ADEQUACY ANALYSIS OF SIMULATION-BASED ASSESSMENT OF SPEECH RECOGNITION SYSTEM

Tetsuji Ogawa, Satoshi Kanba and Tetsunori Kobayashi

Department of Computer Science, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

ABSTRACT

The adequacies of the simulation-based assessment of speech recognition systems under noisy conditions are investigated and discussed. To evaluate the speech recognition systems in various environments, it is desirable to collect the test data uttered in the corresponding environments but it is not realistic since enormous works are required. To conduct evaluations of the speech recognition systems properly, it is promising to simulate evaluation experiments in the target environments as described below: comparatively small test data are collected, and test data of the target environment are generated by computing convolution of the impulse response of the target environment with the collected data. However, it is well known that changes of the acoustic characteristics are caused by Lombard effect, and so it is not necessarily obvious whether the simulation can precisely approximate the experiment in actual environment. This paper clarifies the condition to perform effective simulations of the noisy speech recognition, focusing on the influence of impulse response accuracies and Lombard effects on the speech recognition performance.

Index Terms— Noisy speech recognition, assessment, simulation, impulse response, Lombard effect

1. INTRODUCTION

In order to evaluate speech recognition performances in noisy environments, recognition experiments are desirable to be conducted using a real speech according to the situations that are decided by the product of the various factors such as room acoustics, natures of noises and characteristics of speakers. However, it is not realistic because enormous amount of test data are required. Thus, assuming that the factors that affect speech recognition are perceived individually and the affections operate independently, test data used for evaluations of the recognition performances can be obtained by combining the data under each influence. For example, a room characteristic can be simulated by computing the convolution of the dry source of a speech utterance with an impulse response of the room. Here, the dry source is, in principle, required to be recorded in an anechoic room. However, the speech data recorded through a close-talking microphone in a silent environment, that are called "clean speech" in the following parts, are generally used instead of the dry source. In addition, a noisy speech can be simulated by superposing the noise that is recorded independently according to each environment on the clean speech. However, it is not easy to measure the impulse response precisely because characteristics of the recording devices might degrade the characteristic of a real environment and a transfer characteristic from the mouth of a speaker to a microphone is different from the transfer characteristic that is computed from the measured impulse response according to each utterance. On the other hand, a Lombard effect generally occurs in the case that people speak in the noisy environments. The Lombard effect has phenomena that a speaking volume get louder, a pitch get higher, and so on. It is well known that the characteristics of the Lombard effect specifically differ from those of the neutral speech [1][2][3][4]. Therefore, it is not guaranteed that the speech obtained by the above means can precisely simulate the real speech uttered in the noisy environments. In the present paper, we attempt to experimentally verify the adequacy of evaluating the speech recognition performance using the simulated speech obtained by the above means. The experiments we conducted focus on the influence of the impulse response accuracies and the Lombard effects on the speech recognition performance.

This paper is organized as follows. At first, a motivation to analyze the adequacies of the simulation-based assessments of speech recognition and important issues to be worked out are described in section 2. Aiming at investigating the influence of the impulse response accuracies and the Lombard effect, the recorded speech are described in section 3. Section 4 describes the verification on the possibilities of the simulations from the viewpoint of the speech recognition performance. Finally in section 5, concluding remarks are presented.

2. MOTIVATION TO ANALYZE ADEQUACIES OF SIMULATION-BASED ASSESSMENT

A sound field from a speaker to a distant microphone can be simulated by computing convolutions of clean speeches with an impulse response of a room and superposing a recorded ambient noise on the speech in which the room acoustics have already been simulated. However, there are still important issues to be worked out in simulation-based assessments of speech recognition systems.

The characteristic of a close-talking microphone through which dry sources are recorded and the characteristic of a loudspeaker that play back time stretched pulses (TSPs) might degrade the characteristic of an actual sound field. In addition, the position of a mouth when people speak actually is different from the position of the loudspeaker that play back the TSP according to each utterance. Thus, in order to obtain the precise impulse response (transfer characteristic as well), the above issues have to be resolved.

