IEICE TRANS. INF. & SYST., VOL.E92–D, NO.11 NOVEMBER 2009

PAPER

Influence of Lombard Effect: Accuracy Analysis of Simulation-Based Assessments of Noisy Speech Recognition Systems for Various Recognition Conditions

Tetsuji OGAWA[†a)] ***and*** **Tetsunori KOBAYASHI**[††], ***Members***

SUMMARY The accuracy of simulation-based assessments of speech recognition systems under noisy conditions is investigated with a focus on the influence of the Lombard effect on the speech recognition performances. This investigation was carried out under various recognition conditions of different sound pressure levels of ambient noise, for different recognition tasks, such as continuous speech recognition and spoken word recognition, and using different recognition systems, i.e., systems with and without adaptation of the acoustic models to ambient noise. Experimental results showed that accurate simulation was not always achieved when dry sources with neutral talking style were used, but it could be achieved if the dry sources that include the influence of the Lombard effect were used; the simulation in the latter case is accurate, irrespective of the recognition conditions.

key words: *Lombard effect, simulation, assessment, noisy speech recognition*

1. Introduction

It is important to evaluate the performance of speech recognition systems using real speech data recorded in different situations, which arise from various combinations of room acoustics, ambient noise, speakers, etc.; these factors affect the speech recognition performance of a system. However, collection of such data is not practically feasible, because a large number of combinations of the above factors must be considered. Therefore, during the evaluation of speech recognition systems, it is usually assumed that the effects of these factors are independent of each other and are treated individually. In this case, test data are obtained by simulating different speech materials under the influence of each factor. For example, a room characteristic is simulated by computing the convolution of a dry source, which is recorded by a close-talking microphone under quiet conditions, with the impulse response of the room. Similarly, a noisy speech utterance can be simulated by superposing ambient noise on a clean speech utterance. However, when a person speaks in a noisy environment, it is very likely for the Lombard effect to occur. The Lombard effect is the phenomena in which the volume of the speech increases, the pitch of the speaker increases, and so on. It is well known that the acoustic characteristics of speech utterances under the Lombard effect are significantly different from those of neutral speech utterances. Therefore, many researchers have attempted to improve the performances of speech recognition systems during recognition of the speech utterances under the Lombard effect [1]–[7]. However, it has not been verified whether the noisy speech data obtained by the above method precisely simulate the real noisy speech data in the evaluation of the performances of speech recognition systems.

In the present study, we attempt to analyze the evaluation accuracy of speech recognition systems using noisy speech data simulated by conventional methods. We emphasize the effects of the acoustic changes induced by the Lombard effect on speech recognition performances. We experimentally analyzed the accuracy of the simulation-based assessments of noisy speech recognition systems under the following different recognition conditions: different sound pressure levels (SPLs) of the ambient noise, different kinds of recognition tasks, such as continuous speech recognition (CSR) and spoken word recognition (SWR), and different types of recognition systems, i.e., systems with and without ambient noise adaptation of acoustic models. From the results of these analyses, we not only determined the accuracy of the simulation-based assessments but also the requirements for achieving accurate simulations in each recognition condition. The results of this study can be useful in predicting the performances of noisy speech recognition systems.

The rest of this paper is organized as follows. In Sect. 2, we describe the speech materials and simulation methods used. In Sect. 3, we describe database analysis from the viewpoint of elucidating acoustic changes due to the Lombard effect. In Sect. 4, we describe the speech recognition experiments carried out by us and analyze the accuracy of the use of simulations in evaluating speech recognition performances. Finally, in Sect. 5, we present the concluding remarks.

2. Simulation

2.1 Accuracy of Simulation

In simulation-based assessments of noisy speech recognition, noisy speech utterances recorded by a distant microphone are simulated by computing the convolutions of dry

Manuscript received February 13, 2009.
Manuscript revised June 27, 2009.

[†]The author is with the Waseda Institute for Advanced Study, Tokyo, 169–8050 Japan.
[†]The author is with the Department of Computer Science, Waseda University, Tokyo, 169–8555 Japan.
a) E-mail: ogawa@pcl.cs.waseda.ac.jp

DOI: 10.1587/transinf.E92.D.2244

sources (clean speech utterances recorded by a close-talking microphone) with the impulse response of the target environment and then superposing the ambient noise in the target environment on the abovementioned clean speech utterances in which the room acoustics are simulated using the impulse response. Here, the dry sources usually exhibit the neutral talking style and are free from the effects of the Lombard effect. We attempt to investigate the accuracy of these assessments with the emphasis on the influence of the Lombard effect on speech recognition performances. For this purpose, we compared the performance of the system during recognition of real noisy speech utterances, which were uttered in a real noisy environment, with that during the recognition of simulated noisy speech utterances. If the recognition performance using a simulated noisy speech utterance is found to be equal to that using the corresponding real noisy speech utterance, the simulation-based assessment is said to be accurate.

2.2 Speech Recordings

In the rest of this paper, a talking style of speech under the Lombard effect is defined as "Lombard talking style" and the speech utterances with Lombard talking style are defined as "Lombard speech utterances".

To investigate the effects of the acoustic changes due to the Lombard effect on speech recognition performances, the following three kinds of speech materials are recorded in a general office room, with a subject and recording devices, as shown in Fig. 1. Here, the recording devices used are listed in Table 1.

