Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Manipulating Vocal Signal in Mixed Music Sounds using Side Information based on the Fundamental Frequency

Akinori Ito and Yuto Sasaki

Graduate School of Engineering
Tohoku University
6-6-05 Aramaki aza Aoba, Sendai, 980-8579 Japa
aito@spcom.ecei.tohoku.ac.jp

Received February 2017; revised May 2017

Abstract. *We propose a system that enables a listener of streaming audio to control the volume (magnitude of the signal) of independent part (specifically the vocal signal) in a mixed audio signal in real-time. In the proposed method, fundamental frequency (F0) of the vocal signal is used as side information. F0 information is estimated from the target signal before being mixed with backing track signals. After receiving the mixed music signal, vocal sound manipulation is performed using a comb filter using F0 information. In addition to the F0 information, we added side information considering the ratio between the level of the signal to be manipulated and the backing signal. As an experimental result, we obtained that the proposed method improved the quality of the manipulated signal compared with sending the information of vocal signal using the existing MP3 encoder.*
Keywords: Music signal manipulation, Encoding, Singing voice, Comb filter

1. **Introduction.** With the development of information and communication technologies, listening to music on the Web becomes popular. Naturally, needs for actively manipulating the music have been increased [1], such as controlling volume (magnitude of the signal) of a specified instrument to listen to music or practice specific instrument of a song. If a part of a specific instrument can be manipulated, the system can help listeners enjoying music. For example, you can make your own Karaoke tracks of a music signal by muting the vocal signal, or remix a music by manipulating the instrumental parts.

Most of conventional researches on music sound manipulation have taken an approach based on sound source separation that use only mixed music sound [2, 3, 4, 5]. Although the complete separation of mixed music signal is very difficult problem, quality of the separation can be improved under some assumption, such as we know the melody of the music to be separated. The score-informed source separation method [6, 7] assumes that the score of the music can be exploited when we separare the signal into sounds of individual instruments. Problem of these approaches is that these methods cannot be applied to a real-time application because these methods segregate the music signal into each part considering whole signal and then remix all parts after the segregation.

In most cases except live recording, complete sound sources are made by mixing each sound of instrument and vocal that is recorded individually in a studio. This means that the producer has sound tracks of all individual parts. We can manipulate desired music signals in real-time having these sound tracks of the parts. Parvaix et al. proposed a framework to embed the sounds of individual instruments into the cover music signal

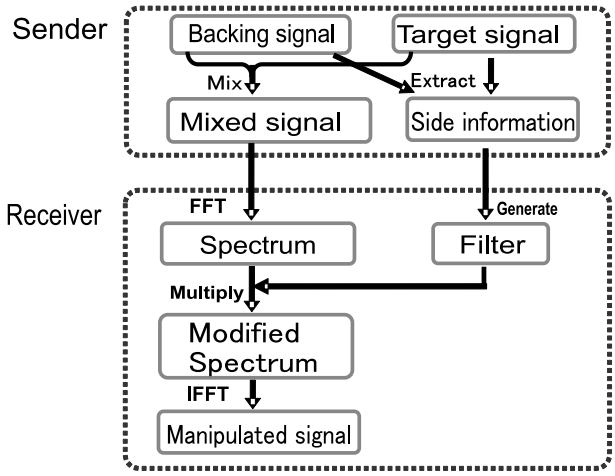

FIGURE 1. Overview of the proposed system

using the watermarking technology [8, 9]. Their method embeds the quantized MDCT (Modified Discrete Cosine Transform) coefficients into the cover signal, which is very similar to sending the original signal itself. Drawback of their method is that the amount of side information is large, and sometimes it is comparable to the information of the cover signal itself.

In this paper, we focus on a method for manipulating vocal track of a mixed music signal using side information. In particular, our aim is to reduce the amount of side information and improve quality of sound after the manipulation. We use F0 information extracted from the target signal as side information [10, 11]. Then, we compose a comb filter from the F0 to manipulate the vocal signal selectively. Next, we designed another comb filter that considers the backing sound signal (sound of instrumental accompaniment). We then evaluate the performance of the system using SNR and PEAQ [12] according to different conditions. Finally, we discuss the amounts of information and quality of the manipulated sound signal of these systems.

