Acoust. Sci. & Tech. **40**, 3 (2019)

INVITED REVIEW

Representation of complex spectrogram via phase conversion

Kohei Yatabe*, Yoshiki Masuyama†, Tsubasa Kusano‡ and Yasuhiro Oikawa§

Department of Intermedia Art and Science, Waseda University,
3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan

Abstract: As importance of the phase of complex spectrogram has been recognized widely, many techniques have been proposed for handling it. However, several definitions and terminologies for the same concept can be found in the literature, which has confused beginners. In this paper, two major definitions of the short-time Fourier transform and their phase conventions are summarized to alleviate such complication. A phase-aware signal-processing scheme based on phase conversion is also introduced with a set of executable MATLAB functions (https://doi.org/10/c3qb).

Keywords: Phase-aware signal processing, Short-time Fourier transform (STFT), Discrete Gabor transform (DGT), Phase conventions, Instantaneous frequency

PACS number: 43.60.Hj, 43.60.Pt [doi:10.1250/ast.40.170]

1. INTRODUCTION

Time-frequency analysis is an indispensable tool in audio signal processing. Among many including cosine and wavelet transform, the short-time Fourier transform (STFT), or discrete Gabor transform (DGT), is one of the most popular choices. A reason behind the popularity should be close connection to the well-understood Fourier transform whose magnitude is translation invariant (unchanged under translation in time domain). STFT inherits such invariance (or robustness) of the Fourier transform to some extent, which allows friendly design of a signal-processing algorithm acting on magnitude of STFT. The unfriendly part of STFT is concentrated into *phase* whose (seemingly) fancy structure had forced ones to abandon it.

Recently, as importance of phase has been recognized widely, phase-aware signal processing is receiving increasing attention from the society. A number of techniques has been introduced over the past few decades to manage information buried inside phase, and the latest concepts of the processing techniques are constructed upon them. However, several definitions and terminologeis for the same concept can be found in the literature, which has confused beginners.

In this paper, two major definitions of STFT (together with DGT) and their phase conventions are summarized. The instantaneous frequency and its calculation based on the window function are also reviewed. Then, a signal processing scheme based on phase conversion [1] is explained with some related topics. A set of MATLAB functions for its calculation is available and can be downloaded[1] as a supporting material.

2. STFT VS DGT

In the literature of audio signal processing, the term "STFT" is often utilized in place of what the literature of time-frequency analysis calls "DGT" (in honor of D. Gabor). In this section, their difference is contrasted by brief definitions and their inversion formulas.

2.1. Terminology in Time-frequency Analysis

For a fixed window function $w \neq 0$, STFT of a time-domain signal $x(t)$ with respect to $w(t)$ is defined as

$$X(f,t) = \int_{\mathbb{R}} x(\tau)\,\overline{w(\tau - t)\,e^{2\pi \mathbb{i} f\tau}}\,\mathrm{d}\tau, \tag{1}$$

where $\overline{z}$ is the complex conjugate of z, and $\mathbb{i} = \sqrt{-1}$ is the imaginary unit. In the case of a discrete signal $x[n]$ with index $n \in \mathbb{Z}$, the discrete STFT is given by

$$X[m,n] = \sum\nolimits_l x[l]\,\overline{w[l-n]\,e^{2\pi \mathbb{i} ml/L}}, \tag{2}$$

where L is the length (total number of elements) of the signal. Note that, in this definition, the window $w[n]$ is translated sample-by-sample, i.e., STFT results in the fully-redundant time-frequency representation.

Obviously, such full redundancy is not necessary for

*e-mail: k.yatabe@asagi.waseda.jp
†e-mail: mas-03151102@akane.waseda.jp
‡e-mail: tsubasa.k@suou.waseda.jp
§e-mail: yoikawa@waseda.jp

[1]Available at Code Ocean: https://doi.org/10/c3qb

representing a signal, and thus it is natural to consider reduction of the number of elements by downsampling. With step-size parameters $a, b > 0$ and countable indices $n, m \in \mathbb{Z}$, STFT in Eq. (1) can be downsampled as

$$X[m,n] = \int_{\mathbb{R}} x(\tau)\, \overline{w(\tau - an)\, e^{2\pi \mathrm{i} bm\tau}}\, \mathrm{d}\tau, \tag{3}$$

which is called the *Gabor transform*[2]. Its discrete counterpart with integer shifting steps $a, b \in \mathbb{N}$,

$$X[m,n] = \sum\nolimits_l x[l]\, \overline{w[l - an]\, e^{2\pi \mathrm{i} bml/L}}, \tag{4}$$

is the so-called DGT (discrete Gabor transform) [2–6].

The term "spectrogram" is also used differently in the literatures. In time-frequency analysis, the squared magnitude of STFT $|X(f,t)|^2$ is only referred to as spectrogram. In this paper, it is used in wider sense as in the acoustics literature: DGT coefficient $X[m,n]$ is called *complex spectrogram*, while $|X[m,n]|$ and $|X[m,n]|^2$ are called magnitude (or amplitude) and power spectrograms, respectively.

