SIP (2018), vol. 7, e21, page 1 of 11 © The Authors, 2018.

doi:10.1017/ATSIP.2018.25

ORIGINAL PAPER

Daily activity recognition based on recurrent neural network using multi-modal signals

AKIRA TAMAMORI[1], TOMOKI HAYASHI[2], TOMOKI TODA[3]AND KAZUYA TAKEDA[2]

Our aim is to develop a smartphone-based life-logging system. Human activity recognition (HAR) is one of the core techniques to realize it. Recent studies reported the effectiveness of feed-forward neural network (FF-NN) and recurrent neural network (RNN) as a classifier for HAR task. However, there are still unresolved problems in those studies: (1) a life-logging system using only a smartphone for recording device has not been developed, (2) only indoor activities have been utilized for evaluation, (3) insufficient investigations/evaluations of RNN. In this study, we address these unresolved problems as follows: (1) we build a prototype system for life-logging and conduct data recording experiment on this system to include both indoor and outdoor activities. The experimental results of HAR on this new dataset showed that RNN-based classifier was still effective. (2) From the results of a HAR experiment, it was demonstrated that a multi-layered Simple Recurrent Unit with a non-linear transform at the bottom layer and a highway-connection was the most effective. (3) We could grasp the reason for the improvement of RNN from FF-NN by observing the posterior probabilities over test data.

Keywords: Human activity recognition, Deep neural network, Recurrent neural network, Multi-modal signal

Received 23 June 2018; Revised 12 November 2018

I. INTRODUCTION

Being faced with an unprecedented super-aging society such as in Japan [1], we consider that a society will be required where sustainable social participation of elderly people is promoted and they can select an active and fresh lifestyle as long as they wish. However, there is a problem to be resolved in order to realize such society: a stay-at-home problem of elderly people. Their physical strength including muscular strength is prone to get weaker and this leads an inconvenience of walking. As a result, they tend to stay-at-home due to psychological and human factors [2]. By using a technology which can sense, record, and understand their daily activities, it may be possible to promote such elderly people to go out. An increase of opportunities for the elderly to go out will bring an increase of opportunities for their social participations. Our aim is to develop such a technology.

Figure 1 shows a concept image of our social implementation in the future. The system interacts with a user through the smartphone. Based on his/her recent activity history, the system will notify a message in order to promote the user to go out and start an activity. When the activity is over, the system will send a message to confirm the user's feeling. Based on the feedback and history, the system again sends a message to promote the next activity, and the cycle will be kept. We consider that human activity recognition (HAR) is one of the core techniques to realize this implementation. The objective of HAR system is to identify human activities from observed signals. The information of identified activity can be utilized for the promotion about going out. Various applications about HAR can be found, such as life-logging [3], monitoring the elderly [4], health care [5], and so on. As shown in Fig. 2, our target system can also be regarded as a life-logging system on the basis of HAR technique. In this study, we develop HAR technique for a smartphone-based life-logging system.

By using deep learning, many researchers have been working on sensor-based HAR technique [7, 8]. Therefore, just simply applying a method for HAR based on deep learning itself is no longer a novel idea, and development of the state-of-the-art deep learning technique for HAR is not our aim. The reason why we still adopt a method based on deep learning in this study is that it outperformed other traditional pattern recognition methods such as k-nearest neighbor, Gaussian mixture model (GMM), a decision tree, and support vector machine (SVM) [6, 9]. Additionally, the reason why we adopt the smartphone as a sensor device in this study is that users may not feel an obtrusiveness from it,

[1]Department of Information Science, Aichi Institute of Technology, 1247 Yachigusa, Yakusa-cho, Toyota, Aichi 470-0392, Japan
[2]Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan
[3]Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan

Corresponding author:

Akira Tamamori
Email: akira-tamamori@aitech.ac.jp

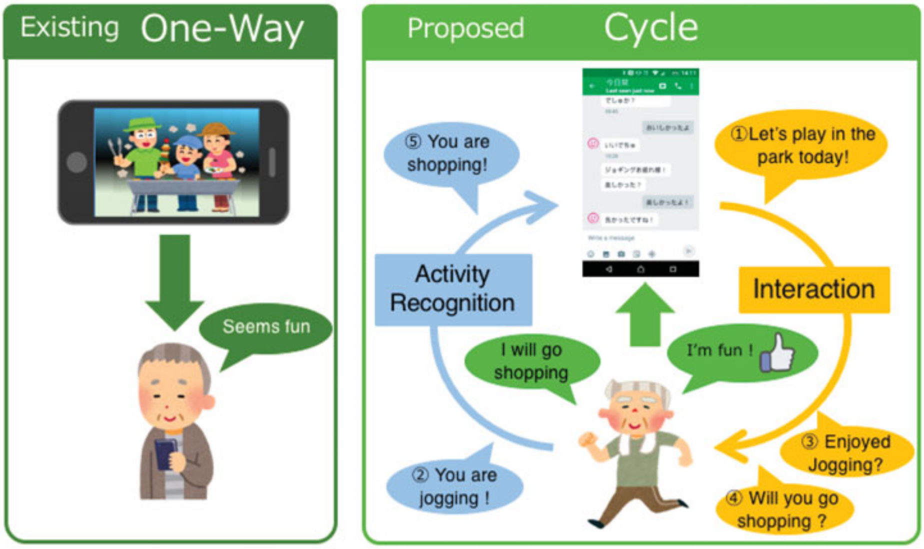


Fig. 1. A concept image of our social implementation.

