PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Speech enhancement by bit-rate extension based on Time-frequency simultaneous-constrained Griffin-Lim algorithm

Haonan Wang[(1)], Takanobu Nishiura[(2)]

[(1)]Graduate School of Information Science and Engineering, Ritsumeikan University, Japan, is0389sf@ed.ritsumei.ac.jp
[(2)]College of Information Science and Engineering, Ritsumeikan University, Japan, nishiura@is.ritsumei.ac.jp

Abstract

Different from the conventional approach of bandwidth extension (BWE), we propose an algorithm to enhance speech by bit-rate extension (BRE). Due to the different resolution of amplitude, low bit-rate (LB) signals contain much more quantization noise than high bit-rate (HB) signals. Theoretically, simply applying conventional speech enhancement algorithms can enhance LB signals. However, even if LB signals are corrupted by quantization noise, the quantization mechanism makes the remaining upper bits of LB signals identical to the corresponding HB signals, and conventional speech enhancement algorithms may corrupt the "correct" upper bits of LB signals. In this paper, we focus on that the "correct" upper bits of LB signals will follow a time-domain constraint, and use it to support enhancement to achieve a better performance. Therefore, after the general amplitude-based speech enhancement, we also designed a time-frequency simultaneously constrained Griffin-Lim algorithm (TFC-GLA) to ensure that the HB signal estimations satisfy the time-domain constraint while maintaining the estimated amplitude in frequency-domain. Objective evaluations show that our proposed BRE algorithm is effective in enhancing speech.

Keywords: Speech enhancement, Bandwidth extension, Bit-rate extension, Phase reconstruction, Griffin-Lim algorithm

1 INTRODUCTION

Sampling rate is an essential measure for determining the quality of a speech signal. It has been a hot topic of speech enhancement by bandwidth extension (BWE) [1, 2]. Generally, the BWE approach uses the correlation between the low and high-frequency bands and estimates the lost spectrum of the high-frequency band. According to the sampling theory [3], the bandwidth of the spectrum corresponds to the sampling rate, which determines the Nyquist frequency. Therefore, BWE actually extends the sampling rate of speech and improves speech quality.

However, bit rate, which is also an extremely critical measure for determining speech quality, has been forgotten somehow. In this paper, we propose an algorithm to enhance speech by bit-rate extension (BRE). The bit rate represents the resolution of quantization [4], and the higher the bit rate, the higher resolution digital amplitude will be. Thus, BRE can make speech capable of representing more precise fluctuations in amplitude. Figure 1 shows the difference between the two approaches.

Due to the different resolution of amplitude, low bit-rate (LB) signals contain much more quantization noise than high bit-rate (HB) signals. Theoretically, simply applying conventional speech enhancement algorithms [5, 6] can enhance LB signals. However, even if LB signals are corrupted by quantization noise, the quantization mechanism makes the remaining upper bits of LB signals identical to the corresponding HB signals, and conventional speech enhancement algorithms may corrupt the "correct" upper bits of LB signals.

Based on the quantization mechanism, we focus on that the "correct" upper bits of LB signals will follow a time-domain constraint, and use it to support enhancement to achieve a better performance. Conventional amplitude-based speech enhancement algorithms cannot ensure enhanced signals satisfying the time-domain constraint because of the noisy phase. Therefore, we need to design a proper phase, which can ensure our estimation of HB signals following both the time-domain constraint and the estimated amplitude. To achieve this, after

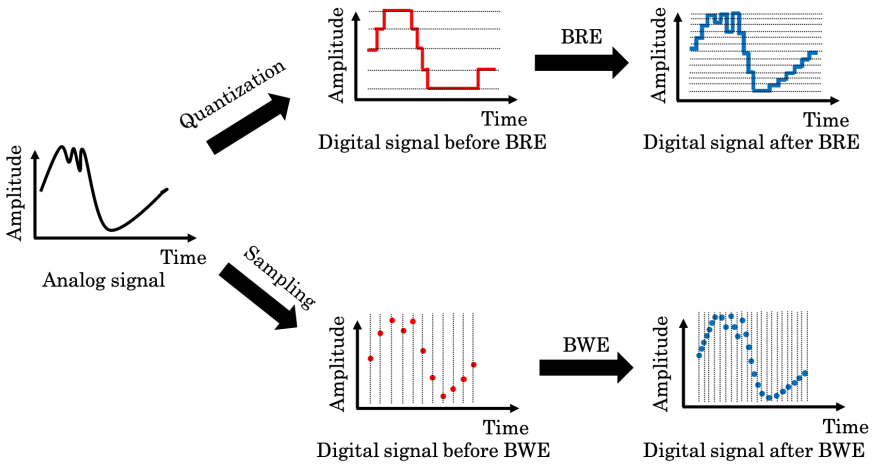


Figure 1. Comparison of BWE and BRE.

