Representation Learning Based on Variational Autoencoders for Imagined Speech Classification

Hajime Yano
Graduate School of System Informatics
Kobe University
Kobe, Japan
hyano@port.kobe-u.ac.jp

Ryoichi Takashima
Graduate School of System Informatics
Kobe University
Kobe, Japan

Tetsuya Takiguchi
Graduate School of System Informatics
Kobe University
Kobe, Japan

Seiji Nakagawa
Center for Frontier Medical Engineering
Chiba University
Chiba, Japan

***Abstract*—Brain computer interfaces based on speech imagery have attracted attention in recent years as more flexible tools of machine control and communication. Classifiers of imagined speech are often trained for each individual due to individual differences in brain activity. However, the amount of brain activity data that can be measured from a single person is often limited, making it difficult to train a model with high classification accuracy. In this study, to improve the performance of the classifiers for each individual, we trained variational autoencoders (VAEs) using magnetoencephalographic (MEG) data from seven participants during speech imagery. The trained encoders of VAEs were transferred to EEGNet, which classified speech imagery MEG data from another participant. We also trained conditional VAEs to augment the training data for the classifiers. The results showed that the transfer learning improved the performance of the classifiers for some participants. Data augmentation also improved the performance of the classifiers for most participants. These results indicate that the use of VAE feature representations learned using MEG data from multiple individuals can improve the classification accuracy of imagined speech from a new individual even when a limited amount of MEG data is available from the new individual.**

***Index Terms*—imagined speech, speech imagery, variational autoencoder, representation learning, magnetoencephalography**

I. Introduction

Brain computer interfaces (BCIs) have been researched and developed as tools of machine control and communication for physically disabled people who have difficulty with speech and gestures. BCIs based on speech imagery have recently attracted attention because they enable more flexible machine control and communication than conventional BCIs based on event-related potentials or motor imagery [1].

There are many electroencephalographic (EEG) studies on speech imagery BCI [1]–[4]. Not only classical machine learning methods, such as the support vector machine, but also neural networks with a large number of parameters have recently been used as classifiers of imagined speech. However, EEG signals are inherently noisy, and it is difficult to collect a sufficient amount of EEG data to train neural networks with high classification performance [3], [4]. In particular, the classifiers are often trained for each individual due to individual differences in brain activities, but the amount of EEG data that can be measured from a single subject is limited due to the physical and mental burden associated with the measurements carried out on the subject. On the other hand, there have been magnetoencephalographic (MEG) studies on brain activities during speech imagery [5], [6]. MEG signals are measured in a less noisy environment, have less distortion during propagation from a signal source to sensors, and have higher spatial resolution than EEG signals [7]. Therefore, classifying imagined speech from MEG signals is easier than from EEG signals, but the amount of MEG data is likely to be smaller than the amount of EEG data because MEG measurements are usually large-scale.

In recent years, there have been many studies on representation learning using large-scale neural networks. Neural networks for various tasks using a portion of neural networks trained only with a large amount of text, speech, or image data without other labels have showed high performance [8]–[10]. Unlike such kinds of data, of course, it is difficult to collect a large amount of brain activity data and to learn feature representations of brain signals during speech imagery using a large-scale neural network. However, by using representations learned from as much brain activity data during different tasks in different conditions as possible, there is a possibility of improving the classification accuracy of imagined speech even when only a small amount of data is available during speech imagery. Among the representation learning methods, variational autoencoders (VAEs) [11] can be used with relatively small network structures and do not necessarily require a large dataset. VAEs have already been used for classifying EEG signals during motor imagery [12] and learning subject-independent EEG representations [13].

This work was supported in part by JSPS KAKENHI (Grant No. JP22K18626)

In this study, to improve the classification accuracy of imagined speech when only a small amount of data is available from each individual, we used VAEs to learn feature representations of MEG signals during speech imagery from multiple individuals. The trained encoders of the VAEs were transferred to convolutional neural networks (CNNs), which classified speech imagery MEG data from a new individual whose data were not used in the VAE training. We also trained conditional VAEs (CVAEs) [14] to generate MEG data for the new individual and augment the training data for the classifiers.

