2018 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2018, Tarragona, Spain, September 2-6, 2018

Image-to-sound transformation using image inpainting technique

Yuya HOSODA,† Arata KAWAMURA‡ and Youji IIGUNI†

†Graduate School of Engineering Science, Osaka University
1–3 Machikaneyama, Toyonaka, Osaka, 560-8531, Japan
‡Faculty of Information Science and Engineering, Kyoto Sangyo University
Kamigamomotoyama, Kitaku, Kyoto, Kyoto, 603-8555, Japan
Email: hosoda@sip.sys.es.osaka-u.ac.jp, kawamura@cc.kyoto-su.ac.jp

Abstract—We propose an image-to-sound transformation that is a technique to synthesize a sound from an image, where the image is treated as a sound spectrogram. The synthesized sound is received at mobile devices and its sound spectrogram is used to reconstruct the original image. The image-to-sound transformation has a trade-off between the synthesized sound quality and the transmission speed. We solve this trade-off by utilizing an image inpainting technique. Simulation results show the effectiveness of the proposed method.

1. Introduction

Mobile devices such as mobile phones receive various information via a carrier signal of radio waves. Sounds can also be used as a carrier signal for transmitting desired information [1]. Advantages of sounds used as the carrier signal are that a transmitter easily changes the transmission region with adjusting an amplifier, and simultaneously provides information for all target mobile devices existing in the transmission region. Since our target is mobile devices in general use, the carrier sounds should be in the audible range, not supersonic range. Acoustic OFDM (Orthogonal Frequency Division Multiplexing) can transmit contents by using sounds in the audible range [2]. The acoustic OFDM adds the OFDM-modulated text data to a sound. Since the transmission speed of the acoustic OFDM is about 1kbps, it is difficult to apply it to transmission of large size contents such as an image.

As a transmission technique of images via a sound in the audible range, the image-to-sound transformation is proposed [3]–[5]. In the image-to-sound transformation, an image is treated as a sound spectrogram where horizontal and vertical positions of the image correspond to the frame and frequency indexes, respectively. Taking the inverse Fourier transform of the image gives a synthesized sound. The sound spectrogram represents only amplitude spectrum, then there is a degree of freedom in selecting the phase spectrum. When we choose the phase spectrum randomly, the synthesized sound is a sound unlike human speech.

Igarashi *et al.* have proposed an image-to-sound transformation using the phase spectrum of a human speech [4]. This method can synthesize a sound like human speech by utilizing the phase spectrum obtained from LTFT (Long Time Fourier Transform). The frame length of LTFT is about 256ms∼1s [4]–[6]. The image is reconstructed from the sound spectrogram of the synthesized sound. This method can transmit both the desired speech and image simultaneously. In addition, Igarashi *et al.* improved the transmission speed by rearranging columns of the image [5]. They also showed the trade-off between the transmission speed and the synthesized sound quality.

We introduce an image inpainting technique to the image-to-sound transformation to solve the trade-off. In the proposed method, the sound spectrogram of an original speech is replaced with the rearranged image as much as possible to improve the transmission speed. To improve the synthesized sound quality, we keep several strong amplitude spectra of the original speech even if they exist in the image region to be replaced. Unfortunately, it causes missing regions in the reconstructed image. Image inpainting techniques are useful for repairing the missing regions of the image [7, 8]. To improve the reconstructed image quality, the proposed method employs a computationally efficient image inpainting technique [8]. Simulation results show that the proposed method can shorten the transmission time without the fatal deterioration on the synthesized sound and the reconstructed image quality.

2. Image-to-Sound Transformation Using LTPS

In the image-to-sound transformation [4, 5], a sound is synthesized from an original speech and an image. Let $s_l(n)$ ($n = 0, 1, \ldots, N-1$) be a segmented original speech signal, where l and n are the frame and time indexes, and N is the frame length for FFT (Fast Fourier Transform), respectively. We put N such as $N \geq \lceil 0.256 F_s \rceil$, where F_s is the sampling frequency and $\lceil \cdot \rceil$ is a ceiling function. LTPS (LTFT Phase Spectrum) $\angle S_l(k)$ is given as

$$\angle S_l(k) = \frac{1}{j} \log \left\{ \frac{S_l(k)}{|S_l(k)|} \right\}, \tag{1}$$

$$S_l(k) = \sum_{n=0}^{N-1} s_l(n) e^{-j2\pi nk/N}, \tag{2}$$

where $j = \sqrt{-1}$, k ($0 \leq k < N$) denotes the frequency index, and $|\cdot|$ denotes the absolute value.