amplitude-based speech enhancement, we designed a time-frequency simultaneously constrained Griffin-Lim Algorithm (TFC-GLA) based on the GLA [7]. To evaluate speech-enhancement quality, we conducted objective evaluations based on the perceptual evaluation of speech quality (PESQ) [8] and signal-to-distortion ratio (SDR) to confirm the speech quality improvement as well as the distortion of estimated HB signals referring to the real HB ones. Results of objective evaluations indicate that our proposed BRE algorithm is effective in enhancing speech.

2 Quantization Mechanism

Quantization is a process of mapping input analog voltage from a large set (often a continuous set) to output values in a countable and finite set, which can be considered as a lossy compression of analog amplitude. There are many ways for quantization and the most widely used one is successive approximation register (SAR) [9, 10]. Our proposed BRE algorithm is based on the condition that the input LB signal being extended is linearly quantized by SAR without an additional process such as dithering [11]. Figure 2 shows what results of quantization at different bit rates look like.

From Figure 2, the upper L bits on the MSB side are identical, while the lower $(H-L)$ bits on the LSB side are lost only for the L-bit LB signal. Our proposed method estimates the lower $(H-L)$ bits while protecting the identical upper L bits.

3 PROPOSED BRE ALGORITHM

In this section, we introduce the proposed algorithm. Figure 3 shows a block diagram of the entire proposed algorithm.

3.1 Quantization Noise Whitening and Temporary Bit-rate Extension by Random Recovery of Lower Bits

To achieve speech enhancement by BRE, it is necessary to estimate the lost lower bits. However, estimating lower bits directly from the remaining upper bits is extremely difficult since there is no clear correlation between these two matrixes made of 0 and 1. Thus, we avoid estimating lower bits directly, instead, filling them with values temporarily to increase the resolution, and apply a speech enhancement algorithm later. Assuming that there is an L-bit signal $x(n)$ and $H(L<H)$-bit signal $y(n)$ corresponding to the same speech signal, the

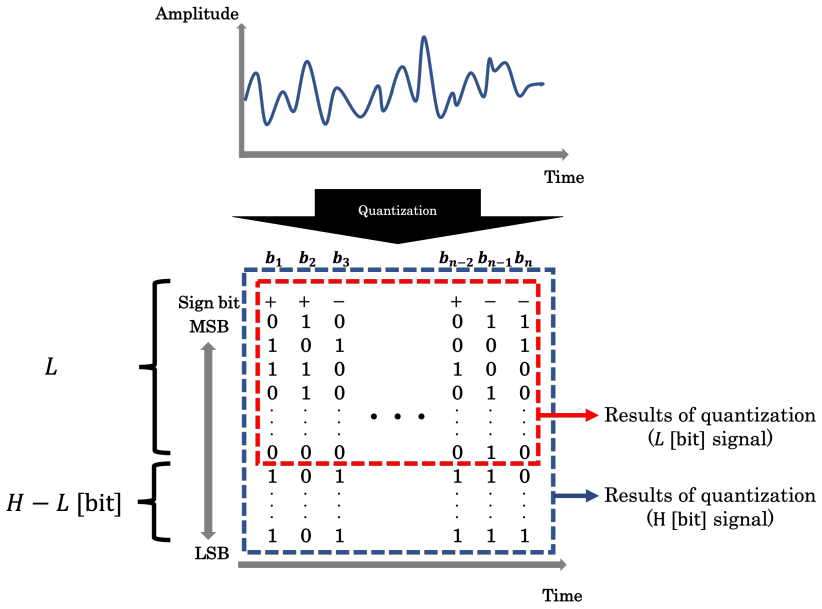


Figure 2. Quantization results in L bits and H bits.

quantization noise $e(n)$ can be represented as

$$e(n) = y(n) - 2^{H-L}x(n). \tag{1}$$

Usually, $e(n)$ has correlations with $x(n)$, which brings some difficulty to the amplitude-based speech enhancement since they are generally based on the assumption of that there is no correlation between the noise and signal. To make this easier, we propose to recover the lost lower bits randomly to whiten the $e(n)$, so that the amplitude estimator may perform better. The procedure of random recovery is shown in Figure 4. Random recovery makes the resolution of an L-bit signal identical to an H-bit signal. In this paper, we call the signal after random recovery a fake HB (FHB) signal.

