Adaptive Post-Filtering Controlled by Pitch Frequency for CELP-based Speech Coder

Hironobu Chiba *, Yutaka Kamamoto†, Takehiro Moriya†, Noboru Harada†
Shigeki Miyabe*, Takeshi Yamada* and Shoji Makino*
*Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
Email:chiba@mmlab.cs.tsukuba.ac.jp
†NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Japan

***Abstract*—Most speech codecs utilize a post-filter that emphasizes pitch structures to enhance perceptual quality at the decoder. Particularly, the bass post-filter used in ITU-T G.718 performs an adaptive pitch enhancement technique for a lower fixed frequency band. This paper describes a new post-filtering method in which the bass the frequency band and the gain are adaptively controlled frame-by-frame depending on the pitch frequency of decoded signal to improve bass post-filter performance. We have confirmed the improvement of the speech quality with the developed method through objective and subjective evaluations.**

I. Introduction

Speech codecs, especially CELP (Code Excited Linear Prediction)-based standard technologies, apply a post-processing to the decoded signal in order to enhance the perceptual quality [1], [2], [3]. One example of the processing is a post-filter that emphasizes the formants and the pitch structures [4], [5]. The ITU-T G.718 decoder, called the bass post-filter for wideband speech sampled at 16 kHz, applies such a post-filter to the lower frequency band, while conventional post-filters have been applied up to the Nyquist frequency for narrow band signal sampled at 8 kHz [6]. This paper presents an adaptive control by which bandwidth for the post-filter is applied depending on the observed pitch frequency (F_0 or fundamental frequency). The conventional method implemented in G.718 uses the fixed cut-off frequency to apply the post-filtering. In contrast, the developed approach adaptively controls the cut-off frequency frame-by-frame. To assess the quality improvements, we conducted ITU-T P.862 PESQ (Perceptual Evaluation of Speech Quality) [7] and ITU-R BS. 1534 MUSHRA (MUlti Stimulus test with Hidden Reference and Anchor) [8] experiments. The obtained test results show that the developed method statistically enhances the decoded speech signal. This method is applicable to most speech codecs based on CELP.

II. Processing of bass post-filter

Figure 1 shows a block diagram of the bass post-filter used in G.718. This post-filter emphasizes the pitch structure of the decoded signal, and $\hat{s}(n)$ is limited to the lower frequency band for each 10-ms frame (hereafter referred to as a processing frame).

First, $s_p(n)$ is calculated by the following equation.

$$s_p(n) = 0.5\hat{s}(n-\tau) + 0.5\hat{s}(n+\tau) \quad (1)$$

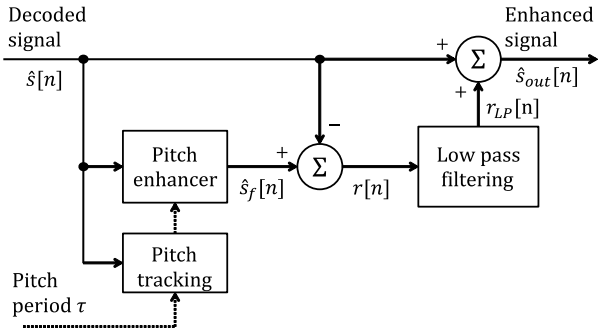

Fig. 1. Block diagram of bass post-filter used in G.718.

where τ is the pitch period obtained by the pitch tracking part. This equation means that the emphasized signal, $s_p(n)$, is generated from a two-sided long-term prediction. In order to allow proper operation of the pitch prediction of equation (1) in all cases, the unavailable data is extrapolated according to the following rule:

$$\hat{s}(n+L) = \hat{s}(n+L-\tau) \quad (2)$$

Here, L is set to a proper value needed by the pitch prediction.

Second, the intermediate signal $s_f(n)$ and the pitch-emphasized over the broad-band frequency signal $r(n)$ is obtained from the following procedures:

$$r(n) = \hat{s}_f(n) - \hat{s}(n) \quad (3)$$
$$\hat{s}_f(n) = (1-\alpha)\hat{s}(n) + \alpha s_p(n) \quad (4)$$

Here, the gain factor α is given by

$$\alpha = \frac{C_p}{0.5(E_p + 10^{0.1\bar{E}_{pp}})} \quad (5)$$

where C_p is the inner product between $\hat{s}(n)$ and $s_p(n)$, E_p is the energy of the predicted signal $s_p(n)$, and $\bar{E}_{pp}$ is the mean prediction error of energy in decibels in the present 5-ms coding frame. Note that α, which is computed by equation (5), is constrained by the following condition.