2.2. Inversion Formula for STFT

One particularly nice property of STFT in Eq. (1) is that it admits the following inversion formula:

$$x(t) = \frac{1}{\langle w, \gamma \rangle} \int_{\mathbb{R}^2} X_w(f,\tau)\, \gamma(t - \tau)\, e^{2\pi \mathrm{i} ft}\, \mathrm{d}f \mathrm{d}\tau, \tag{5}$$

where X_w is X with respect to window w, γ is a window function satisfying $\langle w, \gamma \rangle \neq 0$, and $\langle \cdot, \cdot \rangle$ is the standard inner product[3]. Similar inversion formula is also available for the discrete STFT in Eq. (2). Many signal processing methods rely on such inversion after modifying time-frequency-domain coefficients.

Unlike STFT, the Gabor transform and DGT do not admit such a simple inversion formula. In general, a pseudo-inverse of the transform is required to reconstruct the time-domain signal from its time-frequency representation. Fortunately, the pseudo-inverse of Gabor-type transformations is highly structured, which allows an inversion algorithm requiring much less computation.

2.3. Inversion and Related Topics on DGT

Let DGT be explained with more details for the finite-dimensional settings. When considering finite samples of a signal, its boundary is always a source of trouble. A mathematically sound treatment for the inconvenience is to impose the periodic boundary condition,

$$x[l + L] = x[l], \tag{6}$$

as usual (for, say, the discrete Fourier transform), which can be easily realized by the modulo operation. This treatment preserves important properties of essential operations including translation and convolution.

Some inconvenience arising from the periodicity is then eliminated by zero-padding[4]. For given $a, b \in \mathbb{N}$, the signal length L should be divisible so that there exist $N, M \in \mathbb{N}$ satisfying $aN = bM = L$. This requirement can be fulfilled simply[5] by increasing the length L via zero-padding until L becomes divisible with both a and b. With these settings, DGT is defined as

$$X[m,n] = \sum_{l=0}^{L-1} x[l]\, \overline{w[l - an]\, e^{2\pi \mathrm{i} bml/L}}, \tag{7}$$

where $n \in \{0, \dots, N-1\}$, $m \in \{0, \dots, M-1\}$, and w zero-padded for extending its length to L is also treated periodically (index is read as $[l - an] \bmod L$).

Although inversion of DGT is not simple as STFT owing to its reduced redundancy, structure of the pseudo-inverse of DGT admits a similar inversion formula:

$$x[l] = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} X_w[m,n]\, \gamma[l - an]\, e^{2\pi \mathrm{i} bml/L}, \tag{8}$$

where γ is a *dual window* of w. While STFT can be inverted with almost any synthesis window, DGT can reconstruct the signal only when some appropriate synthesis window γ with respect to the analysis window w is utilized[6]. All procedures for calculating an appropriate synthesis window and the above DGT can be found in the supplemental MATLAB codes (see the footnote in Sect. 1 for the download URL).

When DGT is redundant, there exist infinitely many dual windows γ corresponding to a single window w. Therefore, many attempts to design a dual window with some desirable properties have been made [8–11]. A window is called *tight* when itself is proportional to its dual window, and its useful properties in signal processing have been studied [11–14]. These theories regarding the Gabor transform are based on *frame theory* [15] which also includes wavelet and general filterbanks [2–6].

3. PHASE AND ITS DERIVATIVE

The complex argument of a time-frequency representation is called *phase*, while its (time) derivative is referred to as *instantaneous frequency*. Their definitions can vary

[2]The special case of Eq. (3) utilizing the Gaussian window and $a = b = 1$ was proposed by D. Gabor in 1946.

[3]More precisely, $x, w, \gamma \in L^2(\mathbb{R})$, and $\langle \cdot, \cdot \rangle$ is the L^2 standard inner product. See, for example, [7] for rigorous statements.

[4]The side effect of the periodic boundary condition appears only within the window function including the end point. Therefore, it can be eliminated by sufficiently large zero-padding when the window is compactly supported.

[5]Denoting the least common multiple of a and b as c, such L can be found by $L = \lceil \tilde{L}/c \rceil c$, where $\tilde{L}$ is the signal length before zero-padding, and $\lceil \cdot \rceil$ is the ceiling function.

[6]Similar situations also arise in the continuous setting in Eq. (3).

according to the definition of the corresponding transform, which is briefly summarized in this section.

3.1. Another Definition of STFT

The definitions in Sect. 2 are merely examples, and some other definitions are widely accepted. STFT in Eq. (1) handles the signal $x(\tau)$ and complex sinusoid $e^{-2\pi \mathbb{i} f\tau}$ through the same variable τ, and only the window function $w(\tau - t)$ is translated. That is, the signal and sinusoid share the time origin ($t = 0$), while that of the window is independent. Alternatively, the window can share the time origin with the sinusoid, which results in another definition of STFT:[7]

$$\begin{aligned} X(f,t) &= \int_{\mathbb{R}} x(\tau)\,\overline{w(\tau - t)\,e^{2\pi \mathbb{i} f(\tau - t)}}\,\mathrm{d}\tau, \\ &= \int_{\mathbb{R}} x(\tau + t)\,\overline{w(\tau)\,e^{2\pi \mathbb{i} f\tau}}\,\mathrm{d}\tau, \end{aligned} \tag{9}$$

where the relation between the signal and sinusoid is different from Eq. (1). Computing the fast Fourier transform (FFT) after truncating the signal by a compactly-supported window, which is a usual procedure in audio signal processing, corresponds to this definition[8] as the time index of FFT is treated relatively to the window.

