Consistent Wiener filtering: designing generalized time-frequency masks respecting spectrogram consistency *

◎ Jonathan Le Roux (NTT CS Labs), Emmanuel Vincent (INRIA),
Yuu Mizuno (The University of Tokyo), Hirokazu Kameoka (NTT CS Labs),
Nobutaka Ono, Shigeki Sagayama (The University of Tokyo)

1 Introduction

Wiener filtering has been one of the most widely used methods for source separation for several decades, in particular in audio signal processing. To exploit the short-term stationarity of audio signals, it is very often applied on time-frequency representations [1], especially the short-time Fourier transform (STFT). However, classical Wiener filtering does not take into account the intrinsically redundant structure of STFT spectrograms, and its output is actually in general not the optimal solution. We show here that by ensuring that the output spectrograms are "consistent", i.e., that they correspond to actual time-domain signals, we can obtain a more efficient filtering. As Wiener filtering is widely used as a post-processing for many methods involving the estimation of the power spectrograms of the component signals (non-negative matrix factorization, AR modeling, etc.) or in time-frequency masking in computational auditory scene analysis, it is of tremendous importance to ensure that the information gathered by those algorithms is best exploited. We generalize here the concept of Wiener filtering to time-frequency masks which can involve a manipulation of the phase as well in order to find the true Maximum-Likelihood solution, by focusing on the concept of consistency, which we already exploited in [2] for fast phase restoration and [3] to improve Kameoka et al.'s complex NMF decomposition [4].

2 Wiener filtering and consistency

2.1 Maximum-Likelihood formulation of the Wiener filtering problem

We assume that the observed signal x is the mixture of two signals, a target s_1 and an interference signal s_2, analyzed using an STFT with frame shift R. We further assume that the STFT coefficients S_1 and S_2 of the signals s_1 and s_2 at each time frame t and frequency bin ω are modeled as statistically independent Gaussian random variables with variance σ_1^2 and σ_2^2 respectively. For convenience of notation, we shall write $\nu^{(i)} = 1/\sigma_i^2$. Note that the case of several interference signals $s_2, \dots, s_I$ can be reduced, without loss of generality, to that of two sources only, as we assume in particular that the sources are not correlated. We would then consider a global interference source $\tilde{s}_2 = \sum_{i=2}^{I} s_i$, and the variance $\tilde{\sigma}_2^2$ would be equal to $\sum_{i=2}^{I} \sigma_i^2$.

Denoting by X the spectrogram of the observed signal, classical Wiener filtering consists in maximizing the log-likelihood of the STFT coefficients S_1 and S_2, which can be written, under the constraint that $X = S_1 + S_2$, as a function of $S = S_1$ only:

$$\mathcal{L}(S) = -\frac{1}{2}\Big(\sum_{\omega,t} \nu_{\omega,t}^{(1)} |S_{\omega,t}|^2 + \sum_{\omega,t} \nu_{\omega,t}^{(2)} |X_{\omega,t} - S_{\omega,t}|^2\Big) + C(\nu^{(1)}, \nu^{(2)}), \quad (1)$$

where C is a constant depending only on $\nu^{(1)}$, $\nu^{(2)}$. Introducing the classical Wiener filtering estimate for S_1,

$$\hat{S}_{\omega,t} = \frac{\nu_{\omega,t}^{(2)}}{\nu_{\omega,t}^{(1)} + \nu_{\omega,t}^{(2)}} X_{\omega,t}, \quad (2)$$

the Maximum-Likelihood problem can be reformulated as the minimization of the objective function

$$\psi(S) = \sum_{\omega,t} \alpha_{\omega,t} |S_{\omega,t} - \hat{S}_{\omega,t}|^2 \quad (3)$$

where $\alpha_{\omega,t} = \nu_{\omega,t}^{(1)} + \nu_{\omega,t}^{(2)}$.

