GAUSSIAN MIXTURE MODELS FOR EXTRACTION OF MELODIC LINES FROM AUDIO RECORDINGS

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

ABSTRACT

The presented study deals with extraction of melodic line(s) from polyphonic audio recordings. We base our work on the use of expectation maximization algorithm, which is employed in a two-step procedure that finds melodic lines in audio signals. In the first step, EM is used to find regions in the signal with strong and stable pitch (melodic fragments). In the second step, these fragments are grouped into clusters according to their properties (pitch, loudness...). The obtained clusters represent distinct melodic lines. Gaussian Mixture Models, trained with EM are used for clustering. The paper presents the entire process in more detail and gives some initial results.

1. 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, and these may include tempo and rhythm, melody, some form of harmonic structure, as well as the overall organisation 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 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.

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 realtime. 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]). 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 ([5, 6, 7]). Algorithms for simplified transcriptions, like extraction of melody, have been studied by few, with the notable exception of the work done by Goto [8].

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 GMM clustering of melodic fragments into melodic lines. The paper briefly describes each of these stages and presents some preliminary results.

2. 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.

2.1. Spectral Modeling Synthesis

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, [9]). 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. Psychoacoustic masking

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. Tonal and noise maskers are calculated from the set of sinusoidal components and the residual signal, as described in [10], and components that fall below the global masking threshold removed. On average, the masking procedure halves the total number of sinusoidal components.

2.3. Predominant pitch estimation

After the sinusoidal 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 [8], with some modifications. The method employs the Expectation-Maximisation (EM) algorithm, which treats the set of sinusoidal components within a short time window 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 interval. 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 possible tone models, while searching for one that maximizes the observed PDF (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 in the observed PDF.

2.4. 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 [8]. 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. An example is given in Fig. 1, 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 accompaniment, 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.

We performed informal listening tests by resynthesizing the fragments (on the basis of their pitch and amplitude) and comparing these resynthesized versions with the original signal. 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. keyboard 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 majority of fragments belonging to the main melodic line, which is the main point of interest in this study.

Most errors of the fragment finding procedure are octave-related, when the pitch of a segment is found to be an octave higher or lower than the perceived pitch. Also, areas in which several competing melodic lines

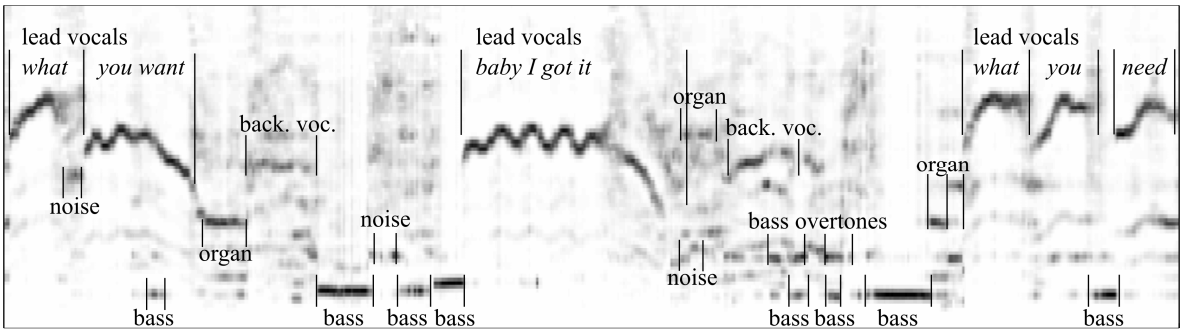

Figure 1. Segmentation into melodic fragments of an excerpt from Respect sung by Aretha Franklin.

with similar loudnesses compete for listener's attention 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 (i.e. bass line in the right part of Fig. 1).

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 the six dimensions of pitch, tempo, timbre, loudness, spatial location, and reverberant environment; sometimes with changes in rhythm; but rarely with changes in contour [12]. 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. 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. [13]). 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 five features, which represent:

- *dominance:* average weight of the tone model that originated the fragment, as calculated by the EM procedure;
- *pitch:* centroid of fragment's frequency with regard to its weight;
- *loudness:* mean loudness calculated according to Zwicker's loudness model [11] for partials belonging to the fragment;
- *pitch stability:* 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:* 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 (i.e. fragments belonging to accompaniment).

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 [14]. 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 constraints) or from different sources (different clusters - negative constraints). They provide additional information to the GMM training algorithm, and seem intuitive 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. 1. We treat melody as a succession of single events (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 in 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. 1). 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 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.

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. 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”, Proc. of ISMIR 2003, Baltimore, Maryland, USA, October 26-30, 2003.

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

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

[7] 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.

[8] 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.

[9] 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.

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

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

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

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

[14] 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.