Noise Robust Voice UI Technology and Its Applications

TSUJIKAWA Masanori, OKABE Koji, HANAZAWA Ken

Abstract

This paper introduces a technology of noise-robust speech recognition to correctly recognize a user's speech even in places with a high level of background noise and immediately respond to it, together with applications of this technology. Voice UI is useful when the user's hands or eyes are busy, but the background noise may cause false operations. The accurate detection of the user's speech using two microphones and the noise reduction with speech patterns suitable for speech recognition; these techniques enable applications of speech recognition in domains where this has previously been regarded as being difficult. In addition, these techniques also enable the usage of voice UI more natural.

Keywords

voice UI, speech recognition, speech detection, noise reduction
response speed, automatic speech translation, natural conversation

1. Introduction

Applications using voice UI for smartphones and tablet devices are becoming increasingly popular. These applications include, for example, one that displays the response to a question made by the user's voice on the screen and one that responds to the question with speech [1]. They allow the user to obtain desired information efficiently without keyboard input.

However, when such an application is used in a place with a high level of background noise, it may be incapable of reacting to the user's speech, or speech recognition errors may lead to false operation. Some applications require a button press before speaking in order to prevent false operation by reacting to noise. These factors narrow the range of voice UI functionality or degrade the efficiency in application use.

This paper describes a technology that NEC has developed for robust speech recognition in a noisy environments (noise-robust speech recognition technology) [2]. In addition, the pilot applications that we produced using this technology are introduced.

2. Noise–Robust Speech Recognition Technology

Fig. 1 shows a voice response system. The user's speech, input through the microphone, is recognized by means of matching with an acoustic model (acoustic information) and a language model (language information). The response according to the recognition results is displayed on the screen and/or output in audio form to loudspeakers.

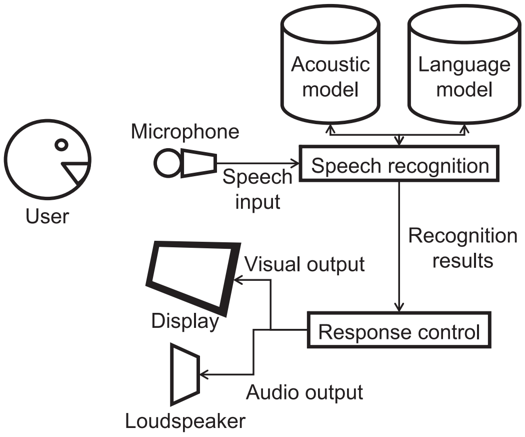


Fig. 1 A voice response system.

When the system is used in a place with a high level of background noise, the microphone picks up the user's speech together with the noise. To prevent false operation of the system due to noise, the system employs a technique to detect when the user is speaking (speech detection) and one to reduce the mixed-in noise (noise reduction).

Fig. 2 shows an example of the effect of these techniques.

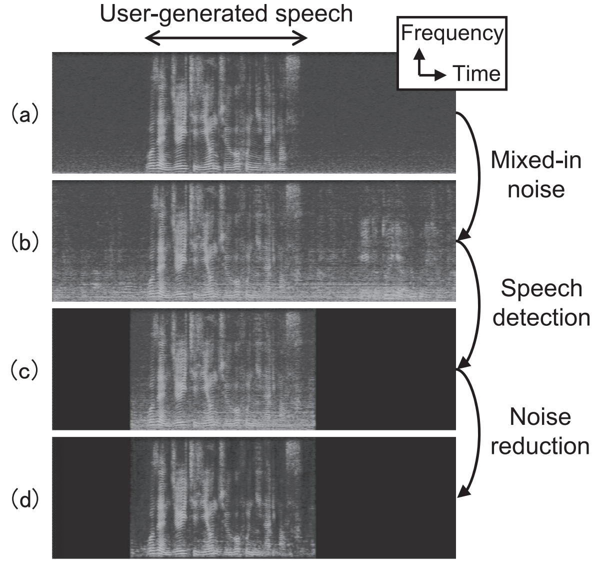


Fig. 2 Example of the effects of speech detection and noise reduction.

Fig. 2 (a) is a spectrogram of a user's speech sample. When noise is mixed in (a), the spectrogram becomes as shown in (b). Speech detection converts from the noisy input (b) to (c) and then noise reduction converts (c) into (d). The degree of reduction of the influence of noise is dependent on the speech detection and noise reduction techniques to be applied.

