MAIN MELODY EXTRACTION FROM POLYPHONIC MUSIC EXCERPTS USING A SOURCE/FILTER MODEL OF THE MAIN SOURCE

DURRIEU Jean-Louis, RICHARD Gaël and DAVID Bertrand
TELECOM ParisTech / CNRS LTCI - 37/39, rue Dareau - 75014 Paris
durrieu@enst.fr

ABSTRACT

In order to extract the audio melody of a polyphonic music mixture, we propose a source/filter model to fit the main source. For MIREX 2008, we design 2 distinct systems. The first system is based on a Gaussian Mixture Model (GMM), allowing only one source-filter couple at each frame for the main part. The second system, while keeping the source/filter formalism, assumes that all the couples are potentially active at any time, which is less optimal in our application. Although the GMM seems more realistic, preliminary tests show that the second model is not as computationally heavy as the GMM, but still keeps good results. Perceptual source separation results using these models however tend to show that the second model better fits the signals in certain circumstances, such as vibrato performances.

1 INTRODUCTION

Extracting the main melody from a polyphonic music signal can be defined as transcribing the notes that are played by an instrument which has to be somehow "dominating" other instruments from the mixture. This instrument can be in the foreground according to different cues, such as its energy or its frequency range.
During ISMIR 2004 and at MIREX 2005 and 2006, the evaluations for audio melody extraction showed there were several possible approaches to the problem [1]. Most of them are perceptually based, to a certain extend involving classifiers. However few works have been done that involve generative models for the observed signals. We propose to follow Ozerov [2] who tackled the problem by first separating the desired source from the rest, and thereafter estimating the pitch on the "monophonic" separated signal.
The general framework is therefore adapted from previous works on source separation, especially voice/music separation [3]. Our algorithms also take advantage of the harmonic nature of the desired source: we included an explicit source/filter model for the main source, while the background music keeps a rather general generative model. A Viterbi smoothing algorithm allows to control the continuity of the melodic line. We propose two different models for the main voice part: one which is directly derived from the Gaussian mixture models (GMM) of the litterature, and one which is based on an instantaneous mixture of all the possible basis elements in a dictionary.
This paper is organized as follows: first we introduce the models we consider for our submissions. The general principles for the estimation of the parameters are then discussed. We also describe the Viterbi algorithm used to retrieve a smooth melody line. At last, we give some preliminary results on the development files for MIREX 2008.

2 SIGNAL MODELS

2.1 Mixture Model

We assume that the observed signal x is the instantaneous mixture of two elementary signals: a signal corresponding to the main source, or main voice, noted v, and one for the background music noted m. We will also refer to v as the "vocal" part, since it will often happen that the main instrument is a human voice in the analyzed signals. The algorithms are well suited for this particular application. We therefore have: $x = v + m$.
This equation also holds for the short time Fourier transform (STFT) X, V and M respectively: $X = V + M$. The models we propose essentially aim at constraining the shapes of these STFT, more specifically for M we use a temporal repetition constraint while for V we focus on the harmonicity of the signal.

2.2 Background Music M

We recall here the model introduced in [4]: the background music signal M is considered to be the instantaneous mixture of R independent Gaussian sources M_r. Each of these sources is centered and characterized by its power spectral density (PSD), i.e. in our case its auto-covariance $\sigma_{M_r}(f)$. For a given frame t, an amplitude coefficient $a_r(t) \neq 0$ is associated to the corresponding source. Let $M_t(f)$ be the STFT of the background signal at frame t and frequency bin

f, then we write its likelihood:

$$\mathrm{p}(M_t(f)) = \mathcal{N}_c\left(M_t(f); 0, \sum_{r=1}^{R} a_r(t)\sigma_{M_r}(f)\right)$$

2.3 Vocal Part V, GMM framework

Our first submission uses the GMM framework as introduced in [3] to model the vocal part V. We proposed a source/filter that allows the algorithm to better fit our purpose: in this generative model, the source element refers to the excitation of the vocal folds and is therefore linked to the fundamental frequency of the sound F_0, while the filter part is characteristic of the vocal tract shape. As in [4], we discretize this space of possibilities such that we consider K possible filter frequency responses: $h_k(f)$ and N_{F0} possible F_0 or notes. The source DSPs are generated through a glottal source model KLGLOTT88, which gives realistic spectral combs $\sigma_{V,f_0}(f)$. As for M, we also include amplitude coefficients $b_{k,f_0}(t)$ for each frame and couple (k, f_0). The likelihood of the vocal part knowing the couple (k, f_0) is given by:

$$\mathrm{p}(V_t(f)|k, f_0) = \mathcal{N}_c\left(V_t(f); 0, b_{k,f_0}(t)h_k(f)\sigma_{V,f_0}(f)\right)$$

and the likelihood for the vocal partfor frame t is given by the weighted sum: $\mathrm{p}(V_t) = \sum_{k,f_0} \pi_{k,f_0}\mathrm{p}(V_t|k, f_0)$, where the a priori probabilities π_{k,f_0} are here assumed to be equal.

