EFFICIENT IMPLEMENTATION OF A SYSTEM FOR SOLO AND ACCOMPANIMENT SEPARATION IN POLYPHONIC MUSIC

Estefanía Cano, Christian Dittmar and Gerald Schuller

Fraunhofer Institute for Digital Media Technology
Ilmenau, Germany
{cano, dmr, shl}@idmt.fraunhofer.de

ABSTRACT

Our goal is to obtain improved perceptual quality for separated solo instruments and accompaniment in polyphonic music. The proposed approach uses a pitch detection algorithm in conjunction with a spectral filtering based source separation. The algorithm was designed to work with polyphonic signals regardless of the main instrument, type of accompaniment or musical style. Our approach features a fundamental frequency estimation stage, a refined harmonic structure for the spectral mask and a post-processing stage to reduce artifacts. The processing chain has been kept light. The use of perceptual measures for quality assessment revealed improved quality in the extracted signals with respect to our previous approach. The results obtained with our algorithm were compared with other state-of-the-art algorithms under SISEC 2011.

Index Terms— Separation, filtering, main melody, harmonics.

1. INTRODUCTION AND PREVIOUS WORK

Due to the immense popularity of karaoke and music video games, as well as the increasing interest in the development of music education tools, the capability to extract main melodies from musical recordings and subsequently obtain accompaniment tracks to play or sing along has gained a lot of attention in the research community.

In this context, some systems have specifically dealt with the problem of singing voice extraction from polyphonic audio. In [1] a system based on classification of vocal/non-vocal sections of the audio file, followed by a pitch detection stage and grouping of the time-frequency tiles was proposed. In [2], voice extraction is achieved by main melody transcription and sinusoidal modeling. A system based on pitch detection and non-negative matrix factorization (NMF) is proposed in [3]. Others have focused on the separation of harmonic from percussive components of an audio track [4], [5]. Similarly, a system is proposed in [6] to specifically address the extraction of saxophone parts in classical saxophone recordings. More general algorithms have also been proposed for main melody separation regardless of the instrument used: Durrieu in [7] proposes a source/filter approach with a two-stage parameter estimation and Wiener filtering based separation. In [8] Lagrange proposes a main melody extraction system based on a graph partitioning strategy - Normalized Cuts, sinusoidal modeling and computational auditory scene analysis (CASA).

The remainder of this paper is organized as follows: Section 2 describes the different stages of the proposed algorithm, Section 3 presents the evaluation scheme and results and in Section 4 we draw some conclusions and present future work.

2. PROPOSED SYSTEM

In this work, we aim to develop a system capable of separating main instruments from music accompaniment, regardless of the type of solo instrument used, musical genre of the track or type of music accompaniment. We focus on commercial recordings of polyphonic music. The algorithm does not require any prior information for processing and has been kept lightweight to allow real-time performance.

The system is composed of 5 building blocks shown in Fig. 1 and further described in the next subsections. For the remainder of this paper the following notation applies: Let $F_{k,n}$ be the Short-Term Fourier Transform (STFT) of a monaural signal $f(t)$ and $M_{k,n} = |F_{k,n}|$ its magnitude spectrogram, with k the frequency index and n the time index. We aim to decompose $M_{k,n}$ into a main melody/solo component $S_{k,n}$ and an accompaniment component $A_{k,n}$. The magnitude spectrogram of the audio signal is modeled as follows: $M_{k,n} = A_{k,n} + S_{k,n}$.

2.1. Pitch Detection

For this system, the pitch detection algorithm described in [9] is used. The performance of this algorithm was tested in the Audio Melody Extraction task within the Music Information

The Thuringian Ministry of Economy, Employment and Technology supported this research by granting funds of the European Fund for Regional Development to the project Songs2See, enabling transnational cooperation between Thuringian companies and their partners from other European regions.