3.2. Phase Difference between Two STFTs

Since the complex sinusoid $e^{2\pi \mathbb{i} f\tau}$ in Eq. (1) can be extracted from that of Eq. (9) as

$$\overline{e^{2\pi \mathbb{i} f(\tau - t)}} = \overline{e^{2\pi \mathbb{i} f\tau} e^{-2\pi \mathbb{i} ft}} = \overline{e^{2\pi \mathbb{i} f\tau}} e^{2\pi \mathbb{i} ft}, \tag{10}$$

the relation between the spectrograms is given as

$$X_{\mathrm{Eq}(9)}(f,t) = e^{2\pi \mathbb{i} ft}\, X_{\mathrm{Eq}(1)}(f,t), \tag{11}$$

where $X_{\mathrm{Eq}(9)}(f,t)$ and $X_{\mathrm{Eq}(1)}(f,t)$ represent $X(f,t)$ in Eqs. (9) and (1), respectively. Therefore, their complex arguments $\phi_{\mathrm{Eq}(9)}(f,t)$ and $\phi_{\mathrm{Eq}(1)}(f,t)$, or phases, are related to each other in the following way:

$$\phi_{\mathrm{Eq}(9)}(f,t) = 2\pi ft + \phi_{\mathrm{Eq}(1)}(f,t). \tag{12}$$

This difference can be easily contrasted by considering

$$x(t) = e^{2\pi \mathbb{i} \xi t} \tag{13}$$

as a signal. For simplicity, let the window function w be nonnegative. STFT of Eq. (13) yields

$$X_{\mathrm{Eq}(1)}(f,t) = \int_{\mathbb{R}} w(\tau - t)\, e^{-2\pi \mathbb{i}(f - \xi)\tau}\,\mathrm{d}\tau \tag{14}$$

which becomes a real value at the frequency of the signal ($f = \xi$) because $e^0 = 1$ and w is real. In other words, its phase is zero and does not evolve along time at $f = \xi$,

$$\phi_{\mathrm{Eq}(1)}(\xi, t) = 0. \tag{15}$$

On the other hand, in the same situation, phase rotates along time if STFT is defined as in Eq. (9):

$$\phi_{\mathrm{Eq}(9)}(\xi, t) = 2\pi \xi t, \tag{16}$$

where 2π ambiguity of phase is ignored in this paper.

3.3. Instantaneous Frequency

Let a derivative of phase with respect to time, $\partial\phi/\partial t$, be considered. It can be obtained as a part of the time derivative of $X(f,t) = A(f,t)\, e^{\mathbb{i}\phi(f,t)}$ as[9]

$$\frac{\partial X}{\partial t} = \frac{\partial A}{\partial t} e^{\mathbb{i}\phi} + \mathbb{i}\, A\, e^{\mathbb{i}\phi} \frac{\partial \phi}{\partial t}. \tag{17}$$

When $A \neq 0$, dividing Eq. (17) by $X = Ae^{\mathbb{i}\phi}$ results in $(\partial X/\partial t)/X = (\partial A/\partial t)/A + \mathbb{i}(\partial\phi/\partial t)$, and thus

$$\frac{\partial \phi}{\partial t} = \mathrm{Im}\left[\frac{\partial X}{\partial t}\frac{1}{X}\right], \tag{18}$$

where $\mathrm{Im}[\cdot]$ denotes the imaginary part. This quantity is called *instantaneous frequency* and can be calculated if $\partial X/\partial t$ is available together with X.

One popular method to obtain the time derivative of X is STFT with a differentiated window $\mathrm{d}w/\mathrm{d}t$. Because

$$\begin{aligned} \frac{\partial X}{\partial t}(f,t) &= \frac{\partial}{\partial t}\int_{\mathbb{R}} x(\tau)\,\overline{w(\tau - t)\,e^{2\pi \mathbb{i} f\tau}}\,\mathrm{d}\tau, \\ &= -\int_{\mathbb{R}} x(\tau)\,\frac{\mathrm{d}\overline{w}}{\mathrm{d}t}(\tau - t)\,\overline{e^{2\pi \mathbb{i} f\tau}}\,\mathrm{d}\tau, \end{aligned} \tag{19}$$

calculating STFT with $\mathrm{d}w/\mathrm{d}t$ yields $-\partial X/\partial t$. Therefore, after calculating the time derivative of a window, the instantaneous frequency can be retrieved through Eq. (18) with two STFTs using w and $\mathrm{d}w/\mathrm{d}t$.

3.4. Difference of Instantaneous Frequencies between Two STFTs

Since phase depends on the definition of STFT as in Sect. 3.2, the instantaneous frequency also differs in accordance with it. The relation between the instantaneous frequencies calculated via Eqs. (9) and (1) is

$$\frac{\partial \phi_{\mathrm{Eq}(9)}}{\partial t}(f,t) = 2\pi f + \frac{\partial \phi_{\mathrm{Eq}(1)}}{\partial t}(f,t), \tag{20}$$

where the left-hand side is more accepted definition of "instantaneous frequency" as it represents the absolute

[7] There exist many other definitions: both signal and window may be translated as the ambiguity function; the window function may be flipped; and its complex conjugation may be omitted.

