DEVELOPMENT OF VAD EVALUATION FRAMEWORK CENSREC-1-C AND INVESTIGATION OF RELATIONSHIP BETWEEN VAD AND SPEECH RECOGNITION PERFORMANCE

Norihide Kitaoka[1], *Kazumasa Yamamoto*[2], *Tomohiro Kusamizu*[2], *Seiichi Nakagawa*[2], *Takeshi Yamada*[3], *Satoru Tsuge*[4], *Chiyomi Miyajima*[1], *Takanobu Nishiura*[5], *Masato Nakayama*[5], *Yuki Denda*[5], *Masakiyo Fujimoto*[6], *Tetsuya Takiguchi*[7], *Satoshi Tamura*[8], *Shingo Kuroiwa*[4], *Kazuya Takeda*[1], *Satoshi Nakamura*[9]

[1]Nagoya University, [2]Toyohashi University of Technology,
[3]University of Tsukuba, [4]University of Tokushima, [5]Ritsumeikan University,
[6]NTT, [7]Kobe University, [8]Gifu University, [9]NiCT/ATR

ABSTRACT

Voice activity detection (VAD) plays an important role in speech processing including speech recognition, speech enhancement, and speech coding in noisy environments. We developed an evaluation framework for VAD in such environments, called Corpus and Environment for Noisy Speech Recognition 1 Concatenated (CENSREC-1-C). This framework consists of noisy continuous digit utterances and evaluation tools for VAD results. By adoptiong two evaluation measures, one for frame-level detection performance and the other for utterance-level detection performance, we provide the evaluation results of a power-based VAD method as a baseline. When using VAD in speech recognizer, the detected speech segments are extended to avoid the loss of speech frames and the pause segments are then absorbed by a pause model. We investigate the balance of an explicit segmentation by VAD and an implicit segmentation by a pause model using an experimental simulation of segment extension and show that a small extension improves speech recognition.

Index Terms— Voice activity detection, Noisy speech recognition, evaluation framework

1. INTRODUCTION

Recently, speech recognition performance has been drastically improved by statistical methods and huge speech databases. Now performance improvement under a realistic environment, such as noisy conditions, has become the focus and some projects for noisy speech recognition evaluation including ours were organized [1, 2, 3, 4, 5, 6, 7].

So far we developed the evaluation frameworks for speech recognition performance itself. But in the noisy speech recognition, not only does the speech recognition method play an important role, but so to does the voice activity detection (VAD). Using a VAD with high performance as a front-end, speech recognizers can drastically reduce false alarms from non-speech periods and deletion errors from speech periods in input speech. In this paper, we introduce a new evaluation framework for VAD under noisy conditions, CENSREC-1-C (CENSREC-1-Concatenated). Each data in CENSREC-1-C includes some connected digit utterances with pauses between them and the task is to detect the speech periods in each data. CENSREC-1-C also provides evaluation tools and baseline results and thus the users can compare their VAD methods with the baseline.

Table 1. Noise environments

	Additive noises	Filter characteristic
Set A	Subway, Babble, Car, Exhibition	G.712
Set B	Restaurant, Street, Airport, Station	G.712

When using VAD in speech recognizer, the detected speech segments are often extended. This is because a speech segments separated by short unvoiced regions have to be connected and the loss of speech frames must be avoided at the beginning and ending of the speech. The extended pause regions are absorbed by a pause model. To assure that this conventional two-stage strategy works well, we investigate the balance of an explicit segmentation by VAD and an implicit segmentation by the pause model using an experimental simulation of segment extension.

2. DATA

The data contained in CENSREC-1-C are constructed by concatenating the same digit string as in CENSREC-1 (formerly called AURORA-2J [5]). The data consist of two major parts: the simulated data by noise-addition and the data recorded in real environments.

2.1. Simulated data

The target evaluation task of the CENSREC-1-C database is voice activity detection in several noise environments. The vocabulary of simulated data included in CENSREC-1-C consists of eleven Japanese digits (“ichi,” “ni,” “san,” “yon,” “go,” “roku,” “nana,” “hachi,” “kyu,” “zero,” and “maru”), a silence (“sil”), and a short pause (“sp”). The recording was conducted in a soundproof booth using a headset microphone, Sennheiser MHD25. The speech data were sampled at 16 kHz, quantized into 16 bit integers, and finally downsampled to 8 kHz. The details of the recording conditions, utterances, and speaking style are the same as in CENSREC-1 (AURORA-2J) [5].

