware of studies that would reliably determine the timbre of predominant voices in polyphonic audio recordings. Many studies, however, make use of timbre related features, when comparing pieces according to their similarity, classifying music according to genre, identifying the singer, etc. (i.e. [15], [16]). The features used in these studies could be applied to our problem, but so far we have not yet made such attempts. To group fragments into melodies, we currently make use of only four features, which represent:

- *pitch* as the centroid of fragment's frequency with regard to its dominance;
- *loudness* as the mean value of the product of dominance and loudness. Loudness is calculated according to Zwicker's loudness model [12] for partials belonging to the fragment. The product of dominance and loudness seems to give better results than if loudness alone would be taken;
- *pitch stability* as the average change of pitch over successive time instances. This could be classified as the only timbral feature used and mostly separates vocal parts from stable instruments;
- *onset steepness* as the steepness of overall loudness change during the first 50 ms of the fragment's start. The feature penalizes fragments that come into picture when a louder sound stops.

To group melodic fragments into melodies, we use a modified Gaussian mixture model estimation procedure, which makes use of equivalence constraints during the EM phase of model estimation [17]. Gaussian Mixture Models (GMMs) are one of the more widely used methods for unsupervised clustering of data, where clusters are approximated by Gaussian distributions, fitted on the provided data. Equivalence constraints are prior knowledge concerning pairs of data points, indicating if the points arise from the same source (belong to the same cluster - positive constraint) or from different sources (different clusters - negative constraint). They provide additional information to the GMM training algorithm, and are very useful in our domain. We use GMMs to cluster melodic fragments into melodies according to their properties. Additionally, we make use of two facts to automatically construct positive and negative equivalence constraints between fragments.

Fragments may overlap in time, as can be seen in Fig. 2. We treat melody as a succession of single notes (pitches). Therefore, we can put negative equivalence constraints on all pairs of fragments that overlap in time. This forbids the training algorithm to put two overlapping fragments into the same cluster and thus the same melodic line. We also give special treatment to the bass line, which may appear quite often in melodic fragments (Fig. 2). To help the training algorithm with bass line clustering, we also put positive equivalence constraints on all fragments with pitch lower than 170 Hz. This does not mean that the training algorithm will not add additional fragments to this cluster; it just causes all low pitched fragments to be grouped together.

The clustering procedure currently only works on entire song fragments (or entire songs), and we are still working on a version that will work within an approx. 5 second long sliding window and dynamically add new fragments to existing clusters or form new clusters as it progresses through a given piece.

We have not yet made any extensive tests of the accuracy of our melody extracting procedure. This is mainly due to the lack of a larger annotated collection of songs that could be used to automatically measure the accuracy of the approach. We have tested the algorithm on a number of examples and are overall satisfied with the performance of the fragment-extracting procedure, and less so with the performance of GMM clustering. GMMs may work perfectly in some cases, like Aretha Franklin's example used for this paper, while for others, problems may occur mainly because fragments belonging to accompanying instruments, which appear close to the lead melodic line are taken to be part of the line.

Results of clustering on a 30 second excerpt of Otis Redding's song Respect, as sung by Aretha Franklin, are given in Table 1.

	lead vocal	back vocals	bass	guitar	brass	keys	noise
C1	0.03	0.24	0.03	0	0.1	0.33	0.35
C2	**0.93**	0.29	0	0	0.1	0	0.05
C3	0.03	0.38	0	0.33	0.3	0	0.3
C4	0	0	0.97	0	0.05	0.33	0.08
C5	0	0.1	0	0.67	0.45	0.33	0.22

Table 1: *GMM clustering of fragments from "Respect"*

152 melodic fragments were found by the fragment finding procedure; all lead vocal and backing vocal parts were correctly discovered. All fragments were hand annotated into one of seven categories (lead vocal, backing vocals, bass, guitar, brass, keyboards, noise). Fragments were then clustered by the GMM algorithm into five clusters, which would ideally represent the melody (lead vocal), bass line, backing vocals, accompaniment and noise.

