USING VOICE SUPPRESSION ALGORITHMS TO IMPROVE BEAT TRACKING IN THE PRESENCE OF HIGHLY PREDOMINANT VOCALS

Jose R. Zapata and Emilia Gomez

Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
{ joser.zapata, emilia.gomez }@upf.edu

ABSTRACT

Beat tracking estimation from music signals becomes difficult in the presence of highly predominant vocals. We compare the performance of five state-of-the-art algorithms on two datasets, a generic annotated collection and a dataset comprised of song excerpts with highly predominant vocals. Then, we use seven state-of-the-art audio voice suppression techniques and a simple low pass filter to improve beat tracking estimations in the later case. Finally, we evaluate all the pairwise combinations between beat tracking and voice suppression methods. We confirm our hypothesis that voice suppression improves the mean performance of beat trackers for the predominant vocal collection.

Index Terms— Beat tracking, source separation, voice suppression, evaluation

1. INTRODUCTION

The *Beat* is a relevant audio descriptor of a piece of music defined as "one of a series of regularly recurring, precisely equivalent stimuli" [1] which represents the perceptually most prominent period at which most people would regularly tap their feet, hands or finger when listening to music.

The location of the beats in music is exploited in higher-level music processing applications such as music retrieval, cover detection, playlist generation, structural analysis, score alignment, rhythm transformations or source separation, among others. For this reason, the Music Information Retrieval (MIR) research community has devoted much effort to finding ways to automate its extraction.

Many approaches for beat tracking from music signals have been proposed in the literature [2–5]. Although current state of the art methods yield satisfactory results for many application contexts (e.g. around 77,7% accuracy according to [6]), and some efforts have also been devoted to their quantitative comparison [7], there is an increasing interest in analyzing the limitations of existing methods in terms of music material and computed descriptors as a way to overcome the glass ceiling in system performance.

In this direction, Holzapfel et al. propose a method for the automatic identification of difficult examples for beat tracking by assigning a "difficulty" score to musical signals based on the mutual agreement between a committee of five beat tracking algorithms [6]. This study was carried out on a music collection of 1360 songs [2, 8]. As a result of this analysis, an annotated audio dataset of 217 difficult excerpts of 40 s from varied musical styles (e.g. classical, chanson, jazz, folk and flamenco) was created. This collection contained, among others, songs with strong and expressive vocals, which resulted in beat estimation errors even in the presence of a rhythmically stable accompaniment.

This paper focuses on beat estimation in this particular context, songs with highly predominant vocals, and is motivated by previous research showing the advantage of source separation techniques as a preprocessing step for automatic tempo estimation [9, 10] and beat tracking [11–13]. We evaluate and discuss how voice suppression techniques improve rhythmic saliency in songs with highly predominant vocals and quiet accompaniment, and thus facilitate the automatic estimation of beat positions.

Using source separation for improving tempo accuracy estimation has been proposed by Alonso [9], based on harmonic + noise decomposition of the audio signal. To improve beat/tempo estimation Gkiokas [11], uses a percussive / harmonic blind source separation and Chordia [10] uses a blind source separation technique using the non-shift-invariant version of Probabilistic Latent Component Analysis (PLCA). In this study we proposed to use source separation for voice suppression in excerpts with highly predominant vocals, in order to improve beat tracking performance. To the best of our knowledge, such an approach has a not been previously considered in the literature.

In this study, we evaluate the performance of five state-of-the-art beat tracking algorithms in combination with seven different voice suppression approaches and a simple low pass filter. We consider an annotated dataset of difficult audio song excerpts with highly predominant vocals. The paper is struc-

Thanks to Colciencias and Universidad Pontificia Bolivariana (Colombia), Music Technology Group at Universitat Pompeu Fabra, SIGMUS project (http://mtg.upf.edu/project/sigmus) for the financial support.

tured as follows. We first present the experimental methodology and tested approaches in sec. 2. Then we present the main results of this work in sec. 3. Finally, we discuss them in sec. 4, giving ideas for future work in this problem.