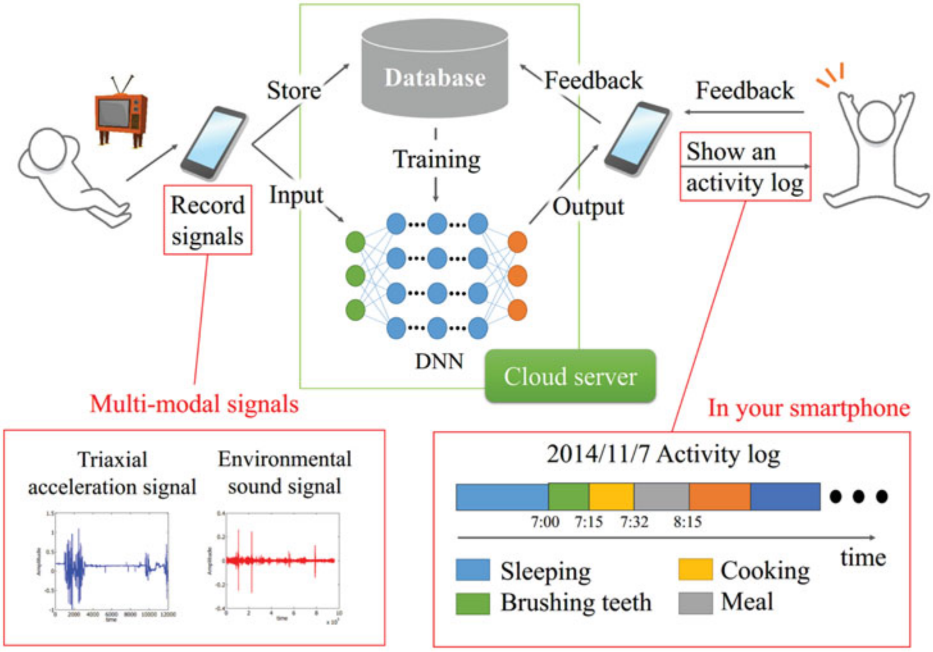


Fig. 2. Overview of our target life-logging system [6]; The system sends the recognition result, the subject's activity to their smartphone. The history of the user's activity can be viewed through a graphical user interface on the smartphone. The subject can send a feedback about his/her feeling about the activity. This feedback information can be used to improve the recognition performance.

compared with many body-worn sensors. Therefore, using a smartphone is suitable for many people to use our target system.

Before proceeding with our social implementation, it is necessary to address unresolved problems in the previous studies [6, 9]:

(i) A life-logging system using only a smartphone for recording device has not been developed yet.
(ii) Only indoor activities have been utilized for evaluation.
(iii) The authors have not demonstrated that a multi-layered long short-term Memory-recurrent neural network (LSTM-RNN) for the classifier outperforms a single layer one.
(iv) The effectiveness of RNN over feed-forward neural network (FF-NN) was evaluated by using only a F1-score.

The purpose of this paper is to strengthen the previous studies [6, 9] by addressing the above problems: For (i) and (ii), a prototype system could be built towards a realization of life-logging system using a smartphone. It was utilized to construct a new dataset which includes not only indoor but also outdoor activities. From the results of the HAR experiment conducted on the newly constructed dataset by using the prototype system, the effectiveness of RNN over FF-NN could be further demonstrated. For (iii), we further investigated better network architecture of RNN which could keep recognition performance even if it was multi-layered. By

Table 1. Data recording conditions of Nagoya-COI database [6]

Number of subjects	1 (long-term) + 18 (short-term)
Recording environment	One-room studio apartment
Instructions	Lead well-regulated life
	Acceleration signal
Recorded signals	Environmental sound signal
	Video

combining a highway connection and a single layer feed-forward network, we could demonstrate better results than those studies conducted on Nagoya-COI database. It was confirmed that the simple recurrent unit (SRU) [10], where a highway connection is incorporated, showed the highest recognition performance. For (iv), posterior probabilities over consecutive time steps were visualized, in order to visually grasp and explain the reason for better recognition performance of RNN than FF-NN. The advantage of RNN compared with FF-NN in a HAR task could be demonstrated.

Our contributions in this study can be summarized as follows:

- Towards a realization of the life-logging system using a smartphone, we showed a concrete framework of the prototype system for recording multi-modal signals.
- By using the newly constructed dataset which includes both indoor and outdoor activities, we could demonstrate that an RNN-based classifier was still effective and outperformed the FF-NN in a HAR task.

This paper is organized as follows. Section II describes a digest of the previous studies [6, 9] and the unresolved problems of these studies. Section III introduces the prototype system towards a realization of our target system. The results of experimental evaluations are given in Section IV. Finally, Section V concludes the paper.

II. A DIGEST OF PREVIOUS STUDY

First, we briefly review the previous studies [6, 9]. Next, we mention the unresolved problems to be addressed in this study.

A) Nagoya-COI daily activity database

The data recording condition is shown in Table 1. The recording environment was a one-room studio apartment. Each subject could freely live in the room and go outside with recording staffs, however, to prevent an idle living such as sleeping all day, they were instructed to lead a well-regulated life. An accelerated signal was recorded with an Android application (HASC Logger [11]). About 300 hours data of indoor activities were annotated. The two types of dataset were constructed: (1) long-term, single subject data of 48 h in length, (2) short-term, multiple subject data with a total length of 250 h. Table 2 lists the recorded daily activities.

Table 2. Recorded daily activities in Nagoya-COI database [6]

Activity name	Length (min)	Activity name	Length (min)
Others	3879	Cleaning	188
Sleeping	2731	Writing	150
Note-PC	2252	Cleaning bath	107
Smartphone	1959	Calling	104
Watching-TV	1873	Tablet	86
Cooking	1827	Light meal	85
Eating	908	Drying clothes	75
Cleaning table	679	Washing	36
Reading	476	Walking	30
Toilet	310	Monologue	5
Tooth brushing	310	Taking a bath	5

B) Experimental evaluations

The authors first conducted the subject-closed experiment on the Nagoya-COI database. In this experiment, the same subject's data were used in both the training and test phases. A total of 56-dimensional features was extracted from the multi-modal signals for each subject in the dataset and used as classifier inputs. The most frequently observed nine activities were used as the target activities, while all of the remaining activities were used as non-target activity [6]. The authors also conducted the subject–subject experiment where the same subject's data were used in both the training and test phases. From the results of the subject-closed experiment, they demonstrated that a classifier based on a FF-NN outperformed other popular classifiers such as GMM and SVM, and LSTM-RNN further outperformed the FF-NN (see Fig. 3). The authors also conducted the subject-open experiment where the data of different subjects were used in the training and test phases. From the results of the subject-open experiment, the adaptation method that all of the layers was re-trained gave the best performance.

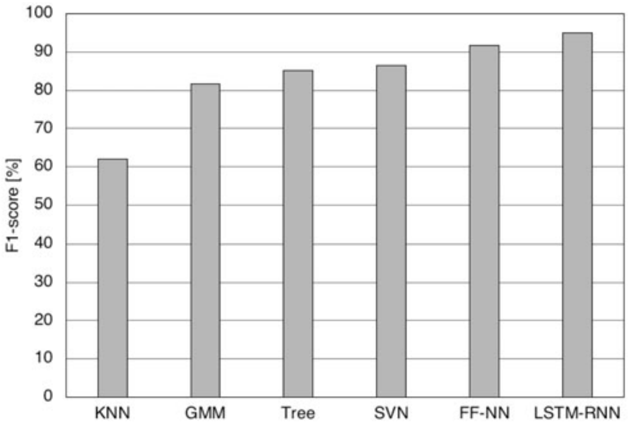


Fig. 3. Performance of daily activity recognition [9]; a comparison with other popular methods. "KNN" and "Tree" represent *k*-nearest neighbor and a decision tree, respectively.

https://doi.org/10.1017/ATSIP.2018.25 Published online by Cambridge University Press

C) Unresolved problems

In the previous studies [6, 9], the Nagoya-COI daily activity database was utilized for evaluation of HAR. We consider that there is a mismatch between the performance evaluation and our aim. Although the target activity should include not only indoor but also outdoor activities, the authors have not evaluated the recognition performance for outdoor activities. Moreover, a concrete life-logging system using only a smartphone for recording device has not been developed. Therefore it is still needed to evaluate the performance of the recognition part which will be incorporated into the target system.

