FAST SIGNAL RECONSTRUCTION FROM MAGNITUDE SPECTROGRAM OF CONTINUOUS WAVELET TRANSFORM BASED ON SPECTROGRAM CONSISTENCY

Tomohiko Nakamura[†] *and Hirokazu Kameoka*[†‡]

[†]Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
[‡]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation,
3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198, Japan
nakamura@hil.t.u-tokyo.ac.jp, kameoka@hil.t.u-tokyo.ac.jp

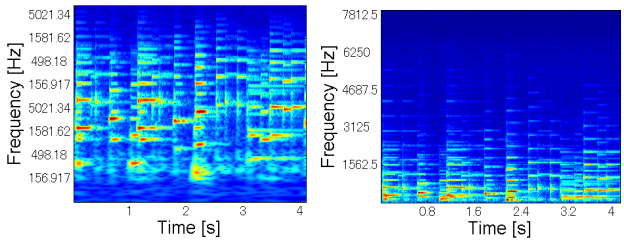
ABSTRACT

The continuous wavelet transform (CWT) can be seen as a filterbank having logarithmic frequency subbands spacing similar to the human auditory system. Thus, to make computers imitate the significant functions of the human auditory system, one promising approach would be to model, analyze and process magnitude spectrograms given by the CWT. To realize this approach, we must be able to convert a processed or modified magnitude CWT spectrogram, which contains no information about the phase, into a time domain signal specifically for those applications in which the aim is to generate audio signals. To this end, this paper proposes a fast algorithm for estimating the phase from a given magnitude CWT spectrogram to reconstruct an audio signal. The experimental results revealed that the proposed algorithm was around 100 times faster than a conventional algorithm, while the reconstructed signals obtained with the proposed algorithm had almost the same audio quality as those obtained with the previous study.

1. INTRODUCTION

The continuous wavelet transform (CWT), also known as the constant-Q transform, is used as a method for time-frequency analysis, which provides a time-frequency representation of a signal with an equal resolution on a log-frequency scale (Fig. 1). The human auditory filterbank is known to have an equal resolution on a log-frequency scale as with the CWT particularly in a high frequency band [1, 2]. Thus, to let computers imitate the significant functions of the human auditory system, one promising approach would be to model, analyze and process spectrograms obtained by the CWT (*CWT spectrogram*). In fact, recent studies (see [3–6]) have shown that multiple fundamental frequency estimation performs very well in the magnitude CWT spectrogram domain. Motivated by this fact, we believe that source separation and sound manipulation can also work well in the magnitude CWT spectrogram domain. However, in order to achieve source separation or sound manipulation, in which the goal is to produce sound, there is a need to reconstruct an appropriate time-domain signal after processing and modifying a magnitude CWT spectrogram. To this end, this paper proposes a method for estimating the phase from a given magnitude CWT spectrogram to reconstruct an audio signal.

The phase estimation algorithm from a magnitude CWT spectrogram has already been proposed by Irino *et al.* [7]. Irino's algorithm consists in iteratively performing the inverse CWT and the CWT followed by replacing the modified magnitude CWT spectrogram with a given magnitude CWT spectrogram. Since the computational speed of the CWT is much slower than the short-time Fourier transform (STFT), this algorithm needs a very long time for computation. In practical situations, the reduction of the computational complexity can be extremely important.



(a) CWT spectrogram. (b) STFT spectrogram.

Figure 1: Examples of the continuous wavelet transform (CWT) and short-time Fourier transform (STFT) spectrograms. While STFT spectrograms have an equal resolution on a linear frequency scale, CWT spectrograms have an equal resolution on a log-frequency scale.

The authors and colleagues have thus far proposed a fast method for estimating the phase from a magnitude STFT spectrogram [8]. When the hop-size is shorter than the frame length, the waveforms in the overlapping segment of consecutive frames must be consistent. This implies the fact that an STFT spectrogram is a redundant representation. Thus, an STFT spectrogram must satisfy a certain condition to ensure that it is associated with a time domain signal. We have referred to this condition as *the consistency condition*. In [8], we have shown that the problem of estimating the phase from a magnitude STFT spectrogram can be formulated as the problem of optimizing the consistency criterion describing how far an arbitrary complex array deviates from this condition.

It became clear that the deviced algorithm is equivalent to the well-known algorithm proposed by Griffin *et al.*, [9]. The formulation derived from the concept of the spectrogram consistency has provided a new insight into the Griffin's algorithm, allowing us to introduce a fast approximate algorithm and give a very intuitive proof of the convergence of the algorithm. Since a CWT spectrogram is also a redundant representation of a signal [10], we may be able to make the best use of the spectrogram consistency concept to develop a fast approximate method for phase estimation from a magnitude CWT spectrogram.

