2018 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP2018)
Honolulu, Hawaii, USA, March 4-7, 2018

Novel Speech Recognition Interface Based on Notification of Utterance Volume Required in Changing Noisy Environment

Takahiro Goto, Takeshi Yamada, Shoji Makino

University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
E-mail: goto@mmlab.cs.tsukuba.ac.jp

Abstract

Speech recognition performance is generally degraded in noisy environments. The degree of performance degradation depends on the nature of the ambient noise. However, current speech recognizers do not feed back information on whether or not the noise environment at that moment is appropriate for speech recognition. This drawback greatly reduces the usability of speech recognition. In this paper, we propose a novel speech recognition interface with notification of the utterance volume required in a changing noise environment. The proposed method first predicts the recognition rate at that moment from the noise signal recorded just previously. It then decides an appropriate utterance volume of normal, loud, or unusable (speech recognition cannot be used) based on the predicted recognition rate and notifies the user of it. To evaluate the effectiveness of the proposed method, we performed an experiment in simulated noise environments. Furthermore, we implemented a speech recognition application with the proposed method on a tablet device and performed an experiment on subjects in real noise environments. The experimental results confirmed that potential recognition errors can be reduced by giving appropriate feed back about the utterance volume.

1. INTRODUCTION

Speech recognition services are becoming more prevalent with the improvement of speech recognition technology and the spread of portable devices such as smartphones. However, serious problems still remain in the further development of speech recognition services. The degradation of recognition performance in noisy environments is an issue to be addressed [1]. Since people using speech recognition for the first time are unaware of the performance degradation caused by ambient noise, they may repeatedly fail to obtain a correct recognition result. To make matters worse, current speech recognizers do not feed back information on whether or not the noise environment at that moment is appropriate for speech recognition. This drawback greatly reduces the usability of speech recognition. This may discourage people from using speech recognition systems. Even for people familiar with the use of speech recognition, it is difficult to instinctively understand the relationship between noise characteristics and recognition performance. Therefore, they tend to speak unnecessarily loud or avoid the use of speech recognition even in a sufficiently quiet environment.

To solve this problem, it is desirable to raise the recognition performance by improving noise-robust methods including noise suppression and acoustic model adaptation [1]. However, the effectiveness of these methods is strongly affected by the target noise environment. Further research is therefore required to enable the wide use of these methods. Another approach is to improve the speech recognition interface (speech input interface), considering the above problem as a user interface problem. This problem can be solved by notifying the user of whether or not the noise environment at that moment is appropriate for speech recognition. However, current speech recognizers do not provide such a function. Although there are speech recognizers that prompt a user to speak again on the basis of a confidence measure obtained by performing speech recognition [2], it is difficult for the user to understand the reason for this.

In this paper, we propose a novel speech recognition interface with notification of the utterance volume required in a changing noise environment. The proposed method first predicts the recognition rate at that moment from the noise signal recorded just previously. It then decides an appropriate utterance volume of normal, loud, or unusable (speech recognition cannot be used), based on the predicted recognition rate and notifies the user of it. When the predicted recognition rate is high, speech recognition can be performed successfully and the user can use speech recognition with a normal utterance volume. When the predicted recognition rate is low, it is too noisy to perform speech recognition. The user can then move to a quieter place or select another input method such as a touch panel keyboard. Furthermore, when the predicted recognition rate has an intermediate value, sufficient recognition performance can be expected by speaking louder than usual. This is based on the fact that a correct recognition result can be obtained by increasing the input SNR (speech to noise ratio) even if the environment is somewhat noisy. In this way, the proposed method reduces potential recognition errors and improves the usability of speech recognition.

To evaluate the effectiveness of the proposed method, we perform an experiment in simulated noise environments. Furthermore, we implement a speech recognition application with the proposed method on a tablet device and perform an experiment on subjects in real noise environments.

