Audio Engineering Society

Convention Paper

Presented at the 137th Convention
2014 October 9–12 Los Angeles, USA

OBRAMUS: A System for Object-Based Retouch of Amateur Music*

Jordi Janer[1], Stanislaw Gorlow[1], and Keita Arimoto[2]

[1]*Universitat Pompeu Fabra, Music Technology Group, 08018 Barcelona, Catalunya, Spain*

[2]*Yamaha Corporation, Music and Sound Processing Group, Iwata, Shizuoka 438-0192, Japan*

Correspondence should be addressed to Jordi Janer (jordi.janer@upf.edu)

ABSTRACT

In the more recent past, the area of semantic audio has become object of special attention due to the increase in attractiveness of signal representations which allow manipulations of audio on a symbolic level. The semantics usually refer to audio objects, such as instruments, or musical entities, such as chords or notes. On this view, we present a system for making minor corrections to amateur piano recordings based on a nonnegative matrix factorization. Acting as middleman between the signal and the user, the system enables a simple form of musical recomposition by altering pitch, timbre, onset and offset of distinct notes. The workflow is iterative, that is the result improves stepwise through user intervention.

1. INTRODUCTION

Semantic audio is an emerging area in audio. Its scope in the standard case is the extraction of symbols from audio that have a meaning, such as beat, genre, etc. One could think of the semantics as being the result of the interplay of some low-level entities with certain properties. A song with misplayed notes, e.g., may have the connotation to be amateurish and of low musical quality. In the present work, we seek to elaborate this idea further, in the sense that we provide a description of a system, which enables the musician to correct the misplayed notes in his song, in order to change the semantics. For this to be possible, we need a suitable signal representation.

Nonnegative matrix factorizations (NMFs) have gained a lot of popularity over the last 15 years since Lee and Seung demonstrated their applicability to real-world problems [1, 2]. So far, typical applications in speech and audio signal processing are speech enhancement, automatic speech recognition, speaker diarization, audio coding, bandwidth extension, dereverberation, automatic music transcription, and finally source separation, to name the most prominent ones. In regard to music retouch, which

*This work was partially funded by the Yamaha Corporation.

is the subject of our paper, the following works can be pointed out: Durrieu *et al.* use an NMF together with a source-filter model to separate the vocal from the instrumental [3, 4]. Rafii *et al.*, on the other hand, look for repetitive patterns to isolate the vocal by means of median filtering [5]. Beyond, since an NMF is by no means "musical", several authors have proposed different approaches in order to obtain musically more meaningful or robust results. Smaragdis *et al.*, e.g., use dictionaries that are trained a priori on clean samples [6]. Ewert *et al.* carry it to extremes and use aligned scores to improve accuracy, and in this way the musical reliability of the factorization [7, 8]. An attempt to apply audio effects to the content of a mix by manipulation of the NMF can be found in [9].

Typical time-frequency (TF) representations for an NMF are the short-time Fourier transform (STFT) [10], for the reason of its analytic properties, or the constant-Q transform (CQT) [11, 12] due to its particular aptitude for stringed instruments with widely harmonic spectra, such as the piano. Ideas for how to model a harmonic structure more explicitly can be found in the works by Vincent *et al.* [13] and Fuentes *et al.* [14]. Another way of improving the factorization is by using, e.g., the β-divergence [15–17] in the cost function and by tuning the β-value to the problem at hand [13]. A value of zero, e.g., corresponds to the Itakura–Saito divergence, which according to Févotte *et al.* yields better results for music in particular as compared to the Kullback–Leibler divergence or, all the same, the Euclidean distance [18].

Also, there exist some commercial solutions that can be mentioned in this context. Celemony's Melodyne[1], for instance, is an award-winning music editor that offers a very satisfactory analysis of polyphonic recordings. It relies on classical signal processing and makes use of the theory of harmony to converge to a meaningful result. Another product worth mentioning here is Zynaptiq's PITCHMAP[2]. It requires no NMF to alter, e.g., the key of a music piece and it works in real time. Another commercial solution, yet for voice removal, that should not go unmentioned is Audionamix' ADX TRAX[3].

