SCORE-INFORMED AND TIMBRE INDEPENDENT LEAD INSTRUMENT SEPARATION IN REAL-WORLD SCENARIOS

Juan J. Bosch[1], Kazunobu Kondo[1], Ricard Marxer[2] and Jordi Janer[2]

[1] Corporate Research & Development Center
Yamaha Corporation
203 Matsunokijima, Iwata-shi,
Shizuoka-ken, 438-0192, Japan

[2] Universitat Pompeu Fabra,
Music Technology Group,
Roc Boronat 138, Barcelona

ABSTRACT

We present a method for lead instrument separation using an available musical score that may not be properly aligned with the polyphonic audio mixture. Improper alignment degrades the performance of existing score-informed source separation algorithms. Several techniques are proposed to manage local and global misalignments, such as a score information confidence measure, and a chroma based MIDI-audio alignment. The proposed separation approach uses time-frequency masks derived from a pitch tracking algorithm, which is guided by the MIDI file's main melody. Timbre information is not needed in the present approach. An evaluation conducted on a custom dataset of stereo convolutive audio mixtures showed significant improvement using the proposed techniques compared to the non score-informed separation.

Index Terms— Timbre independent source separation, score-informed source separation, MIDI-audio alignment, lead instrument separation

1. INTRODUCTION

Audio source separation deals with the problem of recovering the original signals from a mixture by computational means. Its application in the musical domain is a complex task which has been the object of much research on the last two decades. However, the results obtained in real world musical signals evidence that there is still much room for improvement. In order to enhance their performance, musical source separation algorithms may exploit the knowledge of additional information such as the instrumentation [1], score information [2], or the position of the sources in the stereo image.

This work is focused on the separation of the lead instrument in stereo convolutive audio mixtures with the guidance of their score (available in MIDI format) and without any knowledge about the timbre of the solo and background instruments. In real world situations, it is possible to have only the melody line of the instrument to be separated either in available MIDI files, or through manual input of the notes. The proposed score-based separation approach relies only on the main melody line, and uses as a basis the time-frequency masking separation algorithm proposed by Marxer [1]. The main difference to this contribution is that in our scenario, there is no knowledge available about the timbre of the target instrument, and therefore no supervised model is used for the separation. Previous score-informed source separation approaches such as [2] assume that the MIDI and the audio are properly aligned. However, this is not the common case in real world situations, where global and local misalignments between the score and audio can be found.

Global misalignments are here considered to be due to differences in tempo which affect all instruments. These are common in real world scenarios, where a piece is interpreted with different tempi, as in cover versions and remixes.

On the other hand, local misalignments are here understood as the time difference between the score and the real performance of the target instrument at both onset and offset of the notes. These may be produced by: 1) the interpretation of the piece by a human (including variations in the execution), 2) the time envelope of the instrument, mainly the attack and decay, and 3) mixing effects on the instrument to be separated, such as delay, echo or reverb.

Cont [3] or Dixon [4] present real-time audio-score alignment, but with only few exceptions such as Duan [5], offline alignment techniques have been typically used for score informed source separation algorithms. Most of the previous approaches render the MIDI into audio [6], and then perform audio-to-audio alignment based on several techniques. However, the results typically depend on the timbre similarity between the synthesizer used and the target instrument. In our scenario no timbre information is available a priori, and therefore several generic techniques are here proposed to deal with MIDI-to-audio alignment, and the subsequent score-informed separation.

2. SCORE INFORMED SEPARATION: OVERVIEW

This section describes the proposed approach to consider the score information in the separation algorithm. The schema of the whole system is depicted in Fig. 1.

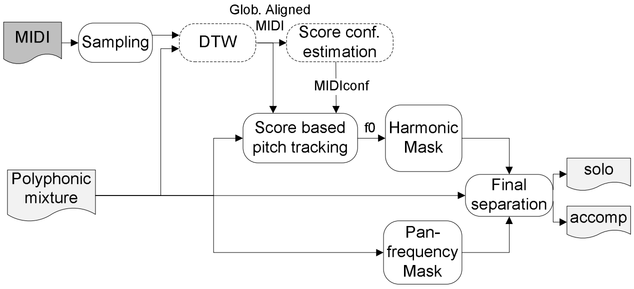

Fig.1. Score-informed source separation schema. The optional processes (score-audio alignment, calculation of the MIDI confidence measure) are marked with a dashed line.

