A COMPUTATIONALLY EFFICIENT SCHEME FOR DOMINANT HARMONIC SOURCE SEPARATION

Mathieu Lagrange
Music Technology Group
McGill University
Canada
mathieu.lagrange@mcgill.ca

Luis Gustavo Martins
Telecom. and Multimedia Unit
INESC Porto / FEUP
Portugal
lmartins@inescporto.pt

George Tzanetakis
Dept. of Computer Science
University of Victoria
Canada
gtzan@uvic.ca

ABSTRACT

The leading voice is an important feature of musical piece and can often be considered as the dominant harmonic source. We propose in this paper a new scheme for the purpose of efficient dominant harmonic source separation. This is achieved by considering a new harmonicity cue which is first compared with state-of-the-art cues using a generic evaluation methodology. The proposed separation scheme is then compared to a generic Computational Auditory Scene Analysis framework. Computational speed-up and performance comparison is done using source separation and music information retrieval tasks.

Index Terms— Audio Source Separation, Auditory Scene Analysis, Harmonicity, Sinusoidal Modeling

1. INTRODUCTION

In videos analysis, spatial and temporal continuity can be considered to define similarities among some elements of a given representation. In the case of auditory scene analysis [1], only the temporal continuity of auditory events can be directly considered to track time/frequency components across time [2, 3]. By contrast, the spectral content is discontinuous in frequency. Fortunately, most auditory events have the property that the frequencies of their spectral components are harmonically related. Indeed, as Alain de Cheveigné states : "Harmonicity is the most powerful among Auditory Scene Analysis (ASA) cues. It is also the cue most often exploited in computational ASA systems and voice-separation systems" [4, 5, 6, 7], and by extension to various Music Information Retrieval (MIR) systems [8, 9].

Most of the source separation algorithms [10] iteratively estimate the dominant fundamental frequency (f0), and then remove the spectral components that are most likely to belong to the source attached to the corresponding f0. By contrast, we focus here on the definition of a similarity function between time/frequency components, that considers the harmonicity cue without relying on the prior estimation or knowledge of the f0's. Although not required this prior knowledge can be embedded easily and improve the grouping capability of the proposed similarity.

Many of the existing similarities that consider the harmonicity cue only make use of the mathematical relationship between the frequencies of the considered components [11, 12, 13]. This approach has several pitfalls. First, the fact that two components have harmonic frequencies is not directly linked to the fact that an audible pitch is perceived. And inversely, the fact that there is an audible pitch does not imply that all of the frequencies of the spectral components of the pitched source will be in perfect harmonic relation.

In this paper, we describe a new way of considering the harmonicity cue by assigning to each time/frequency components a spectral pattern. The correlation of those patterns in a wrapped harmonic space defines the similarity of the two considered components as detailed in Section 3. This similarity called Harmonically Wrapped Peak Similarity (HWPS) is then compared in Section 4 to state-of-the-art harmonic cues previously described in Section 2.

We proposed in [14] to consider the Normalized cuts algorithm to design a Computational Auditory Scene Analysis framework that allows many auditory cues such as the HWPS to be combined in order to express physical or perceptual constraints. This framework has been applied to dominant harmonic source separation in [15]. The main drawback of the approach is that even with the use of efficient computational techniques [16] and careful implementation, the algorithm is computationally demanding. We therefore propose in this paper an efficient algorithm that considers only the HPWS cue. The computational speed-up and performance comparison with the Normalized Cuts approach is studied in Section 5.

2. EXISTING HARMONICITY CUES

A wide variety of sounds produced by humans are harmonic, from singing voice and speech vowels, to musical sounds. As a result the harmonicity cue has been widely studied. However, only few studies have focused on the identification of harmonic relations between peaks without any prior fundamental frequency estimation.