2. PROPOSED METHOD

2.1 Overview

Fig. 1 represents the process flow of the proposed method. First, it predicts the recognition rate at that moment from the

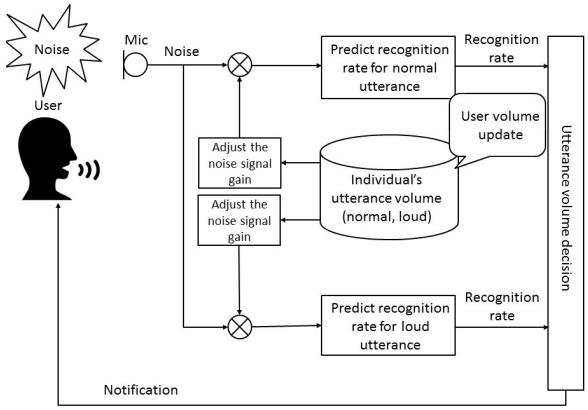


Figure 1: Process flow of the proposed method.

noise signal recorded just previously. The recognition rate is predicted for both a normal utterance volume and a loud utterance volume by adjusting the power of the noise signal. Since the utterance volume of the user depends on the individual, it is updated using previous utterances of the user. The proposed method then decides an appropriate utterance volume based on the predicted recognition rates and notifies the user of it. The utterance volume given to the user is normal, loud, or unusable (speech recognition cannot be used) in consideration of comprehensibility. The proposed method repeatedly performs these steps at regular short intervals.

The recognition rate prediction, the utterance volume decision, and the user volume update are described in detail below.

2.2 Recognition rate prediction

In the prediction of the recognition rate, an acoustic feature is first extracted from a fixed length of the noise signal recorded just previously. Spectral information plays an important role in predicting recognition performance [3]. In this paper, a 24-order mel filter bank output and a logarithmic power are calculated frame by frame from the noise signal. Since the length of the noise signal, the analysis frame length, and the analysis frame period are set to 2 s, 25 ms, and 10 ms, respectively, a feature set is obtained from 197 frames. Finally, a feature vector of 50 dimensions calculated from this feature set, which consists of the frame average and variance of a 24-order mel filter bank output and a logarithmic power, is used as the acoustic feature for recognition rate prediction [4].

The recognition rate is then predicted for each normal utterance volume and loud utterance volume by SVR (support vector regression) [5]. The training for SVR is performed on the basis of the relationship between the recognition rate and the acoustic feature. The teaching data are the word recognition rates obtained by performing speech recognition under various noise conditions. The acoustic feature is calculated using the noise signal corresponding to each noise condition. Note that since the predicted recognition rate depends on the average power of the speakers used for training, the power of the noise signal used for recognition rate prediction is adjusted according to the power of the individual user.

2.3 Utterance volume decision

The decision of the utterance volume is performed by a simple threshold-based method using the two predicted recognition rates and a fixed threshold value. The details are as follows.

- When both the predicted recognition rate for the normal utterance volume and that for the loud utterance volume are higher than the threshold value, speech recognition can be performed successfully. Thus, the utterance volume is decided as normal.
- When both the predicted recognition rate for the normal utterance volume and that for the loud utterance volume are lower than the threshold value, it is too noisy to perform speech recognition. Thus, the utterance volume is decided as unusable.
- When the predicted recognition rate for the normal utterance volume and that for the loud utterance volume are lower and higher than the threshold value, respectively, sufficient recognition performance can be expected by speaking louder than usual. Thus, the utterance volume is decided as loud.

Although the predicted recognition rate was adopted as a difficulty index for speech recognition, other indexes such as the SNR (signal to noise ratio) are also available. In this case, the decision method described above is applicable by replacing the recognition rate with the SNR.