Our system is oriented towards amateur musicians who would like to make some minor corrections to their home recordings, and all that in a simple, computer-assisted manner. To this end, we pursue an NMF-based approach in combination with supervised learning using an instrument-specific dictionary. The audio material under consideration is composed of piano-plus-vocal recordings captured with a single microphone. The user performs note-level manipulations in pitch, time, and timbre. The vocal is removed in a preprocessing step and not further modified. After the corrections have been made to the piano part, the separated vocal is mixed back again. The user interacts with the system through a graphical interface. He provides additional input that improves the decomposition and specifies the corrections to be made. The partial outcome is assessed after each applied modification via integrated playback functionality.

The organization is as follows. Section 2 shows the basic framework of our system. Section 3 gives details on our extensions to the basic framework and its fine structure. Section 4 illustrates how the system operates in a retouch scenario, in which the user makes some modifications to a single-channel piano-and-vocal recording. In section 5, we discuss our approach from various angles. Section 6 concludes the paper and highlights the main points.

2. BASIC FRAMEWORK

In this section, we present the basic system consisting of the following three processing steps: the vocal removal, the decomposition, and the resynthesis. All steps require some user input, be it only through the assessment of the intermediate results.

2.1. Vocal Removal

In the existing literature, the problem of vocal removal has exhaustively been studied. The state of the art, such as [3], achieves good separation in a scenario where the vocal is dominating over the musical accompaniment. In our system, the vocal removal is a pre-process, which is decoupled from the piano analysis and transformation.

We resort to an existing system [19–21]. It estimates the vocal pitch contour through timbre classification [19]. In addition, it captures inharmonic voice components, such as breathiness [20] and fricative sounds [21]. The soloed vocal is side-chained and mixed back into the piano part after resynthesis before the output.

2.2. Decomposition

The decomposition of the instrumental part into objects after vocal removal is based on an NMF approach in the

[1] http://www.celemony.com/
[2] http://www.zynaptiq.com/pitchmap/
[3] http://www.audionamix.com/

TF domain. More precisely, the observed spectrogram is factorized as

$$\mathbf{X} \approx \mathbf{S}\mathbf{A}^{\mathsf{T}} : K \times N \in \mathbb{R}_0^+, \tag{1}$$

where

$$\mathbf{A} = \begin{bmatrix} A_1(n) & A_2(n) & \cdots & A_R(n) \end{bmatrix}_n \tag{2}$$

represents the activation gains or velocity of the spectral note bases

$$\mathbf{S} = \begin{bmatrix} S_1(k) & S_2(k) & \cdots & S_R(k) \end{bmatrix}_k, \tag{3}$$

which is also referred to as "the dictionary". The size of the dictionary, i.e. the number of distinct note objects, is given by the number of columns $R \ll \min(K,N)$, which is equal to the rank of the factorization.

In real-world recordings, many simultaneous notes with high spectrotemporal overlap appear. For this reason, but more particularly because (1) is not unique, the resulting factorization may not capture the underlying sequence of played notes, making it hard to interpret the result from a musical point of view. A better result is achieved when the dictionary is learned from separate notes of the same or similar instrument under the same or similar acoustic conditions. For piano recordings, a dictionary that covers the full piano range typically has $R = 88$ notes, e.g. such an (overcomplete) dictionary would be held fixed during the factorization and the activations would be adapted to best fit the mixture spectrum w.r.t. to the chosen distance metric $D_F^{\mathbf{X}}(\mathbf{S}\mathbf{A}^{\mathsf{T}}, \mathbf{X})$:

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} D_F^{\mathbf{X}}(\mathbf{S}\mathbf{A}^{\mathsf{T}}, \mathbf{X}), \tag{4}$$

where $D_F^{\mathbf{X}}(\cdot,\cdot)$ in this case refers to a Bregman distance that is generated from a convex function F. In literature, (4) is also termed *supervised learning*. If the NMF uses multiplicative updates, the result further improves if the velocity matrix $\mathbf{A}$ is initialized with zeros for notes that are not likely to appear. In this simple yet effective way, the total energy will be distributed among the remaining notes only.