Following the idea proposed in [8], this paper derives an algo-

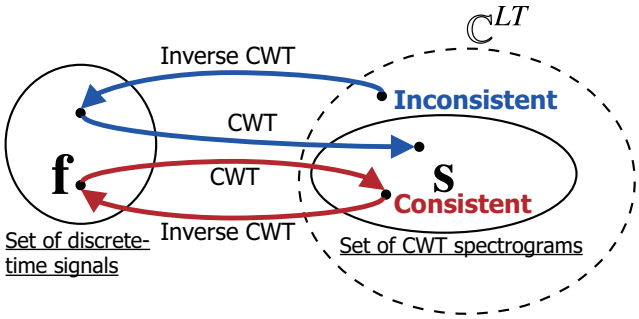


Figure 2: Illustration of the spectrogram consistency concept for the continuous wavelet transform (CWT).

rithm for estimating the phase from a magnitude CWT spectrogram. Sec. 2 formulates the phase estimation problem as an optimization problem based on a consistency condition. Sec. 3 derives an iterative algorithm for phase estimation based on an auxiliary function approach, which turns out to be equivalent to the algorithm proposed by Irino [7]. Our formulation gives a very clear proof of the convergence of the algorithm, though it should be noted that the proof of the convergence has already been given in [10]. Sec. 4 describes a fast approximate method for computing each iterative step of the proposed algorithm.

2. CWT SPECTROGRAM CONSISTENCY

2.1. Consistency condition

The scale parameter of the CWT corresponds to the period (the reciprocal of the center frequency) of the wavelet basis function. Here we consider discretizing the scale parameter such that the center frequencies of the wavelet basis functions are uniformly spaced on a log-frequency scale. Let the indices of the scale parameter and time shift parameter be denoted by $l \in [0, L-1]$ and $t \in [0, T-1]$, respectively, and let the component of a CWT spectrogram associated with the l-th scale parameter $a_l > 0$ (hereafter, the l-th component) be denoted by $\boldsymbol{s}_l = [s_{l,0}, s_{l,1}, \dots, s_{l,T-1}]^\mathsf{T} \in \mathbb{C}^T$. Given a discrete-time signal $\boldsymbol{f} = [f_0, f_1, \dots, f_{T-1}]^\mathsf{T} \in \mathcal{F}$ where $\mathcal{F} := \{\boldsymbol{f}' \in \mathbb{C}^T; \sum_t f'_t = 0\}$, its CWT spectrogram $\boldsymbol{s} = [\boldsymbol{s}_0^\mathsf{T}, \boldsymbol{s}_1^\mathsf{T}, \dots, \boldsymbol{s}_{L-1}^\mathsf{T}]^\mathsf{T} \in \mathbb{C}^{LT}$ is defined as

$$\boldsymbol{s} = W\boldsymbol{f}, \tag{1}$$

where $W \in \mathbb{C}^{LT \times T}$ denotes the CWT matrix, defined as

$$W := \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{L-1} \end{bmatrix}, \quad W_l := \begin{bmatrix} \psi_{l,0} & \psi_{l,1} & \cdots & \psi_{l,T-1} \\ \psi_{l,T-1} & \psi_{l,0} & \cdots & \psi_{l,T-2} \\ \vdots & & \ddots & \vdots \\ \psi_{l,1} & \psi_{l,2} & \cdots & \psi_{l,0} \end{bmatrix}. \tag{2}$$

Here, $\psi_{l,t} := \psi(t\Delta/a_l)/a_l$ is a scaled mother wavelet with the scale of a_l and the time shift of $t\Delta$, where Δ denotes the sampling period of the time shift parameter and $\psi(\cdot) \in \mathbb{C}$ denotes the mother wavelet satisfying the admissibility condition. Each row of $W_l \in \mathbb{C}^{T \times T}$ contains the wavelet basis function of scale a_l with a different time shift parameter. The inverse CWT can be defined by the pseudo-inverse of W:

$$\boldsymbol{f} = W^+\boldsymbol{s}, \quad W^+ := (W^\mathsf{H}W)^{-1}W^\mathsf{H}, \tag{3}$$

where $^\mathsf{H}$ is used to denote the Hermitian transpose. This implicitly means that the inverse CWT is defined as the solution to the following minimization problem:

$$W^+\boldsymbol{s} = \operatorname*{argmin}_{\tilde{\boldsymbol{f}} \in \mathcal{F}} \|\boldsymbol{s} - W\tilde{\boldsymbol{f}}\|_2^2, \tag{4}$$

where $\|\cdot\|_2$ denotes the ℓ^2 norm of a vector.

While the CWT spectrogram of an audio signal (i.e., a complex vector that belongs to the subspace spanned by the column vectors of W) will be mapped to itself by applying the inverse CWT followed by the CWT, a complex vector that does not belong to the subspace will not come back to the same point but will be projected onto the nearest point in the subspace. Thus, we can define a condition for a complex vector to be "consistent" (in the sense that it corresponds to a CWT spectrogram of a signal) as follows:

$$\boldsymbol{0}_{LT} = \boldsymbol{s} - WW^+\boldsymbol{s}, \tag{5}$$

where $\boldsymbol{0}_{LT}$ denotes an LT-dimensional zero vector. It is important to note that when W is replaced with a matrix in which each row is a basis function of the STFT, (5) becomes equivalent to the consistency condition for an STFT spectrogram proposed in [8].

2.2. Phase estimation using spectrogram consistency

When given a magnitude CWT spectrogram $\boldsymbol{a} \in [0, \infty)^{LT}$, we can construct a signal by assigning phase $\boldsymbol{\phi} \in [-\pi, \pi)^{LT}$ to it to obtain a complex spectrogram $\boldsymbol{s}$, and applying the inverse CWT, i.e., $W^+\boldsymbol{s}$. Here, if we assign "inconsistent" phase to the given magnitude spectrogram, the complex spectrogram $\boldsymbol{s}$ will not belong to the signal subspace and so the spectrogram of the constructed signal, $WW^+\boldsymbol{s}$, will be different from $\boldsymbol{s}$. As we want to keep the magnitude spectrogram of the constructed signal consistent with the given magnitude spectrogram, we must find "consistent" phase such that $\boldsymbol{s}$ satisfies the consistency condition.

