Proceedings of the 10th WSEAS International Conference on SYSTEMS, Vouliagmeni, Athens, Greece, July 10-12, 2006 (pp767-770)

A Musical Source Separation System with Lyrics Alignment

Ofer Hadar
Dmitry Bykhovsky
Guy Goldwasser
Etan Fisher

Department of Communications Systems Engineering
Ben-Gurion University of the Negev
Beer-Sheva,
Israel

Abstract: This paper examines the use of the generalized likelihood ratio test (GLRT) for the purposes of audio extraction and manipulation. The GLRT, which was designed originally for distinguishing between harmonic and non-harmonic audio frames, is extended for two music-oriented purposes. The first is to decompose a multiple-source mono recording into separate sources by which the decomposed files may be used to create new interpretations of the original recording. The second is for the purpose of lyrics alignment. The test shows a clear distinction of the singing voice within an orchestrated recording. Furthermore, words and syllables are indicated and can be used to align lyrics to the music automatically.

Key-Words: Source separation, Score alignment, Browsing by lyrics, Music signal processing

1 Introduction

Over the last several years there has been considerable interest in multimedia systems that allow interactive involvement of the user in the watching or listening process. In the audio field, the emergence of such interactive systems is only in the first stage of development. The most advanced digital music players offer only limited digital effects. The most important task at present is to provide the possibility of manipulating a recording in musically meaningful ways which will open up new possibilities in music entertainment, education and authoring applications.

This study employs the Generalized Likelihood Ratio Test (GLRT) developed in [1] for two such applications: Musical source separation and automatic alignment of lyrics to an existing audio file. These applications, combined with new lyrics alignment software, provide the Ben-Gurion university contribution to the Semantic Hi-Fi (SHF) Consortium [2].

In the context of large-scale digital music distribution, the goal of the SHF consortium was to develop a new generation of HIFI systems, offering new functionalities for browsing, interacting, rendering, personalizing and editing musical material. The possibility of adding lyrics to existing audio data has been under consideration for some time now, especially in commercial applications such as karaoke, etc. One of the aims of the present work is to automatically synchronize existing lyrics to pre-recorded audio data.

The current framework for lyrics alignment is the browsing by lyrics (BBL) tool. This is an audio file browsing tool in which the song's lyrics are used to skip back and forth inside a song. BBL also enables manual alignment of lyrics to the audio file.

2 Problem Formulation

The alignment / separation system is based on the harmonic model. Harmonic model representation assumes that sound source (voice or musical instrument) can be described as a sum of time-varying sinusoids [2]. The amplitudes, frequencies and phases of the sinusoids are derived from short time analysis of salient spectral peaks of the music signal.

2.1 The Harmonic Model

Let **y** be a finite audio frame with L samples at times $t_l, l = 1,\ldots,L$. The harmonic model can be written as

$$y(t_l) = b_{c_0} + \sum_{m=1}^{M} b_{cm} \cos \omega_0 m t_l + \sum_{m=1}^{M} b_{sm} \sin \omega_0 m t_l + noise \quad (1)$$

where the coefficients b_{cm}, b_{sm} carry the information on the intensity and phase of the *m*-th harmonic of the signal. ω_0 represents the fundamental component in the audio frame. Equation (1) in matrix notation can be rewritten as

$$\bar{y} = A(\omega_0)\mathbf{b} + n \tag{2}$$

where $\mathbf{A}(\omega_0)$ is the harmonic matrix, **b** is the harmonic coefficient vector and *n* is noise modelled as coloured Gaussian noise with an unknown covariance matrix. The interference included in the noise component may or may not be harmonic. The vector **b** is given by $b = [b_{c0},\ldots,b_{cM},b_{s1},\ldots,b_{sM}]$ and the matrix $\mathbf{A}(\omega_0)$ can be partitioned as: $\mathbf{A}(\omega_0) = [\mathbf{A}_c(\omega_0) \quad \mathbf{A}_s(\omega_0)]$. The elements of $\mathbf{A}(\omega_0)$ are given by

$$A_{clm}(\omega_o) = \cos(\omega_0 m t_l), \quad m = 0,\ldots,M,\ l = 1,\ldots,L$$
$$A_{slm}(\omega_o) = \sin(\omega_0 m t_l), \quad m = 0,\ldots,M,\ l = 1,\ldots,L$$

2.2 Maximum-Likelihood (ML) Parameter Estimation

The conditional probability density function (pdf) of the measurement vector **y** is given by

$$f_{\mathbf{y}|\omega_0,\mathbf{b},\sigma_n^2}(\mathbf{y} \mid \omega_0,\mathbf{b},\sigma_n^2) = \frac{1}{(2\pi\sigma_n^2)^{1/2}} e^{-\frac{1}{2\sigma_n^2}\|\mathbf{y}-\mathbf{A}(\omega_0)\mathbf{b}\|^2} \tag{3}$$

The ML estimator is obtained by maximizing the log-likelihood function,

$$L_{\mathbf{y}}(\hat{\omega}_0,\hat{\mathbf{b}},\hat{\sigma}_n^2) = \log f_{\mathbf{y}|\omega_0,\mathbf{b},\sigma_n^2}(\mathbf{y} \mid \omega_0,\mathbf{b},\sigma_n^2) = \left[-\frac{1}{2}\log(2\pi\sigma_n^2) - \frac{1}{2\sigma_n^2}\|\mathbf{y}-\mathbf{A}(\omega_0)\mathbf{b}\|^2\right] \tag{4}$$

with respect to unknown parameters.