2. EXPERIMENTAL METHODOLOGY

2.1. Music Material

Two datasets have been considered for this study. The first one is varied in terms of genre and tempo, and it has been widely used in the literature [2, 6, 8, 14]. It consists on 1360 beat-annotated musical pieces (*Dataset1360*), with tempi ranging from 50 to 200 bpm, and covering the following musical genres: acoustic, afro-american, jazz/blues, classical, choral, electronic, rock/pop, balkan/greek and samba. This dataset allows us to obtain a baseline evaluation of the considered beat tracking algorithms (see Table 2). The second one (*DatasetSMC*)[1] contains 217 beat-annotated musical pieces which have been found to be difficult for automatic beat tracking according to [6]. It includes the following genres: classical music, romantic music, jazz, blues, chanson, and solo guitar compositions.

The difficulty of the excerpts in the *Dataset1360* and *DatasetSMC* was further assessed from the mean performance of the five considered beat trackers using the method proposed in [6, 15]. From the difficult excerpts, we finally selected 75 examples with highly predominant vocals (*DatasetVocal*).

2.2. Voice Suppression Methods

Voice suppression methods intend to remove the singing voice from a polyphonic music signal by means of source separation techniques. According to [16], there are three main approaches for singing voice separation methods: spectrogram factorization, pitch-based inference and repeating-structure removal. In this study, we consider a set of state-of-the-art algorithms based on those different principles which are accessible for evaluation purposes. Three different spectrogram factorization approaches, explained below, are evaluated. They are based on decomposing a magnitude spectrogram as a set of components that represent features such as the spectral patterns (basis) or the activations (gains) of the active components along time [16–18]. We also evaluate the use of four repeating-structure removal methods [19–21] which rely on pattern recognition to identify and extract accompaniment segments, without manual labeling, which can be classified as repeating musical structures. Finally, we evaluated the use of an low pass filter to remove higher spectral components in order to compare the results of voice suppression algorithms with a simple approach. We provide a brief description of the considered algorithms.

[1] `http://smc.inescporto.pt/research/data/`

Low Pass Filter (LPF): Base on [22], a simple Butterworth double-pole low-pass filter at 261.6 Hz (4800 cent) and Q = 0.707 was used as a baseline approach to remove high spectral components where the voice is assumed to be predominant.[2]

Instantaneous Mixture Model (IMM): Durrieu et al. [17] propose a source/filter signal model of a mixed power spectrum as a decomposition into a dictionary of pre-defined spectral shapes, which provide a mid-level representation of the signal content together with some timbre information. A non-negative matrix factorization (NMF) technique is used for source separation [3].

Low Latency Instrument Separation (LLIS): This method allows voice suppression under real-time constraints, and it is based on time-frequency binary masks resulting from the combination of azimuth, phase difference and absolute frequency spectral bin classification and harmonic-derived masks. A support vector machine (SVM) is used for timbre classification, and for the harmonic-derived masks, a pitch likelihood estimation technique based on Tikhonov regularization is used. We refer to [18] for a detailed description of the algorithm.

Repeating Pattern Extraction Technique (REPET): REPET[4] is a method for separating the repeating background from the non-repeating foreground in a excerpt audio mixture. The approach assumes that musical pieces are often characterized by an underlying repeating structure over which varying elements are superimposed. The system identifies the repeating elements in the audio, compares them to repeating models derived from them, and extracts the repeating patterns via time-frequency masking. REPET with sliding window (REPET win) is an extension of the algorithm to full-track songs applying the algorithm to local sections over time using a fixed sliding window. We refer to [19] for a detailed description of the algorithm.

Adaptive REPET (REPET ada): The REPET method is originally intended for excerpts with a relatively stable repeating background. For full-track songs, the repeating background is likely to vary over time, so the adaptive REPET can be directly adapted along time by locally modeling the repeating background to handle varying repeating structures. This method is detailed in [20].