2.2 Wiener filtering with consistency constraint

If no further constraint is assumed on S, the objective function is obviously minimized for $S = \hat{S}$. However, we need to keep in mind that the STFT is a redundant representation with a particular structure. Denoting by N the number of frequency bins and T the number of frames, STFT spectrograms of time-domain signals are elements of $\mathbb{C}^{NT}$, which we shall call "consistent spectrograms", but one of the fundamental points of this paper is that not all elements of $\mathbb{C}^{NT}$ can be obtained as such [5, 2]. If we assume that inverse STFT is performed in such a way that there is "perfect reconstruction", i.e., that a signal can be exactly reconstructed from its spectrogram through inverse STFT, then we showed in [2] that a necessary and sufficient condition for an array W to be a consistent spectrogram is for it to be equal to the STFT of its inverse STFT. The set of consistent spectrograms can thus be described as the null space $\mathrm{Ker}(\mathcal{F})$ of the $\mathbb{R}$-linear operator $\mathcal{F}$ from $\mathbb{C}^{NT}$ to itself defined by

$$\mathcal{F}(W) = \mathcal{G}(W) - W, \quad (4)$$

where $\mathcal{G}(W) = \mathrm{STFT}(\mathrm{iSTFT}(W))$.

Going back to the Wiener filtering problem, if we now impose that the solution be consistent, the problem amounts to finding a consistent spectrogram S minimizing ψ, or in other words to minimize ψ under the constraint that $\mathcal{F}(S) = 0$. Imposing consistency is not a mere elegance or theory-oriented concern, but a truly fundamental problem. Indeed, the spectrogram of the signal resynthesized from the classical Wiener filter spectrogram $\hat{S}$ is actually different in general from $\hat{S}$, and is no longer maximizing the Wiener log-likelihood

* Consistent Wiener filtering：スペクトログラム無矛盾性を保証する一般化時間周波数マスクの設計法、
ルルー・ジョナトン (NTT)、ヴァンサン・エマヌエル (INRIA)、水野優 (東大情報理工)、亀岡弘和 (NTT)、
小野順貴、嵯峨山茂樹 (東大情報理工)

(or minimizing ψ), so that the final result of the processing that we are listening to is in fact not the optimal solution. What we really want to do is to find a signal in the time domain such that its spectrogram minimizes the Wiener criterion ψ, or, formulating this in the time-frequency domain, to minimize the following "true" objective function

$$\tilde{\psi}(S) = \sum_{\omega,t} \alpha_{\omega,t} \left| \mathcal{G}(S)_{\omega,t} - \hat{S}_{\omega,t} \right|^2, \qquad (5)$$

where $\mathcal{G}(S)$ is again the spectrogram of the signal resynthesized from S by inverse STFT. If S is constrained to be consistent, then the objective functions $\tilde{\psi}$ and ψ are equal, and one possibility to solve our problem is to minimize $\tilde{\psi}$ by minimizing ψ under that constraint. Another one is to solve the problem directly in the time domain, by considering the signal as the parameter. Yet another possibility is to relax the consistency constraint by introducing it as a penalty function: if the weight of the penalty is chosen sufficiently large, or is increased during the course of the optimization, the estimated spectrogram should finally be both consistent and minimizing ψ among the consistent spectrograms. We shall now investigate these three possibilities.

3 Optimization algorithms

3.1 Time-domain formulation

The consistent Wiener filtering optimization problem amounts to minimizing $\sum_{\omega,t} \alpha_{\omega,t} |S_{\omega,t} - \hat{S}_{\omega,t}|^2$ on the subspace of consistent spectrograms, while the problem of estimating the signal whose STFT spectrogram is closest to the modified STFT spectrogram $\hat{S}$ amounts to minimizing $\sum_{\omega,t} |S_{\omega,t} - \hat{S}_{\omega,t}|^2$ on the same subspace [5]. The latter problem can be transformed through Parseval's theorem into the minimization of a simple quadratic form on the time signal parameters, but the weights α make here the computation of the optimal signal cumbersome as they hinder us from simplifying the product of the Fourier matrix and its transpose. If we note A_t the $N \times N$ diagonal matrix with diagonal coefficients $\alpha_{\omega,t}$, F the $N \times N$ Fourier transform matrix, w_t the $N \times L$ matrix which computes the t-th windowed frame of the signal x (of length L), and $\hat{s}_t$ the inverse transform of the t-th STFT frame of $\hat{S}$, then we can show that the optimal signal x is given by

$$\hat{x} = \Big(\sum_t w_t^H F^H A_t F w_t \Big)^{-1} \sum_t w_t^H F^H A_t F \hat{s}_t. \quad (6)$$