On the other hand, the Lombard effect generally occurs when people speak in noisy environments, and it is well known that the characteristics of Lombard speeches specifically differ from those of neutral speeches. However, in general, the simulation-based assessments are performed on the basis of the dry sources with the neutral phonation style and the Lombard effect is not considered.

From the above viewpoints, we focus on two issues: 1) the influence of the accuracies of the impulse responses on the speech recognition performance, which is described in 4.2, and 2) the influence

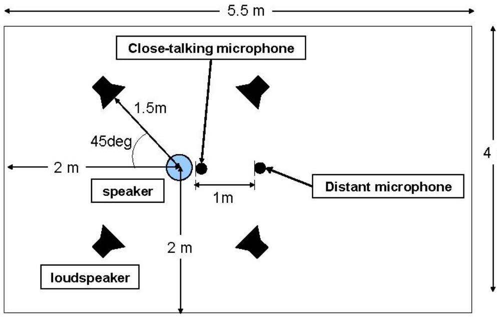

Fig. 1. The arrangement of a speaker and recording devices.

of the Lombard effect on the speech recognition performance, which is described in 4.3.

3. SPEECH MATERIAL

Aiming at investigating the influence of the impulse response accuracies and the Lombard effects on the speech recognition performances, the following three kinds of speeches were recorded in a general office with a reverberation time of 240 ms.

- Clean speech without the Lombard effect: CLEAN
- Clean speech with the Lombard effect: L-CLEAN
- Noisy speech with the Lombard effect: L-NOISY

Here, Fig.1 shows the arrangement of a microphone, loudspeakers and a speaker in the room where above three kinds of speeches were recorded. 100 newspaper article sentences were spoken by every 10 male speaker. The speech was simultaneously recorded through the close-talking microphone and the distant microphone.

CLEAN represents a neutral speech (as compared with a Lombard speech) uttered in the silent environment in which the ambient noise of about 30 dBA exists. L-NOISY was recorded under the condition that the ambient noise of a station concourse was played back from the four loudspeakers. The loudness of the ambient noise was adjusted to be 60 dBA in the speaker's ears. L-CLEAN was recorded as follows. A dummy head (NEIMANN KU-100) was substituted for the speaker and then binaural recordings were performed. The same kind and volume of ambient noise as L-NOISY was recorded through the microphones mounted on the dummy head. When people speak while mounting an open-air head phone (SENNHEISER HD650) and hearing the recorded ambient noise through the head phone, the speech obtained in such a way can be approximated as the pseudo-speech spoken in the noisy environment. Here, utterances of a speaker are required to be fed back to speaker himself through the head phone aiming at the speaker hearing the natural level speech at the ears. In addition, the noise that is floated out from the head phone and recorded through the microphone is required to be removed. However, in this experiment, the influences of the above issues can be negligible because the power ratio between the speeches recorded through the microphone of the dummy head with and without the head phone was mere 0.25 dB, and the power ratio between the speeches recorded with and without the noise floated out from the head phone was below 0.1 dB.

Table 1 shows the ten kinds of speeches generated using CLEAN, L-CLEAN and L-NOISY. C-D and LC-D represent the clean speeches with the neutral phonation style and the Lombard phonation style, respectively, which are recorded actually through the distant microphone under the silent condition. C-DS and LC-DS represent the simulated clean speeches corresponding to C-D and LC-D, respectively. They are obtained by computing the convolutions of the impulse response with the dry sources of CLEAN and L-CLEAN. LN-D represents a noisy speech recorded through the distant microphone under the noisy condition. Therefore, LN-D is affected by the Lombard effect. C-DSN can be obtained by computing the convolution of CLEAN with the impulse response of the room and then superposing an ambient noise. It is expected to simulate the noisy speech but has the neutral phonation style. LC-DSN can be obtained by computing the convolution of L-CLEAN with the impulse response of the room and then superposing the ambient noise. It is expected to simulate the noisy speech and is affected by the Lombard effect. C-DSN-C is generated in the same way as C-DSN but the SNR is adjusted so as to be the same as the SNR of LC-DSN. We also attempted to obtain the simulated noisy speech on the basis of SNR estimation of the real noisy speech. C-DSN-E and LC-DSN-E represent the simulated noisy speeches obtained so that the SNR of C-DSN and LC-DSN are respectively the same as the estimate of the SNR of LN-D.