The simulated speech data of CENSREC-1-C are constructed by concatenating several utterances spoken by one speaker. The number of utterances in concatenated speech data is nine or ten. A one-second silence signal taken from CENSREC-1 is inserted between the utterances. In CENSREC-1, the number of speakers per noise environment is 104 (52 males and 52 females). Thus, in CENSREC-1-C, the number of speech data per noise environment is 104.

The noise environments of CENSREC-1-C are shown in Table 1 and are the same as in CENSREC-1 [5].

Table 2. Recording equipment and conditions in real environments.

Microphone	Electret condenser microphone Sony ECM-77 (with close/remote)
Microphone amplifer	Portable multi-mixer Audio-technica AT-PMX5P
Recorder	Linear PCM Recorder Sony PCM-D1
Sampling frequency	8 kHz (Recording is 48 kHz)
Quantization	16 bits

An additional filtering, the ITU-T G.712 recommendation, is applied to consider the realistic frequency characteristics of terminals and equipment in the telecommunication area. Speech signals and noise signals are passed through the filters. Filtered noise signals are artificially added to the filtered speech signals. To add noises at a desired Signal-to-Noise ratio (SNR; Clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB), we calculate the SNR after filtering both signals.

In CENSREC-1, noises in Test Set A are known in the training phase, whereas those in Set B are not. CENSREC-1-C also has types of test sets, Set A and B as CENSREC-1, but there are no training phases in CENSREC-1-C, so there are no differences between Set A and B in the experimental condition. It should be also mentioned that CENSREC-1-C does not have Set C, a data set with channel distortion.

2.2. Data recorded in real environments

We recorded the speech data in two real noisy environments (in a university restaurant and near a highway) and in two SNR conditions (the low and high SNR conditions) based on the conditions shown in Table 2. In this case, we conducted synchronous recording with a close microphone (head-set microphone) and a remote microphone (50 cm away from speaker's mouth). The data from the close microphone, however, are not used in the test. We defined the low SNR as a crowded university restaurant (avg. 69.7 dBA) and near a main highway (avg. 69.2 dBA), and the high SNR as an uncrowded university restaurant (avg. 53.4 dBA) and near a subsidiary highway (avg. 58.4 dBA). Ten subjects for recording speech were employed (three women and three men about twenty years old, one woman and one man about thirty years old, and one woman and one man over forty years old). The recorded speech consists of four files for one subject (total of all utterances: 38-39). A single file includes 8-10 utterances with continuous numbers consisting of 1-12 digit numbers and two-second intervals for each utterance in each noisy environment and each SNR condition.

One subject tends to put a long time interval between digits in one continuous digit utterance. It is hard for most VAD algorithms to decide whether such an utterance is one continuous digit utterance or not. Therefore the speech data of the subject were not used as the evaluation data, but were included as realistic samples in the database.

The recorded speech data are 1380 utterances (144 files) for nine evaluation subjects in two real noisy environments and two SNR conditions. If the removed subject is included, the recorded speech data are 1532 utterances (160 files) for ten subjects.

The condition of CENSREC-1-C is entirely the same as CENSREC-1, so the recognition experiments can be conducted using the HMMs trained using scripts in CENSREC-1. But the HMM training and recognition script are not distributed as contents of CENSREC-1-C so far.

Table 3. Conditions included performance evaluation for each SNR condition.

SNR cond.	Simulated data	Real data
Clean	Clean	n/a
High SNR	20, 15, 10 dB of SNR	high SNR env.
Low SNR	5, 0, -5 dB of SNR	low SNR env.
Average cond.	Clean and 20, 15, 10, 5, 0, -5 dB of SNR	high and low SNR env.

2.3. Time label file

In addition to concatenated speech data, CENSREC-1-C provides time label files with information on speech periods which are used for the evaluation of the voice activity detection. The correct segmentations for clean data (in the simulation) and the close microphone (in the real environment) were made manually by humans. In addition, the correct segmentations for the remote microphone were designed by adding the 12 samples' delay, which is corresponding to the distance between close and far microphones, to the segmentation results for the close microphone.

