IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 23, 2021, accepted March 24, 2021, date of publication March 29, 2021, date of current version April 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3069267

Audio-Visual Model for Generating Eating Sounds Using Food ASMR Videos

KODAI UCHIYAMA[1] AND KAZUHIKO KAWAMOTO[2], (Member, IEEE)
[1]Graduate School of Science and Engineering, Chiba University, Chiba 263-8522, Japan
[2]Graduate School of Engineering, Chiba University, Chiba 263-8522, Japan

Corresponding author: Kazuhiko Kawamoto (kawa@faculty.chiba-u.jp)

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP19K12039.

ABSTRACT We present an audio-visual model for generating food texture sounds from silent eating videos. We designed a deep network-based model that takes the visual features of the detected faces as input and outputs a magnitude spectrogram that aligns with the visual streams. Generating raw waveform samples directly from a given input visual stream is challenging; in this study, we used the Griffin-Lim algorithm for phase recovery from the predicted magnitude to generate raw waveform samples using inverse short-time Fourier transform. Additionally, we produced waveforms from these magnitude spectrograms using an example-based synthesis procedure. To train the model, we created a dataset containing several food autonomous sensory meridian response videos. We evaluated our model on this dataset and found that the predicted sound features exhibit appropriate temporal synchronization with the visual inputs. Our subjective evaluation experiments demonstrated that the predicted sounds are considerably realistic to fool participants in a "real" or "fake" psychophysical experiment.

INDEX TERMS Multi-modal deep neural network, autonomous sensory meridian response, eating sound generation.

I. INTRODUCTION

We often hear phrases such as "this fried chicken is so crunchy" or "the cheese is melted" in food review shows. These videos generally mention the food texture besides its taste. Food items makes distinctive sounds when they are bitten or chewed. These sounds stimulate our appetite and they are often emphasized in food advertisements. By emphasizing the food texture with relevant sound, the consumer appetite can be significantly influenced, thereby generating more sales. In contrast, a study demonstrated that the chewing sounds activate the feeling of satiety [8]; considering the results of this study, the generation of emphasized eating sounds could achieve similar satisfaction with smaller amounts of food intake to prevent obesity.

In this research, we propose a deep learning model that generates food texture sounds from silent eating videos. To generate eating sounds from visual information, we require a dataset with sounds corresponding to eating behaviors; accordingly, we used food autonomous sensory meridian response (ASMR) videos from YouTube. They are also known as "mukbang" (a portmanteau of the South Korean words for "eating" ["meokneun"] and "broadcast" ["bangsong"] that refers to online broadcasts where individuals eat food and interact with the viewers). These food ASMR videos record the eating behaviors that specialize in texture sounds. They have become popular in America following a similar trend that circulated in South Korea in 2016. Viewers watch food ASMR videos for social reasons, sexual reasons, entertainment, eating reasons, and/or as an escapist compensatory strategy. We trained the proposed model using such videos.

The consistency with the corresponding visual information is an essential characteristic for the generated sounds. The sound generating method should produce a sound event at exactly the same time or soon after the occurrence of the corresponding visual event. For example, an eating sound should be synchronized with the visual event in which the person closes his mouth. Various recent research works have attempted to implement deep generative models with speech synthesis and voice conversion to produce visually aligned sounds that correspond to the visual

The associate editor coordinating the review of this manuscript and approving it for publication was Hiu Yung Wong.

features [4], [20]; nevertheless, the alignment issue is still a challenge.

Our model takes the facial features acquired from a video frame as input and predicts the amplitude spectrogram corresponding to the time series. To generate a waveform from the amplitude spectrogram, we used the Griffin-Lim algorithm [12] to restore the phase and implemented inverse short-time Fourier transform (STFT).

The contributions of this study can be summarized as follows:

- We introduced a learning framework for generating food texture sounds from silent food eating videos.
- We proposed a multi-modal deep neural network architecture comprising a convolutional network, recurrent network, and fully connected layers.
- We showed that the proposed structure outperformed the baseline in eating sound reconstruction.
- We created a dataset consisting of food ASMR videos from YouTube. It contains videos of people grabbing, biting, chewing, and swallowing food items.