REPET with Similarity Matrix (REPET sim): This method [21], generalizes the REPET approach to handle cases where repetitions also happen intermittently or without a fixed period, thus allowing the processing of music pieces with fast-varying repeating structures and isolated repeating elements. Instead of looking for periodicity, this method uses a similarity matrix to identify repeating elements. It then calculates a repeating spectrogram model using the median and extracts repeating patterns using a time-frequency masking.

[2] sox in.wav out.wav lowpass 261.6

[3] `www.durrieu.ch/research/jstsp2010.html` VU output

[4] `music.cs.northwestern.edu/`

Singing Voice Separation (UJaen): The last considered approach, described in [16], factorizes a mixture spectrogram into three separated spectrograms (Percussive, Harmonic and Vocal). Harmonic sounds are modeled by sparseness in frequency and smoothness in time, percussive sounds by smoothness in frequency and sparseness in time and vocal sound are modeled by sparseness in frequency and sparseness in time. A predominant f_0 estimation method is used for the vocal separation, for which the vocal parts were previously labeled by hand. The implementation used in this study had the same source separation method, but was completely unsupervised.

2.3. Beat trackers

We consider five state-of-the-art beat tracking approaches presented in Table 1. The algorithms used consist of two processing steps: First, the extraction of an onset detection function which is a mid-level representation that reveals the main changes in the audio signal in time, like Bandwise Accent Signal (BAS) [5], Complex spectral difference (CSD) [23], Energy Flux (EF) [24], Mel auditory feature (MAF) [25], Harmonic feature (HF) [26], Beat emphasis function (BEF) [27] and Spectral flux (SFX) [23].

Second, a periodicity detection function is used to obtain an estimation of the beat times and finally the phase (position) of the beats are defined in this process. The tracker systems used by the evaluated approaches are: Autocorrelation function (ACF) [3, 28] , Comb bank filter (CBF) [5], Inter-onset interval (IOI) [2, 4], Hidden Markov Model (HMM) [3, 5, 28] and Multiple agents (MA) [2].

Beat Tracker	Onset Detection Function	Tracker System
Beatroot[5] [2]	SFX	IOI, MA
Degara[6] [3]	CSD	ACF, HMM
IBT[7] [4]	SFX	IOI, MA
Klapuri[8] [5]	BAS	CBF, HMM
MultiFeature_Inf [28]	CSD,EF,MAF HF,BEF	ACF, HMM

Table 1. Evaluated Beat trackers.

2.4. Evaluation Measures

For evaluating the beat tracking accuracy against manual annotations, we consider the beat tracking evaluation toolbox[9] which is used on the beat tracking evaluation task at the MIREX evaluation initiative [29].

[4] www.eecs.qmul.ac.uk/~simond/beatroot/

[5] www.gts.tsc.uvigo.es/~ndegara/Publications.html

[6] marsyas.info/ ; IBT off-line mode.

[7] www.cs.tut.fi/~klap/iiro/meter/index.html

[9] www.elec.qmul.ac.uk/digitalmusic/downloads/beateval/

Among all the proposed evaluation metrics, we consider the most permissive continuity measures that Allowed Metrical Level errors, because it considers that beat estimations at double or half of the correct metrical level are valid, and it also accepts off-beat estimations. We compute AMLc (Allowed Metrical Level with continuity required) and AMLt (Allowed Metrical Level with no continuity required) as defined in [5, 30]. Output range between [0 - 100]%.

3. RESULTS

Table 2 shows the average evaluation results of the considered beat tracking systems on a varied dataset *Dataset1360* and Table 3 shows their performance on the *DatasetVocal*. The Beat estimations and evaluation data are publicly available [10].