II. Materials and Methods

A. MEG Data

We used speech sound imagery MEG data measured in [5]. These data consist of MEG recordings from eight participants (seven males and one female, 20–40 years old). In each trial during the MEG measurements, the participants listened to a speech sound twice and then imagined the speech sound as they listened to it, without moving their tongue and mouth. Three speech sounds of three Japanese words—"*amagumo*" (rain cloud), "*ibento*" (event), and "*uranai*" (fortune-telling)—were used. The duration of the speech sounds was 800 ms. The time intervals between the listening and the imagery in each trial were 500 ms. Three kinds of trials corresponding to the three words were conducted repeatedly and randomly. The number of trials for each word and for each participant was at least 100. MEG signals during the trials were measured using a 122-channel whole-head neuromagnetometer (Neuromag-122™, Neuromag, Ltd., Helsinki, Finland) with a sampling frequency of 400 Hz.

The measured MEG signals were downsampled to a sampling frequency of 200 Hz and digitally filtered to remove low-frequency components below 1 Hz. Signals from abnormal MEG sensors were removed and spatially interpolated based on signals from the other normal sensors. -100–900 ms signals after the start of the first listening and the imagery were extracted as MEG epochs during speech listening and imagery, respectively. MEG epochs that included signals with the peak-to-peak amplitude above 1000 fT/cm were removed. Ocular activities were also removed from the MEG signals using independent component analysis. This preprocessing was performed by MNE-Python [15].

B. Network Architecture

An overview of the network architectures used in this study is shown in Fig. 1. VAE or conditional VAE (CVAE) was used to learn representations of the MEG signals during speech listening and imagery. EEGNet was used to classify the MEG data during the speech imagery.

1) EEGNet: EEGNet is a light-weight CNN architecture for EEG classification that uses depthwise separable convolution layers instead of ordinary convolution layers [16]. EEGNet shows competitive performance with ordinary CNNs when a small amount of training data is available. EEGNet consists of an ordinary convolution layer on the time axis, a depthwise convolution layer on the sensor axis, a separable convolution layer, and a fully connected (FC) layer. In this study, the architecture of EEGNet was modified to use two separable convolution layers.

2) Variational Autoencoder (VAE): A VAE consists of an encoder and a decoder: the encoder outputs the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ of the approximate probability distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ of a latent variable $\boldsymbol{z}$ given an input $\boldsymbol{x}$, and the decoder reconstructs $\boldsymbol{x}$ using $\boldsymbol{z}$ sampled from the distribution $q_{\boldsymbol{\phi}}$ [11]. The encoder and decoder are neural networks, and their parameters are denoted by $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively. The loss function for training a VAE is as follows:

$$\mathcal{L}_{\mathrm{VAE}} = \mathbb{E}_{q_{\boldsymbol{\phi}}}[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] + \mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))]. \quad (1)$$

The first term represents the reconstruction error between the original input $\boldsymbol{x}$ and the reconstructed input $\hat{\boldsymbol{x}}$. The second term represents the Kullback-Leibler divergence between the approximated posterior $q_{\boldsymbol{\phi}}$ and the prior probability distribution of $\boldsymbol{z}$, $p(\boldsymbol{z})$.

In this study, the encoder was composed of the convolutional layers of EEGNet and two FC layers. The decoder had a symmetrical structure to the encoder with transposed convolution layers instead of the convolution layers. The number of dimensions of $\boldsymbol{z}$ was set to 32. The mean squared error loss was used as the reconstruction error. The prior $p(\boldsymbol{z})$ was the multivariate standard normal distribution.

3) Conditional VAE (CVAE): In a CVAE, the approximate posterior q_ϕ and the probability distribution of the decoder, $p_{\boldsymbol{\theta}}$, are further conditioned on the label y corresponding to the input $\boldsymbol{x}$ [14]. Therefore, a CVAE is composed of the encoder and decoder of the VAE with an additional class label input. In this study, the class labels were one-hot encoded, and were linearly transformed into 4- or 8-dimensional vectors and combined with the input of the second FC layer of the encoder and the input of the decoder.