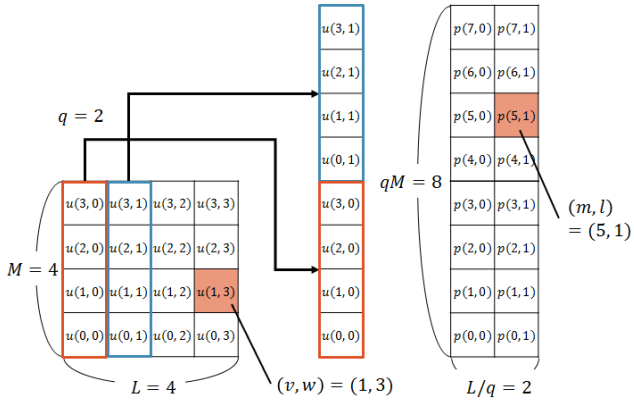

Figure 1: Rearranged image (q=2).

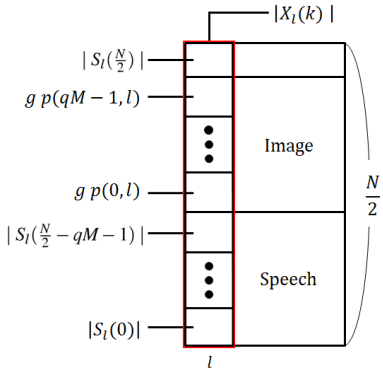

Figure 2: Sound spectrogram consisting of rearranged image and amplitude spectrum of original speech.

The brightness of the original $M \times L$ image at $(v,\ w)$ is represented as $u(v,\ w)$ $(0 \le v < M,\ 0 \le w < L)$. Let the origin (0, 0) be at the bottom left of the image. We assume that the resolution of the image is 8 bits, i.e., $0 \le u(v,\ w) < 256$. In the literature [5], the original image is rearranged such as the single column consists of the q columns of the original image. Here, q is a natural number satisfying $1 \le q \le \lfloor N/(2M) \rfloor$, where $\lfloor \cdot \rfloor$ denotes a floor function. The size of the rearranged image is $qM \times L/q$. We represent $p(m,\ l)$ as the brightness of the rearranged image at $(m,\ l)$, where $0 \le m < qM$, $0 \le l < L/q$. The relation between $(v,\ w)$ and $(m,\ l)$ is given as

$$l = \lfloor Q/(qM) \rfloor, \quad m = Q \bmod qM, \tag{3}$$

$$Q = wL + v, \tag{4}$$

$$w = \lfloor J/M \rfloor, \quad v = J \bmod M, \tag{5}$$

$$J = lqM + m, \tag{6}$$

where "mod" denotes the modulo operator. Figure 1 illustrates the rearranged image when $M = L = 4$ and $q = 2$, e.g., $u(1, 3)$ is corresponded to $p(5, 1)$.

The rearranged image is embedded in a part of the sound spectrogram. In the image-to-sound transformation, the amplitude spectrum of the synthesized sound, $|X_l(k)|$ $(k = 0,\ \ldots,\ N/2)$, is given as [5]

$$|X_l(k)| = \begin{cases} g p(k + qM - \frac{N}{2}, l), & \frac{N}{2} - qM \le k < \frac{N}{2} \\ |S_l(k)|, & \text{otherwise} \end{cases}, \tag{7}$$

where $g\,(> 0)$ is a scaling parameter, and $|X_l(N - k)| = |X_l(k)|$. Figure 2 illustrates the sound spectrogram made from the rearranged image and the amplitude spectrum of the original speech. As shown in this figure, the amplitude spectrum may exist under the image region.

The synthesized sound signal $x_l\,(n)$ is given as

$$x_l(n) = \frac{1}{N} \sum_{n=0}^{N-1} X_l(k) e^{j2\pi kn/N}, \tag{8}$$

$$X_l(k) = |X_l(k)| e^{\angle S_l(k)}. \tag{9}$$

The inverse process of making the sound spectrogram gives the reconstructed image.