If A_t were not present, as in Griffin and Lim's case, then $F^H F$ would simplify to $N\mathrm{Id}$ and we would get the simple weighted overlap-add estimation $x = \sum_t w_t^H \hat{s}_t / \sum_t w_t^H w_t$. However, the simplification cannot be done in the consistent Wiener filtering problem, leading to a very large ($L \times L$) matrix inversion problem. Still, this matrix is band-diagonal (and Hermitian), and solving the system is possible in a reasonable amount of time and using a reasonable amount of memory space. To reduce in particular the memory requirements, we can split in practice the estimation of the time domain signal on overlapping blocks of a few frames, and reconstruct an approximate solution on the whole interval by overlap-add from the locally optimal signals.

3.2 Operator splitting

If we let $f_1 = \psi$ and $f_2 = i_{\mathrm{Ker}(\mathcal{F})}$, where $i_{\mathrm{Ker}(\mathcal{F})}$ is the indicator function of $\mathrm{Ker}(\mathcal{F})$ defined by $i_{\mathrm{Ker}(\mathcal{F})}(S) = 0$ if $\mathcal{F}(S) = 0$ and $i_{\mathrm{Ker}(\mathcal{F})}(S) = +\infty$ if $\mathcal{F}(S) \neq 0$, then finding a consistent spectrogram S which minimizes ψ amounts to finding the global minimum of $f_1 + f_2$. f_1 and f_2 are both proper lower semi-continuous convex functions. This kind of minimization problem has been studied in convex optimization theory, and can be efficiently solved using the so-called Douglas-Rachford splitting algorithm for monotone operators. We shall refer to [6] for more details and references. For every $S \in \mathbb{C}^{NT}$, the function $Z \mapsto \frac{1}{2}\|Z - S\|^2 + f_i(Z)$ achieves its infimum at a unique point denoted by $\mathrm{prox}_{f_i}(S)$. The uniquely-valued operator thus defined is called the proximity operator of f_i. Here, $\mathrm{prox}_{\beta f_1}$ and $\mathrm{prox}_{\beta f_2}$, where $\beta > 0$ is a constant which will be used later on, can be explicitely computed:

$$\mathrm{prox}_{\beta\psi}(S)_{\omega,t} = \frac{\beta\alpha_{\omega,t}\hat{S}_{\omega,t} + \frac{1}{2}S_{\omega,t}}{\beta\alpha_{\omega,t} + \frac{1}{2}}, \qquad (7)$$

$$\mathrm{prox}_{\beta i_{\mathrm{Ker}(\mathcal{F})}}(S) = \mathcal{G}(S). \qquad (8)$$

Eq. (7) is simply obtained by minimizing a second-order function; Eq. (8) is obtained by noticing that the minimum of $\frac{1}{2}\|Z-S\|^2 + i_{\mathrm{Ker}(\mathcal{F})}(Z)$ is an element of $\mathrm{Ker}(\mathcal{F})$ which minimizes $\|Z - S\|^2$, i.e., a consistent spectrogram closest to S. As shown by Griffin and Lim [5], $\mathcal{G}(S)$ is such a spectrogram if we assume, as we shall do, that the inverse STFT is performed using the windowed overlap-add procedure with the synthesis window before normalization equal to the analysis window. Applying the Douglas-Rachford splitting to this problem, we obtain the following algorithm. Let $S^{(0)} \in \mathbb{C}^{NT}$, $\beta > 0$, $(\lambda_p)_{p\in\mathbb{N}}$ be a sequence in $(0, 2)$ such that $\sum_p \lambda_p(2 - \lambda_p) = +\infty$, and define the recursion

$$S^{(p+1)}_{\omega,t} = S^{(p)}_{\omega,t} + \lambda_p \frac{\beta\alpha_{\omega,t}(\hat{S}_{\omega,t} - \mathcal{G}(S^{(p)}))_{\omega,t}) + \frac{1}{2}\mathcal{F}(S^{(p)})_{\omega,t}}{\beta\alpha_{\omega,t} + \frac{1}{2}}, \quad (9)$$