2. **Overview of the system.** Fig. 1 shows a block diagram of the system for manipulating music signal using F0 information of the target. In the sending side, the target (vocal) sound signal $v(t)$ and the backing sound signal $b(t)$ are mixed to make the mixed music signal.

$$i(t) = b(t) + v(t) \tag{1}$$

By converting the signal into the time-frequency domain, the signal can be expressed as

$$I(f, \tau) = B(f, \tau) + V(f, \tau) \tag{2}$$

where f is the frequency and τ is the frame number.

The side information is extracted from the target signal in advance. The side information and the mixed music signal are transmitted to the receiver together. In the receiving side, a filter $G(f, \tau)$ is formed using the side information frame by frame. Using this filter, the mixed music sound signal $I(f, \tau)$ is adjusted by multiplying spectrum of the signal by the filter.

$$O(f, \tau) = G(f, \tau)I(f, \tau) \tag{3}$$

Finally, manipulated waveform $o(t)$ was obtained by transforming manipulated spectrum $O(f, \tau)$ into time domain.

3. **Filter Design.**

3.1. **Theoretical background.** As mentioned before, the mixed music signal $I(f)$ can be expressed as Eq. (2). The goal of our sound manipulation is that, given an amplification factor $A \geq 0$, calculate the manipulated signal

$$O'(f,\tau;A) = B(f,\tau) + AV(f,\tau). \tag{4}$$

It can be easily achieved if we have the vocal signal $V(f,\tau)$, by

$$O'(f,\tau;A) = I(f,\tau) + (A-1)V(f,\tau). \tag{5}$$

However, sending $V(f,\tau)$ itself as side information consumes too large bandwidth compared with the mixed music signal. Therefore, we make several assumptions for making the amount of side information smaller.

As $V(f,\tau)$ is a vocal signal, most part of it has harmonic structure with fundamental frequency $f_0(\tau)$. Then we approximate the vocal signal as

$$V(f,\tau) \approx \sum_{k=1}^{K_0} a_k \varphi(f - kf_0(\tau);\sigma_0) \tag{6}$$

where

$$\varphi(f;\sigma) = \exp\left(-\frac{f^2}{2\sigma^2}\right) \tag{7}$$

and $\sigma_0 \ll f_0$. Here, K_0 is the maximum number of harmonic components and a_k is the magnitude of the k-th harmonic component. This means that we approximate individual line spectrum of the signal using the Gaussian function. Here, we can write

$$I(f,\tau) = B(f,\tau) + \sum_{k=1}^{K_0} a_k \varphi(f - kf_0(\tau);\sigma_0). \tag{8}$$

Next, let $G(f,\tau)$ be

$$G(f,\tau) = 1 + (A-1)G_0(f,\tau) \tag{9}$$

$$G_0(f,\tau) = \sum_{k=1}^{K} \varphi(f - kf_0(\tau);\sigma) \tag{10}$$

where $K \leq K_0$ and $\sigma \ll f_0(\tau)$. G is a frequency response of a comb-like filter that emphasizes the harmonic components of the vocal signal by the amplification factor A and passes other frequency components unchanged. By multiplying $G(f,\tau)$ to $I(f,\tau)$,

$$G(f,\tau)I(f,\tau) = B(f,\tau) + V(f,\tau) + (A-1)G_0(f,\tau)(B(f,\tau) + V(f,\tau)). \tag{11}$$

Now, considering that $\varphi(f;\sigma)$ is almost zero when $|f| > \sigma$ and $\sigma, \sigma_0 \ll f_0(\tau)$,

$$\varphi(f - k_1 f_0(\tau);\sigma_0)\varphi(f - k_2 f_0(\tau);\sigma) \approx 0 \tag{12}$$

when $k_1 \neq k_2$. Therefore,

$$\begin{aligned} G_0(f,\tau)V(f,\tau) &= \left(\sum_{k=1}^{K_0} a_k \varphi(f - kf_0(\tau);\sigma_0)\right)\left(\sum_{k=1}^{K} \varphi(f - kf_0(\tau);\sigma)\right) && (13)\\ &\approx \sum_{k=1}^{K} a_k \varphi(f - kf_0(\tau);\sigma_0)\varphi(f - kf_0(\tau);\sigma) && (14)\\ &= \sum_{k=1}^{K} a_k \varphi\left(f - kf_0(\tau); \frac{\sigma_0\sigma}{\sqrt{\sigma_0^2 + \sigma^2}}\right). && (15) \end{aligned}$$