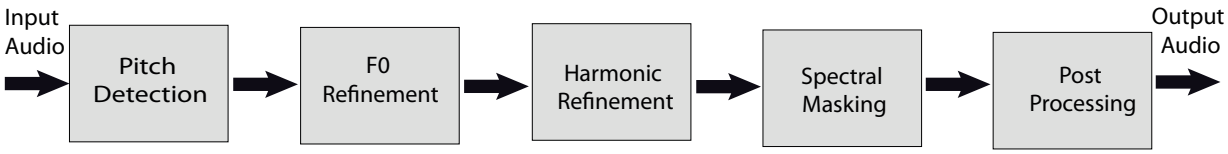

Fig. 1. Block diagram of the proposed system.

Retrieval eXchange (MIREX09) [1] obtaining the best results among all competing algorithms. During pitch extraction, an analysis frame of 46 ms is used in conjunction with a hopsize of 5.8 ms.

2.2. F0 Refinement

In order to improve the f0 (fundamental frequency) estimation delivered by the pitch detection algorithm, a refinement stage is proposed where the magnitude spectrogram is interpolated in a narrow band around each initial f0 value and its constituent harmonics. For a particular time frame *n*:

$$M_{i,n} = M_{k_1,n} + \frac{(f_{k_i} - f_{k_1})M_{k_2,n} - (f_{k_i} - f_{k_1})M_{k_1,n}}{f_{k_2} - f_{k_1}} \quad (1)$$

with interpolation step $i = 1, \ldots, i_{max}$, $f_{k_1} = f0/2^{(25/1200)}$ and $f_{k_2} = f0 \cdot 2^{(25/1200)}$ quarter tone deviations from the initial f0 location in Hz. For each interpolation step a cumulative magnitude sum is obtained and the maximum position is taken as an indicator of the new f0 value.

$$f0_n = \underset{i}{argmax}\,(E_i = \sum_{h=1}^{h_{max}} M_{H_h^i,n}) \cdot (\frac{f_{k_2} - f_{k_1}}{i_{max}}) \cdot f0 \quad (2)$$

with harmonic number $h = 1, \ldots, h_{max}$. The calculated harmonic location for each partial in each interpolation step is given by $H_h^i = f0 \cdot h \cdot k_i$.

2.3. Harmonic Series Refinement

After a refined estimate of the fundamental frequency has been obtained, the location of each harmonic component is also refined. The two underlying principles at this stage are: (1) Each harmonic component is allowed to have an *independent* deviation from the calculated ideal location of the harmonic, i.e., multiple integer of the fundamental frequency, and (2) the acoustic differences between the voice, string and wind instruments need to be considered when harmonic components are located. While no prior information is given to the algorithm regarding the instrument class, the harmonic refinement stage has to be kept consistent. Namely, inharmonicity characteristics differ between instrument families. A well known characteristic of conical bore instruments, for example, is the flattening of upper resonances in relation to the fundamental component due to open end corrections in the tone hole lattice [6]. To keep control of harmonic deviations, each partial is allowed a maximum deviation ρ_{max} from its harmonic location k_h of one quarter tone. This will guarantee that tones will remain perceptually harmonic. For harmonic numbers $h = 2, \ldots, h_{max}$, time frame n and $k_h - \rho_{max} \leq k \leq k_h + \rho_{max}$, we define a detection matrix $d_{k,n}$ such that :

$$d_{k_0,n}^{(h)} = \begin{cases} 1 & \text{for } k_0 = \underset{k}{argmax}\,(M_{k,n}) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

2.4. Spectral Masking

After the complete harmonic series has been estimated, initial binary spectral masks for the solo $\tilde{M}_{S_{k,n}}$ and accompaniment $\tilde{M}_{A_{k,n}}$ are created. At this stage, each time-frequency tile is defined either as part of the solo instrument or part of the accompaniment. To compensate for spectral leakage in the time frequency transform, a tolerance band Δ centered at the estimated location k_0, is included in the masking procedure. Thus, for a frequency range $k_0 - \Delta \leq k \leq k_0 + \Delta$ and time frame *n* we have:

$$\left(\tilde{M}_{S_{k,n}}, \tilde{M}_{A_{k,n}}\right) = \begin{cases} (M_{k,n}, 0) & \text{if } d_{k,n} = 1 \\ (0, M_{k,n}) & \text{otherwise} \end{cases} \quad (4)$$

