PAPER Special Section on Corpus-Based Speech Technologies

# AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition

**Satoshi NAKAMURA**[†a)], **Kazuya TAKEDA**[††], **Kazumasa YAMAMOTO**[†††], **Takeshi YAMADA**[††††], **Shingo KUROIWA**[†††††], **Norihide KITAOKA**[††††††], **Takanobu NISHIURA**[*], **Akira SASOU**[**], **Mitsunori MIZUMACHI**[†***], **Chiyomi MIYAJIMA**[††], **Masakiyo FUJIMOTO**[†], ***and*** **Toshiki ENDO**[†], ***Members***

**SUMMARY** This paper introduces an evaluation framework for Japanese noisy speech recognition named AURORA-2J. Speech recognition systems must still be improved to be robust to noisy environments, but this improvement requires development of the standard evaluation corpus and assessment technologies. Recently, the Aurora 2, 3 and 4 corpora and their evaluation scenarios have had significant impact on noisy speech recognition research. The AURORA-2J is a Japanese connected digits corpus and its evaluation scripts are designed in the same way as Aurora 2 with the help of European Telecommunications Standards Institute (ETSI) AURORA group. This paper describes the data collection, baseline scripts, and its baseline performance. We also propose a new performance analysis method that considers differences in recognition performance among speakers. This method is based on the word accuracy per speaker, revealing the degree of the individual difference of the recognition performance. We also propose categorization of modifications, applied to the original HTK baseline system, which helps in comparing the systems and in recognizing technologies that improve the performance best within the same category.

***key words:*** *noisy speech recognition, evaluation platform, performance differences over speakers, evaluation categories*

## 1. Introduction

The recent progress of speech recognition technology has been brought about by the advent of statistical modeling and large-scale corpora. Furthermore, it is also known that progress has been accelerated by the U.S. DARPA projects initiated in the late '80s in terms of project participants competitively developing speech recognition systems on the same task, using the same training and test corpus.

However, current speech recognition performance must still be improved if the system is to be exposed to noisy environments, where speech recognition applications might be used in practice. Therefore, robustness to acoustic noise is an emerging and crucial factor to be solved for speech recognition systems.

With regard to the noise robustness problem, there have been two evaluation projects, SPINE1, 2 [1] and AURORA [2]. The SPINE (SPeech recognition In Noisy Environment) project was organized by U.S. DARPA, with SPINE1 in 2000 and SPINE2 in 2001. The task included spontaneous English dialog between an operator and a soldier in a noisy field to evaluate spontaneous continuous speech recognition in noisy environments. The results of the project brought many improvements to continuous noisy speech recognition, though the task seems quite special and a little difficult to handle.

On the other hand, the European Telecommunications Standards Institute (ETSI) AURORA group initiated a special session in the EUROSPEECH conference. They are actively working to develop standard technologies under ETSI for distributed speech recognition [3]. In parallel with their standardization activities, they have distributed to academic researchers a noisy connected speech corpus based on TI digits with baseline HTK scripts for further noisy speech recognition research. To date, Aurora 2, a connected digit corpus of telephone band-limited speech with additive noise, and Aurora 3, an in-car noisy digit and word corpus, have been distributed with HTK scripts, which can be used to obtain baseline performance and relative improvements over the baseline results [5], [6]. The advantages of the AURORA are 1) the connected digit task is relatively small compared to spontaneous speech, and 2) the baseline performance can be easily attained by the attached HTK scripts.

The authors voluntarily organized a special working group in October 2001 under the auspices of the Information Processing Society of Japan in order to assess speech recognition technology in noisy environments. The focus of the working group included the planning of comprehensive fundamental assessments of noisy speech recognition, standardized corpus collection, evaluation strategy developments, and distribution of standardized processing modules. To begin with, we decided to follow the Aurora 2, connected digit telephone band-width speech corpus and evaluation,

Manuscript received June 30, 2004.
Manuscript revised September 5, 2004.

[†]The authors are with the ATR Spoken Language Translation Research Laboratories, "Keihanna Science City", Kyoto-fu, 619–0288 Japan.
[††]The authors are with Nagoya University, Nagoya-shi, 464–8603 Japan.
[†††]The author is with Shinshu University, Nagano-shi, 380–8553 Japan.
[††††]The author is with University of Tsukuba, Tsukuba-shi, 305–8573 Japan.
[†††††]The author is with University of Tokushima, Tokushima-shi, 770–8506 Japan.
[††††††]The author is with Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
[*]The author is with Ritsumeikan University, Kusatsu-shi, 525–8577 Japan.
[**]The author is with National Institute of Advanced Industrial Science and Technology, Tsukuba-shi, 305–8568 Japan.
[***]Presently, with Kyushu Institute of Technology, Kitakyushu-shi, 804–8550 Japan.
a) E-mail: satoshi.nakamura@atr.jp
DOI: 10.1093/ietisy/e88–d.3.535

since the task is small enough and the evaluation scheme is quite clear. As for the Japanese Aurora 2, AURORA-2J, we have translated English digits into Japanese digits. Although there are alternative pronunciations for Japanese digit pronunciation, we specified one of those taking account of occurrence frequencies in a real usage. Then we added the same noise as that in Aurora 2 to Japanese digit data. This paper also describes a new baseline HTK script designed for AURORA-2J.

The Aurora 2 provided an Excel spreadsheet to calculate average word accuracy and relative improvements compared to those of the baseline system. This spreadsheet is indeed useful for showing average performances and improvements by the proposed methods. However, these scores do not always reflect the real feelings of users of speech recognition systems. Thus, we further investigate a new method to analyze the speech recognition performance in terms of performance differences among speakers.

Finally as researchers develop new methods and improve performance, comparisons become increasingly more difficult. To avoid this problem, we also propose categorization of modifications that applied to the original HTK baseline system. This categorization will help us to compare the systems and to recognize technologies that improve the performance most in the same category.

In this paper, Sect. 2 describes the AURORA-2J corpus collection and its evaluation scripts. The evaluation schemes and the results are described in Sect. 3. Section 4 describes the categories in which the developed noisy speech recognition system should be fairly compared. Finally, Sect. 5 summarizes the paper and describes future directions.

## 2. AURORA-2J Data Configuration and Baseline Recognition System

### 2.1 Pronunciation of Japanese Digits

The data contained in AURORA-2J is the same as in Aurora 2, but uttered in Japanese. The number of speakers is the same and the digit strings for each speaker are identical. Figure 1 shows a histogram of the number of digits in the training data and test data. Although there are no six digit utterances in the database†, the occurrence frequencies of two-, three-, four-, five-, and seven-digit utterances are almost equal. Table 1 shows the pronunciations of eleven digits in Aurora 2 and AURORA-2J. Speakers were requested to pronounce digits as specified in this table. The occurrence frequency of each digit is also shown in the table, indicating that the occurrences of these digits are also well balanced in the database.