2.3. Resynthesis

Once the mixture spectrogram is factorized, we would go on and try to manipulate and resynthesize the recording. Individual piano notes, alias note objects, can be isolated by means of spectral filtering or "soft masking":

$$\mathbf{X}_i = \left(\mathbf{s}_i \mathbf{a}_i^{\mathsf{T}} \oslash \mathbf{S}\mathbf{A}^{\mathsf{T}}\right) \odot \mathbf{X}. \tag{5}$$

In (5), elementwise multiplication is indicated by $\odot$ and elementwise division by $\oslash$, respectively. The vectors $\mathbf{s}_i$ and $\mathbf{a}_i$ denote the spectral basis and the activation gains of the ith note. The phase relation is $\arg \mathbf{X}_i = \arg \mathbf{X}$. One can alter the pitch and the duration of the separated note by use of pitch-shifting and/or time-stretching methods directly in the frequency domain [22]. Transformed note objects are then mixed back into the residual signal.

The main drawback of this approach is that the quality of the transformed note strongly depends on the accuracy of the separation. If, e.g., the note attack is captured barely, the transformed note will sound less natural. Moreover, phase-vocoder techniques offer good quality in the range of less than one octave. The amount of pitch shifting that can be applied without introducing too many artifacts is hence very limited. In Section 3, we propose techniques to overcome these limitations.

2.4. User Input

Any of the above steps, if carried out fully automatically, will be imprecise, which might lead to an end result that is far from satisfactory. And thus, for the reason that no machine learning algorithm, so far, can reliably replicate neither human behavior nor perception, the user must be made part of the framework. His direct implication is to the effect that the result at each step is satisfactory. This he can assess through critical listening. In case the result is not satisfactory, he should be enabled to make further corrections and listen to the result again. Then he would decide between going over to the next step or going back to the previous state. One example is the fine-adjustment of the pitch contour in vocal removal. That is the reason why we consider the work flow "iterative". It should be noted that the hearing apparatus of the individual, which cannot be replaced by any objective metric, is involved in the whole process.

3. EXTENSIONS

One major problem that we address is that the accuracy of the factorization compromises the extent of possible note transformations with respect to sound quality. The accuracy of the factorization, on the other hand, depends largely on the musical complexity of the recording in the form of overlapping partials. For that reason, we resort to an external note generator in the form of a sample bank, when necessary. The acoustic color of notes stemming from the sample bank is "morphed" into the color of the recorded piano to better fit the mix. As a result,

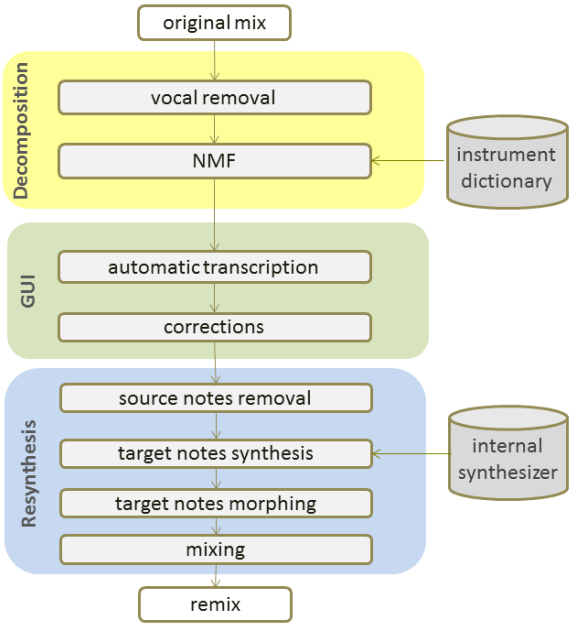


Fig. 1: The system overview (decomposition with vocal removal, user intervention via GUI, and resynthesis).

with our system, single notes can be muted, transformed, or inserted from a sample bank and timbre-adapted to the original mix. The accuracy of the decomposition could be improved, if the corresponding score was available [7, 8]. Therefore, we provide the user with a coarse initial transcription that he can refine further. The complete system is shown in Fig. 1. The extensions to the basic framework are discussed below.