The MIDI file is first sampled at the same frame rate at which the separation algorithm runs. The following two processes can be bypassed or executed. The first is the Dynamic Time Warping (DTW), used to deal with the global misalignments by synchronizing the score to the input audio mixture, as will be introduced in section 3. The second process is the estimation of the score confidence measure (MIDIconf), used to deal with small scale misalignments. This confidence measure is derived for each frame, and used along with the pitch information to guide the predominant instrument pitch tracking. The low-latency separation is based on [1], combining harmonic masks derived from the estimated pitch of the target instrument (f0) and pan-frequency masks, under the assumption that most target sources in the mixture present time-frequency orthogonality.

3. CHROMA BASED DTW AGAINST GLOBAL MISALIGNMENTS

As previously introduced, score information is commonly not properly synchronized with the mixture to be separated due to differences in tempo. This section introduces techniques to deal with such issues based on chroma information, which has been extensively used for a number of tasks such as cover version identification [7] or audio-to-audio alignment. We propose MIDI-to-audio alignment methods, deriving a chroma mask from the score, without using any knowledge about the instrumentation.

The chromagram mask is created from the MIDI score in a similar fashion as proposed by Ellis with spectrogram masks [8], by mapping each MIDI note to its pitch class. As a result of not rendering the score information into audio, a binary mask created by directly translating notes into their pitch class does not account for properties of the instruments that can be relevant to perform a proper alignment. Some usual differences between the chromagram of a real instrument and a binary mask created from the MIDI score come from slow attacks, or longer release times which may also be extended due to reverberation. In order to account for these factors, we investigate the use of a non binary mask, extending the notes length with an ascending value of the energy of the pitch class in the attack, and a descending value in the decay. The chromagram of the audio mixture is created with the chroma toolbox [9], and the alignment is performed using Dynamic Time Warping (DTW). The globally aligned score is used for the pitch tracking, along with a measure of the confidence in the score.

4. SCORE CONFIDENCE MEASURE AGAINST LOCAL MISALIGNMENTS

Local misalignments commonly found in real performances may significantly affect the separation performance if there is full confidence on the score. The MIDI confidence is introduced to deal with such problems, by considering that the information from the score should not be trusted around the transitions of the notes, and fully trusted at sustained portions of the notes or silences. The values of the MIDI confidence are defined in the interval [0,1].

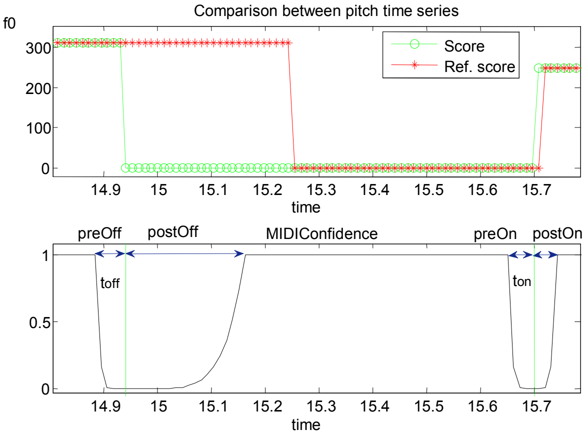

Fig.2. Green circles correspond to pitches derived from the original score, and red asterisks from the manually aligned (reference) score. The MIDI confidence function (below) is calculated from the score transition positions. Misalignments are partially covered in the uncertainty region.

The lowest values representing untrusted information are assigned to the frames around the note onsets and offsets, whereas the highest values are given to the trusted sustained portions. A symmetrical shape of the MIDI confidence is proposed in the onset, as depicted in Fig. 2. However, an asymmetrical distribution is used in the offset in order to deal with the fact that notes are usually present in the mixture for longer times than notated in the score due to the interpretation, the sustain of the instrument itself, or mixing effects (e.g. echo or reverb).

