inter.noise 2014
MELBOURNE AUSTRALIA
16-19 NOVEMBER

Evaluation of clipping-noise suppression of stationary-noisy speech based on spectral compensation

Takahiro FUKUMORI[1]; Makoto HAYAKAWA[1]; Masato NAKAYAMA[2]; Takanobu NISHIURA[2]; Yoichi YAMASHITA[2]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan

[2] College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan

ABSTRACT

Development of communication systems allows people to easily record and distribute their speech. The clipping-noise, however, degrades the sound quality in the speech recording when gain level of input signals is excessive in the maximum range of an amplitude. In this case, it is necessary to suppress the clipping-noise in the observed speech for improving its sound quality. Although a linear prediction method has been conventionally proposed for suppressing the clipping-noise, it has a problem with degradation of the restoration performance by cumulating error when the speech includes a large amount of the clipping-noise. This paper describes a method for the clipping-noise suppression for the stationary-noisy speech based on the spectral compensation in a noisy environment. In this method, to suppress the clipping-noise, the Gaussian mixture models are utilized for modeling the power spectral envelope of the speech on each frame in the lower frequency band. The clean speech signals in a database are also utilized for restoring the clipping speech in the higher frequency band. We carried out evaluation experiments with a speech quality, and confirmed the effectiveness of the proposed method for the speech which includes a large amount of the clipping-noise.

Keywords: Clipping-noise, Spectral envelope, Spectral compensation
I-INCE Classification of Subjects Numbers: 01.4

1. INTRODUCTION

Recent speech communication systems help people to easily record their speech with high-quality. It is necessary for accurately recording the speech to properly set gain level of input signals. In the recording, the clipping-noise is one of the problem which deteriorates the sound quality of a speech signal. It is generated when the amplitude of an input signal unnecessarily exceeds the maximum allowance range (MAR) of an amplitude. In addition, the clipping-noise is also generated due to smaller rated current than the maximum allowance one of an amplifier. The noise also makes listeners uncomfortable due to a loss of the original amplitude in the clipped speech signals. It is required to re-record the speech with the proper gain level if a recorded speech was clipped. It is however necessary to apply a method for the clipping-noise suppression if it is difficult to re-record it in the situation with the speech communication systems in real time.

A conventional method has been proposed for suppressing the clipping-noise by using a linear prediction model (1). The method suppresses the clipping-noise by restoring clipped samples using the linear prediction with the past unclipped samples in the speech. In this method, the restoration performance is however degraded by cumulating prediction error when the clipping-noises are continuously generated in two samples or more of the speech signal. For addressing this problem, it is necessary to process a method without the past speech signals. We have therefore proposed the clipping-noise suppression method that requires no past signals on the basis of the spectral compensation (2). In this method, the spectral envelope of a target speech signal in each analysis frame is approximated to that of an original speech signal to remove the influence of the clipping-noise. In particular, the envelope on the higher frequency band which includes a static characteristic of the speaker is replaced with that of the unclipped speech signal which is prepared in advance. After that, the envelope on the lower frequency which includes a characteristic of a phoneme is approximated with Gaussian mixture models (3).

[1] {cm013061,is033080}@ed.ritsumei.ac.jp

[2] {mnaka@fc,nishiura@is,yama@media}.ritsumei.ac.jp

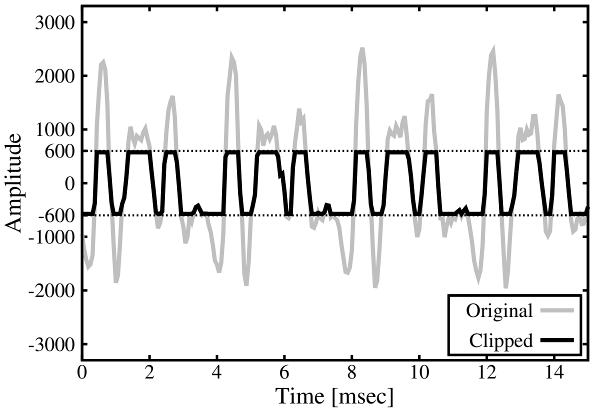


Figure 1 – Waveforms of the original and clipped speech (MAR (A_c): 600)

In this paper, we evaluate the method for the clipping-noise suppression for the stationary-noisy speech in a real noisy environment. We carry out experiments to evaluate the sound quality of the speech signals that are processed by the proposed method.

