EFFECTIVE SEPARATION OF LOW-PITCH NOTES USING NMF WITH NON-POWER-OF-2 SHORT-TIME FOURIER TRANSFORMS

Ta-Chun Chen, Tien-Ming Wang, Ya-Han Kuo and Alvin Su

The Department of Computer Science and Information Engineering,
National Cheng-Kung University
No. 1, Ta-Hsueh Road, Tainan, Taiwan, ROC
titmis@gmail.com, showmin@csie.ncku.edu.tw,
yu08723@gmail.com, alvinsu@mail.ncku.edu.tw

ABSTRACT

Recently, non-negative matrix factorization (NMF), which is applied to decompose signals in frequency domain by means of short-time Fourier transform (STFT), is widely used in audio source separation. Separation of low-pitch notes in recordings is of significant interest. According to time-frequency uncertainty principle, it may suffer from the tradeoff between time and frequency localizations for low-pitch sounds. Furthermore, because the window function applied to the signal causes frequency spreading, separation of low-pitch notes becomes more difficult. Instead of using power-of-2 FFT, we experiment on STFT sizes corresponding to the pitches of the notes in the signals. Computer simulations using synthetic signals show that the Source to Interferences Ratio (SIR) is significantly improved without sacrificing Sources to Artifacts Ratio (SAR) and Source to Distortion Ratio (SDR). In average, at least 2 to 6 dB improvement in SIR is achieved when compared to power-of-2 FFT of similar sizes.

1. INTRODUCTION

1.1. Motivation

Non-negative matrix factorization (NMF) is applied to factorize the spectrogram into basis spectra and temporal activation in music signal analysis [1]. Recently, it is widely used for audio blind source separation [2], music transcription [3, 4], pitch detection, onset detection and analysis/synthesis of bowed-string instrument recordings [5]. In [5], we found that it is difficult to perfectly separate low-pitch notes due to the properties of Fourier transform that is normally used before NMF is performed. Without employing suitable STFT sizes, the spectral leakage of the window functions applied prior to STFT is the main reason that results in ambiguous spectrogram of each low-pitch note. In this paper, an aspect is proposed that the spectrogram of each musical note should be kept as intact as possible in advance. It is found that separation of low-pitch notes may be more effective if the STFT sizes corresponding to the pitches can be applied.

1.2. Related Works

In 1964, Cooley and Tukey reported the fast Fourier Transform (FFT) algorithm [6]. The decimation-in-time power-of-2 FFT then became the most popular tool to achieve the discrete Fourier transform (DFT) for frequency-domain signal processing. Short-time Fourier transform (STFT) [7] is one popular general-purpose transform for the analysis of audio signals in the time-frequency domain, though, with some faults. First, frequency domain resolution may not be high enough to represent low-pitch notes if the STFT size is relatively small. Secondly, the window functions [8] applied prior to STFT create the inevitable main and side lobes which cause the energy spreading to neighbor frequency bins and interference between notes close in their pitches usually occur.

1.3. This Work

Unlike conventional analysis using power-of-2 FFT, this work aims at providing a procedure to effectively decompose low-pitch notes using non-radix FFT and non-negative matrix factorization. NMF introduced in [1] is used for the separation step. In order to establish test sets, synthetic musical signals generated with a MIDI synthesizer are used. We experiment on various window sizes to gain the best separation results on two low-pitch notes. Without reducing SAR and SDR, the proposed method provides 2 to 6 dB improvement in SIR. Brief review of previous works is made in section 2. Section 3 describes the proposed work. In section 4, experiments and results are presented. Conclusion is given in section 5.

2. BACKGROUND

2.1. NMF-BASED MUSIC SIGNAL ANALYSIS

NMF is applied to decompose a matrix into two matrices. In [9], the non-negativity is shown to be a useful constraint for matrix factorization. Given an $k \times n$ matrix $V \in R^{k \times n}$ with non-negative entries, NMF tries to factorize V into an $k \times r$ non-negative matrix $W \in R^{k \times r}$ and an $r \times n$ non-negative matrix $H \in R^{r \times n}$ such that $V \approx \tilde{V} = WH$ where r is a positive integer and $r \le \min(k, n)$. The factorization is achieved by minimizing a specific cost measuring the distance between the above two matrices V and $\tilde{V}$. In [1], multiplicative update rules are introduced to iteratively obtain randomly initialized W and H. The update rules are:

- Kullback-Leibler divergence:

$$H \leftarrow H \otimes \frac{W^T \frac{V}{WH}}{W^T \cdot \mathbf{1}},\ W \leftarrow W \otimes \frac{\frac{V}{WH} H^T}{\mathbf{1} \cdot H^T} \tag{2-1}$$

where division is carried out element-wise, $\otimes$ denotes element-wise multiplication, and $\mathbf{1}$ represents an $M \times N$ matrix of ones, used to compute row and column sums [2].

