

ACOUSTICAL LETTER

Technique of improving speech restoration from spectrograms of short windows for the Griffin–Lim algorithm

Naofumi Aoki*

Graduate School of Information Science and Technology, Hokkaido University, N14 W9, Kita-ku, Sapporo, 060–0814 Japan

(*Received 14 December 2022, Accepted for publication 6 January 2023*)

Keywords: Acoustical archeology, Speech restoration, Griffin–Lim algorithm, Source-filter theory

1. Introduction

The restoration of speech materials recorded in the past might be regarded as a study in acoustical archeology. It may potentially attract the interest of many people learning acoustics [1]. For that purpose, we have devised a tool that may be employed for such education [2]. It especially focuses on speech restoration from spectrogram images found in past documents.

One of the oldest spectrogram images is exemplified in a book titled *Visible Speech* published in 1947 [3] shortly after a device called the sound spectrograph, invented at Bell Telephone Laboratories, was publicized in 1946 [4]. Some spectrogram images in the book can nowadays be found on some websites [5].

Almost all such spectrogram images are so-called amplitude spectrograms that include no phase information. Nevertheless, phase information that is lost in these spectrogram images must be appropriately estimated in some way to enable speech restoration. The Griffin–Lim algorithm is known as one such technique that may restore speech signals from only-amplitude spectrograms [6]. It estimates appropriate phase information from the amplitude spectrogram by ensuring consistency of frequency components between adjacent frames. It iterates pairs of discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) calculations to search for the optimal phase information.

With the Griffin–Lim algorithm, it is not so difficult to restore speech signals from spectrograms calculated using long windows since these spectrograms show the apparent temporal connections of frequency components between adjacent frames. However, it is often difficult to restore speech signals from spectrograms calculated using short windows. In many cases, spectrogram images printed in past documents were calculated with short windows, so some more improvements are required to realize better speech restoration from these spectrogram images.

On the basis of the above discussion, the main theme of this paper was focused on the phase reconstruction from spectrograms calculated using short windows. It may potentially expand the scope of application of the Griffin–Lim algorithm.

*e-mail: aoki@ime.ist.hokudai.ac.jp
[doi:10.1250/ast.44.186]

2. Speech restoration using Griffin–Lim algorithm

The Griffin–Lim algorithm requires initial phase information for speech restoration. In general, it is set to be random. The iteration process of the Griffin–Lim algorithm is expected to search for appropriate phase information. However, this process does not necessarily reach the optimal solution. Compared with spectrograms of long windows, spectrograms of short windows often yield incorrect solutions, since temporal connections of frequency components between adjacent frames are not sufficiently clear. Consequently, speech restoration from spectrograms of short windows often results in uncomfortable speech quality.

3. Proposed technique

The above-described problem may be mitigated by setting the initial phase information more appropriately. Instead of random initialization, the proposed technique assumes a more plausible choice for the initial phase information. Since the vertical projection of a spectrogram image corresponds to the temporal structure of a speech signal, the proposed technique assumes it to be a coarse estimation of the source signal for the speech generation explained in the source-filter theory [7]. The phase information of the source signal is then calculated and set as the initial phase information of the Griffin–Lim algorithm. The flowchart of the proposed technique is illustrated in Fig. 1.

Spectrograms of short windows have a high time resolution that reflects fine temporal structures of speech signals. Periodic vertical stripes in such a spectrogram image may provide a coarse estimation of the source signal. As shown in Fig. 2, the proposed technique calculates the vertical projection of a spectrogram image by summing its amplitude values. Peak pulses are then extracted to form the source signal. A parabolic interpolation is employed in this process. The source signal includes not only periodic pitch pulses for voiced sections but also aperiodic excitations for unvoiced sections.

Figure 3 shows an example of the speech restoration. Figure 3(a) shows the short-window spectrogram used to obtain Figs. 3(c) and 3(e). The long-window spectrogram shown in Fig. 3(b) is the target of the speech restoration. Figure 3(d) is the source signal employed in the proposed technique. It provides a coarse estimation of the harmonic structure shown in the target speech.

The conventional technique often results in an unstable pitch contours as shown in this example. On the other hand,

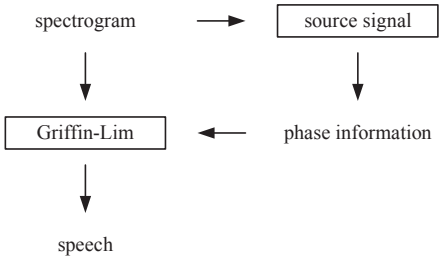

Fig. 1 Flowchart of the proposed technique.

