Audio Melody Extraction Based on Timbral Similarity of Melodic Fragments

Matija Marolt, *Member, IEEE*

***Abstract* — The presented study deals with extraction of melodic line(s) from polyphonic audio recordings. Our approach is based on finding significant melodic fragments throughout the analyzed piece of music and clustering these fragments according to their timbral similarity. Fragments within clusters are taken to represent fragments belonging to different melodic lines. Holes between significant fragments within each cluster are filled-in by a shortest-path approach over all melodic fragments. The paper presents our study in more detail and provides results on real recordings.**

***Keywords* — audio melody extraction, music information retrieval, melodic fragments.**

I. Introduction

ONE of the problems that remain largely unsolved in current computer music researches 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, such as tempo, rhythm, melody, some form of harmonic structure, as well as the overall organization of a piece.

A set of tools that could handle these tasks well would provide good grounds for construction of large annotated musical audio databases. The lack of such data currently represents a major drawback for the computer music community, as it is very difficult to make use of a large variety of machine learning algorithms (requiring large amounts of annotated data) or make any kind of large scale evaluations of various music information retrieval (MIR) approaches on real-world data. It would also bridge the gap between a large number of researches made on parametric (MIDI) data that amongst other include similarity measures, estimation of rhythm or GTTM decomposition. Audio analysis, learning and compositional systems could also make use of such information.

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. [1]). 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 ([2, 3, 4]). Recently, there has been a growing number of studies into algorithms for simplified transcriptions, like extraction of melody [5,13,14].

Our work builds on ideas proposed by Goto with the goal of producing a tool for extraction of melodic lines from audio recordings. The paper describes each phase of our approach and presents some results on real recordings.

Matija Marolt is with the Faculty of Computer and Information Science, University of Ljubljana, Trzaska 25, 1000 Ljubljana, Slovenia; (e-mail: matija.marolt@fri.uni-lj.si).

II. Finding Melodic Fragments

The extraction of 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 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.

A. Spectral Modeling Synthesis and psychoacoustic masking

We first separate the slowly-varying sinusoidal components (partials) of the signal from the rest (transients and noise) by the well known spectral modeling synthesis approach (SMS, [6]). 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.

The obtained sinusoidal components are subjected to a psychoacoustic masking model that eliminates the components masked by stronger ones. Only simultaneous masking is taken into consideration – temporal masking is ignored [7]. On average, the masking procedure halves the total number of sinusoidal components.

B. Predominant pitch estimation

After the sinusoidal components have been extracted, and masking applied, we estimate the predominant pitch(es) in short segments of the signal. Our pitch estimating procedure is based on the PreFEst approach introduced by Goto [5], with some modifications. The method employs the Expectation-Maximization (EM) algorithm, which treats the set of sinusoidal components within a short time window as a probability density function, which is considered to be generated from a weighted mixture of tone models of all possible pitches at this time interval. The EM algorithm iteratively estimates the weights of all possible tone models, while searching for one that maximizes the set of sinusoidal components within the chosen time window. Consequently, each tone model weight represents the dominance of the tone model and thereby the dominance of the tone model's pitch.

C. Melodic fragments

Weights produced by the EM algorithm indicate the dominant pitches in short regions of time across the signal. Melodic fragments are formed by tracking the dominant pitches through time and thereby forming fragments with continuous pitch contours (loudness or other factors are not taken into consideration). The first part of the procedure is similar to pitch salience calculation as described by Goto [5]. 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 and individual noisy frames that might corrupt pitch tracks by looking at the contents of the entire 50 ms window.

After saliences are calculated, melodic fragments are formed by continuously tracking the dominant salient peaks and producing fragments along the way. 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 slowly changing pitches. 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.

Most errors of the fragment finding procedure occur in areas in which several competing melodic lines with similar loudnesses compete for listener's attention. These are problematic, as the EM algorithm tends to switch between lines thereby producing series of broken fragments. Sometimes, such switching also appears between a line and its octave equivalent, which is highly undesirable. Strong vibrato is also a cause of many fragment-tracking errors, as well as masking caused by loud, especially low-pitched or noisy sounds (drums, bass).

III. Forming MElodic Lines

The goal of our work 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 the six dimensions of pitch, tempo, timbre, loudness, spatial location, and reverberant environment; sometimes with changes in rhythm; but rarely with changes in contour [9]. 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 found 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.

Our approach to finding the melodic line is divided into several phases:

1. regions with strong harmonic contents are identified (we call them significant fragments);
2. significant fragments are clustered into several groups according to their similarity;
3. directed acyclic graphs (DAGs) are formed from closely spaced significant fragments from the same cluster, which results in formation of larger melodic regions (groups of significant fragments);
4. empty space between linked pairs of significant fragments is filled by finding best paths in DAGs created from closely spaced fragments inbetween. This results in a finer definition of melody between significant fragments;
5. conflicting paths in groups of significant fragments are resolved by finding shortest path through each group. This defines melody within groups of significant fragments;
6. the dominating cluster is found and main melodic path decoded from paths within groups of significant fragments belonging to the cluster.