Although *vowel lengthening* sometimes occurs in /ni/ and /go/, the two pronunciations are not distinguished in AURORA-2J.

Sometimes, "4" is read as /shi/, "7" is read as /shichi/, and "0" is read as /rei/ in Japanese. However, in order to make perplexity the same as Aurora 2, these pronunciations were not employed in AURORA-2J.

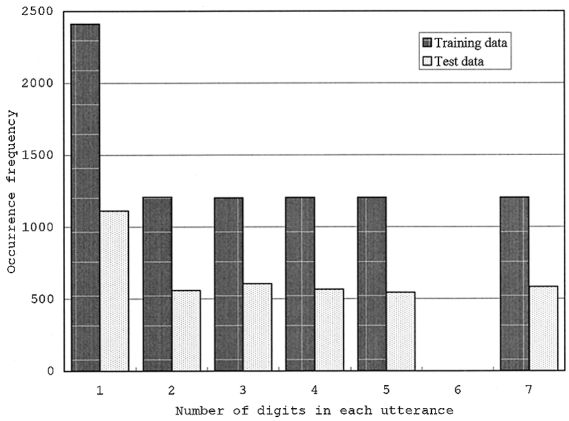



**Fig. 1** Histogram of number of digits in training data and test data.

**Table 1** Pronunciation of digits and occurrence frequency of each digit.

| Digit | AURORA-2 | AURORA-2J | Frequency training / test |
|---|---|---|---|
| 1 | one | /ichi/ | 2545 / 1221 |
| 2 | two | /ni/ | 2531 / 1203 |
| 3 | three | /saN/ | 2521 / 1200 |
| 4 | four | /yoN/ | 2539 / 1182 |
| 5 | five | /go/ | 2491 / 1222 |
| 6 | six | /roku/ | 2545 / 1187 |
| 7 | seven | /nana/ | 2525 / 1192 |
| 8 | eight | /hachi/ | 2515 / 1175 |
| 9 | nine | /kyuH/ | 2492 / 1183 |
| 0(Z) | zero | /zero/ | 2500 / 1190 |
| 0(O) | oh | /maru/ | 2523 / 1204 |

### 2.2 Data Recording

A headset microphone, Sennheiser MHD25, was used for recording with a USB-audio interface (Edirol UA-5) connected to a Windows personal computer. The recording was done in a soundproof booth where speakers read a list of digit strings presented on CRT monitor screen connected to the PC. The speech was sampled at 16 kHz and quantized to 16 bit linear PCM data without any compression.

### 2.3 Filtering and Noise Adding

The AURORA-2J database follows the Aurora 2 database, and was created in exactly the same way. All programs and scripts used here were kindly provided by the AURORA project for both filtering and noise adding.

#### 2.3.1 Filtering

An additional filtering is applied to consider the realistic frequency characteristics of terminals and equipment in the telecommunication area. Two types of "standard" frequency

†The source speech for Aurora 2 is TIdigits [4] which does not include six digit utterances, either. Unfortunately, the document of TIdigits does not mention the reason for this.

characteristics are used: G.712 and MIRS [5]. The recommended G.712 provides the transmission performance characteristics of PCM channels in digital transmission equipment, and its frequency characteristic is flat in the range between 300 and 3,400 Hz. MIRS can be seen as a frequency characteristic that simulates the behavior of a telecommunication terminal, and its frequency characteristics shows a rising characteristic with an attenuation of lower frequency. Speech signals and noise signals are passed through either of these filters.

#### 2.3.2 Noise Adding

Filtered noise signals are artificially added to the filtered speech signals. To add noises at a desired SNR (20 dB, 15 dB, 10 dB, 5 dB, 0 dB and −5 dB), we calculate the SNR after filtering both signals. We also use noise signals as follows:

- Subway
- Babble
- Car
- Exhibition hall
- Restaurant
- Street
- Airport
- Train Station.

These are considered as the most probable cases for telecommunication terminals such as cellphones. And our noise data is the same as that for Aurora 2, which is described in [5] for details. Since the airport noise contains Japanese and English annauncement, influences on the performance are supposed to be the same.

### 2.4 Training/Testing Dataset

The design of the training and testing datasets is the same as that of Aurora 2. Two sets of training data are prepared, such as a clean-training dataset and a multicondition dataset. Total utterances are 8,440 by 110 speakers (55 male and 55 female speakers). For the multicondition training dataset, four types of noise (Subway, Babble, Car, Exhibition) are added to the clean speech in five types of SNR (clean, 20 dB, 15 dB, 10 dB, 5 dB). For each noise and SNR condition, 422 utterances are included. The G.712 filter is applied to all the speech data.

For the testing dataset, we prepare three types of dataset completely the same way as in Aurora 2.

[Testset A] The noise condition is the same as in the multicondition training set. Subway, Babble, Car, and Exhibition noises are used.

[Testset B] The noise condition is different from the multicondition training dataset. Restaurant, Street, Airport, and Station noises are used.

[Testset C] The channel condition is different from the training dataset. The MIRS channel is applied to the speech data. Subway and Street noises are used.

### 2.5 Reference Scripts

The reference back-end scripts using HTK [7] are mostly based on the original Aurora 2 baseline back-end scripts [5], and some modifications were introduced from the Microsoft complex baseline back-end scripts [8]. The reference scripts are written in sh and perl, and there is an additional feature — the number of dimensions of feature vectors can be easily changed.

We chose a setup with the best performance as the baseline among various conventional ones without noise-robust techniques. The various conditions in the reference scripts were determined experimentally, and many experimental investigations have been conducted. For example, although introduction of the auto label production in training phase was considered as in the Microsoft scripts, it was not introduced due to its effect of degrading the recognition performance. We also exermined the case of not using log energy and energy normalization. Experiments were performed on various conditions of log energy; we decided on using log energy in this reference baseline.

As a result, the number of recognition units, the HMM topology, and the training procedure were basically the same as the original Aurora 2 conditions, except for the strategy of increasing the number of Gaussian mixtures per state. The feature vector consists of 12 MFCC and log energy with their corresponding delta and acceleration coefficients and cepstral mean normalization (CMN) was not applied to these features. Thus, each vector contained 39 components in total. These parameters were calculated using HCopy with the same conditions as the Aurora 2 HTK baseline. Table 2 summarizes the speech analysis and HMM conditions.