Furthermore, an LSTM-RNN with single hidden layer was applied. This is because some results have been obtained from a preliminary experiment that the recognition performance degraded when using a multi-layered LSTM-RNN. We consider that this is due to a vanishing gradient problem in optimizing the network parameters and an overfitting problem, and these problem can be further mitigated by introducing a simple and suitable network architecture such as highway connection [12]. This will be beneficial for realization the target system and can be incorporated into it. Lastly, the effectiveness of RNN over FF-NN was evaluated by using only F1-score. We consider that it is insufficient to account for the effectiveness and an additional experiment will be needed.

III. TOWARDS REALIZATION OF TARGET LIFE-LOGGING SYSTEM

In order to construct a new dataset of multi-modal signals which include both indoor and outdoor activities, a prototype system for data recording was built. Figure 4 shows the framework of the prototype system. From a microphone and an acceleration sensor in user's smartphone, HASC Logger [11], an Android application, records both sound and acceleration signals. We have modified the original HASC Logger so that MFCC, zero crossing rate (ZCR) and root mean square (RMS) can be extracted from the raw sound waveform stored temporary in the smartphone, in consideration to user's privacy. The extracted acoustic features and acceleration signals are then uploaded every 11 seconds and stored in the temporary database. After a feature extraction from acceleration signals on the temporary database, the acoustic and acceleration features are concatenated and sent to the recognition engine where the RNN-based classifier is utilized. After loading the network parameters of RNN, the engine is driven and the recognition results from the engine are stored in the activity database. From the temporary database, the acoustic and acceleration features are uploaded once a day to Nagoya-COI Data Store [13] which is a data storage server specially developed by the Center of Innovation Program (Nagoya-COI). It should be noted that multiple users can use this system in parallel. The system can accept the data uploaded asynchronously from each user. The monitoring controller checks whether the smartphone is active or not during an operation. This feature has been introduced to support a stable and continuous operation of the system. Although we have not developed a notification/browsing application, it is necessary for our social implementation in future.

IV. EXPERIMENTAL EVALUATION

A) Experiment on Nagoya-COI database

A subject-closed experiment was conducted in order to compare the architecture of RNN where the same subject's data was used in both the training and test phases. A subject-open experiment was also conducted in order to

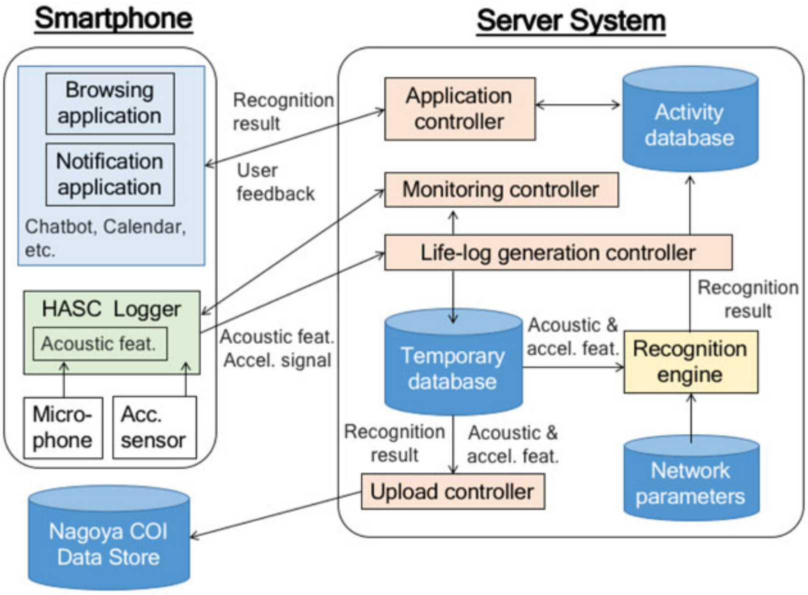


Fig. 4. Framework of the prototype system.

https://doi.org/10.1017/ATSIP.2018.25 Published online by Cambridge University Press

compare the recognition performance of FF-NN with RNN using a leave-one-subject-out validation. In this experiment, the data of different subjects was used in the training and test phases. Moreover, a subject-adaption experiment was conducted. Finally, we visualized the posterior probabilities on test data, in order to grasp the reason why better recognition performance of RNN than FF-NN was obtained.

1) Experimental conditions

In this paper, the following variants of RNN were applied:

- Simple RNN (SRNN) [14]: It calculates hidden vector sequences and output vector sequences through a linear transform and an activation function.
- Simple Recurrent Unit (SRU) [10]: This can be viewed as a special case of Quasi-RNNs [15]. The forget gate and the reset gate at current time step do not require the hidden vector at previous time step. The highway connection is also introduced between the input and output.
- Minimal Gated Unit (MGU) [16].
- Gated Recurrent Unit (GRU) [17].

Those architectures are simpler than LSTM-RNN in terms of the number of network parameters. In fact, the number of weight matrices of SRNN (SRU), MGU, GRU, are about 25, 50, and 75% of a vanilla LSTM-RNN, respectively. The details of these variants are described in Appendix.

Table 3 and 4 list the number of activities in the subject-closed experiment and the subject-open experiment, respectively. The feature vectors were extracted from the dataset by the same manner as the previous studies [9]; A total of 56-dimensional features was used as classifier inputs; The environmental sound signal and the acceleration signal were synchronized, the features were extracted from each frame. The frame size and shift size were set to both 1 second. From the windowed environmental sound signals, the 41-dimension feature vectors are extracted: 13-order Mel-Frequency Cepstral Coefficients (MFCC) with its 1st and 2nd order derivative coefficients, ZCR and RMS. MFCC is a feature reflecting human aural characteristics. ZCR and RMS represent volume and pitch, respectively. From the windowed acceleration signals, the 15-dimension feature vectors are extracted: the mean, variance, energy and entropy in the frequency domain for each axis, and the correlation coefficients between these axes.

From our preliminary results, the number of units per one hidden layer of RNN was set to 512. The length of the unfold of RNN was set to 60 frames. The loss function of the network was a cross-entropy loss. The optimization algorithm was back-propagation through time via Adam [18] and the learning rate was fixed to 0.001. The minibatch size was set to 128. For regularization, we applied the standard dropout [19] and added a L2 loss term to the loss function. All of the networks were trained using the open source toolkit, TensorFlow [20] with a single GPU.