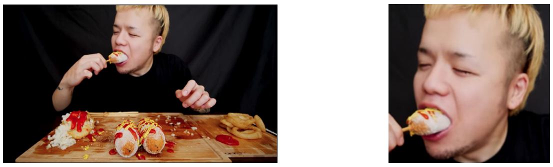
The demonstration video can be found at the following link: https://youtu.be/xFA7nU4K8aE

II. RELATED RESEARCH

Videos with audios include the audio information associated with the visual information of the videos. The correlation between the visual and audio information has been studied in the field of multi-modal deep learning.

A. VISUALLY ALIGNED SOUND SYNTHESIS

Owens *et al.* [13] proposed a method for predicting sound features from videos considering the interactions between various objects, and demonstrated that sound contains important information for recognizing the material properties and interactions. Chen *et al.* [4] proposed a method to generate sound considering the sound class and used the perceptual loss to align the semantic information. In this study, we conducted subjective experiments to evaluate the generated eating sounds. To generate a plausible eating sound, we utilized the example-based synthesis method proposed in [13]. Zhou *et al.* [20] collected an unconstrained dataset (VEGAS) that included 10 types of sounds recorded in the wild and proposed a recurrent neural network-based method to directly generate a waveform from videos. Chen *et al.* [5] exploited conditional generative adversarial networks to generate cross-modal audio visuals of musical performances. Chen *et al.* [6] designed an audio forwarding regularizer that could control the irrelevant sound component, thereby preventing the model from learning incorrect mapping between the video frames and the sound emitted by the out-of-screen objects. Akbari *et al.* [3] tried to reconstruct natural sounding speech using a neural network that takes as input the face region of the talker and estimates bottleneck features extracted from the auditory spectrogram by a pre-trained autoencoder.


(a) Food ASMR video

(b) Face segment

FIGURE 1. Example of dataset [1].

B. SOUND SEPARATION

Gao and Grauman [11] proposed a model to detect each musical instrument in a video clip of multiple sounds and divide the sound emitted from each instrument. Gan *et al.* [10] improved the performance of time-frequency mask estimation for sound source separation using a context-aware graph network to extract information from the time series of the performer key points. In the research on human speech segmentation, a method to predict complex ratio masks with the amplitude and phase information was proposed to extract speech from the spectrogram of synthetic speech [2], [9]. The human speaking and eating processes are similar in terms of human mouth movement. In this study, the proposed model and a visual input procedure are constructed based on the aforementioned studies.

III. FOOD ASMR DATASET

We collected a dataset with sounds corresponding to eating behaviors. We used food ASMR videos from YouTube, as shown in Figure 1, and created a dataset.[1] This is the first dataset that includes food ASMR videos and focuses on eating sounds (up to our knowledge). From these videos, we selected those that did not contain any noise such as human voice; we only included food texture sounds. Moreover, because the food does not occlude much of a scene, we can also observe what happens to the food after it is eaten. To train the model, we used segmented images of faces extracted by MTCNN [19] for each frame, as shown in Figure 1. There are two main types of food texture sounds: sounds that are “generally” pleasant to humans (e.g., fried food, vegetables, and fruit) and those that are unpleasant to humans (e.g., slurping noise). We focused on the former type of sound in this study. The recorded food videos included fried food such as fried chicken, corn dog, and fried potato, which have the texture of “crispy” onomatopoeias. We collected 45 different types of videos which are approximately 5-16 min and divided each of these videos into 3-second segments. We collected 3588 segment data from the videos in total. Additionally, in each clip, the visible face in the video and audible sound in the soundtrack belonged to a single eating person.

[1] https://github.com/KodaiUchiyama/Food-ASMR-Dataset

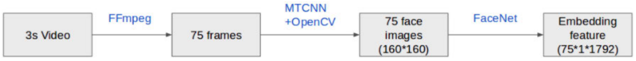


FIGURE 2. Video representation.

IV. EATING SOUND GENERATION FRAMEWORK

The proposed model takes the detected facial features as input and predicts the amplitude spectrogram corresponding to the time series of the input.