The goal is to define a similarity measure between two frequency components (i.e. peaks) that should be high for harmonically related peaks and low for peaks that are not harmonically related. Many existing approaches [12, 11] use the mathematical properties of the harmonically related frequencies to build such a similarity measure. Srinivasan [11] considers an harmonicity map that can be pre-computed to calculate an harmonic similarity between two spectral bins, independently of their amplitudes. The map is computed as:

$$\text{hmap}(i, j) = 1 \text{ if } i|j \text{ or } j|i \quad (1)$$

where i, j are bin indices of the Fourier transform. The map is then smoothed to allow increasing level of inharmonicity using a Gaussian function. It is also normalized so that the sum of its elements is unity. The standard deviation of the Gaussian function is set to

This work was funded by the National Science and Engineering Research Council (NSERC), the Social Sciences and Humanities Research Council (SSHRC) of Canada, and the Portuguese Foundation for Science and Technology (FCT).

be 10% of its center frequency, as detailed in [11]. The similarity $W_s(p_l, p_m)$ between two sinusoidal peaks of indexes l,m in the analysis frame is then defined as:

$$W_s(p_l, p_m) = \text{shmap}(M_l, M_m) \quad (2)$$

where M_l, M_m are the corresponding bin indices of the two peaks and shmap is the smoothed harmonicity map.

Virtanen estimates for each spectral peak precise floating-point frequency parameters [12]. These parameters can be estimated using phase-based estimators [17]. If two peaks p_l and p_m are harmonically related, the ratio of their frequencies f_l and f_m is a ratio of two small positive integers a and b (which correspond to the harmonic rank of each peak, respectively).

By assuming that the fundamental frequency cannot be below the minimum frequency found by the sinusoidal modeling front-end (i.e. $f_{\text{min}} = 50$ Hz), it is possible to obtain an upper limit for a and b, respectively $a < \left\lfloor \frac{f_l}{f_{\text{min}}} \right\rfloor$ and $b < \left\lfloor \frac{f_m}{f_{\text{min}}} \right\rfloor$. Calculating all the ratios for possible a and b and choosing the closest to the ratio of the frequencies, Virtanen uses this error to define a harmonic similarity measure:

$$W_v(p_l, p_m) = 1 - \min_{a,b} \left| \log \left(\frac{f_l / f_m}{a/b} \right) \right| \quad (3)$$

3. THE HARMONICALLY WRAPPED PEAK SIMILARITY (HWPS)

Unlike most existing methods that only consider each peak in isolation the HWPS takes into account spectral information in a global manner to calculate harmonicity. The basic mechanism behind the HWPS measure is to assign each peak a spectral pattern. A harmonically wrapped frequency space is used to make the spectral patterns pitch invariant and the degree of matching between them is used as a similarity measure between the peaks.

Shifted Spectral Pattern

Our approach relies on a description of the spectral content using estimates of the frequency and amplitude of local maxima in the power spectrum, i.e. the peaks. We therefore propose to assign to each peak, p_l, a given spectral pattern, $\tilde{F}_l$, defined as the set of frequencies (in Hz) $F_l = \{f_i\}$ within the analysis frame k and shifted as follows:

$$\tilde{F}_l = \{\tilde{f}_i | \tilde{f}_i = f_i - f_l\} \forall i \in [1, L_k] \quad (4)$$

where L_k is the highest peak index of frame k. The spectral pattern is essentially a shift of the set of peak frequencies such that the frequency of the peak corresponding to the pattern maps to 0 (when i is equal to l).

Wrapped Frequency Space

This shifting allows some components of the spectral patterns of peaks belonging to the same harmonic source to be aligned. This alignment would be perfect in a wrapped frequency axis with modulus the fundamental frequency of the harmonic source.

To estimate whether two peaks p_l and p_m are in the same harmonic source, we propose to measure the correlation between the two spectral patterns in such a wrapped frequency space. The frequencies of each spectral pattern are then wrapped as follows:

$$\hat{f}_i = \text{mod}\left(\frac{\tilde{f}_i}{h(f_l, f_m)}, 1 \right) \quad (5)$$