Table 3. Target activities in subject-closed experiment conducted on Nagoya-COI database [9]

Activity name	Length (min)	Activity name	Length (min)
Cleaning	39	Sleeping	1257
Cooking	108	Smartphone	198
Meal	120	Toilet	61
Note-PC	141	Watching-TV	109
Reading	164	Other	582

Table 4. Target activities in subject-open experiment conducted on Nagoya-COI database [9]

Activity name	Length (min)	Activity name	Length (min)
Cleaning	67	Sleeping	2731
Cooking	1826	Smartphone	1959
Meal	908	Toilet	310
Note-PC	2257	Watching-TV	1873
Reading	476	–	–

Throughout these experiments, a hold-out validation method for evaluation was adapted because the number of samples for each activity class is different. In this validation method, 10 test samples are randomly selected for each class and the rest is used for training, and this procedure was repeated 10 times. For evaluation, the following averaged F1-score was adopted:

$$F = \frac{1}{10}\sum_{r=1}^{10}\left(\frac{1}{C}\sum_{c=1}^{C}F_c^{(r)}\right), \qquad (1)$$

where C is the number of classes to be recognized and $F_c^{(r)}$ is the F1-score of the class c at the r-th repetition, respectively.

2) Results of experiment

Figure 5 shows the recognition performance when the number of weight matrices and gates in LSTM-RNN was reduced; we applied SRNN, MGU, and GRU for comparison. The "Frame" represents a frame level accuracy and "Sample" represents a sample level accuracy, which is the recognition accuracy obtained using the majority vote of the frame recognition results in each sample. The number

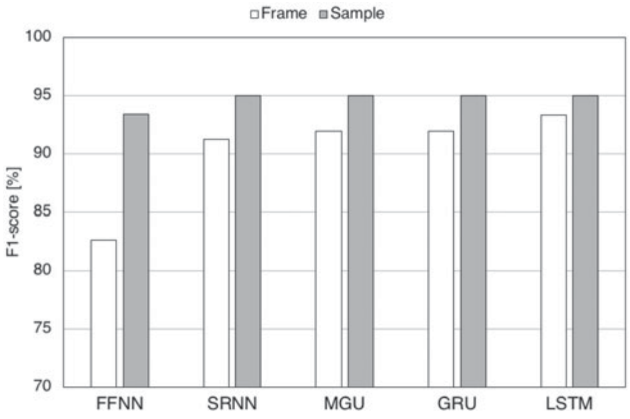


Fig. 5. Performance of daily activity recognition in comparison with architectures of LSTM-RNN: the number of parameters, i.e., weight matrices and gates. The number of hidden layers was set to 1.

https://doi.org/10.1017/ATSIP.2018.25 Published online by Cambridge University Press

of hidden layers was set to 1. At first, it can be observed that even SRNN, which has the most simple architecture, exceeded FF-NN in performance. This result means that the recurrent architecture of RNN has a high discrimination ability in modeling sequential data. While a large difference between the variants of LSTM-RNN in terms of sample level F1-score could not be observed, the best result was given by the LSTM-RNN in terms of frame level F1-score. These results suggest that the number of parameters of LSTM-RNN is a little excessive in this HAR task.

Figure 6 shows the recognition performance when the highway connection was incorporated and a non-linear transform of input at the bottom layer was applied. It can be observed from the figure that the recognition performance of LSTM-RNN degraded when the number of layers was increased. On the one hand, the recognition performance was improved from the single layer LSTM-RNN when the multi-layered FFNN-LSTM was applied. On the other hand, only applying highway connections to LSTM-RNN did not achieve a significant improvement. It can be considered that the cause of the degradation due to the multi-layered LSTM-RNN was that the input feature was not embedded in a space suitable for discrimination, and the vanishing gradient problem was not the main cause. Furthermore, by applying highway connections to FFNN-LSTM, it was possible to obtain an improvement. This is because the attenuation of the gradient was further suppressed. In addition to the above factors, i.e., a non-linear embedding and highway connections, due to the reduction of the number of network parameters, the FFNN-SRU could obtain the best performance.

Figure 7 shows the results of a subject-open experiment to compare FF-NN with FFNN-SRU, which gave the best performance in the subject-closed experiment. In this figure, "Sample" level accuracy is shown. We can see that FFNN-SRU outperformed FF-NN even in the subject-open setting. However, the performance in the subject-open evaluation was still lower than the subject-closed evaluation even if FFNN-SRU was applied. As already discussed in [6], this is because there are large differences in subject behavior and the orientation of the smartphone was not normalized. Therefore, we consider it is necessary for future work to extract a new feature which is robust and independent to the variation of the orientation of the smartphone.

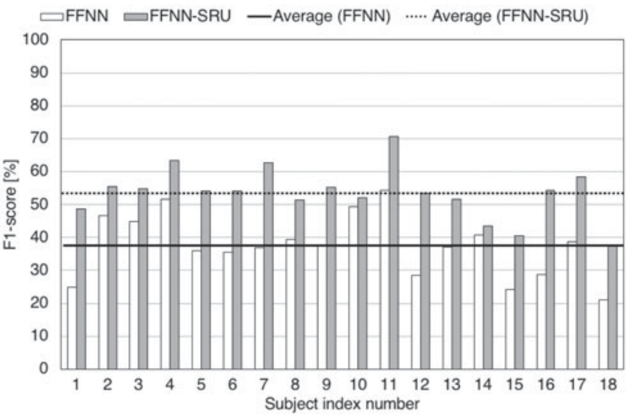


Fig. 7. Performance of daily activity recognition; leave-one-subject-out evaluation.

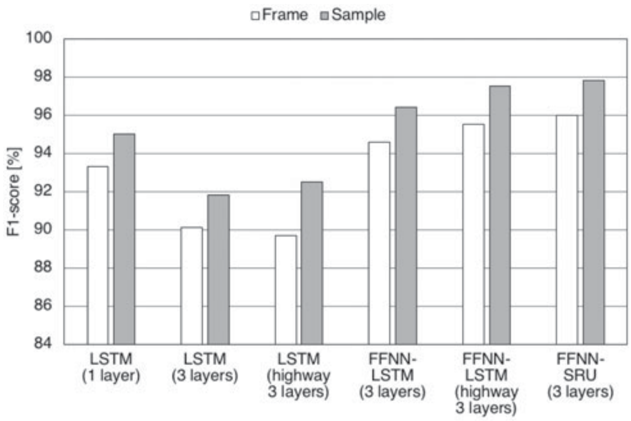


Fig. 6. Performance of daily activity recognition in comparison with architectures of RNN: highway connection and non-linear transform of input. "*n* layer (s)" means that the number of hidden layer is set to *n*.