2. FORMULATION OF CLIPPED SPEECH SIGNAL

This section is described the effect of the clipping-noise in speech. Clipped speech loses the higher or lower amplitude when the absolute one is over the maximum allowance range (MAR). The clipping process is derived from Eq. (1).

$$s_{\mathrm{c}}(n) \quad = \quad \begin{cases} A_{\mathrm{c}} & (s(n) > A_{\mathrm{c}}) \\ s(n) & (|s(n)| \leq A_{\mathrm{c}}) \\ -A_{\mathrm{c}} & (s(n) < -A_{\mathrm{c}}) \end{cases}, \tag{1}$$

where $s(n)$ and $s_c(n)$ are an original speech and a clipped speech signal at time n, respectively. A_{c} indicates the MAR of the clipped speech signal. The clipping-noise is generated when the absolute value of the input speech signal $s(n)$ exceeds the MAR A_{c}. Figure 1 shows an example of the clipped speech under the condition that the MAR A_{c} is set as 600. The clipping ratio (CR) has been conventionally proposed as the evaluation index for the amount of a clipping-noise. The CR C_i of the clipped speech in each frame is derived from Eq. (2).

$$C_i \quad = \quad \frac{A_{\mathrm{c}}}{\sqrt{\frac{1}{N_i}\sum_{n=0}^{N_i-1} s_i(n)^2}}, \tag{2}$$

where $s_i(n)$ is the original speech signal in the i-th frame before clipping, and N_i is also the number of samples in $s_i(n)$. The CR expresses the ratio between the MAR and the root mean square of a speech signal before clipping. The CR becomes lower under the condition with the larger gain level of the clipping-noise.

3. CONVENTIONAL METHOD (LINEAR PREDICTION METHOD)

A linear prediction method (1) has been proposed as the conventional method for the clipping-noise suppression. This method is used the linear prediction model as follows.

$$S(n) \quad = \quad \sum_{i=1}^{p} a_i \cdot s(n-i) + \varepsilon(n), \tag{3}$$

where $s(n)$ is the input speech signal at time n, and $\varepsilon(n)$ is the difference between the original amplitude and the predicted amplitude of the speech signal. $E[\varepsilon(n)]$ becomes zero under the condition that the original speech is a random signal . Equation (3) shows that the obtained amplitude $S(n)$ is predicted by using the amplitude p of the past speech signals from $s(n-1)$ to $s(n-p)$. $a_i (1 \leq i \leq p)$ are called prediction coefficients, and they are calculated so that the expectation value $E[\varepsilon(n)]$ becomes the minimum. The linear prediction method

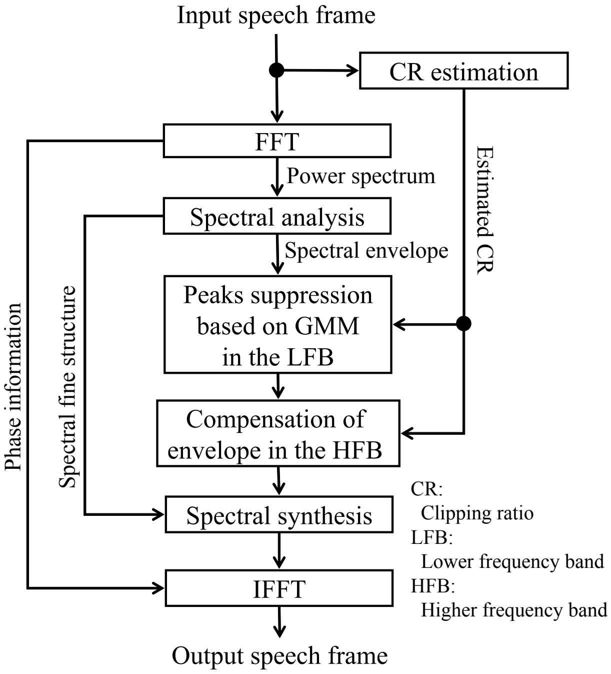


Figure 2 – Flowchart of the proposed method

restores the clipped amplitude with the prediction coefficients which are calculated by using the unclipped speech section. In the method, the restoration performance is however degraded by cumulating prediction error when the clipping-noises are continuously generated in two samples or more of the speech signal.

4. PROPOSED METHOD

This section describes a method for the suppression of the clipping-noise in an observed speech signal based on the spectral compensation. The previous study (2) has clarified some characteristics of the spectral envelope in the clipped speech. There are new some peaks in the spectral envelope of the clipped speech signal in the lower frequency band (LFB). On the other hand, the power of the clipping-noise rises and its spectral envelope becomes a flat shape in the higher frequency band (HFB). The proposed method attempts to suppress the clipping-noise by transforming the spectral envelope on each frequency band on the basis of the difference of characteristics in each LFB and HFB. Figure 2 shows the flowchart of the proposed method.

