Experimental Evaluation of Postfilter-Based Nonnegative Matrix Factorization with Statistical Model Parameter Estimation *

☆ Murota Yuki, Daichi Kitamura, Shunsuke Nakai, Hiroshi Saruwatari,
Satoshi Nakamura (Nara Institute of Science and Technology)
Kazunobu Kondo, Yu Takahashi (Yamaha Corporation)

1 Introduction

In recent years, music signal separation has been a very active area of signal-processing research. This technique is suitable for many potential applications, e.g., controlling each source in a music signal in interactive 3D audio systems and realizing automatic music transcription for each instrument player.

Common methods used for audio signal separation, which were mainly developed for speech enhancement, are nonlinear filtering algorithms such as Wiener filtering (WF) [1] and the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [2]. In particular, the MMSE-STSA estimator and its extended algorithms are optimal Bayesian estimators based on the a priori speech statistical model, resulting in a great improvement of sound quality. However, these methods cannot be effectively applied to music signals because it is difficult to deal with nonstationary interference signals. Also, the performance has strong dependence on the selection of the a priori statistical model, which cannot be determined automatically.

As another means of signal separation, nonnegative matrix factorization (NMF) has been actively studied [3], in which an input spectrogram is decomposed into the product of a spectral basis matrix and its activation matrix. In particular, supervised NMF (SNMF) [4],which includes a priori training with some sample sounds of a target instrument, can extract the target signal to some extent. Although SNMF can deal with nonstationary source signals, it has an inherent drawback that the decomposition includes *approximation* only valid for linear combinations of spectrograms, and is not strictly valid for time- (or time-frequency-) domain complex-valued mixtures. In addition, it is difficult to implement the statistical model for time sequences.

Motivated by the *complementarity* between the properties of the MMSE-STSA estimator and SNMF, in this paper, we propose a new approach based on the generalized Bayesian estimator with automatic prior estimation suitable for music signal separation. This method consists of three parts, namely, the generalized MMSE-STSA estimator [5] with a flexible target signal prior, the SNMF-based interference spectrogram estimator, and a new closed-form parameter estimation for the statistical model of the target signal time sequence based on higher-order statistics.

Compared with the conventional methods, the proposed method has the following advantages: (I) The target signal extraction is carried out via the generalized MMSE-STSA estimator so that the mixing of the time-frequency-domain complex-valued signals can be properly considered without any approximation. (II) Thanks to the SNMF-based spectrogram estimator, we can dynamically estimate the nonstationary power spectra of the interference signal. (III) The statistical model of the hidden target signal can be detected automatically only using the observable data, and can be used for optimal Bayesian estimation with online target-signal prior adaptation.

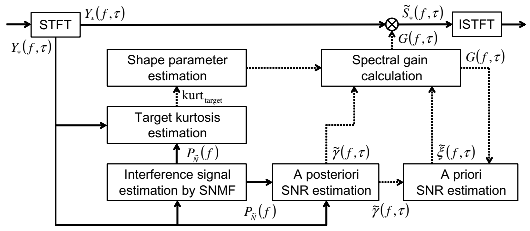

Fig. 1 Block diagram of proposed method.

2 Conventional method

2.1 Generalized MMSE-STSA estimator [5]

We apply short-time Fourier analysis to the observed signal, which is a mixture of target and interference signals, to obtain the time-frequency-domain complex-valued signal

$$Y_{\mathrm{R}}(f,\tau)+iY_{\mathrm{I}}(f,\tau)=(S_{\mathrm{R}}(f,\tau)+iS_{\mathrm{I}}(f,\tau))+(N_{\mathrm{R}}(f,\tau)+iN_{\mathrm{I}}(f,\tau)), \tag{1}$$

where $Y_*(f,\tau)$ is the observed signal, $S_*(f,\tau)$ is the target signal, $N_*(f,\tau)$ is the interference signal, and $*=\{\mathrm{R},\mathrm{I}\}$ denote the real and imaginary parts of the signal, respectively. Also, f is the frequency bin and τ is the frame index. The MMSE-STSA estimator provides the amplitude spectrum of the target signal based on the MMSE criterion under the assumption that the amplitude spectrum of the target signal obeys a Rayleigh distribution. The generalized MMSE-STSA estimator [5] provides the amplitude spectrum of the target signal under the assumption of a chi distribution.