[8] In contrast to the usual implementation, directly implementing STFT in Eq. (1), or DGT in Eq. (7), requires rotation of the time index after truncation by the window (see the supplemental code).

[9] The existence of the partial derivative can be found in [16]. This paper assumes a sufficiently smooth and rapidly-decaying window w so that the time derivative of $X(f,t)$ exists.

frequency. In contrast, the second term on the right-hand side represents the frequency *relative* to the (angular) frequency axis $2\pi f$, and therefore some authors call it "relative instantaneous frequency" [17].

3.5. Instantaneous Frequency for DGT

As differentiation is a concept for continuous functions, the instantaneous frequency has been explained with STFT in the preceding sections. For discrete signals, the instantaneous frequency $D_t\phi$ can be computed numerically through DGT in the same manner:

$$D_t\phi = -\mathrm{Im}\left[\frac{X^D}{X}\right] = -\mathrm{Im}\left[\frac{X^D\overline{X}}{|X|^2}\right], \tag{21}$$

where $X^D[m,n]$ is the DGT coefficient calculated with the differentiated window corresponding to the window used for $X[m,n]$ (note that $X^D[m,n]$ corresponds to $-\partial X/\partial t$ as in Eq. (19), which results in the negative sign). For avoiding division by zero, the denominator may require some treatment [18]. The derivative of a discrete window can be calculated numerically, where the spectral method is an appropriate choice for the numerical differentiation [19].

Similar to Eq. (9), DGT can also be defined as

$$\begin{aligned} X[m,n] &= \sum_{l=0}^{L-1} x[l]\,\overline{w[l-an]}\,e^{2\pi\mathrm{i}bm(l-an)/L}, \\ &= \sum_{l=0}^{L-1} x[l+an]\,\overline{w[l]}\,e^{2\pi\mathrm{i}bml/L}, \end{aligned} \tag{22}$$

whose relation to DGT in Eq. (7) is similar to Eq. (11):

$$X_{\mathrm{Eq}(22)}[m,n] = e^{2\pi\mathrm{i}bman/L}\,X_{\mathrm{Eq}(7)}[m,n]. \tag{23}$$

While the instantaneous frequencies for DGT may differ based on their unit, the relation can be written as

$$D_t\phi_{\mathrm{Eq}(22)}[m,n] = m + D_t\phi_{\mathrm{Eq}(7)}[m,n], \tag{24}$$

whose unit is assumed to be the same as the index m. When the continuous derivative is related to its discrete approximation in a different way, the instantaneous frequency can be defined in other units (e.g., normalized angular frequency) as well.

3.6. Some Topics Related to Phase Derivatives

As another concept related to phase, the *group delay* is defined as the derivative of phase with respect to frequency. It is utilized together with the instantaneous frequency to obtain a sparse time-frequency representation, namely *reassigned spectrogram* [19–27], and the close relationship between these partial derivatives and the log-magnitude spectrogram $\log(|X|)$ has been discussed [25]. Their applications to acoustical signal processing have been proposed recently [28–31]. Not only the instantaneous frequency but also the group delay can be computed in the similar way as Eq. (21), which allows their computation for general filterbanks [19,27], while approximating them by the finite difference method is also a popular strategy [20–22]. Some higher order partial derivatives of phase are recently investigated [32–34] in the context of *synchrosqueezing* which is another method for obtaining a sparse time-frequency representation with emphasis on auditory nerve models and modal decomposition [35–37].

4. SIGNAL PROCESSING WITH PHASE CONVERSION

In the previous sections, the basic concepts related to the Fourier-type time-frequency representation and phase derivative have been briefly reviewed. In this section, a signal-processing scheme based on those concepts is introduced as a part of the main topic of this paper.

4.1. Neighborhood Relation of Spectrogram

As spectrogram is a structured representation, adjacent time-frequency bins exhibit a specific relation. For a complex sinusoid $x[l] = e^{2\pi\mathrm{i}b\xi l/L}$, a simple formula can be derived as in Sect. 3.2. Its DGT in terms of Eq. (22) has the following relationship when the index of spectrogram m coincides with ξ:

$$X_{\mathrm{Eq}(22)}[\xi,n+1]\,e^{-2\pi\mathrm{i}b\xi a/L} = X_{\mathrm{Eq}(22)}[\xi,n], \tag{25}$$

which resembles the phase evolution in Eq. (16). Note that the usual procedure for calculating DGT, "truncate signal and compute FFT" (without index rotation), corresponds to this representation, where the phase of a sinusoidal signal varies along time. That is, the amplitude of this spectrogram is smooth ($|X_{\mathrm{Eq}(22)}[\xi,n+1]| = |X_{\mathrm{Eq}(22)}[\xi,n]|$), yet the complex spectrogram is not smooth owing to the phase evolution.