C. Imagined Speech Classification

EEGNet models that classified the MEG during the speech imagery were trained for each participant. These models are referred to as "within-participants models." The trained within-participants models were evaluated using 10-fold cross-validation. Each split set of the cross-validation was composed of successive MEG epochs. The temporally first 80% and the remaining 20% of the MEG epochs not included in the evaluation set were used for training and validating the models, respectively.

EEGNet models were also trained with MEG data from multiple participants. These models are referred to as "between-participants models." The trained between-participants models were evaluated using leave-one-out cross-validation with respect to the participant. The temporally first 80% and the remaining 20% of the MEG epochs not included in the evaluation set were used for training and validating the models, respectively.

The macro F1 score was used as the evaluation index for the trained EEGNet. This index is the average of the F1 scores

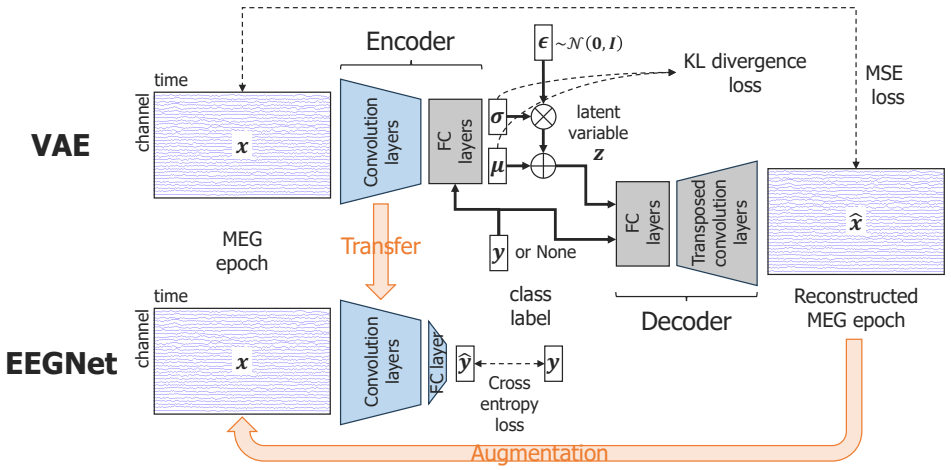

Fig. 1. Overview of the network architectures used in this study.

calculated for each class, i.e. macro-$F1 = \frac{1}{K}\sum_{k=1}^{K} F1_k$. The F1 score for the class k is defined as $F1_k = 2TP_k/(2TP_k + FP_k + FN_k)$ where TP_k, FP_k, and FN_k are the number of true positives, false positives, and false negatives for the class k, respectively. The EEGNet models were trained for up to 50 training epochs, and the model with the maximum macro F1 score at the validation was selected and evaluated using the evaluation set.

D. Transfer Learning

First, VAE models were trained with MEG data during the speech imagery from multiple participants. The training and validation data of the VAE models were the same as those of the between-participants model. Furthermore, MEG data during the first speech listening and the speech imagery were used for training the VAE models. Owing to the increase in the training data, it was expected that better feature representations of the MEG data would be extracted. The VAE models were trained for up to 500 training epochs, and the model with the minimum validation loss was selected.

The convolution layers of the trained VAE model were transferred to those of an EEGNet model, and the weights of only the FC layers or all the layers of the EEGNet model were fine-tuned. The data for fine-tuning, validating, and evaluating the transfer-learned model were from a target participant whose data were not used in the training and the validation of the VAE model, and were split in the same way as the within-participants model.

Since the transferred layers were trained using the data from the participants except for the target participant, feature representations extracted by the transferred layers may not be suitable for classifying the imagined speech of the target participant. Therefore, the VAE models were adapted to the data from the target participant before the transfer.