When $\sigma \gg \sigma_0$,

$$G_0(f,\tau)V(f,\tau) = \sum_{k=1}^{K} a_k \varphi\left(f - kf_0(\tau); \frac{\sigma_0\sigma}{\sqrt{\sigma_0^2+\sigma^2}}\right) \quad (16)$$

$$\approx \sum_{k=1}^{K} a_k \varphi\left(f - kf_0(\tau); \sigma_0\right) \quad (17)$$

$$= V(f,\tau) - \sum_{k=K+1}^{K_0} a_k \varphi\left(f - kf_0(\tau); \sigma_0\right). \quad (18)$$

Therefore, assuming the second term of (18) to be small (that means that the components of the higher harmonics is sufficiently smaller than the lower components),

$$G(f,\tau)I(f,\tau) \approx B(f,\tau) + V(f,\tau) + (A-1)V(f,\tau) + (A-1)G_0(f,\tau)B(f,\tau) \quad (19)$$

$$= B(f,\tau) + AV(f,\tau) + (A-1)G_0(f,\tau)B(f,\tau) \quad (20)$$

$$= O'(f,\tau;A) + (A-1)G_0(f,\tau)B(f,\tau) \quad (21)$$

Therefore, if $G_0(f,\tau)B(f,\tau)$ is sufficiently small, $I(f,\tau)G(f,\tau)$ can be an approximation of the objective signal, $O'(f,\tau;A)$. This assumption holds only if $V(f,\tau)$ and $B(f,\tau)$ do not share the same harmonic components.

To reduce the error of the manipulated signal, we can consider the generalized filter, as follows:

$$G(f,\tau) = 1 + (A-1)\left(\sum_{k=1}^{K} \beta_k(\tau)\varphi(f - kf_0(\tau);\sigma)\right) \quad (22)$$

Here, $0 \le \beta_k(\tau) \le 1,\ k = 1,\ldots,K$ are the coefficients introduced to reduce the errors. When using the non-uniform $\beta_k(\tau)$, we need to transmit the information of $\beta_k(\tau)$ in addition to $f_0(\tau)$.

3.2. **Filter with uniform β_k.** The simplest way of realizing $G(f,\tau)$ is to use uniform β_k, i.e., $\beta_k = 1$ for $k = 1,\ldots,K$. In this case, we use the following filter function.

$$G(f,\tau) = 1 + (A-1)\sum_{k=1}^{K} \varphi(f - kf_0(\tau);\sigma) \quad (23)$$

This is a simple comb filter. If we fix the values of σ and K, information we transmit is only $f_0(\tau)$ for each frame (A is given at the receiver side). Figure 2 shows an example of the simple comb filter.

3.3. **Selective β_k.** The approach based on the simple comb filter assumes that $B(f,\tau)$ is sufficiently small around the harmonic comonents of $V(f,\tau)$. However, in reality, $B(f,\tau)$ and $V(f,\tau)$ often share the same harmonic components, which causes degradation of the manipulated signal. Our idea for improving the quality of manipulated signal is to avoid manipulating frequency components if $B(f,\tau)$ is not negligible compared with $V(f,\tau)$.

$$\beta_k(\tau) = \begin{cases} 1 & \text{if } |V(kf_0(\tau),\tau)| \ge |B(kf_0(\tau),\tau)| \\ 0 & \text{otherwise} \end{cases} \quad (24)$$

$$G(f,\tau) = 1 + (A-1)\left(\sum_{k=1}^{K} \beta_k(\tau)\varphi(f - kf_0(\tau);\sigma)\right) \quad (25)$$

When using this filter, the information to be transmitted is $f_0(\tau)$ and $\boldsymbol{\beta}(\tau) = (\beta_1(\tau),\ldots,\beta_K(\tau))$, which is K bit larger than that of the simple comb filter. Figure 3 shows an example of the selective comb filter.

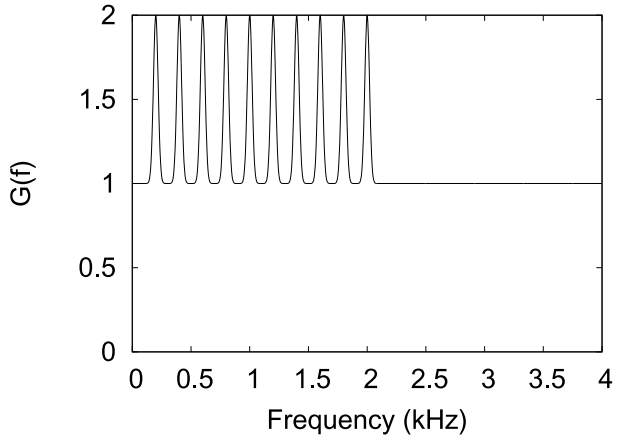

FIGURE 2. An example of a filter with uniform β_k (f_0=200Hz, σ=20Hz, $K = 10$)