4.1 Estimation of the clipping ratio

The CR is initially estimated for compensation of the LFB and HFB in the "CR estimation" shown in Fig. 2. The preliminary experiments have confirmed a higher correlation between the CR and the logarithmic clipping incidence (LCI). The LCI L_i logarithmically shows the incidence of the clipped signals in the speech as follows.

$$L_i = \log_{\mathrm{e}} \frac{N_i}{D_i}, \tag{4}$$

where D_i is also the number of samples whose absolute amplitudes are the same as A_c in the ith analysis frame of the clipped speech signal. The LCI becomes lower under the condition with lower CR. The proposed method then estimates the CR using the LCI as follows.

$$\hat{C}_i = \alpha \cdot L_i, \tag{5}$$

where $\hat{C}_i$ is the estimated CR, and α is also the regression coefficient. As stated above, the compensation strength of the clipped speech signal is controlled on the basis of the estimated CR.

4.2 Peaks suppression of the spectral envelope in the lower frequency band

In the "Peaks suppression based on GMM in LFB" shown in Fig. 2, the peaks of the spectral envelope in the LFB are controlled with the approximated ones on the basis of Gaussian mixture models (GMMs) (3) which are expressed as follows.

$$S_\ell(k) \;=\; \sum_{m=1}^{M} w_m \cdot N(k\,|\,\mu_m,\,\sigma_m^2) \quad (w_1 > w_2 > \cdots > w_M)\,, \tag{6}$$

where $S_\ell(k)$ is the normalized spectral envelope in the LFB, $N(k|\,\mu_m,\,\sigma_m^2)$ is Gaussian function, M is the mixture number of Gaussian functions, and w_m, μ_m, and σ_m^2 are the weight, mean, and variance of each Gaussian function, respectively. The first and second formants which have large powers are approximated using two Gaussian functions with the higher weights when the spectral envelope in the LFB is approximated with GMM. In the proposed method, the spectral envelope of the clipped speech is multiplied by the peaks suppression function on the basis of the Gaussian functions with the $M-2$ lower weights as follows.

$$W(k) \;=\; \prod_{m=3}^{M}\left[1-\beta\cdot\exp\left\{-\frac{(k-\mu_m)^2}{2\sigma_m^2}\right\}\right] \quad (0<\beta<1), \tag{7}$$

where $W(k)$ is the peaks suppression function, and β is also the suppression coefficient based on the estimated CR. The peaks generated by the clipping-noise are suppressed by multiplying the spectral envelope using the peak suppression function $W(k)$.

4.3 Spectral compensation with the clean speech in the higher frequency band

In the "Compensation of envelope in the HFB" shown in Fig. 2, the clipped spectral envelope in the HFB is compensated with that of the clean speech prepared in advance as follows.

$$S_\mathrm{h}(k) \;=\; \eta\cdot S_\mathrm{a}(k)+(1-\eta)\cdot S_\mathrm{c}(k) \quad (0<\eta<1), \tag{8}$$

where $S_\mathrm{h}(k)$ is the spectral envelope in the HFB after the compensation, $S_\mathrm{a}(k)$ is the spectral envelope of the clean speech, $S_\mathrm{c}(k)$ is the spectral envelope of the clipped speech, and η is also the compensation coefficient on the basis of the estimated CR. The higher CR gives the smaller compensation amount. The clean spectral envelope is also prepared in each phoneme of the target speaker because the characteristics of the envelope in the HFB greatly depend on the speaker and phoneme.

5. EVALUATION

The objective and subjective experiments were carried out to evaluate the performance of the clipping-noise suppression using the proposed method for the stationary-noisy speech in a noisy environment. The sound quality of the speech signals was evaluated in these experiments under the conditions that are shown in Tab. 1. As the objective index for evaluating the sound quality, the logarithmic spectral distance (LSD) (4) was employed and it is expressed as follows.

$$\mathrm{LSD} \;=\; -\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(20\log_{10}\frac{|S_\mathrm{r}(k)|}{|S_\mathrm{d}(k)|}\right)^2}, \tag{9}$$

where $S_\mathrm{r}(k)$ and $S_\mathrm{d}(k)$ are the spectra of an original speech and a degraded speech, respectively. k also indicates the frequency bin index. The LSD becomes higher under the condition with the higher sound quality. On the other hand, the mean opinion score (MOS) (5) for five subjects was used as the subjective index for evaluating the sound quality. The subjects evaluated how the speech signal was degraded with five grades (5: imperceptible, 4: perceptible but not annoying, 3: slightly annoying, 2: annoying, 1: very annoying).