2.3. Filter bank interpretation

To give a deeper insight into the consistency condition, we focus on the filter bank interpretation of the CWT. The CWT of a signal can be thought of as the output of a filter bank consisting of subband filters whose impulse responses are given by the the scaled mother wavelets. Now, by applying the T-point discrete Fourier transform (DFT) to each block of (5), (5) can be written equivalently as

$$\boldsymbol{0} = \hat{\boldsymbol{s}} - \hat{W}\hat{W}^+\hat{\boldsymbol{s}}, \tag{6}$$

where

$$\hat{W} = \begin{bmatrix} \hat{W}_0 \\ \hat{W}_1 \\ \vdots \\ \hat{W}_{L-1} \end{bmatrix}, \quad \hat{W}_l = F_T W_l F_T^\mathsf{H}, \quad \hat{W}^+ = (\hat{W}^\mathsf{H}\hat{W})^{-1}\hat{W}^\mathsf{H}, \tag{7}$$

$F_T \in \mathbb{C}^{T \times T}$ is the DFT matrix and $\hat{\cdot}$ denotes the DFT of a variable. Since W_l is a circulant matrix, W_l is diagonalized by F_T and F_T^H. The diagonal elements of $\hat{W}_l$ represent the frequency response of the l-th subband filter associated with the scale parameter a_l. The k-th diagonal element of (6) is explicitly written as

$$0 = \hat{s}_{l,k} - \frac{1}{C_k} \sum_{l'} \hat{\psi}_{l,k}\hat{\psi}^*_{l',k}\hat{s}_{l',k}, \tag{8}$$

where $k \in [0, T-1]$ denotes the angular frequency index, C_k is a normalization constant, and * is used to denote the complex conjugate.

If the subbands of the filter bank overlap each other (more precisely, if there exists a pair of channels such that the product of their frequency responses is non-zero at every frequency), i.e. $\forall k, \exists l \neq l', \hat{\psi}_{l,k}\hat{\psi}_{l',k} \neq 0$, (5) becomes a nontrivial condition for a complex vector $\boldsymbol{s} \in \mathbb{C}^{LT}$ to correspond to a consistent CWT spectrogram. Otherwise, all the elements of $\mathbb{C}^{LT}$ trivially satisfy (5), implying that the consistency condition cannot be used as a criterion for phase estimation. Therefore, care must be taken in choosing the quantization intervals of the scale parameter and the type of the mother wavelet function. The Morlet [11], the log-normal wavelet [4] and the wavelets used in the auditory wavelet transform [7] satisfy the above requirement when the quantization intervals of the scale parameter are appropriately chosen. We hereafter assume to use a filter bank that satisfies $\forall k, \exists l \neq l', \hat{\psi}_{l,k}\hat{\psi}_{l',k} \neq 0$.

The requirement for the subbands of the CWT to overlap each other is analogous to the requirement for the short time frames of the STFT to overlap. The consistency condition of STFT spectrograms can be understood as implying that the waveforms within the overlapping segment of consecutive frames must be consistent [8]. The consistency condition of CWT spectrograms, on the other hand, can be interpreted as implying that the outputs of adjacent channels within the overlapping subbands must be consistent.

3. PHASE ESTIMATION BASED ON CWT SPECTROGRAM CONSISTENCY

3.1. Formulation of phase estimation problem

Assume that we are given a magnitude CWT spectrogram, arranged as a non-negative vector $\boldsymbol{a} \in [0, \infty)^{LT}$. We would like to estimate the phase of the given magnitude spectrogram such that it meets the consistency condition. To allow for any vector $\boldsymbol{a} \in [0, \infty)^{LT}$ as the input, we consider finding a phase estimate $\boldsymbol{\phi} \in [-\pi, \pi)^{LT}$ that minimizes the consistency criterion

$$\mathcal{I}(\boldsymbol{\phi}) := \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - WW^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})\|_2^2, \tag{9}$$

where $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ denotes the estimated CWT spectrogram defined by

$$\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) := \boldsymbol{a} \odot \begin{bmatrix} e^{\mathrm{j}\phi_0} \\ e^{\mathrm{j}\phi_1} \\ \vdots \\ e^{\mathrm{j}\phi_{LT-1}} \end{bmatrix}. \tag{10}$$

$\odot$ denotes the element-wise product. $\mathcal{I}(\boldsymbol{\phi})$ describes how far $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ deviates from the consistency condition. Namely, the more consistent $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ becomes, the smaller $\mathcal{I}(\boldsymbol{\phi})$ becomes.

