NOVEL SOUND MIXING METHOD FOR VOICE AND BACKGROUND MUSIC

Wataru Owaki, Kota Takahashi

Department of Communication Engineering and Informatics,
The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan

ABSTRACT

A novel sound mixing method, which can mix a voice and background music (BGM) without hiding the voice, is proposed. The proposed method can be used without having to consider the levels of input signals. We implemented this effective method by precise nonlinear processing on the time-frequency plane. We use the name "smart mixer" to refer to the type of sound mixer used in this method. In this paper, we describe the basic concept of the smart mixer and its application to mixing a voice and BGM. Experiments on three sets of real audio signals were performed and the results are discussed.

Index Terms— Sound mixing, time-frequency plane, psychoacoustic model

1. INTRODUCTION

In recent years, sound has become increasingly utilized in everyday life. In a car, both the sound system and the navigation system may transmit information. Sometimes this occurs at the same time, making it difficult to hear both sounds.

When the aim is to only hear a voice, a simple process can be used to mute background music (BGM) while the voice is spoken. However, to hear not only the voice but also the BGM, the process must consist of a complex combination of functions involving equalizers and compressors. Therefore, such complex mixing has only been carried out in a limited range of applications such as broadcasting and professional music production.

Voices and BGM can be distinguished by comparing their time-frequency representations generated by a suitable signal processing structure, such as the outputs of a short-time Fourier transform (STFT). We have developed a new type of sound mixer based on this structure named a smart mixer. The proposed method consists of two key elements. The first element is a newly developed process to determine the contribution of each time-frequency component to the smart mixer. The second element is the adoption of a psychoacoustic model used for this determination. In this paper, we propose the implementation of a smart mixer to ensure voice audibility.

2. RELATION TO PRIOR WORK

In the field of broadcasting, a method of automatic gain control called ducking [1] is used. Ducking simply reduces the gain of BGM when the power of a voice exceeds a threshold. In this approach, the volume attenuation of BGM occurs in principle. On the other hand, the proposed method applies gain reduction to the minimum number of components in the time-frequency distribution essential to hear a voice; therefore, the volume attenuation of BGM is less than that in the case of ducking.

In the field of music production, a technique for coordinating the frequency characteristics of multichannel signals called mirrored equalization [1] is used. The characteristics to which this technique is whereas both a voice and BGM are time-variant. Thus, it tends to be suboptimal in each short time duration. On the other hand, the proposed method can be applied to time-variant characteristics.

Cross-adaptive digital audio effects [2], [3] focus on automating the adjustment of parameters in the conventional mixing method. Often its analysis and processing units are independent, and these representations differ from each other. In automatic equalization using this method [4], the analysis units are based on precise frequencies generated by an STFT, but the processing units are based on five rough band filter banks. On the other hand, the proposed method analyzes and processes the same time-frequency representation. Thus, it can be used for precise processing without degrading the resolution.

Perceptual irrelevant component elimination [5] uses a psychoacoustic model for speech enhancement. This method focuses on the situation of hearing sounds from a digital TV or mobile phone in a noisy environment. In this situation, the masking effect of environmental noise is a problem. To eliminate the masking effect, a multiband dynamic range compressor is used to control the level of the masking threshold in this method. This method can only process the voice because the environmental noise cannot be processed. On the other hand, the proposed method can process both a voice and the BGM.

This work was supported by JSPS KAKENHI Grant Numbers 20500111, 23500147, 26330188.

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

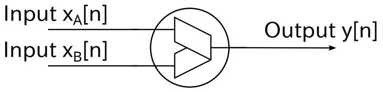


Fig. 1. Symbol for the smart mixer.