L-NOISY (Lombard noisy speech) denotes the real noisy speech utterances with the Lombard talking style, recorded by the distant microphone. The subjects spoke in a noisy environment, where noise from a station concourse [8] of 60 or 70 dB(A) (at the subjects' ears) was played on four loudspeakers.

N-CLEAN (neutral clean speech) denotes the clean speech utterances with neutral talking style recorded by the close-talking microphone, which are free from the influences of the Lombard effect. The subjects spoke under quiet conditions with the ambient noise of approximately 30 dB(A).

L-CLEAN (Lombard clean speech) denotes the clean speech utterances with the Lombard talking style recorded by the close-talking microphone. The subjects spoke while listening to the ambient noise that was played through headphones. Initially, ambient noise with the same SPL as that used in the L-NOISY recording was played and recorded by microphones mounted on a dummy head, which was substituted for the subject. The subjects spoke while they heard this ambient noise through open-air headphones.

Each of these speech materials consisted of 50 Japanese newspaper article sentences read as continuous speech utterances and 100 phonetically balanced word

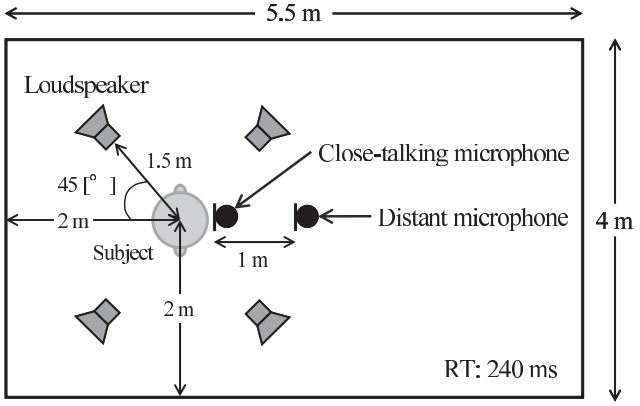


Fig. 1 Arrangement of a subject and recording devices.

Table 1 Recording devices.

Device	Manufacturer / Model number
close-talking microphone	SONY / F-710
distant microphone	SONY / ECM-77B
analog terminal	Thinknet / DF-3000
pre-amplifier	Thinknet / MA-2016
loudspeaker	BOSE / 1200VI
dummy head	NEIMANN / KU-100
headphone (open-air)	SENNHEISER / HD650

speech utterances. Ten male subjects uttered the entire set of these speech materials. These speech utterances were simultaneously recorded by the directional close-talking microphone, placed near the subject's mouth, and the omni-directional distant microphone placed at a distance of 100 cm from the subject.

During the recording of L-CLEAN, it was expected that some problems may arise due to the use of headphones to play the ambient noise. One problem is that the subjects may not be able to hear their own utterances at the original volume. As a compensation for this effect, it is required that the utterances are fed back to the subjects through the headphones so that the subjects can hear their voice at the original volume. However, the use of open-air headphones in place of closed ones eliminates the need for such feedback. Another problem is that the noise played through open-air headphones may leak and get recorded at the microphone. As a compensation for this effect, noise reduction of the leaked noise would have to be performed. It is assumed that the noise played through closed headphones does not leak. To analyze the first problem, we measured the speech powers recorded at the microphones of the dummy head with and without the open-air headphones. We found that the difference in the two speech powers was approximately 0.25 dB, which is negligible. To analyze the second problem, we measured the noise signals that leaked from the open-air headphones at the microphone. We found that this leaked noise was also negligible because its amplitude toward background noise amplitude was below 0.1 dB. On the basis of these results, we selected the open-air headphones for use in our experiments without applying any compensa-

Table 2 Evaluation items. "*" denotes computing a convolution; x denotes SPLs of ambient noise played (in dB(A)), which could be 60 or 70 dB(A).

	Notation	Recorded sound	Talking style	Speech (distant microphone)	Noise (distant microphone)	SNR
dry source	NC	N-CLEAN	neutral	—	—	—
	LCx	L-CLEAN	Lombard	—	—	—
clean speech	NC-D	N-CLEAN	neutral	direct input	—	—
	NC-DS	N-CLEAN	neutral	dry source * impulse response	—	—
	LCx-D	L-CLEAN	Lombard	direct input	—	—
	LCx-DS	L-CLEAN	Lombard	dry source * impulse response	—	—
noisy speech	LNx-D	L-NOISY	Lombard	direct input	direct input	——
	NCx-DSN	N-CLEAN	neutral	dry source * impulse response	superposition	——
	LCx-DSN	L-CLEAN	Lombard	dry source * impulse response	superposition	——
	NCx-DSN-C	N-CLEAN	neutral	dry source * impulse response	superposition	= LCx-DSN

tion; that is, we did not apply feedback of subjects' voices to themselves through the headphone or carry out the noise reduction of the leaked noise.