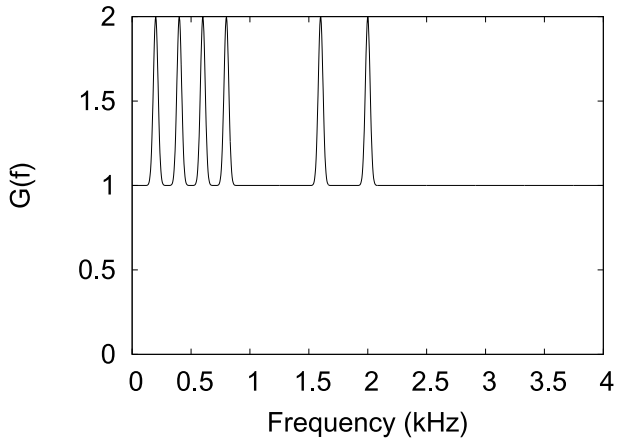

FIGURE 3. An example of a filter with selective β_k (f_0=200Hz, σ=20Hz, $K = 10$)

3.4. **Optimum β_k.** We can optimize β_k so that the manipulated signal to be as similar to the target signal as possible. The target signal is

$$O'(f,\tau;A) = I(f,\tau) + (A-1)V(f,\tau), \tag{26}$$

whereas the manipulated signal is

$$\begin{aligned} G(f,\tau)I(f,\tau) &= I(f,\tau) + (A-1)G_0(f,\tau)I(f,\tau) &(27)\\ &= I(f,\tau) + (A-1)\left(\sum_{k=1}^{K}\beta_k(\tau)\varphi(f-kf_0(\tau);\sigma)\right)I(f,\tau). &(28)\end{aligned}$$

Comparing (26) and (28), it is obvious that $G(f,\tau)I(f,\tau)$ coincides $O'(f,\tau;A)$ when

$$V(f,\tau) = \left(\sum_{k=1}^{K}\beta_k(\tau)\varphi(f-kf_0(\tau);\sigma)\right)I(f,\tau). \tag{29}$$

However, it is impossible to achieve this relation for all f. Therefore, we determine $\beta_k(\tau)$ so that this relation holds at the multiple of $f_0(\tau)$, i.e.

$$V(nf_0(\tau),\tau) = \left(\sum_{k=1}^{K}\beta_k(\tau)\varphi(nf_0(\tau)-kf_0(\tau);\sigma)\right)I(nf_0(\tau),\tau). \tag{30}$$

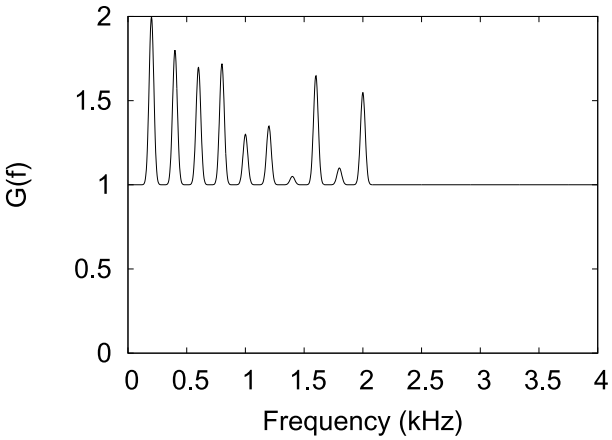

FIGURE 4. An example of a filter with optimum β_k (f_0=200Hz, σ=20Hz, $K = 10$)

$$V(nf_0(\tau),\tau) = \sum_{k=1}^{K} \beta_k(\tau)\varphi(nf_0(\tau) - kf_0(\tau);\sigma)(V(nf_0(\tau),\tau) + B(nf_0(\tau),\tau)) \tag{31}$$

for $n = 1, ..., K$. Since $\varphi(nf_0(\tau) - kf_0(\tau);\sigma)$ is almost zero when $n \neq k$, we obtain

$$V(nf_0(\tau),\tau) \approx \beta_n(\tau)(V(nf_0(\tau),\tau) + B(nf_0(\tau),\tau)) \tag{32}$$

and thus

$$\beta_n(\tau) = \frac{V(nf_0(\tau),\tau)}{V(nf_0(\tau),\tau) + B(nf_0(\tau),\tau)}. \tag{33}$$