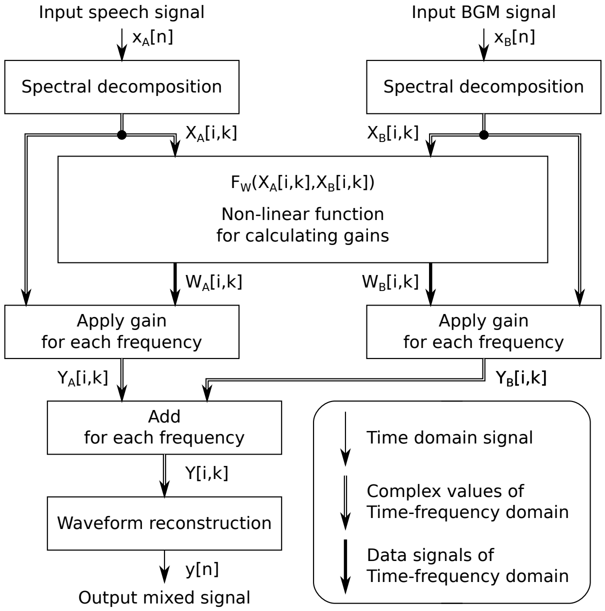


Fig. 2. Block diagram of the proposed smart mixer.

3. PROPOSED METHOD

3.1. Smart mixer

A smart mixer, i.e., our novel sound mixer using a time-frequency representation, can efficiently directly modify specific components in a time-frequency distribution. We denote the smart mixer by the symbol shown in Fig. 1.

3.2. Application of smart mixer to mixing of voice and BGM

A block diagram of the smart mixer is shown in Fig. 2. The proposed mixer has two monaural inputs $x_{\mathrm{A}}[n]$ and $x_{\mathrm{B}}[n]$ in descending order of priority and one monaural output $y[n]$, where n is the sampling time. In the case of mixing a voice and BGM, the voice is the priority input signal and the BGM is the nonpriority input signal. $X_A[i,k]$, $X_B[i,k]$, and $Y[i,k]$ are complex numbers obtained by an STFT at the point (i,k) on the time-frequency plane of each signal, in the ith (STFT) frame, and in the kth frequency bin, respectively. $W_A[i,k]$ and $W_B[i,k]$ are the gains for $X_A[i,k]$ and $X_B[i,k]$, respectively. $F_W(X_A[i,k],X_B[i,k])$ is a nonlinear function used to calculate $W_A[i,k]$ and $W_B[i,k]$. Then, $Y_A[i,k]$ and $Y_B[i,k]$ are generated by applying gains $W_A[i,k]$ and $W_B[i,k]$ to $X_A[i,k]$ and $X_B[i,k]$, respectively. Finally, the output signal $y[n]$ is obtained by applying an inverse short-time Fourier transform (ISTFT) to $Y[i,k]$.

3.2.1. *Mixing method based on psychoacoustic model*

To achieve smart mixing, we focused on the characteristics of the perception of sound. Differences in the perception of a voice and BGM are related to time transition patterns of their time-frequency distributions. One method for extracting important time-frequency components contributing to audibility is the psychoacoustic model included in MPEG-1 audio [6], [7]. There is a close relationship between audio data compression techniques and the proposed mixing method. Therefore, the proposed method uses tonal masker distributions and audible component distributions based on the time-frequency distribution considered in the psychoacoustic model of MPEG-1 audio to determine whether components should be enhanced, suppressed, or unchanged.

The tonal masker criterion $C_{\mathrm{tm}}[X,i,k]$ (Boolean) is given in (1) and (2), where $L_{\mathrm{dB}}[X,i,k]$ is the dB value of $X[i,k]$.

$$C_{\mathrm{tm}}[X,i,k] = \begin{cases} 1 & \left(\begin{array}{c} L_{\mathrm{dB}}[X,i,k] > L_{\mathrm{dB}}[X,i,k'] \\ \wedge\, L_{\mathrm{dB}}[X,i,k] > L_{\mathrm{dB}}[X,i,k'']+7, \\ \forall k' \in k+\{-1,1\}, \forall k'' \in k+\Delta_k \end{array}\right) \\ 0 & \text{(otherwise)} \end{cases}, \quad (1)$$

$$\Delta_k = \begin{cases} -2,2 & (2<k<63) \\ -3,-2,2,3 & (63 \le k < 127) \\ -6,\cdots,-2,2,\cdots,6 & (127 \le k < 255) \\ -12,\cdots,-2,2,\cdots,12 & (255 \le k \le 500) \end{cases}. \quad (2)$$