Figure 2 shows the recognition grammar in the reference scripts, where '|' denotes alternatives, '<>' denotes one or more repetitions, and '[ ]' encloses options. This grammar generates arbitrary repetitions of digits optionally followed by short pauses, and tereminal silences are also allowed. Since this reference script aims at the evaluation of acoustic models, a very simple language model is actually used to avoid the effect the language model has on the recognition performance.

**Table 2** Speech analysis and HMM conditions.

| | |
|---|---|
| Sampling frequency | 8 kHz |
| Pre-emphasis | $1 - 0.97z^{-1}$ |
| Analysis window | Hamming |
| Window length | 25 ms |
| Frame shift | 10 ms |
| # Mel FB channels | 23 |
| Feature parameters | 12 MFCC +Δ + ΔΔ<br>+ log energy +Δ + ΔΔ |
| # recog. units | 13 (11 digits + silence + short pause) |
| # HMM states | 16 for digits, 3 for silence,<br>1 for short pause<br>(shared with middle state of sil.) |
| # mixtures | 20 for digits<br>36 for silence and short pause |

```
$digit = one | two | three | four |
         five | six | seven | eight |
         nine | zero | oh ;
(
  [sil] < $digit [sp] > [sil]
)
```

**Fig. 2** Grammar written in EBNF.

## 3. Evaluation Schemes

To evaluate a speech recognition system, the average recognition performance of speakers is often used. Since AURORA-2J is the platform for evaluating noisy speech recognition, we define the calculation method of the average recognition rates that enables us to evaluate the noise-reduction methods not only by the overall performance, but also by the performance for each noise condition.

Indeed, these recognition rates are good measurements; these values are obtained by averaging the performance among the speakers, and the result is impossible to tell the any difference in the recognition performances among speakers. In real use, however, individual differences reflect the real feelings of users of speech recognition systems. To overcome this, we propose a new method to compensate for the speaker variability.

Section 3.1 describes the ETSI standard DSR front-end ES 202 050 [9]. The front-end is used as an example for the evaluation. In Sect. 3.2, we explain the details of the average recognition performance evaluation measures, a tool for calculating the measures, and an example evaluation result. In Sect. 3.3, we introduce a new method to analyze the effect of speaker variability on the recognition performance.

### 3.1 ETSI ES 202 050

Figure 3 shows a block diagram of ES 202 050 on the terminal side. Cepstral features are computed from an input signal after noise reduction and waveform processing, then blind equalization is applied to the features. Finally, the features are compressed and further processed for channel transmission. In this paper, however, only the feature extraction part in Fig. 3 is used.

Noise reduction is based on Wiener filter theory and performed in two stages as shown in Fig. 4. In the first stage, the linear spectrum of each frame is estimated, then frequency domain Wiener filter coefficients are computed by using the current frame spectrum and the spectra of the noise frames detected by using the VAD (Voice Activity Detection). In the PSD (Power Spectral Density) Mean block, the spectrum is smoothed along the frame index. Finally, the input signal is de-noised by filtering the time domain Mel-warped Wiener filter, which is converted from the Wiener filter for the linear frequency domain. In the second stage, additional and dynamic noise reduction is performed according to the signal-to-noise ratio of the output signal in

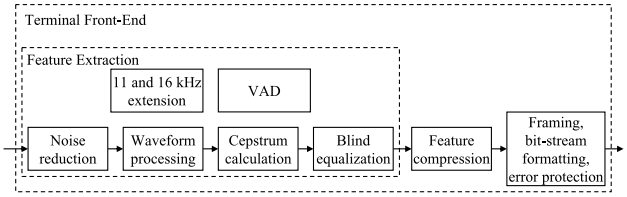



**Fig. 3** Block diagram of ES 202 050 on the terminal side.

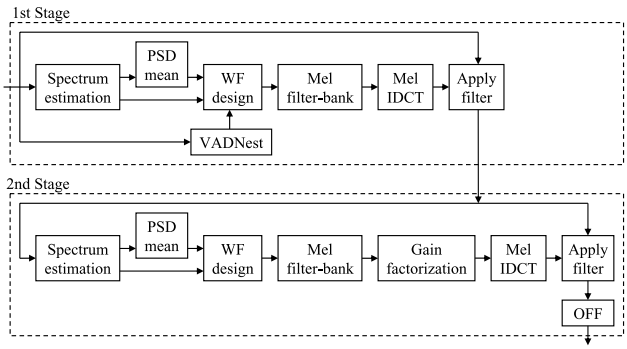



**Fig. 4** Block diagram of the noise reduction part.

the first stage. After this, the DC offset of the de-noised signal is removed.

### 3.2 Evaluation Based on Average Recognition Performance

The AURORA-2J provides a common Microsoft Excel spreadsheet. This spreadsheet automatically calculates the relative performance toward the baseline performance when recognition results are inputted into the spreadsheet. Using this, we can compare the recognition performance objectively among various noise-reduction methods.

Table 3 shows the details of AURORA-2J's baseline performance and the evaluation results by the ES 202 050 front-end, while Table 4 shows a summary of the evaluation. These tables are obtained by the provided spreadsheet explained above.

In Table 3, *%Acc* and *Relative performance* are calculated by the following equations:

$$\%Acc = \frac{H - I}{N} \times 100(\%), \tag{1}$$

$$\textit{Relative performance} = \frac{\%Acc - \%Acc\ of\ baseline}{100 - \%Acc\ of\ baseline} \times 100(\%), \tag{2}$$

where $H$, $I$ and $N$ indicate the number of correct words, the number of inserted words and the total number of words, respectively. Note that the average *%Acc* for each type of noise is calculated by averaging the *%Acc* for 0 dB to 20 dB SNRs. The calculation method of *%Acc* is the same as that of Aurora 2 used as a standard evaluation environments of noisy speech recognition [5]. Then, the average *%Acc* of each testset is calculated by averaging the *%Acc* of the noises included in the testset. The overall average *%Acc* is

**Table 3** AURORA-2J baseline performance and evaluation results of ES 202 050 front-end. The upper is the AURORA-2J baseline performance, the middle is word accuracy of ES 202 050 front-end and the lower is the relative performance toward the baseline performance.