3. PERFORMANCE EVALUATION MEASURES

CENSREC-1-C includes evaluation tools. Users can easily evaluate and compare their VAD algorithm with baseline results described in Section 4. We adopt two types of evaluation measures in CENSREC-1-C: frame-level measures and utterance-level measures. The former are intuitive and well-known, whereas the latter are originally proposed in this paper as new speech recognition-oriented measures.

3.1. Frame-level evaluation

The frame-level performance evaluation of VAD algorithms is based on FRR (False Rejection Rate) and FAR (False Acceptance Rate) defined by

$$\mathrm{FRR} = \frac{N_{FR}}{N_s} \times 100\ [\%] \text{ and} \quad (1)$$

$$\mathrm{FAR} = \frac{N_{FA}}{N_{ns}} \times 100\ [\%], \quad (2)$$

where N_s, N_{ns}, N_{FR}, and N_{FA} are the total number of speech frames, the total number of non-speech frames, the number of speech frames detected as non-speech frames, and the number of non-speech frames detected as speech frames, respectively. When multiple speech data files are evaluated, the average FRR and FAR are used. They are defined by

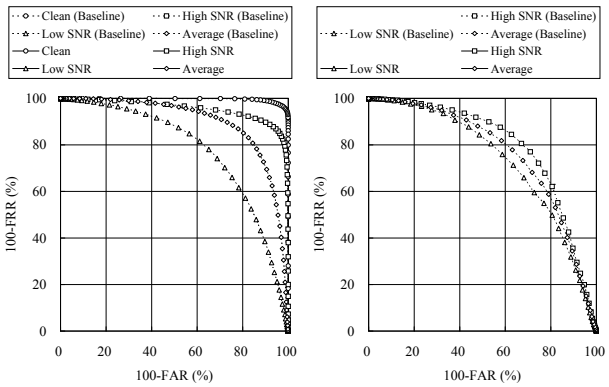
$$\overline{\mathrm{FRR}} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{FRR}_m \text{ and} \quad (3)$$

$$\overline{\mathrm{FAR}} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{FAR}_m, \quad (4)$$

where M is the number of speech data files, and FRR_m and FAR_m are the FAR and FRR for the mth data file, respectively. Note that the frame length is arbitrary[1].

In principle, the performance of VAD algorithms is evaluated using ROC (Receiver Operating Characteristic) curves, where the x-

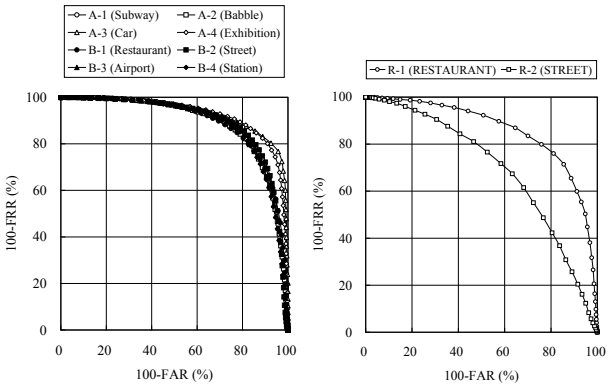
[1]It is not appropriate to determine the frame length *a priori*, since the VAD performance depends on the frame length. Typical speech processing (ex. recognition, enhancement, and coding) requires VAD results at intervals of several ms (or several tens of ms). In this case, there are no serious differences in the FRR and the FAR.



(a) Simulated data (b) Real data

Fig. 1. ROC curve of each SNR for baseline VAD algorithm

These figures will include the results of the other method indicated by the solid lines, to compare with the baseline results indicated by the dashed lines. Now the legends include all the lines but the figures do not include the solid lines.



(a) Simulated data (b) Real data

Fig. 2. ROC curve of each noise for baseline VAD algorithm

axis and y-axis are $100 - \mathrm{FAR}$ and $100 - \mathrm{FRR}$, respectively. In CENSREC-1-C, two types of ROC curves are used:

- ROC curves obtained from VAD results for each SNR condition
- ROC curves obtained from VAD results for each noise type.