The MIDI confidence measure (MIDIConf) is used to weight the score derived probability considered in the pitch tracking algorithm described in section 5 with a factor related to the vicinity to a note transition. The best separation results have been observed when the length of the uncertainty region is adjusted to the characteristics of the mixture: time envelope of the lead instrument, characteristics of the performance and production effects. In our tests, no assumptions are made for each mixture, and heuristically determined values and curve shapes have been selected in the score confidence function:

$$MIDIConf\left(t\right)=\begin{cases}\left(\left(\left|t-t_{on}\right|\right)/T_1\right)^{deg}, & \left|t-t_{on}\right|<T_1\\ \left(\left(\left|t_{off}-t\right|\right)/T_1\right)^{deg}, & t_{off}-t<T_1\\ \left(\left(\left|t-t_{off}\right|\right)/T_2\right)^{deg}, & t-t_{off}<T_2\\ 1, & otherwise\end{cases} \tag{1}$$

The selected values are: deg = 6.322, T_1 = preOff = preOn = postOn = 6 frames (70 ms), and T_2 = postOff = 20 frames (232 ms), t_{on} is the time of the nearest onset, and t_{off} the time of the nearest offset. If several conditions are met simultaneously, the minimum confidence value is selected.

5. SCORE-BASED LEAD INSTRUMENT PITCH TRACKING

The time series of the pitch of the target instrument derived from the aligned score cannot be used directly to create the harmonic masks for the separation due to the pitch fluctuation in real performances and non abrupt pitch transitions between successive notes (e.g. slides, glissandos). The score-derived pitch (f_{score}) should however be used as a guidance to the lead instrument pitch tracking algorithm, along with the score confidence measure. In our strategy, a dynamic programming algorithm is used to estimate the sequence of pitches corresponding to the target instrument. Four candidate pitches (f_{cand}) and their likelihood are firstly estimated for each audio frame, and a two step Viterbi algorithm is employed to select either one of the candidate frequencies, or none of them in the case that the frame is predicted as not having the presence of target instrument. For each node, a set of probabilities is computed in natural logarithmic terms (maximum probability is zero), based on: score information (P_{midi}), pitch likelihood (P_{fo}), and frequency continuity (P_{jump}). Assuming probabilistic independence, the node probability is $P_{fo}+P_{midi}$, and the transition probability P_{jump}. An incremental forward pass and a backtracking pass are conducted to find the most likely sequence of states in both steps.

In the first Viterbi step, the best sequence of f0 candidates is selected, following equations (2) to (9). The frequency distance in semitones between f_1 and f_2 is $\Delta(f_1,f_2)$, as presented in (2). Equation (3) represents the natural logarithm of a Gaussian (G), in which μ represents the mean, and σ the variance. Both values are heuristically determined for all the equations using (3). Equation (4) represents the difference in octaves (N_{oct}) between f_{score} and f_{cand} . Finally, P_{midi} is calculated as the maximum of P_1 and P_2 . P_1 in (5) gives higher probabilities to the pitches around f_{score}, and P_2 in (6) considers octave errors in the pitch estimation, giving higher probabilities to the pitch candidates around the lowest multiples of f_{score}.

$$\Delta(f_1,f_2)=12\left|\log_2\left(f_1/f_2\right)\right| \tag{2}$$

$$\ln G(x,\mu,\sigma)=-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2 \tag{3}$$

$$N_{oct}\ =\ \text{round}\left(\Delta(f_{score},f_{cand})/12\right) \tag{4}$$

$$P_1=\ln G\left(\Delta(f_{score},f_{cand}),0,5\right) \tag{5}$$

$$P_2=\max\left(\ln G\left(\Delta(f_{score},f_{cand}),12,0.8\right),-0.8\right)\cdot\left(1+N_{oct}\right) \tag{6}$$

$$P_{midi}=\max\left(P_1,P_2\right) \tag{7}$$

P_{fo} is proportional to the pitch likelihood of the candidates, where x represents the likelihood of each candidate divided by the maximum of all candidate likelihood values:

$$P_{fo}=\ln G\left(x,1,0.4\right) \tag{8}$$

P_{midi} and P_{fo} are node probabilities, and P_{jump} is a transition probability, inversely proportional to the distance between consecutive candidate pitches. Only distances between 0.5 and 6.5 semitones are considered:

$$P_{jump}=\ln G(min(6,max(0,\Delta(f_{cand},f_{cand}')-0.5)),0,4) \tag{9}$$

After the first Viterbi step, one f0 is found per frame. In the second Viterbi step, the best path is found in a matrix with two states corresponding either to the found pitch (“f0” node), or to no pitch (“0” node). P_{fo} is calculated similarly as in the first step, and P_{midi} is now defined as:

$$\begin{aligned}P_{midi}&=\begin{cases}-10, & f_{score}=0\\ 0, & otherwise\end{cases}\text{, in the “f0” node}\\ P_{midi}&=\begin{cases}-10, & f_{score}\neq 0\\ 0, & otherwise\end{cases}\text{, in the “0” node}\end{aligned} \tag{10}$$

The transition probability is defined in (11), where Δf is the difference in semitones between consecutive pitches:

$$P_{jump}=\ln G(max(0,\Delta f-0.5),0,6) \tag{11}$$

The score confidence measure can be used to modify P_{midi} in both steps. In our tests, we modify it only in the second step as follows: $P_{midiConf} = P_{midi} \cdot MIDIconf$. This leads to an increasing probability of the "f0" node before a note start, and a decreasing probability when the note ends.

The following process is the creation of a harmonic mask in (12) to mute a source, derived from the f0 in each frame:

$$m_i^h[f] = \begin{cases} 0, (\text{f0}_i \cdot \text{h}) - \text{L}/2 < f < (\text{f0}_i \cdot \text{h}) + \text{L}/2, \forall \text{h} \\ 1, \textit{otherwise} \end{cases} \quad (12)$$

where f0_i is the pitch of the ith frame, and L the width in bins to be removed around the partial position [1]. This mask can additionally be combined with a pan-frequency mask. In the case that no pitch is selected in the second Viterbi step, no separation is conducted. The length of the uncertainty region in the MIDI confidence function influences thus the start and end of the separation. Finally, the output signal is estimated from the filtered spectra of each frame, resynthesised using the ISTFT (Inverse Short Time Fourier Transform). The solo signal is estimated similarly, using the inverse mask.

6. EVALUATION

The quality of the separation achieved with the proposed techniques is evaluated with the objective measures provided by the BSS_EVAL toolbox [10]. These measures are: SDR (Source to Distortion Ratio), SIR (Source to Interference Ratio), ISR (source Image to Spatial Distortion Ratio) and SAR (Signal to Artifacts Ratio). An additional measure (%f0) is used to evaluate the alignment as a previous step to the separation. It is calculated as the proportion of frames in the aligned score which have the same f0 as in the ground truth score. It is important to note that the ground truth score is not the score given to the musicians, but the same score in which the position and length of the notes have been manually adjusted to the human interpretation for each of the solo instruments, considering also the instrument decay, and mixing effects.

Two kinds of experiments have been conducted on a dataset of song excerpts created for this research. The first set of experiments (S1) deals with the separation of the lead instrument in a mixture given the score of the song. The second set of experiments (S2) deals with the separation of the lead instrument in a mixture, given the score of another version of the same song, in which lead and accompaniment instruments are arranged differently, and with different orchestration. Tests were executed with a sampling rate of 44.1 kHz, window size of 4096 and hop size of 512 samples.

6.1. Datasets

We created the datasets used for both sets of experiments S1 and S2. Two songs have been composed, interpreted and produced, in several versions and with several solo instruments. The first song: "Smile" presents two versions, and the second song "Harusora" three versions. The scores of each of the versions are different but still similar, with different arrangements, accompaniment instruments and tempi (between 110 and 128 bpm). 13 excerpts with duration between 15 and 30 seconds have been extracted from the versions of the songs. The lead instrument is centrally panned, and played by a human in order to be more realistic. The accompaniment is spatially distributed and has been produced with several sound libraries. Five mixtures per excerpt are considered, corresponding to the instruments playing the main melody score: guitar, lead guitar, violin, saxophone and voice. This gives a total of 65 excerpts to be separated with guidance of different scores depending on the dataset.