Here, we can define the transmission time P[s] given as

$$P = \frac{L}{q} \frac{N}{F_s}, \tag{10}$$

and the ratio of the amplitude spectrum of the original speech $|S_l(k)|$ to $|X_l(k)|$ is given as

$$h = 1 - \frac{2qM}{N}. \tag{11}$$

When h becomes large, the synthesized sound quality is improved.

We see from (10) and (11) that when q increases, P and h become small, i.e., the transmission speed becomes high and the synthesized sound quality is degraded. Inversely, when q decreases, the transmission speed becomes low and the synthesized sound quality is improved. This is the trade-off between the transmission speed and the synthesized sound quality.

3. Image-to-Sound Transformation Using Image Inpainting Technique

We utilize an image inpainting technique to improve both the transmission speed and the synthesized sound quality, simultaneously. The proposed method uses the minimum transmission time P, i.e., $q = \lfloor N/(2M) \rfloor$. To improve the synthesized sound quality, we keep several original speech spectra whose amplitudes are greater than $255g$.

We put $|S_l(k)|$ $(N/2 - qM \le k < N/2)$ as $|S_l(k_1)| \ge |S_l(k_2)| \ge \ldots \ge |S_l(k_i)| \ldots \ge |S_l(k_{qM})|$, where k_i denotes the frequency index corresponded to the amplitude big in the i th.

The proposed method designs $|X_l(k)|$ $(0 \le k \le N/2)$ as

$$|X_l(k)| = \begin{cases} g p(k + qM - \frac{N}{2}, l), & \frac{N}{2} - qM \le k < \frac{N}{2} \\ & \text{where } k \ne k_1, \ldots, k_i \\ |S_l(k)|, & \text{otherwise} \end{cases}, \tag{12}$$

where $|X_l(N - k)| = |X_l(k)|$. We define the ratio of $|S_l(k)|$ to $|X_l(k)|$ in the image region $(N/2 - qM \le k < N/2)$ given as

$$r = \frac{i}{qM}, \tag{13}$$

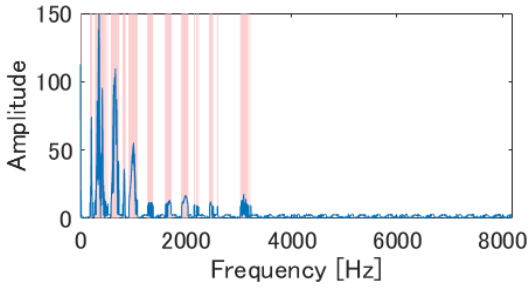

Figure 3: Proposed amplitude spectrum which maintains several large original speech spectra.

where $r = 0$ ($i = 0$) implies that $|S_l(k)|$ is not used in the image region. In this case, the proposed method is corresponding to the conventional method [5]. On the other hand, when $r = 1$ ($i = qM$), there are no image components in the sound spectrogram. In this case, the synthesized sound is identical to the original speech, but it is impossible to reconstruct the original image.

Figure 3 shows $|X_l(k)|$ with $r = 0.125$, where the horizontal and vertical axis denote the frequency and the amplitude, respectively. The shaded amplitudes denote $|S_l(k)|$ within the image region. Such amplitudes do not represent the brightness of the original image, hence the reconstructed image has missing regions. Figure 4 illustrates the reconstructed image with missing regions caused from (12), where (a) and (b) denote the original image and the reconstructed image, respectively.

When the maintained amplitude $|S_l(k_i)|$ is greater than $255g$, we easily find the missing regions from the brightness of the reconstructed image. Under the condition that the missing regions are known, we repair the missing regions by using an image inpainting technique. Let $\tilde{u}(v, w)$ be the brightness of the reconstructed image when using (12). The brightness of the repaired image $\hat{u}(v, w)$ is given as [8]

$$\hat{u}(v,w) = \begin{cases} G(v,w), & (v,w) \in \Omega \\ \tilde{u}(v,w), & \text{otherwise} \end{cases}, \tag{14}$$

$$G(v,w) = \text{Gauss}(\, g(v,w)\,), \tag{15}$$

$$g(v,w) = \begin{cases} \tilde{u}\left(\arg\min_{v'w' \in \Phi} d(vw, v'w')\right), & (v,w) \in \Omega \\ \tilde{u}(v,w), & \text{otherwise} \end{cases}, \tag{16}$$

$$d(vw,\ v'w') = \sqrt{(v-v')^2 + (w-w')^2}, \tag{17}$$

where Ω denotes a set of positions included in missing regions, and Φ denotes a set of positions included in non-missing regions. The operator Gauss$(\cdot)$ is a Gaussian filter.

