# Time stretching tailored for percussive sound based on adaptive time-frequerncy representation*

◎Natsuki Akaishi (TUAT), Nicki Holighaus (ÖAW), Kohei Yatabe (TUAT)

## 1 Introduction

Time stretching is a process that modifies the time scale of a signal without altering its pitch. One of the most widely used techniques is the phase vocoder (PV). However, PV-based methods often suffer from "percussion smearing," which degrades the quality of percussive components. We attribute this artifact to a time-scale mismatch between the temporally smeared magnitude spectrogram and the localized, newly generated phase. In this study, to address this mismatch, we propose a signal-adaptive PV algorithm that reduces percussion smearing. Our method employs an adaptive time-frequency (T-F) representation based on the nonstationary Gabor transform (NSDGT), providing temporally localized magnitudes for percussive components. Experimental result demonstrates that the proposed method effectively preserves the quality of percussive sounds.

## 2 PV-based time stretching

PV-based time stretching is performed in three steps, as shown in the top row of Fig. 1. First, the magnitude of the T-F representation of the original signal is obtained using the discrete Gabor transform (DGT). The phase of the time-stretched signal is then generated by the PV, and the resulting signal is synthesized using the inverse DGT. Because the PV assumes a sinusoidal model for phase generation, it can properly stretch sinusoidal components. However, it is not well suited for generating appropriate phase information for percussive components, leading to the sound degradation.

This degradation arises from a mismatch between the magnitude and phase of T-F representation. Although an ideal phase could, in principle, localize percussive components in the time domain, the magnitude itself is temporally spread, as shown in the center of the top row of Fig. 1. As a result, even with accurate phase generation, percussive components cannot be fully localized and become temporally smeared in the stretched signal.

## 3 Proposed method

In this study, to reduce the mismatch between magnitude and phase, we propose a signal-adaptive PV algorithm that contains a preprocessing method for the magnitude spectrogram used in phase generation (Fig. 1, center). The proposed method obtains a magnitude spectrogram whose temporal localization is enhanced only in regions containing percussive components. This is achieved using the NSDGT, which provides an adaptive T-F representation through variable window lengths and hop sizes.

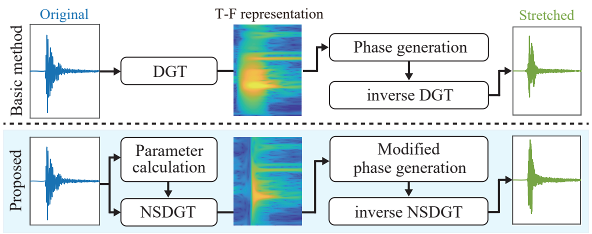


Fig. 1 Flowcharts of standard time stretching based on PV (top) and the proposed method (bottom). In the T-F representations, the vertical and horizontal axes correspond to frequency and time, respectively.

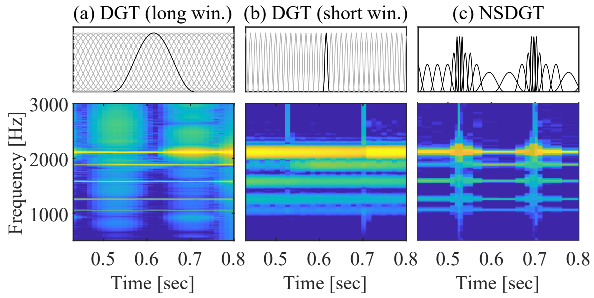


Fig. 2 Comparison of windows and spectrograms for the DGT and NSDGT. The upper boxes show the shifted windows, with representative DGT windows highlighted in black. The lower boxes show the corresponding spectrograms for (a) DGT with a long window, (b) DGT with a short window, and (c) NSDGT.

### 3.1 NSDGT

The NSDGT [1] generalizes the DGT by incorporating variable hop sizes, windows, and a varying number of frequency channels per time-frame to yield a flexible, invertible T-F representation. In this study, only the hop sizes and windows are varied, whereas the number of frequency channels remains fixed to simplify phase generation process.

Fig. 2 illustrates the compatibility of the NSDGT. As shown in (a) and (b), the DGT exhibits a trade-off between T-F localization: a long window provides good frequency localization for sinusoidal components, whereas a short window provides good temporal localization for percussive components. In contrast, the NSDGT in (c) employs adaptive windowing, using long windows for sinusoidal regions, short windows for percussive regions [1], and intermediate window lengths for mixed regions, thereby improving the overall T-F localization.

---

*適応的な時間周波数表現に基づく打撃音に特化したタイムストレッチング．赤石夏輝 (農工大), Nicki Holighaus (オーストリア科学アカデミー), 矢田部浩平 (農工大).

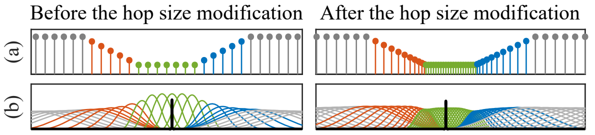


Fig. 3 (a) Window lengths and (b) shifted windows around an impulse. The left and right columns correspond to constant and adaptive hop sizes, respectively. Windows before, during, and after the impulse are highlighted in red, green, and blue, respectively.

