# SCORE-INFORMED LEADING VOICE SEPARATION FROM MONAURAL AUDIO

**Cyril Joder, Björn Schuller**
Institute for Human-Machine Communication, Technische Universität München, Germany
cyril.joder@tum.de, schuller@tum.de

## ABSTRACT

Separating the leading voice from a musical recording seems to be natural to the human ear. Yet, it remains a difficult problem for automatic systems, in particular in the blind case, where no information is known about the signal. However, in the case where a musical score is available, one can take advantage of this additional information. In this paper, we present a novel application of this idea for leading voice separation exploiting a temporally-aligned MIDI Score.

The model used is based on Nonnegative Matrix Factorization (NMF), whose solo part is represented by a source-filter model. We exploit the score information by constraining the source activations to conform to the aligned MIDI file. Experiments run on a database of real popular songs show that the use of these constraints can significantly improve the separation quality, in terms of both signal-based and perceptual evaluation metrics.

## 1. INTRODUCTION

Extracting the main melody from a musical signal can be of interest, for example, for the remixing of a musical piece or the creation of a 'play-along' version of a recording, in the context of karaoke or classical concertos. Whereas this task is quite natural to the human ear, the automated solving of such a separation problem is notoriously difficult.

In the past, many works have considered the separation of musical sources as a blind audio source separation problem, assuming only general knowledge about the sources, such as temporal and harmonicity priors [16] or timbre information [14]. On the other hand, audio source separation approaches which integrate specific information about the content of each recording (see [15] for example) have recently received a large interest. In the case of music, valuable information about the sources can be found in the score when it is available. Hence, the topic of score-informed source separation, exploiting a temporally aligned score, has recently emerged. The score information is used to initialize the parameters of a model, which are then re-estimated in order to precisely match the data. Several kinds of models have been proposed, such as a sinusoidal model [12], Nonnegative Matrix Factorization (NMF) [4] or Probabilistic Latent Component Analysis (PLCA) [6]. The model of [10] exploits MIFI syntheses of the score, and operates a trade-off between fidelity to the synthesized sound and to the actual data to be separated. In [1], a multipitch estimator is used to model the spectral shape of each note. The authors of [9] employ a parametric NMF model which estimates a constant harmonic structure for each source. The specific problem of extracting the main melody part has also been addressed in [7] with an NMF-like probabilistic model, where each note is represented as a harmonic template.

In the present work, we exploit the physically-motivated source-filter NMF model proposed in [2], which is specifically designed for the extraction of the leading voice. We take advantage of the aligned score through time and pitch constraints. These constraints are similar to the ones already applied in [8] to a source-filter NMF model. However, while the latter work exploits the information given by a multipich estimator, we make use of actual MIDI transcriptions of the pieces. We evaluate the benefit of the score-based information on a database of real data, composed of nine multi-track recordings of popular songs. The scores are constituted by real-life MIDI scores, which are synchronized using a state-of-the-art alignment algorithm [11]. Several signal-based and perceptual evaluation criteria are used and the results show that both the interferences and the separation artifacts are reduced thanks to the score information. Furthermore, the use of time-frequency constraints applied on the leading voice components allows for a multi-pass approach for the removing of the reverberated voice, which improves the perceived quality of the separated accompaniment.

The rest of this paper is organized as follows: in Section 2 we present the source-filter NMF model used as baseline system for blind leading voice separation. Section 3 explains how the aligned MIDI score is exploited in the proposed methods for score-informed leading voice separation. We finally report the performed experiments in Section 4, before drawing some conclusions.

## 2. BASELINE SYSTEM: BLIND SEPARATION

As baseline system for the blind separation of the leading voice, we use the model proposed in [2]. Let us now detail

the main features of this model.