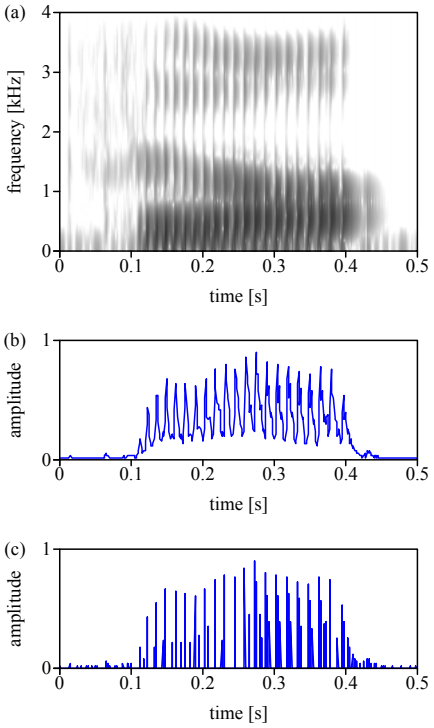

Fig. 2 Estimation of a source signal: (a) short-window spectrogram, (b) its vertical projection, and (c) peak pulses.

the proposed technique may potentially mitigate that problem. The unstable pitch contour is fixed so that the restored speech more closely resembles the original reference speech. This indicates that the initial phase information may control the speech quality.

4. Evaluation

To evaluate the proposed technique compared with the conventional technique of random initialization, an experiment was performed. Forty speech samples obtained from ATR speech database were randomly selected [8]. 20 male (MHT) and 20 female (FKN) speech samples were examined. The speech data were resampled at rates of 8 kHz and 16 kHz. Then, the spectrograms were calculated using short windows. The window size and the shift size of spectrograms were set at 4 ms and 1 ms, respectively. The Hanning window was employed. The number of samples in DFT was set at 1,024.

The number of iterations used in the Griffin–Lim algorithm was set at 100. The signal-to-noise ratio (SNR)

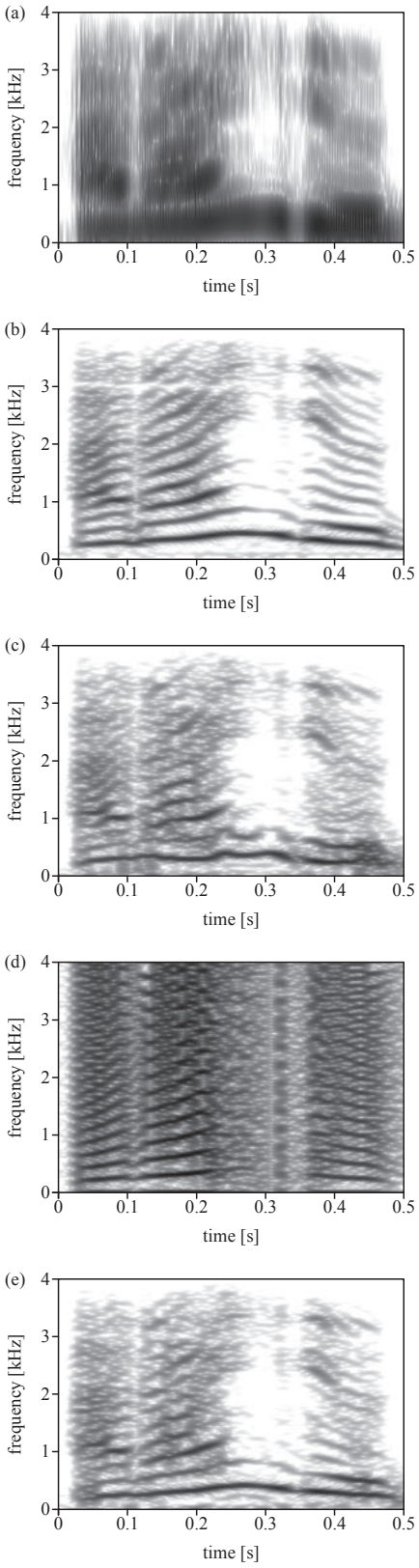

Fig. 3 Spectrograms of the speech restoration: (a) original reference speech (short window), (b) original reference speech (long window), (c) speech restored by the conventional technique, (d) estimated source signal, and (e) speech restored by the proposed technique.