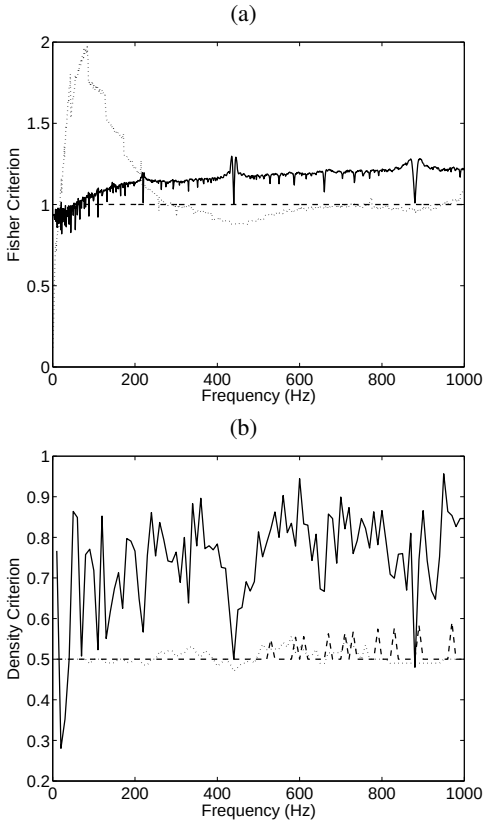

Fig. 1. Fisher (a) and Density (b) criteria versus the f0 of the second source for the Srinivasan cue (dotted line), the Virtanen cue (dashed line) and the HPWS cue (solid line). The f0 of the first source is set to 440 Hz.

where h is the wrapping frequency parameter and mod is the real modulo function. As stated before, this wrapping operation would be perfect with the prior knowledge of the fundamental frequency. Without such prior, we consider a conservative approach which tends to under estimate the fundamental frequency with:

$$h(f_l, f_m) = \min(f_l, f_m) \quad (6)$$

Discrete Cosine Similarity

These two harmonically wrapped spectral patterns ($\hat{F}_l$ and $\hat{F}_m$) are then quantized using an amplitude weighted histograms of 20 bins, where the contribution of each peak in the histogram is equal to its amplitude. The spectral pattern is also folded into an octave to form a pitch-invariant chroma profile. The HWPS measure between the peaks p_l and p_m is then defined as the cosine distance between those two corresponding histograms.

4. EVALUATION OF THE HWPS

In this section, we study the properties of the HWPS cue and contrast it with existing state-of-the art harmonicity cues using a generic evaluation methodology.

A good similarity metric between two peaks is a metric which is high for peaks of the same harmonic source and low for peaks that

	W_s	W_v	W_h	W_h(f0)
$\mathcal{F}$	1.44 (0.31)	1.00 (0.01)	1.22 (0.05)	2.27 (0.37)
$\mathcal{D}$	0.50 (0.01)	0.55 (0.11)	0.80 (0.12)	0.94 (0.16)

Table 1. Results for the separation of two harmonics sets of peaks. Mean and standard deviation values of the Fisher and Density criteria are computed for the Srinivasan, Virtanen, HWPS, and HWPS with prior knowledge of the two f0's.

do not belong to the same source. We can define the Fisher criterion $\mathcal{F}$ (loosely based on the Fisher discriminant commonly used in statistical analysis) as the sum of all the inter-class similarities divided by the sum of all the intra-class similarities. Since the Fisher criterion is not scale invariant, it may not be the best choice to compare the performance of distinct cues. Nevertheless, it is still an interesting way of evaluating the performance of a metric with respect to different scenarios.

Given so, we also define a Density criterion $\mathcal{D}$, computed as the number of peaks that have the closest neighboring peak in the feature space belonging to the same set. This criterion is scale invariant and closer to the one considered by clustering algorithms. We represent the partitioning of a set of elements X using an indicator function:

$$\begin{array}{rccc} E\colon & X & \rightarrow & \mathbb{N} \\ & x & \mapsto & i \end{array}$$

where i is the partition index in which x belongs, we can define:

$$\mathcal{D}(X) = \frac{1}{(\#\, X)^2} \quad \#\left\{(a,b) \,|\, d(a,b) = \min_{c \in X} d(a,c) \right. \\ \left. \wedge E(a) = E(b)\right\} \tag{7}$$

To evaluate the capabilities of the harmonic cues, we consider two synthetic sets of peaks with harmonically related frequencies and exponentially decaying amplitude envelope. The first set of peaks has a fundamental frequency of 440 Hz whereas the f0 of the second set is iteratively changed to values from 10 to 5000 Hz, using a 10 Hz step. Table 1 presents the performance of the evaluated harmonic cues in the $[100, 5000]$ Hz range, using the Fisher and Density criteria. The last column shows the performance of the HWPS with prior knowledge of the f0's of the two sets of peaks.