3.1. Morphing

As mentioned previously, the sound quality of a filtered (through soft-masking)and pitch-shifted, or transposed, note object can suffer from severe artifacts, which are due to modeling errors or the behavior of the soft mask at a low signal-to-interference ratio (SIR), or both. This is the case when a low-energy note interferes with high-energy partials of a more dominant note, such as a bass note. To overcome this limitation, we resort to an internal sample bank or a tone generator. As the sample bank may not necessarily match with the instrument in the mixture, we then apply a timbre filter to a note that stems from outside the mix, i.e. from the sample bank, e.g., to make the remix sound more homogeneous. The computation of the timbre filter is as follows. With the equivalent-rectangular-bandwidth (ERB) scale given by [23]

$$\mathrm{ERBS}(f) = 21.4 \cdot \log_{10}(1 + 0.00437 \cdot f), \qquad (6)$$

where f is in Hz, the transfer function[4] writes

$$H_i(f) = \frac{\int_{f'} X_i^{\mathrm{mix}}(f')}{\int_{f'} X_i^{\mathrm{out}}(f')} \quad \forall f' \qquad (7)$$
$$\text{s.t. } \lfloor \mathrm{ERBS}(f') \rfloor = \lfloor \mathrm{ERBS}(f) \rfloor.$$

Note that both notes have the same pitch, as indicated by the subscript i. This is to make sure that the spectra have the same support (harmonics) in order to avoid divisions by zero, or by very small numbers in general. In (7), the spectrum $X_i^{\mathrm{mix}}(f)$ hence belongs to the note that is to be corrected and so removed, while $X_i^{\mathrm{out}}(f)$ is the spectrum of the corresponding note in the sample bank. The actual filtering operation or "morphing" simply consists in the multiplication of the spectrum of the note that is inserted from the sample bank into the mix:

$$\tilde{X}_{i+j}^{\mathrm{mix}}(f) = H_i(f) \cdot \bar{X}_{i+j}^{\mathrm{out}}(f), \qquad (8)$$

where

$$\bar{X}_{i+j}^{\mathrm{out}}(f) = \frac{\int_f X_i^{\mathrm{out}}(f)}{\int_f X_{i+j}^{\mathrm{out}}(f)} \cdot X_{i+j}^{\mathrm{out}}(f). \qquad (9)$$

The weighting in (9) makes sure that the inserted note is approximately as loud as the removed note. Note that $\lfloor \cdot \rfloor$ in (7) denotes the floor function.

3.2. Automatic Transcription

As seen in the results obtained in score-informed separation, it is to expect that a preliminary transcription step can improve the quality of the note decomposition and separation of piano recordings. Here we propose a new method for piano note transcription by combining *i)* a NMF-based note decomposition with *ii)* a transient estimation per note using the phase information.

The resulting transcription is computed by detecting onset candidates on the note energy curve and selecting only those with a transient estimation above a threshold.

From (1) we obtain in $\mathbf{A}$ an activations curve over time for all 88 notes. Next we perform source separation for all 88 notes using soft masking, see (5), obtaining one separated complex spectrogram per note i.

[4] Only the amplitude spectrum is considered.

Then for each note i, we compute a transient curve from the separated magnitude spectrograms. The transient estimation function $\Gamma(\mathbf{X}_i)$ provides a scalar value per frame computed from the complex spectrogram $\mathbf{X}_i$. For piano recordings, maxima on the transient estimation function shall indicate note onsets. $\Gamma(\mathbf{X}_i)$ is a measure of the temporal center of gravity of energy of spectral peaks within the windowed signal for one frame. See [24] for details. To increase robustness, the transient curve is weighted by the note energy curve $e_i(n)$, avoiding spurious transients from other notes. We obtain a transient curve $t_i(n)$ for all 88 notes as depicted in figure 2.

$$t_i(n) = \frac{\Gamma(\mathbf{X}_i)e_i(n)}{\frac{1}{N}\sum_{n=1}^{N} e_i(n)} \tag{10}$$

Next step is detecting a set of peaks from the energy curve $e_i(n)$ as onset candidates K_i. Only peaks above a threshold β_e are considered. Then we select only those candidates k_i^t with a transient curve above a threshold $t_i(k_i) > \beta_t$. The onset detection output is a list of frame indexes for each note K_i^t for a note i. Threshold values β_e and β_t are empirically determined.