We describe these steps in more detail in the following sections

A. Finding and clustering significant fragments

To cluster melodic fragments according to their timbral features, we first find fragments, for which the features can be reliably estimated – significant fragments. These are located according to their relative loudness, which is calculated as:

$$l_f^r = \sum_{t \in t_f} \frac{l_f(t) w_f(t)}{L(t)} \quad (1)$$

where $l_f(t)$ is loudness of fragment f at time t calculated by Zwicker's loudness model [8] for partials belonging to the fragment, $w_f(t)$ weight of the tone model that originated the fragment and $L(t)$ overall loudness of the signal. The significant fragments are picked by first using Grubbs' test for eliminating outliers in relative loudness and then picking fragments with relative loudness above one standard deviation from the mean of fragments without outliers. The obtained set of fragments is subjected to another sieve, which removes all fragments with loudness smaller than one standard deviation below the mean loudness of fragments. We call fragments in the resulting set significant fragments, as they represent regions with strongly defined melody.

Significant fragments are then clustered according to their similarity. To calculate fragment similarity, we trained a feedfoward neural network on a set of examples from ISMIR2004 melody extraction competition [11]. Inputs of the network consist of a set of 20 features describing musical properties of fragments and include:

- mean frequency;
- dominance: average weight of the tone model that originated the fragment, as calculated by the EM procedure;
- mean loudness
- percentages of fragment length covered by fragments an octave above and below
- ratio of first three even to odd harmonics
- spectral centroid and bandwidth of all partials of the fragment
- spectral irregularity and ratio of even to all harmonics [12]
- several cepstral coefficients
- tristimulus and inharmonicity [12]

These features were picked out of a larger set of features in a training/evaluation procedure for the network. The trained network is used to calculate similarity between all significant fragments and the obtained similarity matrix is used as the basis for clustering.

Clustering is performed with k-means algorithm; fragments are clustered into two to five clusters. The optimal number of clusters is then picked according to the silhouette criterion. Finally, significant fragments that lie on cluster borders are assigned to all the neighboring clusters as well, so that one fragment may belong to several clusters. The results of the procedure are illustrated in parts A and B of Figure 1.

B. Forming melodic lines

Significant fragments within each cluster are linked into directed acyclic graphs according to fragment proximity in time and frequency. Graphs represent groups of significant fragments that are close in time, frequency and timbral similarity, as all nodes in a graph belong to the same cluster. Such groups of significant fragments form a rough approximation of larger chunks of melodic lines within a piece (see Figure 1C).

After groups of significant fragments are formed, each linked pair of significant fragments within each group is connected with a new directed acyclic graph through all the fragments in-between the two significant fragments. Graph edges are formed according to time and frequency proximity of fragments. Weights of edges are calculated according to a linear combination of:

- frequency difference between nodes;
- loudness difference between nodes;
- time gap between nodes;
- penalty for interrupting a continuous sequence of fragments;
- timbral similarity to all significant fragments within the group;
- timbral similarity between the two nodes.

The features and their respective weights within the linear combination were obtained by optimizing the entire melody extraction procedure with a genetic algorithm on the ISMIR 2004 dataset.

Shortest path algorithm is then applied to each DAG between a pair of significant fragments to find the optimal path through all fragments between the two nodes. This results in a finer definition of melody between two significant fragments and is illustrated in Figure 1D.

Cost of the shortest path between two significant fragments is taken as the cost of the path between the two fragments in the DAG defining the group that the two fragments belong to. Shortest path algorithm is again applied to each group of significant fragments to remove the ambiguities that may arise due to competing paths within groups. This results in a clear definition of melody within each group of significant fragments (see Figure 1E).