Figure 1(a) shows the evolution of the Fischer criterion with respect to the f0 of the second set for the three harmonic cues. The Srinivasan cue shows the expected behavior, with minimal performance when the two sources have close f0's (around 440 Hz). Another local minima is found around 880 Hz and the performance globally increases with the second f0. Since the Virtanen and HPWS cues consider more precise frequency estimates, a finer behavior can be noticed around frequencies multiple of the first peak f0. Differently from the cue proposed by Virtanen, the HWPS performance increases with the frequency difference between the two f0's, as desired. As shown in Figure 1(b), the HWPS performs well as far as the Density criterion is concerned, except at frequency locations multiple of 440 Hz, the f0 of the reference set.

5. PROPOSED ALGORITHM

The HWPS estimates the degree of harmonic relationship between two frequency components and more precisely the likelihood that those two components belong to a dominant harmonic source. This cue can be combined with other CASA cues as proposed in [15].

Focusing on the task of dominant harmonic source separation using the HWPS cue only, we can substantially reduce the computational effort by considering the following algorithm: for each frame the precise frequencies and magnitudes of spectral peaks are estimated [17]. An harmonicity factor computed as follows:

$$h_l = \sum_{i \neq l} a_i W_h(p_l, p_i) \tag{8}$$

is assigned to each selected peak p_l. This factor indicates the likelihood that the considered peak belongs to a dominant harmonic source, therefore only the peaks with the highest factor value are considered for resynthesis.

We evaluate the performance of this method, termed Harmonically Enhanced Spectral Representation (HESR), against the Ncut approach using the different tasks considered in [15], namely singing voice separation, dominant pitch estimation, and voice detection.

The Ncut algorithm is based on a graph representation of the set of peaks. The global normalized cut criterion is used to partition the graph. Within this framework, a perceptual grouping can be achieved by appropriately defining the similarity between peaks. We consider here a combination of the HWPS criterion together with amplitude and frequency proximity's as described in [15]. Among a texture window of 10 analysis frames, the set of peaks is divided between 5 clusters and the peaks of the 2 clusters that are the more dense in the feature space are considered for resynthesis. The HESR algorithm, per frame, resynthetizes the 10 peaks with the higest harmonicity factor h_l among the 20 extracted ones.

A 2.2 GHz intel machine was used for the experiments and both algorithms are implemented in C++ using the Marsyas framework [1]. The computation times are calculated for the processing of the 10 songs. The Ncut algorithm is roughly 2 times real-time and could therefore be close to real-time with a parallel implementation of the Singular-Value Decomposition algorithm considered for the clustering step. Even though, when considered as a front-end for MIR tasks, one would like to save computation time for further processing. The proposed approach is 8 times faster, making it a more practical option.

For the separation experiment, we use a dataset consisting of 10 polyphonic music signals of different genres for which we have the original vocal and music accompaniment tracks before mixing, as well as the final mix. The signal-to-distortion ratio (SDR) is considered as a simple measure of the distortion caused by the separation algorithm [9] and the mean value of the segmental SDR's achieved using all the signals is computed. The proposed algorithm only considers the HWPS cue and selects the frequency components of interest in a straightforward fashion. Consequently, the separation performance drops by approximately 3 dB, see Table 2.

For the melody extraction experiment, we use single channel polyphonic music signals from the MIREX audio melody extraction dataset [2]. It consists of 23 clips of various styles including instrumental and MIDI tracks. We estimate the pitch contour from the dominant melodic voice using using the Praat pitch estimation [18] on the resynthetized signals using both methods. The HESR achieved better results for this experiments, which confirms that the proposed harmonicity criterion is able to select some components of the dominant harmonic source.