2.4 Vocal Part V, IMM framework

We propose a second model which was derived from the first one in order to find a solution that would be more efficient to compute. We came up with a formulation that keeps the source/filter model within an instantaneous mixture framework, which we will refer to as the instantaneous mixture model (IMM). Moreover, we assume the amplitude coefficient for one couple (k, f_0) at frame t can be written as the product of 2 separate amplitude coefficients, one for the filter and one for the source part of the vocal model: $b_{k,f_0}(t) = c_k(t)d_{f_0}(t)$. The likelihood is then expressed as follows:

$$\mathrm{p}(V_t(f)) = \mathcal{N}_c\left(V_t(f); 0, S_V(f,t)\right), \text{ where:}$$

$$S_V(f,t) = \left(\sum_k c_k(t)h_k(f)\right)\left(\sum_{f_0} d_{f_0}(t)\sigma_{V,f_0}(f)\right)$$

3 PARAMETER ESTIMATION

We use a maximum likelihood (ML) criterion in order to estimate the different parameters, for each of the models. The principle is the same as what is described in [4], where one can find updating rules for the parameters of the IMM. To estimate the parameters in the case of the GMM, we use an EM algorithm. However, apart from the expectation step of the EM algorithm, the updating rules are quite similar for both models.

4 MELODY ESTIMATION: SMOOTHING AND PRUNING FALSE POSITIVES

We could use different strategies to infer the melody line from the estimated parameters. The most straight-forward being to choose the f_0 that maximizes the a posteriori probability for the GMM and the amplitude coefficients for the IMM. However, this leads to rather poor results and a better strategy is to use a Viterbi smoothing of the melody line, assuming some sort of Markov model in the sequence of f_0, hence obtaining a trade-off between the smoothness of the melody and its global energy in the signal.
We also parameterized the transitions between the possible f_0, thus penalizing jumps in fundamental frequencies for the main melody, without disabling jumps for one note to the other. We tried several parameter settings in order to find the right way of parameterizing this weighting step. Other designs for this part may involve supervised techniques and other processes in order to take into account onsetting times.
Another mechanism permits to prune some spurious notes by checking whether the energy of the signal to which they correspond is high enough. The source separation framework we use allows, through Wiener filtering, to obtain the separated signals. Computing the energy for each frame of the separated main melody and thereafter thresholding, we can discriminate between spurious notes and true positives.

5 PRELIMINARY RESULTS

On the development set provided by the MIREX teams, i.e. the ISMIR 2004 and MIREX 2005 development data sets, we obtain results that are at the state of the art, compared to the submissions from MIREX 2006, as can be seen in table 1.
A surprising result is that, even though the computational time of the GMM is much higher, and despite the precision of the model compared to the IMM, the results seem to be significantly lower.

6 CONCLUSIONS

We proposed two new approaches to the audio melody extraction task. These approaches are based on generative models for both the vocal (or main melody) part and the background music part. The differences between the two proposed algorithms lie in the models: one uses a GMM, while the other one is based on an instantaneous mixture of the elements of the basis (IMM). The results are of the same

Participant	Raw Pitch Acc.	Raw Chr. Acc.	Overall Acc.
IMM	**82.3**%	**83.2**%	75.5%
GMM	64.5%	71.3%	56.8%
Dressler	80.0%	82.9%	**77.3**%
Ryynänen	75.5%	78.2%	72.1%
Poliner	72.6%	75.7%	69.3%
Sutton	59.1%	62.5%	55.7%
Brossier	48.3 %	61.7 %	39.8 %

Table 1. Results for MIREX 2006 participants (test set) and the proposed methods (development set).

order as the state of the art, especially the overall accuracy that clearly benefits from the pruning step we describe in this paper.
Note also that this technique is well suited for source separation, which makes it possible to adapt this main melody detection to the task of multipitch detection, by iteratively applying the algorithm to the separated residual corresponding to the background music.

7 REFERENCES

[1] G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Tr. Audio, Speech, Lang. Proc.*, 14(4):1247–1256, May 2007.

[2] Alexey Ozerov. *Adaptation de modles statistiques pour la sparation de sources mono-capteur. Application la sparation voix / musique dans les chansons*. PhD thesis, University of Rennes 1, 2006.

[3] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs. *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, 15(5):1564–1578, 2007.

[4] Jean-Louis Durrieu, Gael Richard, and Bertrand David. Singer melody extraction in polyphonic signals using source separation methods. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 169–172, March 31 2008-April 4 2008.