2.5. Post-Processing

In contrast to the previous stages, which are performed on a frame by frame basis, the post-processing stage evaluates each tone as a whole. Two specific events are here addressed: (1) attack frames and (2) crosstalk of transients in solo signals.

The pitch detection algorithm requires a few processing frames before a valid f0 value can be detected. To compensate for this inherent delay, a region of 70 ms before the start

[1] Mirex: http://www.music-ir.org/mirex/wiki/2009:Main_Page

of each tone n_0, is searched for harmonic components that correlate with the harmonic structure of the tone. The binary masks $\tilde{M}_{S_{k,n}}$ and $\tilde{M}_{A_{k,n}}$ are modified accordingly to include the attack frames found for each tone.

Due to overlapping of spectral information from different sources, percussion hits are often detected as being part of a tone. Bearing in mind that percussion onsets are evident in the spectrogram as vertical events occurring in a defined time interval [4], a final analysis is performed, where sudden magnitude peaks occurring simultaneously in several harmonic components, are detected. As the perceptual impact of the percussion onsets is stronger for higher harmonics, this analysis is only performed for harmonics higher than h_{min}. For each harmonic $h \geq h_{min}$, a reference time trajectory, $t_h^{(n)} = \psi\left(M_h^{(n)}, L\right)$ is obtained, where ψ is a median filter of length L [5]. Let γ_p be the maximum magnitude variation allowed and min_h the minimum number of harmonics where the event should simultaneously occur, then if $\tilde{M}_{S_{k,n}} \geq t_h^{(n)}\gamma_p$ for at least min_h harmonics, then $\tilde{M}_{S_{k,n}} = t_h^{(n)}$ and $\tilde{M}_{A_{k,n}} = M_{k,n} - \tilde{M}_{S_{k,n}}$. The new spectral masks are no longer binary.

2.6. Re-synthesis

Finally, the complex valued spectrogram is masked and independent solo and accompaniment tracks are re-synthesized by means of the inverse Short-Term Fourier Transform (ISTFT). Thus, the solo track is $S_{k,n} = F_{k,n} \otimes \tilde{M}_{S_{k,n}}$ and the accompaniment track is $A_{k,n} = F_{k,n} \otimes \tilde{M}_{A_{k,n}}$, where $\otimes$ denotes elementwise multiplication. The output solo and accompaniment tracks are then $s(t) = ISTFT(S_{k,n})$ and $a(t) = ISTFT(A_{k,n})$.

3. EVALUATION

3.1. Experiments

Two different sets of experiments were conducted in this study: (1) The goal of the first experiment was to evaluate the contribution of each processing stage into the quality of the extracted signals. For this purpose, different versions of the algorithm were created where the different processing stages were bypassed and quality of resulting signals was evaluated. (2) The second set of experiments tested the full algorithm under SiSEC 2011 [10] conditions in the *Professionally Produced Music Recordings* task. For the SiSEC 2011, a dataset of multi-track recordings was made available and a common evaluation scheme was established to test different algorithms. The PEASSS Toolkit, described in Section 3.3, was used for evaluation of results. The goal of this experiment was to compare the performance of our algorithm with other state-of-the-art approaches.