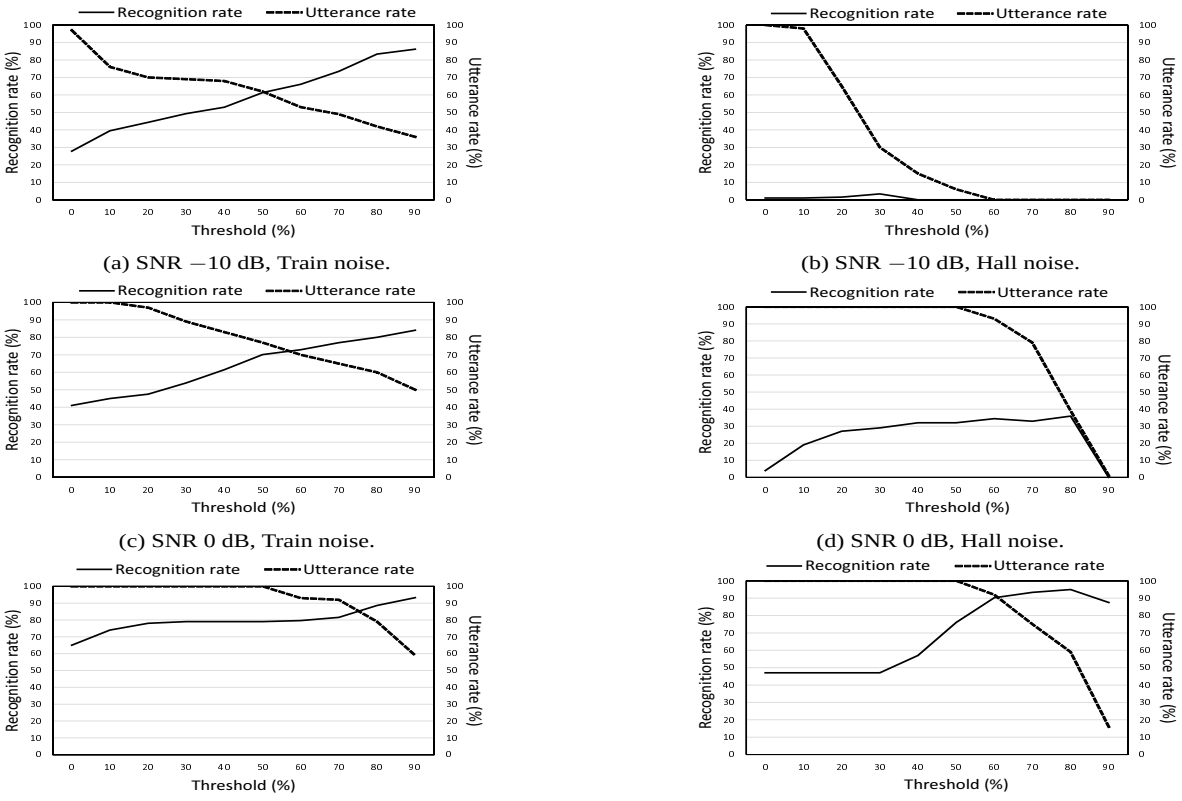
2.4 User volume update

In the user volume update, the power of the current user utterance is first calculated, which is assigned to normal or loud based on the notification to the user. The power for each of the the normal utterance volume and loud utterance volume is then updated as a weighted average power of the current utterance and the past several utterances. This makes it possible to compensate the power difference between the speakers used for training and an individual user as well as the power difference between normal and loud for the individual user, as described in Section 2.2.

3. EVALUATION

3.1 Experimental conditions

To evaluate the effectiveness of the proposed method, we performed a simulation experiment. One hundred utterance opportunities are set up at 5 s intervals for a noise signal with a long time duration. The proposed method predicts the recognition rate at each utterance opportunity and notifies the user of whether the appropriate utterance volume is normal, loud, or unusable. It is assumed that the user strictly follows the notification and decides to either use speech recognition with a normal utterance volume, use speech recognition with a loud utterance volume, or not use speech recognition. In this experiment, assuming that the power of each user is known, the power of the noise signal used to predict the recognition rate was adjusted in advance. The threshold value was changed from 0% to 90% in 10% steps (a common threshold value was set for all 100 utterance opportunities).