### 3.2 Window length calculation

The proposed method designs windows for NS-DGT such that percussive components are naturally localized according to the stretching factor. Fig. 3 conceptually shows the windows, where the center of the percussive event is represented by an impulse. Before the impulse, the window length is gradually decreased; around the impulse, short windows are used; and after the impulse, the window length is gradually increased back to its original value.

The minimum window length $S$ used in percussive regions (Fig. 3, green) is calculated as

$$S = \left\lfloor \left(1 - r^2(1 - 1/\alpha)\right) W \right\rfloor , \tag{1}$$

where $W$ is the original window length, $\alpha \geq 1$ is the stretching factor, $\lfloor \cdot \rfloor$ denotes the floor operator, and $r \in (0, 1]$ is a compression rate. The compression rate $r$, originally introduced in [2], enables adaptive control of the degree of temporal compression, thereby preventing sound quality degradation.

To obtain a natural T-F representation, the window lengths are smoothly varied around each percussive event. The adaptation is applied to the time frames satisfying $I - N_{\mathrm{h}} \leq n' \leq I + N_{\mathrm{h}}$, where $I$ denotes the center of the percussive event and $N_{\mathrm{h}}$ the number of frames covering it. We define the window length at the $n'$-th time frame as

$$W_{n'} = \max(S,\ W - 2a_{\min}|N_{\mathrm{h}} - n'|) , \tag{2}$$

where $a_{\min}$ is the minimum hop size.

### 3.3 Hop size modification

As shown in the left column of Fig. 3, conventional fixed hop sizes often provide insufficient coverage of percussive events, leading to excessive window amplitudes and unstable signal synthesis. While a straightforward remedy would be to uniformly reduce the hop size, this approach significantly increases the computational cost. To address this limitation, the proposed method smoothly varies the hop size (Fig. 3, right), increasing the density of windows exclusively around percussive events. This enables a localized, dense observation of percussive components while maintaining low computational complexity. Consequently, as demonstrated in the lower-right panel of Fig. 3, the resulting window amplitudes remain nearly constant.

In accordance with this hop-size modification, the subsequent phase reconstruction is approprately adjusted by interpolating the phase derivatives.

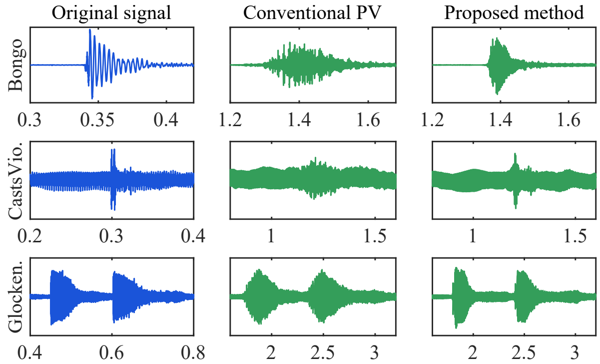


Fig. 4 Comparison of the stretched signals for "Bongo," "CastanetsViolin," and "Glockenspiel" (from top to bottom). The left column shows the original signals (blue), while the center and right columns display the stretched signals (green) obtained by the conventional PV and the proposed method, respectively. The horizontal axis of each panel represents time in seconds.

## 4 Experiment

The proposed method was compared with the conventional PV algorithm. Experiments were conducted using the real-world signals "Bongo", "CastanetsViolin", and "Glockenspiel" from the TSM Toolbox [3]. The sampling frequency was 22 500 Hz. A DGT with a Hann window of length $2^{11}$ and a synthesis hop size of $2^9$ was used. The number of frequency channels was set to $2^{15}$. For the proposed method, the maximum window length and minimum hop size were set to $W = 2^{11}$ and $a_{\min} = 2^7$, respectively. The stretching factor was set to $\alpha = 4$. Audio examples are available on our demo page[1].

Examples of the stretched signals are presented in Fig. 4. Percussive smearing caused by the conventional PV is clearly visible in all cases. In contrast, the proposed method yields a more natural waveform by preserving transient attacks and harmonic signal components, demonstrating a significantly higher fidelity to the original signal.

## 5 Conclusion

In this study, we proposed a time-stretching method for mitigating percussive smearing by obtaining a temporally localized T-F representation using the NSDGT. Experimental result showed that the proposed method yields more natural time-stretched signals than a conventional method.

### References


[1] P. Balazs, M. Dörfler, F. Jaillet, N. Holighaus and G. Velasco, "Theory, implementation and applications of nonstationary Gabor frames," *J. Comput. Appl. Math.*, **236**(6), 1481–1496 (2011).

[2] N. Akaishi, K. Yatabe and Y. Oikawa, "Improving phase-vocoder-based time stretching by time-directional spectrogram squeezing," *IEEE ICASSP*, pp. 1–5 (2023).

[3] J. Driedger and M. Müller, "TSM Toolbox: MATLAB implementations of time-scale modification algorithms," *DAFx*, pp. 249–256 (2014).


[1] https://natsukiakaishi.github.io/selebi/