4. SPEECH RECOGNITION EXPERIMENT

In this section, we focus on the speech recognition performance, and evaluate the adequacies of the simulation-based assessments used in noisy speech recognition. Here, if the recognition performance of the simulated speech becomes close to the performance of the real speech, the simulation can be regarded to be adequate from a viewpoint of the recognition performance.

4.1. Experimental setup

Experimental comparisons were conducted using continuous speech of newspaper article sentences (ASJ-JNAS) and phoneme balanced sentences (ASJ-PB), which were recorded using a close-talking microphone. Training and test data were represented by 25-dimensional parameters (12-dimensional MFCCs, 12-dimensional delta MFCCs and delta power), sampled every 10ms. Cepstrum mean normalization was applied to each utterance to remove the difference of input circumstances.

Acoustic models were trained with 20406 sentences from the ASJ-JNAS and ASJ-PB database. We adopted state-tied triphones in which the number of states was 2000 and the distribution function in each state was represented by a 16-mixture Gaussian distribution with diagonal covariances.

We used trigram language models that were constructed using a lexicon of 20K vocabulary.

For the evaluation, we used 1000 newspaper article sentences. 100 utterances were spoken by every 10 male speaker actually in the room shown in Fig.1.

4.2. Experiment 1: Influences of impulse response accuracies

4.2.1. What has to be evaluated for adequacy analysis ?

In this part, we evaluated the adequacies of the simulations in which the sound field without ambient noises is approximated by computing the convolutions of the dry sources with the impulse response. This aims at investigating the influence of the impulse response accuracies on the speech recognition performances. To that end, we focused on the issues described in 2: the influence of the compensations of the characteristics of the recording devices on the recognition performance and the influence of the gain adjustments of the

Table 1. Evaluation items. ("*" means computing the convolution.)

speech data	original speech	phonation	utterance	ambient noise	remarks of SNR
C-D	CLEAN	neutral	recorded directly	——	——
C-DS	CLEAN	neutral	dry source * impulse response	——	——
LC-D	L-CLEAN	Lombard	recorded directly	——	——
LC-DS	L-CLEAN	Lombard	dry source * impulse response	——	——
LN-D	L-NOISY	Lombard	recorded directly	recorded directly	——
C-DSN	CLEAN	neutral	dry source * impulse response	superposition	——
C-DSN-C	CLEAN	neutral	dry source * impulse response	superposition	same as LC-DSN
LC-DSN	L-CLEAN	Lombard	dry source * impulse response	superposition	——
C-DSN-E	CLEAN	neutral	dry source * impulse response	superposition	estimate of LN-D
LC-DSN-E	L-CLEAN	Lombard	dry source * impulse response	superposition	estimate of LN-D

Table 2. Conditions of simulations. "recording device" represents whether the characteristics of recording devices was compensated (on) or not (off). "impulse response" represents the method of adjusting the impulse response gain.

	recording device	impulse response
S-1	off	common to all utterances
S-2	on	common to all utterances
S-3	on	according to each utterance

impulse response according to each utterance on the recognition performance. Here, we attempted to compensate the characteristics of the recording devices by computing the convolutions of the simulated speeches with the inverse filter of the impulse response from the speaker to the close-talking microphone. In addition, it is difficult to measure the impulse response precisely according to each utterance. In order to reduce the errors, we attempted to adjust the gain of the impulse response according to each utterance so that the power of the simulated speech corresponded with the power of the real speech recorded through the distant microphone. We also attempted to adjust the impulse response gain among all utterances on the basis of the power ratio between the real Gaussian white noise and the simulated one. Table.2 shows conditions of the simulations we used. The compensation of the characteristics of the recording devices was not applied only under the condition of S-1. Under the conditions of S-1 and S-2, the impulse response gain was adjusted in the same way for all utterances using a reference signal of the Gaussian white noise. On the other hand, under the condition of S-3, the impulse response gain was adjusted according to each utterance.

The performances of C-D and LC-D that can be recorded directly through the distant microphone have to be compared to the performances of their simulated speeches, C-DS and LC-DS, respectively. Here, if the performances of C-DS and LC-DS are close to those of C-D and LC-D, respectively, the impulse response used in such a case has a good accuracy and the simulation of the sound field based on the impulse response can be considered to be adequate.