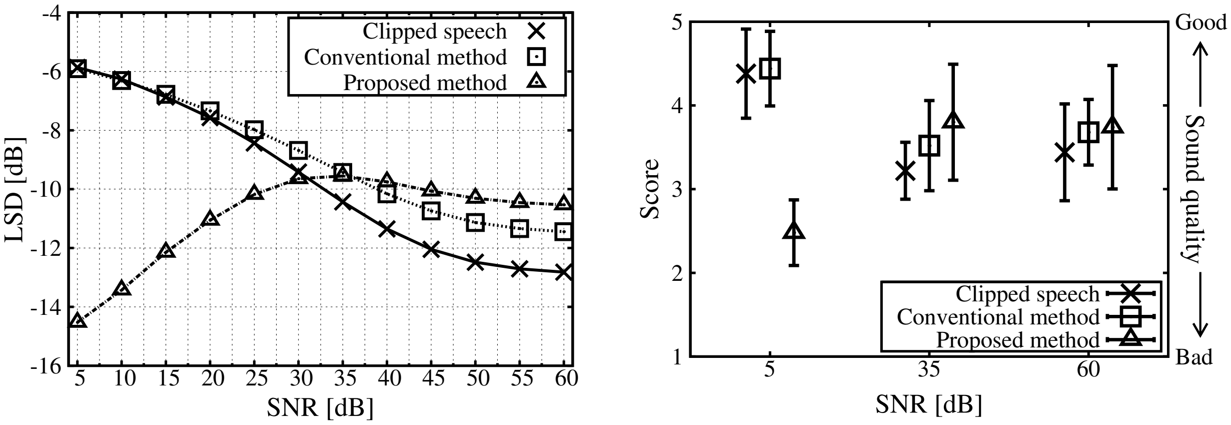
The experimental results are shown in Fig. 3. Horizontal axises in these two figures represent SNR between a clean speech sample and a stationary-noise, and vertical axises in Figs.3 (a) and 3 (b) represent LSD and MOS, respectively. In Fig. 3, the propose method achieved the higher LSD and MOS under the higher SNR condition (higher than 35 dB). These results indicated that the clipping-noise was suppressed using the proposed method in comparison with the conventional one. On the other hand, the performance using the proposed method degraded under the lower SNR conditions. It may be caused by simultaneously suppressing the clipping-noise and the white noise when the spectral envelope of the speech is compensated by the proposed method. We considered that the suppression performance would be improved by switching the conventional and proposed methods, depending on the SNR condition.

6. CONCLUSIONS

In this paper, we evaluate the method for the clipping-noise suppression for the stationary-noisy speech based on the spectral compensation in a noisy environment. We carry out evaluation experiments to evaluate

Table 1 – Experimental conditions

Number of speaker	Two female and three male speakers
Content of speech	Isolated vowels (/a/, /i/, /u/, /e/, /o/)
Sampling	16 kHz / 16 bit
Clipping ratio	0.5
FFT length	1024 points
Frame length	32 ms (512 points)
Shift length	4 ms (64 points)
Noise	White noise
SNR	5 ∼ 60 dB



(a) Result of objective evaluation

(b) Result of subjective evaluation

Figure 3 – Experimental results for noisy speech

the sound quality of the speech signal that is processed by the proposed method. As a result, we confirmed that the clipping-noise was efficiently suppressed under the lower SNR condition using the proposed method in comparison with the conventional one. In the future, we intend to propose the method by switching the conventional and proposed methods, depending on the SNR condition.

ACKNOWLEDGEMENTS

This work was partly supported by a Grant-in-Aid for Scientific Research funded by MEXT and a Grant-in-Aid for JSPS Fellows funded by JSPS.

REFERENCES

1. A. Dahimene, M. Noureddine and A. Azrar, “A simple algorithm for the restoration of clipped speech signal,” Informatica, vol. 32, pp. 183-188, 2008.

2. M. Hayakawa, M. Morise, M. Nakayama and T. Nishiura, “Restoring Clipped Speech Signal Based on Spectral Transformation of Each Frequency Band, ” Acoustics 2012, Paper Number: 4aSP10, May 2012.

3. P. Zolfaghari and T. Robinson, “Formant analysis using mixture of Gaussians,” Proc. ICSLP, pp. 1229-1232, 1996.

4. T. T. Vu, M. Unoki and M. Akagi, “An LP-based blind model for restoring bone-conducted speech,” Proc. ICCE2008, pp. 212-217, 2008.

5. ITU-T Recommendation P. 800, “Methods for subjective determination of transmission quality,” 1996.