then $S^{(p)} \to \check{S}$ and $\mathcal{G}(\check{S})$ is a solution of the consistent Wiener filtering problem, i.e., it is both consistent and minimizing ψ. If we assume that $\lambda_p = 1$ and write $\gamma = \frac{1}{2\beta}$, then the update becomes

$$S^{(p+1)}_{\omega,t} = \frac{\alpha_{\omega,t}(\hat{S}_{\omega,t} - \mathcal{F}(S^{(p)}))_{\omega,t}) + \gamma\mathcal{G}(S^{(p)})_{\omega,t}}{\alpha_{\omega,t} + \gamma}, \quad (10)$$

which, as we shall see later, is very close to the update obtained when introducing consistency as a penalty function.

3.3 Consistency as a penalty function

For an array of complex numbers $W \in \mathbb{C}^{NT}$, $\mathcal{F}(W)$ represents the relation between W and the STFT of its inverse STFT. Instead of enforcing consistency through the "hard" constraint $\mathcal{F}(W) = 0$, which may be difficult to handle, we can relax that constraint by using any vector norm of $\mathcal{F}(W)$ to derive a numerical criterion which can be used to quantify how far an array of complex numbers is from being consistent. We consider

here in particular the L^2 norm of $\mathcal{F}(W)$, which leads, as shown in [2], to a criterion which is related to that used by Griffin and Lim to derive their iterative STFT algorithm [5]. Introducing the consistency penalty in (3), the new objective function to minimize reads

$$\psi_\gamma(S) = \psi(S) + \gamma \sum_{\omega,t} \left|\mathcal{G}(S)_{\omega,t} - S_{\omega,t}\right|^2. \qquad (11)$$

An efficient iterative optimization algorithm for ψ_γ can be derived through the auxiliary function method [7]. A function $\psi_\gamma^+(S,\bar{S})$ verifying $\psi_\gamma(S) = \min_{\bar{S}} \psi_\gamma^+(S,\bar{S})$, $\forall S$, is called an auxiliary function for $\psi_\gamma(S)$, and $\bar{S}$ an auxiliary variable. The minimization of ψ_γ can be performed indirectly by alternating minimizations of ψ_γ^+ w.r.t. S and $\bar{S}$.

Assuming here again that the synthesis window before normalization in the inverse STFT is equal to the analysis window, it results from [5] that $\mathcal{G}(S)$ is the closest consistent spectrogram to S in a least-squares sense:

$$\sum_{\omega,t} \left|\mathcal{G}(S)_{\omega,t} - S_{\omega,t}\right|^2 = \min_{\bar{S}\in \mathrm{Ker}(\mathcal{F})} \sum_{\omega,t} \left|\bar{S}_{\omega,t} - S_{\omega,t}\right|^2, \ \forall S. \qquad (12)$$

If we now define the function $\psi_\gamma^+ : \mathbb{C}^{NT} \times \mathrm{Ker}(\mathcal{F}) \to \mathbb{R}$ such that $\forall S \in \mathbb{C}^{NT}, \forall \bar{S} \in \mathrm{Ker}(\mathcal{F})$,

$$\psi_\gamma^+(S,\bar{S}) = \psi(S) + \gamma \sum_{\omega,t} \left|S_{\omega,t} - \bar{S}_{\omega,t}\right|^2, \qquad (13)$$

we easily see from (12) that ψ_γ^+ is an auxiliary function for ψ_γ. This leads to an iterative optimization scheme in which, starting at step p from a spectrogram $S^{(p)}$, $\bar{S}$ is first updated to $\mathcal{G}(S^{(p)})$, and the new estimate $S^{(p+1)}$ is simply estimated as the minimum of a second-order form with diagonal coefficients, altogether resulting in the following update equation:

$$S_{\omega,t}^{(p+1)} \leftarrow \frac{\alpha_{\omega,t}\hat{S}_{\omega,t} + \gamma \mathcal{G}(S_{\omega,t}^{(p)})}{\alpha_{\omega,t} + \gamma}. \qquad (14)$$

We note that this update is very close but slightly different from the update (10) obtained through the application of the Douglas-Rachford splitting.