2.3 Evaluation Items

Table 2 lists two dry sources, four clean speech data, and four noisy speech data, which are obtained using the N-CLEAN, L-CLEAN, and L-NOISY recordings. Here, x denotes the SPL of the ambient noise heard by the subjects. The utterance notations are represented by three terms as (X)-(Y)[-(Z)], where (X) is either NC (N-CLEAN), LC (L-CLEAN), or LN (L-NOISY), which are the recorded sounds, and (Y) is either D, DS, or DSN, which indicates whether the evaluation was performed with or without simulation. D represents real speech data, which were directly recorded by the distant microphone. DS represents simulated clean speech data, which were synthesized by computing the convolution of the dry sources with the impulse response from the subject to the distant microphone. DSN represents simulated noisy speech data, which were synthesized by superposing the ambient noise on the simulated clean speech data, i.e., (X)-DS. The term (Z) could only be C, which indicates that while superposing the noise, the SNR was adjusted. Therefore, the notation of NCx-DSN-C represents the simulated noisy speech data, which were obtained by computing the convolution of the dry source with neutral talking style (N-CLEAN) with the impulse response. Then, the ambient noise of x dB(A) was superposed such that SNRs in this superposition would be the same as those in the superposition of LCx-DSN.

The first aspect of this study is the determination of the effect of the talking style on speech recognition performance. For this purpose, the recognition performances using the clean speech data with neutral talking style, i.e., NC, NC-D, and NC-DS, are compared with those using clean speech data with the Lombard talking style, i.e., LCx, LCx-D, and LCx-DS. The second aspect deals with the main purpose of this study: analysis of the accuracy of the assessments of noisy speech recognition systems. This analysis is carried out by comparing the recognition performances using the simulated noisy speech data, NCx-DSN, LCx-DSN, and NCx-DSN-C, to the recognition performance using the real noisy speech data LNx-D. In this case, if the recognition performance of the simulated noisy speech data is equivalent to that of LNx-D, the accurate simulation-based assessment of speech recognition performance is achieved by using such simulated noisy speech data.

2.4 Simulation Method and Requirements

2.4.1 Transfer Function Measurement

The transfer function from the subject to the distant microphone was obtained as the impulse response by the time stretch pulse (TSP) method. TSPs were played through a loudspeaker, which was placed at the position of the subject shown in Fig. 1. An up-type 131072-point TSP sampled at 32 kHz was used. Synchronous addition of eight microphone inputs was carried out in order to improve the SNR. The measured impulse response included the influence of the frequency characteristic of the loudspeaker.

2.4.2 Simulation of Room Acoustics

The effect of a sound field that does not contain ambient noise (i.e., room acoustics) was simulated. Speech utterances detected at the distant microphone were approximated by computing the convolution of the impulse response obtained previously (as described in Sect. 2.4.1) with the dry sources.

It should be noted that the position of the subject's mouth and that of the loudspeaker playing the TSPs did not completely coincide for each utterance. Therefore, the spectra of the real speech utterances recorded at the distant microphone did not precisely conform to those of the approximated speech utterances, in which the influence of the sound field was simulated. However, it is not practically feasible to measure the impulse response for each utterance. Thus, we simply adjusted the speech powers of the simulated and the corresponding real distant-talking speech utterances.

2.4.3 Compensation of Characteristics of Recording Devices

Figure 2 illustrates the method for the simulation of room

acoustics. In this figure, G_{CM}, G_{SP}, G_{SF}, and G_{DM} represent the frequency characteristics of the close-talking microphone, the loudspeaker playing the TSPs, the sound field, and the distant microphone, respectively. G_{SF} and G_{DM} influence both real and simulated speech utterances. On the other hand, G_{CM} and G_{SP} affect only the simulated speech utterances. Therefore, we attempted to compensate for the distortions in the impulse response caused by both G_{CM} and G_{SP}. We measured the impulse response from the subject to the close-talking microphone and designed its inverse filter. Then, we computed the convolution of the speech utterances recorded from the same close-talking microphone with this inverse filter.

2.4.4 Superposition of Ambient Noise

We superposed the ambient noise recorded by the distant microphone on the clean speech utterances; the influence of the surrounding environment (room acoustics) on these utterances was simulated, as described in Sect. 2.4.2. The ambient noise was adjusted to be 60 dB(A) or 70 dB(A) at the subject's ears. Under the quiet condition, since the ambient noise of 30 dB(A) was observed, this ambient noise was recorded and then superposed on the simulated clean speech utterances.

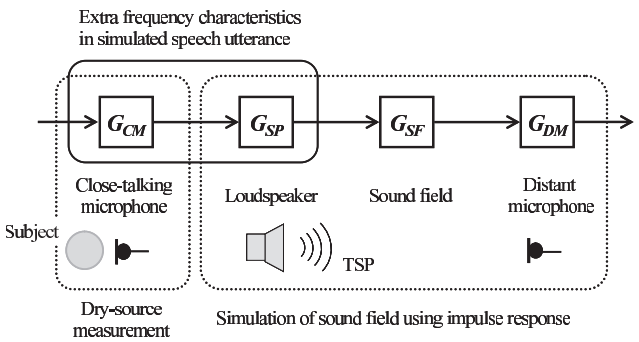
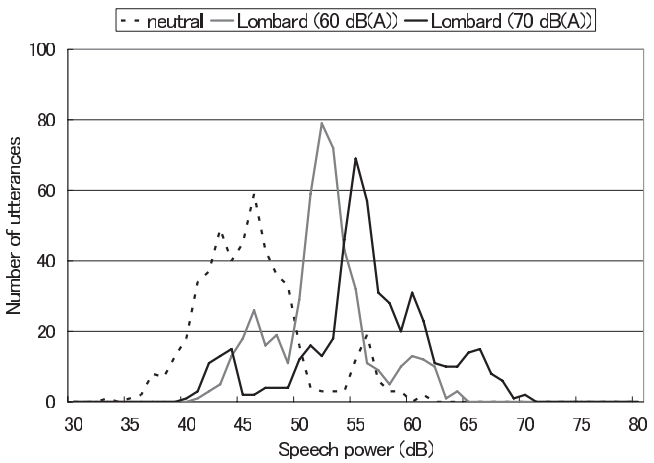