2.2. SPECTRAL LEAKAGE

When analysing a musical signal, one usually divides the signal into short segments and STFT is applied. The STFT size, however, may not match the periods of the sub-signal. Hence, the possible discontinuities at the segment boundaries cause the spectral leakages [8].
Two methods can be used to reduce the spectral leakage effects. The first one is to find an appropriate segment size that exactly matches the period of the signal. The other is to apply a proper window function to the segment to eliminate the boundary discontinuities.
Applying a window function has its side effects, too. The window function alters the signal behaviours in frequency domain. This is critical to us because NMF usually operates in frequency domain. The spreading effect of the window function in frequency domain causes troubles in separating tones with close pitches [10]. Equivalent Noise Bandwidth (ENBW) [8] is used to evaluate the total leakage of a window function.

$$ENBW = \frac{\sum_n w^2(nT)}{\left[\sum_n w(nT)\right]^2}. \qquad (2\text{-}2)$$

Larger ENBW represent more spectral leakage. For example, rectangle window has the smallest ENBW of 1, while ENBWs of hamming window and hanning window are 1.36 and 1.5, respectively. The -3 dB main lobe bandwidth is another important indicator [8]. The -3dB bandwidth of rectangle window is 0.89 bin. The -3 dB bandwidths of hamming window and hanning window are 1.3 and 1.44 bins, respectively. Figure 1 shows Frequency response of 3 window functions.
It implies that rectangular window has the best spreading characteristics. Therefore, we use rectangle window in this paper. However, using rectangle window without choosing the appropriate STFT size, the spectral leakages will be large because rectangular window has relative larger side lobes. In our experience, spectral leakage is also a key point to NMF-based separation.

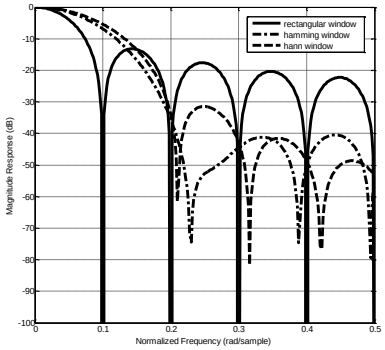


Figure 1: *Frequency responses of 3 window functions.*

3. OPTIMAL STFT SIZE DETERMINATION

In section 2, it concludes that we have to get the exact period of the signal for the STFT analysis with least spectral leakage. Multiple f_0 estimation [11] can be adopted to achieve the task. In this work, musical scores of the signal are regarded as a priori knowledge for the sake of not introducing unexpected artefact. There are, however, some slight inevitable pitch differences between the score information and the associate audio signal. A likely frequency range, normally a semitone, of the respective pitch has to be considered. In this section, a systematic flow is proposed to determine the optimal STFT size.
According to the given possible pitch period, the possible optimal STFT size, N, can be define as

$$\mathrm{N} = \bigcup_{z \in Z} \bigcup_{f=0.97* f_0^z}^{1.03* f_0^z} Round(\frac{F_s}{f}), \qquad (3\text{-}1)$$

where f_0^z is the fundamental frequency of *z*-th pitch of the signal, and F_s is sampling rate. The set *Z* and *N* stand for all pitches and corresponding likely periods in samples, respectively. For each period n ∈*N*, it is then considered as STFT size to transform the signal into the spectrogram, *V*. To evaluate the amount of the interferences among these notes, it is necessary to separate the note first. In this paper, the score is given in advance such that the harmonics can be constrained [5]. NMF, as showed in section 2.1, can be adopted to separate the notes adequately [5]. In music analysis, *V* is used to represent signal spectrogram. For example, the column vector V_j of *V* is the spectrum of the *j*-th time frame. Hence, the column vector W_i of *W* represents the template of the *i*-th note contained in the signal, and the element H_{ij} of *H* indicates the intensity of the *i*-th note which appears in the *j*-th time frame. The initial template *W* can then be defined as