E. Data Augmentation

CVAE models were trained with MEG data during the speech imagery from multiple participants except for a target participant in the same way as the VAE models. The CVAE models were adapted to the MEG data from the target participant. The training data from the target participant were fed into the CVAE models, and one to four times the number of the training data were generated. The generated data were merged with the original training data. EEGNet models were trained using the augmented training data.

III. Results and Discussion

A. Transfer Learning

Fig. 2 shows the classification results of the EEGNet transfer-learned from the VAE model trained using the MEG data only during the speech imagery. The classification performance in six participants was improved when only the FC layers or the entire network were fine-tuned (w/ trans. (FC FT) or w/ trans. (entire FT), respectively) in comparison with when the entire network was trained without the transfer (w/o trans.). In four of these participants, the classification performance was further improved by adapting VAE to the data from each participant before the transfer (w/ adapt. & trans. (FC FT) or w/ adapt. & trans. (entire FT)). These results indicate that transfer learning using VAE models trained using MEG data from multiple participants and adaptation of the VAE models to a target participant are effective in classifying MEG signals during speech imagery.

Table I shows the classification results of the between-participants model. The macro F1 scores of the between-participants models tended to be slightly higher than those of the within-participants models. The standard deviation of the macro F1 scores of the between-participants models was smaller than that of the within-participants models. This may be due to the increase in the training data.

Fig. 3 shows the classification results of the EEGNet transfer-learned from the VAE models that were trained using the MEG data during both the speech imagery and listening. Despite the increase in the number of training samples for the VAE models, there was little improvement in the classification

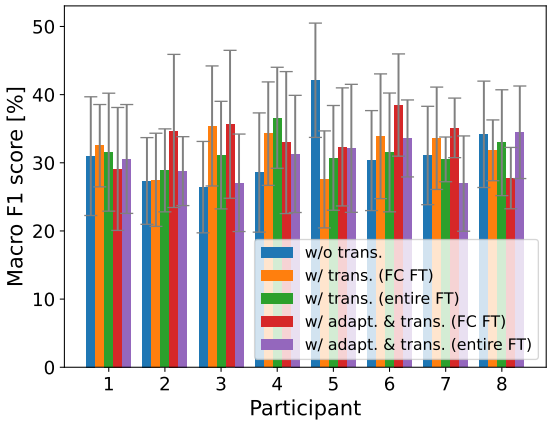

Fig. 2. Macro F1 scores of imagined speech classification with transfer learning using the VAE model trained with only speech imagery MEGs. Each error bar indicates the standard deviation.

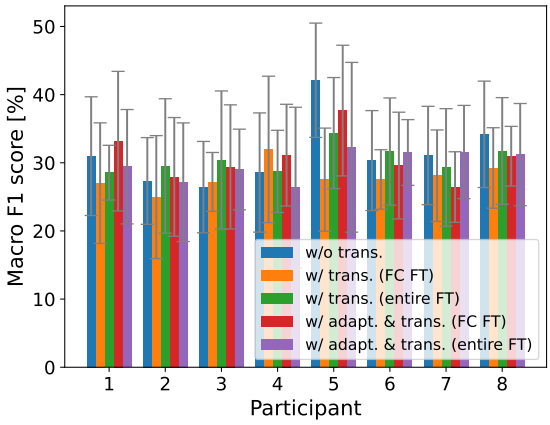

Fig. 3. Macro F1 scores of imagined speech classification with transfer learning using the VAE model trained with MEGs during speech imagery and listening to speech. Each error bar indicates the standard deviation.