AURORA-2J baseline performance

| Clean Training (%Acc) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | | | | B | | | | | C | | | Overall |
| | Subway | Babble | Car | Exhibition | Average | Restaurant | Street | Airport | Station | Average | Subway M | Street M | Average | Average |
| Clean | 99.72 | 99.58 | 99.82 | 99.60 | 99.68 | 99.72 | 99.58 | 99.82 | 99.60 | 99.68 | 99.82 | 99.67 | 99.75 | 99.69 |
| 20 dB | 96.90 | 80.80 | 89.59 | 95.90 | 90.80 | 84.86 | 88.51 | 82.17 | 82.29 | 84.46 | 91.50 | 92.26 | 91.88 | 88.48 |
| 15 dB | 76.27 | 56.83 | 58.16 | 75.41 | 66.67 | 61.10 | 65.39 | 57.80 | 55.01 | 59.83 | 70.80 | 75.39 | 73.10 | 65.22 |
| 10 dB | 47.16 | 38.63 | 38.86 | 41.65 | 41.58 | 40.50 | 42.59 | 41.93 | 37.98 | 40.75 | 43.51 | 47.28 | 45.40 | 42.01 |
| 5 dB | 25.27 | 23.16 | 20.79 | 21.97 | 22.80 | 21.06 | 23.79 | 26.16 | 22.25 | 23.32 | 25.91 | 25.03 | 25.47 | 23.54 |
| 0 dB | 12.28 | 8.16 | 10.38 | 11.97 | 10.70 | 9.89 | 13.75 | 12.68 | 9.84 | 11.54 | 13.72 | 13.60 | 13.66 | 11.63 |
| -5 dB | 7.43 | 4.35 | 7.25 | 7.90 | 6.73 | 1.90 | 8.56 | 4.77 | 5.46 | 5.17 | 8.81 | 8.74 | 8.78 | 6.52 |
| Average | 51.58 | 41.52 | 43.56 | 49.38 | 46.51 | 43.48 | 46.81 | 44.15 | 41.47 | 43.98 | 49.09 | 50.71 | 49.90 | 46.17 |

| Multicondition Training (%Acc) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | | | | B | | | | | C | | | Overall |
| | Subway | Babble | Car | Exhibition | Average | Restaurant | Street | Airport | Station | Average | Subway M | Street M | Average | Average |
| Clean | 99.79 | 99.64 | 99.67 | 99.75 | 99.71 | 99.79 | 99.64 | 99.67 | 99.75 | 99.71 | 99.69 | 99.55 | 99.62 | 99.69 |
| 20 dB | 99.63 | 99.67 | 99.70 | 99.57 | 99.64 | 98.62 | 99.46 | 98.90 | 97.99 | 98.74 | 99.51 | 99.40 | 99.46 | 99.25 |
| 15 dB | 99.26 | 99.40 | 99.37 | 98.83 | 99.22 | 96.90 | 97.58 | 96.45 | 94.11 | 96.26 | 99.17 | 98.37 | 98.77 | 97.94 |
| 10 dB | 98.25 | 97.43 | 97.94 | 97.38 | 97.75 | 86.83 | 89.57 | 91.29 | 84.94 | 88.16 | 96.90 | 93.71 | 95.31 | 93.42 |
| 5 dB | 93.89 | 89.78 | 92.16 | 92.32 | 92.04 | 68.56 | 71.28 | 77.72 | 76.18 | 73.44 | 87.47 | 80.86 | 84.17 | 83.02 |
| 0 dB | 74.85 | 62.48 | 64.96 | 73.68 | 68.99 | 31.87 | 48.22 | 49.36 | 51.90 | 45.34 | 52.32 | 50.57 | 51.45 | 56.02 |
| -5 dB | 30.46 | 25.12 | 23.17 | 29.56 | 27.08 | -3.78 | 18.65 | 16.70 | 16.69 | 12.07 | 21.31 | 14.96 | 18.14 | 19.28 |
| Average | 93.18 | 89.75 | 90.83 | 92.36 | 91.53 | 76.56 | 81.22 | 82.74 | 81.02 | 80.39 | 87.07 | 84.58 | 85.83 | 85.93 |

Word accuracy of ES 202 050 front-end

| Clean Training (%Acc) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | | | | B | | | | | C | | | Overall |
| | Subway | Babble | Car | Exhibition | Average | Restaurant | Street | Airport | Station | Average | Subway M | Street M | Average | Average |
| Clean | 98.43 | 98.37 | 98.57 | 98.40 | 98.44 | 98.43 | 98.37 | 98.57 | 98.40 | 98.44 | 98.62 | 98.64 | 98.63 | 98.48 |
| 20 dB | 97.61 | 98.43 | 98.69 | 97.72 | 98.11 | 96.04 | 96.43 | 97.94 | 98.03 | 97.11 | 97.64 | 96.67 | 97.16 | 97.52 |
| 15 dB | 94.17 | 96.07 | 97.94 | 95.80 | 96.00 | 92.17 | 94.23 | 96.00 | 96.17 | 94.64 | 93.80 | 93.65 | 93.73 | 95.00 |
| 10 dB | 86.18 | 89.18 | 95.17 | 90.34 | 90.22 | 81.46 | 88.27 | 89.02 | 89.36 | 87.03 | 85.69 | 86.91 | 86.30 | 88.16 |
| 5 dB | 66.53 | 69.20 | 83.21 | 73.19 | 73.03 | 59.90 | 72.46 | 73.22 | 77.78 | 70.84 | 62.57 | 68.83 | 65.70 | 70.69 |
| 0 dB | 36.97 | 31.92 | 48.05 | 37.70 | 38.66 | 23.43 | 44.01 | 40.98 | 49.24 | 39.42 | 32.70 | 40.24 | 36.47 | 38.52 |
| -5 dB | 9.06 | -2.15 | 12.65 | 6.26 | 6.46 | -6.91 | 13.42 | 8.56 | 11.97 | 6.76 | 9.40 | 16.17 | 12.79 | 7.84 |
| Average | 76.29 | 76.96 | 84.61 | 78.95 | 79.20 | 70.60 | 79.08 | 79.43 | 82.12 | 77.81 | 74.48 | 77.26 | 75.87 | 77.98 |