Fig. 2 Compensation of characteristics of recording devices. G_{CM}, G_{SP}, G_{SF}, and G_{DM} represent the frequency characteristics of close-talking microphone, loudspeaker, sound field, and distant microphone, respectively.

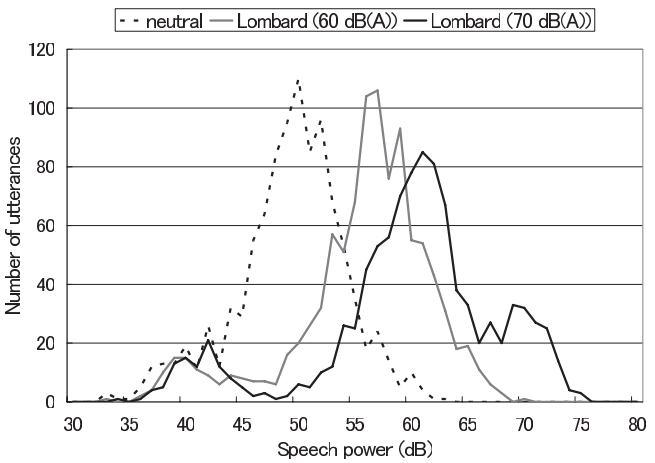
3. Database Analysis

We attempted to verify that speech utterances recorded as dry sources with Lombard talking style actually exhibit acoustic characteristics different from those of speech utterances with neutral talking style. This verification was carried out using the speech powers and vowel formants in the voiced parts of all the neutral (NC in Table 2) and the Lombard speech utterances (LC60 and LC70 in Table 2) recorded by the close-talking microphone. It is well known that these physical quantities affect speech recognition performances. Figures 3 (a) and 3 (b) show the number of utterances as a function of the average speech powers of the continuous speech utterances and the spoken word utterances, respectively; the speech powers were averaged over all frames in an utterance. Figures 4 (a) and 4 (b) show the positions of the first two formant frequencies of the five Japanese vowels, /a/, /i/, /u/, /e/, and /o/ in the continuous speech utterances and the spoken word utterances, respectively. In both these figures, each frequency is averaged over 1000 continuous speech utterances and 500 spoken word utterances by 10 male subjects. These figures show that the acoustic characteristics of the neutral speech utterances and the Lombard speech utterances are significantly different, regardless of the recognition tasks. It was observed that the power distributions and the vowel formant patterns shifted with an increase in the noise SPLs in the case of both the continuous speech utterances and the spoken word utterances.

The results of this analysis reveal that the real noisy speech utterances, which include the influence of the Lom-

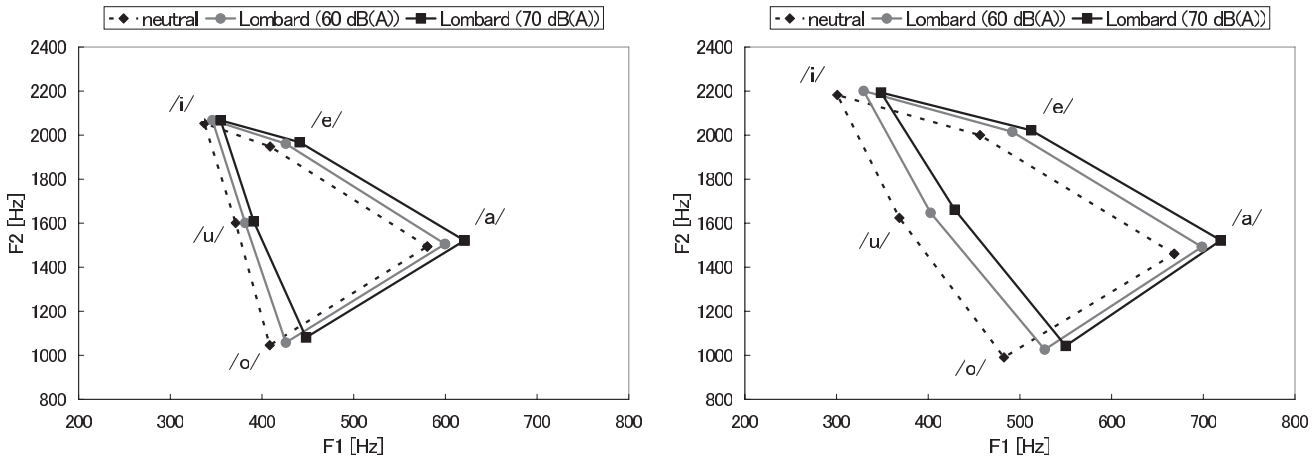


(a) Continuous speech utterances.