$$W_k = \sum_p Rand(p) \cdot f\left(p \cdot f_0^k - \sigma, p \cdot f_0^k + \sigma; x\right), \qquad (3\text{-}2)$$

where $f(\cdot)$ is a uniform distribution for the interval [$p \cdot \mu_k - \sigma$, $p \cdot \mu_k + \sigma$], f_0^k is the fundamental frequency corresponding to W_k and *p* is partial index. Due to the characteristics of rectangular window described in section 2, the spreading width of its main lobe is less than a bin. Therefore, σ is set as $\min(1,\ 0.03 * f_0^k)$. $Rand(\cdot)$ is a random function for generating the initial partial magnitudes. KL divergence shown in equation 2-1 is used.

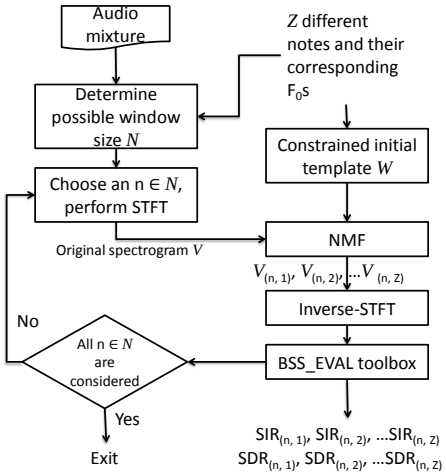


Figure 2: *Proposed system flow of optimal STFT size determination.*

Inverse STFT is then applied to convert the separated spectrograms back to time-domain signals. In order to evaluate the amount of the interferences, a Matlab toolbox called BSS_EVAL [12] is used for the objective measure. The Sig-

nal to Interference Ratio (SIR), the logarithmic ratio of target signal and estimated interference error, is introduced to determine the optimal STFT size which leads to the least interference. The details of the evaluation metrics and results are presented in section 4.The system flow chart of proposed method is depicted in Figure 2.

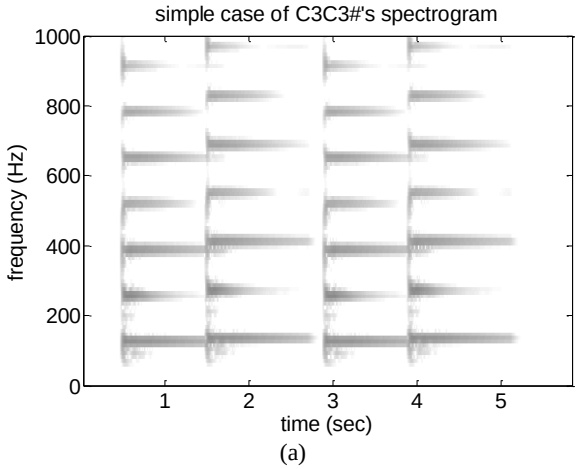
4. EXPERIMENTAL RESULTS

Experiments on synthetic musical signals obtained by means of MIDI synthesizer are introduced. The synthetic signals are first converted using STFT with windows of different types and sizes. NMF with harmonic constraints described in the previous section is adopted for the separation of low-pitch notes. To evaluate the performance, the separated notes are compared to the original tracks. Sampling rate is set as 44100 Hz. The hop size is 5.8 ms (256 samples). Rectangular window is used. Hamming window and Hanning window are also used for comparison. Objective measures reported in [12] for evaluation of source separation methods are used. The Signal to Interference Ratio (SIR) and the Signal to Distortion Ratio (SDR) are computed for each note. All those metrics are expressed in dB. The SIR quantifies the degree of influent energy from other sources and is the most important measurement in our case to determine the performance of separation of low-pitch notes. The SDR is related to the distortion of the estimated signal and is used to quantify the degree of preservation of the target source. They are defined as