| Multicondition Training (%Acc) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | | | | B | | | | | C | | | Overall |
| | Subway | Babble | Car | Exhibition | Average | Restaurant | Street | Airport | Station | Average | Subway M | Street M | Average | Average |
| Clean | 99.79 | 99.73 | 99.79 | 99.75 | 99.77 | 99.79 | 99.73 | 99.79 | 99.75 | 99.77 | 99.88 | 99.70 | 99.79 | 99.77 |
| 20 dB | 99.63 | 99.58 | 99.76 | 99.38 | 99.59 | 99.32 | 99.43 | 99.31 | 99.38 | 99.36 | 99.60 | 99.58 | 99.59 | 99.50 |
| 15 dB | 99.26 | 99.40 | 99.61 | 99.01 | 99.32 | 98.56 | 98.94 | 98.39 | 98.40 | 98.57 | 99.39 | 99.06 | 99.23 | 99.00 |
| 10 dB | 98.28 | 98.31 | 98.30 | 97.13 | 98.01 | 94.53 | 96.58 | 94.90 | 95.25 | 95.32 | 97.97 | 96.95 | 97.46 | 96.82 |
| 5 dB | 94.14 | 92.14 | 94.72 | 92.01 | 93.25 | 82.16 | 88.85 | 86.22 | 88.77 | 86.50 | 91.80 | 88.63 | 90.22 | 89.94 |
| 0 dB | 78.94 | 68.77 | 80.14 | 76.09 | 75.99 | 52.13 | 69.07 | 67.97 | 71.34 | 65.13 | 70.13 | 63.18 | 66.66 | 69.78 |
| -5 dB | 44.15 | 25.67 | 43.13 | 43.20 | 39.04 | 8.35 | 33.74 | 28.30 | 37.67 | 27.02 | 33.80 | 29.29 | 31.55 | 32.73 |
| Average | 94.05 | 91.64 | 94.51 | 92.72 | 93.23 | 85.34 | 90.57 | 89.36 | 90.63 | 88.98 | 91.78 | 89.48 | 90.63 | 91.01 |

Relative performance of ES 202 050 front-end

| Clean Training (Relative performance) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | | | | B | | | | | C | | | Overall |
| | Subway | Babble | Car | Exhibition | Average | Restaurant | Street | Airport | Station | Average | Subway M | Street M | Average | Average |
| Clean | -460.71% | -288.10% | -694.44% | -300.00% | -435.81% | -460.71% | -288.10% | -694.44% | -300.00% | -435.81% | -666.67% | -312.12% | -489.39% | -446.53% |
| 20 dB | 22.90% | 91.82% | 87.42% | 44.39% | 61.63% | 73.84% | 68.93% | 88.45% | 88.88% | 80.02% | 72.24% | 56.98% | 64.61% | 69.58% |
| 15 dB | 75.43% | 90.90% | 95.08% | 82.92% | 86.08% | 79.87% | 83.33% | 90.52% | 91.49% | 86.30% | 78.77% | 74.20% | 76.48% | 84.25% |
| 10 dB | 73.85% | 82.37% | 92.10% | 83.44% | 82.94% | 68.84% | 79.57% | 81.09% | 82.84% | 78.09% | 74.67% | 75.17% | 74.92% | 79.39% |
| 5 dB | 55.21% | 59.92% | 78.80% | 65.64% | 64.89% | 49.20% | 63.86% | 63.73% | 71.42% | 62.05% | 49.48% | 58.42% | 53.95% | 61.57% |
| 0 dB | 28.15% | 25.87% | 42.03% | 29.23% | 31.32% | 15.03% | 35.08% | 32.41% | 43.70% | 31.55% | 22.00% | 30.83% | 26.42% | 30.43% |
| -5 dB | 1.76% | -6.80% | 5.82% | -1.78% | -0.25% | -8.98% | 5.31% | 3.98% | 6.89% | 1.80% | 0.65% | 8.14% | 4.39% | 1.50% |
| Average | 51.04% | 60.60% | 72.74% | 58.42% | 61.12% | 47.98% | 60.67% | 63.17% | 69.44% | 60.39% | 49.87% | 53.86% | 51.84% | 59.09% |

| Multicondition Training (Relative Performance) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | | | | | B | | | | | C | | | Overall |
| | Subway | Babble | Car | Exhibition | Average | Restaurant | Street | Airport | Station | Average | Subway M | Street M | Average | Average |
| Clean | 0.00% | 25.00% | 36.36% | 0.00% | 15.34% | 0.00% | 25.00% | 36.36% | 0.00% | 15.34% | 61.29% | 33.33% | 47.31% | 21.74% |
| 20 dB | 0.00% | -27.27% | 20.00% | -44.19% | -12.86% | 50.72% | -5.56% | 37.27% | 69.15% | 37.90% | 18.37% | 30.00% | 24.18% | 14.85% |
| 15 dB | 0.00% | 0.00% | 38.10% | 15.38% | 13.37% | 53.55% | 56.20% | 54.65% | 72.84% | 59.31% | 26.51% | 42.33% | 34.42% | 35.95% |
| 10 dB | 1.71% | 34.24% | 17.48% | -9.54% | 10.97% | 58.47% | 67.21% | 41.45% | 68.46% | 58.90% | 34.52% | 51.51% | 43.01% | 36.55% |
| 5 dB | 4.09% | 23.09% | 32.65% | -4.04% | 13.95% | 43.26% | 61.18% | 38.15% | 52.85% | 48.86% | 34.56% | 40.60% | 37.58% | 32.64% |
| 0 dB | 16.26% | 16.76% | 43.32% | 9.16% | 21.38% | 29.74% | 40.27% | 36.75% | 40.42% | 36.79% | 37.35% | 25.51% | 31.43% | 29.55% |
| -5 dB | 19.69% | 0.73% | 25.98% | 19.36% | 16.44% | 11.69% | 18.55% | 13.93% | 25.18% | 17.34% | 15.87% | 16.85% | 16.36% | 16.78% |
| Average | 12.81% | 18.42% | 40.11% | 4.81% | 20.09% | 37.47% | 49.80% | 38.33% | 50.61% | 43.79% | 36.39% | 31.77% | 33.88% | 36.08% |

**Table 4** Summary of evaluation results of ES 202 050 front-end.

| %Acc | | | | |
|---|---|---|---|---|
| | A | B | C | Overall |
| Clean Training | 79.20 | 77.81 | 75.87 | 77.98 |
| Multicondition training | 93.23 | 88.98 | 90.63 | 91.01 |
| Average | 86.22 | 83.39 | 83.25 | **84.49** |

| Relative performance | | | | |
|---|---|---|---|---|
| | A | B | C | Overall |
| Clean Training | 61.12% | 60.39% | 51.84% | 59.09% |
| Multicondition training | 20.09% | 43.79% | 33.88% | 36.08% |
| Average | 40.61% | 52.09% | 42.86% | **47.59%** |

the weighted averages of testsets A, B, and C with weights proportional to the sizes of the testsets. Furthermore, the average *Relative performance* is calculated by using the average *%Acc* and the average *%Acc* of baseline performance; this is not calculated as the average of *Relative performance* in the individual noise condition. Table 3 contains all the *%Acc* for every noise condition, the various average *%Acc* for the baseline and evaluated methods, and their relative performance. The summary shown in Table 4 contains the average *%Acc* values and their relative performance. Also in the summary, the results of clean training and multicondition training are averaged.