Figure 8 shows the results of subject adaptation experiment to compare FF-NN, LSTM-RNN and FFNN-SRU. We can see that RNN-based classifier outperformed FF-NN even when subject adaptation task. Moreover, FFNN-SRU still performed better than LSTM-RNN.

3) Visualization of posterior probability on test data

In this section, we try to visually grasp the reason for the improvement of recognition performance, by seeing how the problem of the width of the temporal context (see Section II.C) could be resolved. A visualization of posterior probabilities over consecutive time steps on test data is shown in Fig. 9. Both FF-NN and RNN were the same ones as evaluated in Section A. In those figures, the horizontal axis represents the number of samples and the vertical axis represents the posterior probability, and the vertical dotted line is the boundary of the sample. Since the test data consists of a holdout set in the sample unit, the continuity of time is not maintained between samples. Comparing with FF-NN, misclassifications could be reduced significantly by

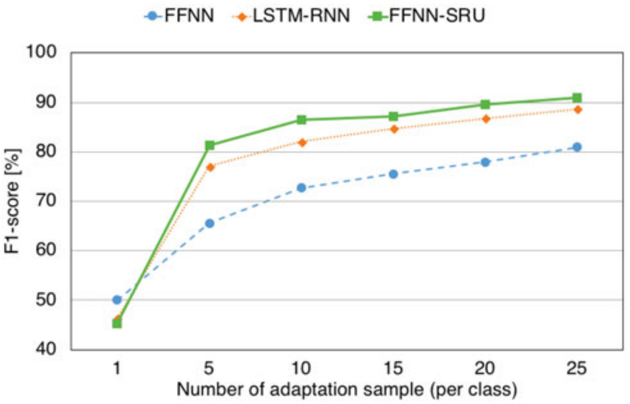
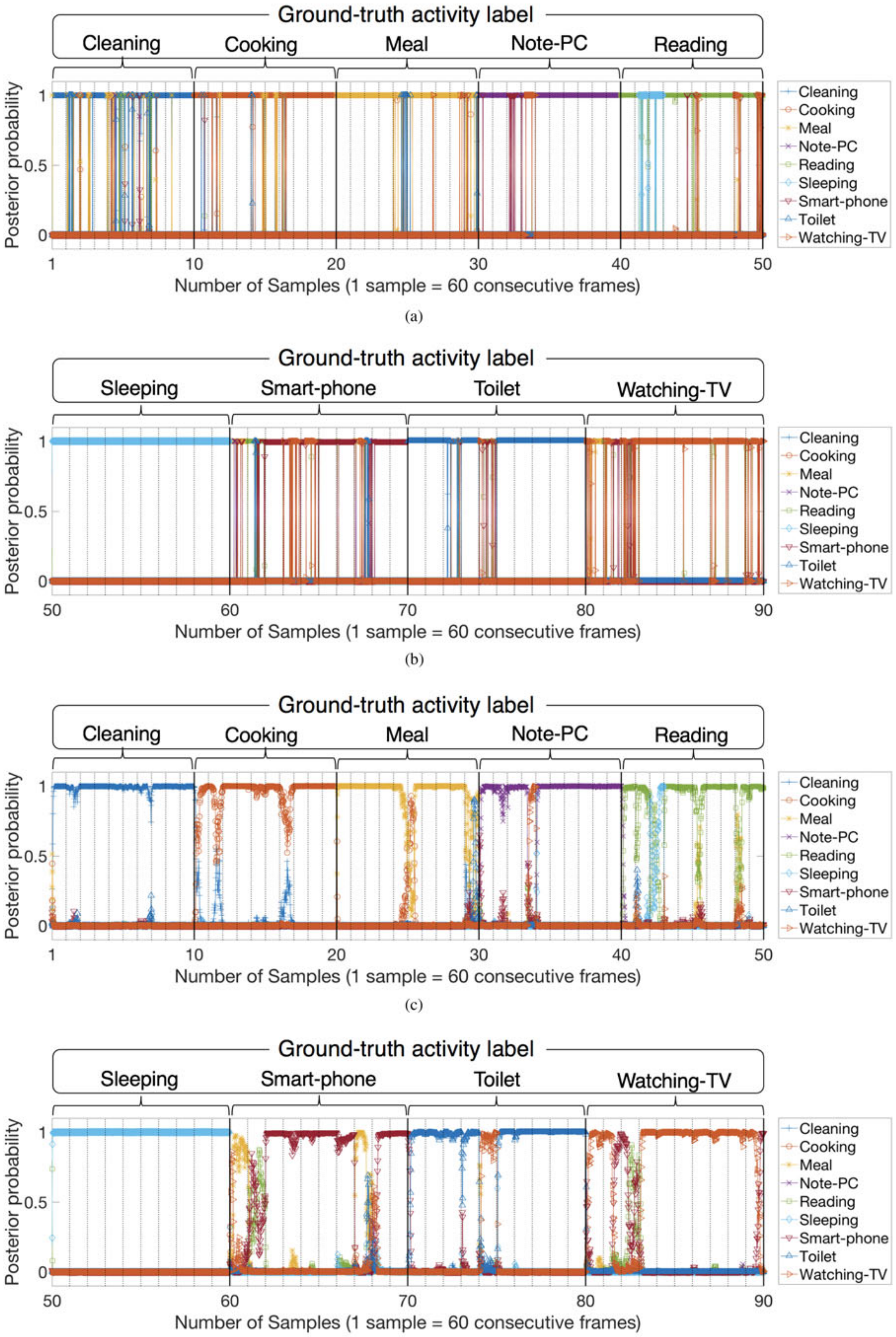


Fig. 8. Performance of daily activity recognition; subject adaptation.

https://doi.org/10.1017/ATSIP.2018.25 Published online by Cambridge University Press



(a)

(b)

(c)

(d)

Fig. 9. Visualization of posterior probability over consecutive frames of test data. (a) FF-NN: From “Cleaning” to “Reading” (b) FF-NN: From “Sleeping” to “Watching-TV” (c) LSTM-RNN: From “Cleaning” to “Reading” (d) LSTM-RNN: From “Sleeping” to “Watching-TV”.