(a) SNR −10 dB, Train noise.

(b) SNR −10 dB, Hall noise.

(c) SNR 0 dB, Train noise.

(d) SNR 0 dB, Hall noise.

(e) SNR 10 dB, Train noise.

(f) SNR 10 dB, Hall noise.

Figure 2: Results of the simulation experiment.

As the speech data, we used 400 phonemically balanced words of two males and two females (100 words per speaker) from the Tohoku University-Matsushita spoken word database [6]. The simulation was carried out for each speaker, and one of the 100 words was assigned to one utterance opportunity so that there was no overlap for the 100 utterance opportunities. As the noise data, we used four types of noise from the Denshikyo noise database [7]: car noise, exhibition noise, train noise, and hall noise. The noisy speech data were generated by artificially adding the noise data to the speech data. The SNR was set to 10, 0, or −10 dB for the normal utterance volume and to 18, 8, or −2 dB for the loud utterance volume. The power difference between the normal utterance volume and the loud utterance volume was determined by measuring the utterances of several subjects. The SNR was defined as the ratio of the average power of the 100 words to the long-time average power of the noise signal. The SNR at each utterance opportunity therefore varied with the time characteristics of the noise signal.

The speech recognition task was isolated word recognition. The dictionary size was set to 106. The acoustic models were gender-independent monophone models with 16 Gaussians per state, which were trained with clean speech data from the ASJ-JNAS database [8]. The SVR used for recognition rate prediction was trained using pair data comprising the acoustic feature and the word recognition rate for each of the noisy speech data mentioned above. In this experiment, epsilon-SVR of LIBSVM [9] was used as the SVR tool, where the kernel function of the SVR was an RBF (radial basis function) and the cost parameter was set to 750.

3.2 Experimental results

The relationship between the recognition rate and the utterance rate when changing the threshold value for one speaker is shown in Fig. 2. The trend of the experimental results was the same among the four speakers. In Fig. 2, the results for the train noise and hall noise are shown. The utterance rate is defined as the ratio of the number of utterance opportunities when the user actually spoke to the total number of utterance opportunities. The recognition rate is also defined as the ratio of the number of utterances correctly recognized to the number of utterances actually spoken.

First, we describe the experimental results for the train noise. Those for the car noise showed the same tendency. When the threshold value is set to 0%, the speaker speaks with the normal utterance volume at all the utterance opportunities, for which the recognition rate is about 40% at the SNR of 0 dB. This corresponds to the result without the notification by the proposed method. As the threshold value is increased, it can be seen that the recognition rate steadily improves while the utterance rate decreases. This is because the number of potential recognition errors is reduced by providing the user with loud and unusable notifications when the prediction recognition rate is intermediate and low, respectively. It can be seen that similar results were obtained at the

Figure 3: Notification icons used in the experiment.

SNR of 10 dB and −10 dB.

Next, we describe the experimental results for the hall noise, which showed the same tendency as the exhibition noise. When the SNR is −10 dB, it can be seen that as the threshold is increased to over 10%, the utterance rate drops sharply. This is because the hall noise makes speech recognition more difficult than the train noise. The predicted recognition rate is therefore lower than that of the train noise and the proposed method notifies the user that most utterance opportunities are unusable. This means that the proposed method works correctly. Also, when the SNR is 0 dB, as the threshold value is increased, the recognition rate improves. However, since the baseline recognition rate is extremely low, the improvement is limited. Finally, when the SNR is 10 dB, a significant improvement is observed as with the train noise. From the above, it is confirmed that the proposed method can reduce potential recognition errors.