$$\alpha = \begin{cases} 0.5 & (\alpha > 0.5) \\ 0.0 & (\alpha < 0.0) \end{cases} \quad (6)$$

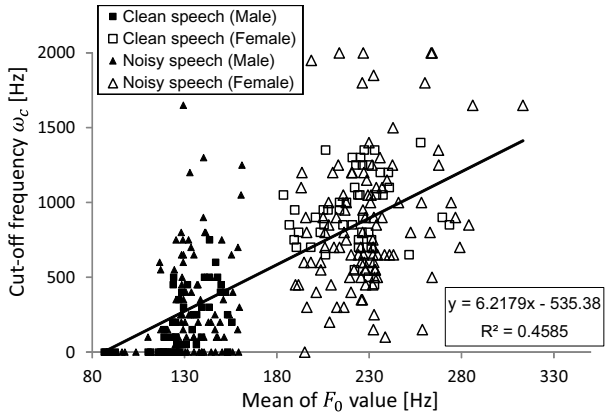

Fig. 2. Relationship between the cut-off frequency that achieves maximum PESQ score and the mean of F_0 value.

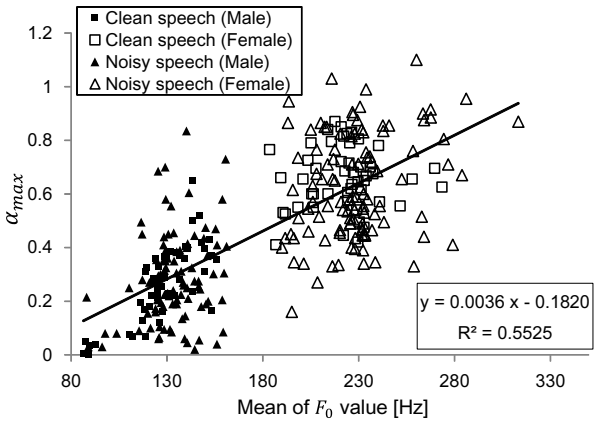

Fig. 3. Relationship between the upper limit of bass post-filter gain that gave the maximum PESQ score and the mean of F_0 value.

Then, the signal $r_{LP}(n)$ is obtained by applying a low-pass filter to $r(n)$, the same as G.718. In other words, $r_{LP}(n)$ is pitch-emphasized signal applied only for low frequency region.

$$r_{LP}(n) = b(0)r(n) + \sum_{k=1}^{16} b(k)(r(n-k) + r(n+k)) \qquad (7)$$

Here, $b(n)$ is a linear phase symmetric FIR (finite impulse response) low-pass filter.

Finally, decoded signal emphasized only for low-frequency region $\hat{s}_{out}(n)$ is computed by the following equation using $\hat{s}(n)$ and $r_{LP}(n)$:

$$\hat{s}_{out}(n) = \hat{s}(n) + r_{LP}(n). \qquad (8)$$

III. Enhancement of bass post-filter

A. Adaptive control of parameters of bass post-filter depending on pitch frequency

By means of emphasizing the harmonic structure of decoded signal as a time-variant filter using pitch frequency in a current frame, the bass post-filter can reduce the inter-harmonic noise in the decoded signal. However, the parameters of this filter do not depend on pitch frequency directly. Generally, the perceptual quality of a high-pitch decoded signal is worse than a low-pitch one in the low bit-rate CELP case. Higher-pitch speech tends to produce more quantization noise due to the CELP algorithm. Therefore, we propose that some parameters in the bass post-filter should depend on pitch frequency.

To obtain the rules for how to adaptively control the parameters in each processing frame, we investigated the correspondence estimation between a certain parameter of the filter that gave the best PESQ score, where we defined the mean opinion score estimated by PESQ as the PESQ score, and a pitch frequency value averaged over each input file. We used 100 files of clean speech and 200 files of noisy speech for the pilot study. The duration of each file was 8 s and each sentence-pair was spoken by males and females in various languages. The SNR (signal to noise ratio) of noisy speech using car noise or office noise was set to 15 or 20 dB. The mean value of the pitch frequency of each file averaged over an item was computed with STRAIGHT [9] as F_0.

B. Preferable cut-off frequency of pitch enhancement in relation to F_0

The cut-off frequency of the low-pass filter defined by equation (7) is fixed to 500 Hz in G.718. Even though pitch characteristics differ between female and male speech, the same cut-off frequency is applied. The number of harmonic pulses may affect human perception, so we try to control the cut-off frequency of the low-pass filter in bass post-filter depending on the pitch frequency.