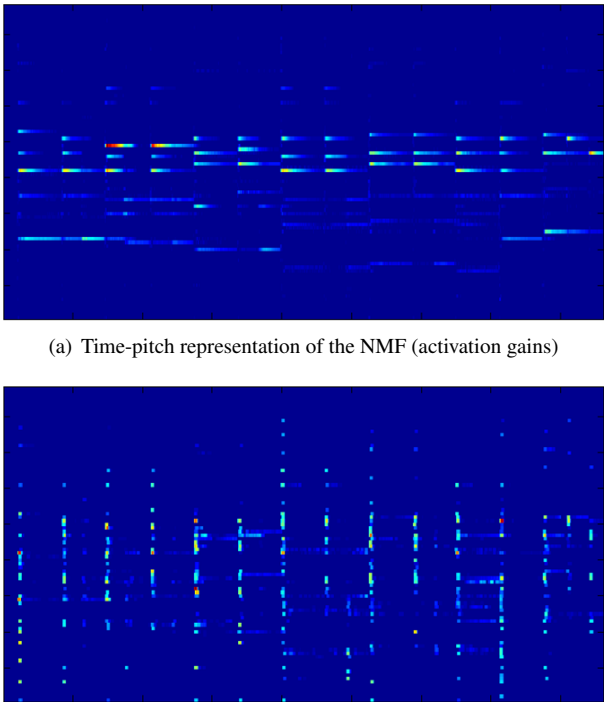
Once we detected the onsets for each note, we obtain the final note transcription by computing also note duration and dynamics. The note duration is set by looking at the frame index in which the energy curve $e_i(n)$ is below the energy threshold β_e, with a maximum duration of 2.5 seconds. A value for note dynamics is also obtained from the maximum of the note energy curve $e_i(n)$ within the note duration.

4. USE CASE: RETOUCH

Now, we illustrate the interplay of all the components of our system in a retouch scenario, in which an (amateur) musician corrects the misplayed notes before uploading a recording with his performance to a video sharing site, such as YouTube or Nico Nico Douga.

4.1. Graphical User Interface

The user interacts with the system by use of a graphical user interface (GUI). It is built around a piano-roll-like representation of the note activation gains obtained after the decomposition. The result of the NMF is captured in the static background image, while single note objects or regions of notes are drawn as transparent rectangles atop



(a) Time-pitch representation of the NMF (activation gains)

(b) Intermediate result of note onset detection

Fig. 2: Note activation gains and the resulting onsets for automatic transcription.

of the background. This allows the user to mark distinct note objects and to specify the desired transformation in a very intuitive manner. The GUI also provides means to improve the decomposition as explained further below.

4.2. Workflow

Fig. 3 shows the interplay of sequential processing and user interaction including the signal flow, which is from top to bottom. Only the corrections on the note transcription and the desired note transformations require user intervention. The rest is automated.

4.3. Analysis

In one of the initial steps, the user invokes the NMF that yields a piano-roll view of the recorded signal. Since the NMF assumes that the input signal is piano only, a vocal removal algorithm is run first. Any algorithm that is apt at this task can be used, see Section 2.1. The signal after vocal removal would ideally contain only the piano part, see Fig. 4.