A. VIDEO REPRESENTATION

To generate the input features, we converted the frame rate of the dataset to 25 FPS. We also divided the video into a frame set of 75 frames (3 seconds) each. We use MTCNN [19] to extract the face segments and resized them to 160 × 160.

We used a pretrained face recognition model [16] and obtained 1792 × 1 face features per frame. The procedure for creating the input image features is illustrated in Figure 2. The face features were used as the input because they eliminate the irrelevant variations between images and retain the information necessary for recognizing millions of faces. A related study [15] showed that face features are effective in interpreting facial expressions. In this study, experiments were also conducted using RGB images as input data; however, they did not improve the prediction accuracy.

B. AUDIO REPRESENTATION

For the audio features, we downsampled the sample rate of sound to 16 kHz, and the stereo audio was converted to mono. We computed the STFT of the 3 s audio segments to obtain the amplitude spectrogram. The STFT was computed using a Hann window of length 25 ms with a hop length of 160 and fast fourier transform size of 512, resulting in an output audio feature of 300 × 256 scalars. However, the overall distribution of the magnitude values was not Gaussian because several entries in the spectrogram were close to zero, which can impede the learning process. Therefore, the following equation (1) was applied to each time-frequency unit, i.

$$M_i = \log(M_i + C), \quad C = 10^{-7} \tag{1}$$

The constant, C, was set empirically to ensure that the sound features approach a normal distribution. Finally, we applied the sigmoid function to each time-frequency unit, i; subsequently, the values of each unit were normalized in the range of 0 to 1.

C. GENERATING WAVEFORMS

We used two methods to generate a waveform from the amplitude spectrogram.

1) RAW SOUND METHOD

We generated a waveform based on phase reconstruction and inverse STFT from the amplitude spectrogram. We used the Griffin/Lim method [12] for phase restoration from the amplitude spectrogram; random values were used as the initial values of the phase. The waveform obtained by this raw sound method was used for evaluating the information obtained by the deep neural network (DNN) model.

TABLE 1. Convolutional layer architecture.

	conv1	conv2	conv3	conv4	conv5	conv6
Num Filters	256	256	256	256	256	256
Filter Size	7 × 1	5 × 1	5 × 1	5 × 1	5 × 1	5 × 1
Dilation	1 × 1	1 × 1	2 × 1	4 × 1	8 × 1	16 × 1
Context	7 × 1	9 × 1	13 × 1	21 × 1	37 × 1	69 × 1

2) EXAMPLE-BASED SYNTHESIS METHOD

This method synthesized the waveform data from the amplitude spectrograms predicted by the DNN model, and the sound generated by this method included a texture sound plausible for a human ear. We used this method for the subjective evaluation experiments. The procedure can be summarized as follows. a) We decomposed the feature vector of the amplitude spectrogram, $\vec{s}_1, \vec{s}_2, \cdots, \vec{s}_N, (N = 300)$, predicted by the DNN model into 25 samples along the time scale. b) For each feature vector of the decomposed amplitude spectrogram, $\vec{s}_1, \vec{s}_2, \cdots, \vec{s}_{25}$, the sound feature with the L2 distance as the nearest neighbor was selected from the training data. c) We obtained the final waveform data by synthesizing the waveform data corresponding to the selected sound features.

D. NETWORK ARCHITECTURE

In the researches on human speech separation, various methods have been proposed to predict complex ratio masks with the amplitude and phase information to extract speech from synthetic speech [4], [9]. In this study, the model was constructed based on such studies. The proposed model takes the facial features obtained by pre-processing the input video frames and outputs the spectrogram of the corresponding sound features.

As shown in Figure 3, the model structure consists of a convolutional neural network (CNN), a single-layer bidirectional long-short-term memory (LSTM) with 400 cells, and three fully connected layers. Details of the CNN module are listed in Table 1. The dimensions of the input face features do not represent the spatial information; they retain the information required for face recognition in a similar manner as [7]. To prevent over fitting, we discarded 20% of the neurons in all three fully connected layers during the learning iterations [17]. The convolutional network of our model processes the input face features; it consists of dilated convolutions, as presented in Table 1. Note that "spatial" convolutions and dilations in the convolutional network are performed over the temporal axis (not over the 1792-D face features channel).