In Table 3, we can see that the *%Acc* by ES 202 050 is considerably higher than that of the baseline performance: in clean training, the average *%Acc* under the 15-dB SNR condition was 95%. Furthermore, even under the 10-dB SNR condition, the average *%Acc* was close to 90%. In multicondition training, ES 202 050 also shows good performance, where even under 5-dB SNR condition, the average *%Acc* was close to 90%. From the relative performance, we can also see that the performance of ES 202 050 is considerably good.

### 3.3 Performance Analysis Based on Word Accuracy Per Speaker

Averaged accuracies described in Sect. 3.2 are calculated from the recognition results of a large number of speakers because the recognition performance strongly depends on speaker variability. However, these measures cannot indicate the performance difference among speakers.

Here, we propose a performance analysis method based on word accuracy per speaker. In the method, the following values are calculated, then compared among recognizers and noise-reduction methods.

- The maximum, minimum, average values and standard deviations of the word accuracy per speaker;
- The histogram of the word accuracy per speaker;
- The rate of speakers whose word accuracy exceeds $x$%.

The comparison enables us to analyze how the recognition performance is affected by speaker variability.

Table 5 shows the maximum, minimum, average value and standard deviation of the word accuracy per speaker, where each value was calculated from the overall performance of each individual speaker.

**Table 5** Maximum, minimum, average values and standard deviations of the word accuracy per speaker.

| | | Baseline | ES 202 050 |
|---|---|---|---|
| Clean training | Maximum | 61.24 | 88.04 |
| | Minimum | 26.33 | 62.08 |
| | Average | 45.51 | 77.82 |
| | Std. dev. | 6.06 | 5.14 |
| Multicondition training | Maximum | 93.98 | 96.03 |
| | Minimum | 74.30 | 79.44 |
| | Average | 85.83 | 90.99 |
| | Std. dev. | 3.64 | 3.01 |

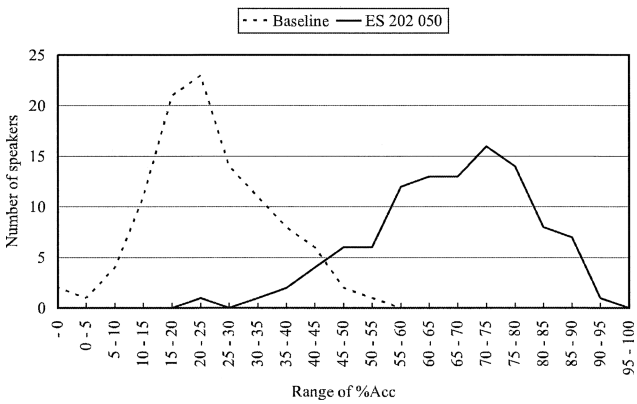



**Fig. 5** Histogram of the word accuracy per speaker.

We can see that the maximum and minimum values of ES 202 050 were considerably higher than that of the baseline front-end. The maximum value was 96.03% for the multicondition training and even 88.04% for the clean training. The standard deviation of ES 202 050 also became slightly smaller than that of the baseline front-end. However, there was still a great difference between the maximum and the minimum value.

Figure 5 shows histograms of the word accuracy per speaker under clean training, where the noise was the subway for testset A and the SNR value was set to 5 dB. In Fig. 5, the horizontal axis is the range of word accuracy and the vertical axis is the number of speakers.

The distribution center of ES 202 050 is clearly was skewed to the right side, while that of the baseline front-end is skewed to the left. However, the tail of the distribution of ES 202 050 spreads to the left side, which means that there were many speakers whose word accuracy was still low.

Finally, Figs. 6 and 7 show the rate of speakers whose word accuracy was more than $x$% for the baseline front-end and ES 202 050, respectively.

We can see that the rate of speakers decreased remarkably when the target word accuracy, $x$, was close to 100%. This rate could be regarded as a measure for judging whether a system is eligible for realistic service. For example, when the target word accuracy and the rate of speakers are set to 90%, ES 202 050 satisfies this condition for the SNR of 20 dB in clean training and 10 dB in multicondition training. This fact confirms that the availability of ES 202

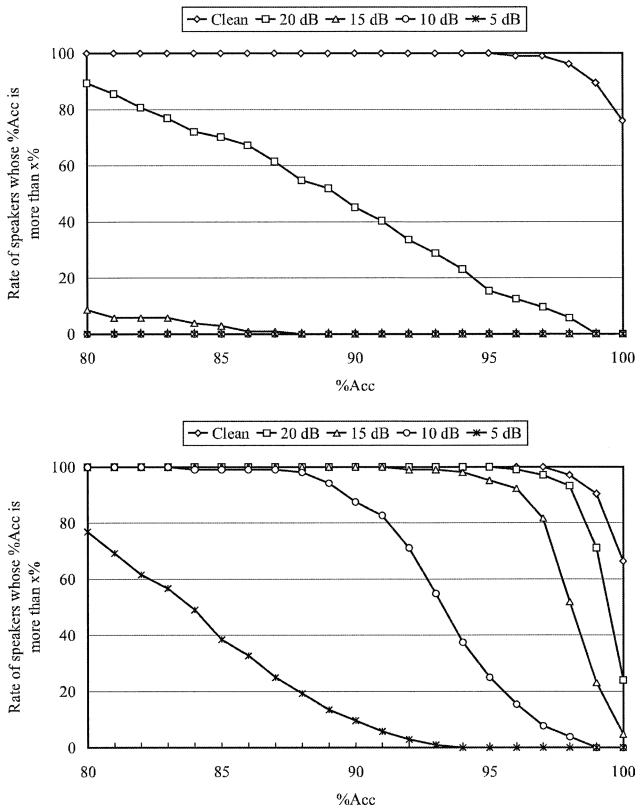



**Fig. 6** Rate of speakers whose word accuracy exceeds $x\%$ for the baseline front-end. The upper is in clean training and the lower is in multicondition training.