2.2 SNMF [4]

The mixture model of NMF approximately assumes the additivity of amplitude (or power) spectrograms as

$$\sqrt{Y_{\mathrm{R}}^2(f,\tau)+Y_{\mathrm{I}}^2(f,\tau)}\approx\sqrt{S_{\mathrm{R}}^2(f,\tau)+S_{\mathrm{I}}^2(f,\tau)}+\sqrt{N_{\mathrm{R}}^2(f,\tau)+N_{\mathrm{I}}^2(f,\tau)}\,. \tag{2}$$

SNMF [4]consists of two processes, namely, training of the target sound and separation of the observed signal. In the training process, the supervised bases are trained as a *dictionary* of the target sound. Then, the observed spectrogram is decomposed into the target and other spectrograms using the supervised bases in the separation process.

3 Proposed method

3.1 Generalized MMSE-STSA estimator

Figure 1 depicts a block diagram of the proposed method. In the next part, we describe these processes in detail.

[*] "統計モデルパラメータ推定を用いたポストフィルタに基づく非負値行列因子分解の実験的評価," by 室田勇騎, 北村大地, 中井駿介, 猿渡洋, 中村哲 (奈良先端大), 近藤多伸, 高橋祐 (ヤマハ株式会社).

In the generalized MMSE-STSA estimator, the amplitude spectrum of the target signal is estimated on the basis of the MMSE criterion under a certain target prior. The processed signal $\tilde{S}(f,\tau)$ via the generalized MMSE-STSA estimator is given by

$$\tilde{S}_*(f,\tau) = G(f,\tau)Y_*(f,\tau), \tag{3}$$

$$G(f,\tau) = \frac{\sqrt{\nu(f,\tau)}}{\gamma(f,\tau)} \cdot \frac{\Gamma(\rho+0.5)}{\Gamma(\rho)} \cdot \frac{\Phi(0.5-\rho,1,-\nu(f,\tau))}{\Phi(1-\rho,1,-\nu(f,\tau))}, \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function, $\Phi(a,b;k) = F_1(a,b;k)$ is the confluent hypergeometric function, and

$$\nu(f,\tau) = \tilde{\gamma}(f,\tau)\tilde{\xi}(f,\tau)\left(1+\tilde{\xi}(f,\tau)\right)^{-1}. \tag{5}$$

Here, $\tilde{\xi}(f,\tau)$ and $\tilde{\gamma}(f,\tau)$ are the estimated a priori and a posteriori SNRs, respectively, which are defined as

$$\tilde{\xi}(f,\tau) = \alpha\tilde{\gamma}(f,\tau-1)G^2(f,\tau) + (1-\alpha)\max[\gamma(f,\tau)-1,0], \tag{6}$$

$$\tilde{\gamma}(f,\tau) = (Y_{\mathrm{R}}^2 + Y_{\mathrm{I}}^2)/P_{\tilde{N}}(f), \tag{7}$$

where $P_{\tilde{N}}(f)$ is the estimated interference power spectral density and α is the forgetting factor.

In the generalized MMSE-STSA estimator, the a priori statistical model of the target signal amplitude spectrum is set to

$$p(x) = \frac{2}{\Gamma(\rho)}\left(\frac{\rho}{\mathrm{E}[x^2]}\right)^{\rho} x^{2\rho-1}\exp\left(\frac{\rho}{\mathrm{E}\,[x^2]}x^2\right), \tag{8}$$

where $p(x)$ is the p.d.f. of signal x in the amplitude domain and ρ is the shape parameter. Here, $\rho = 1$ gives a Rayleigh distribution that corresponds to a Gaussian distribution in the time domain, and a smaller value of ρ corresponds to a super-Gaussian distribution signal.

In the generalized MMSE-STSA estimator, to calculate $\tilde{\gamma}(f,\tau)$, dynamic estimation is required if the interference signal is nonstationary, and estimation of the shape parameter ρ, which depends on the type of target signal, is also required. To solve these problems, we propose the use of SNMF as the interference signal estimator and estimate the shape parameter ρ using higher-order statistics of the target signal.