3.2. Iterative algorithm with auxiliary function approach

Unfortunately, the optimization problem of minimizing $\mathcal{I}(\boldsymbol{\phi})$ with respect to $\boldsymbol{\phi}$ is difficult to solve analytically. However, we can invoke the auxiliary function approach to derive an iterative algorithm that searches for the estimate of $\boldsymbol{\phi}$, as with [8]. To apply the auxiliary function approach to the current optimization problem, the first step is to construct an auxiliary function $\mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ satisfying $\mathcal{I}(\boldsymbol{\phi}) = \min_{\tilde{\boldsymbol{s}}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. We refer to $\tilde{\boldsymbol{s}}$ as an auxiliary variable. It can then be shown that $\mathcal{I}(\boldsymbol{\phi})$ is non-increasing under the updates $\boldsymbol{\phi} \leftarrow \mathrm{argmin}_{\boldsymbol{\phi}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ and $\tilde{\boldsymbol{s}} \leftarrow \mathrm{argmin}_{\tilde{\boldsymbol{s}}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. The proof of this shall be omitted owing to space limitations. Thus, $\mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ should be designed as a function that can be minimized analytically with respect to $\boldsymbol{\phi}$ and $\tilde{\boldsymbol{s}}$. Such a function can be constructed as follows.

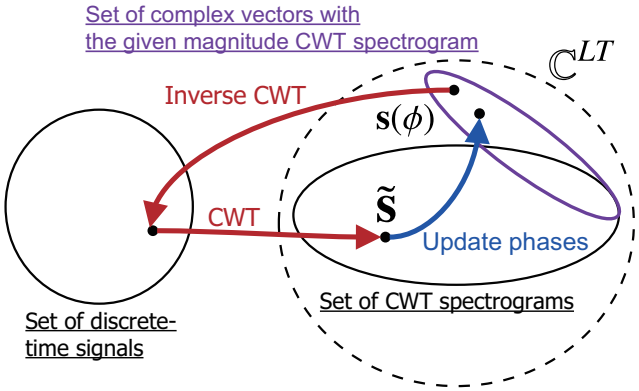


Figure 3: Illustration of the iterative phase estimation algorithm. The red and blue arrows correspond to (14) and (15).

Recall that the operator WW^+ is an orthogonal projection onto the subspace spanned by the column vectors of W and so $WW^+\boldsymbol{s}$ indicates the closest point in the subspace from $\boldsymbol{s}$. Thus, we can show that

$$\mathcal{I}(\boldsymbol{\phi}) = \min_{\tilde{\boldsymbol{f}} \in \mathcal{F}} \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - W\tilde{\boldsymbol{f}}\|_2^2 \tag{11}$$

$$= \min_{\tilde{\boldsymbol{s}} \in \mathcal{W}} \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - \tilde{\boldsymbol{s}}\|_2^2, \tag{12}$$

where $\mathcal{W}$ denotes the set of consistent CWT spectrograms (the subspace spanned by the column vectors of W). Therefore, we can confirm that

$$\mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}}) := \|\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}) - \tilde{\boldsymbol{s}}\|_2^2, \quad \tilde{\boldsymbol{s}} \in \mathcal{W}, \tag{13}$$

satisfies $\mathcal{I}(\boldsymbol{\phi}) = \min_{\tilde{\boldsymbol{s}} \in \mathcal{W}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. (13) can thus be used as an auxiliary function for $\mathcal{I}(\boldsymbol{\phi})$. We can thus monotonically decrease $\mathcal{I}(\boldsymbol{\phi})$ by iteratively performing $\tilde{\boldsymbol{s}} \leftarrow \mathrm{argmin}_{\tilde{\boldsymbol{s}}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ and $\boldsymbol{\phi} \leftarrow \mathrm{argmin}_{\boldsymbol{\phi}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$. Here, $\tilde{\boldsymbol{s}} \leftarrow \mathrm{argmin}_{\tilde{\boldsymbol{s}}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ and $\boldsymbol{\phi} \leftarrow \mathrm{argmin}_{\boldsymbol{\phi}} \mathcal{I}^+(\boldsymbol{\phi}, \tilde{\boldsymbol{s}})$ can be written explicitly as

$$\tilde{\boldsymbol{s}} \leftarrow WW^+\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi}), \tag{14}$$

$$\boldsymbol{\phi} \leftarrow \angle\tilde{\boldsymbol{s}}, \tag{15}$$

respectively, where $\angle$ denotes an operator that gives the arguments of the components of a complex vector as a real vector in $[-\pi, \pi)^{LT}$.

(14) means applying the inverse CWT followed by the CWT to $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$. Here, when $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ is already a complex vector corresponding to a consistent spectrogram, this update simply becomes $\tilde{\boldsymbol{s}} \leftarrow \boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$. (15) means replacing the phase estimate $\boldsymbol{\phi}$ with the phase of $\tilde{\boldsymbol{s}}$. A schematic illustration of these updates is shown in Fig. 3. $\mathcal{I}(\boldsymbol{\phi}) = 0$ indicates that $\boldsymbol{s}(\boldsymbol{a}, \boldsymbol{\phi})$ lies in the intersection of the set of consistent CWT spectrograms and the set of complex vectors that are equal to $\boldsymbol{a}$ up to a phase factor.

3.3. Relation to previous work

The present algorithm is equivalent to an algorithm proposed by Irino [7]. In addition, when W is replaced with a matrix in which each row is a basis function of the STFT, the present algorithm becomes equivalent to the phase estimation algorithm for a magnitude STFT spectrogram proposed in [8].

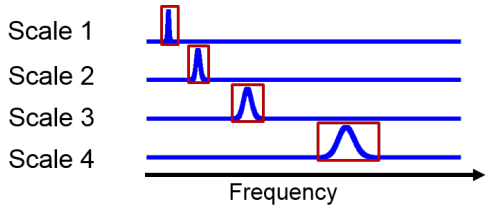


Figure 4: Example of the frequency responses of different subband filters (i.e., the scaled mother wavelets). The mother wavelet is the log-normal wavelet [4].