Figure 5 illustrates spectrograms of quantization noises before and after the random recovery where $L = 8$ and $H = 16$. We can see that the quantization noise has been whitened by random recovery and the LB signal is temporarily extended to HB. Now, we can apply a conventional speech enhancement algorithm to estimate the clean spectrum of HB signals. In this paper, we chose spectral subtraction [5], a widely used speech enhancement algorithm, as the amplitude estimator.

3.2 Theoretical Limited Region of High Bit-rate Signals

In conventional research, simply synthesizing the estimated spectrum $|\hat{Y}(\tau,k)|$ with a noisy phase $\phi(\tau,k)^G$ may improve speech quality. However, as shown in Figure 2, the difference between LB and HB signals only exists in the lost $H-L$ lower bits on the LSB side, which means the L upper bits are identical to the original HB signal, and $\phi(\tau,k)^G$ is not capable of protecting the "correct" upper bits. Since the combination of 0 or 1 for $(H-L)$ lower bits are finite, it is possible to calculate a limited region from an $x(n)$ where the original HB signal is strictly inside. Assuming the lower and upper limits of this region are $l(n)$ and $u(n)$, the relation among $l(n), u(n)$, and $x(n)$ is as follows,

$$\begin{aligned} l(n) &= \min\{2^{H-L}x(n), 2^{H-L}x(n) + (2^{H-L}-1)\mathrm{sgn}(x(n))\}, \\ u(n) &= \max\{2^{H-L}x(n), 2^{H-L}x(n) + (2^{H-L}-1)\mathrm{sgn}(x(n))\}, \end{aligned} \tag{2}$$

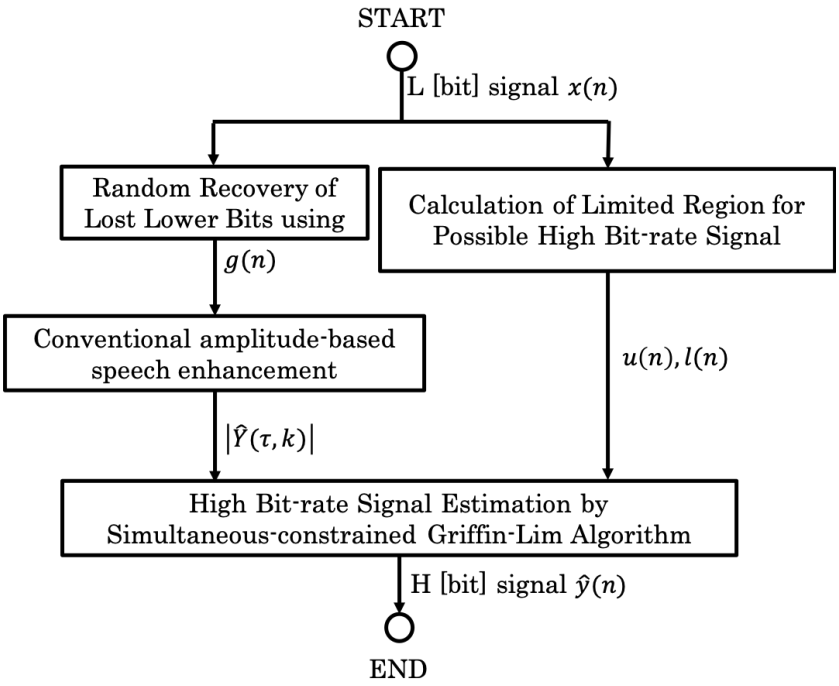


Figure 3. Block-diagram of entire proposed algorithm to enhance speech by bit-rate extension (BRE), where $x(n)$ is input low bit-rate (LB) signal, $g(n)$ is signal after random recovery, $|\hat{Y}(\tau,k)|$ is estimated spectrum, $\hat{y}(n)$ is output of estimated high bit-rate (HB) signal. $u(n)$ and $l(n)$ are upper and lower limits of all possible $\hat{y}(n)$.

where sgn is an operator for extracting the sign of its input value, whose operation is defined as

$$\mathrm{sgn}(x) = \begin{cases} 0 & ,x = 0, \\ 1 & ,x > 0, \\ -1 & ,x < 0. \end{cases} \tag{3}$$

For any LB signal $x(n)$ and bit-rate needed to be extended, it is possible to calculate the upper limit $u(n)$ and lower limit $l(n)$ of $y(n)$ where

$$\forall n,\ l(n) \leq y(n) \leq u(n), \tag{4}$$

and Equation (4) will be the time-domain constraint during the HB signal estimation.