3.2 Interference signal estimation by SNMF

The following equation represents the decomposition model of SNMF using the trained supervision components $F(f,k)$:

$$\begin{aligned} A(f,\tau) &= \sqrt{Y_{\mathrm{R}}^2(f,\tau) + Y_{\mathrm{I}}^2(f,\tau)} \\ &\approx \textstyle\sum_k F(f,k)V(k,\tau) + \sum_n H(f,n)U(n,\tau), \end{aligned} \tag{9}$$

where $F(f,k)$ is a nonnegative element of the supervised basis matrix trained in advance, which involves spectral patterns of the target signal as column vectors, $V(f,k)$ is a nonnegative element of an activation matrix that corresponds to $F(f,k)$, $H(f,n)$ represents a nonnegative element of the other basis matrix, which involves residual spectral patterns that cannot be expressed by $\sum_k F(f,k)V(k,\tau)$, and $U(n,\tau)$ is a nonnegative element of the activation matrix that corresponds to $H(f,n)$. Moreover, k is the basis index of $F(f,k)$, and n is the basis index of $H(f,n)$. The supervised basis matrix can be trained using sample sounds of the target signal in the training process. Hence, ideally, $\sum_k F(f,k)V(k,\tau)$ represents the target signal components and $\sum_n H(f,n)U(n,\tau)$ represents the other components different from the target signals after the decomposition.

We can use $\sum_n H(f,n)U(n,\tau)$ (or $A(f,\tau) - \sum_k F(f,k)V(k,\tau)$) as the estimated amplitude spectrogram of the interference signal for the generalized MMSE-STSA estimator. Thus, $(\sum_n H(f,n)U(n,\tau))^2$ is regarded as a good estimate of $P_{\tilde{N}}(f)$ in (7) in the time-frequency grids even if the interference sounds are nonstationary, which is common in actual music signals.

3.3 Target signal prior estimation

3.3.1 Shape parameter and kurtosis

Generally, we cannot obtain any a priori statistical model (8) from the training data (e.g., several octave notes of the target instrument) in SNMF because the statistical time structure is quite different from that of the target signal $S_*(f,\tau)$. Also, the target signal component $\sum_k F(f,k)V(k,\tau)$ in SNMF cannot be used because its accuracy is not sometimes enough. Therefore, we inversely calculate the parameter of the target amplitude spectrogram in a data-driven manner, utilizing two observable statistics of the input signal and interference spectrogram estimated by SNMF.

Regarding the chi distribution $p(x)$ in (8), the shape parameter ρ can be written as follows:

$$\rho = (\mu_4/\mu_2^2 - 1)^{-1}, \tag{10}$$

where μ_4/μ_2^2 is called the *kurtosis* and μ_m is the mth-order moment of the amplitude spectrum. μ_m is defined as

$$\mu_m = \int_0^{\infty} x^m p(x)dx. \tag{11}$$

From this relation, the shape parameter of the subjective target signal can be estimated by obtaining its kurtosis value. In general, however, it is difficult to directly estimate the kurtosis of such a *hidden* target signal because of the contamination by additive interference signals. In the following subsections, a new algorithm of target kurtosis estimation is proposed for estimating of the shape parameter of the target p.d.f.

To cope with the mathematical problem that the mixing of the target and interference signals is additive but generally their higher-order moments are not additive, we introduce the *cumulant*, which holds the additivity for additive variables. Meanwhile, in transformation from a waveform to its amplitude spectrum, the exponentiation operation is conducted but the cumulant does not have a straightforward relationship. In this case, we use the moment instead of the cumulant. Thus, we propose *moment-cumulant transformation*.

3.3.2 Moment-cumulant transformation

In this section, we derive some formulas regarding moment-cumulant transformation. They explicitly represent the relations between the moment and cumulant in each order, which are useful for estimating the kurtosis of the target amplitude spectrum.