Using the formulation appearing in [3], Maximum Likelihood estimation of the harmonic model parameters is obtained. Maximization of the log-likelihood function (4) with respect to **b** yields

$$\hat{\mathbf{b}} = [\mathbf{A}^T(\omega_0)\mathbf{A}(\omega_0)]^{-1} A^T(\omega_0)\mathbf{y} \tag{5}$$

Accordingly, the ML estimator of fundamental frequency is obtained by

$$\hat{\omega}_{o_{ML}} = \max_{\omega_0} \|P_{\mathbf{A}}(\omega_0)\mathbf{y}\|^2 \tag{6}$$

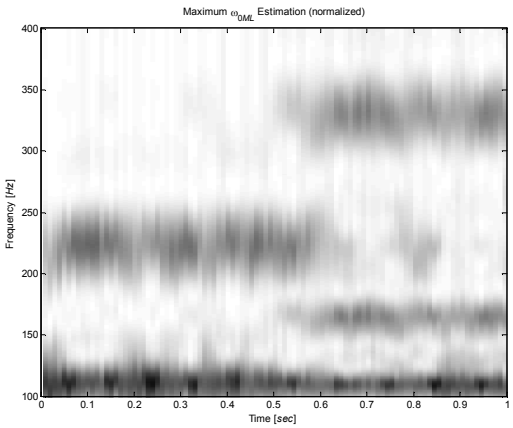
where $P_A(\omega_0) = \mathbf{A}(\omega_0)[\mathbf{A}^T(\omega_0)\mathbf{A}(\omega_0)]^{-1}\mathbf{A}^T(\omega_0)$ is the projection matrix into the subspace spanned by the columns of $\mathbf{A}(\omega_0)$.

3 Multiple Source Separation

3.1 ML Source Separation

Figure 1 shows the log-likelihood function of a short segment of polyphonic music containing a bass section, a string section and percussion. The musical sources can be clearly identified from the figure. The bass section appears around 110 Hz and the string instruments at 165, 220 and 330 Hz. The residual non-harmonic data (not appearing in the figure) contains a resonating cymbal.

Our suggested solution for the separation has three main stages:



1) Calculation of the log-likelihood functions over a predefined range of frequencies.
2) Peak extraction from the log-likelihood function using a frame-to-frame peak matching algorithm [6].
3) Extraction of the audio data for each harmonic source according to the method suggested in [5].

The frame-to-frame peak matching algorithm was taken from [6]. Figure 2 illustrates the "birth-death" procedure for the above musical segment. The data in the figure may be used for separation, alignment, etc.

3.2 Harmonic Sharing

Several approaches are examined for separating musical sound sources. The first is top-down analysis, in which analysis priority is given to higher harmonic sources (musical notes). The second is temporal alignment of the analysis based on previous knowledge of the score. This algorithm has been discussed previously and appears in [4]. The third approach is harmonic sharing. Since a large number of notes in a composition share the same harmonics, this fact must be employed when considering a feasible separation approach. Priority is given to lower harmonics, i.e. the fundamental

harmonic, which decreases towards the overtones of the note. Also, for smoothness and precision, the re-synthesis of notes can be based on 'memory' of a previous harmonic structure [5]. Transient detection and the handling of non-harmonic sounds are performed using the GLRT presented in [1]. This algorithm is also used in the context of lyrics alignment.

Figure 1 – *Pitch frequency estimation*

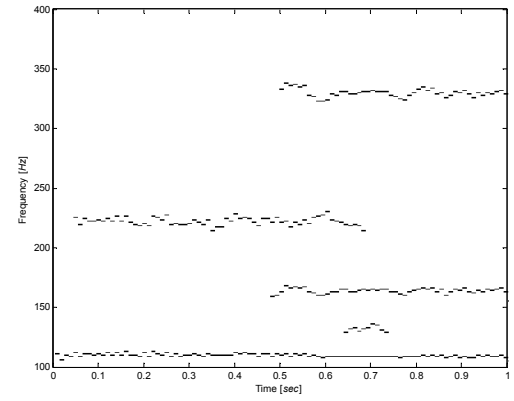


Figure 2 – *The results of "birth-death" algorithm*

Figure 3 – *GLRT lyrics detection test*

Figures 1 and 2 represent an audio example of the decomposition of a polyphonic electric piano. The separated notes were then spatially reconstructed thus producing an enhanced stereophonic interpretation of the recording. This example and some other examples can be heard at: http://www.cse.bgu.ac.il/hadar/Etan/SHF.htm .