### 2.1 Signal Model

Let $\mathbf{S}$ be the matrix representing the short-time power spectrum of the musical recording, which is assumed to be of single-channel nature. We suppose that this matrix can be decomposed as:

$$\mathbf{S} = \mathbf{S}^V + \mathbf{S}^A, \tag{1}$$

where $\mathbf{S}^V$ and $\mathbf{S}^A$ are the short-time power spectrum of the leading voice and of the musical accompaniment, respectively. Furthermore, a source-filter model is assumed for the solo part. Thus, the matrix $\mathbf{S}^V$ can be written as the element-wise product of a 'source' matrix $\mathbf{S}^{F_0}$ by a 'filter' matrix $\mathbf{S}^{\Phi}$:

$$\mathbf{S}^V = \mathbf{S}^{\Phi} \odot \mathbf{S}^{F_0}, \tag{2}$$

where $\odot$ denotes the element-wise product.

For the contributions $\mathbf{S}^A$, $\mathbf{S}^{F_0}$ and $\mathbf{S}^{\Phi}$, an NMF model is assumed. Each of these terms is modeled as the product of two nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$. The former is a dictionary of spectrum templates (in columns) and the latter contains the corresponding activation amplitudes over time. Finally, we have:

$$\mathbf{S} = \left(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right) \odot \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \mathbf{W}^A\mathbf{H}^A \tag{3}$$

The 'source spectral shapes' of matrix $\mathbf{W}^{F_0}$ are set to fixed harmonic combs with logarithmically-spaced fundamental frequencies, with 20 $F_0$ values per semitone (in order to take into account tuning variations or vibratos) between 100 Hz and 800 Hz. This range is sufficient for most popular songs. In order to estimate smooth filters, the elements of the filters dictionary $\mathbf{W}^{\Phi}$ are modeled as the combination of overlapping Hann windows (in the frequency domain). As default parameters, the size of the filter dictionary is 10 and the rank of the accompaniment decomposition is set to 40.

### 2.2 Separation Strategy

The leading voice separation procedure consists of several steps. First, the matrices of eq. (3) (except for $\mathbf{W}^{F_0}$) are estimated from the processed signal using an NMF optimization algorithm based on the Itakura-Saito divergence. From this first result, only the estimated main source activation matrix $\mathbf{H}^{F_0}$ is kept. It can be interpreted as the instantaneous 'power' of the corresponding fundamental frequencies. Since the signal of interest is the *leading* voice, it is assumed to correspond to the dominant pitch. However, in order to avoid spurious 'jumps' in the case where the voice stops or if another instrument has a higher energy in a distant pitch, a tracking algorithm is used to estimate the whole sequence of $F_0$ values for the leading voice.

In the second step, another estimation of the model (3) is performed, in which $\mathbf{H}^{F_0}$ is constrained so that only the values around the tracked pitch are allowed to be non-zero. Finally, the leading voice is reconstructed by Wiener filtering. The estimate $\hat{\mathbf{S}}^V$ of the short-time power spectrum is then given by:

$$\hat{\mathbf{S}}^V = \frac{\left(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right) \odot \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right)}{\left(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right) \odot \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \mathbf{W}^A\mathbf{H}^A} \tag{4}$$

and the time-domain signal is retrieved by inverse Fourier transform (using the phase of the original mixture) and overlap-add.

## 3. EXPLOITATION OF THE SCORE INFORMATION

We now explain how we exploit the additional information given by the aligned musical score.

### 3.1 Information Conveyed by the Score

By musical score, we designate a set of notes characterized by their pitch, onset time and duration. In this work, we additionally assume that the notes in the score corresponding to the leading voice can be discriminated from the other notes. This is the case in most score MIDI files, where the instruments correspond to different *tracks*. Hence, the score can provide valuable information for the leading voice separation task. However, the score employs a temporal scale (expressed in beats), whose correspondence with the actual time in second is in general both unknown and variable. Fortunately, systems for accurate music-to-score alignment have been proposed to overcome this problem [5, 11, 13].

The aligned score then provides the pitch, onset and offset time of the notes played in the musical piece. Nevertheless, some limitations have to be taken into account. In particular, the pitches of the score are expressed in semitones, which constitutes a coarser frequency resolution than the short-time Fourier transform representation of the audio. Furthermore, there can be various sources of imprecision or mismatch between the score and the actual recording. For example, vibratos can strongly alter the fundamental frequency of a note. There may also be transcription errors or different interpretations of the music. In particular, the synchronization between the instruments or voices in polyphonic music may not be perfect, yielding a temporal indeterminacy and thus a possible imprecision in the alignment. For these reasons, the information conveyed by the aligned score cannot be fully trusted at a precise level. Nevertheless, it can be used at a coarser level, so as to narrow the search for the voice components in the spectrogram. In this work, we use only the 'voice track' of the aligned score. Indeed, in most cases of popular music, no reference musical score exists and the available transcriptions can often resemble 'lead sheet' scores, which focus on the main melody and only describe the global harmony of the accompaniment.