(b) Spoken word utterances.

Fig. 3 Number of utterances as a function of powers of speech utterances. The dotted line represents neutral speech utterance, and the solid lines represent Lombard speech utterances.



(a) Continuous speech utterances. (b) Spoken word utterances.

Fig. 4 Vowel formant positions of speech utterances. The dotted line represents neutral speech. The solid lines represent Lombard speech.

bard effect (LN60-D and LN70-D), and the simulated noisy speech utterances, which were simulated using dry sources with neutral talking style (NC60-DSN and NC70-DSN), have different acoustic characteristics in the presence of ambient noise of 60 and 70 dB(A).

4. Speech Recognition Experiment

4.1 Experimental Overview

The first stage of the experiments is the investigation of the effect of the talking style on the speech recognition performances. This investigation was conducted using the clean speech utterances. Here, the accuracy of the room acoustics approximated using the impulse response was also investigated.

The second stage of the experiments is the determination of the accuracy of simulation-based assessments of noisy speech recognition and the requirements for accurate simulation. The experiments were conducted on eight recognition systems listed in Table 3, and their results were compared. Here, the influence of the Lombard effect on speech recognition performances was investigated as a function of the following factors: SPLs of the ambient noise (e.g., 60 dB(A) or 70 dB(A)), recognition tasks (e.g., CSR or SWR), and the type of recognition systems (e.g., systems with or without ambient noise adaptation of acoustic models).

4.2 Experimental Setup

4.2.1 Acoustic Feature Extraction

Acoustic feature parameters used were 25-dimensional parameters consisting of 12-dimensional MFCCs, 12-dimensional ΔMFCCs, and a Δ power. Cepstral mean normalization (CMN) was applied to each utterance in order to eliminate the differences in input conditions. The experimental conditions for acoustic feature extraction are shown in Table 4.

Table 3 Evaluated systems.

	Task	Adaptation	Noise level
1)	CSR	—	60 dB(A)
2)	CSR	—	70 dB(A)
3)	CSR	MLLR	60 dB(A)
4)	CSR	MLLR	70 dB(A)
5)	SWR	—	60 dB(A)
6)	SWR	—	70 dB(A)
7)	SWR	MLLR	60 dB(A)
8)	SWR	MLLR	70 dB(A)

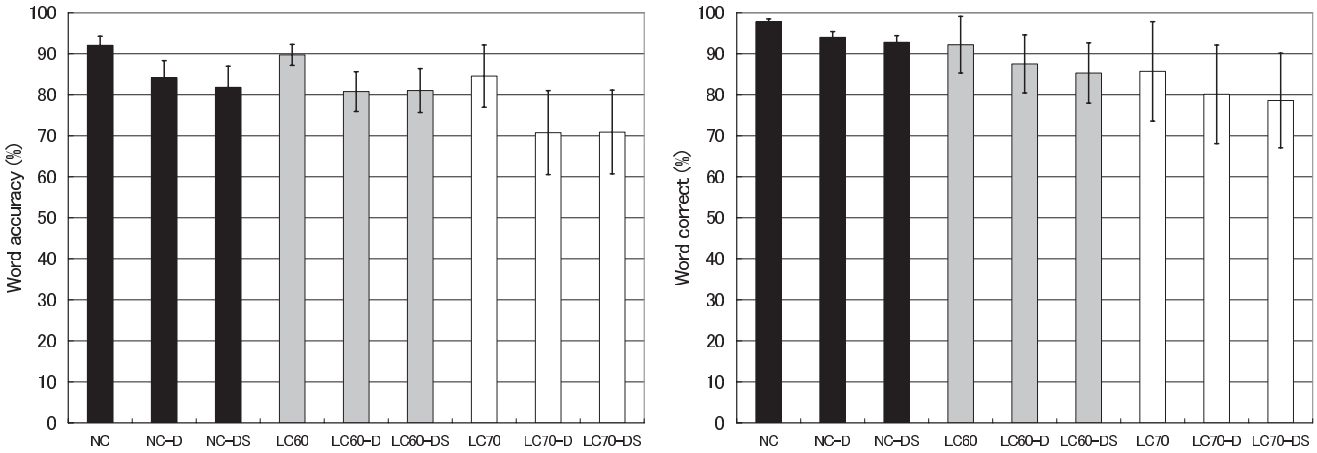
Table 4 Experimental conditions for acoustic feature extraction.

sampling frequency	16 kHz
frame length	25 ms
frame shift	10 ms
analysis window	Hamming window
pre-emphasis	$1–0.97z^{-1}$

4.2.2 Setup for CSR

Acoustic models were trained with 20406 sentences spoken by 133 male speakers, taken from the ASJ database [9]; this database included Japanese newspaper article sentences (ASJ-JNAS) and phoneme-balanced sentences (ASJ-PB) recorded by close-talking microphones. We used the state-tied triphone HMMs as the acoustic models with 2000 states. The distribution function in each state of the models was represented by a 16 mixture Gaussian distribution with diagonal covariances.

Ambient noise adaptation using maximum likelihood linear regression (MLLR) was applied to those models. The adaptation data consisted of 475 phoneme-balanced clean speech utterances by 95 male speakers taken from the ASJ-



(a) Word accuracies in CSR. (b) Word corrects in SWR.