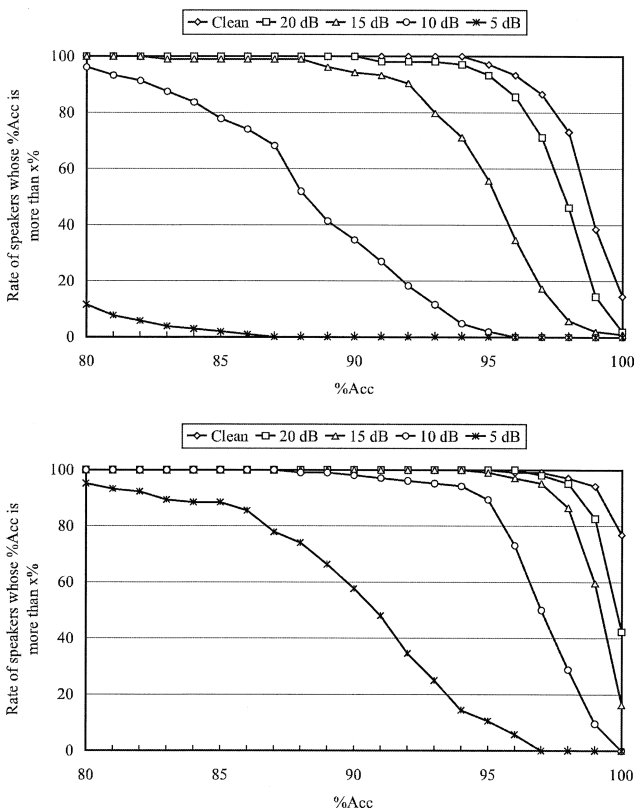



**Fig. 7** Rate of speakers whose word accuracy exceeds $x\%$ for ES 202 050. The upper is in clean training and the lower is in multicondition training.

050 is considerably better than that of the baseline front-end.

## 4. Evaluation Categories

Strictly speaking, the purpose of AURORA project is to develop and evaluate front-ends for recognizers. However, in some papers reported so far, many modifications to the original baseline back-end were introduced, such as using extra data not included in the AURORA database, increasing the number of mixtures, using HMMs that were not whole-word models, and so on. The recognition results using these methods cannot be fairly compared with methods using the original back-end HTK system. Therefore, we propose new evaluation categories in this paper, where one method is compared with other methods only within the same category. These categories were designed to show how much the user's method modified the baseline sysytem from a viewpoint of changes in training method, adaptation process, model topology, decoder, and computational cost in the recognition phase except for the change of front-end process. According to the degree of modification to the back-end system from the original baseline, users declare the category to which they belong from the following categories:

**Category 0.** No changes to the back-end HTK scripts. Changes to only front-end processing, i.e. changes to feature vectors, can be included in this category.

**Category 1.** If the HMM topology is the same as the baseline scripts, any training process will be allowed. Discriminative training can be introduced in this category. The computational cost in the recognition phase should be the same as it was.

**Category 2.** If the HMM topology is the same, adaptation processes can be introduced using some testing data. Speaker or environment adaptation, and PMC with one state noise model can be allowed in this category. An increase in the computational cost will be caused only by the adaptation process.

**Category 3.** Changes in the standard HMM topology. A different number of mixtures and states can be allowed. However, the recognition unit should comprise whole-word models. PMC with more than one state noise model can be included in this category.

**Category 4.** Any process will be allowed as long as the computational cost is under the CPU time that the baseline scripts used. For example, a complex structure model can be used with low-dimensional feature vectors.

**Category 5.** Any process with any computational cost will be allowed.

**Category B.** The use of any training data not included in AURORA is allowed — not only speech data, but also environment noise data. Of course, the evaluation data

is AURORA. This category essentially differs from Categories 1–5.

## 5. Conclusion

In this paper, we introduced AURORA-2J, an evaluation framework for Japanese noisy speech recognition. AURORA-2J consists of a Japanese connected-digits corpus and its evaluation scripts. The data collection, baseline scripts, and its baseline performance were described.

We also proposed a new performance analysis method based on the word accuracy per speaker to consider performance differences among speakers. To compare the performances of various noise-reduction methods, we proposed categorizing modifications of the baseline back-end system. These data and evaluation tools are available to the public. See the AURORA-J Web site[†] to find contact information for obtaining the data. Tools used for the evaluation described in Sect. 3 can also be downloaded from the same Web site.

We plan to develop a series of frameworks for noisy speech recognition, named CENSREC (Corpus and Environments for Noisy Speech RECognition). AURORA-2J is also regarded as a part of the series and has been given an alternative name, CENSREC-1. In the near future, we will develop the frameworks AURORA-2.5J with digit-string utterances spoken by speakers listening to the same noises as AURORA-2J through headphones, which can be seen as noise-free AURORA-2J with Lombard effects, and AURORA-2J with those recorded in a car, to evaluate the systems under more realistic environments. We also plan to develop a series of frameworks with word utterances in parallel with the AURORA-J series, and all the frameworks including the AURORA-J series will be distributed as a series of the CENSREC.

## Acknowledgements

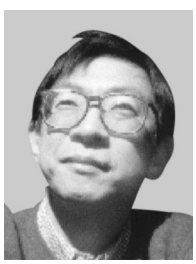
The authors wish to thank Dr. David Pearce of the AURORA group for his help with these activities. This work was supported in part by the Telecommunications Advancement Organization of Japan. The present study was conducted using the AURORA-2J database developed by IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

### References

[1] http://elazar.itd.nrl.navy.mil/spine/
[2] http://eurospeech2001.org/ese/NoiseRobust/index.html, http://www.elda.fr/proj/aurora1.html, http://www.elda.fr/proj/aurora2.html
[3] ETSI standard document, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm," ETSI ES 201 108 v1.1.2 (2000-04), 2000.
[4] R.G. Leonard, "A database for speaker independent digit recognition," Proc. ICASSP-84, vol.3, pp.328–331, 1984.
[5] H.G. Hirsh and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," ISCA ITRW ASR2000, pp.181–188, Sept. 2000.
[6] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & What next," ASRU 2001, pp.131–134, 2001.
[7] S. Young, D. Kershow, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book, 2000.
[8] J. Droppo, L. Deng, and A. Acero, "Evaluation of SPLICE on the AURORA 2 and 3 tasks," ICSLP 2002, pp.29–32, 2002.
[9] ETSI standard document, "Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," ETSI ES 202 050 V1.1.1, 2002.