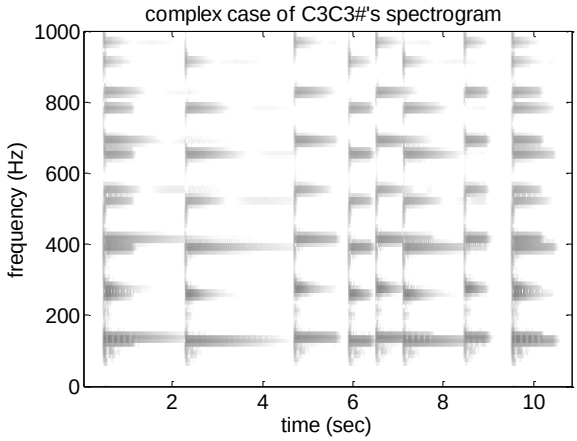
$$SIR = 10\log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf}\right\|^2} \quad (4\text{-}1)$$

$$SDR = 10\log_{10}\frac{\left\|s_{target}\right\|^2}{\left\|e_{interf} + e_{artif}\right\|^2} \quad (4\text{-}2)$$

where the estimated signal $\hat{s}$ can be decomposed into a target signal s_{target}, interference error e_{interf}, and artifact error e_{artif}.
Some signals generated by a MIDI synthesizer are used for the experiments. Two spectrograms of signals containing C3 and C3# notes are depicted in Figure 3. The fundamental frequencies are 130.81 Hz and 138.59 Hz respectively. In the simple case, these notes are slightly overlapped with each other and exposed asynchronously. In the complex case, more overlaps among the notes are created.

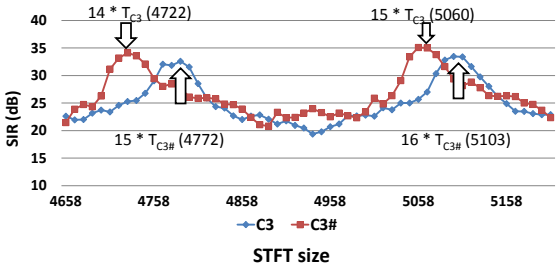


(a)

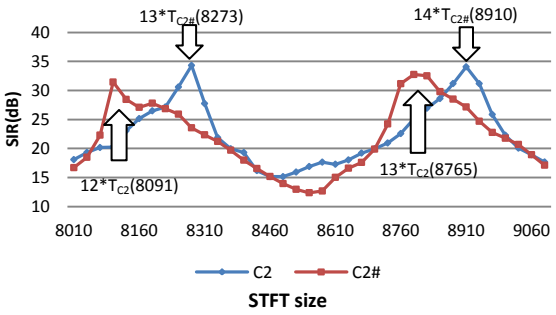


(b)

Figure 3: *The spectrograms of (a) simple and (b) complex cases with notes C3 and C3#.*



(a)



(b)

Figure 4: (a) *SIRs of the simple case for C3, C3# notes with different STFT size in the section [4627, 5207].*(b) *SIRs of the complex case for C2, C2# notes with different STFT sizes in the section [8010, 9060].*

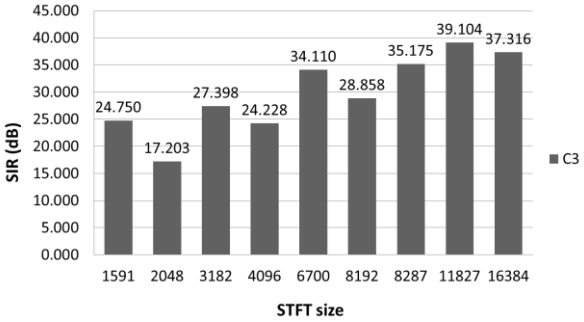
It should be noticed that the fundamental periods of C3 and C3# are $T_{c3} = F_s/f_{c3} = 44100/130.81 \approx 337$ and $T_{c3\#} = F_s/f_{c3\#} = 44100/138.59 \approx 318$, respectively. In the following experiments, the possible appropriate STFT sizes are chosen such that they are over ten times of the periods of the notes. Therefore, the possible optimal STFT sizes for C3 and C3# notes should fall within the interval [4903, 5207] and the interval [4627, 4913], calculated as $0.97\times15\times337 \fallingdotseq 4903$, $1.03\times15\times337 \fallingdotseq 5207$, $0.97\times15\times318 \fallingdotseq 4627$, and $1.03\times15\times318 \fallingdotseq 4913$, respectively. For C2 and C2# notes, the possible optimum STFT sizes should locates within the interval [8010, 9060]. With the proposed separa-

tion procedure described in section 3, the SIRs of each source note with various STFT sizes are shown in Figure 4.