The audible criterion $C_{\mathrm{ATH}}[X,i,k]$ (Boolean) is applied by comparing the local power on the time-frequency plane and the table of the absolute threshold of hearing (ATH) [7], $\mathrm{ATH}[f]$, which is expressed by (3) and (4).

$$\mathrm{ATH}[f] = 3.64\,(f/1000)^{-0.8} - 6.5e^{(f/1000-3.3)^2} + 10^{-3}\,(f/1000)^4 \quad (\mathrm{dB\ SPL})\,, \quad (3)$$

$$C_{\mathrm{ATH}}[X,i,k] = \begin{cases} 1 & \left(\begin{array}{c} L_{\mathrm{dB}}[X,i,k] + p_{\mathrm{ATH}} > \\ \mathrm{ATH}[k] + L^{A}_{\mathrm{dB}}[X,i] - p_{\mathrm{range}} \\ \wedge L^{A}_{\mathrm{dB}}[X,i] > 0 \wedge L_{\mathrm{dB}}[X,i,k] - p_{\mathrm{ATH}} > 0 \end{array}\right) \\ 0 & \text{(otherwise)} \end{cases}. \quad (4)$$

Here, $L^A[X,i]$ is the local average power of the total frequency band of $X[i,k]$, which is averaged over the averaging time p^A_{t} (default: 60 ms), $L^A_{\mathrm{dB}}[X,i]$ is the dB value of $L^A[X,i]$, p_{ATH} (default: -69) is a parameter used to tune the gap between the listening level using a dynamic range of input signals, and p_{range} (default: 96) is a parameter used to control the dynamic range of the input signals under consideration. With this criterion, the depth of the effect is correlated with the gap between the average levels of input signals.

By combining these two criteria, the time-frequency representations are classified into four determining values expressed in Boolean variables: D_{tm}, D_{tn}, D_{nm}, and D_{na}. The four determining values for X are given by (5)-(8).

Table 1. Table of situations of the four determining values for the input signals for switching between the four processes.

		Priority input : voice (X_A)			
		D_na	D_nm	D_tn	D_tm
Nonpriority input	$\neg D_\mathrm{na}$		S_nm	S_tn	S_tm
: BGM (X_B)	D_na	S_na			

$$D_\mathrm{tm}[X,i,k] = \textstyle\sum_{k'''\in\{0,\pm1\}} \begin{pmatrix} C_\mathrm{tm}[X,i,k+k'''] \\ \wedge C_\mathrm{ATH}[X,i,k+k'''] \end{pmatrix}, \quad (5)$$

$$D_\mathrm{tn}[X,i,k] = \textstyle\sum_{k'''\in\Delta_k} \begin{pmatrix} C_\mathrm{tm}[X,i,k+k'''] \\ \wedge C_\mathrm{ATH}[X,i,k+k'''] \end{pmatrix} \wedge\neg D_\mathrm{tm}[X,i,k], \quad (6)$$

$$D_\mathrm{nm}[X,i,k] = C_\mathrm{ATH}[X,i,k] \wedge\neg D_\mathrm{tm}[X,i,k] \wedge \neg D_\mathrm{tn}[X,i,k], \quad (7)$$

$$D_\mathrm{na}[X,i,k] = \neg C_\mathrm{ATH}[X,i,k] \wedge\neg D_\mathrm{tm}[X,i,k] \wedge \neg D_\mathrm{tn}[X,i,k]. \quad (8)$$

Furthermore, the process of calculating the gains is switched depending on the situation of the four determining values for the input signals shown in Table 1 and the power difference between these signals. This is because these relationships determine whether the important components of the voice are audible after mixing.