V. EXPERIMENTS AND RESULTS

A. IMPLEMENTATION DETAILS

The experiments were conducted using the Food ASMR DATASET. We split the food ASMR DATASET into training sets (35 kinds of videos) and testing sets (10 kinds of videos)

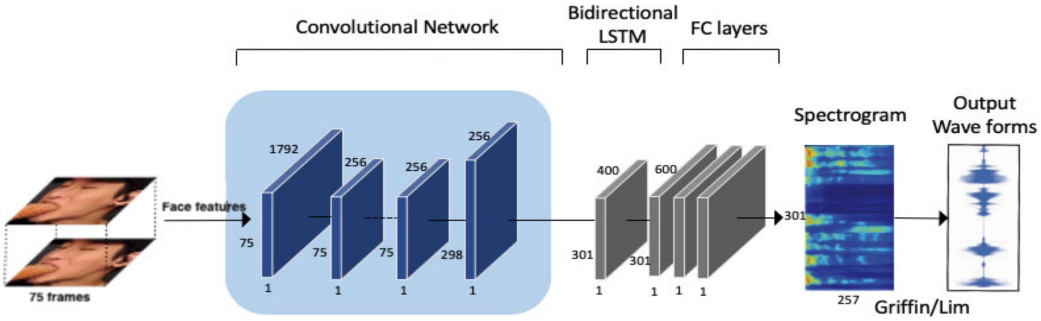


FIGURE 3. Proposed model architecture. The convolutional network takes the segments of the faces detected in each frame of the video as inputs. The convolutional network extracts the face features for each segment using a pretrained face recognition model; it then learns each visual feature using a dilated convolutional network. The visual features are further processed using a bidirectional LSTM and three fully connected layers. The network outputs a spectrogram of eating sounds; it is converted back to waveforms using phase restoration and inverse STFT.

and performed 4-fold cross validation to ensure that the same kind of video is not represented in both testing and training sets, which enables the model to avoid data bias and overfitting. In each fold, we trained the proposed model with 90% of the data and validated it on the rest 10%.

The proposed model was implemented in TensorFlow. We used a batch size of four samples and trained using an Adam optimizer with a learning rate of 1.0×10^{-3}.

The mean square error of the predicted sound features and the ground truth features were used as the loss functions. To generate a waveform, we used the 1) raw sound method and 2) example-based synthesis method.

B. EXPERIMENT RESULTS

1) QUALITATIVE VISUALIZATION

Figure 4 shows the results of the predicted amplitude spectrograms in the test data. Figure 5 shows the sound wave shapes generated from the predicted amplitude spectrograms. We compared the predicted spectrograms with the ground truth spectrograms. The pronunciation timing features were found to be consistent although this method failed in capturing high frequency information which is mainly due to the nature of the task. This can be attributed to the timing of the high-frequency band sounds in the eating behaviors. The ‘‘crispy’’ sound in the high frequency band is emitted when a person starts eating a food item. Therefore, the model experiences difficulties in learning the sounds in the high frequency band because the duration of the first eating action is less than that of eating from the entire eating behavior. Note that the accuracy of the spectrogram predicted by the example-based synthesis method depends on the spectrogram of the raw sound method; thus, it is necessary to improve the accuracy of the raw sound method to enhance the quality of the example-based synthesis method.