### 3.2 Time and Pitch Constraints

We propose to exploit the score information through two types of constraints applied in the model (3). The first approach only makes use of the information regarding whether the leading voice is present or not in each frame.

This corresponds to the case where the pitch of the aligned score is not sufficiently reliable. This *temporal constraint* consists in forcing all the activations of the source element (contained in the matrix $\mathbf{H}^{F_0}$) to be equal to zero when the voice is known to be absent. A time tolerance window is allowed, in order to overcome the possible temporal imprecision of the score alignment. The voice is then considered as absent in a frame when no note of the aligned score is present inside a temporal window of length $\theta_t$. The value of this tolerance threshold is a trade-off between two goals. If it is short, one may 'miss' the voice in the case where the score is imperfectly aligned. On the other hand, a long tolerance window may result in the extraction of another instrument as the leading voice, when the latter is absent.

The second approach takes advantage of both, time and pitch information, on the aligned score. As previously, the constraint used consists in forcing zero values of the source activation matrix where the source is known to be absent. This implies the use of an additional tolerance threshold $\theta_f$ on the fundamental frequency, in order to limit the pitch imprecisions. Hence, a component $\mathbf{H}^{F_0}_{i,j}$ or the source activation matrix (corresponding to fundamental frequency $i$ in frame $j$) is allowed to be non-zero only if there is a note in the aligned score, of pitch $p$, onset time $t_1$ and offset time $t_2$, such that:

$$|p - i| \leq \theta_f \text{ and } t_1 - \theta_t \leq j \leq t_2 + \theta_t. \quad (5)$$

## 4. EXPERIMENTS

### 4.1 Database and Settings

The database used in this work is composed of nine separated-track versions of well-known popular songs, for which a MIDI transcription was found on the internet. The list of the songs is displayed in Table 1. Unfortunately, these data cannot be shared due to copyright restrictions. In all these pieces, the source of interest is a human voice. Some of the songs contain vocal harmonies (several vocal parts), which introduce an ambiguity about the determination of the main source. In some others, mistakes are found in the MIDI score, where some vocal parts are not transcribed. In the corresponding pieces, only an excerpt where these problems do not occur has been used. All the files were converted to mono signals with 44.1 kHz sampling rate. For each piece, the file to be processed was created by linearly mixing the leading voice with the accompaniment. This procedure is much simpler than the mixing phase of professionally processed music, which often involves additional filtering or dynamic range compression. However, it was necessary to ensure that the final mixtures perfectly correspond to the separated tracks.

The MIDI scores were aligned by the method presented in [11]. This results in very accurate alignment, and the imprecision between the recording and the synchronized MIDI are most of the time not noticeable. As this system is reported to detect almost all the notes within a 300 ms window around their actual position, we set the threshold $\theta_t$ to this value. The frequency tolerance threshold $\theta_f$ is heuristically set to 3 semitones.

| | Title | Original Artist |
|---|---|---|
| 1 | A Day in the Life | The Beatles |
| 2 | Genie in a Bottle | Christina Aguilera |
| 3 | I Heard it Through the Grapevine | Marvin Gaye |
| 4 | Is This Love | Bob Marley |
| 5 | Long Train Running | The Doobie Brothers |
| 6 | Sgt Pepper's Lonely Hearts Club Band | The Beatles |
| 7 | She's Leaving Home | The Beatles |
| 8 | Stop Me If You Think You've Heard This One Before | The Smiths |
| 9 | With a Little Help From My Friends | The Beatles |

**Table 1**. List of the songs in the database.

In the experiments, we compare the separation obtained with both systems described in Section 3 with the baseline system of Section 2. We also introduce an additional method, which exploits the temporal information of the aligned score for a post-processing of the baseline system. In this approach, a 'temporal mask' is applied on the leading voice estimate: when the voice is considered as absent (in the sense of Subsection 3.2), the corresponding signal frames are shifted to the accompaniment estimate.