3.3 Time-frequency Simultaneous-constrained Griffin-Lim Algorithm

So far, we have an amplitude estimation as the frequency-domain constraint as well as the time-domain constraint. Next, we need to estimate a proper phase $\hat{\phi}^Y(\tau,k)$ that will satisfy the following condition,

$$\begin{aligned} \hat{y}(n) = \mathrm{ISTFT}\{|\hat{Y}(\tau,k)|\hat{\phi}^Y(\tau,k)\}, \\ \forall n,\ l(n) \leq \hat{y}(n) \leq u(n), \end{aligned} \tag{5}$$

where $\mathrm{ISTFT}\{*\}$ means performing short-time inverse Fourier transform (ISTFT) of the input complex spectrum matrix $*$ [12, 13].

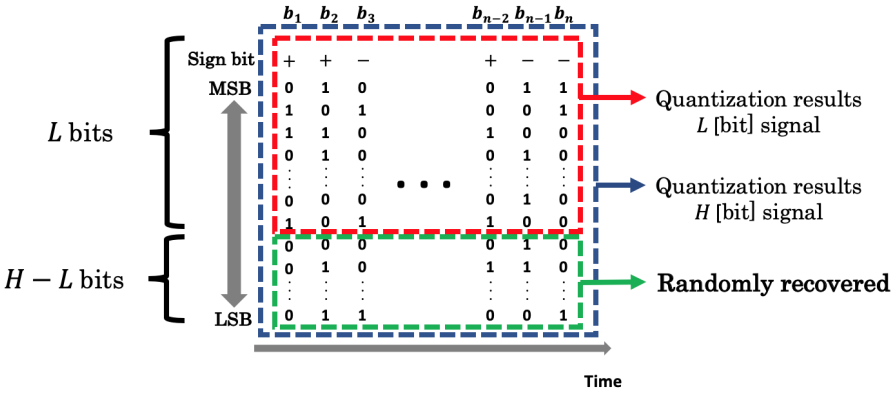
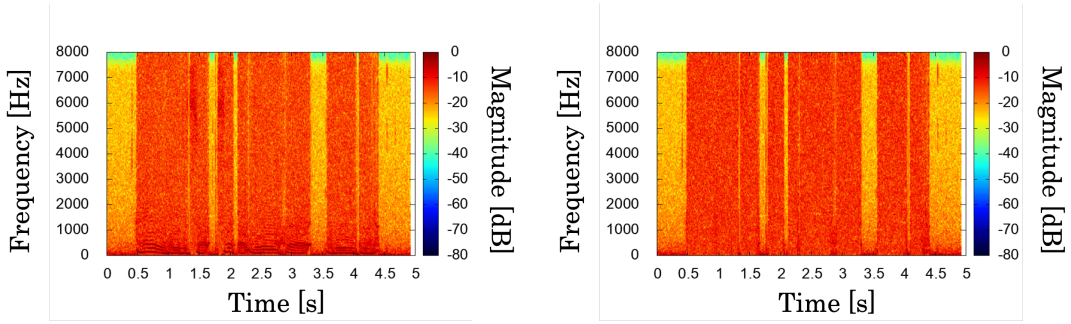


Figure 4. Random recovery of lost $(H-L)$ lower bits on LSB side.



(a) Original.

(b) After random recovery.

Figure 5. Comparison of spectrogram of quantization noise.

It is reported that inserting a time-domain constraint can assist GLA in recovering the phase [14] . Inspired by this discovery, we designed a TFC-GLA to estimate $\hat{\phi}^Y(\tau,k)$ satisfying Equation (5). TFC-GLA is shown as Algorithm 1, where α is the penalty ratio while constraining $\hat{y}(n)$, c is the convergence criterion, $\phi^R(\tau,k)$ is a random phase for initialization, and $\|X\|_F$ is the Frobenius norm of X given by

$$\|X^{n\times m}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} X_{i,j}^2}. \tag{6}$$

With Algorithm 1, we carry out a mutual projection between time and frequency domains. When $|\hat{Y}(\tau,k)|\hat{\phi}^Y(\tau,k)$ is projected into the time domain as $\hat{y}(n)$, we add a penalty to force $\hat{y}(n)$ to become closer to the limit region given by Equation (2). On the other hand, When $\hat{y}(n)$ is projected into the frequency domain as $|\hat{Y}(\tau,k)|\hat{\phi}^Y(\tau,k)$, we force its spectrum to be identical to $|\hat{Y}(\tau,k)|$. Eventually, by iterating the projections, $\hat{y}(n)$ will satisfy Equation (4) while its spectrum will be our desired one.