The above neighborhood relation depends on the definition of DGT. For the index-rotated DGT in Eq. (7), successive time-frequency bins are identical at the frequency of the sinusoid *in terms of complex number*:

$$X_{\mathrm{Eq}(7)}[\xi,n+1] = X_{\mathrm{Eq}(7)}[\xi,n], \tag{26}$$

where the same sinusoid $x[l] = e^{2\pi\mathrm{i}b\xi l/L}$ is considered. This relation indicates that enforcing smoothness in time direction with DGT defined in Eq. (7) can enhance sinusoidal components of a signal [38]. Similarly, some signal processing methods specific to such phase relation can be considered for each definition of DGT.

4.2. Signal Processing via Phase Conversion

Spectrograms computed by differently-defined DGTs can be converted to each other afterward. According to Eq. (23), multiplying $e^{2\pi\mathrm{i}bman/L}$ to $X_{\mathrm{Eq}(7)}[m,n]$ results in $X_{\mathrm{Eq}(22)}[m,n]$. This conversion of the phase can be effortlessly inverted by multiplying $e^{-2\pi\mathrm{i}bman/L}$ again because

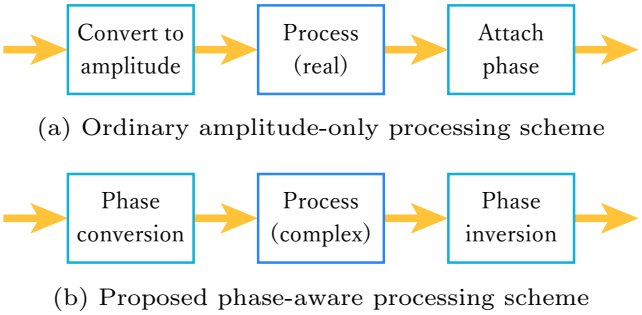

Fig. 1 Comparison of the ordinary amplitude-based and the proposed phase-conversion-based frameworks (the input and output are complex spectrograms).

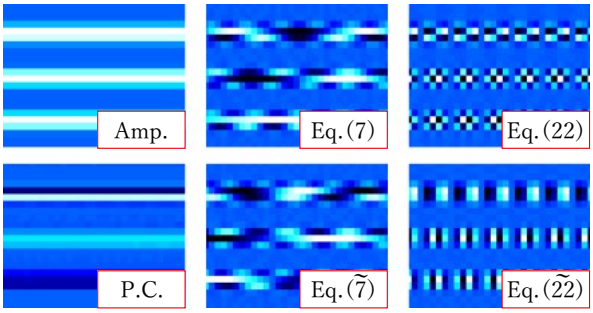

Fig. 2 Spectrograms of sum of 3 sinusoids. While amplitude is the same for all figures, their phase differs depending on the definition of DGT. For visual convenience, the real parts of the complex spectrogram are shown (except the top-left figure which illustrates the amplitude common for every other figures). Four representations on the right, calculated by Eqs. (7) and (22) (those with tildes were calculated with a zero-phased window), exhibit periodic patterns corresponding to the phase evolution. In contrast, the phase-corrected spectrogram on the bottom left is smooth along time.

$e^{-2\pi \mathrm{i} bman/L} e^{2\pi \mathrm{i} bman/L} = 1$. Conversion from $X_{\mathrm{Eq}(22)}[m,n]$ to $X_{\mathrm{Eq}(7)}[m,n]$ can be accomplished in the similar manner, and some other definitions of DGT can also be realized through such multiplication of a phase factor.

Since it is invertible, combination of signal processing with the phase conversion can be easily developed. As a simple example, spectrogram calculated by Eq. (22) can be converted to that by Eq. (7) which allows to utilize the neighborhood relation in Eq. (26). After some signal processing based on that relation is applied, its result is retrieved through inversion of the phase and the inverse DGT. We proposed such signal processing framework based on phase conversion [1] which can be generally stated as follows: (1) multiply $e^{\mathrm{i}\theta[m,n]}$ to the given complex spectrogram X for modifying the phase by a predefined bin-wise scalar $\theta[m,n] \in \mathbb{R}$; (2) apply some signal processing method; and (3) invert the converted phase by multiplying $e^{-\mathrm{i}\theta[m,n]}$.

The ordinary and phase-conversion-based frameworks are schematically contrasted in Fig. 1. Ordinarily, the given spectrogram is converted into amplitude (or an amplitude-based feature), which allows us to adopt a processing method defined in the real number system. We generalized it to the phase conversion so that complex-domain processing can directly handle both amplitude and phase, which makes the processing phase-aware. Note that, when θ is chosen as $\theta[m,n] = -\mathrm{Arg}(X[m,n])$, the first step (phase conversion) obtains

$$e^{\mathrm{i}\theta} X = e^{-\mathrm{i}\mathrm{Arg}(X)}(|X|\, e^{\mathrm{i}\mathrm{Arg}(X)}) = |X|, \tag{27}$$

which illustrates that the ordinary amplitude-based scheme is a special case of the proposed framework.

4.3. Phase Correction Based on Instantaneous Frequency

In the proposed scheme (bottom row of Fig. 1), the constant $\theta[m,n]$ for the phase conversion can be arbitrary and should be chosen so that the converted spectrogram becomes convenient for the subsequent processing. In [1], we proposed to utilize the instantaneous frequency of the inputted signal for that, which is referred to as *phase correction*.