Table 5. Data recording conditions

Number of subjects	10
Recording environment	Indoor and outdoor in daily life
Instructions	perform normal daily activities
Recorded signals	Acceleration signal Environmental sound signal

Table 7. Target activities of newly constructed dataset

Activity name	Activity name
Car-driving	Other
Cleaning	Reading
Cooking	Shopping
Cycling	Talking
Meal	Walking
Office	Watching-TV

LSTM-RNN. In fact, at a frame level accuracy, 81.57% (FF-NN) and 90.28% (LSTM-RNN) were obtained. Many misclassifications from the "Cleaning" class shown in Fig. 9(a) were suppressed over almost all the frames in Fig. 9(c). This result clearly shows an advantageous characteristic of RNN in the prediction that past frames influences the prediction of a future frame. This characteristic, a sequential prediction, does explain why the significant improvement of recognition performance was obtained by applying RNN.

B) Experiment on newly constructed dataset

1) Experimental conditions

By using the prototype system, we conducted an additional data recording. The number of subjects was 10, e.g., university students, housewife, and office workers. A smartphone holder was attached to the rear pocket of the subject's trousers. They were instructed to put the smartphone into the holder with a fixed orientation every time. They were also instructed so that they record the beginning and end time of the activity if they performed one of the activities as listed in Table 7. An Android application, which was originally developed by us for this experiment, was utilized to record both times efficiently. Therefore, the recorded data were firstly annotated by each subject. Tables 5 and 6 show the data recording conditions and the recorded daily activities, respectively. Using the information of annotation, we again tagged all of the activities of subjects. It should be noted that the signals of "Walking" were recorded in the indoor and outdoor environment. Compared with the Nagoya-COI database, "Car driving", "Shopping", "Cycling" activities are newly added. The "Other" contains signals recorded in a restaurant (eating out) and tooth-brushing.

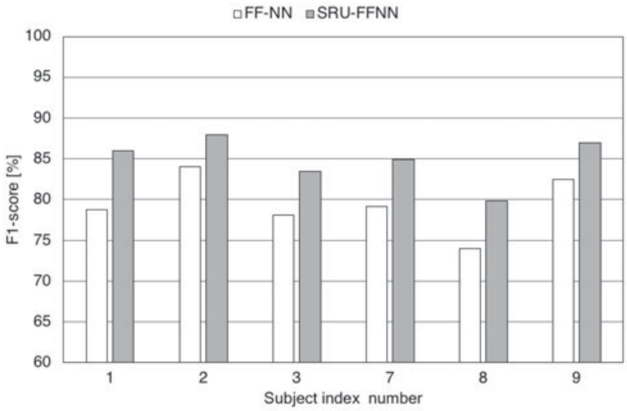


Fig. 10. Performance of daily activity recognition on newly constructed dataset.

In this experiment, a subject-closed experiment was conducted on the newly constructed dataset. The subject #1, #2, #3, #7, #8, and #9 were selected for evaluation because at least one of "Cycling", "Car-driving", and "Shopping", were recorded by these subject, all of which are outdoor activities. The SRU-FFNN was applied for an RNN-based classifier, where it consists of three hidden layers and a single layer of FF-NN at the bottom layer. The same FF-NN as the previous study [6] was also applied for comparison. The conditions of network training were set to the same as the previous experiment. For evaluation, a hold-out validation method was applied.

Table 6. Recorded daily activities using prototype system for each subject; the number in each cell represents the activity length in hours.

Activity name	sub. #1	sub. #2	sub. #3	sub. #4	sub. #5	sub. #6	sub. #7	sub. #8	sub. #9	sub. #10
Car-driving	4.1	–	–	–	–	–	2.5	3.1	2.3	–
Cleaning	–	–	0.77	–	–	1.9	0.3	2.9	–	–
Cooking	–	–	0.38	0.17	–	5.8	–	2.9	–	–
Cycling	–	2.3	0.05	–	0.54	–	–	–	–	–
Meal	4.1	–	3.8	1.9	–	5.7	–	5.1	–	–
Office	33	15	25	1.7	–	26	22	–	29	36
Other	–	0.8	7.5	–	0.43	–	8.0	–	3.5	2.1
Reading	6	4.2	–	1.1	–	–	–	–	–	–
Shopping	–	–	2.7	–	–	–	3.0	1.8	0.4	0.24
Talking	1.1	–	–	–	0.39	–	3.9	–	–	–
Walking	3.3	3.0	5.5	0.36	0.42	18	8.3	–	8.0	2.2
Watching-TV	5.2	–	–	–	–	4.4	3.0	3.9	–	–

https://doi.org/10.1017/ATSIP.2018.25 Published online by Cambridge University Press

Table 8. Confusion matrix of subject #8. Diagonal elements represent recall, that of the right-end column represent precision, and that of the lower end row represent F1-score.

Recall	Car-driving	Cleaning	Cooking	Meal	Shopping	Watching-TV	Precision
Car-driving	94.4	3.4	0.0	2.2	0.0	0.0	100
Cleaning	8.6	71.4	17.4	2.9	0.0	0.0	57.8
Cooking	0.0	0.0	74.3	22.9	0.0	2.8	70.8
Meal	0.0	0.0	4.9	73.8	0.0	21.3	80.4
Shopping	0.0	5.0	5.0	0.0	90.0	0.0	89.4
Watching-TV	0.0	2.1	19.6	4.3	0.0	73.9	86.2
F1-score	94.7	65.0	72.3	76.9	91.8	78.1	79.8

2) Results of experiment

Figure 10 shows the results of the experiment for each subject. In this figure, "Sample" level F1-score is shown. We can see that the classifier based on RNN was still more effective than that of FF-NN even if the dataset includes the outdoor activities. Figure 8 shows the confusion matrix of subject #8. This seems intuitive; for example, we can infer the "Car-driving" is likely to be far from other indoor activities such as "Cooking" and "Watching-TV", because we can feel or imagine its unique environmental sound and acceleration (vibration) in the car. The "Shopping" includes some body movements such as walking, standing, and ascending/descending a staircase. Not only those movements, but also some announcements broadcasted and background music played in the shop can be observed.

V. CONCLUSION

In this study, towards the realization of a smartphone-based life-logging system, the unresolved problems in the previous studies were addressed. First, a prototype system was successfully built and utilized to construct a new dataset which includes not only indoor but also outdoor activities such as "Cycling", "Car-driving", and "Shopping". From the results of a subject-closed experiment conducted on the new dataset, it was confirmed that RNN-based classifier for HAR was still effective for indoor and outdoor activities. Next, we investigated variants of LSTM-RNN for HAR from the viewpoint of the number of matrices, gates and layers with the help of highway connection. From the results of a subject-closed, open, and adaptation experiments, it could be demonstrated that the SRU with 3 hidden layers and a non-linear conversion of input by a single layer FF-NN was the most effective. Finally, we visualized the posterior probabilities of RNN over consecutive frames on test data, and could partly explain the reason for the improvement from FF-NN.