The separation quality is measured by the criteria described in [3]. They comprise three signal-based and four perceptual metrics, namely the Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifacts Ratio (SAR), Overall Perceptual Score (OPS) and Target-, Interference- and Artifacts-related Perceptual Scores (TPS, IPS and APS respectively).

### 4.2 Results

The results of the evaluations, averaged over the nine pieces, are compiled in Table 2. One can first notice that the system exploiting the time-frequency constraint obtains the best results according to almost all the measures used. In particular, the average OPS of the leading voice estimates improves from 21.5 to 32.5 and the average SDR increases by 1.5 dB. This indicates that the proposed approach does improve the leading voice separation quality, since both the interferences and artifacts are reduced compared to the baseline system.

The use of the temporal indications of the score, which indicate when the leading voice is active, results in an improvement of the quality of both leading voice and accompaniment estimates. As expected, the interferences of the accompaniment in the leading voice estimates are greatly reduced with the 'temporal mask' post-processing of the baseline system. Hence, the average SIR increases from 8.4 dB to 11.0 dB. Moreover, the constraints detailed in Subsection 3.2, which forces the voice to be active only where the main melody is actually present, leads to a further improvement for most of the songs. The use of these constraints results in a more precise estimation of the spectral components of the NMF and, as a consequence, in a reduction of the artifacts. For instance, the average APS on the accompaniment parts increases from 59.0 to 63.8 with

| | SDR (dB) | | SIR (dB) | | SAR (dB) | | OPS | | TPS | | IPS | | APS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LV | Ac | LV | Ac | LV | Ac | LV | Ac | LV | Ac | LV | Ac | LV | Ac |
| Baseline | 5.8 | 9.1 | 8.4 | 12.4 | 15.2 | 19.3 | 21.5 | 37.0 | 39.5 | 62.9 | 50.8 | 55.6 | 31.4 | 50.6 |
| Baseline + Temporal Mask | 6.7 | 10.0 | 11.0 | 12.9 | 15.8 | 20.5 | 29.5 | **43.6** | 41.4 | 68.1 | 58.3 | 57.6 | 35.1 | 59.0 |
| Time Constraint | 7.0 | 10.3 | 11.5 | 13.3 | 16.1 | 20.8 | 31.6 | 43.3 | 45.4 | 67.9 | **58.9** | 57.5 | 37.4 | 62.6 |
| Time-Frequency Constraint | **7.3** | **10.5** | **11.9** | **13.7** | **16.9** | **21.5** | **32.5** | 42.9 | **46.4** | **68.3** | 57.9 | **58.1** | **39.9** | **63.8** |

**Table 2**. Average evaluation criteria, measured on the leading voice (LV) and accompaniment (Ac) parts. In boldface are the best value of each column.

time and frequency constraints.

A more precise representation of the SDR values for every tested song is displayed in Figure 1. This figure confirms that the use of the the information conveyed in the musical score is valuable. Indeed, in terms of SDR, the baseline system is outperformed by all the other approaches. An observation which can seem surprising is that in many of the pieces, the addition of the frequency constraint does not improve the SDR measure. This is explained by the efficiency of the tracking algorithm used for the determination of the fundamental frequency of the leading voice. Hence, when the leading voice is strongly dominant in the recording, this tracking does not need to be constrained. On the other hand, the constraint has a visible effect on recording no. 4: *Is This Love*. Indeed, this song contains background vocals which can incidentally be tracked as main voice, when the lead singer is not dominant (for example in the case of breaths). Hence, the global average SDR slightly increase from 8.7 dB to 8.9 dB.

The OPS criterion measured on the database is displayed in Figure 2. In general, this metric exhibits the same tendencies as the SDR. However, there are some noticeable differences concerning the accompaniment estimates. Indeed, for the first three songs, the best OPS is obtained with the original system. A more specific analysis reveals that these correspond to cases where the score does not perfectly match the performance.