The aim of phase correction is to reduce the amount of mismatch between the frequencies[10] of a sinusoid ξ and time-frequency bins m. As the neighborhood relation in Eq. (26) is appropriate only when $\xi = m$, let a sinusoid whose frequency deviates from m be considered:

$$x[l] = e^{2\pi \mathrm{i} b(m+\delta)l/L}, \tag{28}$$

where $\delta \in \mathbb{R}$ represents the deviation. Its neighborhood relation in terms of DGT in Eq. (7) can be written as

$$X_{\mathrm{Eq}(7)}[m,n+1]\, e^{-2\pi \mathrm{i} b\delta a/L} = X_{\mathrm{Eq}(7)}[m,n], \tag{29}$$

which holds for any m in contrast to Eq. (26) (only appropriate for $m = \xi$). Since δ is the amount of mismatch to the index m, it resembles the relative instantaneous frequency $D_t\phi_{\mathrm{Eq}(7)}$ in Eq. (24). Setting it as $\theta[m,n]$ and applying the phase conversion should cancel the mismatch $\delta[m,n]$ and make the complex spectrogram smooth along time. Therefore, phase conversion based on cumulative sum of $D_t\phi_{\mathrm{Eq}(7)}[m,n]$ was proposed[11] and named as phase correction [1].

The effect of the phase correction is illustrated in Fig. 2, where the spectrograms consisting of three sinusoids were calculated with different definitions, phase correction was calculated by Eq. (21), and the real parts

[10] Here, frequencies are considered in the unit of index.

[11] The reason for using cumulative sum of the relative instantaneous frequency $\theta[m,n] = \sum_{k=1}^{n} D_t\phi_{\mathrm{Eq}(7)}[m,k]$ is that $D_t\phi_{\mathrm{Eq}(7)}$ only represents the adjacent (local) relation, while the phase correction is based on the global relation $\theta[m,n]$ given for all m,n.

are shown for visibility[12]. While amplitude is the same for all expressions, phase is different in accordance with the definition of DGT. By applying phase correction, phase evolution was canceled that can be seen from the figure at the bottom left. As sinusoids become smooth along time, phase correction has been applied to some applications enhancing/distinguishing sinusoidal components such as speech enhancement [1] and harmonic/percussive source separation [39].

4.4. Application: Complex Low-rankness

An interesting example of its property is that phase correction allows a complex spectrogram consisting of sinusoidal components to be low-rank [40]. While low-rank modeling of amplitude spectrograms including NMF (nonnegative matrix factorization) has been studied widely [41], there exists few research considering low-rankness of complex spectrograms [42]. This is because usual complex spectrograms are not low-rank owing to the phase evolution. Low-rankness of the phase-corrected spectrogram might be another direction of low-rank modeling in acoustics as it is phase-aware.

5. SIGNAL PROCESSING WITH SPECTROGRAM CONSISTENCY

Spectrogram consistency is a popular property utilized in phase-aware signal processing. In this section, we briefly mention that phase correction in the previous section can be combined with spectrogram consistency.

5.1. Spectrogram Consistency in Nutshell

In most cases, DGT maps a time-domain signal into higher-dimensional representation, i.e., dimensionality of spectrogram NM is higher than that of the signal L. Such redundancy indicates that only the L-dimensional subspace within the NM dimensions directly corresponds to time domain. In other words, any component in the remaining $(NM - L)$-dimensional subspace does not affect the time-domain signal. Spectrograms are said to be *consistent* when they do not contain any component in that $(NM - L)$-dimensional subspace. While DGT of any time-domain signal is a consistent spectrogram, it can easily be inconsistent after some time-frequency-domain signal processing. Since inconsistent spectrogram contains ineffective components, constraining signal processing methods to output consistent spectrogram has been preferred by many researchers.

5.2. Phase Recovery Based on Consistency

Phase recovery is a branch of signal processing which aims to estimate better phase corresponding to the given spectrogram [43,44]. One popular application of spectrogram consistency is phase recovery which, in such case, is formulated as a problem of finding a spectrogram, whose amplitude is close to the given one, within the consistent L-dimensional subspace. To do so, DGT and its inverse are repeatedly applied in an iterative procedure so that the result is constrained to be consistent [45–47]. Such combination of forward and inverse DGT corresponds to the projection onto that subspace.

5.3. Phase Correction with Consistency

Time-frequency-domain signal processing can be constrained to be consistent by applying it with the pair of forward and inverse DGT. However, nonlinear transform (such as taking absolute value) may complicate derivation and/or analysis of a signal-processing algorithm. As a simpler alternative, phase correction can be utilized to construct a phase-aware and consistent signal processing method [1,39,40].

By handling variables in time domain and processing complex spectrogram through DGT, the processed result can be searched within the consistent subspace. Phase correction allows to take advantage of the structure of spectrograms including time-directional smoothness and low-rankness which have been rarely considered in the context of spectrogram consistency. As phase correction is linear transform, it is easier than nonlinear transform in both theoretical and practical senses. Some recent finding suggests effectiveness of such phase-correction-based consistent signal processing [1,39,40], which should be worth investigating further.