First, the characteristic function $\phi_x(it)$ of the random variable x is defined as

$$\phi_x(it) = \int_{-\infty}^{\infty} e^{itx}P(x)dx. \tag{12}$$

Then, we can define the mth-order moment $\mu_m(x)$ and the mth-order cumulant $\kappa_m(x)$ of x as follows:

$$\mu_m(x) = \left.\frac{\partial^{(m)}\phi_x(it)}{\partial it^{(m)}}\right|_{t=0}, \tag{13}$$

$$\kappa_m(x) = \left.\frac{\partial^{(m)}\log\phi_x(it)}{\partial it^{(m)}}\right|_{t=0}. \tag{14}$$

Next, polynomial forms of interrelations between the moment and cumulant are derived below. From (13), the mth-order moment $\mu_m(x)$ can be rewritten as

$$\begin{aligned}\mu_m(x) &= \left.\frac{\partial^{(m)}\exp(\log\phi_x(it))}{\partial it^{(m)}}\right|_{t=0}\\ &= \textstyle\sum_{\pi(m)}\prod_{B\in\pi(m)}\kappa_{|B|}(x),\end{aligned} \tag{15}$$

where we use a *combinational form of Faà di Bruno's formula*,

$$\frac{\partial^{(m)}f(g(x))}{\partial x^{(m)}} = \textstyle\sum_{\pi(m)}f^{(|\pi(m)|)}(g(x))\prod_{B\in\pi(m)}[g(x))]^{(|B|)}, \tag{16}$$

where $\pi(m)$ runs through the list of all partitions of a set of size m, $B\in\pi(m)$ means that B is one of the blocks into which the set is partitioned, and $|B|$ is the size of the set B.

In the same manner, from (14), the mth-order cumulant $\kappa_m(x)$ is given by

$$\begin{aligned}\kappa_m(x) &= \textstyle\sum_{\pi(m)}\log^{(|\pi(m)|)}(\phi_x(it))\prod_{B\in\pi(m)}[\phi_x(it))]^{(|B|)}\Big|_{t=0}\\ &= \textstyle\sum_{\pi(m)}(-1)^{|\pi(m)|-1}(|\pi(m)|-1)!\prod_{B\in\pi(m)}\mu_{|B|}(x).\end{aligned} \tag{17}$$

3.3.3 Estimation of target signal kurtosis from observations

In this section, we estimate the amplitude-domain kurtosis of the target signal. First, we express the target kurtosis by using complex-valued variables as intrinsic parameters. After that, we show that the kurtosis can be represented in the amplitude-spectrogram domain. Hereafter, we ignore the indexes f, τ, k, and n of each signal. Only the statistics of $(Y_\mathrm{R}+iY_\mathrm{I})$ and $(N_\mathrm{R}^2+N_\mathrm{I}^2)^{1/2}$ obtained by SNMF are observable, and $(S_\mathrm{R}^2+S_\mathrm{I}^2)^{1/2}$ is a hidden value to be estimated. First, we assume to obtain the following mth-order moments from the data:

$$\mu_m(Y_\mathrm{R}) = \mathrm{E}[Y_\mathrm{R}^m], \tag{18}$$

$$\mu_m(Y_\mathrm{I}) = \mathrm{E}[Y_\mathrm{I}^m], \tag{19}$$

$$\mu_m(N_\mathrm{R}) = \mathrm{E}[N_\mathrm{R}^m], \tag{20}$$

$$\mu_m(N_\mathrm{I}) = \mathrm{E}[N_\mathrm{I}^m]. \tag{21}$$

Generally, the cumulant has the property of additivity for additive independent variables, i.e., $\kappa_m(a+b) = \kappa_m(a)+\kappa_m(b)$. Using this relation and (17), we can estimate the cumulant of the real part of the target signal $S_\mathrm{R}{=}Y_\mathrm{R}-N_\mathrm{R}$ as

$$\begin{aligned}\kappa_m(S_\mathrm{R}) &= \kappa_m(Y_\mathrm{R}) - \kappa_m(N_\mathrm{R})\\ &= \textstyle\sum_{\pi(m)}(-1)^{|\pi(m)|-1}(|\pi(m)|-1)!\prod_{B\in\pi(m)}\mu_{|B|}(Y_\mathrm{R})\\ &\quad-\textstyle\sum_{\pi(m)}(-1)^{|\pi(m)|-1}(|\pi(m)|-1)!\prod_{B\in\pi(m)}\mu_{|B|}(N_\mathrm{R}).\end{aligned} \tag{22}$$