Figure 2 shows relationship between the cut-off frequency that gives the highest PESQ score and the mean of F_0 value averaged over each speech item. Each plot corresponds to each speech item. From these results, we can observe that it is possible to improve the perceptual quality of high-pitch speech such a female's voice by utilizing a low-pass filter with a cut-off frequency higher than the 500 Hz used in G.718. Therefore, we drew a linear regression line that gives a suitable cut-off frequency depending on pitch frequency F_0 as follows:

$$\omega_c = 6.2179F_0 - 535.38 \qquad (9)$$

Here, R^2 is the coefficient of determination of this regressive line. It is expected that sound quality can be improved by the adaptive frequency band to emphasize the pitch using the regressive line in the bass post-filter.

C. Preferable gain of pitch enhancement in relation to F_0

According to equation (5), the conventional method limits the upper bound of the gain α to $\alpha_{max} = 0.5$. However, we found that the preferable gain limit was also dependent on the pitch frequency when the best PESQ scores were searched among the pitch-frequencies by the brute-force method. So we investigated the upper gain limit that gives the highest PESQ score at a certain pitch frequency.

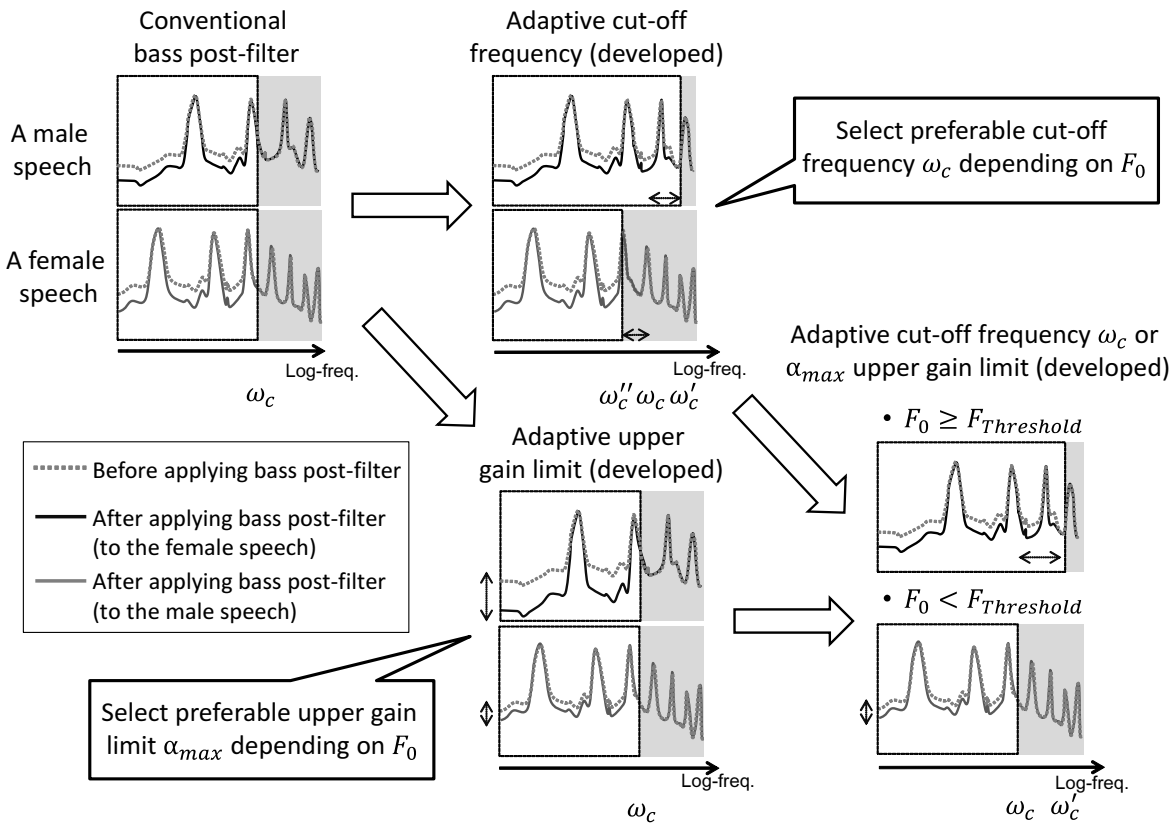

Fig. 4. Effectiveness of conventional and developed bass post-filter.