6. CONCLUSION

In this paper, several time-frequency representations, STFT and DGT with variation on their definitions, were summarized with emphasis on difference of phase and instantaneous frequency among them. Signal processing with phase conversion and its instantaneous-frequency-based version, or phase correction, was introduced based on the neighborhood relation of spectrogram. Their collaboration with spectrogram consistency was also mentioned shortly. Some research has suggested that time-frequency representation is a worthwhile topic for discussion also in deep-learning-based methods [47,48], and we hope that this paper (together with the supplemental MATLAB code) will be helpful for developing audio

[12]Equation numbers indicate the definitions of DGTs. Those with tilde were calculated with the same but zero-phased window (index was rotated before FFT so that peak of the window becomes the first element, see the supplemental code). Making a window zero-phased removes the linear phase component caused by the mismatch of indexing between the signal and FFT (its frequency-domain counterpart is well-known: index is rotated so that the negative frequency follows after the Nyquist frequency).

signal processing with consideration of such representation in time-frequency domain.

REFERENCES

[1] K. Yatabe and Y. Oikawa, "Phase corrected total variation for audio signals," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), pp. 656–660 (2018).

[2] H. G. Feichtinger and T. Strohmer, Eds., *Gabor Analysis and Algorithms: Theory and Applications* (Birkhäuser, Boston, 1998).

[3] H. G. Feichtinger and T. Strohmer, Eds., *Advances in Gabor Analysis* (Birkhäuser, Boston, 2003).

[4] O. Christensen, *Frames and Bases: An Introductory Course* (Birkhäuser, Boston, 2008).

[5] O. Christensen, *An Introduction to Frames and Riesz Bases* (Birkhäuser, Boston, 2016).

[6] P. G. Casazza and G. Kutyniok, Eds., *Finite Frames: Theory and Applications* (Birkhäuser, Boston, 2013).

[7] K. Gröchenig, *Foundations of Time-Frequency Analysis* (Birkhäuser, Boston, 2001).

[8] T. Werther, Y. C. Eldar and N. N. Subbanna, "Dual Gabor frames: Theory and computational aspects," *IEEE Trans. Signal Process.*, **53**, 4147–4158 (2005).

[9] S. Li, Y. Liu and T. Mi, "Sparse dual frames and dual Gabor functions of minimal time and frequency supports," *J. Fourier Anal. Appl.*, **19**, 48–76 (2013).

[10] N. Perraudin, N. Holighaus, P. L. Søndergaard and P. Balazs, "Designing Gabor windows using convex optimization," *Appl. Math. Comput.*, **330**, 266–287 (2018).

[11] T. Kusano, Y. Masuyama, K. Yatabe and Y. Oikawa, "Designing nearly tight window for improving time-frequency masking," *arXiv 1811.08783* (2018).

[12] Z. Cvetkovic, "On discrete short-time Fourier analysis," *IEEE Trans. Signal Process.*, **48**, 2628–2640 (2000).

[13] J. Kovacevic and A. Chebira, "Life beyond bases: The advent of frames (Part I)," *IEEE Signal Process. Mag.*, **24**, 86–104 (2007).

[14] J. Kovacevic and A. Chebira, "Life beyond bases: The advent of frames (Part II)," *IEEE Signal Process. Mag.*, **24**, 115–125 (2007).

[15] I. Daubechies, A. Grossmann and Y. Meyer, "Painless non-orthogonal expansions," *J. Math. Phys.*, **27**, 1271–1283 (1986).

[16] P. Balazs, D. Bayer, F. Jaillet and P. L. Søndergaard, "The pole behavior of the phase derivative of the short-time Fourier transform," *Appl. Comput. Harmon. Anal.*, **40**, 610–621 (2016).

[17] N. Perraudin, N. Holighaus, P. Majdak and P. Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **26**, 1083–1094 (2018).

[18] H. Kawahara, T. Irino and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), pp. 5420–5423 (2011).

[19] S. Fenet, R. Badeau and G. Richard, "Reassigned time-frequency representations of discrete time signals and application to the constant-Q transform," *Signal Process.*, **132**, 170–176 (2017).

[20] K. Kodera, C. D. Villedary and R. Gendrin, "A new method for the numerical analysis of non-stationary signals," *Phys. Earth Planet. Inter.*, **12**, 142–150 (1976).

[21] D. J. Nelson, "Cross-spectral methods for processing speech," *J. Acoust. Soc. Am.*, **110**, 2575–2592 (2001).

[22] S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *J. Acoust. Soc. Am.*, **119**, 360–371 (2006).

[23] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Process.*, **43**, 1068–1089 (1995).

[24] G. K. Nilsen, "Recursive time-frequency reassignment," *IEEE Trans. Signal Process.*, **57**, 3283–3287 (2009).

[25] F. Auger, E. Chassande-Mottin and P. Flandrin, "On phase-magnitude relationships in the short-time Fourier transform," *IEEE Signal Process. Lett.*, **19**, 267–270 (2012).

[26] F. Auger, P. Flandrin, Y. Lin, S. McLaughlin, S. Meignen, T. Oberlin and H. Wu, "Time-frequency reassignment and synchrosqueezing: An overview," *IEEE Signal Process. Mag.*, **30**, 32–41 (2013).