4 BBL

BBL is an audio file browsing tool in which the song's lyrics are used to skip back and forth inside a song. BBL also enables manual alignment of lyrics to the audio file in a simple user friendly way.

The song's lyrics are aligned by rows or sentences. The basic commands are *play*, *pause*, *stop*, *skip forward/backward*, *FF*, and *RW*. The play button triggers a timer that counts cycles of 250 milliseconds. The pause button pauses the timer and the stop button stops the lyrics (and the music) and returns the clock to its starting position (initial time = zero). *RW* and *FF* skip sentences until another command is chosen and skip_RW / skip_FF skip a single line at a time.

4.1 Graphical User Interface (GUI)

The graphical user interface (GUI) shown in Fig. 4 has three parts. The first is the lyrics bar containing the song lyrics and user messages. The second part includes the basic commands needed for browsing by lyrics. The third part (the bottom part) includes commands that can align lyrics by listening to the melody.

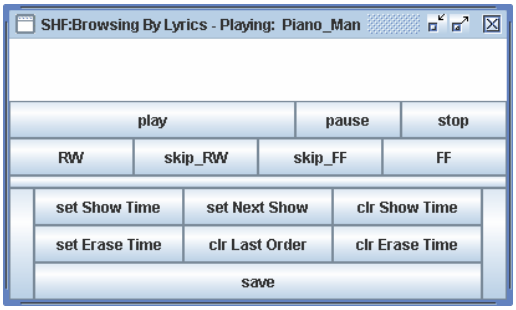


Figure 4 – *BBL GUI*

4.2 Lyrics Structure

The lyrics file format exists as a simple txt file. Each line consists of three objects:
1). Time for lyrics on (display time).
2). Time for lyrics off (erase time).
3). The.
The time structure is h:mm:ss,milisecond with a resolution of 250 milisecond (as in the Windows Media Player format).

4.3 Lyrics Alignment

At present, lyrics alignment achieved by the user. BBL has been developed from the point of view that anyone can choose any song with any lyrics, and can easily combine them. It is required, however, that the user be familiar with the song and its lyrics. Experience shows that using BBL, songs can be aligned faster than by other tools (such as WMP [8]). Alignment can take place both on lyrics with no alignment and lyrics that have previously been aligned in order to edit or to do corrections).

4.4 Automatic Lyrics Alignment

Figure 3 shows the GLRT pitch decision for a short segment of a song with vocals. The sharp transitions correspond to the beginning and end of lyrics being sung. This indicates the GLRT can be used for lyrics alignment and voice detection within music. Key words and syllables can be used to align lyrics to the music automatically.

5 Conclusions

This study demonstrated the use of a recently published GLRT for the purposes of audio extraction and manipulation. The first was to decompose a multiple-source mono recording into separate sources. The second was for the purpose of lyrics alignment. The test showed a clear distinction of the singing voice within an orchestrated recording. These results, combined with the BBL lyrics alignment software provide an interesting tool compatible with state-of-the-art audio listening technologies.

6 Acknowledgements

This work is supported by the IPF6 "Semantic Hi-Fi" consortium.

7 References

[1] E. Fisher, J. Tabrikian, and S. Dubnov, Generalized likelihood ratio test for voiced/unvoiced decision using the harmonic plus noise model, in *Proc. ICASSP*, 2003.

[2]SemanticHIFI, http://shf.ircam.fr/.

[3] J. Tabrikian, S. Dubnov and Y. Dickalov, Maximum A-Posteriori Probability Pitch Tracking in Noisy Environments Using Harmonic Model, *IEEE Trans. Speech, Audio Processing*, Vol. 12, pp. 76-87, 2004.

[4] S. Shalev-Shwartz, J. Keshet and Y. Singer, Learning to Align Polyphonic Music, *ISMIR*, 2004.

[5] A. Ben-Shalom and S. Dubnov, Optimal Filtering of an Instrument Sound in a Mixed Using Harmonic Model and Score Alignment, *Proceedings of International Computer Music Conference*, November 2004, Miami.

[6] R.J. McAulay, T.F. Quatieri, Speech Analysis/Synthesis Based on a Sinusoidal Representation, *IEEE Transactions on Acoustics, Speech and Signal Processing.*, pp. 744-754, 1986.

[7] W. Ye, M. Kan, T. L. Nwe, A. Shenoy and J. Yin, LyricAlly: Automatic Synchronization of Acoustic Musical Signals and Textual Lyrics, *In Proceedings of ACM Multimedia,* 2004.

[8] Microsoft Corp., Add Lyrics to Music Files, http://www.microsoft.com/windows/windowsmedia/knowledgecenter/howto/addlyrics.aspx, 2005.