4.2.2. *Experimental results and discussions*

Figure 2 shows the word accuracies of C-D, C-DS, LC-D and LC-DS under the simulation conditions of S-1, S-2 and S-3. Figure 2 shows that the adequate simulation of the room acoustics can be achieved only under the condition of S-3. Therefore, in order to simulate the transfer characteristics from the speaker to the distant microphone precisely, both the compensations of the characteristics of the recording devices and the adjustment of the impulse response gain according to each utterance are indispensable.

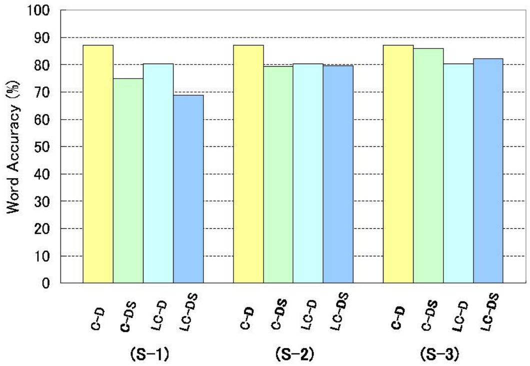

Fig. 2. Influences of the impulse response accuracies.

4.3. Experiment 2: Influences of Lombard effect

4.3.1. *What has to be evaluated for adequacy analysis ?*

In this part, we investigated the influence of the Lombard effect on the speech recognition performance. To that end, it is only necessary to compare the recognition performance of the real noisy speech, LN-D, with the performances of the simulated noisy speeches, C-DSN, C-DSN-C and LC-DSN. The simulated speeches that give equivalent performance to that of LN-D can be considered to be adequate from a viewpoint of the speech recognition performance. We also evaluated the recognition performance of the simulated speeches, C-DSN-E and LC-DSN-E, which were obtained on the basis of SNR estimation of LN-D.

4.3.2. *Experimental results and discussions*

Figure 3 shows the word accuracies for the real noisy speech and the simulated noisy speeches shown in Table 1 under the condition of S-3. Each thick bar represents the average recognition performance for 10 male speakers and the error bars represent the maximum and the minimum performances. Figure 4 shows the word accuracies for the real and the simulated noisy speeches under the conditions of S-1 and S-2. Under the S-1 and the S-2 conditions, the recognition performances of the simulated noisy speeches significantly differ from the recognition performance of the real noisy speech, and the adequate simulation can not be realized at all. Thus, in order to realize the adequate simulation of noisy speech recognition including the

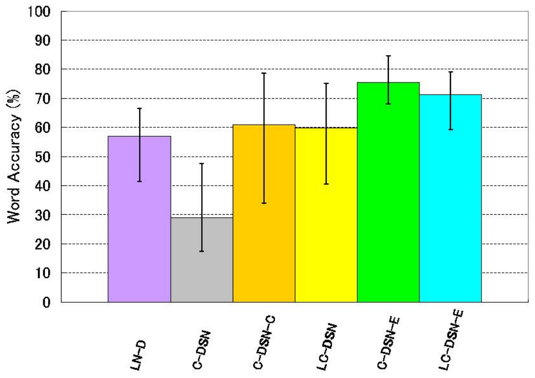

Fig. 3. Word accuracies for the real and the simulated noisy speech under the condition of S-3. Each thick bar represents the average performance for 10 male speakers. The error bars represent the maximum and the minimum performances.

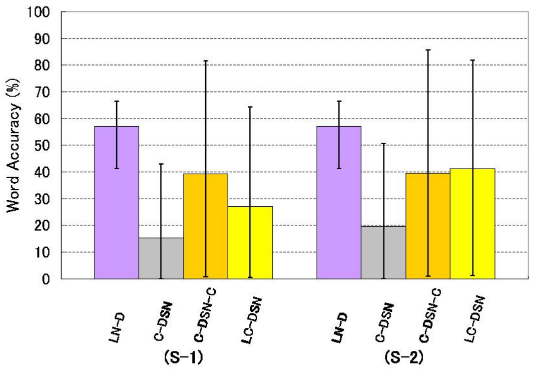

Fig. 4. Word accuracies for the real and the simulated noisy speech under the conditions of S-1 and S-2. Each thick bar represents the average performance for 10 male speakers. The error bars represent the maximum and the minimum performances.