The former is used for comparing the performance of user's VAD algorithm with that of the baseline (or target) algorithm. The ROC curves for each of the four SNR conditions described in Table 3 are obtained by averaging over all the noise types. On the other hand, the latter is used not for comparative evaluation but for evaluating the robustness of your VAD algorithm against the noise types. Only the "Average cond." in Table 3 is considered. The results for the baseline VAD algorithms described in Section 4 are shown in Figures 1 and 2.

3.2. Utterance-Level Performance Evaluation

A voice activity detector used as a front end for a speech recognition system generally detects the endpoints of speech utterances, e.g., words, connected words, or sentences. In the case of CENSREC-1-C, an utterance corresponds to a connected digit string. To evaluate the utterance-level VAD performance, two evaluation measures "`Corr`" (Correct rate of utterance boundary detection) and "`Acc`" (Accuracy of utterance boundary detection) are used:

$$\mathrm{Corr} = \frac{N_c}{N} \times 100\ [\%] \text{ and} \quad (5)$$

$$\mathrm{Acc} = \frac{N_c - N_f}{N} \times 100\ [\%], \quad (6)$$

where N is the total number of speech utterances, N_c is the number of correctly detected utterances, and N_f is the number of incorrectly detected utterances. "`Corr`" assesses how many speech utterances can be detected by VAD algorithms, whereas "`Acc`" also takes into account the number of over-detected utterances. These are regarded as the recognition performance by the 'ideal' recognizer which can perfectly recognize correctly detected speech.

When evaluating VAD algorithms for more than one speech data file, average `Corr` and average `Acc` are used,

$$\overline{\mathrm{Corr}} = \frac{\sum_{m=1}^{M} N_{c,m}}{\sum_{m=1}^{M} N_m} \times 100\ [\%] \text{ and} \quad (7)$$

$$\overline{\mathrm{Acc}} = \frac{\sum_{m=1}^{M} N_{c,m} - \sum_{m=1}^{M} N_{f,m}}{\sum_{m=1}^{M} N_m} \times 100\ [\%], \quad (8)$$

where M is the total number of speech data, and N_m, $N_{c,m}$, and $N_{f,m}$ correspond to N, N_c, and N_f of the m-th speech data, respectively.

If a detected speech segment is shorter than an actual speech segment, phoneme information at the beginning or the end of the utterance is missing, resulting in a speech recognition error. On the other hand, even if a detected speech segment has additional non-speech ranges before and after the speech utterance, it would not cause significant damage to speech recognition performance. We assume that, if a detected speech segment includes all speech intervals of an utterance without overlapping either the preceding or succeeding speech utterance, it can be a candidate for correct detection.

There are chances that the baseline VAD algorithm in CENSREC-1-C may detect more than one speech segment for each utterance because the baseline algorithm allows the overlap of detected speech segments. In such a case, only one of the detected segments for an utterance is counted as a correct detection (N_c) and others are counted as false detections (N_f). Such over-detection does not affect `Corr` but decreases `Acc`.

All other kinds of detection results are counted as error detection (N_f), including the following cases: (a) a detected segment only includes non-speech range, (b) either or both endpoints of a detected segment exist at speech intervals, and (c) a detected segment covers two or more speech utterances.

For this evaluation measure, better evaluation results can be obtained by adding appropriate margins, e.g., a few hundred ms, at both ends of detected segments because even a short loss of speech segments at both ends causes a detection error.

The final result should be obtained using the threshold with the maximum average `Corr`. This means that the threshold is selected to maximize the detection rate and then the precision (here, `Acc.`) is evaluated. The threshold can be different among simulated data and real data, but should be the same in one of these data sets. Tables 4 and 5 show the results for the baseline VAD algorithms described in Section 4.