Figure 5 shows the block diagram of the proposed method, where ILTFT denotes the inverse LTFT. Figure 5 (a) shows the sound synthesis procedure, and (b) shows the image reconstruction procedure. The synthesized sound signal $x_l(n)$ has the high quality due to maintaining the several large amplitude spectra of the original speech. The

(a) **(b)**

Figure 4: (a) Original image. (b) Reconstructed image.

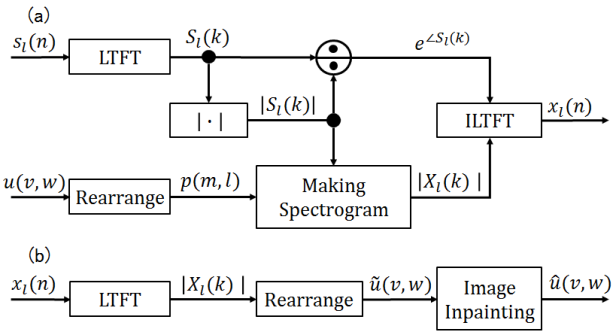

Figure 5: Block diagrams of proposed method. (a) Sound synthesis process. (b) Image reconstruction procedure.

reconstructed image $\tilde{u}(v, w)$ is repaired by using (14)–(17), and the repaired image $\hat{u}(v, w)$ is finally obtained.

4. Simulation

We carried out computer simulations to confirm the capability of the proposed method. 15 standard images ($M = L = 256$, resolution 8bit) were used as the original image. 5 male and 5 female speech signals taken from Japanese speech corpus ASJ-JNAS [9] were used as the original speech. The sampling frequency is 16kHz, $N = 16384$ (about 1s), and we empirically put $g = 4/255$.

Figure 6 (a)–(c) show examples of the simulation results, where (a) shows the conventional method [5] with $h = 0$, (b) shows the conventional method [5] with $h = 0.25$, and (c) shows the proposed method with $h = 0$ and $r = 0.125$. Each top panel shows the original speech, the middle panel shows the synthesized sound, and the bottom panel shows the reconstructed image. We see from (a) that the synthesized sound causes degradation like noise addition, where the transmission time is $P = 8.19$s. The synthesized sound of (b) is obtained by utilizing the original speech spectrogram of 25% in the whole sound spectrogram, and achieves the good synthesized sound quality, but the transmission time becomes long as $P = 11.26$s. The reconstructed images of (a) and (b) are identical with the original image. On the other hand, the proposed method (c) gives the good synthesized sound quality without the degradation of the transmission speed, where $P = 8.19$s. In the reconstructed image, the missing regions are repaired without discomfort.

