CIAIR speech corpus for real world speech recognition

Nobuo Kawaguchi, Kazuya Takeda, Shigeki Matsubara, Ikuya Yokoo, Taisuke Ito
Kiyoshi Tatara, Tetsuya Shinde and Fumitada Itakura
Center for Integrated Acoustic Information Research
Nagoya University, Nagoya 464-8603, Japan.
Tel. +81-52-789-3629, Fax. +81-52-789-3172
takeda@nuee.nagoya-u.ac.jp

Abstract

This paper describes the speech corpora designed for research on development of real world speech applications. The corpora built at Center for Integrated Acoustic Information Research (CIAIR), Nagoya university, include a multimedia database for the in-car speech, a corpus for lavalier microphone speech in real world environments, and a corpus for whispered speech. This paper also presents studies on recognition of speech in each corpus using HMM based acoustic models.

1 INTRODUCTION

Evolution of computer networking and internet technologies to provide multiple, ubiquitous and flexible connections has enabled access to infinite information resources around the world. Advanced technologies for indexing and searching the distributed information can be used to obtain a quick response for complex queries. However, the existing technology for interfaces to the information access systems is still a barrier thus preventing their use by the ordinary people independent of the background of users. Development of interfaces based on speech recognition in real world environments has an important role in providing the ubiquitous access to information.

Advances in technology for large vocabulary continuous speech recognition have led to development of systems that can be used in office like environments. However, these systems have limitations in developing speech interfaces that can be used in real life situations such as driving in a car or walking on a street. Construction of large corpora of speech in real world environments is important for development of systems capable of recognition of noisy speech in these environments. Large corpora of speech and acoustic data are being built at Center for Integrated Acoustic Information Research (CIAIR). In this paper, we describe these corpora and present the studies on recognition of speech in real world environments.

2 IN-CAR SPEECH CORPUS

Research on development of speech interfaces that can be used while driving a car is important for the following reasons: (1) It is difficult to use a keyboard or a touch panel, (2) Communication in a moving environment is essential for ubiquitous access to information, and (3) The in-car speech is contaminated with multiple sources of noise. It is expected that the technology developed for recognition of the in-car speech can be used for other real world environments as well. Therefore, our focus is on collection of the in-car speech data for spoken dialogues while driving a car.

Table 1: Specifications of recording devices.

Type of Data	Specifications
Sound Input	16ch,16bit,16kHz
Sound Output	16ch,16bit,16kHz
Video Input	3ch, MPEG1
Control Signal	Status of Accelerator and Brake, Angle of Steering wheel Engine RPM, Speed : 16bit,1kHz
Location	D-GPS

2.1 Data collection vehicle

The DCV is a car specially designed for the collection of multimedia data. The vehicle is equipped with eight network-connected personal computers (PCs). Three PCs have a 16-channel analog-to-digital and digital-to-analog conversion port that can be used for recording and playing back data. The data can be digitized using 16-bit resolution and sampling frequencies up to 48 kHz. One of these three

Figure 1: Visual signal captured by the three cameras. (a) the driver's face.(left upper), (b) the driver, the operator and the back view (right upper) and (c) front view (right bottom).

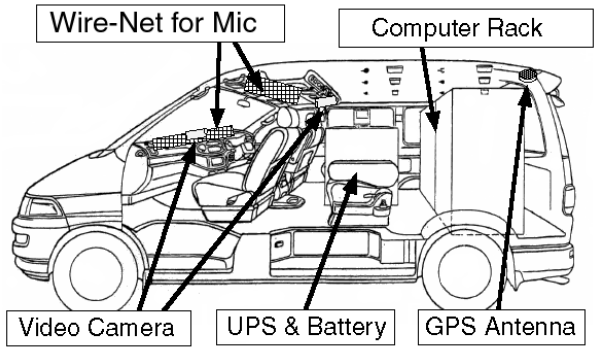

Figure 2: Configuration of DCV

PCs can be used for recording audio signals from 16 microphones. The second PC can be used for audio playback on 16 loudspeakers. The third PC is used for recording five signals associated with the vehicle: the angle of the steering wheel, the status of the accelerator and brake pedals, the speed of the car and the engine speed. These vehicle-related data are recorded at a sampling frequency of 1 kHz in 2-byte resolution. In addition, location information obtained from the Global Positioning System (GPS) is also recorded at the sampling frequency of 1 Hz.

Three other PCs are used for recording video images (Figure 1). The first camera captures the face of the driver . The second camera captures the conversation between the driver and the experiment navigator. The third camera captures the view through the windshield. These images are coded into MPEG1 format. The remaining two PCs are used for controlling the experiment. The multimedia data on all systems are recorded synchronously. The total amount of data is about 2 GB for about a 60-minute drive during which three dialogue sessions are recorded. The recorded data is directly stored on the hard disks of the PCs in the car.