The statistics of the squared variable of S_R is given by

$$\mu_m(S_\mathrm{R}^2) = \mu_{2m}(S_\mathrm{R}) = \textstyle\sum_{\pi(2m)}\prod_{B\in\pi(2m)}\kappa_{|B|}(S_\mathrm{R}). \tag{23}$$

In the same manner, we can estimate the statistics of the squared variable of S_I. Given $\mu_m(S_\mathrm{R}^2)$ and $\mu_m(S_\mathrm{I}^2)$, we can calculate the cumulant of the power spectrum $S_\mathrm{R}^2{+}S_\mathrm{I}^2$ as

$$\begin{aligned}&\kappa_m(S_\mathrm{R}^2+S_\mathrm{I}^2)\\ &= \kappa_m(S_\mathrm{R}^2)+\kappa_m(S_\mathrm{I}^2)\\ &= \textstyle\sum_{\pi(m)}(-1)^{|\pi(m)|-1}(|\pi(m)|-1)!\prod_{B\in\pi(m)}\mu_{|B|}(S_\mathrm{R}^2)\\ &\quad+\textstyle\sum_{\pi(m)}(-1)^{|\pi(m)|-1}(|\pi(m)|-1)!\prod_{B\in\pi(m)}\mu_{|B|}(S_\mathrm{I}^2),\end{aligned} \tag{24}$$

and the mth-order moment of the power spectrum is given by

$$\mu_m(S_\mathrm{R}^2+S_\mathrm{I}^2) = \textstyle\sum_{\pi(m)}\prod_{B\in\pi(m)}\kappa_{|B|}(S_\mathrm{R}^2+S_\mathrm{I}^2). \tag{25}$$

Furthermore, the mth-order moment of the amplitude spectrum $(S_\mathrm{R}^2+S_\mathrm{I}^2)^{1/2}$ is

$$\mu_m((S_\mathrm{R}^2+S_\mathrm{I}^2)^{\frac{1}{2}}) = \mu_{\frac{m}{2}}(S_\mathrm{R}^2+S_\mathrm{I}^2). \tag{26}$$

Using (18)–(26), we can estimate the resultant kurtosis of the target amplitude spectrum as

$$\begin{aligned}\mathrm{kurt}_\mathrm{target} &= \frac{\mu_4((S_\mathrm{R}^2+S_\mathrm{I}^2)^{\frac{1}{2}})}{\mu_2^2((S_\mathrm{R}^2+S_\mathrm{I}^2)^{\frac{1}{2}})}\\ &= \frac{\mathcal{N}(\mu_m(Y_\mathrm{R}),\mu_m(Y_\mathrm{I}),\mu_m(N_\mathrm{R}),\mu_m(N_\mathrm{I}))}{\mathcal{D}(\mu_m(Y_\mathrm{R}),\mu_m(Y_\mathrm{I}),\mu_m(N_\mathrm{R}),\mu_m(N_\mathrm{I}))},\end{aligned} \tag{27}$$

where

$$\begin{aligned}&\mathcal{N}(\mu_m(Y_\mathrm{R}),\mu_m(Y_\mathrm{I}),\mu_m(N_\mathrm{R}),\mu_m(N_\mathrm{I}))\\ &= \mu_4(Y_\mathrm{R})+\mu_4(Y_\mathrm{I})-\mu_4(N_\mathrm{R})-\mu_4(N_\mathrm{I})\\ &\quad+6\mu_2^2(N_\mathrm{R})+6\mu_2^2(N_\mathrm{I})+2\mu_2(Y_\mathrm{R})\mu_2(Y_\mathrm{I})+2\mu_2(N_\mathrm{R})\mu_2(N_\mathrm{I})\\ &\quad-6\mu_2(Y_\mathrm{R})\mu_2(N_\mathrm{R})-6\mu_2(Y_\mathrm{I})\mu_2(N_\mathrm{I})\\ &\quad-2\mu_2(Y_\mathrm{R})\mu_2(N_\mathrm{I})-2\mu_2(Y_\mathrm{I})\mu_2(N_\mathrm{R}),\end{aligned} \tag{28}$$