2.1 Two-microphone Speech Detection

Two-microphone speech detection detects user's speech by using two microphones to spatially distinguish from background noise. It is especially effective when this noise contains the speech of a person other than the user. One of the two kinds of speech detection methods described below is selected according to the usage setting.

(1) Speech detection using phase difference

Fig. 3 (a) shows an example of microphones positioned to distinguish a user's speech from noise using the phase difference of the sounds picked up by the two microphones. The sounds simultaneously input to microphones 1 and 2 can be regarded as the user's speech and those not simultaneously input can be regarded as noise. The advantage of this method is the possibility of positioning the two microphones within a small space.

The method we have developed is based on the ratio between the output of a filter that enhances the audio arriving from the user's direction (the frontal direction in the case of Fig. 3 (a)) and the output of a filter that eliminates the audio arriving from that direction. If the ratio between the two outputs is larger than a threshold value, it is judged that the user's speech is arriving. This method processes the cancellation of the audio arriving from the user's direction in two steps, i.e. in the complex spectral domain and in the amplitude spectral domain. This feature makes it possible to detect user's speech robustly even when the user deviates from the expected direction.

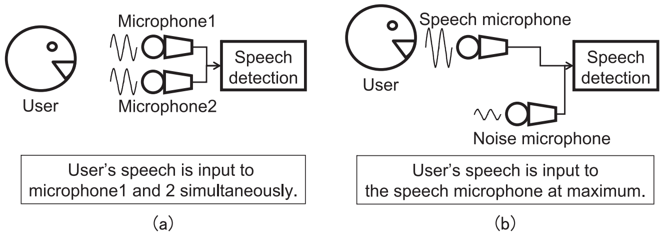

Microphone1
Microphone2
User
Speech detection
User's speech is input to microphone1 and 2 simultaneously.
(a)
Speech microphone
User
Speech detection
Noise microphone
User's speech is input to the speech microphone at maximum.
(b)

Fig. 3 Examples of microphone positions for two-microphone speech detection.

(2) Speech detection using amplitude difference

Fig. 3 (b) shows an example of microphones positioned to distinguish a user's speech from noise using the amplitude difference between the sounds arriving at the two microphones. The sound input at a higher level to the speech microphone can be regarded as the user's speech and that input at a higher or equivalent level to the noise microphone can be regarded as noise. The advantage of this method is a relatively high degree of freedom in the positions of the two microphones.

The method we have developed is based on the ratio between the inputs of the speech and noise microphones. If the ratio is larger than a threshold value, it is judged that the user's speech is arriving. This method calculates the ratio at each frequency subband and uses the highest ratio for speech detection. This makes it possible to detect user's speech robustly even when the noise level is equivalent to that of the user's speech.

2.2 Noise Reduction

Noise reduction in the amplitude spectral domain is used frequently because of its effectiveness as preprocessing for speech recognition. This technique obtains the speech spectrum by eliminating the noise spectrum estimated from the noisy speech spectrum.

The method we have developed [2)] is based on the ratio between the speech and noise spectra estimated from the noisy speech spectrum. When the ratio is small, that is, when the noise level is higher, the noise reduction filter coefficient becomes closer to zero. The noise is eliminated by multiplying the filter coefficient by the noisy speech spectrum. This method calculates the filter coefficient with the estimated speech spectrum compensated using pre-trained speech patterns. This enables noise reduction suitable for speech recognition.

3. Development of Pilot Applications Using Noise-Robust Speech Recognition Technology

3.1 Automatic Speech Translation System for Smooth Conversations

We prototyped an automatic speech translation system that makes conversations smooth by making use of our noise-robust speech recognition technology, in particular the speech detection method based on phase difference ((1) in section 2.1). Target language switching according to speech arrival direction and speech detection automation make it possible to omit a button press before speaking. This is expected to improve the smoothness of conversations.

(1) Outline of a prototype automatic speech translation system

Imagine that two speakers with different mother tongues are facing each other across a hotel lobby or a counter in a commercial facility to hold a conversation through a tablet device running an automatic speech translation application, as shown in **Fig. 4** . The two persons hold a conversation by viewing speech recognition results and their translation results.

The prototyped system assigns the direction of user A determined from the two microphones on the tablet as direction 1 and the direction of user B as direction 2 and only detects sounds from these directions. Sounds arriving from other directions are ignored to reduce false detections due to noise. In addition, if direction 1 differs greatly from direction 2, the speech detector for direction 1 can reject speech from direction 2. This property can be used to identify the language, e.g. the speech from direction 1 is in Japanese and that from direction 2 is in English.