3.2. Dataset

For the first experiment, three audio segments were selected from multi-track recordings. Table 1 presents the main characteristics of these signals. Signals 1 and 3 are publicly available [2], but due to copyright issues, Signal 2 is not publicly available. For the second experiment, seven songs from the SiSEC 2011 [10] dataset were used. Table 3 shows the signals used for this experiment, all publicly available in the campaign website [3]

3.3. Quality Measures

The use of objective measures based on energy ratios between the signal's components, i.e., Signal to Distortion Ratio (SDR), Image to Spatial Distortion Ratio (ISR), Signal to Interference Ratio (SIR) and Signal to Artifacts Ratio (SAR), has been the standard approach in the Sound Separation community to test the quality of extracted signals. However, due to the fact that these measures do not directly correlate to perceptual attributes, the results could potentially be misleading. In the attempt to provide a more suitable tool for testing separation results, the PEASS Toolkit - Perceptual Evaluation Methods for Audio Source Separation - has been developed [11]. This system proposes a family of four objective measures with the aim of predicting a set of subjective scores. The system makes use of auditory-motivated metrics to assess the perceptual salience of the target distortion, interference and artifacts. The family of objective measures is composed of the Overall Perceptual Score (OPS), the Target-related Perceptual Score (TPS), the Interference-related Perceptual Score (IPS) and the Artifacts-related Perceptual Score (APS).

3.4. Discussion

For the two experiments, the following processing parameters were used: $h_{max} = 25$, $\Delta = 1$, $h_{min} = 8$, and $\gamma_p = 1.2$.

3.4.1. Experiment 1: Contribution of processing blocks

The results of this experiment are presented in Table 2. The final signals can be accessed in our result website [4]. As shown in Table 2, for both the solo and accompaniment tracks, the highest OPS is obtained with the full algorithm in two of the three signals. However, in the case of *Scenaric*, where the highest OPS is not reached for the solo, and in the case of *Natmin*, where the highest OPS is not reached for the accompaniment, the scores of the full and the best variants are very close. These results also suggest that each one of the different processing stages contributes somehow to a perceptual gain

[2] http://bass-db.gforge.inria.fr/BASS-dB/?show=browse&id=mtracks

[3] http://sisec.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings

[4] http://www.idmt.fhg.de/eng/business%20areas/music_performance_applications.htm

Signal Num.	Name	Solo Instrument	Accompaniment	Duration [sec]
1	Natmin	Electric Guitar	Acoustic Guitar 1 & 2, Distorted Guitar, Bass, Bongos	14
2	Track 36	Alto Saxophone	Piano, Bass, Drums	30
3	Scenaric	Electric Guitar	Drums, Synth. Bass, Synth. Lead, Rhythm Guitar 1 & 2	14

Table 1. Data set for Experiment 1: signals used to evaluate the contribution of the different processing blocks of the algorithm

	1. Natmin				2. Track 36				3. Scenaric			
	OPS	**TPS**	**IPS**	**APS**	**OPS**	**TPS**	**IPS**	**APS**	**OPS**	**TPS**	**IPS**	**APS**
1. S_Full	44.194	34.184	77.056	45.043	19.766	20.815	57.596	18.138	22.936	24.914	57.286	25.729
2. S_noF0	35.823	30.552	69.473	41.702	15.555	18.615	47.364	15.815	23.559	25.024	56.525	28.155
3. S_noHS	42.622	29.287	74.922	45.503	17.211	17.749	52.436	16.108	22.913	13.326	60.763	21.629
4. S_noPP	10.542	10.760	46.873	6.236	14.414	17.759	54.366	9.748	8.574	15.847	33.639	6.636
5. A_Fulll	12.087	42.824	49.971	8.357	17.992	40.164	57.734	15.263	16.875	24.029	39.754	26.836
6. A_noF0	12.126	44.548	45.681	10.139	16.780	46.947	50.570	17.673	16.403	21.609	35.166	30.216
7. A_noHS	12.558	45.2057	47.575	10.180	17.352	48.110	51.909	18.115	14.802	25.608	30.001	30.658
8. A_noPP	12.234	51.391	56.699	6.518	16.962	39.606	55.169	14.827	15.534	26.2541	43.2442	19.474