The datasets for S1 have been created by modifying the score with tempo changes. Three sets have been created: D1 contains the scores without any modification; D2_X contains the scores modified to have a number of beats per minute (bpm) equal to X = {85, 145} which represent maximum tempo changes in the interval: 66-132%. Finally, D3 contains multiple tempi in each of the songs, within the same maximum change percentage interval. In the case of S2, the dataset consists of the scores of one version of the song being used for the separation of another version, and with two different tempi {85, 145} (bpm) not corresponding to any of the mixtures.

6.2. S1: Separation using own score at different tempi

The following notation has been used for the experiments in Table 1: Exp: name of the experiment (e.g: E1), C: use of MIDI Confidence – F (full confidence), V (Variable confidence); Mask: type of chroma mask used – NM (No mask), Mel (mask derived from the melody score), All (mask derived from all instruments score), Wall (mask derived from all instruments score, with an extra weight on the melody information), B (Binary mask), N (Non-binary mask). The evaluation measures: %f0 and SDR (dB) are computed as a mean of the values of all excerpts and configurations in each dataset.

E1 represents the baseline separation, with no MIDI information: P_{midi} is not considered in the Viterbi algorithm, and in every frame the predominant pitch is used for the separation. If the separation is conducted with the ground truth alignment, the following upper bound for the separation performance is obtained: SDR-solo = 6.31dB, and SDR-accomp = 10.46dB. The results show that with full confidence on the original MIDI (E2), we gain around 1dB in the solo and accompaniment with respect to E1 if the score is properly aligned (D1). Using a varying MIDI confidence (E3) results on an increase of more than 1dB in the original dataset (D1) compared to E2. If the score information considered is not properly aligned, worse results

			S1 - D1		S1 - D2		S1 - D3		S2				
E	C	Mask	%f0	SDR(dB)	%f0	SDR(dB)	%f0	SDR(dB)	%f0	SDR(dB)	SIR(dB)	SAR(dB)	ISR(dB)
E1	-	-	-	7.69/3.72	-	7.69/3.72	-	7.69/3.72	-	7.20/3.24	16.96/7.57	7.68/4.59	13.04/14.44
E2	F	NM	75.0	8.92/4.76	36.0	6.68/2.51	49.6	7.66/3.50	37.8	6.79/2.83	9.85/11.95	9.86/1.06	20.75/6.39
E3	V	NM	76.3	10.35/6.20	36.0	7.41/3.25	49.6	8.83/4.67	37.8	7.58/3.62	11.86/11.91	9.68/2.92	19.15/8.51
E4	V	All-B	83.9	10.32/6.16	83.4	10.30/6.15	76.3	10.05/5.90	75.8	9.72/5.77	17.23/13.50	10.81/6.11	18.90/13.83
E5	V	Wall-N	85.4	10.41/6.25	85.2	10.40/6.25	83.3	10.35/6.20	87.6	9.99/6.04	18.13/13.64	11.00/6.53	18.83/14.71
E6	V	Mel-N	82.9	10.32/6.17	82.0	10.29/6.13	81.8	10.27/6.12	85.1	9.92/5.97	17.96/13.55	10.96/6.41	18.83/14.54
E7	V	Mel-B	76.1	10.36/6.20	74.8	10.23/6.08	75.2	10.24/6.08	77.9	9.90/5.95	17.35/13.90	10.99/6.31	19.23/13.92
E8	V	Wall-B	79.1	10.45/6.30	78.8	10.45/6.29	78.0	10.38/6.23	81.3	10.00/6.04	17.71/13.96	11.06/6.46	19.22/14.27

Table 1. Results of experiments S1 and S2. The values for the BSS Eval measures represent the extracted accompaniment / solo. The SDR for each datasets is provided in S1. In the case of S2, details are provided for each BSS Eval measure.

compared to not using MIDI are obtained (D2 and D3 with E3). However, the use of the proposed MIDI-audio alignment methods provides robustness against differences in tempo, including not constant tempos, achieving very similar separation quality compared with the use of ground truth scores. The best results are obtained with weighted chromagram masks (E5 and E8), however, relying just on the melody track for alignment (E6) provides only slightly worse results. Generally, a better alignment produces better separation results except when the MIDI confidence is used. In that case, local misalignments in the low confidence area do not degrade the separation quality. The combination of MIDI confidence and the alignment is thus a complex matter which will be the object of further study.