However, keep in mind that it is not reasonable to expect

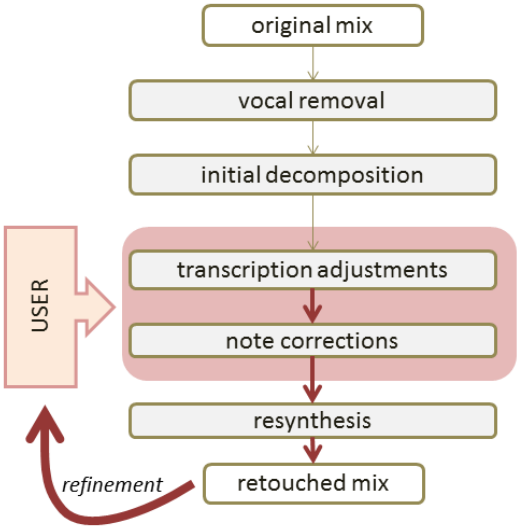


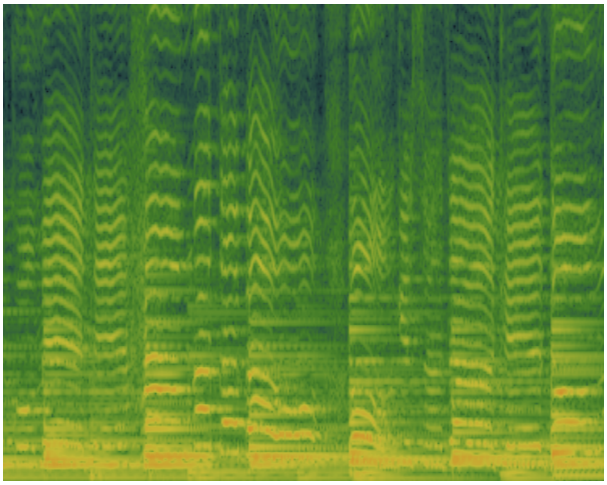
Fig. 3: Iterative workflow with user intervention.

that the initial decomposition is comparable with a MIDI score. This is partly because it is run without any a priori information on the notes played, see Fig. 5(a). The notes played by the left and the right hand are bordered by two translucent rectangles. The notes activated outside of the rectangles spurious and are mostly due to octave errors.

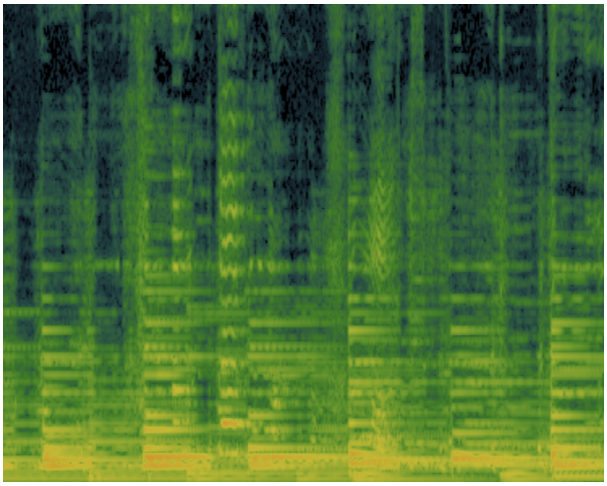
To reduce the number of spurious notes, the user would rerun the NMF specifying the active note regions or in a more extreme case specifying each active note. In order to facilitate this process and to reduce the amount of time that the user would spend on fine-adjusting each distinct note, the system provides an automatic transcription after every new run of the NMF. Thus, the user can choose to either run a region-informed or a transcription-informed NMF subsequent to the previous decomposition. In that case, the activation gains outside of the marked areas are set to zero during initialization, see Section 2.2 for further details. For comparison see Fig. 5(b), which shows the result of the region-informed NMF.

4.4. Transformation and Resynthesis

Once the decomposition is deemed reliable enough, one would resort to the provided note transcription and apply the desired corrections by use of the GUI. At the present stage, the system allows the user to move around, delete, insert, and change the duration of distinct notes. Wrong or misplayed notes are labeled as "source" objects while



(a) Spectrogram of the piano-and-vocal recording

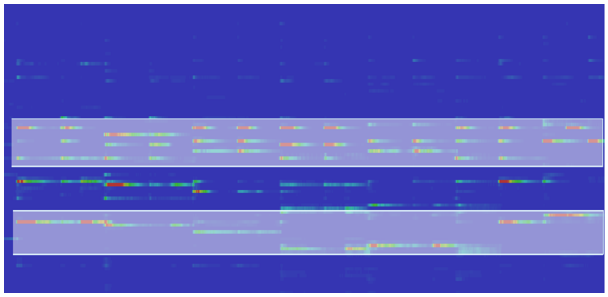


(b) Spectrogram after vocal removal

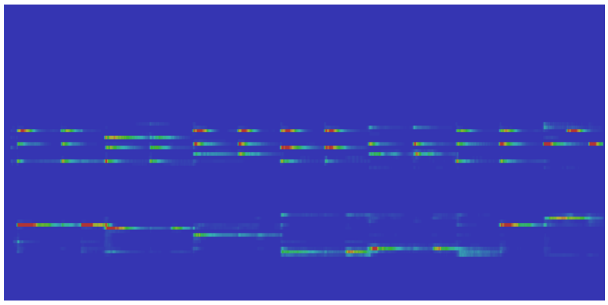
Fig. 4: The spectrogram before and after vocal removal.