$$\begin{aligned}&\mathcal{D}(\mu_m(Y_\mathrm{R}),\mu_m(Y_\mathrm{I}),\mu_m(N_\mathrm{R}),\mu_m(N_\mathrm{I}))\\ &= \mu_2^2(Y_\mathrm{R})+\mu_2^2(Y_\mathrm{I})+\mu_2^2(N_\mathrm{R})+\mu_2^2(N_\mathrm{I})+2\mu_2(Y_\mathrm{R})\mu_2(Y_\mathrm{I})\\ &\quad-2\mu_2(Y_\mathrm{R})\mu_2(N_\mathrm{R})-2\mu_2(Y_\mathrm{R})\mu_2(N_\mathrm{I})-2\mu_2(Y_\mathrm{I})\mu_2(N_\mathrm{R})\\ &\quad-2\mu_2(Y_\mathrm{I})\mu_2(N_\mathrm{I})+2\mu_2(N_\mathrm{R})\mu_2(N_\mathrm{I}).\end{aligned} \tag{29}$$

Next, the sums of the 4th-order moments $\mu_4(Y_\mathrm{R})+\mu_4(Y_\mathrm{I})$ and $\mu_4(N_\mathrm{R})+\mu_4(N_\mathrm{I})$ are represented by the amplitude-domain kurtosis of the observed signal spectrum and the interference signal spectrum as

$$\begin{aligned}\mu_4(Y_\mathrm{R})+\mu_4(Y_\mathrm{I}) &= \left(\mu_2^2(Y_\mathrm{R})+\mu_2^2(Y_\mathrm{I})+2\mu_2(Y_\mathrm{R})\mu_2(Y_\mathrm{I})\right)\frac{\mu_4(A)}{\mu_2^2(A)}\\ &\quad-2\mu_2(Y_\mathrm{R})\mu_2(Y_I),\end{aligned} \tag{30}$$

$$\begin{aligned}\mu_4(N_\mathrm{R})+\mu_4(N_\mathrm{I}) &= \left(\mu_2^2(N_\mathrm{R})+\mu_2^2(N_\mathrm{I})+2\mu_2(N_\mathrm{R})\mu_2(N_\mathrm{I})\right)\frac{\mu_4(\sum_n HU)}{\mu_2^2(\sum_n HU)}\\ &\quad-2\mu_2(N_\mathrm{R})\mu_2(N_\mathrm{I}).\end{aligned} \tag{31}$$

If we assume that the real and imaginary parts are i.i.d., $\mu_2(Y_\mathrm{R})$ equals $\mu_2(Y_\mathrm{I})$ and $\mu_2(N_\mathrm{R})$ equals $\mu_2(N_\mathrm{I})$. Under this assumption and (15) and (17), we obtain the following relation for the SNMF output:

$$\mu_2(Y_\mathrm{R}) = \mu_2(Y_\mathrm{I}) = \frac{1}{2}\mu_2(A), \tag{32}$$

$$\mu_2(N_\mathrm{R}) = \mu_2(N_\mathrm{I}) = \frac{1}{2}\mu_2(\textstyle\sum_n HU). \tag{33}$$

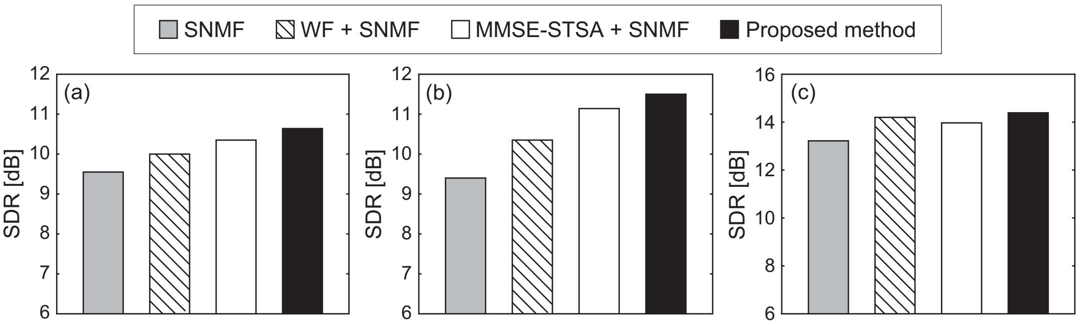

Fig. 2 Average SDRs for extraction of (a) clarinet signal, (b) oboe signal, and (c) cello signal for each method.