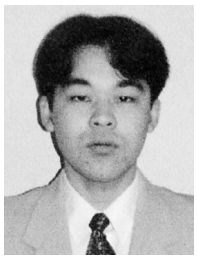
[†]http://sp.shinshu-u.ac.jp/AURORA-J/

**Satoshi Nakamura** received his B.S. degree in electronic engineering from Kyoto Institute of Technology in 1981 and a Ph.D. degree in information science from Kyoto University in 1992. Between 1981–1993, he worked with the Central Research Laboratory, Sharp Corporation, Nara, Japan. From 1986–1989, he worked with ATR Interpreting Telephony Research Laboratories and from 1994–2000, he was an associate professor of the Graduate School of Information science, Nara Institute of Science and Technology, Japan. In 1996, he was a visiting research professor of the CAIP center of Rutgers University of New Jersey USA. He is currently the head of Acoustics and Speech Research Department at ATR Spoken Language Translation Laboratories, Japan. He also serves as an honorary professor at the University of Karlsruhe, Germany since 2004. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He received the Awaya award from the Acoustical Society of Japan in 1992, and the Interaction2001 best paper award from the Information Processing Society of Japan in 2001. He served as an associate editor for the Journal of the IEICE Information from 2000–2002. He is currently a member of the Speech Technical Committee of the IEEE Signal Processing Society. He is a member of the Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ), and IEEE.

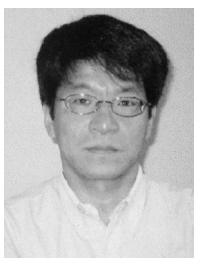


**Kazuya Takeda** received the B.S. degree, the M.S. degree, and the Dr. of Engineering degree from Nagoya University, in 1983, 1985, and 1994 respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated in a project for constructing voice-activated telephone extension system. He has joind Graduate School of Nagoya University in 1995. Since 2003, he is a professor at Graduate School of Infomation Science at Nagoya University. He is a member of the IEEE and the ASJ.
E-mail: takeda@is.nagoya-u.ac.jp

**Kazumasa Yamamoto** received his B.E., M.E. and Dr. Eng. degrees in information and computer sciences from Toyohashi University of Technology, Toyohashi, in 1995, 1997 and 2000, respectively. Since 2000 he has been a research associate in the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University, Nagano. He has been engaged in research on speech recognition. He is a member of the ASJ.
E-mail: kyama@sp.shinshu-u.ac.jp



**Takeshi Yamada** received his B.E. degree from Osaka City University in 1994, and M.E. and Dr. Eng. degrees from Nara Institute of Science and Technology in 1996 and 1999, respectively. Since 1999, he has been on the faculty of the University of Tsukuba, where he is currently a lecturer in Graduate School of Systems and Information Engineering. His research interests include robust speech recognition, sound scene recognition, microphone array signal processing and sound field control and reproduction. He is a member of the IEEE, the IPSJ and the ASJ.
E-mail: takeshi@cs.tsukuba.ac.jp



**Shingo Kuroiwa** received his B.E., M.E. and Dr. Eng. degrees in electro-communications from the University of Electro Communications, Tokyo, Japan, in 1986, 1988, and 2000, respectively. From 1988 to 2001 he was a researcher at the KDD R & D Laboratories. Since 2001, he has been with the Faculty of Engineering, Tokushima University, Tokushima, Japan, where he is currently an Associate Professor. His current research interests include speech recognition, speaker recognition, natural language processing, and information retrieval. He is a member of the IPSJ, the ASJ.
E-mail: kuroiwa@is.tokushima-u.ac.jp

**Norihide Kitaoka** received his B.E. and M.E. degrees from Kyoto University in 1992 and 1994, respectively, and a Dr. Eng. degree from Toyohashi University of Technology in 2000. He joined DENSO CORPORATION, Japan in 1994. He then joined the Department of Information and Computer Sciences at Toyohashi University of Technology as a research associate in 2001 and has been a lecturer since 2003. His research interests include speech processing, speech recognition and spoken dialog. He is a member of the IPSJ, the ASJ and the Japan Society for Artificial Intelligence (JSAI). E-mail: kitaoka@slp.ics.tut.ac.jp

**Takanobu Nishiura** received his B.E. degree from Nara National College of Technology in 1997, and M.E. and Ph.D. degrees from Nara Institute of Science and Technology (NAIST) in 1999 and 2001, respectively. He received the TELECOM System Technology Award for Student from The Telecommunications Advancement Foundation (TAF) in 2000. From 2001 to 2004, he was a research associate of Wakayama University. He is currently an associate professor of Ritsumeikan University, and a visiting researcher at ATR Spoken Language Translation Research Laboratories. His current research interests include acoustic sound signal sensor using a microphone array. He is a member of the IEEE and the ASJ.
E-mail: nishiura@is.ritsumei.ac.jp

**Akira Sasou** received his B.E., M.E. and Dr. Eng. degrees in electronic engineering from Tokyo Denki University, Tokyo, Japan, in 1994, 1996 and 1999, respectively. In 2000, he joined the Electrotechnical Lab., which was reorganized to the National Institute of Advanced Industrial Science and Technology (AIST) in 2001. He is a member of ASJ.
E-mail: a-sasou@aist.go.jp

**Mitsunori Mizumachi** received his Bachelor's degree from Kyushu Institute of Design in 1995, and his Master's degree and Ph.D. from Japan Advanced Institute of Science and Technology in 1997 and 2000, respectively. In 2000, he joined ATR Spoken Language Translation Research Laboratories. He is currently a research associate of Kyushu Institute of Technology. His research interests include spatial signal processing and the behavior of auditory system. He is a member of the ASJ, ASA, and IEEE.
E-mail: mitsunori.mizumachi@atr.jp

**Chiyomi Miyajima** received her B.E. degree in computer science and M.E. and Dr. Eng. degrees in electrical and computer engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1996, 1998, and 2001, respectively. From 2001 to 2003, she was a Research Associate of the Department of Computer Science, Nagoya Institute of Technology. Currently she is a Research Associate of the Graduate School of Information Science, Nagoya University. Her research interests include automatic speaker recognition and multi-modal speech processing. She is a member of the ASJ and the Japanese Association of Sign Linguistics (JASL).
E-mail: miyajima@is.nagoya-u.ac.jp

**Masakiyo Fujimoto** graduated from the Department of Electronics and Informatics, Ryukoku University in 1997, received his master's degree in 2001 and is a research student in the doctoral program. He is also an intern researcher at ATR Spoken Language Translation Research Laboratories. He received the Awaya Award from ASJ in 2003. He is engaged in research on noise-robust speech recognition. He is a member of ASJ and ISCA.
E-mail: masakiyo.fujimoto@atr.jp

**Toshiki Endo** received B.E. and M.E. degrees in Electrical Engineering from Keio University, Kanagawa, Japan, in 1996 and 1998, respectively. In 1998 he joined Kokusai Denshin Denwa Co., Ltd., (currently KDDI) where he has been engaged in research on Telecommunication services and transmission technology for voice IP packets. He joined ATR Spoken Language Translation Research Laboratories in 2002. His main research interests include robust speech recognition against noise and data loss. He is a member of the ASJ. E-mail: toshiki.endo@atr.jp