Measure	Bea.	Deg.	IBT	Kla.	MF_inf
AMLc (%)	53,50	69,89	63,96	69,79	71.99
AMLt (%)	70,83	77,72	73,76	77,70	80.16

Table 2. Evaluation Results in *Dataset1360*

We observe that the beat tracking performance drastically decreases for songs with highly predominant vocals for all the considered methods. This confirms our hypothesis and the observations of previous research work [6], which identified the difficulty of these examples. To get an idea of the best algorithmic performance currently achievable, we define an "Oracle" beat tracker whose performance is equal to the best performance obtained for each excerpt by any of the considered algorithms. For the *DatasetVocal*, the Oracle tracker would yield 33.95% and 52.65% accuracy for the AMLc and AMLt measures respectively. Evidently, there is still much room for improvement for this type of music material.

Regarding the advantage of using voice suppression techniques, we observe that all beat trackers increases their mean performance (AMLc and AMLt measures) over *DatasetVocal* when using *UJaen* and *IMM* as a preprocessing step, although the accuracy increase is small. In Addition, Degara's beat tracking approach (with one of the highest performance in *Dataset1360*) statistically improves its performance for all the evaluated voice suppression algorithms ($p<0.05$). Moreover, all beat trackers improve their accuracy (AMLt measure) using *LLIS* as a preprocessing step. Finally, the three best performing methods on *Dataset1360* experience an increase of the performance on *DatasetVocal* using very simple (*LPF*) and fast (*REPET*) approaches. This leads us to one of the most critical aspects of using voice suppression over large collections: the computational cost (The runtime is provided in Table 3). Although these approaches vary in terms of optimization level, we observe large differences in runtime (e.g. *IMM* is almost 50 times slower than *UJaen* algorithm).

[10] sites.google.com/site/tempoandbeattracking/

Measure	BT name	Original	LPF	Repet	Repet ada	Repet sim	Repet win	LLIS	UJaen	IMM
AMLc (%)	**Beatroot**	10.54	10.97	9.14	7.59	10.12	9.37	10.49	**14.68**	11.97
	Degara	16.88	24.17*	24.94*	25.38*	24.13*	24.86*	23.91*	26.23*	**26.83***
	IBT	24.70	17.07	16.51	20.00	18.39	22.02	19.75	24.79*	**26.46***
	Klapuri	22.61	24.52	25.61	22.53	24.53	22.74	23.37	**29.44**	26.31
	MultiF_inf	21.32	24.64	28.11	27.26	23.87	22.92	27.38	**29.47**	28.75
AMLt (%)	**Beatroot**	25.39	25.19	24.84	19.72	26.17	23.38	27.23	**31.89**	27.36
	Degara	28.70	37.64*	38.45*	37.67*	37.99*	38.28*	37.86*	40.13*	**42.24***
	IBT	27.55	37.45	27.35	32.38	29.55	33.17	32.78	39.40*	**40.49***
	Klapuri	36.60	39.12	38.43	36.41	38.70	34.49	39.43	**43.96**	43.02
	MultiF_inf	34.88	37.34	40.99	39.59	38.45	35.85	41.51	41.75	**42.81**
Process Time [=] Min			0.37	3.42	15.54	14.30	6.74	221.54	293.51	14723.12

Table 3. Mean AMLc and AMLt performance results in the original and the processed audio files from *DatasetVocal* per beat tracking system (* indicates statistically significant improvements with $p<0.05$)

Audio Files	Measure	LPF	Repet	Repet ada	Repet sim	Repet win	LLIS	UJaen	IMM
Improved (%)	AMLc	8	1.33	4	4	4	9.33	10.66	6.66
	AMLt	12	4	14.66	8	5.33	12	13.33	13.33
Degraded(%)	AMLc	1.33	4	2.66	2.66	1.33	8	4	1.33
	AMLt	5.33	8	2.66	1.33	4	2.66	5.33	2.66

Table 4. Percentage of songs that improves and degraded in all beat trackers in each voice suppression system

In Table 4 we present the total number of songs for which all beat trackers obtained improved performance when using voice suppression algorithms. We observe that the performance is improved for the majority of songs (with the exception of the *REPET* method). We also observe that the better the performance of the voice suppression algorithm, the greater the increase in beat tracking performance.