One of the main sources of deviation is the length of the notes in the MIDI score. Indeed, whereas the note onsets can be relatively well defined, determining the offsets is a notoriously hard problem, which can even be ill-posed. The score often indicates how the notes are to be *played*, which can actually be different from how the notes are *heard* in the recording, mainly because of the reverberation phenomenon (which is often increased by artificial effects). This phenomenon is strongest in song no. 1 *A Day in the Life*. In this piece, with the proposed constraints, the voice is 'cut' at the end of some musical phrases, because it is considered as absent while it can still be heard in the recording. This phenomenon is not prominent from the 'signal' point of view: indeed, the SIR criteria measured on the accompaniment estimates of this piece are 16.3 dB with the time-frequency constraint and 15.6 dB with the baseline system. However, this results in intermittent 'bursts' of voice in the accompaniment part, which is more strongly penalized by the perceptual measures. Hence, the value of the IPS degrades from 60.4 to 50.9.

In the songs no. 2 *Genie in a Bottle* and no. 3 *I Heard it Through the Grapevine*, this note length problem is also visible. Besides, the lead singer sometimes adds 'ornaments' to the transcribed score, in particular through 'vocalises', which are common in the *soul music* style. Hence, both time and frequency priors indicated in the MIDI file can be misleading at some point. As previously, this does not have a large influence on the signal-based measures, since the SIR of the accompaniment estimate only decreases from 13.3 dB to 12.6 dB. However, the perceptual importance of these separation errors is greater: the OPS drops from 45.3 to 34.1.

### 4.3 Constrained Second Pass

In order to reduce the problem caused by the reverberation of the leading voice, we experimented with the use of a second pass of the separation algorithm. Indeed, the reverberation often introduces 'polyphony', in the sense that several notes of the leading voice can be present at the same time in the recording. Since the separation model is inherently monophonic, because it is motivated by the physics of voice production, a multi-pass approach is needed for the handling of several simultaneous notes.

Hence, after the first separation with the time-frequency constraint, we apply the same algorithm on the accompaniment estimate, where some reverberated leading voice is supposed to remain. In this second pass however, the time tolerance for the offset is modified, so that each note is allowed to be active for 800 ms after its annotated extinction in the synchronized MIDI score. The threshold for the onset time is unchanged. The estimated reverberated voice is then added to the voice estimate of the first separation.

Figure 3 displays the influence of this approach on the OPS criteria. While it has little effect on the OPS of the leading voice estimate, it can recover from some of the previously described problems of the accompaniment. Indeed, on six of the nine tested pieces, the two-pass separation visibly increases the OPS. Furthermore, the use of this approach leads to an improvement on every piece compared to the baseline system, except for song no. 9 *With a Little Help From My Friends*, where the score is equivalent.

More thorough analysis reveals that the second pass actually slightly degrades the leading voice estimates, according to many evaluation metrics. In particular, it adds some interference in the vocal track, since in many places, the reverberated voice is dominated by the accompaniment. Thus, the average SIR decreases from 11.9 dB to