Using this relation, $\beta_n(\tau)$ becomes a complex number, which requires twice as much side information as a real number. To reduce the side information, we use a real number as $\beta_n(\tau)$. If we assume that $V(f,\tau)$ and $B(f,\tau)$ are not correlated, we obtain

$$\beta_n(\tau) = \frac{|V(nf_0(\tau),\tau)|}{|V(nf_0(\tau),\tau)| + |B(nf_0(\tau),\tau)|}. \tag{34}$$

Fig. 4 shows an example of $G(f,\tau)$ generated by the proposed method. This formulation means that we need to transmit $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \ldots, \beta_K(\tau))$ in addition to F0 as side information. As the information to be transmitted increases, we need to investigate the trade-off between the bit-rate of the side information and signal quality.

4. **Experiment.**

4.1. **Overview.** We conducted an experiment to compare the sound quality of the manipulated signal. We used the following three types of filters:

(A) Simple comb filter: $\beta_k = 1$
(B) Comb filter of component selection: Eq. (24)
(C) Comb filter of component magnitude calculation: Eq. (34)

We used four music clips as the materials of the experiment. Table 1 shows the music signals used in the experiment.

We optimized the window width of short-time Fourier transform, and component width σ *a posteriori*. The window width was changed from 20 ms to 120 ms, and σ was changed from 20 Hz to 360 Hz. We carried out experiments for all combination of window width and σ, and chose the parameter values that gave the best result.

Frame shift was set to the half of the window size. The F0 of the vocal signal was estimated frame by frame using the fund function of the seewave package [13]. We used

TABLE 1. Music clips

No.	Title	Length (s)	Gender of the singer
001	Senbon Zakura	6.8	Female
002	Ihoujin	11.0	Female
003	Konayuki	5.6	Male
004	Yuki no Hana	7.0	Female

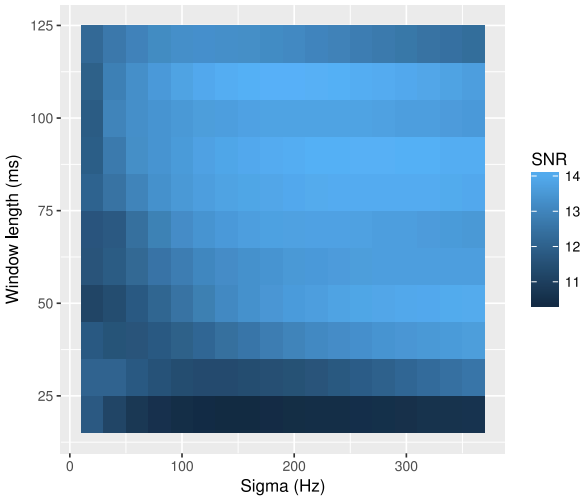

FIGURE 5. Optimization result of σ and window length using filter (C)

TABLE 2. Optimum parameter values

Filter	σ (Hz)	window (ms)	SNR (dB)
(A)	260	90	14.15
(B)	260	90	14.15
(C)	200	110	14.18

signal-to-noise ratio (SNR) as an evaluation metric of manipulated signals, defined as follows.

$$SNR = 10 \log_{10} \frac{\sum_t (b(t) + Av(t))^2}{\sum_t (b(t) + Av(t) - o(t))^2} \tag{35}$$

4.2. **Optimum parameter values.** We optimized the parameter value *a posteriori*, where the target amplitude $A = 2.0$. The number of component K was set to 20. The metric for evaluation was the average SNR.

Figure 5 shows the optimization result of σ and the window length using filter (C). We can see that the optimum σ and the window length are relatively large. Table 2 shows the optimum parameters for all the filters. The optimum parameter values for filter (C) is slightly different from the other filters, but the result of filter (C) for $\sigma = 260$ Hz and window length 90 ms (SNR 14.17 dB) was almost same as the optimum one (SNR 14.18 dB). Thus we use $\sigma = 260$ Hz and window length 90 ms in the later experiments.