Algorithm 1 Time-frequency simultaneous-constrained GLA (TFC-GLA)

Input: $l(n), u(n), \alpha, |\hat{Y}(\tau,k)|, c, \phi^R(\tau,k)$
Output: $\hat{y}(n)$
1: $\hat{\phi}^Y(\tau,k) \Leftarrow \phi^R(\tau,k)$
2: $\hat{y}(n) \Leftarrow \mathrm{ISTFT}\{|\hat{Y}(\tau,k)|\hat{\phi}^Y(\tau,k)\}$
3: $\hat{y}'(n) \Leftarrow \begin{cases} \hat{y}(n) - \alpha(\hat{y}(n) - u(n)) & , \ \hat{y}(n) > u(n) \\ \hat{y}(n) - \alpha(\hat{y}(n) - l(n)) & , \ \hat{y}(n) < l(n) \end{cases}$
4: $|\hat{Y}'(\tau,k)|\hat{\phi}^Y(\tau,k) \Leftarrow \mathrm{STFT}\{\hat{y}'(n)\}$
5: **if** $\| |\hat{Y}'(\tau,k)| - |\hat{Y}(\tau,k)| \|_F \geq c$ **then**
6: **go to** 2
7: **else**
8: return $\hat{y}(n)$
9: **end if**

4 OBJECTIVE EVALUATION

To evaluate the performance of our proposed BRE algorithm, we conducted objective evaluations of speech quality based on PESQ [8] and SDR. We chose 20 Japanese speech signals from the ATR speech database [15] sampled at 16 kHz and quantized at 16 bits. We created the corresponding 8-bit signal for each original 16-bit signal by using the following equation.

$$x(n) = \begin{cases} (2^{H-L})\lfloor \dfrac{y(n)}{2^{H-L}} \rfloor & , y(n) \geq 0, \\ (2^{H-L})\lceil \dfrac{y(n)}{2^{H-L}} \rceil & , y(n) < 0, \end{cases} \tag{7}$$

where $\lfloor * \rfloor$ is the operator to calculate the maximum integer less than $*$, $\lceil * \rceil$ is the minimum integer greater than $*$, and $H = 16, L = 8$. Equation (7) is the theoretical relationship between the LB and HB signals of the same analog waveform. Furthermore, the accuracy of magnitude estimation mentioned in Section **??** is also critical to the performance of the proposed BRE algorithm. Thus, we also gave the real amplitude $|Y(\tau,k)|$ to confirm their theoretical maximum performances.

Evaluation target signals were as follow.

- **LB**: **LB** represents 8-bit signals that have been linearly extended to 16-bit. We can consider **LB** as an 8-bit signal since there is no perceptual difference and effect on SDR results after linearly extended to 16-bit.
- **R**: Randomly recovered as described in Section 3.1.
- **R+D**: Only apply the SS and the phase remains noisy.
- **R+D+GLA**: Apply GLA to **R+D**.
- **R+D+TFC-GLA**: Apply Algorithm 1 to **R+D**.
- **R+D***: Theoretical maximum performance of **R+D** by providing it with the real amplitude while the phase is still noisy.
- **R+D*+GLA**: Theoretical maximum performance of **R+D+GLA** by providing it with the real amplitude to GLA

- **R+D*+TFC-GLA**: Theoretical maximum performance of **R+D+TFC-GLA** by providing it with the real amplitude to Algorithm 1

The results of PESQ and SDR are shown in Figures 6-7, respectively. From the results, we can confirm that, random recovery of lost lower bits of **LB** does not improve speech quality. Moreover, results also indicate that only applying SS can improve speech quality, but with the combination of TFC-GLA, we can achieve a more accurate HB signal estimation according to both PESQ and SDR.