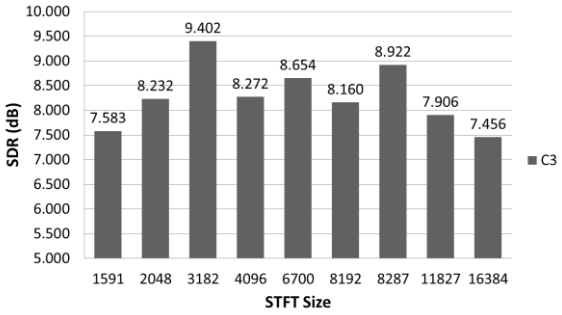
In Figure 4(a), it is interesting to notice that the STFT sizes corresponding the local maximum SIRs for the C3 note are related to the periods of C3# and vice versa. For example, the two SIR peaks for the C3 note are at 4722 and 5060 that are 14 and 15 times of the fundamental period of C3#, respectively. So, the optimal STFT size related to one source, C3 for example, leads to the best SIR of the other source (C3#). Furthermore, after source separation with NMF is used, SIR can vary massively (15dB in this case) by changing the STFT size .It deduces that: a) it is important to choose the analysis STFT size cautiously before performing source separation in frequency domain and b) the STFT size of the maximum SIR rather than the normal power-of-2 STFT size has to be chosen.

Figure 5 to 6 show the SIR and SDR results of separation of C3 and C3# notes in the two audio signals shown in Figure 3 with different STFT sizes. The power-of-2 STFT normally used in the literatures (2048, 4096, 8192, and 16384) are listed as well. In all cases, the SIR performs the best on the STFT size that is integer multiple period of the source. The performance is also excellent when the size is larger than 10,000, but the corresponding SDR significantly drops due to the lack of time-domain localization. On the other hand, the SIR of the non-power-of-2 STFT cases whose size is related to the sources performs much better in comparison with power-of-2 STFTs of similar sizes. For example, in Figure 5 (a), one may see that SIR(1519) > SIR(2048), SIR(3182) > SIR(4096), SIR(6700) > SIR(8192), and SIR(11827) > SIR(16384). Generally speaking, there are 2 to 6 dB improvement in SIR comparing to power-of-2 STFT of the similar size, without sacrificing performances in SDR and SAR. That means it is advantageous on both source separation performance and time-domain localization simultaneously if a proper STFT size is determined. Figure 7 to 8 demonstrates the result of the complex case. It is found that one should still choose STFT size close to the signal period. For the experiments of using C2 and C2# notes, similar results are obtained and not shown in this paper. One can conclude that our finding is consistent.

It is worth to notice that rectangular window is not adoptable while power-of-2 FFT is used. Except for the case that the analysis size is integer multiple of the period of the signal, spreading effect rectangular window will produce the poor performances. Without displaying redundant figures, only the evaluation results of the C3 note in the complex case are depicted in Figure 9.The best SIRs and SDRs obtained according to the proposed system flow in Figure 2 are listed with rectangular window **r**, Hamming window **h** and Hanning window **n**. **r_N**, **h_N** and **n_N** shown under the figures represent the STFT size corresponding to rectangular, Hamming, and Hanning window, respectively. The respective SIRs/SDRs are shown. Notice that both Hamming and Hanning windows give better performance than rectangular window when STFT size is large. Even the score information is given, it is not always possible to obtain the "exact" period of the signal in the unit of sound sample. The spectral spreading caused by the side lobe of rectangular window becomes serious. Furthermore, signal cannot be truly periodic in a long time frame. The large side lobs of rectangular window may also cause troubles.

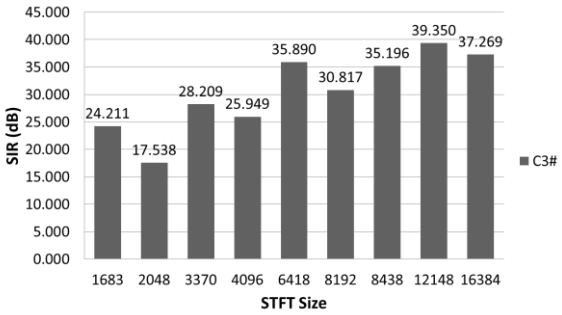


(a)