Figure 3 shows relationship between the upper limit of bass post-filter gain that achieves the highest PESQ score and the mean of the F_0 value averaged over each speech item. From these results, we can observe the tendency of gain limit and F_0. We can imagine that it is possible to improve the perceptual quality of low-pitch speech such a male's voice by utilizing a smaller value of the upper gain α_{max} than the 0.5 used in G.718. Therefore, we drew a regression line that gives suitable gain limit depending on pitch frequency F_0 as follows:

$$\alpha_{max} = 0.0036F_0 - 0.1820 \tag{10}$$

It is expected that sound quality is improved by the adaptive upper limit of gain in pitch emphasis using the regressive line.

D. Implementation of adaptive control of cut-off frequency and gain depending on pitch frequency

For each 10 ms of a processing frame, pitch frequency $\hat{F}_0$ is computed using the corrected pitch period used at the decoder, τ, as

$$\hat{F}_0 = \frac{F_s}{\tau} \tag{11}$$

where sampling frequency F_s is 16 kHz. Utilizing this $\hat{F}_0$, we implemented adaptive cut-off frequency $\hat{\omega}$ and maximum gain $\hat{\alpha}_{max}$ on the basis of equations (9) and (10). Note that $\hat{\alpha}_{max}$ is set to 0 if $\hat{\alpha}_{max}$ is a negative value. Moreover, cut-off frequency quantized at 50 Hz intervals is set using the following equation:

$$\hat{\omega}_c = 50\lfloor \frac{6.2179\hat{F}_0 - 535.38}{50} \rfloor \tag{12}$$

Here, $\lfloor \cdot \rfloor$ is a flooring function, and the range of cut-off frequency is constrained by $0 \leq \hat{\omega}_c \leq 2000$. Then, the bass post-filter is bypassed in the processing frame in which $\hat{\omega}$ is 0. Utilizing $\hat{\omega}$ selected with equation (12), the impulse response of the low-pass filter is selected from the look-up table. Because of the above implementation, the increased complexity comprises only a few operations per frame at the decoder side and additional ROM for a set of tables of the low-pass filter coefficients.

IV. Experimental evaluation

A. Objective evaluation with PESQ

In order to assess the quality, we carried out the subjective and objective evaluations at 8 kbit/s for 16 kHz sampling

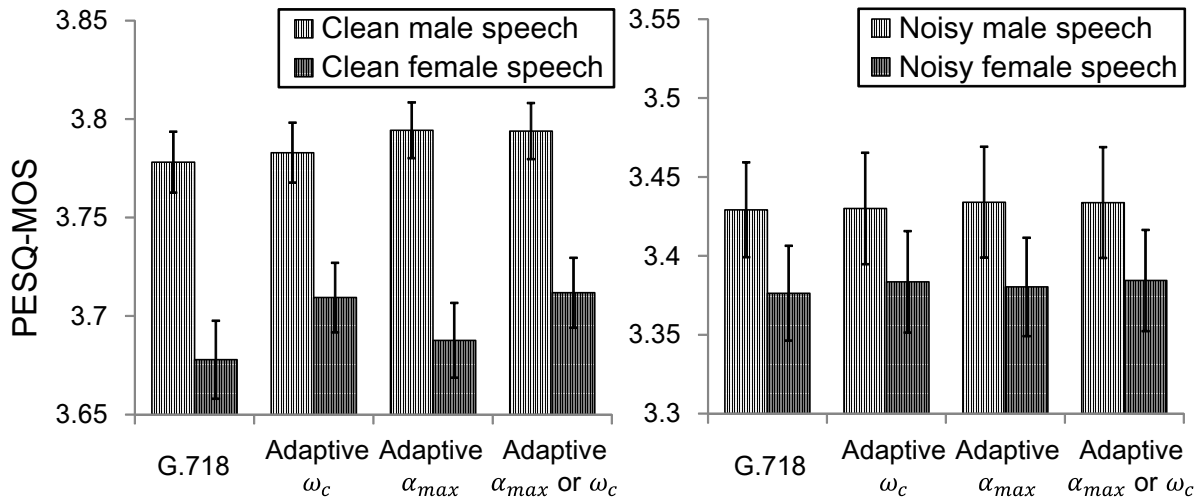

Fig. 5. Objective evaluation of speech quality with wideband PESQ.

input. First, we carried out the objective evaluation with PESQ. Tested files were 231 clean speech sentence-pairs and 200 noisy speech sentence-pairs (SNR: 15 dB and 20 dB; car and office noises). This data set is different from the one used in the experiments in section 3. We evaluated four types of bass post-filter: 1) Fixed cut-off frequency and fixed upper gain limit (G.718), 2) adaptive cut-off frequency and fixed upper gain limit (Adaptive ω_c), 3) fixed cut-off frequency and adaptive upper gain limit (Adaptive α_{max}), and 4) adaptive cut-off frequency or adaptive upper gain limit (Adaptive α_{max} or ω_c). In three of the developed methods either the cut-off frequency or upper gain limit is adaptive; the other not selected is fixed. The adaptive parameter is decided depending on the value of pitch frequency F_0 in a current frame. In this paper, the adaptive cut-off frequency is used if the value of F_0 is more than 220 Hz; otherwise, the adaptive upper gain limit is used.