Results of the clustering procedure are given in Table 1. It shows percentages of fragments belonging to the seven annotated categories in the five clusters. Ideally, lead vocal fragments (melody) would all be grouped into one cluster with no additional fragments. Most (93%) were indeed grouped into cluster 2, but the cluster also contains some other fragments, belonging to backing vocals, brass and a small amount of noise. The majority of bass fragments were put into cluster 4, together with some low pitched keyboard parts, while other clusters contain a mixture of accompaniment and backing vocals. As our goal lies mainly in the discovery of the (main) melodic line, results are satisfying, especially if we take into consideration that practically no timbre based features were taken into consideration when clustering. Most of the melody is represented by fragments in cluster 2, with some additional backing vocal fragments, which could actually also be perceived as part of the melody.

The effect of negative and positive constraints on the clustering procedure was also assessed; somewhat surprisingly, constraints did not have a large impact on the clustering procedure. Small improvements were achieved mostly in separation of accompaniment from lead vocal and bass lines.

4. CONCLUSIONS

The presented approach to melody extraction is still in an initial phase, but we are satisfied with first obtained results. Currently, we are in the process of annotating a larger number of pieces, which will be used for improving the feature set used in GMM training, as so far, we settled for a very small number of parameters, mainly because of the small set of examples we worked with. We plan to concentrate on timbral features, which are expected to bring improvements, especially with mismatches in parts where accompaniment becomes dominant. The larger database will also enable us to test and compare several different clustering strategies.

5. REFERENCES

[1] S. Dixon, "Automatic Extraction of Tempo and Beat from Expressive Performances", *Journal of New Music Research*, 30 (1), pp 39-58, 2001.

[2] J. Seppänen, "Tatum grid analysis of musical signals", in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 21--24, 2001.

[3] A. Sheh and D.P.W. Ellis, "Chord Segmentation and Recognition using EM-Trained Hidden Markov Models", in Proceedings of ISMIR 2003, Baltimore, Maryland, USA, 2003.

[4] T. De Mulder, J.P. Martens, M. Lesaffre, M. Leman, B. De Baets, H. De Meyer, "An Auditory Model Based Transcriber of Vocal Queries", in Proceedings of ISMIR 2003, Baltimore, Maryland, USA, October 26-30, 2003.

[5] T. Heinz, A. Brueckmann, "Using a Physiological Ear Model for Automatic Melody Transcription and Sound Source Recognition", in *114th AES Convention 2003,* Amsterdam, The Netherlands, 2003.

[6] Klapuri, A, "Automatic transcription of music," in *Proceedings Stockholm Music Acoustics Conference*, Stockholm, Sweden, Aug. 6-9, 2003.

[7] Marolt M, "Networks of Adaptive Oscillators for Partial Tracking and Transcription of Music Recordings," *Journal of New Music Research*, 33 (1), 2004.

[8] J.P. Bello, *Towards the Automated Analysis of simple polyphonic music: A knowledge-based approach*, Ph.D. Thesis, King's College London - Queen Mary, University of London, 2003.

[9] M. Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models", in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.V-3365-3368, May 2001.

[10] X. Serra and J. O. Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition", *Computer Music Journal* 14(4), pp. 14–24, 1990.

[11] T. Painter, A. Spanias, "Perceptual Coding of Digital Audio", *Proceedings of the IEEE*, Vol. 88 (4), 2000.

[12] E. Zwicker, H. Fastl, *Psychoacoustics: Facts and Models,* Berlin: Springer Verlag, 1999.

[13] M. Goto and S. Hayamizu, "A Real-time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals", *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pp.31-40, August 1999.

[14] D.J. Levitin, "Memory for Musical Attributes from Music", *Cognition, and Computerized Sound*, Perry R. Cook (ed.). Cambridge, MA: MIT Press, 1999.

[15] J.-J. Aucouturier and F. Pachet, "Representing Musical Genre: A State of the Art", *Journal of New Music Research,* Vol. 32, No. 1, pp. 83–93, 2003.

[16] T. Zhang, "Automatic singer identification", *Proceedings of IEEE Conference on Multimedia and Expo*, vol.1, pp.33-36, Baltimore, July 6-9, 2003.

[17] N. Shental, A.B. Hillel, T. Hertz, D. Weinshall, "Computing Gaussian Mixture Models with EM using Side-Information", in *Proceedings of International Conference on Machine Learning, ICML-03,* Washington DC, August 2003.