2) QUALITY AND ACCURACY MEASUREMENT OF THE RECONSTRUCTIONS

Lip2Audspec [3] is a method for reconstructing natural sounding speech given visual input, which applies a neural network to estimate bottleneck features extracted from the auditory spectrogram by a pre-trained autoencoder. We consider it our baseline because it is relevant in terms of generating sounds from silent videos which recorded the area around the faces. We adapted food ASMR dataset to this baseline model and compared the accuracy of predicted spectrograms and generated audio waveforms for performance comparisons. We also performed the 4-fold cross validation to train the baseline model using the same cross validation partitions with the proposed model. In Table2, we show the 4-fold average of L2 distance between the predicted spectrogram and the ground truth spectrogram in the experiments. Note that we consider the original audio the ground truth. The sound quality is measured with the 4-fold average of

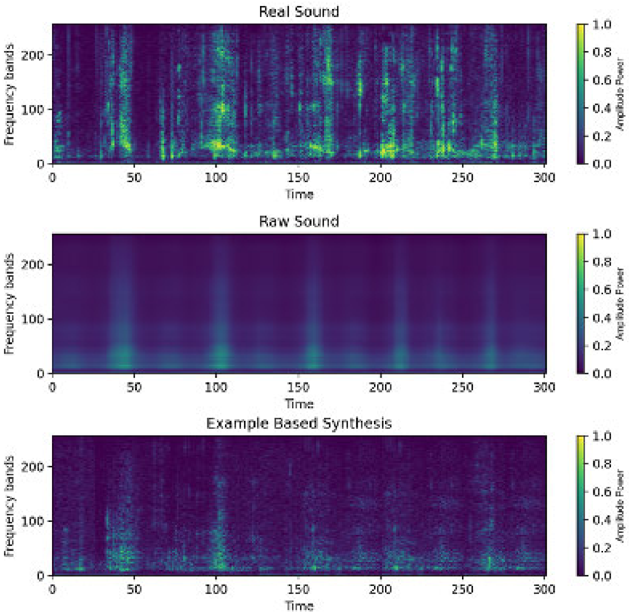


FIGURE 4. Predicted spectrogram. The top row shows the ground truth spectrogram. The results of the raw sound method and example-based synthesis method are demonstrated in the middle and bottom rows, respectively.

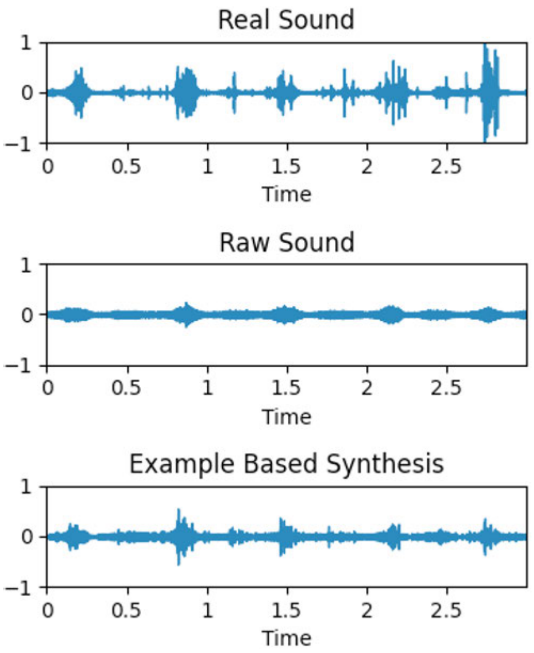


FIGURE 5. Generated waveform. The top row shows the ground truth waveform. The waveforms produced using the raw sound and example-based synthesis methods are presented in the middle and bottom rows, respectively.

TABLE 2. Average L2 distances and PESQ (higher is better) between the predicted and the ground truth spectrograms using simulated data.

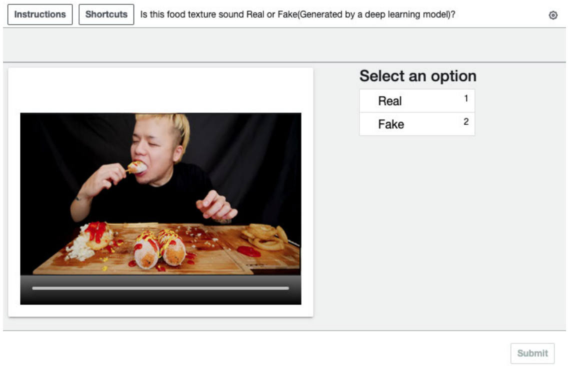
	Lip2Audspec	Ours
L2 distances	188.30	**179.34**
PESQ	1.19	**1.27**

perceptual evaluation of speech quality (PESQ) metric [14] (higher is better). PESQ was originally designed to quantify degradation due to codecs and transmission channel errors. As it can be seen from the Table2, the predicted spectrograms and the reconstructed audios using the proposed method have both higher accuracy and quality compared to the baseline. We argue that because the face features given as the input to our proposed model can eliminate the irrelevant features such as food around the mouth and retain the lip-movement information necessary for generating eating sound.