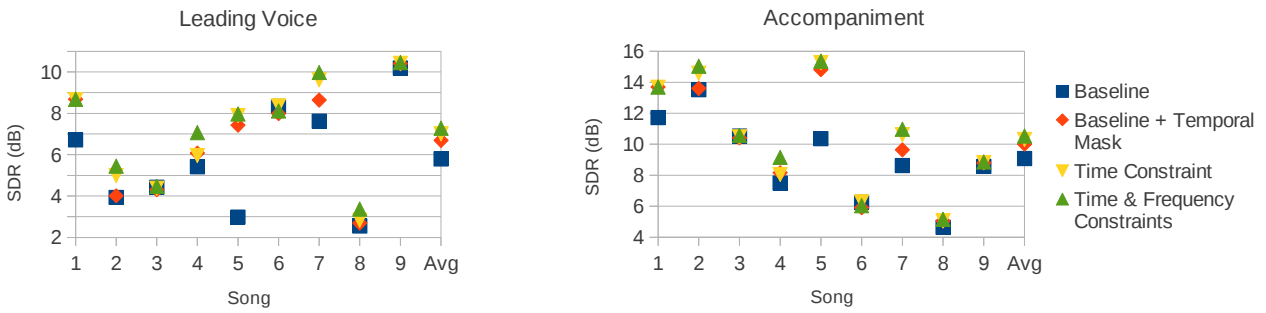


**Figure 1**. Signal to Distortion Ratio (SDR) measured on each of the tested songs and average.

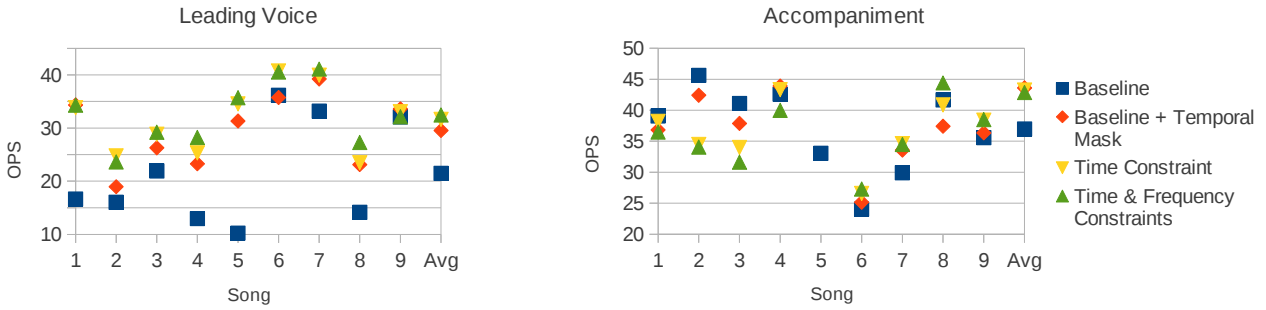


**Figure 2**. Overall Perceptual Score (OPS) measured on each of the tested songs and average. For the accompaniment of song no. 5, the extraction is nearly perfect, since all proposed methods obtain an OPS of 99 (not represented here).

9.6 dB. However, the artifacts are somewhat reduced and the voice seems to be better preserved. Hence, the average TPS improves from 46.4 to 53.7.

The opposite effect is observed on the accompaniment estimates, since more artifacts are measured. Hence, the average APS decreases from 63.8 to 55.3. However, these artifacts, which are very limited in terms of signal energy (the average SAR is 21.2 dB), are counterbalanced by the reduction of the interferences: the average SIR increases from 13.7 dB to 15.0 dB.

## 5. CONCLUSION

In this work, we exploited of a time-aligned MIDI file to perform a score-informed separation of the leading voice from a musical recording. The source-filter model assumed for the leading voice allowed for a natural use of the score information, by means of time and frequency constraints on the source components. We evaluated the usefulness of these constraints on a database of real recordings of popular songs and corresponding MIDI scores.

The results show that the score-guided constraints applied to the model not only reduce the interferences of the accompaniment in the leading voice separated track, but also allow for a more accurate estimation of the spectral shapes of all the components. Hence, this results in a reduction of the separation artifacts on both leading voice and accompaniment estimates. These improvement can be measured with perceptual metrics as well as signal energy-based criteria. Furthermore, a two-pass approach is made possible by the time-frequency constraints on the voice components. This allows for the removal of the remaining reverberated voice in the accompaniment estimate, while limiting the artifacts introduced when the voice has been correctly eliminated.

However, some problems are observed when the score does not exactly match the performance.In these cases, the score-based constraints can prevent the system from estimating the right components. Although these problems do not generally represent much in terms of signal energy, they can have some perceptual importance. Thus, future work for the improvement of the separation could involve musically-motivated modifications of the constraints, for example allowing more frequency deviation in the beginning and at the end of the notes, in order to account for *glissandi*. One could also investigate a 'soft constraint' approach, which would penalize source activations which are far from the score indication, without completely forbidding them. The influence of the separation parameters (number of components for the accompaniment, size of the filter dictionary) could also be more thoroughly investigated. In particular, the search for a relation between the optimal parameters and some features extracted from the musical score could be interesting. Finally, another perspective can be the exploitation of the score information for the extraction of the unvoiced components of the leading voice, which were not taken into account in this work.