4 Experimental evaluation

4.1 Settings and implementations

The sampling rate was 16 kHz. All the spectrograms considered were built with a frame length $N = 1024$, a frame shift $R = 512$ if not specified, and with a sine window for both analysis and synthesis.

The time-domain method was implemented as follows: the analytical solution is computed separately on blocks of 64 STFT frames; the blocks have a 50 % overlap, and the resulting short-time signals are cross-faded on a small region (here 16 frames) around the center of the overlap regions in order to throw away portions of signal near the block boundaries, as we can expect them to suffer from boundary effects. The above values for the block size and the amount of overlap and cross-fade were determined experimentally so as to minimize computation and memory costs while still obtaining solutions with a true Wiener criterion very close to that of the analytical solution computed on the whole interval.

Table 1 *Performance comparison results*

	Time (s)	$\tilde{\psi}$	SNR (dB)
Wiener	0.1	1.91×10^6	15.2
Griffin-Lim	148.5	3.85×10^8	9.9
Time domain	794.8	2.76×10^4	17.8
Splitting	133.7	2.90×10^4	16.1
Penalty	6.8	2.90×10^4	17.2

For both the splitting algorithm and the penalty-based algorithm, heuristically, the larger γ, the slower the convergence, but the better the solution. For the penalty function algorithm, we noticed experimentally that the criterion $\tilde{\psi}$ monotonically decreased through the update (14) with γ fixed when starting from a point obtained through updates with a smaller γ. We thus designed an automatic update scheme for γ: starting from a very small value γ_0 (typically 10^{-5}) for γ, we update S through (14) while slightly increasing γ by δ (initially set to γ_0 as well) until the decrease of $\tilde{\psi}$ becomes slower than 1 %, in which case we update δ to 2δ and restart the S updates. The algorithm stops after two increases of δ without significant improvement of $\tilde{\psi}$, which typically occurred after around 200 iterations. The monotonical decrease behavior was not as obvious for the splitting algorithm, and we thus ran it for 4000 iterations with a large γ experimentally fixed to 10^4.

4.2 Separation under oracle conditions

We evaluate here the performance of the proposed methods in terms of computation time and final value of the "true" Wiener criterion $\tilde{\psi}$ for the separation of a 5.5 s mixture of two female speakers under oracle conditions, i.e., assuming that the true power spectrograms of both sources are known. For comparison, we also give the results for the classical Wiener filter output $\hat{S}$ ("Wiener") and for the spectrogram whose magnitude is closest to the magnitude of the classical Wiener filter, computed through Griffin and Lim's phase reconstruction algorithm ran for 4000 iterations ("Griffin-Lim"). This way of obtaining a consistent spectrogram through post-processing of the classical Wiener filter magnitude seems indeed a natural method to attempt.

The results are summarized in Table 1. Although the performance of the classical Wiener filter is already very good, we can see that the proposed methods all lead to significant improvements in both the true Wiener criterion $\tilde{\psi}$ and the signal-to-noise ratio (SNR), while simply reconstructing the phase as a post-processing does not solve the problem (higher $\tilde{\psi}$, lower SNR). The increase in SNR may not seem straightforward, but it can be understood as a result of the fact that with our methods the spectrogram of the resynthesized signal is closer to the intended ML solution. Computation of the analytical solution in the time domain is very costly, but enables us to see that the solution obtained in much less time with the penalty-based algorithm is close to optimal. We will use this algorithm for the noise reduction experiments below.

We also studied the influence of the frame shift on performance. Results are summarized in Table 2. We can see that the SNR increases with the amount of overlap between frames, especially for the analytical solu-

Table 2 *Evolution of SNR (dB) w.r.t. overlap*

	50 %	75 %	87.5 %
Wiener	15.2	15.6	15.6
Griffin-Lim	9.9	12.4	12.5
Time domain	17.8	19.4	20.3
Splitting	16.1	16.5	16.7
Penalty	17.2	17.7	17.9

Table 3 *SNR (dB) of the denoised speech*

Initial SNR		-10 dB	0 dB	10 dB
Oracle	Wiener	9.6	14.6	20.5
	Penalty	10.5	15.6	21.4
Variance	Wiener	8.8	13.8	19.8
	Penalty	8.9	14.0	20.1
Subtraction	Wiener	-3.0	6.5	15.3
	Penalty	4.1	10.3	17.0

tion. This could be expected as consistency constraints become stronger when overlap increases. Computation time of course also increases with the amount of overlap, roughly linearly with the total number of spectrogram frames for all the methods.