Fig. 5 Speech recognition performances for real and simulated clean speech utterances. Each bar represents the average performance for 10 male subjects. The error bars on top of these bars denote the 90% confidence intervals.

PB. The ambient noise, which was the same as that used in the evaluation, was superposed on these utterances at SNRs of 5 dB, 10 dB, 15 dB, and 20 dB. The adapted models were constructed for each value of SNR. The number of regression classes in these adaptations was four. It should be noted that both the general acoustic models and the adapted acoustic models had the neutral talking style.

We used trigram language models that were constructed using a lexicon with a vocabulary size of 20 K. The vocabulary set comprises the most frequently appearing words in the articles in Mainichi Newspaper issues dating from January 1991 to September 1994.

In evaluation, we used 500 newspaper article sentences for each evaluation item listed in Table 2, where each of the 10 male subjects spoke 50 utterances in the room shown in Fig. 1.

4.2.3 Setup for SWR

The construction of acoustic models and their MLLR-based ambient noise adaptation were carried out under the same conditions as those used in the CSR experiment.

In evaluation, we used 1000 phoneme-balanced word utterances for each evaluation item listed in Table 2, where each subject spoke 100 words in the room shown in Fig. 1. In this case, the vocabulary set consisted of 216 words.

4.3 Experimental Results of Clean Speech Data

Figures 5 (a) and 5 (b) show the performance of the systems during the recognition of the clean speech utterances in the CSR and SWR experiments, respectively.

The acoustic models used in this experiment were trained with the speech data having neutral talking style as well as those used in general speech recognition assessments, while the test data had the Lombard talking style. Therefore, the acoustic discrepancies between the acoustic models and the test data increased, and the recognition performances decreased with increasing SPLs of the ambient noise.

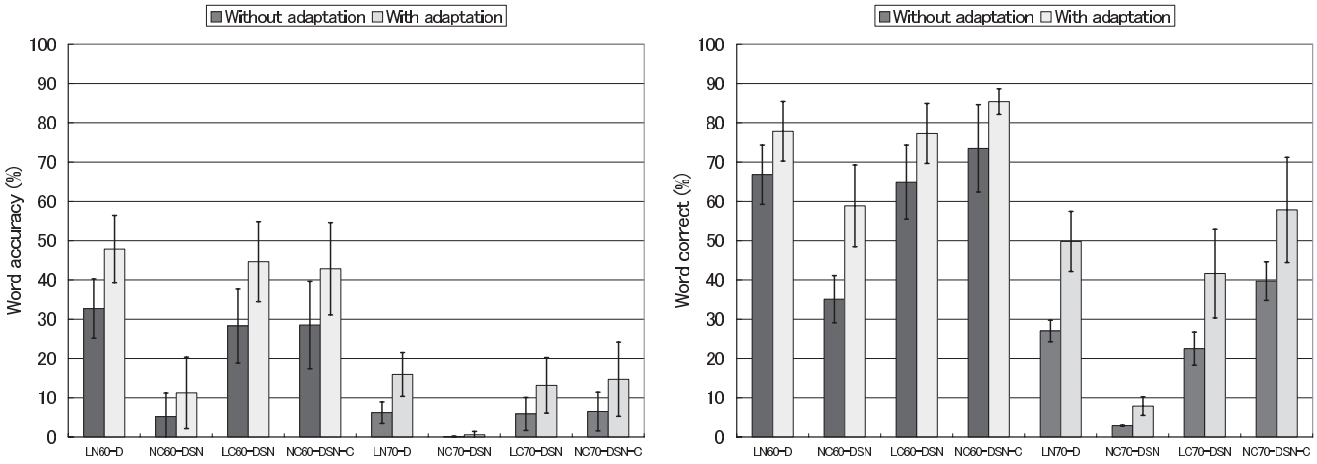
In addition, the performance of NC-D was comparable to that of NC-DS, and the performances of LC60-D and LC70-D were comparable to those of LC60-DS and LC70-DS, respectively. Thus, we can conclude that the impulse response used in this experiment was sufficiently accurate to approximate the room acoustics, the effect of which on speech recognition performances is independent of the Lombard effect.

4.4 Experimental Results of Noisy Speech Data

Figures 6 (a) and 6 (b) show recognition performances obtained by the CSR and SWR experiments, respectively, for each evaluation item listed in Table 2. The bars on the left-hand side and right-hand side in each item denote the performances of the recognition system trained with clean speech data and the recognition system adapted to noisy speech data, respectively. In the case of noise-adapted systems, the evaluation was carried out at each SNR of the adaptation data (e.g., 5 dB, 10 dB, 15 dB, and 20 dB), and the performances were averaged over all these SNR values.

4.4.1 Results of CSR Experiments

The evaluation items LN*x*-D and LC*x*-DSN have the Lombard talking style, while NC*x*-DSN has the neutral talking style. Therefore, NC*x*-DSN has a lower volume of speech signals and thus a lower SNR than both LN*x*-D and LC*x*-DSN. In fact, NC*x*-DSN showed a significantly low performance as compared to LN*x*-D, irrespective of the SPLs of the ambient noise and ambient noise adaptation. Moreover, the performance of LC*x*-DSN was almost the same as that



(a) Word accuracies in CSR. (b) Word corrects in SWR.