Table 2. Resulting measures for all signals in Experiment 1. Signal variants: Full - all processing blocks included, noF0 - F0 refinement block removed, noHS - Harmonic refinement block removed, noPP - Post-Processing block removed. Signals 1-4 are the solo variants and signals 5-8 the accompaniment variants.

in the extracted signals. The biggest perceptual improvement for the solo signals (Signals 1-4) is obtained with the Post-Processing block as for the noPP variants, the lowest OPS and APS scores are always obtained. For the accompaniment tracks (Signals 5-8), the lowest APS scores are always obtained for the noPP but as can be seen, this is not the case for the OPS. Our goal is to build an algorithm that will have solid performance regardless of the signal used. This represents a big challenge given the great variability of signals. In this sense, we aim not to obtain the highest OPS scores at all times with the full algorithm, but to make sure that for those cases where the highest OPS score is not reached with full processing, the difference in terms of perceptual quality with respect to the best variant is minimum. The accompaniment tracks show in general lower OPS scores compared to its corresponding solo tracks. There are also some cases where the variant that obtains the highest scores for the solo signal, also obtains the lowest scores for the accompaniment or vice-versa, e.g., TPS of Track 36_noHS, IPS of Natmin_noPP, OPS of Natmin_Full, TPS of Scenaric_noF0. This invalidates the assumption that better solo signal extraction necessarily translates into better accompaniment extraction and suggests that the effects of some of the processing stages might be increasing quality of the solo tracks at the expense of loosing quality in the accompaniment tracks and vice-versa. Such effects need to be further studied to guarantee a quality balance between the solo and the accompaniment at all times under the condition $M_{k,n} = A_{k,n} + S_{k,n}$

3.4.2. Experiment 2: SISEC 2011

For the *Professionally Produced Music Recordings* task in SiSEC 2011, a total of seven algorithms were submitted [10]. For sake of simplicity, we only present in Table 3 our results and compare them to the system proposed by Durrieu [12]—both have the same final goal: *main melody and accompaniment separation*. For particular details of the algorithms and the full table of results, we refer the reader to the campaign's website. It can be seen that our algorithm presents OPS values comparable to those obtained by Durrieu. Furthermore, both algorithms show in general high IPS values that suggest a successful isolation of the main melody. However, consistently lower values for the TPS and APS are obtained with our approach. This suggests that artifacts are still perceptually evident and work in detriment of the target source quality. It is important to point out that the main melody of all the signals in this dataset was the voice. Informal experiments have shown that our approach shows consistently lower results when dealing with the voice. This can also be seen in Table 2, where instrumental signals are used and more homogeneous scores are obtained. Another important fact in this analysis is time efficiency of the algorithms. While Durrieu's approach reports an average processing time of 600 sec per excerpt, our algorithm requires an average of 8 sec per excerpt. This represents 1/75 of the processing time.

4. CONCLUSIONS

We presented a system for the separation of main instruments from accompaniment in real world music recordings. With the use of the PEASS toolkit, we found that our refinements mostly resulted in an improved separation quality. The performance of the algorithm was compared to state-of-the-art approaches under the SiSEC 2011. Results show that our algorithm achieves a good balance between performance and efficiency. Further work needs to be conducted to improve

		Test Set							Development Set		
		Tamy [V G]		Bea. [V]	Phil. [V]	Nine [V]	Hur. [V]	Total	An. [V]	NZ [V]	Total
Cano	**OPS**	22.8	25.7	18.3	26.1	32.1	13.3	24.1	25.0	37.7	31.3
	TPS	7.0	55.7	5.0	1.7	0.4	11.1	30.4	1.6	0.3	1.0
	IPS	65.7	53.7	58.6	75.2	62.5	61.1	59.1	67.3	69.6	68.4
	APS	13.4	45.6	13.2	3.2	0.7	19.8	27.9	3.5	0.5	2.0
Durrieu	**OPS**	33.5	30.2	27.3	22.3	17.1	11.6	26.3	29.6	24.7	27.8
	TPS	40.0	79.6	33.1	35.4	22.5	13.2	54.2	43.0	39.7	42.2
	IPS	71.8	34.5	57.1	56.1	58.0	51.8	46.7	65.2	66.7	65.0
	APS	37.9	57.7	35.7	34.0	25.4	21.0	44.3	35.1	27.5	32.5