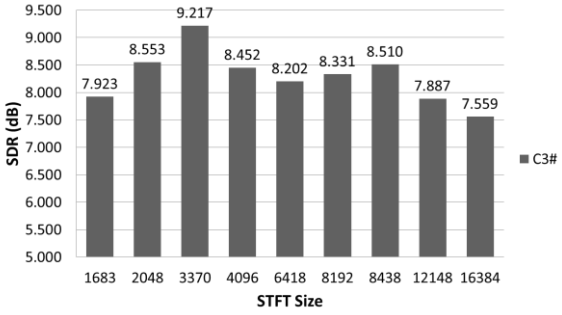


(b)

Figure 5: *The (a) SIR and (b) SDR of* the *note C3 of the simple case with respect to different STFT sizes.*

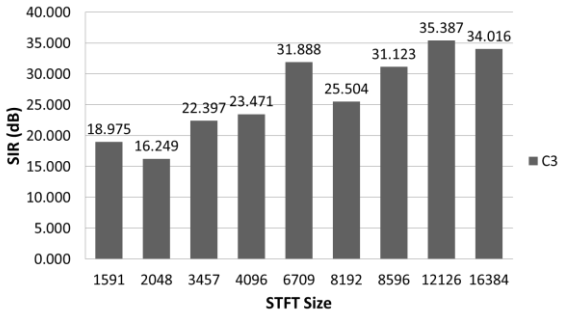


(a)

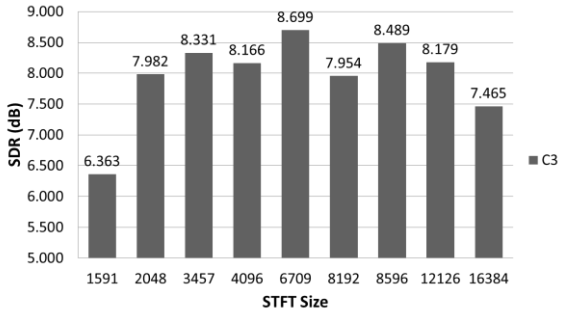


(b)

Figure 6: The*(a) SIR and (b) SDR of the note C3# of the simple case with respect to STFT size.*

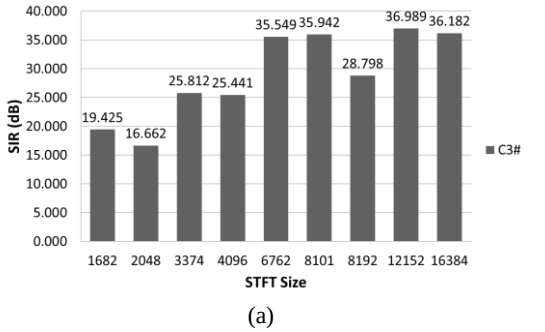


(a)

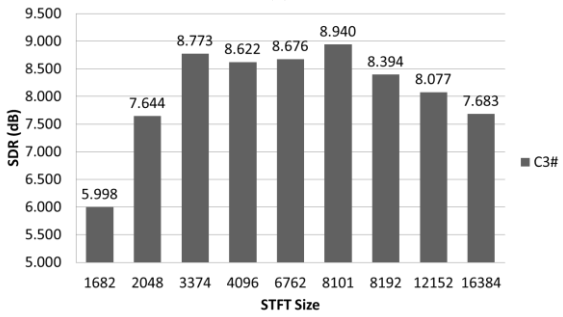


(b)

Figure 7: *The (a) SIR and (b) SDR of the note C3 of the complex case with respect to STFT size.*

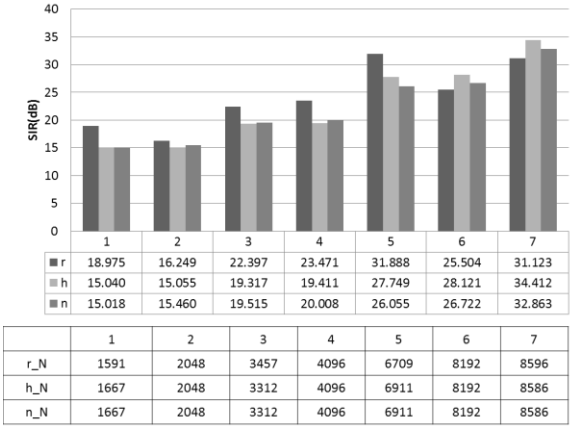


(a)