[27] N. Holighaus, Z. Průša and P. L. Søndergaard, "Reassignment and synchrosqueezing for general time-frequency filter banks, subsampling and processing," *Signal Process.*, **125**, 1–8 (2016).

[28] Z. Průša, P. Balazs and P. L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **25**, 1154–1164 (2017).

[29] D. Fourer and G. Peeters, "Fast and adaptive blind audio source separation using recursive Levenberg–Marquardt synchrosqueezing," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), pp. 766–770 (2018).

[30] A. Hiruma, K. Yatabe and Y. Oikawa, "Separating stereo audio mixture having no phase difference by convex clustering and disjointness map," *Proc. Int. Workshop Acoust. Signal Enhance.* (*IWAENC*), pp. 266–270 (2018).

[31] A. Hiruma, K. Yatabe and Y. Oikawa, "Detection of clean time-frequency bins based on phase derivative of multichannel signals," *Proc. Int. Congr. Acoust.* (*ICA*), (2019).

[32] T. Oberlin, S. Meignen and V. Perrier, "Second-order synchrosqueezing transform or invertible reassignment? Towards ideal time-frequency representations," *IEEE Trans. Signal Process.*, **63**, 1335–1344 (2015).

[33] D. Pham and S. Meignen, "High-order synchrosqueezing transform for multicomponent signals analysis — With an application to gravitational-wave signal," *IEEE Trans. Signal Process.*, **65**, 3168–3178 (2017).

[34] R. Behera, S. Meignen and T. Oberlin, "Theoretical analysis of the second-order synchrosqueezing transform," *Appl. Comput. Harmon. Anal.*, **45**, 379–404 (2018).

[35] I. Daubechies and S. Maes, "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," *Wavelets Med. Biol.*, pp. 527–546 (1996).

[36] I. Daubechies, J. Lu and H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Appl. Comput. Harmon. Anal.*, **30**, 243–261 (2011).

[37] T. Oberlin, S. Meignen and V. Perrier, "The Fourier-based synchrosqueezing transform," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), pp. 315–319 (2014).

[38] I. Bayram and M. E. Kamasak, "A simple prior for audio signals," *IEEE Trans. Audio Speech Lang. Process.*, **21**, 1190–1200 (2013).

[39] Y. Masuyama, K. Yatabe and Y. Oikawa, "Phase-aware harmonic/percussive source separation via convex optimization," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), (2019).

[40] Y. Masuyama, K. Yatabe and Y. Oikawa, "Low-rankness of complex-valued spectrogram and its application to phase-aware audio processing," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), (2019).

[41] D. Kitamura, "Nonnegative matrix factorization based on complex generative model," *Acoust. Sci. & Tech.*, **40**, 155–161 (2019).
[42] K. Yatabe and D. Kitamura, "Determined blind source separation via proximal splitting algorithm," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), pp. 776–780 (2018).
[43] Y. Wakabayashi, "Speech enhancement using harmonic-structure-based phase reconstruction," *Acoust. Sci. & Tech.*, **40**, 162–169 (2019).
[44] Y. Masuyama, K. Yatabe and Y. Oikawa, "Model-based phase recovery of spectrograms via optimization on Riemannian manifolds," *Proc. Int. Workshop Acoust. Signal Enhance.* (*IWAENC*), pp. 126–130 (2018).
[45] Y. Masuyama, K. Yatabe and Y. Oikawa, "Griffin–Lim like phase recovery via alternating direction method of multipliers," *IEEE Signal Process. Lett.*, **26**, 184–188 (2019).
[46] K. Yatabe, Y. Masuyama and Y. Oikawa, "Rectified linear unit can assist Griffin–Lim phase recovery," *Proc. Int. Workshop Acoust. Signal Enhance.* (*IWAENC*), pp. 555–559 (2018).
[47] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, "Deep Griffin–Lim iteration," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), (2019).
[48] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, "Data-driven design of perfect reconstruction filterbank for DNN-based sound source enhancement," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), (2019).

Kohei Yatabe received his B.E., M.E., and Ph.D. degrees from Waseda University in 2012, 2014, and 2017, respectively. He is currently an assistant professor of the Department of Intermedia Art and Science, Waseda University. His research interests include optical and acoustical signal processing.

Yoshiki Masuyama received the B.E. degree from Waseda University in 2019. He is currently pursuing a M.E. degree in the Department of Intermedia Art and Science, Waseda University. His research interests include musical acoustics and acoustical signal processing using convex optimization technique.

Tsubasa Kusano received the B.E. and M.E. degrees from Waseda University in 2017 and 2018, respectively. He is currently pursuing a Ph.D. degree in the Department of Intermedia Art and Science, Waseda University. His research interests include signal processing based on functional data analysis and time-frequency analysis.

Yasuhiro Oikawa received his B.E., M.E., and Ph.D. degrees in Electrical Engineering from Waseda University in 1995, 1997, and 2001, respectively. He is a professor of the Department of Intermedia Art and Science, Waseda University. His main research interests are communication acoustics and digital signal processing of acoustic signals. He is a member of ASJ, ASA, IEICE, IEEE, IPSJ, VRSJ, and AIJ.