TABLE I
MACRO F1 SCORE OF THE BETWEEN-PARTICIPANTS MODELS FOR IMAGINED SPEECH CLASSIFICATION

Tartget participant	Macro F1 score [%]
1	36.1
2	33.2
3	34.7
4	37.2
5	35.9
6	37.6
7	31.4
8	36.2
Mean $\pm$ SD	35.3 ± 2.1

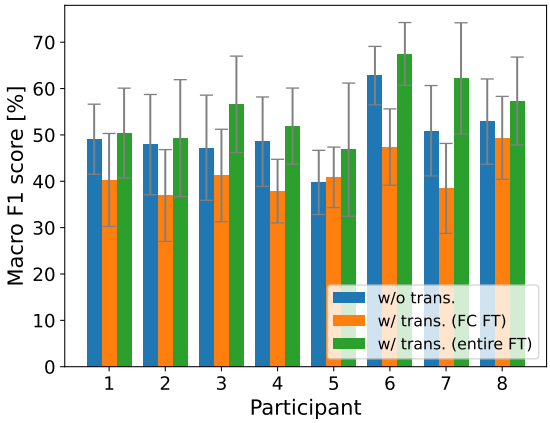

Fig. 4. Macro F1 scores of classifications of MEG while listening to speech with transfer learning. Each error bar indicates the standard deviation.

performance. It is possible that the MEG data during speech listening negatively affected the feature representations learned by the VAE models for the imagined speech classification.

To demonstrate the difference in the ease of learning representations between the MEG data during the speech imagery and listening, we trained EEGNet models to classify the MEG data during speech listening, and also conducted transfer learning from the VAE models trained using the MEG data during both the speech imagery and listening. Fig. 4 shows the results of classifying the MEG data during speech listening. The macro F1 scores were approximately 50%, and were higher than when classifying the MEG data during the speech imagery. The classification performance was improved by fine-tuning the entire network after the transfer. The classification performance was degraded when only the FC layers were fine-tuned because the feature representations obtained from the transferred encoder were deviated from the feature representations of the MEG data from the target participant, whose data were not used for training the VAE model.

This explains why the transfer learning from the VAE model trained using the MEG data during both the speech imagery and listening did not improve the performance of the imagined speech classification. In the training of the VAE models, the feature representations of the MEG data during speech listening may also be easier to be learned than those of the MEG data during speech imagery. Additionally, the macro F1 scores of the classification of MEG data during the speech imagery were lower than those of MEG data during speech listening, and closer to the chance rate (33.3%). The MEG activity associated with speech imagery might have been smaller than the MEG activity associated with speech listening, and been obscured by background MEG activity, which could have prevented the EEGNet model from training sufficiently.

B. Data Augmentation

Fig. 5 shows the classification performance of the EEGNet with data augmentation. For many participants, the classification performance was improved when the data were augmented. There was a tendency for the classification performance to improve as the data augmentation factor increased.

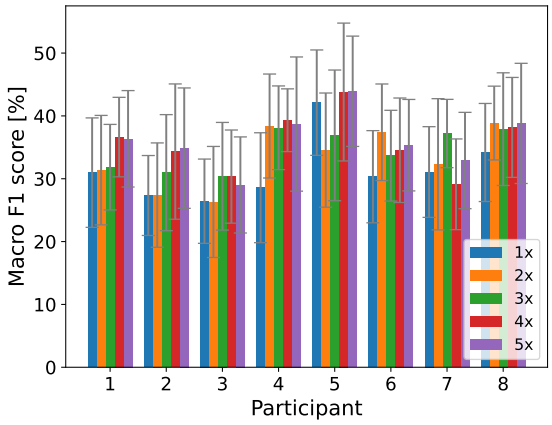

Fig. 5. Macro F1 scores of imagined speech classification with data augmentation. Each error bar indicates the standard deviation.

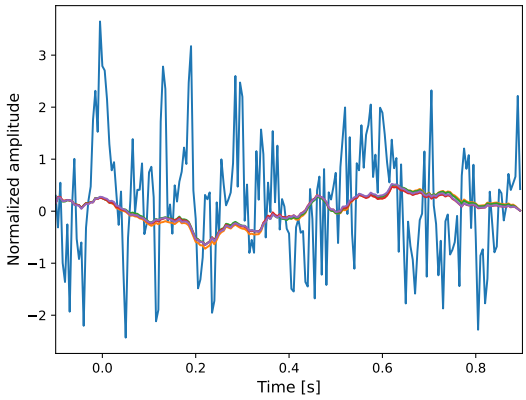

Fig. 6. Examples of an original MEG signal (blue) and MEG signals reconstructed by the CVAE model (other colors) at a left temporal MEG channel.