If we apply voice suppression methods not only to music with highly predominant vocal but for *Dataset1360*, we only get small improvements in accuracy for the combination of all *REPET+Degara*, *LLIS+Klapuri* and *REPET sim+ IBT*. None of these improvements are statistically significant, though. We then conclude that while voice suppression might be beneficial for excerpts with highly predominant vocals, these algorithms do not provide enhancements for varied datasets.

4. DISCUSSION AND FUTURE WORK

Voice suppression allows beat trackers to achieve higher estimation accuracy than the Oracle in some song excerpts with highly predominant vocals, as they enhance the signal and allow a better mid-level representation for beat tracking. Although the highest increase is yielded by the *IMM* voice suppression method, this approach needs a very high computation time (around 196 min per song) to process the audio. Other methods such as *LLIS* and *UJaen* yield similar results in less computation time (around 3.9 min per song). This fact makes them more suitable to process large music collections.

We have demonstrated that voice suppression techniques help to push up the glass ceiling of state-of-the-art beat tracking algorithms in music with highly predominant vocals. Nevertheless, this approach would decrease beat tracking performance in the contrary situation, i.e. a cappella, choral or music where the voice carries relevant rhythmic information. Future work has to be devoted to automatically selecting the candidate material where voice suppression would have a positive effect on beat tracking.

Beat trackers with higher mean performance in this evaluation seem to benefit more from voice suppression in difficult songs with highly predominant vocals. Moreover, voice suppression can be used as a pre-processing stage without having to modify the beat tracking algorithm.

Most of the voice suppression algorithms use spatial information to improve their performance. This evaluation was carried out on excerpts of mono files. For future experiments, we will consider full length stereo songs in order to evaluate voice suppression methods in more realistic setting.

Finally, we plan to investigate if there is a suitable methodology to combine different voice suppression methods with alternative beat tracking algorithms as a way to maximize the performance increase.

5. ACKNOWLEDGEMENTS

Thanks to the authors of the algorithms for making them available and the reviewers for your helpful recommendations.

6. REFERENCES

[1] G. Cooper and L. B. Meyer, *The rhythmic structure of music*, University Of Chicago Press, Chicago, Apr. 1960.

[2] S. Dixon, "Evaluation of the Audio Beat Tracking System BeatRoot," *J. of New Music Research*, vol. 36, no. 1, pp. 39–50, 2007.

[3] N. Degara, E. Argones Rua, A. Pena, S. Torres-Guijarro, M. E. P. Davies, and M. D. Plumbley, "Reliability-Informed Beat Tracking of Musical Signals," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 20, no. 1, pp. 290–301, Jan. 2012.

[4] J. Lobato Oliveira, F. Gouyon, L. G. Martins, and L. Reis, "IBT: A Real-Time tempo and beat tracking system," in *Proc. 11th Int. Soc. on Music Info. Retrieval Conf.*, Utrecht, 2010, pp. 291–296.

[5] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 1, pp. 342–355, 2006.

[6] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. Lobato Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 20, no. 9, pp. 2539–2548, 2012.

[7] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms," *J. of New Music Research*, vol. 36, no. 1, pp. 1–16, Mar. 2007.

[8] F. Gouyon, *A Computational Approach to Rhythm Description*, Ph.D. thesis, Pompeu Fabra University, Barcelona, Audio Visual Institute, 2005.

[9] M. Alonso, G. Richard, and B. David, "Accurate tempo estimation based on harmonic + noise decomposition," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 161–175, Oct. 2007.

[10] P. Chordia and A. Rae, "Using source separation to improve tempo detection," in *Proc. 10th Int. Soc. on Music Info. Retrieval Conf.*, Kobe, Japan, 2009, pp. 183–188.

[11] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stajylakis, "Music tempo estimation and beat tracking by applying source separation and metrical relations," in *proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, kyoto, Mar. 2012, pp. 421–424.