The core equations for the gains in each time-frequency bin are given by (9)-(11), which are expressed using the limiting function $F_\mathrm{lim}(g,l,u)$. The limiting function $F_\mathrm{lim}(g,l,u)$ has upper limit u and lower limit l.

$$W_\mathrm{A}[i,k] = F_\mathrm{lim}\left(\sqrt{g_\mathrm{A}\cdot\frac{L^f[X_\mathrm{A},i,k]+L^f[X_\mathrm{B},i,k]}{L^f[X_\mathrm{A}i,k]}},\ 1, l_\mathrm{af}\right), \quad (9)$$

$$W_\mathrm{B}[i,k] = F_\mathrm{lim}\left(\sqrt{g_\mathrm{B}\cdot\frac{L^f[X_\mathrm{A},i,k]+L^f[X_\mathrm{B},i,k]}{L^f[X_\mathrm{B},i,k]}},\ 1/l_\mathrm{af}, 1\right), \quad (10)$$

$$l_\mathrm{af} = F_\mathrm{lim}\left(\sqrt{L^A[X_\mathrm{B},i]/L^A[X_\mathrm{A},i]}, 1, l_\mathrm{pb}\right). \quad (11)$$

$L^f[X,i,k]$ is the average power of the frequency band with a frequency range of one octave, which was selected on the basis of the results of preliminary experiments. l_af is a parameter used to limit the gain coefficients to avoid extreme enhancement in the entire spectrum envelope. l_pb (default: 4.8) is a parameter used to limit the gain coefficients to avoid the rapid modulation of the total power. g_A and g_B are gains used to tune the balance of whether the output signal is similar to either of the input signals at each time-frequency component after mixing. It was found in a preliminary experiment that setting gains g_A and g_B to the values in Table 2 produced good overall results, where g_tm (default: 4) and g_nm (default: 0.8) are the gains for each corresponding component.

Table 2. Gains g_A and g_B set under different situations.

Situations	S_tm	S_tn	S_nm	S_na
g_A	g_tm	1	g_nm	$\frac{L^f_\mathrm{A}[i,k]}{\left(L^f_\mathrm{A}[i,k]+L^f_\mathrm{B}[i,k]\right)}$
g_B	0	0	g_nm	$\frac{L^f_\mathrm{B}[i,k]}{\left(L^f_\mathrm{A}[i,k]+L^f_\mathrm{B}[i,k]\right)}$

Table 3. Characteristics of input signals. The voices are taken from SRV-DB [8] and the BGM is taken from the RWC-MDB-P-2001-M01 of the RWC music database [9].

Set		Contents	Standard deviation	Volume difference
1	Voice	PF00 (Female)	500	12 dB
	BGM	No.5	2000	
2	Voice	PM00 (Male)	1000	12 dB
	BGM	No.15	4000	
3	Voice	PF01 (Female)	375	24 dB
	BGM	No.94	6000	

Table 4. Parameters used in experiments.

	Method C (Conventional)	Method S (Proposed)
Sampling frequency (F_S)	44100 Hz	44100 Hz
Samples of FFT (N_FFT)	128	1024
Samples of frame shift	16	384
Type of analyze window	Hanning	Hanning
Samples of analyze window	127	1023
Type of synthesize window	Hanning	Hanning
Samples of synthesize window	63	767

Furthermore, the gains of each frequency bin are averaged over the averaging time p_t (default: 60 ms). As a result of these modifications, the musical noise is reduced.

4. EXPERIMENTS AND DISCUSSION

The characteristics of the three sets of input signals used in this study are shown in Table 3 [8],[9]. Each signal has a total duration of 6 s; the first half only contains a voice and the second half contains a voice and BGM. The sound levels of the voice and BGM were adjusted so that voice could not be heard in the case of simple addition. The parameters used in the proposed method were fixed by performing preliminary experiments using the three sets of signals. The parameters used in the experiments are shown in Table 4.