In future, we will conduct a large data recording experiment and investigate the effectiveness and validity of our data collection scheme. We will also develop a notification/feedback application for the target system. Development of a useful feature for HAR which is robust to the variation of the orientation of smartphone will also be a future work.

ACKNOWLEDGMENT

This research is supported by the Center of Innovation Program (Nagoya-COI) from Japan Science and Technology Agency.

REFERENCES

[1] Cabinet Office, Government of Japan: Aging society white paper in fiscal, 2018.

[2] Holzinger, A.; Searle, G.; Nischelwitzer, A.: On some aspects of improving mobile applications for the elderly, in Int. Conf. on Universal Access in Human-Computer Interaction, Springer, 2007, 923–932.

[3] Gurrin, C.; Smeaton, A.F.; Doherty, A.R.: Lifelogging: personal big data. *Found. Trends. Inf. Ret.*, **8** (1) (2014), 1–125.

[4] Rajasekaran, M.P.; Radhakrishnan, S.; Subbaraj, P.: Elderly patient monitoring system using a wireless sensor network. *Telemed. E Health*, **15** (1) (2009), 73–79.

[5] Liang, Y.; Zhou, X.; Yu, Z.; Guo, B.: Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare. *Mobile. Netw. Appl.*, **19** (3) (2014), 303–317.

[6] Hayashi, T.; Nishida, M.; Kitaoka, N.; Toda, T.; Takeda, K.: Daily activity recognition with large-scaled real-life recording datasets based on deep neural network using multi-modal signals. *IEICE Trans. Fund. Electron. Comm. Comput. Sci.*, **E101.A** (1) (2018), 199–210.

[7] Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L.: Deep learning for sensor-based activity recognition: a survey. *Pattern. Recognit. Lett.*, (2018), in press. https://www.sciencedirect.com/science/article/abs/pii/S016786551830045X.

[8] Nweke, H.F.; Teh, Y.W.; Al-garadi, M.A.; Alo, U.R.: Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. *Expert. Syst. Appl.*, **105** (2018), 233–261.

[9] Tamamori, A.; Hayashi, T.; Toda, T.; Takeda, K.: An investigation of recurrent neural network for daily activity recognition using multi-modal signals, in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC), 2017, 1334–1340.

[10] Lei, T.; Zhang, Y.; Artzi, Y.: Training RNNs as Fast as CNNs. arXiv preprint arXiv:1709.02755, 2017.

[11] Human Activity Sensing Consortium: HASC Logger. http://hasc.jp/hc2011/hasclogger.html.

[12] Srivastava, R.K.; Greff, K.; Schmidhuber, J.: Training very deep networks, in Cortes, C.; Lawrence, N.D.; Lee, D.D.; Sugiyama, M.; Garnett, R.: Eds., Advances in Neural Information Processing Systems 28, Curran Associates, Inc., Red Hook, NY, 2015, 2377–2385.

https://doi.org/10.1017/ATSIP.2018.25 Published online by Cambridge University Press

[13] Nagoya-COI Data Store. http://ds.coi.nagoya-u.ac.jp/.

[14] Elman, J.L.: Finding structure in time. *Cogn. Sci.*, **14** (2) (1990), 179–211.

[15] Bradbury, J.; Merity, S.; Xiong, C.; Socher, R.: Quasi-recurrent neural networks, in Proceedings of the 3rd Int. Conf. on Learning Representations (ICLR), 2017.

[16] Zhou, G.B.; Wu, J.; Zhang, C.L.; Zhou, Z.H.: Minimal gated unit for recurrent neural networks. *Int. J. Autom. Comput.*, **13** (3) (2016), 226–234.

[17] Cho, K. *et al.*: Learning phrase representations using rnn encoder-decoder for statistical machine translation, in Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing, ed. A. for Computational Linguistics, Doha, Qatar, 2016, 1724–1734.

[18] Kingma, D.P.; Ba, J.: Adam: a method for stochastic optimization, in Proc. of the 3rd Int. Conf. on Learning Representations (ICLR), 2014.

[19] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15** (1) (2014), 1929–1958.

[20] TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/.

[21] Sak, H.; Senior, A.W.; Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in Proc. Interspeech 2014, 2014, 338–342.

[22] Nair, V.; Hinton, G.: Rectified linear units improve restricted Boltzmann machines, in Proc. of the 27th Int. Conf. on Machine Learning, ICML-10, AAAI Press, 2010, 807–814.

[23] Pundak, G.; Sainath, T.N.: Highway-lstm and recurrent highway networks for speech recognition, in Proc. Interspeech 2017, 2017, 1303–1307.

APPENDIX: VARIANTS OF LSTM-RNN

In this appendix, we review the LSTM-RNN and its variants.

Simple recurrent neural network

Given input vector sequences, simple RNN (SRNN) [14] calculates hidden vector sequences and output vector sequences through a linear transform and an activation function. A hidden vector $\boldsymbol{h}_t$ is calculated from $\boldsymbol{h}_{t-1}$ and $\boldsymbol{x}_t$ through an activation function, a hyperbolic tangent function in this paper.

$$\boldsymbol{h}_t = \phi(\boldsymbol{W}_h[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_h), \tag{A.1}$$

where ϕ represents an activation function; we applied a hyperbolic tangent function in this paper. $\boldsymbol{W}_h$ and $\boldsymbol{b}_h$ represent the weight matrix and bias term, respectively.

RNN with long short-term memory

The LSTM-RNN has the special architecture, LSTM memory blocks. The hidden units in SRNN are replaced with it. The LSTM memory block contains memory cell which stores past information of the state, and gates which control the duration of storing. The LSTM-RNN can capture long-term context by representing information from past inputs as hidden vector and propagating it to future direction. The following sections describe several variants of LSTM-RNN with much simpler architecture.

The architecture can be written as:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i), \tag{A.2}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f), \tag{A.3}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o), \tag{A.4}$$

$$\tilde{\boldsymbol{c}}_t = \phi(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c), \tag{A.5}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \tag{A.6}$$

$$\boldsymbol{h}_t = \phi(\boldsymbol{c}_t) \odot \boldsymbol{o}_t, \tag{A.7}$$

where $\boldsymbol{c}_t$ is the state of memory cell. $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{o}_t$ represent input gate, forget gate, and output gate for memory cell, respectively, and $\odot$ is a Kronecker product operator (element-wise product). Thanks to these gates, LSTM-RNN can capture long-term context by representing information from past inputs as hidden vector and propagating it to future direction.