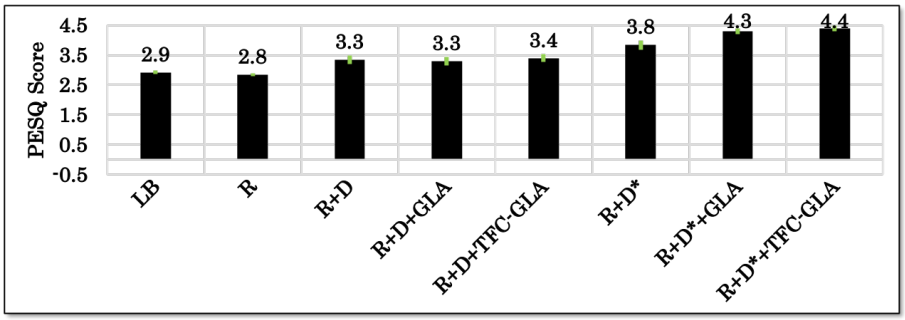


Figure 6. Results of objective evaluation based on PESQ.

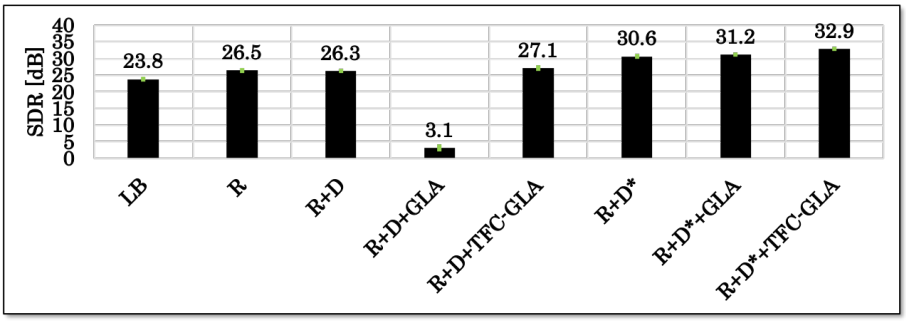


Figure 7. Results of objective evaluation based on SDR.

5 CONCLUSIONS & FUTURE WORKS

We proposed an algorithm of speech enhancement by BRE based on our TFC-GLA. By applying TFC-GLA, we can estimate an HB signal by enhance a LB signal while keeping its upper bits protected. The results of objective evaluations indicate that our proposed BRE algorithm is effective in improving speech quality compared with the LB signal. Conventional amplitude-based speech enhancement can also improve speech quality, but with the combination of TFC-GLA, we can achieve a more accurate HB signal estimation based on the higher SDR results. As for the future work, we will consider applying a more accurate amplitude estimation algorithm.

ACKNOWLEDGEMENTS

This work was partly supported by the Center of Innovation Program (COI) and Japan Society for the Promotion of Science (JSPS) KAKENHI Project Number 18K19829 and 19H04142.

REFERENCES

[1] Z. Ling, Y. Ai, Y. Gu, and L. Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 883–894, May 2018.

[2] H. Liu, C. Bao, and X. Liu, "Spectral envelope estimation used for audio bandwidth extension based on rbf neural network," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 543–547, May 2013.

[3] R. J. Marks, *Introduction to Shannon Sampling and Interpolation Theory*, pp. 1–6. New York, NY: Springer New York, 1991.

[4] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, pp. 2325–2383, Oct 1998.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, Apr 1979.

[6] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *1979. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211, Apr 1979.

[7] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," in *1983 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, pp. 804–807, Apr 1983.

[8] I. Rec, "P. 862: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union*, 2005.

[9] W. Kester, A. Engineeri, i. E. s. Analog Devices, and i. Analog Devices, *Data Conversion Handbook*. Analog Devices series, Elsevier Science, 2005.

[10] W. Ball and H. Coxeter, *Mathematical Recreations and Essays*. Dover Recreational Math Series, Dover Publications, 1987.

[11] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of nonsubtractive dither," *IEEE Transactions on Signal Processing*, vol. 48, pp. 499–516, Feb 2000.

[12] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, pp. 235–238, June 1977.

[13] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, pp. 1558–1564, Nov 1977.

[14] K. Yatabe, Y. Masuyama, and Y. Oikawa, "Rectified linear unit can assist griffin-lim phase recovery," in *2018 16th International Workshop on Acoustic Signal Enhancement*, pp. 555–559, Sept 2018.

[15] 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 Communication*, vol. 9, no. 4, pp. 357–363, 1990.