Fig. 6 Speech recognition performances for real and simulated noisy speech utterances. Each bar represents the average performance for 10 male subjects. The error bars denote 90% confidence intervals. The left-hand side and the right-hand side bars in each category represent the performances of HMMs trained using clean speech utterances and HMMs adapted using noisy speech utterances, respectively.

of LN*x*-D, irrespective of the noise levels and noise adaptation. Therefore, the simulation of noisy speech utterances is accurate from the viewpoint of speech recognition performances if the simulation is conducted using the dry sources that include the influence of the Lombard effect.

The recognition performance of NC*x*-DSN-C was almost the same as those of LC*x*-DSN and LN*x*-D, regardless of the noise levels and noise adaptation. Since both NC*x*-DSN-C and NC*x*-DSN have neutral talking style, the difference in their recognition performances is attributed to the difference in their SNRs. In addition, since NC*x*-DSN-C and LC*x*-DSN have the same SNR, the difference in their recognition performances is attributed to the difference in the acoustic characteristics due to the different talking styles (i.e., neutral and Lombard talking styles). Thus, on the analysis of the effect of presenting the ambient noise to the subjects on speech recognition performances during CSR, it is found that the increase in the SNR due to the increase in the utterance power has a greater impact on the performances than the changes in the speech spectra induced by the Lombard effect. In addition, if the SNRs of the real noisy speech utterances are accurately estimated, the performances of the noisy speech recognition systems can be accurately simulated using the dry sources with not only the Lombard talking style but also the neutral talking style. However, a convenient method for the precise estimation of SNRs of real noisy speech data [11] has not been proposed.

4.4.2 Results of SWR Experiments

The recognition performance of LC*x*-DSN was comparable to that of LN*x*-D, while the performance of NC*x*-DSN was significantly low as compared to that of LN*x*-D, irrespective of the noise levels and noise adaptation. Therefore, it can be conducted that as in the case of CSR, recognition performances of real noisy speech utterances can be accurately simulated in SWR, if the speech utterances were simulated using dry sources with the Lombard talking style.

However, unlike the results of the CSR experiments, the performance of NC*x*-DSN-C was not comparable to those of LN*x*-D and LC*x*-DSN. In this experiment, the performance of NC*x*-DSN-C was higher than that of LN*x*-D by approximately 10 points, regardless of the noise levels and noise adaptation.

Table 5 Requirements of accurate simulation.

Task	Talking style of dry source	SNR	Lang. model constraint
CSR	Lombard	——	——
	neutral	precise SNR estimation	strong
SWR	Lombard	——	——

4.5 Requirements of Accurate Simulation

We now discuss the requirements for carrying out accurate simulation of noisy speech recognition. Table 5 summarizes the requirements in the case of each recognition task.

In CSR task we evaluated, accurate simulation could be achieved by using not only the dry sources with Lombard talking style but also those with neutral talking style. The latter approach is possible under the assumption that the SNRs are precisely estimated. In this experiment, since the acoustic models are trained using the clean speech data with neutral talking style, the recognition systems are expected to potentially show discrepancies between the acoustic models used and the test data, which include the influences of the Lombard effect (e.g., LN*x*-D and LC*x*-DSN). The CSR systems use language models as well as the acoustic models. Therefore, the difference in acoustic likelihoods due to the

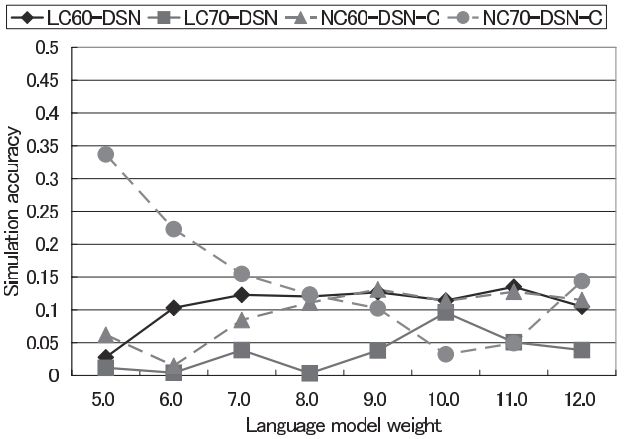


Fig. 7 Simulation accuracy in terms of speech recognition performance as a function of language model weights in the presence of ambient noise of 60 or 70 dB(A) (x denotes noise SPLs (60 or 70 dB(A)).

difference in the talking styles (e.g., neutral and Lombard talking style) might be compensated by the likelihoods from the language models. As a result, the inconsistencies between the acoustic models and the test data may not have a considerable effect on the recognition performance.