4.3. **Result of different filters for different amplitide.** Figure 6 shows the relationship between the amplitude A and SNR. As shown in the figure, the SNR was better when A was near to 1. The SNR should be infinity when $A = 1$. The result of filter (C) was slightly better than the other two filters, especially when $A > 0$.

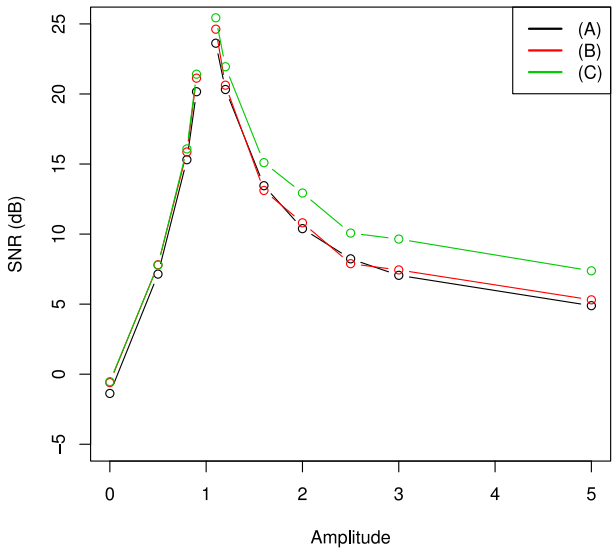

FIGURE 6. Amplitude and SNR

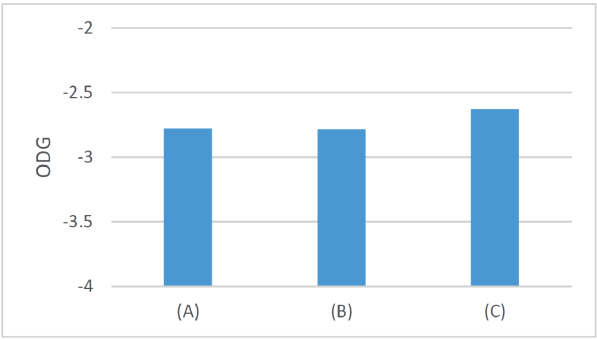

FIGURE 7. Comparison of the three methods using PEAQ ($A = 2.0$)

4.4. **Evaluation using PEAQ.** Next, we compared the three methods using PEAQ (Perceptual Evaluation of Audio Quality) [12]. PEAQ is an objective evaluation index of audio signal that simulates perceptual quality of audio. We used peaqb [14] as the measurement tool. Figure 7 shows the result. The Y-axis of the figure is ODG (Objective Difference Grade), which is the value calculated by PEAQ. The ODG value is between -4 and 0, which is designed to be combatible with the value of the Subjective Difference Grade (SDG). The SDG value is average of subjective evaluation for difference between the original and processed signal, where 0 means "Imperceptible", -1 is "Perceptible, but not annoying", -2 is "Slightly annoying", -3 is "Annoying" and -4 is "Very annoying". As shown, the results by filter (A) and (B) were almost identical, while that by filter (C) was slightly better. The reason why the results by (A) and (B) were worse was that those signal contained musical noise caused by the signal processing, while the noise was smaller in the signal generated by filter (C). The musical noise [15] is the noise induced by signal processing that sounds like instrumental sounds, which appears when manipulation of the signal is too large, In filter (C), strength of manipulation of a specific harmonic component was controlled by changing the filter coefficient, which suppressed the musical noise.

4.5. **Comparison with existing encoders.** Next, we compared the bitrate and signal quality to the existing low-bitrate audio encoder. In this framework, the target signal

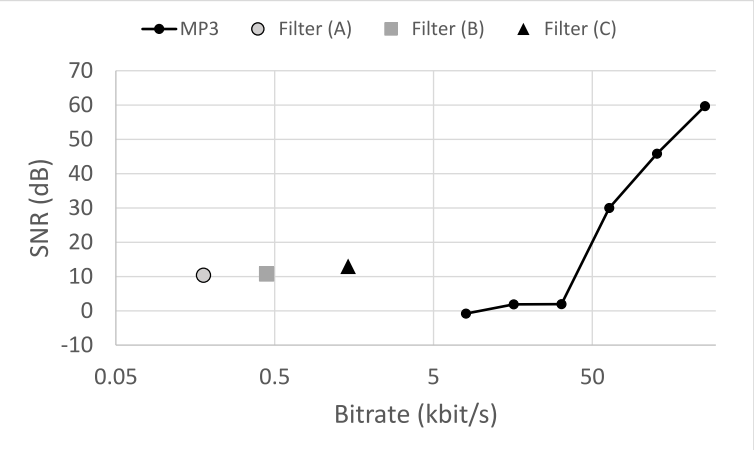

FIGURE 8. Comparison with other encoders (target emphasis $\times 2.0$).