We then conducted a Voicing Detection evaluation, where we tried to identify whether a given time frame contains a "melody"

[1] http://marsyas.sourceforge.net

[2] http://www.music-ir.org/mirex2005/index.php/Main_Page

	Ncut	HESR
Separation Performance (dB)	4.25	1.07
Pitch Estimation Accuracy (%)	46	64
Voice Detection Accuracy (%)	86	83
Computational Cost ($\times$ real-time)	1.91	0.23

Table 2. Performance comparison of the Normalized Cut approach and the HESR one. Computational cost is expressed in terms of real-time and separation performance in terms of SDR.

pitch or not. The goal of this experiment was to determine whether the proposed algorithm can be used to achieve a good voicing detection accuracy in monaural polyphonic recordings. The same dataset of the 10 polyphonic music pieces for which we have the original separate vocal track was used for this experiment. The voicing regions were manually labeled from the original vocal track and were used as the ground truth. A supervised learning approach was used to train voicing/no-voicing classifiers for two configurations, as presented in table 2: Ncut refers to using Mel-Frequency Cepstral Coefficients (MFCC) [19] calculated over the automatically separated voice signal, and HESR refers to MFCC calculated over the HESR mixed voice and music signal. The experiments were performed using the Weka machine learning framework, where a support vector machine was trained using the sequential minimal optimization (SMO) [20]. As shown, the HESR MFCC accuracy compares well to the slightly superior value achieved when using the Ncut and MFCC approach. Not presented in the table, but also interesting to compare, is the accuracy of the MFCC feature using the same classifier but when applied directly to the original mixed signal (i.e. without any Ncut separation or HESR processing). For this case we get 69% accuracy, a quite inferior value in comparison to the two values discussed above.

Those experiments show that the proposed approach, even though not capable of achieving comparable results for the source separation task, is able to achieve convincing results as a front-end for MIR tasks such as pitch estimation and voicing detection at a low computational cost.

6. DISCUSSION

In this paper we proposed a computationally efficient algorithm for dominant source separation that considers a harmonicity criterion for selecting relevant frequency component in the mixture. This harmonicity criterion is based on a harmonic similarity which is shown to have good statistical properties.

The efficiency of the proposed algorithm is well suited for being integrated in feature extraction algorithms for very large datasets or with real-time constraints as a pre-processing step allowing to focus on the dominant harmonic content. More specifically, we believe that the harmonicity factor could be useful for enhancing spectral representations used for chords or key estimation [8] and plan to investigate further in this direction.

7. REFERENCES

[1] A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1990.

[2] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on sinusoidal representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34(4), 1986.

[3] M. Lagrange, Sylvain Marchand, and Jean-Bernard Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Trans. on Acoustics, Speech, and Language Processing*, vol. in press., 2007.

[4] D.F. Rosenthal and H.G. Okuno, Eds., *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, 1998.

[5] F.R. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2004.

[6] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, 2006.

[7] E. Vincent, "Musical source separation using time-frequency priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(1), pp. 91–98, 2006.

[8] E. Gòmez, *Tonal Description of Music Audio Signals*, Ph.D. thesis, 2006.

[9] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*, Springer, 2006.

[10] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, Wiley, 2006.

[11] S.H. Srinivasan and M. Kankanhalli, "Harmonicity and dynamics based audio separation," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[12] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proc. ICASSP*, 2000, vol. 2, pp. 765–768.

[13] L. G. Martins and A.J.S Ferreira, "PCM to MIDI transposition," in *Proc. of Audio Engineering Society (AES)*, 2002.

[14] M. Lagrange and G. Tzanetakis, "Sound source tracking and formation using normalized cuts," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[15] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for predominant melodic source separation," *IEEE Trans. on Acoustics, Speech, and Language Processing*, vol. to appear, 2007.

[16] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22(8), pp. 888–905, 2000.

[17] M. Lagrange and S. Marchand, "Estimating the instantaneous frequency of sinusoidal components using phase-based methods," *to appear in the Journal of the Audio Engineering Society*, 2007.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 4.5.06)," Retrieved December 13, 2006, from http://www.praat.org/.

[19] S. Davis and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," *IEEE Transcactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[20] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, 2005.