the correct notes are labeled as "target" objects. After all the source objects and their corresponding target objects have been specified by the user, the system performs the the following processing steps:

1. Removal of source objects from the mix
2. Insertion of target objects from the sample bank
3. Morphing of target objects to match the mix
4. Combining 1 and 3 to form the remix
5. Mixing the vocal back into 4



(a) Initial decomposition and active note regions



(b) Region-informed decomposition

Fig. 5: An example of a user-refined decomposition by indication of active note regions.

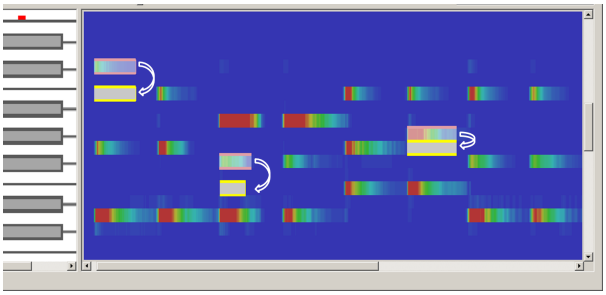


Fig. 6: Source objects (marked red) are transformed into target objects (marked yellow).

4.5. Playback

The GUI is further equipped with playback functionality that lets the user select and listen not only to the original and the final mix, but also to the removed vocal, the part of the piano, the source objects, the background, and the target object before and after morphing. In this manner, one can easily detect the source of error in case the final mix is not satisfactory and can focus on the problem.

5. DISCUSSION

One may ask the question why we address amateurs and not professionals in the first place. The answers are very simple. Firstly, amateur recordings are less sophisticated from a musical point of and exhibit more errors than the recordings of professional musicians. And so, there is a greater need for simple tools that would allow for minor corrections. Secondly, amateurs are more likely to accept certain imperfections in the retouched recording. In that respect, the presented system is “good enough”. Still, it should be noted that the present context is not as simple as it may seem: the piano is polyphonic, the score is not available, there is the interference with a loud vocal, and we should neither forget about the effect of a reverberant environment.

Given the high level of difficulty, we would like to point out the importance of user intervention. The more input the user provides to the system at various levels, such as the transcription, the more accurate in terms of quality is the final result. As each modification that is made by the user can change the result for the better or for the worse, it is necessary to let the user listen to each partial result, so he can decide whether to carry on with the next step.

If during the decomposition the spectral particularity of an instrument is taken into account, more reliable results can be expected. So, if further optimized, the NMF-like approach would allow for a finer decomposition w.r.t. a tool such as Melodyne or similar. However, this does not go without the incorporation of music theory such as the theory of harmony in Western music, which would help the NMF make musically more meaningful decisions. It could also assist the user in the refinement of objects. A system with the described optimizations could also serve as a post-production tool for film and music. In the latter case at least, the processing techniques would need to be extended to stereo.

It is wrong to think that the system is applicable to piano recordings only. Dictionaries for different instruments or different playing styles, room acoustics, etc. are storable and could be selected by the user according to the given mix. These can also be adapted to the mix through NMF updates. Apart from retouch, other use cases include:

Instrument replacement
The bass line (left hand), e.g., is replaced by a bass guitar, a contrabass, or any other instrument.

Mono-to-stereo upmix
The separated bass line (left hand) and the melody line (right hand) of a mono recording are panned in stereo, so as to stretch the sound in space.

Automatic accompaniment
Using the transcription function, a group of objects representing, e.g., the melody line is synthesized as, e.g., a string ensemble and added to the remix.