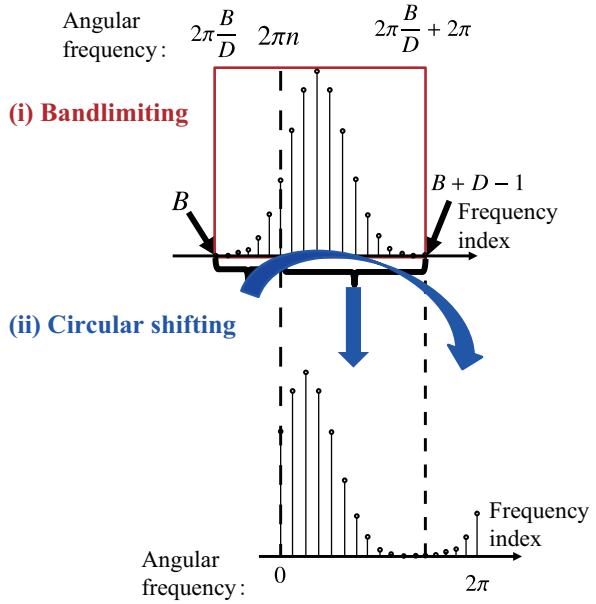


Figure 5: A circularly shifted version of $G_{l,B},\ldots,G_{l,B+D-1}$ [12].

4. FAST PHASE ESTIMATION ALGORITHM

4.1. Fast approximate continuous wavelet transform

The CWT and the inverse CWT are computationally expensive compared to the STFT and the inverse STFT. Here we briefly describe the fast approximate method for computing CWT proposed in [12]. The proposed fast approximate CWT uses the fact that the dominant part of the frequency response of each subband filter is concentrated around its center frequency (as shown in Fig. 4), as is common in many types of mother wavelets including the Morlet and log-normal wavelets [4].

According to the filter bank interpretation of the CWT, the CWT of an input signal, $\boldsymbol{s}_l = [s_{l,0},\ldots,s_{l,T-1}]^{\mathsf{T}} = W_l\boldsymbol{f}$, can be computed by multiplying the DFT of the entire signal, i.e, $\hat{\boldsymbol{f}} = [\hat{f}_0,\ldots,\hat{f}_{T-1}]^{\mathsf{T}} = F_T\boldsymbol{f}$, by the frequency response of the l-th subband, i.e., $\hat{W}_l = \mathrm{diag}(\hat{\psi}_{l,0},\ldots,\hat{\psi}_{l,T-1})$, and then computing the inverse DFT of $\hat{W}_l\hat{\boldsymbol{f}}$. This can be confirmed from

$$\boldsymbol{s}_l = W_l\boldsymbol{f} = F_T^{\mathsf{H}}F_TW_lF_T^{\mathsf{H}}F_T\boldsymbol{f} = F_T^{\mathsf{H}}\hat{W}_l\hat{\boldsymbol{f}}. \tag{16}$$

Note that the second equality follows from the fact that the DFT matrix F_T is a unitary matrix, i.e., $F_T^{\mathsf{H}}F_T = \boldsymbol{I}_T$. Here, if we can assume that the elements of $\{\hat{\psi}_{l,k}\}_k$ are dominant within and near 0 outside the range $k\in[B,B+D-1]$ $(0\le B, 0<D\le T)$, we can approximate $\boldsymbol{s}_l$ reasonably well by using the elements of $\{\hat{\psi}_{l,k}\hat{f}_k\}_k$ only within that range and neglecting the remaining elements. This implies the possibility of computing an approximation of $\boldsymbol{s}_l$ with a lower computational cost.

For simplicity of notation, let us put $G_{l,k} = \hat{\psi}_{l,k}\hat{f}_k$. We are concerned with computing an approximation of the full-band inverse DFT of $G_{l,k}$:

$$s_{l,t} = \sum_{k=0}^{T-1} G_{l,k}e^{j\frac{2\pi kt}{T}}. \tag{17}$$

As mentioned above, $G_{l,0},\ldots,G_{l,T-1}$ can be approximately viewed as a band-limited spectrum. In general, the inverse DFT of a band-limited spectrum can be computed by taking the inverse DFT over the finite support. In the time domain, this process corresponds to downsampling the signal given by the "full-band" inverse DFT. The proposed method uses this idea to approximate the inverse DFT of the full-band spectrum $G_{l,0},\ldots,G_{l,T-1}$. Now, if we choose D such that T/D becomes an integer, we can approximate the downsampled version of $s_{l,t}$ by

$$\tilde{s}_{l,d} = \sum_{k=B}^{B+D-1} G_{l,k}e^{j\frac{2\pi kd}{D}} = \sum_{k=B}^{B+D-1} G_{l,k}e^{j\frac{2\pi k(T/D)d}{T}}. \tag{18}$$