3) SUBJECTIVE EVALUATION EXPERIMENT

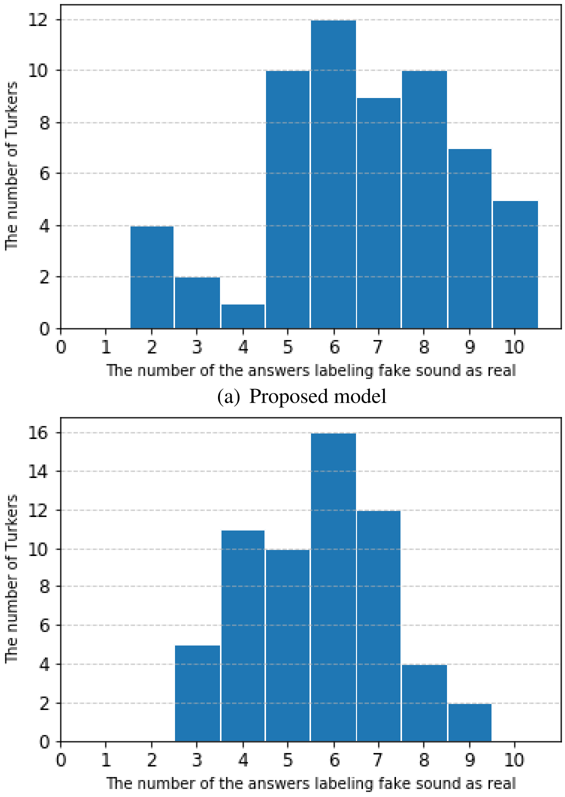
It is difficult to quantify the human perception of synthesized sound using objective measures. Therefore, we evaluated the realism of the generated sounds on Amazon Mechanical Turk. We wanted to determine whether the generated sounds can trick people into believing that they were real. For this test, we used the generated results of the model that showed the best performances in cross-validation. 60 turkers were shown 20 videos (10 real and 10 synthesized). They were asked to label each video as real (originally belonging to this video) or fake (synthesized by the example-based synthesis method). Figure 6 shows an example of the evaluation experiment. A total of 66% of the generated audio was rated as real. The distribution of turker scores is depicted in Figure 7.

We also evaluated the results of the baseline model trained with the food ASMR dataset. The 60 turkers were shown 20 videos (10 real and 10 synthesized). Each video was 3 s long. A total of 56.5% of the generated audio was rated as real. The distribution of turker scores is depicted in Figure 7. Subjective evaluations of the results also show that our



Is this food texture sound Real or Fake(Generated by a deep learning model)?

FIGURE 6. Example of the evaluation experiment. Turkers were shown 3 s videos for each question. They were asked to label each video as real (original sound) or fake (synthesized by the model).



(a) Proposed model

(b) Baseline model

FIGURE 7. Distribution of answers labeling fake sound as real of turkers in subjective evaluation. The top figure shows the results of the proposed model. The bottom figure shows the results of the baseline model: Lip2audspec.

proposed model received higher rating in terms of realness than the baseline model.

VI. CONCLUSION

We examined a method for generating food texture sounds from silent food eating videos. To train the proposed model, we created a dataset using food ASMR videos from YouTube. The proposed multimodal deep learning model took facial features from video frames as input and predicted the amplitude spectrograms corresponding to the time series. The waveforms were generated from the spectrograms based on phase restoration and inverse STFT. We showed that the predicted amplitude spectrogram confirmed that the sound timing was accurately captured. We also showed that the proposed structure outperformed the baseline in eating sound reconstruction. An example-based synthesis method was used to generate food texture sounds for subjective evaluation. We demonstrated that the sounds predicted by our model are realistic enough to fool the participants in a ''real'' or ''fake'' subjective evaluation; additionally, it demonstrated efficient temporal synchronization with the visual inputs. In the future, to build the end to end model, we want to train a model such as Wavenet [18] to generate waveforms from spectrograms. In addition, because the phase information is important for predicting impact sounds such as food texture sounds, we want to examine a model that uses DNN to predict the phase information. We envision that our work will open up the new research on studying the generation of food texture sounds.