Note that in this evaluation measure a detected segment is judged correct if the segment includes a whole true speech segment. However, detection with a too long extension may degrade recognition performance of actual speech recognition (see Section 5). So this measure tends to give optimistic evaluation results. Users have to consider this utterance-level evaluation and frame-level evaluation simultaneously. Or we have to improve this measure considering the

Table 4. Utterance-level evaluation of baseline VAD algorithm for the simulated data (Top: Correct rate; bottom: Accuracy)

Simulated Data	Correct Rate [%]										
		A					B					Overall
		Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Average
Correct Rate [%]	Clean	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90	99.90
	20 dB	96.60	96.10	96.39	95.80	96.22	97.10	97.10	96.89	96.20	96.82	96.52
	15 dB	94.01	93.31	93.99	92.91	93.56	95.70	96.50	95.49	94.51	95.55	94.55
	10 dB	91.31	90.01	86.97	88.01	89.08	94.51	93.81	92.69	88.71	92.43	90.75
	5 dB	93.31	87.61	81.16	81.72	85.95	84.72	78.92	84.37	72.83	80.21	83.08
	0 dB	48.35	42.56	76.15	78.02	61.27	44.76	56.54	62.32	47.45	52.77	57.02
	-5 dB	9.39	21.88	65.83	64.54	40.41	27.17	37.26	34.97	28.37	31.94	36.18
	Average	76.12	75.91	85.77	85.84	80.91	77.69	80.00	80.95	75.42	78.52	79.71

Simulated Data	Accuracy [%]										
		A					B					Overall
		Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Average
Accuracy [%]	Clean	99.80	99.90	99.80	99.80	99.83	99.80	99.90	99.80	99.80	99.83	99.83
	20 dB	94.81	94.71	94.89	94.51	94.73	95.60	96.30	95.79	95.40	95.77	95.25
	15 dB	89.91	89.61	90.68	89.51	89.93	93.11	94.01	91.88	91.91	92.73	91.33
	10 dB	84.52	83.92	78.66	74.53	80.41	85.71	87.11	81.86	78.62	83.33	81.87
	5 dB	61.64	69.43	65.03	63.64	64.94	62.04	64.64	66.23	56.04	62.24	63.59
	0 dB	5.09	7.99	55.91	44.66	28.41	5.79	27.57	32.16	21.18	21.68	25.04
	-5 dB	-27.57	-9.99	28.86	16.58	1.97	-19.58	-3.40	-1.80	-3.90	-7.17	-2.60
	Average	58.31	62.22	73.40	69.03	65.74	60.35	66.59	66.56	62.72	64.06	64.90

Table 5. Utterance-level evaluation of baseline VAD algorithm for the real data (Top: Correct rate; bottom: Accuracy)

Real Data	Correct Rate [%]			
		Remote Microphone		
		Restaurant	Street	Average
Correct Rate [%]	High SNR	74.20	39.42	56.81
	Low SNR	56.52	41.45	48.99
	Average	65.36	40.44	52.90

Real Data	Accuracy [%]			
		Remote Microphone		
		Restaurant	Street	Average
Accuracy [%]	High SNR	21.45	-15.65	2.90
	Low SNR	-43.48	-33.91	-38.70
	Average	-11.02	-24.78	-17.90

results in Section 5, to reflect the recognition performance more.

4. BASELINE VAD ALGORITHM

We provide a power-based VAD algorithm as the baseline.

4.1. VAD algorithm

The baseline algorithm works as follows:

1. Speech Framing
   The speech signal is divided into overlapping frames. In this baseline, the frame length and the frame shift are set at 5 ms and 2 ms, respectively.
2. Computing the logarithmic frame energy
   The logarithmic frame energy (POW_i) is computed using the following equation:

$$\mathrm{POW}_i = 10 \cdot \log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1} s_i^2(n)\right), \quad i = 0, 1, \ldots, M-1, \tag{9}$$

   where $s_i(n)$ indicates the speech signal of frame i. M and N indicate the number of frames and the number of points of the speech signal in a frame, respectively.
3. Calculating the threshold
   The initial threshold (THR_{int}) is determined by using the threshold selection method [8]. This method classifies each frame into two classes based on the logarithmic frame energy, speech class (C_1) and non-speech class (C_2).
   The subjective function, $\eta(s)$, is

$$\eta(s) = \frac{\sigma_B^2(s)}{\sigma_T^2}, \tag{10}$$

$$\sigma_T^2 = \sigma_W^2(s) + \sigma_B^2(s). \tag{11}$$

   The parameter, s, which maximizes the subjective function $\eta(s)$, is calculated, where $\sigma_W^2(s)$ and $\sigma_B^2(s)$ are the average variance within the class and the average variance between the classes, respectively [8]. Then, using the initial threshold, the threshold, THR, is determined by the following equation:

$$\mathrm{THR} = \mathrm{THR}_{int} + k \cdot \alpha, \tag{12}$$

$$\alpha = \frac{(\mathrm{POW}_h - \mathrm{POW}_l)}{K}, \tag{13}$$

   where, POW_l is the average of the logarithmic frame energies that are less than the initial threshold and POW_h is the average of the logarithmic frame energies that are equal to or greater than the initial threshold. In the baseline, K is set to 40.
4. Voice Activity Detection
   The VAD is performed as follows:
   (a) Detection of the start-frame of voice
   If the logarithmic frame energy of the frame is greater than the threshold, this frame is set as a start-frame candidate of the voice.
   (b) Detection of the end-frame of voice
   After the candidate frame of voice start is decided, the candidate frame of voice end is decided. The logarithmic frame energy of the candidate frame of voice end is less than the threshold and the non-voiced section,

i.e., the section that is less than the threshold, is longer than 500 ms. If the length of voice section is less than 500 ms, the candidate end-frame of voice must be re-selected.

(c) Detection of the speech section

If the candidate speech section is more than 100 ms, this speech section is adopted. However, if the candidate speech section is less than 100 ms, this speech section is ignored. After detection of the speech section, the process of detecting the candidate of voice start is re-started.

5. Output of the results

The detected frame-level voice section is changed to point-level by using the following equation:
Start-point = Start-frame × Sampling frequency × Frame length (second)
End-point = End-frame × Sampling frequency × Frame length (second) -1

4.2. Baseline performance

4.2.1. Frame-level performance

Results are shown in Figure 1 and Figure 2.

Figure 1 shows that the performance is high under the high SNR conditions for both simulated and real data, but the performance is rapidly degraded as the SNR becomes worse. Figure 2 shows that the difference in noises also affects the VAD performance.

4.2.2. Utterance-level performance

Results are shown in Tables 4 and 5. The threshold with the highest correct rate ($k = 10$ for simulated data and $k = 17$ for real data) was used. Both ends of the voice sections were extended by 300 ms.

These results show that the algorithm obtains good performance in the high SNR condition for both real and simulated data, but the performance degrades along with the SNR degradation and it obviously impacts the speech recognition.

5. RELATIONSHIP BETWEEN VAD AND RECOGNITION ACCURACY

As described previously, the data contained in this database are constructed by concatenating the test data in CENSREC-1. Therefore, we can perform recognition experiments with this database by using the acoustic models trained with the training data in CENSREC-1. In this section, as an example of recognition experiments, we investigated the relationship between VAD accuracy and recognition performance, i.e. how longer/shorter detection than the true speech segment affects the recognition accuracy.

Generally, in order to avoid frame dropping at the beginning or ending of a true speech segment, or even inside the segment, the detected segment is often extended because the loss of speech information makes recognition impossible. As a result of this process, extra pauses are attached, especially to the beginning and ending of the segments. In usual speech recognition technique, a noise model is applied to absorb such pause segments. Although this two-stage strategy (VAD and pause model) is considered to work well, it is not supported by evidence-based investigation. Therefore, we evaluate the performance of this approach by simulating the VAD performance on CENSREC-1-C.

In this experiment, the length of the detected speech segments are varied artificially, and we examine the change of recognition performance.

5.1. Experimental conditions

We vary the extention of speech segments in the range of -200 ms (with the beginning and ending of speech segments cut off 200 ms, respectively) to $+200$ ms (similarly expanded) every 10 ms for the true segments. "0 ms" corresponds to an ideal VAD result. We used it as VAD results.

Experimental conditions for the recognition process were almost the same as that of the baseline of CENSREC-1 [5] except for the feature parameters, which were 12MFCC $+12\Delta$MFCC $+12\Delta\Delta$MFCC $+\Delta$pow $+\Delta\Delta$pow. Same as in the baseline scripts of CENSREC-1, HMMs were whole digit models having 16 emitting states with 32 Gaussian pdfs in each state. The recognition experiments were performed for each detected segment.