Fig. 3 shows the time-frequency distributions indicating the results of the proposed method. To hear a voice without it being obscured by the BGM, the spectrum transition pattern of the voice must be visible after mixing. For the interval where the BGM exists, the spectrum transition pattern of the voice is not visible in the time-frequency plane in the case of simple addition (Fig. 3(c)). On the other hand, it can be seen when the proposed method is used (Fig. 3(d)). Moreover, the spectrum transition pattern of the BGM also remains clearly visible.

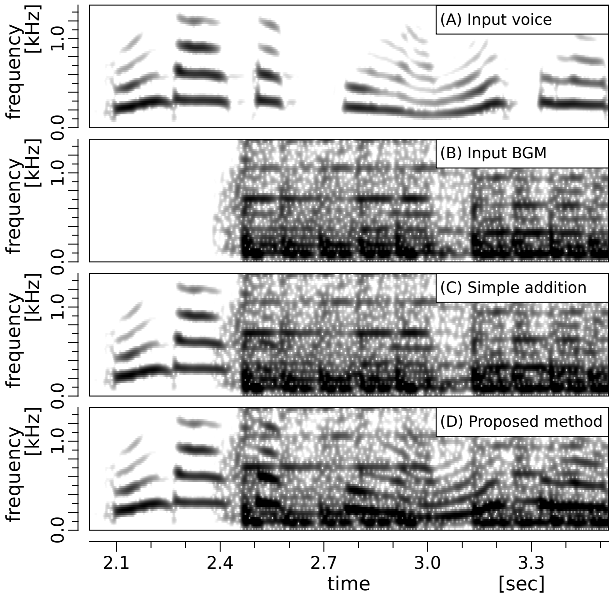


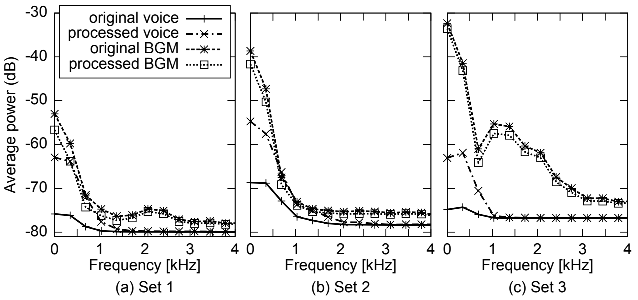
Fig. 3. Time-frequency distributions of input signal set 1 (a: voice, b: BGM), the mixing result for method B (simple addition) (c), and the mixing result for method S (the proposed method) (d).

Table 5. Descriptions of mean opinion scores (MOSs) used in the hearing experiment.

Point	Impairment
5	Degradation is inaudible
4	Degradation is audible but not annoying
3	Degradation is slightly annoying
2	Degradation is annoying
1	Degradation is very annoying

To verify the processing results, we conducted a subjective hearing experiment. The evaluated sounds were nine mixing results: the three sets of signals shown in Table 3 were processed by three mixing methods (B: simple addition, C: time-invariant equalization (the conventional method), S: the proposed method). Signals C, obtained by equalizing narrow-band spectrograms of input signals to output signals, were generated by method S, as shown in Fig. 4. The energy increments of sets 1, 2, and 3 from signals B to signals C are 0.4 dB, 0.7 dB, and -1.7 dB, and the increments from signals B to signals S are 0.1 dB, -0.1 dB, and -0.8 dB, respectively. The increments of the perceptual SNR of sets 1, 2, and 3 from signals B to signals C are -0.1 dB, -0.4 dB, and 0.6 dB, and the increments from signals B to signals S are -0.2 dB, 0.7 dB, and 0.1 dB, respectively.

The evaluation categories were Q1: voice audibility, Q2: sound quality of the voice, and Q3: sound quality of the BGM. These categories were evaluated as five-point mean opinion scores (MOSs), as shown in Table 5. The subjects were six males, who listened to the signals with headphones. The results of the hearing experiment are shown in Fig. 5.