The results of the PESQ evaluation are shown in Figure 6, where the value of the bar graph represents the mean of the PESQ score, and the error bar is the 95% confidence interval. These results confirmed that that Adaptive ω_c and Adaptive α_{max} improved the PESQ score for female or male speech, compared with G.718 (in which these parameters are fixed), respectively. Moreover, Adaptive α_{max} or ω_c improved the PESQ score of both female and male speech. The above PESQ evaluation shows that the developed method statistically improves the quality for clean speech materials. In contrast to the improvements described above, we cannot observe significant enhancements in any of the conditions for noisy speech as shown in Fig. 6.

B. Subjective evaluation with MUSHRA

We performed a subjective evaluation with MUSHRA [8] for the developed method Adaptive α_{max} or ω_c. Six test items for the MUSHRA experiment were chosen from the outliers of the PESQ evaluation.

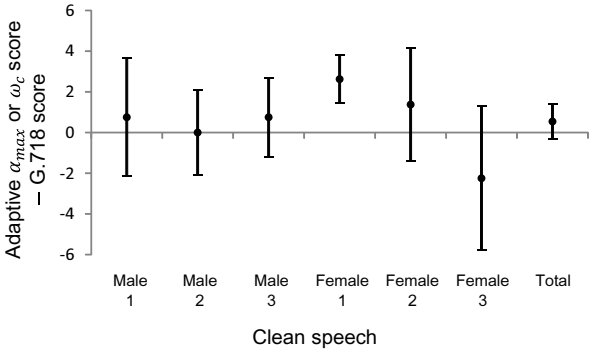

Fig. 6. Subjective evaluation of speech quality with MUSHRA.

In the experimental results, as represented in Fig. 6, we observed that one item out of six was enhanced statistically, no item was significantly degraded, and the averaged score was also positive compared with G.718.These subjective and objective evaluations confirmed that new method improves the quality of decoded signal that has been encoded by G.718.

V. Conclusion

This paper described a new post-filtering method in which the frequency band and gain are adaptively controlled frame by frame depending on the estimated pitch frequency of the decoded signal to improve the performance of the bass post-filter. We have observed simple linear dependence of the cut-off frequency and gain limit that give the best PESQ scores depending on the pitch frequency.

We have experimentally drawn the regression lines and adaptively controlled the cut-off frequency and gain limit on the basis of them. We have confirmed the improvement of the

speech quality with the developed method through an objective evaluation by wideband PESQ and a subjective evaluation by MUSHRA, compared with the ITU-T G.718 standard. By means of the adaptive cut-off frequency band, we observed significant quality enhancement in PESQ for female clean speech. By means of the adaptive limit of gain of post-filter, we observed significant quality enhancement in PESQ for male clean speech. For other conditions, including for noisy speech signals, we never encountered quality degradation in PESQ. We have also found a female speech item that is significantly enhanced in subjective listening tests. This method is applicable not only for G.718 but also for most speech codecs based on CELP for 16 kHz sampled input.

REFERENCES

[1] W. B. Kleijn and K. K. Paliwal, "Speech Coding and Synthesis," Elsevier Science, 1998.

[2] A. M. Kondoz, "Digital Speech: Coding for Low Bit Rate Communication Systems, Second Edition," John Wiley and Sons Ltd., 2004.

[3] J. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," IEEE Trans. on Speech and Audio Proc., vol. 3, no. 1, pp.59-71, Jan. 1998.

[4] ITU-T Recommendation G.729, Mar. 1996.

[5] 3GPP TS 26.171, Adaptive Multi-Rate Wideband (AMR-WB), Apr. 2001.

[6] ITU-T Recommendation. G.718, Jun. 2008.

[7] ITU-T Recommendation. P.862.2, Jul. 2007.

[8] ITU-R Recommendation. BS.1534-1, Mar. 2003.

[9] H. Kawahara, H. Katayose, A. de Cheveigne and R. D. Patterson, "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," Proc. EUROSPEECH, vol. 6, pp.2781-2784, Sep. 1999.