REFERENCES

[1] *ASMR Cheese Corn Dog Eating Show (2019) Youtube Video, Added by Ogui Rasukaru*. Accessed: Feb. 23, 2021. [Online]. Available: https://www.youtube.com/watch?v=mngxuleohxm

[2] T. Afouras, J. S. Chung, and A. Zisserman, ''The conversation: Deep audio-visual speech enhancement,'' in *Proc. Interspeech*, Sep. 2018, pp. 3244–3248.

[3] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, ''Lip2Audspec: Speech reconstruction from silent lip movements video,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2516–2520.

[4] K. Chen, C. Zhang, C. Fang, Z. Wang, T. Bui, and R. Nevatia, ''Visually indicated sound generation by perceptually optimized classification,'' in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCV)*, in Lecture Notes in Computer Science, vol. 11134. Cham, Switzerland: Springer, 2019.

[5] L. Chen, S. Srivastava, Z. Duan, and C. Xu, ''Deep cross-modal audio-visual generation,'' in *Proc. Thematic Workshops ACM Multimedia (Thematic Workshops)*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 349–357.

[6] P. Chen, Y. Zhang, M. Tan, H. Xiao, D. Huang, and C. Gan, ''Generating visually aligned sound from videos,'' *IEEE Trans. Image Process.*, vol. 29, pp. 8292–8302, 2020.

[7] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, ''Synthesizing normalized faces from facial identity features,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3386–3395.

[8] R. S. Elder and G. S. Mohr, ''The crunch effect: Food sound salience as a consumption monitoring cue,'' *Food Qual. Preference*, vol. 51, pp. 39–46, Jul. 2016.

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, ''Looking to listen at the cocktail party,'' *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–11, Aug. 2018.

[10] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba, ''Music gesture for visual sound separation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jun. 2020, pp. 10475–10484.

[11] R. Gao and K. Grauman, ''Co-separating sounds of visual objects,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3878–3887.

[12] D. Griffin and J. Lim, ''Signal estimation from modified short-time Fourier transform,'' *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[13] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, ''Visually indicated sounds,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2405–2413.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, ''Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.

[15] E. M. Rudd, M. Günther, and T. E. Boult, ''Moon: A mixed objective optimization network for the recognition of facial attributes,'' in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2016, pp. 19–35.

[16] F. Schroff, D. Kalenichenko, and J. Philbin, ''FaceNet: A unified embedding for face recognition and clustering,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, ''Dropout: A simple way to prevent neural networks from overfitting,'' *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, ''WaveNet: A generative model for raw audio,'' in *Proc. ICLR*, 2016, pp. 1–15.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, ''Joint face detection and alignment using multitask cascaded convolutional networks,'' *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[20] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, ''Visual to sound: Generating natural sound for videos in the wild,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3550–3558.

KODAI UCHIYAMA was born in Tokyo, Japan, in 1995. He received the B.E. degree in informatics and imaging systems from Chiba University, Japan, in 2018, where he is currently pursuing the master's degree with the Graduate School of Science and Engineering. His research interests include multi-modal deep learning, speech separation, sound localization, and sound generation.

KAZUHIKO KAWAMOTO (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Chiba University, Japan, in 1997, 1999, and 2002, respectively. From 2002 to 2005, he was an Assistant Professor with the Tokyo Institute of Technology, Japan. From 2005 to 2009, he was an Associate Professor with the Kyushu Institute of Technology, Japan. He is currently a Professor with the Graduate School of Engineering, Chiba University. His research interests include computer vision, pattern recognition, and statistical signal processing.