CENSREC-1 includes two types of training conditions [5]: clean training in which HMMs are trained using only clean speech, and multi-condition training in which HMMs are trained under multiple noisy environments. We first compare these two conditions. Then we compare the cases with/without a pause model, which is expected to absorb the extra pauses.

In recognition systems for noisy speech, noise suppression methods are generally used. Hence, we used the spectral subtraction-based methods (SS) described in [9] with a subtraction gain of $\alpha = 1.0$, a frame length of 32 ms, and a frame shift length of 6.25 ms. The noise spectrum is estimated using the first 30 frames in each speech file.

5.2. Results and discussions

Results are shown in Figs. 3 – 8, in which "sp" indicates the use of a pause mode. In the "baseline" condition, we used clean HMMs with a pause model. SS is not used in the "baseline." In each figure we can find the performance difference between "baseline" and the other conditions: HMMs, pause model, and SS. The X-axis indicates the extension length of speech segments. For example, -200 and 200 mean the deletion and extension of speech segments, respectively. The recognition accuracy is averaged over the kinds of noises for each SNR. The results are shown for the conditions of clean, 10 dB, and 0 dB SNR for simulation data.

From all 8 figures, we find that the intuition that the deletion degrades the recognition performance more than the extension is correct. For the clean data, we obtained the maximal performance at around $+40$ ms. Small extension yields better results. As SNR degrades, the extension with maximal performance approaches zero and the small extension and deletion lead to severe degradation. This suggests that the more the SNR degrades, the more accurate VAD must be. The degradation tendencies are almost the same between simulation data and real data. The results from the simulation data, which are easily obtained, are also valid for the real environments.

From Figs. 3 and 6, we find that the absolute performance by the multi-training HMMs is much better than by the clean training HMMs, but the degradation tendencies caused by the change of extension are almost the same.

Figs. 4 and 7 shows that the degradation caused by extension was much larger without a pause model. The pause model matched the silences and reduce the false alarms. In the low SNR cases, however, the difference between with/without a pause model is small because the simple pause model could not match the various loud noises. But the maximal points approach $+40$ as clean condition when using SS as shown in Figs. 5 and 8.

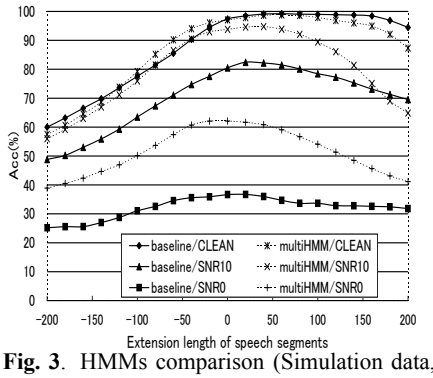


Fig. 3. HMMs comparison (Simulation data, with sp)

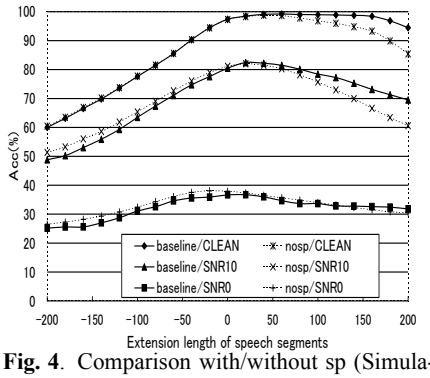


Fig. 4. Comparison with/without sp (Simulation data, clean HMMs)

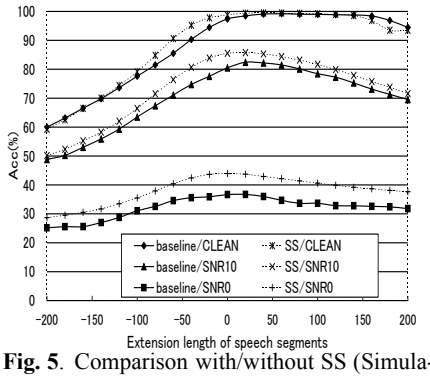


Fig. 5. Comparison with/without SS (Simulation data, clean HMMs with sp)