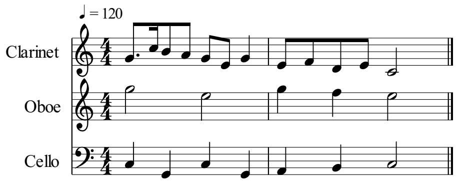

Fig. 3 Scores of each part.

Finally, using (30)–(33), (27) is rewritten as follows:

$$\mathrm{kurt}_{\mathrm{target}} = \frac{\mu_4(A) - \mu_4\left(\sum_n HU\right) + 4\mu_2^2(\sum_n HU) - 4\mu_2(A)\mu_2(\sum_n HU)}{\mu_2^2(A) + \mu_2^2(\sum_n HU) - 2\mu_2(A)\mu_2(\sum_n HU)}. \tag{34}$$

The shape parameter of the target p.d.f. can be estimated using (10) and (34). Note that all the estimates can be obtained from the result of SNMF without using any waveforms; this implies a good applicability to the combination with SNMF.

4 Experiments

4.1 Experimental conditions

In this experiment, we compared four methods, i.e., simple SNMF [4], WF with SNMF-based interference estimation (**WF+SNMF**) [6], the MMSE-STSA estimator with SNMF-based interference estimation (**MMSE-STSA+SNMF**), and the proposed method. WF and the MMSE-STSA estimator utilized the interference spectrogram estimated by SNMF. We used three instrumental signals, namely, an oboe, clarinet, and cello, as the target sounds (each melody part is depicted in Fig. 3). These signals were artificially generated by a MIDI synthesizer, and the observed signals were produced by mixing two sources selected from three signals with the same power. In estimation of the interference signal using SNMF, we used the same MIDI sounds of the target instruments as supervision for the training process. The training sounds contained two octave notes that cover all the notes of the target signal in the observed signal. The spectrograms were computed using a 92-ms-long rectangular window with a 46-ms overlap shift. Moreover, the number of trained bases was 100 and the number of other bases was 50. In the proposed method, the forgetting factor α and amplitude compression parameter β were set to 0.1 and 1.0, respectively.

4.2 Experimental results and discussion

We used the signal-to-distortion ratio (SDR) defined in [7] as the evaluation score. SDR indicates the overall quality of the separated target sound, showing high separation and less artificial distortion.

Figure 2 shows the average SDRs for each method and each target instrument. From the SDR results, we can confirm that the separation performance of the proposed method is better than those of the other methods. This result indicates the efficacy of introducing the flexible a priori statistical model of the target signal. The simple MMSE-STSA estimator also assumes the fixed a priori model of the Gaussian distribution but the assumption is not appropriate for a music target signal. In contrast, the proposed method almost always uses more spiky p.d.f. ($\rho < 1$), which enhances the true sparseness of the target music signal.

5 Conclusions

In this paper, we propose a new approach for addressing music signal separation based on the generalized Bayesian estimator with automatic prior adaptation. From the experimental evaluation, it is found that the proposed method outperforms competitive methods, namely, simple NMF, WF, and the MMSE-STSA estimator with a fixed Gaussian prior.

References

[1] P. C. Loizou, *"Speech Enhancement Theory and Practice"* CRC Press, Taylor & Francis Group, FL, 2007.

[2] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.32, no.6, pp.1109–1121, 1984.

[3] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," *Proc. Advances in Neural Information Processing Systems*, vol.13, pp.556–562, 2001. 2001.

[4] P. Smaragdis, B. Raj, M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," *Proc. 7th International Conference on Independent Component Analysis and Signal Separation*, pp.414–421, 2007.

[5] C. Breithaupt, M. Krawczyk, R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.4037–4040, 2008.

[6] E. M. Grais, H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," *Proc. IEEE 17th International Conference on Digital Signal Processing*, pp.1–6, 2011

[7] E. Vincent, R. Gribonval, C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol.14, no.4, pp.1462–1469, 2006.