Table 3. Results of Experiment 2 - SiSEC 2011: the algorithm presented by Durrieu [12] is compared with the proposed algorithm (Cano). In the Table, [V] represents a voice signal and [G] a guitar signal. Short names are used to represent the different signals: **Tamy**- Tamy: Que Pena Tanto Faz, **Bea**- Bearlin: Roads, **Phil**- Glen Philips: The Spirit of Shackleton, **Nine**- Nine Inch Nails: The Good Soldier, **Hur**- Shannon Hurley: Sunrise, **An**- Another Dreamer: The Ones We Love, **NZ**- Ultimate NZ Tour.

the quality of voice signals. The development of quality measures is still work in progress. Furthermore, the concept of quality is strongly dependent on the application considered. For this reason, the reader is always advised to contrast the numerical results by listening to the resulting signals.

5. REFERENCES

[1] Yipeng Li and DeLiang Wang, "Separation of Singing Voice from Music Accompaniment for Monaural Recordings," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1475–1487, 2007.

[2] Matti Ryynänen, Tuomas Virtanen, Jouni Paulus, and Anssi Klapuri, "Accompaniment Separation and Karaoke Application Based on Automatic Melody Transcription," in *IEEE International Conference Multimedia Expo*, Hannover, Germany, 2008, pp. 1417–1420.

[3] Tuomas Virtanen, Annamaria Mesaros, and Matti Ryynänen, "Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music," in *ISCA Tutorial and Research Workshop on Statistical and Peceptual Audition (SAPA)*, Brisbane, Australia, 2008, pp. 17–22.

[4] Nobutaka Ono, Kenichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka, and Shigeki Sagayama, "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram," in *EUSIPCO*, Lausanne, Switzerland, 2008, pp. 1–4.

[5] Derry Fitzgerald, "Harmonic/Percussive Separation Using Median Filtering," in *13th International Conference on Digital Audio Effects (DAFx -10)*, Graz, Austria, 2010, pp. 1–4.

[6] Estefanía Cano and Corey Cheng, "Melody Line Detection and Source Separation in Classical Saxophone Recordings," in *12th International Conference on Digital Audio Effects (DAFx-09)*, Como, Italy, 2009, pp. 1–6.

[7] Jean-Louis Durrieu, Gaël Richard, and Bertrand David, "An Iterative Approach to Monaural Musical Mixture De-Soloing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, pp. 105–108.

[8] Mathieu Lagrange, Luis Gustavo Martins, Jennifer Murdoch, and George Tzanetakis, "Normalized Cuts for Predominant Melodic Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 278–290, 2008.

[9] Karin Dressler, "Pitch Estimation by pair-wise Evaluation of Spectral Peaks," in *Proceedings of the AES 42nd Conference on Semantic Audio*, Ilmenau, 2011, pp. 278–290.

[10] Shoko Araki, Francesco Nesta, Emmanuel Vincent, Zbynk Koldovsk, Guido Nolte, Andreas Ziehe, and Alexis Benichoux, "The 2011 signal separation evaluation campaign (sisec2011): - audio source separation -," in *Latent Variable Analysis and Signal Separation*, Fabian Theis, Andrzej Cichocki, Arie Yeredor, and Michael Zibulevsky, Eds., vol. 7191 of *Lecture Notes in Computer Science*, pp. 414–422. Springer Berlin / Heidelberg, 2012.

[11] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[12] Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180–1191, 2011.