Lombard effect, the precise impulse responses are required to be measured.

In the following part, the investigation of the influence of the Lombard effect on the speech recognition performance is conducted under the condition of S-3 (described in Fig.3). C-DSN has low volumes of speech signals compared to those of LN-D and LC-DSN because LN-D and LC-DSN have the Lombard phonation style while C-DSN has the neutral phonation style. Thus, C-DSN can not obtain the efficient SNR. Actually, the recognition performance of C-DSN was degraded significantly in comparison to that of LN-D. On the other hand, LC-DSN can obtain almost the same performance as LN-D. Thus, the real noisy speech can be adequately simulated from a viewpoint of the speech recognition performance if the simulation is performed on the basis of the clean speech recorded in such situation that people speak while hearing the same ambient noise through the head phone as in the real environment. Next, we attempted to adjust the SNR of C-DSN so as to be the same as the SNR of LC-DSN (named as C-DSN-C). C-DSN-C achieved the equivalent recognition performance to LC-DSN and LN-D. This result shows that if the SNR can be estimated precisely, the adequate simulation of the noisy speech can be performed in the speech recognition performance regardless of the phonation style such as the neutral and the Lombard speech. Also, the difference between C-DSN-C and LC-DSN represents the influence of the Lombard effect on the speech recognition performance. Thus, for the influence of giving the noise to the speakers on the speech recognition performances, the increase of the SNR arising from the increase of the utterance power is more important factor than the changes of speech spectra caused by the Lombard effect.

On the basis of the observations mentioned above, we tried to estimate the SNR of the real noisy speech, LN-D, and then to adjust the SNR of the simulated noisy speeches so as to be the same as the estimated SNR. We used the SPQA package[5] to estimate the SNR of the real noisy utterance. C-DSN-E that has the neutral phonation style and LC-DSN-E that has the Lombard phonation style represent the simulated speeches obtained on the basis of SNR estimation of LN-D. As we can see in Fig.3, C-DSN-E and LC-DSN-E gave the higher performance of about 10% in comparison to LN-D because SPQA package gave the higher SNR than the SNRs of C-DSN-C and LC-DSN. Since it is generally difficult to estimate the SNR of the real noisy speech precisely including this case, the simulation based on it can not achieve the adequate performance.

From the discussion above, if the SNR of the real noisy speech can be estimated precisely, the adequate simulation can be realized for both the neutral speech based simulation and the Lombard speech based simulation. However, in general, precise SNR estimation is difficult. Therefore, in order to conduct the adequate simulation from the viewpoint of the speech recognition performance, it is desirable to simulate the noisy speech on the basis of the clean speech recorded while giving the noise through the head phone to speakers.

5. CONCLUSIONS

The adequacies of the simulation-based assessments used in noisy speech recognition was investigated and discussed. Especially, we focused on the influence of the impulse response accuracy and the Lombard effect on speech recognition performances. As the results, the adequate simulation of noisy speech recognition requires to measure the impulse response precisely and to use the dry source uttered while hearing the noise through a head phone. However, the simulated speech can give the equivalent performance regardless of the phonation style such as the neutral and the Lombard speech under the same SNR. Therefore, for the influence of giving noises to speakers on the speech recognition performances, the increase of a SNR arising from the increase of an utterance power was more important factor than the changes of speech spectra caused by the Lombard effect.

6. REFERENCES

[1] J. C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Comunication*, vol. 20, pp. 13–22, 1996.

[2] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, pp. 151–170, 1996.

[3] A. Wakao *et.al.*, "Variabitily of Lombard effects under different noise conditions," in *Proc. ICSLP*, pp. 2009–2012, 1996.

[4] H. Boril *et.al.*, "Design and collection of Czech Lombard speech database," in *Proc. Interspeech*, pp. 1577–1580, 2005.

[5] "Speech Quality Assurance (SPQA) Package," http://www.nist.gov/speech/tools/.