By comparing (17) and (18), we can confirm that

$$s_{l,(T/D)d} \simeq \tilde{s}_{l,d} \quad (d\in[0,D-1]), \tag{19}$$

if we assume $G_{l,k}\simeq 0$ outside the range $k\in[B,B+D-1]$. Since $\tilde{s}_{l,d}$ can be rewritten as

$$\tilde{s}_{l,d} = \sum_{k=0}^{D-1} G_{l,k+B}e^{j(\frac{2\pi k}{D}+2\pi\frac{B}{D})d} = e^{j2\pi\frac{B}{D}d}\sum_{k=0}^{D-1} G_{l,k+B}e^{j\frac{2\pi kd}{D}}, \tag{20}$$

we notice that $\tilde{s}_{l,d}$ can be computed by multiplying the inverse DFT of $G_{l,B},\ldots,G_{l,B+D-1}$ by $e^{j2\pi\frac{B}{D}d}$. Note that this is equivalent to computing the inverse DFT of a circularly shifted version of $G_{l,k}$ (see Fig. 5):

$$\tilde{G}_{l,k} = \begin{cases} G_{l,k+nD} & (k=0,\ldots,B-(n-1)D-1) \\ G_{l,k+(n-1)D} & (k=B-(n-1)D,\ldots,D-1) \end{cases}, \tag{21}$$

where n is an integer such that

$$n-1 < \frac{B}{D} \le n. \tag{22}$$

We consider invoking the fast Fourier transform (FFT) algorithm for computing the inverse DFT and so we assume the size D to be a power of 2. Since $D<T$, the computational cost for computing $\tilde{s}_{l,0},\ldots,\tilde{s}_{l,D-1}$ is obviously lower than that for computing $s_{l,0},\ldots,s_{l,T-1}$.

4.2. Fast phase estimation algorithm

The processes of bandlimiting and circular shifting can be represented by a matrix K:

$$K := \underbrace{\begin{bmatrix} 0_{B_0\times(D-B_0)} & I_{B_0} \\ I_{D-B_0} & 0_{(D-B_0)\times B_0} \end{bmatrix}}_{\text{(ii) Circular shifting}} \underbrace{\begin{bmatrix} 0_{D\times B} & I_D & 0_{D\times(T-D-B)} \end{bmatrix}}_{\text{(i) Bandlimiting}} \tag{23}$$

where I_D and $0_{D\times B}$ are the $D \times D$ identity matrix and the $D \times B$ zero matrix. The downsampled version of $\boldsymbol{s}_l$ obtained with the abovementioned fast approximate CWT can be described as

$$\check{\boldsymbol{s}}_l = F_D^{\mathsf{H}} K \hat{W}_l F_T \boldsymbol{f}. \tag{24}$$

Similarly to the inverse CWT, the fast approximate version of the inverse CWT can be defined by the pseudo-inverse matrix of $F_D^{\mathsf{H}} K \hat{W}_l F_T$. It is important to note that convergence of the phase estimation algorithm in which the CWT and inverse CWT steps are replaced with the fast approximate versions is still guaranteed.

4.3. Time and space complexity

The computational costs for the CWT and the fast approximate CWT mainly depend on the number of the points for the inverse DFT. Since the computational complexity of the full band inverse DFT is $O(T \log_2 T)$, the total computational complexity of the CWT is $O(T \log_2 T + LT \log_2 T)$. By contrast, the computational complexity of the band-limited DFT is $O(D \log_2 D)$ and so the total computational complexity is $O(T \log_2 T + \sum_{l=0}^{L-1} D_l \log_2 D_l)$.

The space complexity of the proposed algorithm is small compared to Irino's algorithm [7]. When the signal length T is long enough, the space complexity depends primarily on the size of the CWT spectrogram. While the size of the CWT spectrogram of Irino's algorithm is LT, that of the proposed algorithm is only $\sum_l D_l$.

5. EXPERIMENTAL EVALUATIONS

5.1. First experiment: Computation time and audio quality

5.1.1. Experimental conditions

To evaluate the computation time and the audio quality of the reconstructed signals by the phase estimation algorithms, we conducted an objective experiment, and compared the proposed algorithm with the Irino's algorithm [7].

We used the magnitude CWT spectrograms of 16kHz-sampled acoustic signals of the 113 male and 115 female speeches in the ATR Japanese speech database A-set [13]. The FFT performs faster for acoustic signals with a power of 2 length than those with the other length, and the used signals were filled by 0 till each length reached a power of 2. Phases were initialized randomly, and both the algorithms were finished at 1000 iterations. As the mother wavelet, we used the log-normal wavelet [4], which is defined in the Fourier-transformed domain:

$$\hat{\psi}(\omega) := \begin{cases} \exp\left(-\dfrac{(\log \omega)^2}{4\sigma^2}\right) & (\omega > 0) \\ 0 & (\omega \geq 0) \end{cases} \tag{25}$$

where ω is an angular frequency and σ is a standard deviation. σ was set at 0.02 and the analysis frequencies ranged 27.5 to 7040 Hz with 20 cents interval (i.e. uniform interval in the log-frequency domain). In the proposed algorithm, we computed the elements within $\pm 3\sigma$ around the central frequencies in the log-frequency domain. The used computer had the Intel Xeon CPU E31245 (3.3 GHz) and a 32 GB RAM.

We employed the perceptual evaluation of speech quality (PESQ) [14] as the evaluation measure for audio quality, which is the world-standard objective evaluation measure for speech quality. It ranges −0.5 to 4.5 and speech quality is higher as the PESQ becomes larger. As an evaluation measure of the computation speed, the computation time per iteration was used.