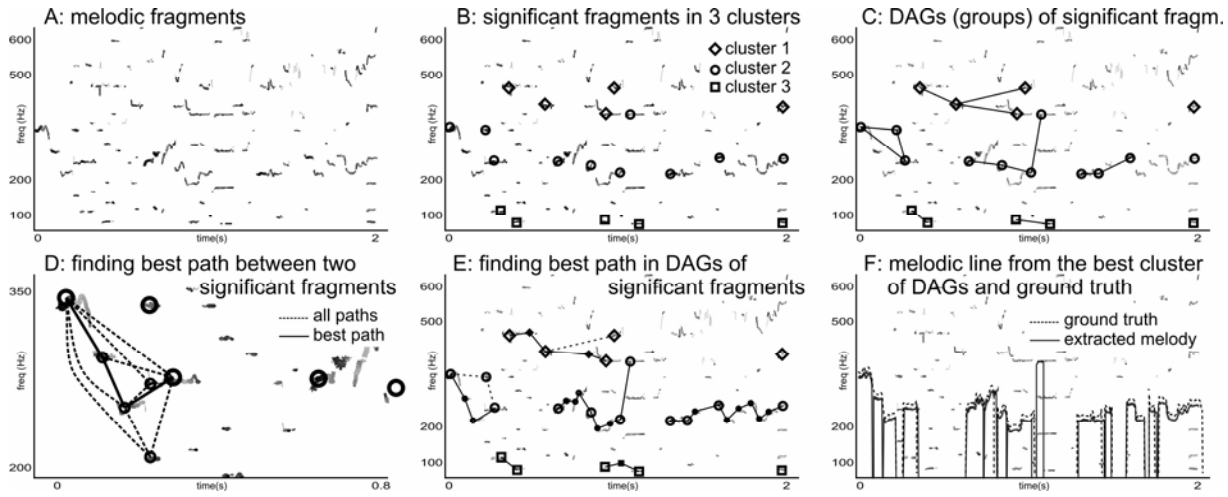

Fig. 1. Steps in finding the melodic line; A: finding melodic fragments, B: significant fragments are extracted and clustered, C: directed acyclic graphs are formed between closely spaced significant fragments, D: DAGs are formed and best paths found between each pair of linked significant fragments, E: best paths are found between DAGs of significant fragments, F: best cluster is found and melodic line extracted (ground truth on the plot is offset from extracted melody for clarity)

Finally, we search for the most dominant cluster of fragments to represent the main melodic line. Several criteria, including fragment loudness, coverage of melody over time and cluster consistency are involved in the search procedure and the cluster with the highest score is taken to represent the main melody, which is extracted from DAGs of groups of significant fragments (Figure 1F).

IV. Results

Due to lack of standard evaluation databases, it is difficult to provide results for our algorithm that could be compared to other approaches. The Music Information Retrieval Evaluation eXchange (MIREX) events within ISMIR conferences are starting to provide standard evaluation data, but so far, only one melody extraction contest has been held and we have used the data from the contest to train our algorithm, so providing results on the set would be irrelevant.

In Table 1, we give results on two approx. 30 second long excerpts used as test data for preparing algorithms for MIREX 2005. Results for two songs are given: S1: Natalie Cole – Almost like being in love and S2: Madonna – Frozen.

TABLE 1: RESULTS ON TWO SONGS.

	num. frags.	frag. %	prec.	recall	oct. prec	oct. recall
S1	390	84.8	84.8	86.6	84.8	86.6
S2	704	53.9	49.3	54.0	57.5	63.0

Columns in Table 1 are: num. frags.: number of melodic fragments found, frag %: percentage of ground truth melody covered by found melodic fragments, precision and recall of the extracted melodic line compared to ground truth, as well as precision and recall of the extracted melodic line compared to ground truth, when both lines are converted to a range of one octave, eliminating octave errors.

Especially for song 2, results are not very good; this mostly occurs due to poor performance of the procedure for discovering melodic fragments, which is currently the weakest point of our approach. The piece contains strong beats and bass line, which interfere with the sung melody and consequently cause problems to the pitch searching algorithm resulting in broken (and many) melodic fragments, as well as missed (masked by louder sounds) fragments. Out of the found fragments, in both cases, melody has been correctly identified. Such behavior is typical for our approach. Sometimes clustering of significant fragments also fails and divides the melody into several clusters, so that only parts of the melody are found to belong to the main melodic line, but such cases are not so frequent. Octave errors may occur with instruments that have strong harmonics.

V. Conclusion

We are currently quite satisfied with the part of our approach that groups melodic fragments into melodies, less so with the discovery of melodic fragments. We plan to include several improvements in our future work; besides improving or replacing the detection of melodic fragments, we plan to test a two-phase approach, where the found melodic line is refined by another search and merge of melodic fragments in the vicinity of the line, that may have been missed or broken. Also, the clustering procedure currently only works on entire song excerpts (or entire songs). We are working on a version that will work within an approx. 5 second sliding window and that will dynamically process new fragments and reform existing clusters or form new clusters as it progresses through a given piece. Such procedure will more accurately reflect the nature of human perception of music, mimicking short-term musical memory.

References

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

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

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

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

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

[6] 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.

[7] T. Painter, A. Spanias, “Perceptual Coding of Digital Audio”, Proceedings of the IEEE, Vol. 88 (4), 2000.

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

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

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

[11] ISMIR 2004 Audio Description Contest, Available: http://ismir2004.ismir.net/ISMIR_Contest.html

[12] A. Eronen, "Automatic Musical Instrument Recognition", M.Sc. Thesis, Tampere University of Technology, Finland.

[13] M. Marolt, "Gaussian Mixture Models For Extraction Of Melodic Lines From Audio Recordings", Proceedings of ISIMIR 2004, Barcelona, Spain.

[14] J, Eggink, G.J. Brown, "Extracting Melody Lines From Complex Audio", Proceedings of ISMIR 2004, Barcelona, Spain.