Fig. 6 shows examples of an original MEG signal and MEG signals reconstructed by the CVAE model. The reconstructed MEG signals did not have high-frequency components and showed little variation with sampling. This is likely because, when the CVAE training was stopped to prevent overfitting, the decoder had not yet acquired the ability to reconstruct the high-frequency components. Additionally, the MEG signals contain many low-frequency components, which may include useful information for classifying imagined speech. In our study, the EEGNet was trained with the augmented low-frequency MEG signals, which were embedded with class-specific features by the CVAE. This likely enabled the EEGNet to stably extract the low-frequency features related to speech imagery, resulting in improved classification performance. Furthermore, there was no significant improvement in the macro F1 score when the training data was increased from four to five times. It is considered that while additional augmentation of low-frequency MEG data stabilizes the training of the EEGNet, the improvement of the classification performance eventually saturates.

IV. Conclusion

In this study, we used VAE models to learn feature representations from the MEG signals of multiple participants during speech imagery. Transfer learning from the VAE encoders to EEGNet encoders improved performance of the MEG signal classification during the speech imagery. The EEGNet classifiers were also trained using augmented MEG data that contained samples generated by a conditional VAE model. The augmentation of the training data improved the classification performance of the classifiers. These results indicate that feature representations of MEG data from multiple individuals learned using VAE models can improve the classification performance of imagined speech of a new individual even when a limited amount of data is available from the new individual.

References

[1] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, “Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features,” J. Neural Eng., vol. 15, no. 1, December 2017, Art. no. 016002.

[2] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, “Single-trial classification of vowel speech imagery using common spatial patterns,” Neural Netw., vol. 22, no. 9, November 2009, pp. 1334–1339.

[3] F. Li et al., “Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network,” J. Neural Eng., vol. 18, no. 4, August 2021, Art. no. 0460c4.

[4] A. Hernandez-Galvan, G. Ramirez-Alonso, and J. Ramirez-Quintana, “A prototypical network for few-shot recognition of speech imagery data,” Biomed. Signal Process. Control, vol. 86, September 2023, Art. no. 105154.

[5] S. Uzawa, T. Takiguchi, Y. Ariki, and S. Nakagawa, “Spatiotemporal properties of magnetic fields induced by auditory speech sound imagery and perception,” in 39th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), 2017, pp. 2542–2545.

[6] D. Dash, P. Ferrari, and J. Wang “Deconding imagined and spoken phrases from non-invasive neural (MEG) signals,” Frontiers Neurosci., vol. 14, April 2020, Art. no. 290.

[7] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, “Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain,” Rev. Modern Phys., vol. 65, no. 2, April 1993, pp. 413–497.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: a framework for self-supervised learning of speech representations,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, “Masked autoencoders are scalable vision learners,” in IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16000–16009.

[11] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in 2nd Int. Conf. Learn. Representations (ICLR), 2014.

[12] M. Dai, D. Zheng, R. Na, S. Wang, and S. Zhang, “EEG classification of motor imagery using a novel deep learning framework,” Sensors, vol. 19, no. 3, January 2019, Art. no. 551.

[13] L. Bollens, T. Francart, and H. V. Hamme, “Learning subject-invariant representations from speech-evoked EEG using variational autoencoders,” IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2022, pp. 1256–1260.

[14] K. Sohn, X. Yan, and H. Lee, “Learning structured output representation using deep conditional generative models,” in Adv. Neural Inf. Process. Syst. (NIPS), 2015.

[15] A. Gramfort et al., “MEG and EEG data analysis with MNE-Python,” Frontiers Neurosci., vol. 7, December 2013, Art. no. 267.

[16] V. J. Lawhern et al., “EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces,” J. Neural Eng., vol. 15, no. 5, July 2018, Art. no. 056013.