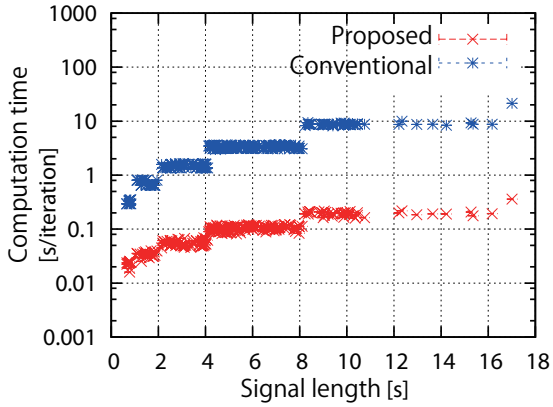


Figure 6: The averaged computation time per iteration with standard errors with respect to various signal lengths.

5.1.2. Results

The averaged PESQ with standard errors were 4.20 ± 0.08 for the Irino's algorithm and 4.1 ± 0.1 for the proposed algorithm. The result indicates that the speech qualities of the reconstructed signals were high enough for practical use. The difference between the Irino's and proposed algorithms was negligible practically[1].

Fig. 6 shows the results for the computation speed with respect to the signal length, since the computational complexity of the algorithms primarily depends on the signal length. The proposed algorithm was around 100 times faster than the Irino's algorithm in the computation time. For example, the averaged computation time per iteration by the Irino's algorithm was around 10 s/iteration for the 15 s signal. In contrast, that by the proposed algorithm was around 0.1 s/iteration.

5.2. Second experiment: Relation between approximation accuracy and audio quality

5.2.1. Experimental conditions

The proposed algorithm includes the approximation, and we next evaluated the relation between the approximation accuracy and the audio quality of the reconstructed signals. We used the 5 s from 30 s of 102 music audio files with 16 kHz sampling frequency in the RWC music genre database [15]. As the mother wavelet, the log-normal wavelet with $\sigma = 0.02$ was chosen. The approximation accuracy of the proposed algorithm corresponds to the calculated range by the downsampling step, and we used the elements within $\pm P\sigma$ ($P = 1, 2, 3, 5$) around the central frequencies in the log-frequency domain. The number of iterations was set at 500 for the proposed algorithm and at 100 for the Irino's algorithm. The used computer had the Intel Core i3-2120 CPU (3.30 GHz) and a 8 GB RAM. The other experimental conditions were the same as in Sec. 5.1.1.

An evaluation measure for audio quality was the objective differential grade (ODG) by the perceptual evaluation of audio qual-

[1]Audio samples are available at `http://hil.t.u-tokyo.ac.jp/~nakamura/demo/fastCWT.html`.

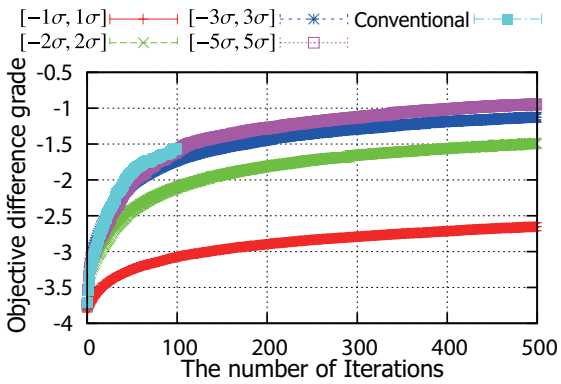


Figure 7: Evolution of the averaged objective difference grades with standard errors by perceptual evaluation of audio quality with respect to the number of iterations for the proposed algorithms with various approximations ($[-P\sigma, P\sigma]$ ($P = 1, 2, 3, 5$)) and the Irino's algorithm [7].

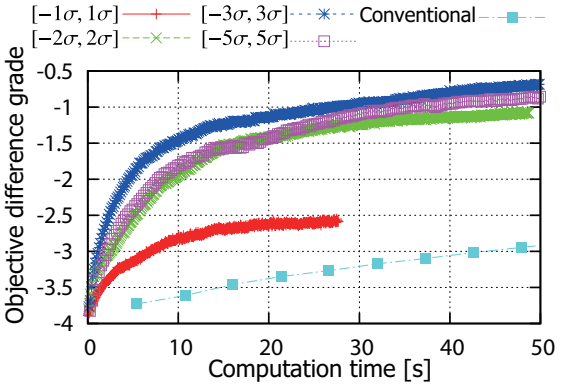


Figure 8: Evolution of the objective difference grades by perceptual evaluation of audio quality with respect to the computation time for the proposed algorithms with various approximations ($[-P\sigma, P\sigma]$ ($P = 1, 2, 3, 5$)) and the Irino's algorithm [7].

ity (PEAQ) [16]. It ranges −4 to 0, and the acoustic quality is higher as the ODG becomes larger.

5.2.2. *Results*

Fig. 7 illustrates the averaged ODGs with standard errors. The ODGs by the proposed algorithms with $P = 3, 5$ were larger than −2.0 after 100 iterations, and the results for $P = 3, 5$ shows high audio quality[2]. The results does not significantly differ from that by the Irino's algorithm in audio quality. We can thus say that the proposed algorithm with around $P \geq 3$ reconstructs the acoustic signals with almost the same audio quality as the Irino's algorithm.