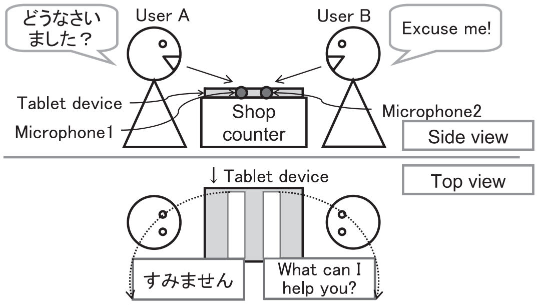


Fig. 4 Example of a conversation using a tablet device running an automatic speech translation application.

(2) Evaluation of speech recognition

We conducted a speech recognition evaluation to confirm the effects of speech detection using two microphones and language identification according to speech arrival direction.

We recorded the evaluation utterances using a mockup having the shape and size of a 7-inch tablet device and two microphones installed at a distance of 3 cm. Assuming that the frontal direction of the microphones is 0 degree, we installed the loudspeaker for playing Japanese utterances in the -45-degree direction and that for playing the English utterances in the 45-degree direction. The distances between each loudspeaker and each microphone were 20 cm and 40 cm respectively. For both Japanese and English utterances, we played a total of 20 travel conversation utterances by 4 persons reading 5 utterances each and recorded them. To simulate the noise that would be present in real use, we added noise data recorded with the same microphones on the speech data to create the evaluation utterances.

Fig. 5 shows the results of these speech recognition evaluations. They show that, compared to the conventional method of performing speech detection and language identification using a single microphone, our proposed method can achieve higher speech recognition accuracy. The main cause of the low speech recognition accuracy of the conventional method is the false detection of noise. Our method also achieves speech recognition accuracy equivalent to the case in which the time frame and language of the user's speech is known in advance. These results confirm the effectiveness of speech detection using two microphones and language identification according to speech arrival direction. Our system not only improves smoothness of conversations with the

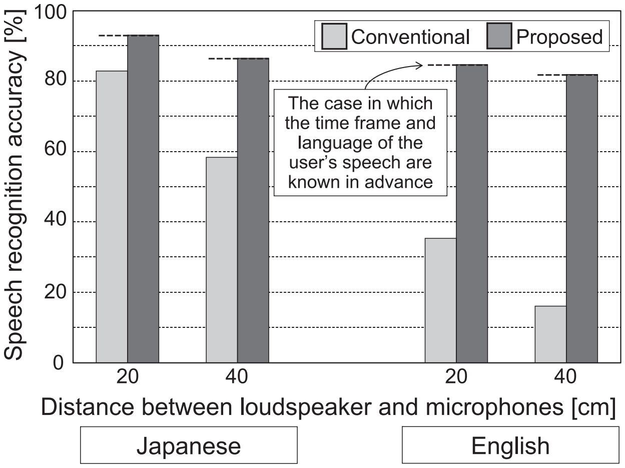


Fig. 5 Speech recognition evaluation results.

omission of button operations but also ensures speech recognition accuracy at a practical level of 80%.

3.2 Trial Service of Voice Response System in the Entertainment Domain

We implemented a trial service of voice response system that provides the experience of a conversation with a character to apply the noise-robust speech recognition technology to the entertainment domain. This trial service was conducted in an indoor theme park known as Sanrio Puroland.

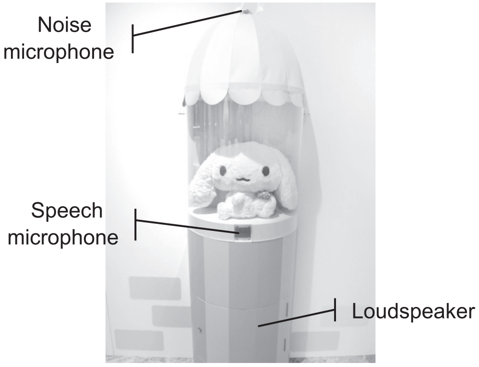
(1) Outline of the trial service

This trial service consisted of a simple conversation experience with "Cinnamon," the main character of the amusement park's "Cinnamoroll" group. When a guest spoke to the Cinnamon puppet, it answered according to the recognized content of the guest's speech so that the guest could experience conversation with the character.