In our framework, vocals removal is a separate process and it is assumed that the remaining instrumental is more or less clean. However, this is not necessarily true, since the result depends on the accuracy of the pitch detection and on the accuracy with which unvoiced phonemes and reverberation tails are suppressed. These artifacts are yet negligible, because the isolated vocal is mixed back into the rectified remix. For some audio samples refer to the group's website.[5]

6. CONCLUSION

We presented an audio system that enables manipulation of individual notes in a piano-plus-vocal recording. The system was originally designed for amateurs, but it may be used in a professional context as well. The framework is NMF-based, and so it is different from Melodyne and similar tools. We highlighted the main limitations of the approach, which are subject to the accuracy of the NMF variant used. Without any doubt, a decomposition that is reliable in a musical sense will allow for a better quality at the output. This includes a better suppression of note attacks and so-called "phantom" notes that are the result of octave errors. One simple way to improve accuracy is to leave the user with the option to intervene and correct the automatic transcription if necessary. He is, thus, part of the system—a sort of supervisory authority. Another thinkable direction for future work is the incorporation of music theory in the decomposition stage to avoid note phantoms. Nevertheless, one should pay attention to the fact that misplayed notes could be mistaken for phantom notes in that case.

The integration of an internal sample bank together with timbre filtering extended the potential of the system in a sustained manner. The quality is sufficient for a note not to be perceived as "extraneous" to the background. Thus, our system could be used to create remixes and mashups of existing music, as well. Whereas notes taken out of a sample bank yield perfect quality, the morphing will not turn, e.g., an electric guitar into a piano. In a more open scenario, the user must hence be given access to a broad range of different sample banks to choose from. Future work can also address the problem of recommending the instrument in the sample bank which comes close to the mix. Joint estimation of the vocal and various objects of different character such as notes, chords, or drum hits in a single NMF is another problem to tackle [25].

[5] `http://www.mtg.upf.edu/`

7. ACKNOWLEDGMENT

The authors thank Martí Umbert for helping them find a sonorous acronym for the system.

8. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[2] ——, "Algorithms for non-negative matrix factorization," in *Proc. NIPS 2000*, Nov. 2000, pp. 556–562.

[3] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proc. EUSIPCO 2009*, Aug. 2009, pp. 15–19.

[4] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 564–575, Mar. 2010.

[5] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (REPET): A simple method for music/voice separation," *IEEE Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 73–84, Jan. 2013.

[6] P. Smaragdis, "Convolutive speech bases and their applications to supervised speech separation," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[7] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proc. ICASSP 2012*, Mar. 2012, pp. 25–30.

[8] J. Driedger, H. Grohganz, T. Prätzlich, S. Ewert, and M. Müller, "Score-informed audio decomposition and applications," in *Proc. MM 2013*, Oct. 2013, pp. 541–544.

[9] R. Sarver and A. Klapuri, "Application of nonnegative matrix factorization to signal-adaptive audio effects," in *Proc. DAFx-11*, Sept. 2011, pp. 249–252.

[10] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[11] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[12] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Am.*, vol. 92, no. 5, pp. 2698–2701, Nov. 1992.

[13] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[14] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1854–1866, Sept. 2013.

[15] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1859–1886, Aug. 2002.

[16] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 3, pp. 780–791, Mar. 2007.

[17] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, Sept. 2011.

[18] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[19] R. Marxer, J. Janer, and J. Bonada, "Low-latency instrument separation in polyphonic audio using timbre models," in *Proc. LVA/ICA 2012*, Mar. 2012, pp. 314–321.

[20] R. Marxer and J. Janer, "Modelling and separation of singing voice breathiness in polyphonic mixtures," in *Proc. DAFx-13*, Sept. 2013, pp. 1–4.

[21] J. Janer and R. Marxer, "Separation of unvoiced fricatives in singing voice mixtures with semi-supervised NMF," in *Proc. DAFx-13*, Sept. 2013, pp. 1–4.

[22] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Audio, Speech, Language Process.*, vol. 7, no. 3, pp. 323–332, May 1999.

[23] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *Acta Acust. United Ac.*, vol. 82, no. 2, pp. 335–345, Mar./Apr. 1996.

[24] J. Janer, R. Marxer, and K. Arimoto, "Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation," in *Proc. ICASSP 2012*, Mar. 2012, pp. 281–284.

[25] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1192–1204, Oct. 2011.