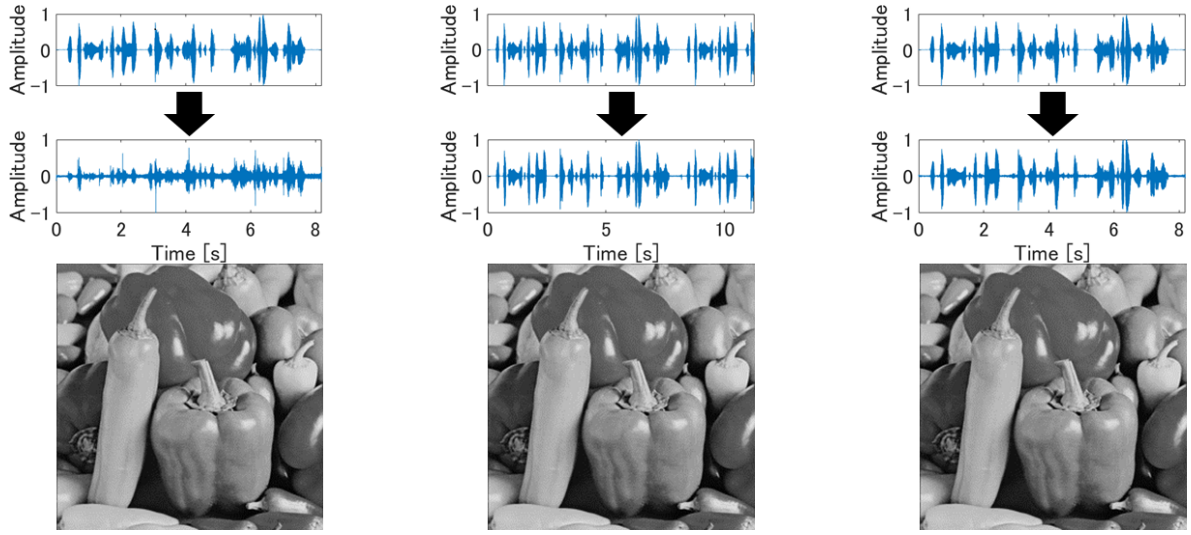

Figure 6: Top panel shows original speech, middle panel shows synthesized sound, and bottom panel shows reconstructed image. (a) Conventional method [5] ($h = 0$). (b) Conventional method ($h = 0.25$). (c) Proposed method ($h = 0$, $r = 0.125$).

Table 1: Averaged evaluation results.

	SNR[dB]	PSNR[dB]	SSIM	P[s]
Conv.(h = 0)	0.40	60.00	1.00	8.19
Conv.(h = 0.25)	14.56	60.00	1.00	11.26
Prop.(r = 0.125)	14.86	35.54	0.98	8.19

Next, we performed objective evaluations which are SNR (Signal-to-Noise Ratio) for evaluating the synthesized sound quality and PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural SIMilarity) for evaluating the reconstructed image quality. The averaged evaluation results for 150 samples are shown in Table 1, where "Conv." and "Prop." denote the conventional method [5] and the proposed method, respectively. We see from Table 1 that the proposed method gives the best SNR in comparison to the conventional methods. Although PSNR of the proposed method was around 15dB lower than that of the conventional method, the difference between SSIM was slight.

5. Conclusion

This paper proposed the image-to-sound transformation with the image inpainting technique. The proposed method gives the high synthesized sound quality and the high transmission speed. Simulation results showed that a 256×256 image can be transmitted around 8s, i.e., about 65kbps. The proposed method achieved SNR of 14dB for the synthesized sound, SSIM of 0.98 for the reconstructed image.

References

[1] NTT DOCOMO, INC., "Air Stamp," http://www.air-stamp.jp/ (accessed Dec. 26. 2017).

[2] H. Matsuoka, Y. Nakashima, and T. Yoshimura, "Acoustic OFDM system and performance analysis," IEICE Trans. Fundamentals, vol.E91-A, no.7, pp.1652–1658, 2008.

[3] P.B.L. Meijer, "An experimental system for auditory image representations," IEEE Trans. Biomed. Eng., vol.39, no.2, pp.112–121, 1992.

[4] H. Igarashi, A. Kawamura, and Y. Iiguni, "Image to speech transformation using long term phase spectrum," IEICE Trans. Fundamentals (Japanese Edition), vol.J99-A, no.8, pp.270–279, 2016.

[5] A. Kawamura, H. Igarashi, and Y. Iiguni, "An efficient image to sound mapping method using speech spectral phase and multi-column image," IEICE Trans. Fundamentals, vol.E100-A, no.3, pp.893–895, 2017.

[6] M. Kazama, S. Goto, and M. Tohyama, "On the significance of phase in the short term Fourier spectrum for speech intelligibility," J. Acoust. Soc. Am., vol.127, no.3, pp.1432–1439, 2010.

[7] A. Telea, "An image inpainting technique based on the fast marching method," Journal of Graphics Tools, vol.9, no.1, pp.25–36, 2004.

[8] H. Yoshida, "Image inpainting method based on painting texture," The Journal of the Institute of Image Electronics Engineers of Japan (Japanese Edition), vol.44, no.1, pp.77–84, 2015.

[9] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," Journal of the Acoustical Society of Japan, vol.20, no.3, pp. 199–206, 1999.