6.3. S2: Separation using version score at different tempi

The accuracy of the alignment and detailed values of the BSS measures and is summarized in Table 1. The limiting factor is the artifacts produced by the separation (SAR). A further observation is that in the case that the arrangements are different, using all tracks with equal weight to create the chroma mask provides slightly worse results than performing the alignment using only the melody line against the mixture. In a similar fashion as in the set of experiments S1, the best separation results are obtained with the weighted versions of the chromagram mask (in E5 and E8), which also provide the best alignment accuracy (E5).

7. CONCLUSIONS

This work presents several techniques to improve the quality of the lead instrument separation results, with the guidance of the score of the musical audio, and with independence of the timbre of the instruments present. The results show that even in the case where only the score of the target instrument is known, the separation is considerably improved. Results are further improved with the MIDI confidence function, by dealing with local misalignments. Additionally, the proposed chroma based MIDI-to-audio alignment techniques provide robustness against global misalignments due to differences in the tempi, with similar separation results compared to the manually adjusted score.

Further work includes a more complete evaluation considering subjective aspects of the separation, and the investigation of a method to set score confidence values in relation to the alignment, thus being adapted to the characteristics of the mixture. The implementation of a completely online algorithm (or with a small latency) is also foreseen, by substituting the offline alignment DTW algorithm with an online version [12], or other approaches.

16. REFERENCES

[1] R. Marxer, J. Janer, and J. Bonada, "Low-latency Instrument Separation in Polyphonic Audio Using Timbre Models," in *Proc. 10th Int.Conf., LVA/ICA*, 2012, pp. 314-321

[2] C. Raphael, “A classifier-based approach to score-guided source separation of musical audio,” *Comput. Music J.*, vol. 32, no. 1, pp. 51–59, 2008.

[3] A. Cont, “A coupled duration-focused architecture for real-time music-to-score alignment,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 974–987, Jun. 2010.

[4] S. Dixon, “Live tracking of musical performances using on-line time warping,” in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Madrid, Spain, 2005, pp. 92–97.

[5] Z. Duan; B.Pardo, "Soundprism: An Online System for Score-Informed Source Separation of Music Audio," *IEEE Journal of Sel. Topics in Sig. Proces.*, vol.5, no.6, pp.1205-1215, Oct. 2011.

[6] J. Ganseman, G. Mysore, P. Scheunders, and J. Abel, “Source separation by score synthesis,” in *Proc. Int. Comput. Music Conf. (ICMC)*, New York, Jun. 2010.

[7] J. Serrà, E. Gómez, P. Herrera, and X. Serra. “Chroma binary similarity and local alignment applied to cover song identification,” *IEEE Trans. on Audio, Speech and Lang. Proces.*, vol. 16, no. 6, pp. 1138-1152, 2008.

[8] D. P. W. Ellis (2008). "Aligning MIDI scores to music audio", http://www.ee.columbia.edu/~dpwe/resources/matlab/alignmidiwav ,web resource, retrieved 20.02.2012.

[9] M. Müller and S. Ewert, “Chroma Toolbox: MATLAB Implementations for Extracting Variants of Chroma-Based Audio Features,” in *Proc. ISMIR*, 2011.

[10] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Proces*. vol. 14, no. 4, pp. 1462–1469, 2006.

[11] A. Cont, D. Schwarz, N. Schnell and C. Raphael, “Evaluation of real-time audio-to-score alignment,” in *Proc. ISMIR*, 2007.

[12] R. Macrae and S. Dixon, “Accurate real-time windowed time warping,” in *Proc. ISMIR*, 2010, pp. 423–428.