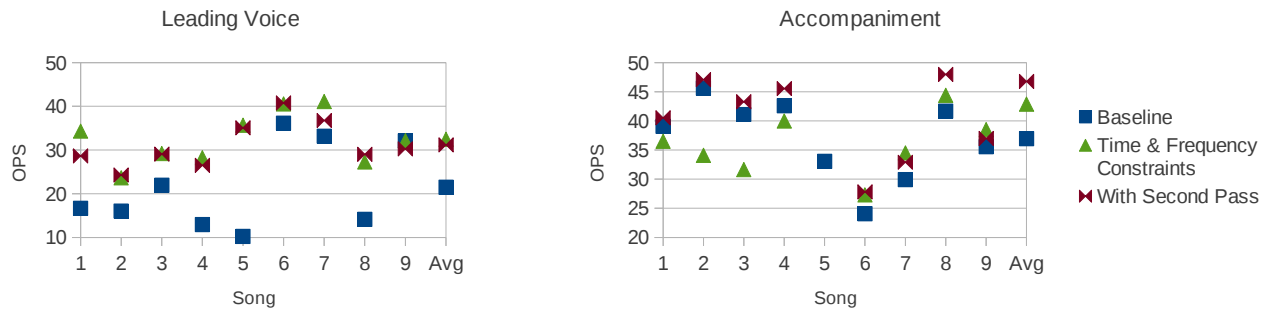


**Figure 3**. Influence of the second pass of leading voice separation on the OPS criterion. The same remark as in Figure 2 holds for song no. 5.

## 6. REFERENCES


[1] Zhiyao Duan and Bryan Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE J. Select. Topics Signal Processing*, 5(6):1205–1215, October 2011.

[2] Jean-Louis Durrieu, Bertrand David, and Gaël Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE J. Select. Topics Signal Processing*, 5(6):1180 –1191, October 2011.

[3] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Trans. Audio, Speech, Language Processing*, 19(7):2046 – 2057, September 2011.

[4] Sebastian Ewert and Meinard Müller. Using score-informed constraints for NMF-based source separation. In *Proc. IEEE ICASSP*, Kyoto, Japan, 2012.

[5] Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proc. IEEE ICASSP*, pages 1869–1872, Taipei, Taiwan, 2009.

[6] Joachim Ganseman, Paul Scheuners, Gautham J. Mysore, and Jonathan S. Abel. Evaluation of a score-informed source separation system. In *Proc. ISMIR*, pages 219–224, Utrecht, Netherlands, August 2010.

[7] Yushen Han and Christopher Raphael. Desoloing monaural audio using mixture models. In *Proc. ISMIR*, pages 145–148, Vienna, Austria, 2007.

[8] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proc. ISMIR*, pages 327–332, Kobe, Japan, 2009.

[9] Romain Hennequin, Bertrand David, and Roland Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. IEEE ICASSP*, pages 45–48, Prag, Czech Republic, 2011.

[10] Katrsutoshi Itoyama, Masataka Goto, Kazumori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals. In *Proc. IEEE ICASSP*, pages 57–60, Honolulu, Hawaii, USA, 2007.

[11] Cyril Joder, Slim Essid, and Gaël Richard. A conditional random field framework for robust and scalable audio-to-score matching. *IEEE Trans. Audio, Speech, Language Processing*, 19(8):2385–2397, November 2011.

[12] Yipeng Li, John Woodruff, and DeLiang Wang. Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Trans. Audio, Speech, Language Processing*, 17(7):1361–1371, September 2009.

[13] Bernhard Niedermayer and Gerhard Widmer. A multi-pass algorithm for accurate audio-to-score alignment. In *Proc. ISMIR*, pages 417–422, Utrecht, the Netherlands, 2010.

[14] Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, and Rémi Gribonval. Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Trans. Audio, Speech, Language Processing*, 15(5):1564–1578, July 2007.

[15] Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Trans. Audio, Speech, Language Processing*, 20(4):1118–1133, May 2012.

[16] Emmanuel Vincent. Musical source separation using time-frequency source priors. *IEEE Trans. Audio, Speech, Language Processing*, 14(1):91–98, January 2006.