[12] M. Malcangi, "Source Separation and Beat Tracking: A System Approach to the Development of a Robust Audio-to-Score System," *Computer Music Modeling and Retrieval*, vol. 3310, pp. 71–82, 2005.

[13] J. R. Zapata and E. Gómez, "Improving Beat Tracking in the presence of highly predominant vocals using source separation techniques: Preliminary study," in *Proc. 9th Int. Symposium on Computer Music Modeling and Retrieval*, London, 2012, pp. 583–590.

[14] J. R Zapata, A. Holzapfel, M. E. P. Davies, J. Lobato Oliveira, and F. Gouyon, "Assigning a confidence threshold on automatic beat annotation in large datasets," in *Proc. 13th Int. Soc. for Music Info. Retrieval Conf.*, Porto, 2012, pp. 157–162.

[15] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. Lobato Oliveira, and F. Gouyon, "On the automatic identification of difficult examples for beat tracking: towards building new evaluation datasets," in *proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, kyoto, 2012, pp. 89–92.

[16] E. Gómez, F. Cañadas, J. Salamon, J. Bonada, P. Vera, and P. Cabañas, "Predominant Fundamental Frequency Estimation vs Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing," in *13th Int. Soc. for Music Info. Retrieval Conf.*, Porto, 2012.

[17] J-L Durrieu, B. David, and G. Richard, "A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation," *IEEE J. of Selected Topics in Signal Proc.*, vol. 5, no. 6, pp. 1180–1191, Oct. 2011.

[18] R. Marxer, J. Janer, and J. Bonada, "Low-Latency Instrument Separation in Polyphonic Audio Using Timbre Models," in *Latent Variable Analysis and Signal Separation*, Tel Aviv, Israel, 2012, pp. 314 – 321, Springer Berlin / Heidelberg.

[19] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 21, no. 1, pp. 71–82, Jan. 2013.

[20] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *2012 IEEE Int. Conf. on Acoustics, Speech and Signal Proc.* Mar. 2012, pp. 53–56, IEEE.

[21] Z. Rafii and B. Pardo, "Music/Voice Separation using the Similarity Matrix," in *Proc. 13th Int. Soc. for Music Info. Retrieval Conf.*, Porto, 2012, pp. 583–588.

[22] M. Goto and S. Hayamizu, "A Real-time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals," in *IJCAI-99 Workshop on Computational Auditory Scene Analysis*, Stockholm, 1999, pp. 31–40.

[23] S. Dixon, "Onset Detection Revisited," in *Proc. of the 9th Int. Conf. on Digital Audio Effects*, Montreal, 2006, pp. 133–137.

[24] J Laroche, "Efficient Tempo and Beat Tracking in Audio Recordings," *J. of the Audio Engineering Soc.*, vol. 51, no. 4, pp. 226–233, 2003.

[25] D. P W Ellis, "Beat Tracking by Dynamic Programming," *J. of New Music Research*, vol. 36, no. 1, pp. 51,60, Mar. 2007.

[26] S. Hainsworth and Malcolm M., "Onset Detection in Musical Audio Signals," in *Int. Computer Music Conf. (ICMC)*, Singapore, 2003, pp. 136–166.

[27] M. E. P. Davies, MD M. D. Plumbley, and Douglas Eck, "Towards a musical beat emphasis function," in *IEEE Workshop on Applications of Signal Proc. to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2009, pp. 61–64, IEEE.

[28] J. R. Zapata, M. E. P. Davies, and E. Gomez, "MIREX 2012: Multi Feature Beat Tracker (ZDG1 AND ZDG2)," in *the Music Info. Retrieval Eval. eXchange (MIREX 2012)*, Porto, 2012.

[29] M. E. P. Davies, N. Degara, and M Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Tech. Rep. October, C4DM-TR-09-06, Queen Mary University of London, Centre for Digital Music, 2009.

[30] S. W. Hainsworth and M.D. Macleod, "Particle Filtering Applied to Musical Tempo Tracking," *J. of Advances in Signal Proc.*, vol. 15, pp. 2385–2395, 2004.