LSTM-RNN WITH RECURRENT PROJECTION LAYER

The LSTM-RNN with Recurrent Projection Layer (LSTMP-RNN), has been proposed to reduce the number of parameters in LSTM-RNN [21], where a linear transform (projection layer) is inserted after an LSTM layer. The output of the projection layer then goes back to the LSTM layer. In the LSTMP-RNN, equation (A.7) is replaced with the following equation:

$$\boldsymbol{h}_t = \boldsymbol{W}_p(\phi(\boldsymbol{c}_t) \odot \boldsymbol{o}_t), \tag{A.8}$$

where $\boldsymbol{W}_p$ is the projection matrix with the size of $P \times N$, N is the dimension of the output vector of $\phi(\boldsymbol{c}_t) \odot \boldsymbol{o}_t$, and P is the dimension after a linear transformation. If the projected dimension, P, is set to satisfy $P < N$, If the projected dimension is set properly, the number of parameters in LSTM block can be reduced significantly.

LSTM-RNN WITH NON-LINEARLY TRANSFORMED INPUT

In this study, we utilize a network architecture; the original input of LSTM-RNN is transformed by a non-linear function beforehand. Concretely, we compute the $\tilde{\boldsymbol{x}}_t$ in equation (A.9) from the input $\boldsymbol{x}_t$, and then the computations from equations (A.2) to (A.5) are applied:

$$\tilde{\boldsymbol{x}}_t = \psi(\boldsymbol{W}_h\boldsymbol{x}_t + \boldsymbol{b}_h), \tag{A.9}$$

where $\psi(\cdot)$ is a non-linear function. In this paper, we use a Rectified Linear Unit (ReLU) function [22]. We refer to this architecture as "FFNN-LSTM".

LSTM-RNN WITH HIGHWAY CONNECTION

In this study, we newly apply LSTM-RNN with highway connection [23] to resolve a vanishing gradient problem of multi-layer LSTM-RNN. By using the highway connection, it is expected that the less attenuated gradient will be propagated to the lower layer in the optimization. The difference between the original LSTM-RNN and the LSTM-RNN with highway connection can be written as:

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{W}\boldsymbol{x}_t, \tag{A.10}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \phi(\boldsymbol{c}_t) + (\mathbf{1} - \boldsymbol{o}_t) \odot \tilde{\boldsymbol{x}}_t. \tag{A.11}$$

The linear transform in equation (A.10) is applied to match the dimension of the input with the output gate.

Gated recurrent unit and minimal gated unit

Gated Recurrent Unit (GRU) [17] have the following architecture:

$$\boldsymbol{u}_t = \sigma(\boldsymbol{W}_u[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_u), \tag{A.12}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_r), \tag{A.13}$$

$$\tilde{\boldsymbol{h}}_t = \phi(\boldsymbol{W}_h[\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_h), \tag{A.14}$$

$$\boldsymbol{h}_t = (\mathbf{1} - \boldsymbol{u}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{u}_t \odot \tilde{\boldsymbol{h}}_t, \tag{A.15}$$

where $\mathbf{1}$ is the vector of 1's. Compared with LSTM-RNN, the memory cell $\boldsymbol{c}_t$ and output gate $\boldsymbol{o}_t$ are removed. The input gate $\boldsymbol{i}_t$ and forget gate $\boldsymbol{f}_t$ is then renamed as update gate $\boldsymbol{u}_t$ and the reset gate $\boldsymbol{r}_t$, respectively. In Minimal Gated Unit (MGU) [16], the reset gate in Gated Recurrent Unit (GRU) [17] is integrated with the update gate. The percentage of parameter size reduction is roughly 75% for GRU and 50% for MGU from LSTM-RNN. It is expected that this parameter reduction leads to a regularization effect.

Simple recurrent unit

Recently, simple recurrent unit (SRU) [10] has been proposed. The architecture is:

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{W}\boldsymbol{x}_t, \tag{A.16}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f\boldsymbol{x}_t + \boldsymbol{b}_f), \tag{A.17}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{b}_r), \tag{A.18}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + (\mathbf{1} - \boldsymbol{f}_t) \odot \tilde{\boldsymbol{x}}_t, \tag{A.19}$$

$$\boldsymbol{h}_t = \boldsymbol{r}_t \odot \phi(\boldsymbol{c}_t) + (\mathbf{1} - \boldsymbol{r}_t) \odot \tilde{\boldsymbol{x}}_t. \tag{A.20}$$

Compared with LSTM-RNN and the above variants, the computation of the forget gate $\boldsymbol{f}_t$ and the reset get $\boldsymbol{r}_t$ does not require the hidden vector $\boldsymbol{h}_{t-1}$. The highway connection is also introduced at the output.

Akira Tamamori received his B.E., M.E., and D.E. degrees from Nagoya Institute of Technology, Nagoya, Japan, in 2008, 2010, 2014, respectively. From 2014 to 2016, he was a Research Assistant Professor at Institute of Statistical Mathematics, Tokyo, Japan. From 2016 to 2018, he was a Designated Assistant Professor at the Institute of Innovation for Future Society, Nagoya University, Japan. He is currently a lecturer at Aichi Institute of Technology, Japan. He is also a Visiting Assistant Professor at Speech and Language Processing Laboratory, Japan. His research interests include speech signal processing, machine learning, and image recognition. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), the Acoustical Society of Japan (ASJ), and International Speech Communication Association (ISCA).

Tomoki Hayashi received his B.E. degree in engineering and M.E. degree in information science from Nagoya University, Japan, in 2013 and 2015, respectively. He is currently a Ph.D. student at the Nagoya University. His research interest includes statistical speech and audio signal processing. He received the Acoustical Society of Japan 2014 Student Presentation Award. He is a student member of the Acoustical Society of Japan, and a student member of the IEEE.

Tomoki Toda received his B.E. degree from Nagoya University, Japan, in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2003 to 2005. He was then an Assistant Professor (2005–2011) and an Associate Professor (2011–2015) at NAIST. Since 2015, he has been a Professor in the Information Technology Center at Nagoya University. His research interests include statistical approaches to speech and audio processing. He has received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Kazuya Takeda received his B.E. and M.E. degrees in Electrical Engineering and his Doctor of Engineering degree from Nagoya University, Nagoya Japan, in 1983, 1985, and 1994, respectively. From 1986 to 1989 he was with the Advanced Telecommunication Research laboratories (ATR), Osaka, Japan. His main research interest at ATR was corpus-based speech synthesis. He was a Visiting Scientist at MIT from November 1987 to April 1988. From 1989 to 1995, he was a researcher and research supervisor at KDD Research and Development Laboratories, Kamifukuoka, Japan. From 1995 to 2003, he was an associate professor of the faculty of engineering at Nagoya University. Since 2003 he has been a professor in the Department of Media Science, Graduate School of Information Science, Nagoya University. His current research interests are media signal processing and its applications, which include spatial audio, robust speech recognition, and driving behavior modeling.