3.3 Implementation and evaluation

Assuming that speech recognition is used on a portable device in various noise environments, we implemented a speech recognition application with the proposed method on a tablet device (ASUS Nexus 7). The speech recognition application displays the recognition result on the device screen using Google Cloud Speech API, which realizes client-server based speech recognition. We used Android Studio as the development environment and Java as the programming language, and implemented the application following the specifications described in Section 2. In this implementation, the SVR used for recognition rate prediction was trained with the same data described in Section 3.1. The threshold in the decision of the utterance volume was adjusted to 70%.

For the notification to the user, notification icons are used as shown in Fig. 3. The time interval for updating the notification is an important parameter related to the understandability of the notification. When the time interval is short, it is difficult for the user to decide an action since the notification changes frequently. Conversely, when the time interval is long, an appropriate notification is not given since it cannot follow the changes in the noise environment. In this implementation, the time interval for updating the notification was set to 0.3 s based on the result of a preliminary experiment.

We verified the effectiveness of the proposed method using the above tablet device. The experiment was conducted in a soundproof room. We placed a loudspeaker facing the corner of the wall and reproduced the train noise and hall noise mentioned in Section 3.1. The power of the noise was adjusted so that the SNR was about −10 or 0 dB at the microphone of the tablet device. Each of five subjects spoke ten words chosen randomly from the 100 words in Section 3.1 at a distance of about 0.8 m from the tablet device, which was placed on a desk. The subjects repeatedly spoke until all ten words were correctly recognized, following the notifications given by the proposed method. The recognition rate and the total number of utterances were used for evaluation.

Table 1: Recognition rate and the total number of utterances without and with the proposed method.

	Without proposed method	With proposed method
Train noise（about 0 dB）	78% / 64 times	82% / 61 times
Train noise（about −10 dB）	63% / 80 times	71% / 70 times
Hall noise（about 0 dB）	63% / 80 times	71% / 70 times
Hall noise（about −10 dB）	37% / 135 times	64% / 78 times

The recognition rate and the total number of utterances for the five subjects without and with the proposed method are shown in Table 1. From the table, it is confirmed that the recognition rate improved as a whole by using the proposed method, as in the simulation experiment.

4. CONCLUSIONS

In this paper, we proposed a novel speech recognition interface with notification of the utterance volume required in a changing noise environment. To evaluate the effectiveness of the proposed method, we performed an experiment in simulated noise environments. Furthermore, we implemented a speech recognition application with the proposed method on a tablet device and performed an experiment on subjects in real noise environments. The experimental results confirmed that the potential recognition errors are reduced by giving appropriate feed back about the utterance volume.

Acknowledgment

This research was supported by KAKENHI (17K00224).

References

[1] J. Li, L. Deng, Y. Gong, R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” IEEE/ACM Trans. Audio Speech Language Processing, Vol. 22, No. 4, pp. 745–777, 2014.

[2] J. Hui, “Confidence measures for speech recognition: A survey,” Speech Communication, Vol. 45, No. 4, pp. 455–470, 2005.

[3] T. Yamada, T. Nakajima, N. Kitawaki, S. Makino, “Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice,” IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No. 6, pp. 2006–2013, Nov. 2006.

[4] E. Morishita, T. Yamada, S. Makino N. Kitawaki, “Performance estimation of noisy speech recognition based on short-term noise characteristics,” Proc. Tunisian-Japan Symposium on Science, Society & Technology, TJASSST 2011, Nov. 2011.

[5] A.J. Smola, B. Scholkopf, “A tutorial on support vector regression,” Statistics and Computing, Vol. 14, No. 3, pp. 199–222, 2004.

[6] http://research.nii.ac.jp/src/TMW.html.

[7] http://research.nii.ac.jp/src/JEIDA-NOISE.html.

[8] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi, “JNAS Japanese speech corpus for large vocabulary continuous speech recognition research,” J. ASJ (E), Vol. 20, No. 3, pp. 199–206, May 1999.

[9] C. Chih-Chung, L. Chih-Jen, “LIBSVM: A library for support vector machines,” ACM Trans. Intelligent Systems and Technology, Vol. 2, No. 3, pp. 1–27, 2011.