(a) Set 1
(b) Set 2
(c) Set 3

Fig. 4. Narrow-band spectrogram (bandwidth = 344.53 Hz, $F_{\mathrm{S}} = 44100$ Hz, $N_{\mathrm{FFT}} = 128$) analysis of original signals and signals processed by the proposed method.

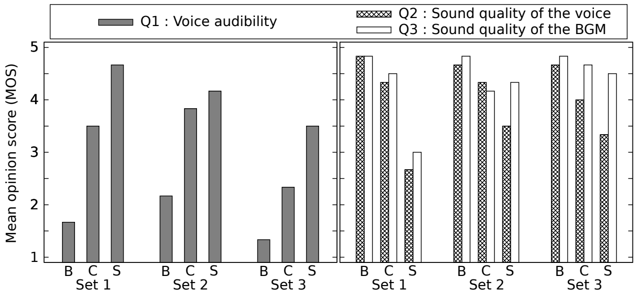


Fig. 5. Results of the hearing experiment.

The MOS of Q1 exceeded 3 for all sets of input signals for method S. Furthermore, the results of Q1 for methods S and C yielded significant p-values ($p = 0.027$, Mann-Whitney test). In contrast, the results of Q2 and Q3 in method S were lower. In particular, the MOSs for set 1 were less than 3. Notably, the results of Q1 for this set were satisfactory. Therefore, by controlling the degree of the effect of the proposed method using other characteristics of the signals, the reduction in what can be moderated.

Furthermore, we conducted a speech intelligibility experiment. The evaluated sounds were a combination of the nine speech signals: three BGMs and three mixing methods (B, C, and S). The subjects were nine males, who listened to the signals with headphones. The results of word intelligibility for methods B, C, and S were 71.7 %, 71.7 %, and 83.3 %, and the results of sentence intelligibility were 56.7 %, 66.7 %, and 76.7 %, respectively. Therefore, the effectiveness of the proposed method for ensuring that a voice remains audible was shown.

5. CONCLUSION

A novel sound mixing method, which can mix a voice and BGM without hiding the voice, was proposed. We developed a new type of sound mixer called a "smart mixer", which performs nonlinear processing on each time-frequency domain. The implementation method used in the smart mixer to maintain voice audibility was based on the psychoacoustic model included in MPEG-1 audio. The effectiveness of the proposed method was shown by the results of hearing experiments using real sound signals with six male subjects.

6. REFERENCES

[1] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*, Electronics & Electrical. Focal Press, 2008.

[2] J.D. Reiss, “Intelligent systems for mixing multichannel audio,” in *Digital Signal Processing (DSP), 2011 17th International Conference on*, July 2011, pp. 1–6.

[3] U. Zölzer, *DAFX: Digital Audio Effects*, Wiley, 2011.

[4] E. Perez-Gonzalez and J. Reiss, “Automatic equalization of multichannel audio using cross-adaptive methods,” in *Audio Engineering Society Convention 127*, Oct 2009.

[5] Hoon Heo, Mingu Lee, Seokjin Lee, and Koeng-Mo Sung, “Development of multiband dynamic range compressor regarding noise characteristics,” in *Audio Engineering Society Convention 130*, May 2011.

[6] ISO/IEC, JTC1/SC29/WG11 MPEG, “Information technology—coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s—part 3: Audio,” IS11172-3 1992 (”MPEG-1”).

[7] T. Painter and A. Spanias, “Perceptual coding of digital audio,” *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, April 2000.

[8] K.Takahashi, K.Tsutaki and T.Yoshihara, “Recording system for controlling speaking rate (ReCoK5) and public domain speech database with speaking rate variations (SRV-DB),” in *IEICE Tech. Rep., vol. 108, no. 338, SP2008-117*, Dec 2008, pp. 227–232, (in Japanese).

[9] M.Goto, H.Hashiguchi, T.Nishimura and R.Oka, “RWC music database: Popular, classical, and jazz music databases,” in *Proceedings of 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, Oct 2002, pp. 287–288.