$v(t)$ is compressed by the encoder at the sender side, and the compressed signal is sent as side information. At the receiver side, the mixed signal $i(t)$ as well as the uncompressed target signal $v'(t)$ is received, and the manipulated signal $o(t)$ is simply calculated as

$$o(t) = i(t) + (A-1)v'(t). \tag{36}$$

In the experiment, we employed the MP3 [**?**] encoder. The signal was compressed at 8k, 16k, 32k, 64k, 128k and 256 kbit/s monophonic. Other conditions were same as that of the previous experiments. For all the filters in the proposed method, we quantized F0 value into 8 bit/frame using Lloyd's algorithm. In filter (B), β_k were sent as 20 bit/frame side information, and the side information of filter (C) was 80 bit/frame. The frame rate is the half of the window width shown in Table 2. Thus, for example, bitrate of (C) is $80/(0.110/2) \approx 1455$ bit/s.

Figure 8 shows the result for $A = 2.0$ emphasis. This result revealed that the proposed method achieved reasonable signal quality with much smaller amount of side information compared with the MP3 encoder.

5. **Conclusion.** In this paper, we proposed a design method of comb filter for mixed music signal manipulation using side information. The three methods were compared; the simplest filter was a simple comb filter, while the other two filters exploit information of the target signal and the backing signal. As experimental results, SNR results of the three methods were almost same, while the PEAQ value of the method (C) was the best. The proposed method was also proved to be more efficient from bitrate-quality trade-off point of view than using the existing MP3 encoder.

REFERENCES

[1] M. Goto, Active music listening interfaces based on signal processing, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. IV-1441, 2007.

[2] C. Chafe and D. Jaffe, Source separation and note identification in polyphonic music, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 11, pp. 1298-1292, 1986.

[3] K. Yoshii, M. Goto and H. G. Okuno, INTER:D: a drum sound equalizer for controlling volume and timbre of drums, *Proc. The 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, 2005. EWIMT 2005.*, pp. 205-212, 2005.

[4] Y. Kitano, H. Kameoka, Y. Izumi, N. Ono and S. Sagayama, A sparse component model of source signals and its application to blind source separation, *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4122-4125, 2010.

[5] K. Itoyama, M. Goto, K. Komatani, T. Ogata and H. Okuno, Instrument Equalizer for Query-by-Example Retrieval: Improving Sound Source Separation based on Integrated Harmonic and Inharmonic Models, *Proc. Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 133-138, 2008.
[6] J. Woodruff, B. Pardo and R. Dannenberg, Remixing Stereo Music with Score-Informed Source Separation, *Proc. Int. Society of Music Information Retrieval Conf. (ISMIR)*, pp. 314-319, 2006.
[7] R. Hennequin, B. David and R. Badeau, Score informed audio source separation using a parametric model of non-negative spectrogram, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 45-58, 2011.
[8] M. Parvaix, L. Girin and J.-M. Brossier, A Watermarking-Based Method for Informed Source Separation of Audio Signals with a Single Sensor, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp.1464-1475, 2010.
[9] M. Parvaix and L. Girin, Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1721-1733, 2011.
[10] Y. Sasaki, S.-J. Hahm and A. Ito, Manipulating vocal signal in mixed music sounds using small amount of side information, *Proc. Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, pp. 298-301, 2010.
[11] A. Ito and Y. Sasaki, Manipulation of vocal signal in mixed music signal using side information of F0 and backing spectrum, *Proc. 12th Int. Conf. on Signal Processing (ICSP)*,pp 605-609, 2014.
[12] International Telecommunication Union, Method for objective measurements of perceived audio quality, ITU-R BS. 1387-1, November, 2001.
[13] J. Sueur, Seewave – an R package for sound analysis and synthesis, http://rug.mnhn.fr/seewave/ (Accessed 5 January, 2017).
[14] G. Gottardi, Peaqb, https://github.com/akinori-ito/peaqb-fast (Accessed 2 April, 2015).
[15] M. Berouti, R. Schwartz and J. Makhoul, Enhancement of speech corrupted by acoustic noise, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. 208-211, 1979.