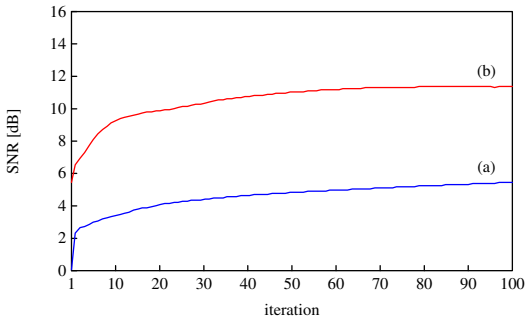

Fig. 4 Convergence curves of the SNR obtained from (a) the conventional technique and (b) the proposed technique.

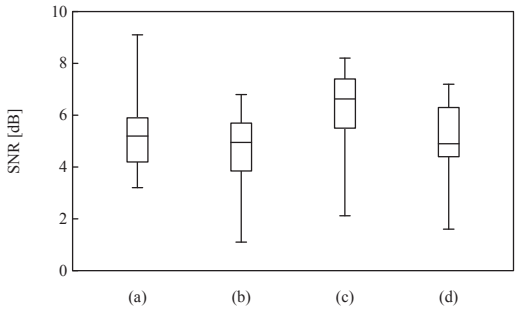

Fig. 5 Differences in the SNR between the proposed technique and the conventional technique: (a) male speech (8 kHz), (b) male speech (16 kHz), (c) female speech (8 kHz), and (d) female speech (16 kHz).

between the original reference speech and the restored speech was employed as a metric of the evaluation. Since the time-domain waveforms do not reflect the quality of phase reconstruction, the SNR was calculated in the spectrogram domain as follows.

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_i\sum_k S(i,k)^2}{\sum_i\sum_k (S'(i,k)-S(i,k))^2} \quad (1)$$

where $S(i,k)$ is the amplitude spectrogram of the original reference speech and $S'(i,k)$ is the amplitude spectrogram of the restored speech. The two-dimensional coordinates denoted by i and k correspond to the frame and the frequency of spectrogram images. These spectrograms were calculated using long windows. The window size and the shift size of spectrograms were set at 32 ms and 1 ms, respectively. The Hanning window was employed. The number of samples in DFT was set at 1,024.

Figure 4 shows an example of the convergence curves of both techniques. As shown in this figure, the proposed technique outperforms the conventional technique. Figure 5 shows how much the SNR is improved by the proposed technique. Since all the cases are positive, it indicates that the proposed technique results in better speech quality in this evaluation than that with the conventional technique. The average improvement of the proposed technique from the conventional technique is about 6 dB. The improvement of the 8 kHz sampling speech seems to be slightly better than that of the 16 kHz sampling speech.

5. Conclusions

The evaluation indicates that the proposed technique may potentially outperform the conventional technique. It indicates that plausible initial phase information may play an important role as a guide when searching for appropriate phase information in the iteration process of the Griffin–Lim algorithm.

The main concern of the proposed technique is its limitation. This technique assumes that the precise information of amplitude spectrograms is available. For the Griffin–Lim algorithm, the correctness of spectrograms is a critical factor in speech restoration. Therefore, the algorithm may perform poorly if the information of spectrograms is significantly degraded.

Printed spectrograms are inevitably degraded by the quantization of gray level, print quality, and scan quality. Such imperfections of spectrograms may have a strong impact on speech restoration. Devising a technique for speech restoration that takes into account such problems still remains as a future work.

References

[1] P. Feaster, *Pictures of Sound* (Dust-to-Digital, Atlanta, 2013).
[2] H. Otake, N. Aoki, K. Ozeki and Y. Dobashi, "A study on the voiced/unvoiced decision of digital pattern playback," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 1365–1366 (2021). (in Japanese).
[3] R. K. Potter, G. A. Kopp and H. C. Green, *Visible Speech* (Van Nostrand, New York, 1947).
[4] W. Koenig, H. K. Dunn and L. Y. Lacy, "The sound spectrograph," *J. Acoust. Soc. Am.*, **18**, 19–49 (1946).
[5] https://www.sciencephoto.com/media/442447/view/spectrogram-of-the-word-visible-speech (accessed on 24 Dec. 2022).
[6] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust. Speech Signal Process.*, **32**, 236–243 (1984).
[7] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals* (Pearson, London, 1978).
[8] ATR Digital Speech Database (set B).