4.3 Noise reduction experiments

We performed noise reduction experiments on speech by a female speaker mixed with Gaussian white noise, under three conditions: 1) the power spectrograms of both the speech and noise are known ("oracle"); 2) the power spectrogram of speech is known, while only the variance of the noise is known ("variance"); 3) only variance of the noise is known, and the power spectrogram of speech is estimated by spectral subtraction ("subtraction") [8]. We compare here the penalty-based algorithm with the classical Wiener filter for three initial SNR settings: -10 dB, 0 dB and 10 dB. Average results on 10 different noise signals for each SNR are summarized in Table 3. We can see that using the penalty-based algorithm leads to significant improvements in particular in the spectral subtraction condition, which is also the most realistic. Perceptually, although musical noise is still strong, the residual noise present in the Wiener filter estimate is much weaker in the penalty-based one. Fig. 1 shows the spectrograms of the noisy speech for 0 dB SNR (top), of the denoised speech obtained using classical Wiener filter (middle) and of the denoised speech obtained using the penalty-based updates (bottom). We believe that the very important improvements obtained with our method constitute a major result.

5 Conclusion

We presented a new framework for Wiener filtering and more generally time-frequency masking which takes into account the consistency of spectrograms to compute the true optimal solution to the Wiener filtering problem. We presented three methods to find optimal or near optimal solutions, investigated their performance in comparison with previous works, and showed in particular that our method combined with spectral subtraction outperforms classical Wiener filtering. Future works include the reduction of computation time by combining this work with the fast approximations investigated in [9].

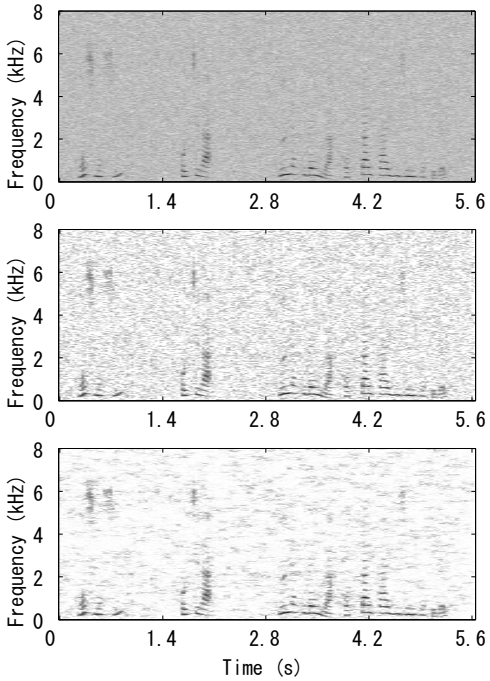

Fig. 1 *Example of speech denoising.*

References

[1] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Kluwer, 2004, pp. 91–115.

[2] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. SAPA*, Sep. 2008, pp. 23–28.

[3] J. Le Roux, H. Kameoka, E. Vincent, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF under spectrogram consistency constraints," in *Proc. ASJ Autumn Meeting*, no. 2-4-5, Sep. 2009.

[4] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. ICASSP*, Apr. 2009, pp. 3437–3440.

[5] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. ASSP*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[6] P. L. Combettes and J.-C. Pesquet, "A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery," *IEEE J. STSP*, vol. 1, no. 4, pp. 564–574, Dec. 2007.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. NIPS*2000*. The MIT Press, 2001, pp. 556–562.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. 27, pp. 113–120, Apr. 1979.

[9] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast phase estimation algorithms based on spectrogram consistency," in *Proc. ASJ Spring Meeting*, no. 3-5-2, Mar. 2010.