In the present study, we also investigated the influence of the constraint of a language model on the accuracy of the simulation in terms of the CSR performance. Figure 7 shows the simulation accuracies for the simulated noisy speech data, LCx-DSN and NCx-DSN-C, in the presence of the ambient noise of 60 or 70 dB(A) as a function of language model weights. Here, the simulation accuracy is defined as follows;

$$\mathrm{SA} = \frac{|\mathrm{WA}_{\mathrm{real}} - \mathrm{WA}_{\mathrm{sim}}|}{\mathrm{WA}_{\mathrm{real}}} \quad (1)$$

where $\mathrm{WA}_{\mathrm{real}}$ denotes the word accuracy of real noisy speech utterances, i.e., LNx-D; and $\mathrm{WA}_{\mathrm{sim}}$ denotes the word accuracy of simulated noisy speech utterances, i.e., LCx-DSN and NCx-DSN-C. If the simulation accuracy is equal to zero, the simulation is considered to be correct. This figure shows the following results. (1) The simulation could be accurately performed, irrespective of the language model weights and the noise SPLs, when using the dry sources with the Lombard talking style. (2) The simulation accuracy was affected by the experimental conditions when using the dry sources with the neutral talking style. Accurate simulation was achieved when using a system with a large language model weight. In contrast, when a small language model weight was used, the simulation accuracy significantly degraded in the presence of ambient noise of 70 dB(A); in this case, it is very likely for the acoustic changes due to the Lombard effect to occur.

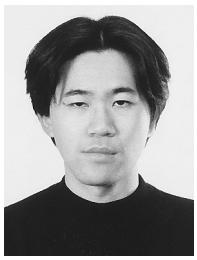
On the other hand, SWR systems use only the acoustic models for computing likelihoods. Therefore, the influence of the spectral differences between the neutral and the Lombard talking styles on the speech recognition performance was not negligible. In fact, the recognition performance of simulated noisy speech data with the Lombard talking style was significantly different from that of simulated noisy speech data with neutral talking style, even when their SNRs were the same. From the above results, it can be concluded that simulation-based assessments of noisy SWR systems could not achieve accurate simulation with the test data synthesized using dry sources with neutral talking style and thus require dry sources with the Lombard talking style.

5. Conclusion

We investigated the accuracy of simulation-based assessments of noisy speech recognition with the emphasis on the influence of the acoustic variations induced by the Lombard effect on speech recognition performance. The recognition performances using real noisy speech data could be accurately simulated by compensating the influence of recording devices on the simulation and using the dry sources with Lombard talking styles, irrespective of the noise levels, recognition tasks, and recognition systems. Although the SNRs in test data used are the same, the influence of the talking styles of the dry sources on recognition performances in different recognition tasks was different. CSR could achieve accurate simulation when using a system with a strong constraint of a language model, irrespective of the talking styles. In contrast, SWR required simulation using the dry sources with the Lombard talking style for accurate simulation.

References

[1] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," Speech Commun., vol.20, pp.13–22, Nov. 1996.
[2] J.H.L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," Speech Commun., vol.20, pp.151–170, Nov. 1996.
[3] A. Wakao, K. Takeda, and F. Itakura, "Variability of Lombard effects under different noise conditions," Proc. ICSLP, vol.4, pp.2009–2012, Oct. 1996.
[4] H. Boril and P. Pollak, "Design and collection of Czech Lombard speech database," Proc. Interspeech, pp.1577–1580, Sept. 2005.
[5] H. Goy, K. Pichora-Fuller, P. van Lieshout, G. Singh, and B. Schneider, "Effect of within- and between-talker variability on word identification in noise by younger and older adults," Proc. Interspeech, pp.418–421, Aug. 2007.
[6] A. Ikeno and J.H.L. Hansen, "Lombard speech impact on perceptual speaker recognition," Proc. Interspeech, pp.441–444, Aug. 2007.
[7] H. Boril, P. Fousek, and H. Hoge, "Two-stage system for robust neutral/Lombard speech recognition," Proc. Interspeech, pp.1074–1077, Aug. 2007.
[8] JEIDA noise database. http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html
[9] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, K. Shikano, T. Kobayashi, and S. Itahashi, "The design of the newspaper based Japanese large vocabulary continuous speech recognition corpus," Proc. ICSLP, pp.3261–3264, Nov. 1998.
[10] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, "Speech database user's manual," ATR Technical Report TR-I-0028, 1988.
[11] Speech Quality Assurance (SPQA) Package, http://www.nist.gov/speech/tools



Tetsuji Ogawa received the B.S., M.S. and Ph.D. degrees in electric, electronics, and computer engineering from Waseda University, Tokyo, Japan, in 2000, 2002, and 2005, respectively. He was a Research Associate (2004–2007) and a Visiting Lecturer (2007) at Waseda University. He has been an Assistant Professor at Waseda Institute for Advanced Study since 2007. His research interests include stochastic modeling for pattern recognition, speech enhancement, and speech recognition. He is a member of Acoustic Society in Japan.

Tetsunori Kobayashi received the B.S., M.S., and Ph.D. degrees in electrical engineering from Waseda University, Tokyo, Japan, in 1980, 1982, and 1985, respectively. He was a Lecturer (1985–1987) and an Associate Professor (1987–1991) at Hosei University, Tokyo, Japan. In 1991, he joined Waseda University as an Associate Professor. Since 1997, he has been a Professor. He was a visiting scientist at Spoken Language System Group, Laboratory for Computer Science, Massachusetts Institute of Technology (1994–1995). His research interests include perceptual computing and intelligent robotics. He is a member of Information Processing Society of Japan, Institute of Electrical and Electronics Engineers, and Acoustic Society in Japan.