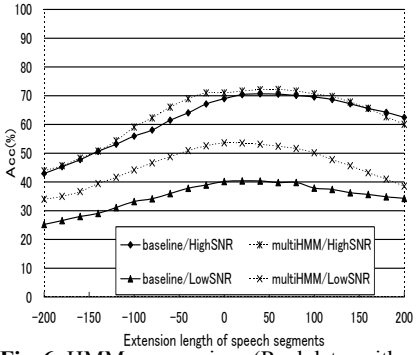


Fig. 6. HMMs comparison (Real data, with sp)

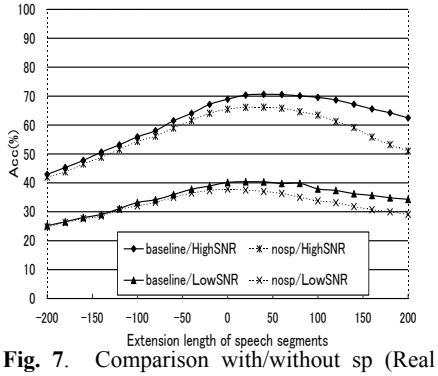


Fig. 7. Comparison with/without sp (Real data, clean HMMs)

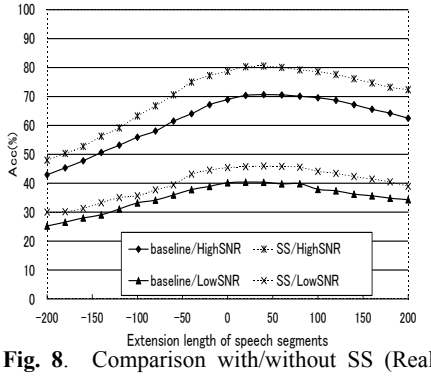


Fig. 8. Comparison with/without SS (Real data, clean HMMs with sp)

Note that we used 16-state whole-word HMMs, and an HMM corresponds to at least 16 frames (= 175 ms) in speech data. If we use 3-state phone models, each model corresponds to much shorter speech segments (at least 3 frames (= 45 ms)), and thus insertion errors may readily occur and compromise the accuracy. Thus, we carefully discuss the tendencies described in this section considering the recognition units [2].

6. CONCLUSION

In this paper, we introduced CENSREC-1-C, a new evaluation framework for VAD, under noisy conditions. This framework provides not only the database but also the evaluation measures. The framework was released in September 2006, and many studies are being conducted using it in Japan. We will propose another evaluation measure based on speech recognition in the near future. We also evaluate the effect of the detected segment extension using CENSREC-1-C and showed that the small extension improves speech recognition.

7. ACKNOWLEDGEMENT

The authors wish to thank the members of the Speech Resources Consortium in the National Institute of Informatics (NII-SRC), Japan, for their generous assistance in these activities. The present study was conducted using the CENSREC-1-C database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

8. REFERENCES

[1] http://elazar.itd.nrl.navy.mil/spine/

[2] H. G. Hirsh, D. Pearce, “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” ISCA ITRW ASR2000, September, 2000

[3] Aurora document no. AU/345/01, “Large vocabulary evaluation of front-ends- baseline recognition system description,” Mississippi State University, Aug. 2001.

[4] AURORA-J/CENSREC Web site: http://sp.shinshu-u.ac.jp/CENSREC/

[5] S. Nakamura *et al.*, “AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition,” *IEICE Transactions on Information and Systems*, Vol. E88–D, No. 3, pp. 535–544, 2005.

[6] S. Nakamura, M. Fujimoto, and K. Takeda, “CENSREC2: Corpus and Evaluation Environments for In Car Continuous Digit Speech Recognition,” Proc. ICSLP ’06, pp. 2330-2333, Sept. 2006.

[7] M. Fujimoto, K. Takeda, and S. Nakamura, “CENSREC-3: An Evaluation Framework for Japanese Speech Recognition in Real Driving-Car Environments,” *IEICE Transactions on Information and Systems*, (accepted).

[8] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Trans. Sys., Man, and Cybernetics, Vol. SMC-9, No. 1, pp. 62–66, 1979.

[9] N. Kitaoka and S. Nakagawa. “Evaluation of spectral subtraction with smoothing of time direction on the AURORA 2 task”, ICSLP-2002, pp. 465–468, 2002.

[2] The constraints such as lexicons and language models affect the tendencies, of course.