The voice response system used dozens of greetings and simple question-and-answers prepared in advance.

Since the location of the trial was in an indoor park, the environmental noise did not include the sound of wind or rain. However, it was still a severe condition for speech recognition because the background music always existed and the voices and footsteps of other guests also made the recognition difficult especially when the park was crowded.

The installed voice response system was based on a speech detection method making use of amplitude difference ((2) in section 2.1). The signals from the speech microphone and the noise microphone were input in two channels to the PC and the response was output from the loudspeaker. All the processes were executed on a single PC as a stand-alone system, including the speech recognition and the voice response. **Fig. 6** shows an appearance of the voice response system. The puppet was set in an acrylic cylinder. The speech microphone was installed on the front and the noise microphone was installed on the top of the cylinder, respectively. The loudspeaker was installed in the bottom of the cylinder. The height of the speech microphone was set to the face level of children and was installed to keep a distance from the noise microphone to get better noise robustness. The PC is installed separately in the backyard.



(c) 2012 SANRIO CO.,LTD.

Fig. 6 Appearance of the voice response system.

(2) Results of the trial service

The voice response system was operated on all the business days of the theme park from April to September 2012. The number of operating days was about 140 and the number of responses made by the system was about 136,000. Because of the nature of a theme park, the numbers of responses on weekends and holidays were several times larger than those on weekdays. On weekdays and holidays, the number of recorded responses was as high as 2,000 per day.

Considering that the operating time on weekends and holidays was about 10 hours, the system responses more than three times per minute on the average, which suggests how popular this trial service was among the guests.

The system was started at opening time of the park and shut down at closing time, without any helps by staff. All the staff had to do was only execute the startup and shutdown procedures, and no difficulties were reported.

We sampled the collected log and evaluated the response speed and accuracy. For the response speed, the responses were output immediately after the guests' speech, which means that the trial succeeded in real-time processing and provided stress-free conversation experiences to guests. For the response accuracy, we evaluated about 1,000 samples and found that 70% of them were accurate. Though this accuracy is not very high, the real-time processing made it possible to obtain correct responses by demanding to the guests immediately to speak again when the response was incorrect or nothing. As a result, we consider that the system was able to provide guests with genuine conversation experiences in most of the cases.

On the other hand, the problem of false operation due to the false detection of background noise, which had been worried at the beginning, rarely occurred. This proves the noise robustness of speech detection using two microphones.

Finally, the feedbacks given by guests and staff are represented. Guests evaluated the system very favorably, saying, "It's so cute that I can't stop talking with it," "I was astonished because I felt I was really talking with Cinnamon," etc. Staff also evaluated it favorably, for example, saying, "I was very happy because the high recognition accuracy made the guests glad."

4. Conclusion

This paper has introduced the noise-robust speech recognition technology NEC has developed. It has also introduced applications of this technology, including an automatic speech translation system enabling smooth conversations and a trial service of voice response system in the entertainment domain. In the future, we expect that noise-robust speech recognition technology will make further progress and that its applications will expand further.

References

1) Apple "Siri''
http://www.apple.com/ios/siri/

2) Masanori Tsujikawa, et al., "In-Car Speech Recognition Using Model-Based Wiener Filter and Multi-Condition Training," Proc. of Interspeech 2008, Sep. 2008.

Authors' Profiles

TSUJIKAWA Masanori
Assistant Manager
Information and Media Processing Laboratories
Central Research Laboratories

OKABE Koji
Information and Media Processing Laboratories
Central Research Laboratories

HANAZAWA Ken
Principal Researcher
Information and Media Processing Laboratories
Central Research Laboratories

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.7 No.3 Smart Device Solutions

◇ Papers for Special Issue

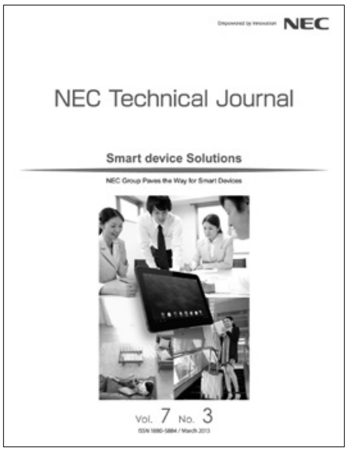
Service platforms

Smart devices

Solutions

Technical researches

◇ General Papers



Vol.7 No.3
March, 2013

Special Issue TOP