In a viewpoint of the computation speed, the computation time becomes shorter as P is smaller. Fig. 8 shows the result for one of the acoustic signals (RWC-MDB-G-2001 No. 1), and the ODGs by the proposed algorithms quickly become higher than those by the Irino's algorithm. The similar result for the speech signal (fafsc110 in the ATR Japanese speech database A-set [13], the 7 s signal) was shown in Fig. 9. Therefore, we conclude that the proposed algorithm with around $P = 3$ provides the reconstructed signals with high audio quality in a reasonable computation time.

[2] c.f.) When the used audio signals were converted into MPEG-3 files with 160 kbps, the averaged ODGs with standard errors were -3.68 ± 0.03.

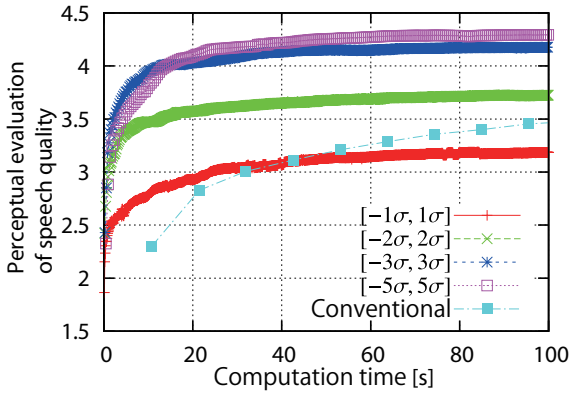


Figure 9: Evolution of the perceptual evaluation of speech quality and the computation time with respect to the proposed algorithms with various approximations ($[-P\sigma, P\sigma]$ ($P = 1, 2, 3, 5$)) and the Irino's algorithm [7].

5.3. Demonstration of phase estimation

We demonstrate pitch transposition of acoustic signals to confirm effectiveness of the proposed algorithm for sound manipulation. When the analysis frequencies are located uniformly in the log-frequency domain and $D_0 = D_1 = \cdots = D_{L-1}$ in the proposed algorithm, we simply shift the components of the CWT spectrograms to the lower or higher analysis frequency components, and the blank components by the move are filled by zero. However, the shifts cause the mismatches of phases, and the use of the original and zero phases leads to failure of the pitch transposition, hence we need to use the phase estimation for synthesizing the pitch-transposed acoustic signals. By the proposed algorithm, we obtained the synthesized signals [3] as we expected.

6. CONCLUSION

We have proposed a fast and convergence-guaranteed algorithm of the phase estimation by using the fast approximate CWT [12]. The phase estimation problem has been formulated based on the consistency condition, and the iterative algorithm has been derived by applying the auxiliary function method, which is the same as the Irino's algorithm [7]. Furthermore, we show the requirement on scale factors and mother wavelets for the phase estimation by using the consistency condition. The experimental results have shown that the proposed algorithm was about 100 times faster than the algorithm provided in [7]. The audio quality of the reconstructed signals for music and speech data was high enough for practical use, and the difference between the results by the proposed algorithm and the algorithm provided in [7] was negligible.

[3] The synthesized signals are available at `http://hil.t.u-tokyo.ac.jp/~nakamura/demo/fastCWT.html`.

We plan to combine the phase estimation with source separations for magnitude CWT spectrograms for music acoustic signal manipulation such as conversions of chords, keys and scales. To increase convenience, developing the online version of the proposed algorithm is important.

7. ACKNOWLEDGMENTS

This work was supported by JSPS Grant-in-Aid for Young Scientists B Grant Number 26730100.

8. REFERENCES

[1] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–61, 1940.

[2] R. D. Patterson, "Auditory filter shape," *J. Acoust. Soc. Am.*, vol. 55, no. 4, pp. 802–809, 2005.

[3] M. N. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Independent Component Analysis and Blind Signal Separation*, pp. 700–707. Springer, 2006.

[4] H. Kameoka, *Statistical Approach to Multipitch Analysis*, Ph.D. thesis, The University of Tokyo, Mar. 2007.

[5] M. Muller, D. P. W. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 6, pp. 1088–1110, 2011.

[6] J. P. de León, F. Beltrán, and J. R. Beltrán, "A complex wavelet based fundamental frequency estimator in single-channel polyphonic signals," in *Proc. Digital Audio Effects*, 2013.

[7] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3549–3554, 1993.

[8] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects*, Sep. 2010, pp. 397–403.

[9] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[10] D. M. Lopes and P. R. White, "Signal reconstruction from the magnitude or phase of a generalised wavelet transform," in *Proc. Eur. Signal Process. Conf.*, 2000, pp. 2029–2032.

[11] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *B. Am. Meteorol. Soc.*, vol. 79, no. 1, pp. 61–78, 1998.

[12] H. Kameoka, T. Tabaru, T. Nishimoto, and S. Sagayama, "(Patent) Signal processing method and unit," Nov. 2008, in Japanese.

[13] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Commun.*, vol. 9, no. 4, pp. 357–363, 1990.

[14] "ITU-T recommendation P.862, Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[15] M. Goto, "Development of the RWC Music Database," in *Proc. Int. Congress Acoust.*, 2004, pp. I–553–556.

[16] "ITU-T recommendation BS.1387-1, Perceptual evaluation of audio quality (PEAQ): Method for objective measurements of perceived audio quality," Sep. 2001.