(b)

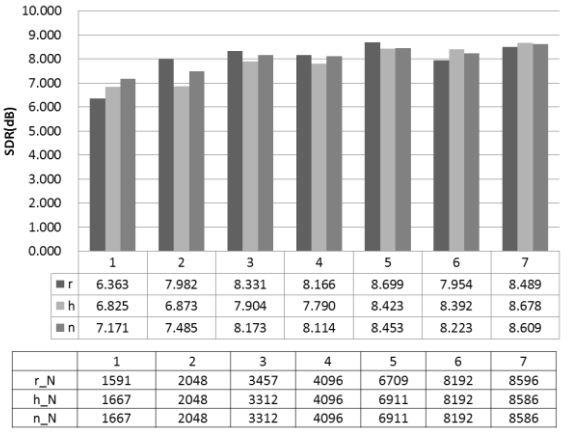
Figure 8: *The (a) SIR and (b) SDR of the note C3# of the complex case with respect to STFT size.*



	1	2	3	4	5	6	7
r	18.975	16.249	22.397	23.471	31.888	25.504	31.123
h	15.040	15.055	19.317	19.411	27.749	28.121	34.412
n	15.018	15.460	19.515	20.008	26.055	26.722	32.863

	1	2	3	4	5	6	7
r_N	1591	2048	3457	4096	6709	8192	8596
h_N	1667	2048	3312	4096	6911	8192	8586
n_N	1667	2048	3312	4096	6911	8192	8586

(a)



	1	2	3	4	5	6	7
r	6.363	7.982	8.331	8.166	8.699	7.954	8.489
h	6.825	6.873	7.904	7.790	8.423	8.392	8.678
n	7.171	7.485	8.173	8.114	8.453	8.223	8.609

	1	2	3	4	5	6	7
r_N	1591	2048	3457	4096	6709	8192	8596
h_N	1667	2048	3312	4096	6911	8192	8586
n_N	1667	2048	3312	4096	6911	8192	8586

(b)

Figure 9: *The (a) SIR and (b) SDR of the note C3 of the complex case with respect to STFT size.**r_N*,*h_N**and* ***n_N*** *shown under the figures represent the STFT size corresponding to the selected SIRs/SDRs in the case of rectangular ,hamming window and hann window respectively.*

5. CONCLUSIONS

Sound sources separation of low-pitch notes is of significant interest in analyzing recordings containing musical instruments such as cello, bass or even piano. Frequency-domain analysis using STFT usually suffers the trade-off between time- and frequency-domain localizations. In this paper, we propose a procedure to determine the proper STFT size related to the fundamental period of one target source. By applying the synthetic signals in our experiments, there are generally 2 to 6 dB improvement in SIR when compared to normal power-of-2 STFT of the similar size, without sacrificing performances in SDR and SAR. The proposed work is currently effective on separating two notes that are close in their pitches. If more than two close low-pitch notes appear simultaneously, it performs similarly to STFT using power-of-2 FFT. A recursive separation algorithm for three low-pitch notes or more is under development to solve this problem.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2001.

[2] E. Battenberg and D. Wessel, "Accelerating nonnegative matrix factorization for audio source separation on multi-core and many-core architectures," presented at the in 10th International Society for Music Information Retrieval Conference, Kobe, Japan, 2009.

[3] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 538-549, 2010.

[4] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," presented at the Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, 2003.

[5] T. M. Wang, Y. L. Chen, W. H. Liao, and A. Su, "Analysis and Trans-Synthesis of Acoustic Bowed-String Instrument Recordings – A Case Study Using Bach Cello Suites," in *International Conference on Digital Audio Effects (Dafx)*, IRCAM, Paris, French, 2011.

[6] J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Math. Computat.*, vol. 19, p. 4, 1965.

[7] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, pp. 235-238, 1977.

[8] F. j. HARRIS, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, pp. 51-83, 1978.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, p. 4, 1999.

[10] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing (3rd Edition)*: Prentice Hall, 2009.

[11] W.-C. Chang, W.-Y. Su, C. Yeh, A. Roebel, and X. Rodet, "Multiple-F0 tracking based on a high-order HMM model," in *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[12] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1462-1469, 2006.