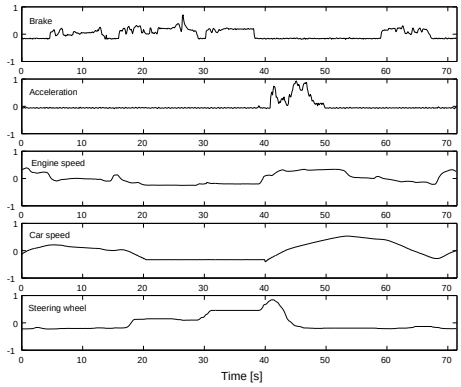

Figure 3: Vehicle-related signals. Brake and accelerator pedals, rotational speed of the engine motor and speed (from top to bottom).

Figure 2 shows the arrangement of equipment in the DCV, including the PCs, a power generator with batteries, video controller, microphone amplifiers and speaker amplifiers. An alternator and a battery are installed for stabilizing the power supply. Wire nets are attached to the ceiling of the car so that the microphones can be arranged in arbitrary positions.

Table 2: Speech materials recorded in the experiment.

item	approx. time
prompted dialogue	5 min
natural dialogue	5 min
dialogue with system	5min
dialogue with WOZ	5 min
P.B. sentence (driving)	10 min
P.B. sentence (idling)	5 min

2.2 Speech materials

The collected speech materials are listed in Table 2. The task domain of the dialogues is the restaurant guidance around the Nagoya University campus. In dialogues with a human

operator and the wizard-of-OZ (WOZ) system, we have prompted the driver to issue natural and varied utterances related to the task domain, by displaying a 'prompt panel'. On the panel, a keyword, such as *fast food*, *bank*, *Japanese food*, or *parking*, or a situation sentence, such as 'Today is an anniversary. Let's have a party.', 'I am so hungry. I need to eat!' or 'I am thirsty. I want a drink!', are displayed. In these modes, therefore, the driver takes the initiative in the dialogue. The operator also navigates the driver to a predetermined destination while they are having a dialogue, in order to simulate the common function of a car-navigation system. All responses of the operator are given by synthetic speech in the WOZ mode. In addition, fully natural dialogues are also conducted between the driver and a distant operator via cellular phone. In such natural dialogues, the driver asks for the telephone number of a shop from the yellow pages information service. These natural dialogues are collected both when idling the engine and while driving the car.

All utterances have been phonetically transcribed and tagged with time codes. Tagging is performed separately for utterances by the driver and by the operator so that timing analysis of the utterances can be carried out. On average, there are 380 utterances and 2768 morphemes in the data for a driver.

In addition to the dialogues, speech of the text read aloud and isolated word utterances have also been collected. Each subject read 100 phonetically balanced sentences while idling the engine and 25 sentences while driving the car. A speech prompter is used to present the text while driving. The speech data of the read text is mainly used for training acoustic models. The set of isolated word utterances consists of digit strings and car control words.

2.3 Data collection using an ASR system

Since dialogue between man and machine is one of our final goals, we are collecting man-machine dialogues using a prototype spoken dialogue system that has speech recognition capabilities. The task domain of the prototype system is restaurant information. Drivers can retrieve information and make a reservation at a restaurant near the campus by conversing with the system. The automatic speech recognition module of the system is based on a common dictation software platform known as Julius 3.1 (Kawahara et al., 1999).

A trigram language model with a 1500-word vocabulary is trained using about 10000 sentences. The main body of the training sentences is extracted from the human-human dialogue collected in the early stage of the experiment. The other sentences are generated from a finite state grammar that accepts permissible utterances in the task domain. State clustered triphone hidden Markov models consisting of 3000 states are used as acoustic models. The number of mixtures for each state is 16. The models are trained using 40,000 phonetically balanced sentences uttered by 200 speakers recorded in a soundproof room with a close-talking microphone (Itou et al., 1999). The same microphone as in this recording is used for speech input in the prototype dialogue system. A preliminary evaluation of the speech recognition module of the system has given a word accuracy of about 70% under real driving conditions.

The dialogue is controlled by transitions among 12 states, each of which corresponds to the database query results. When a set of particular conditions defined for a transition is satisfied, the predefined state transition occurs, invoking associated actions, i.e., generating speech responses. Up to today, 75% of the man-machine dialogues have been correctly completed by the system.

2.4 In-car speech recognition through multiple regression

Based on the constructed corpus, a new multichannel method for noisy speech recognition is proposed based on the multiple regression of the log spectra (MRLS). The basic idea of the proposed method is to approximate the log power spectrum of the close-talking microphone speech by a linear combination of the log power spectra of the distant microphones. The approximation is given by the following procedure.

Suppose that $X_0(k)$ is the spectrum of the speech obtained by the close-talking microphone at the k^{th} spectral channel, and $X_i(k)$, $i=1,\ldots, N$, are the spectra of the speech obtained by the distant microphones located at N different positions. The spectral regression

is given by

$$\log |X_0(k)| = \sum_{i=1}^{N} \bar{w}_i(k) \log |X_i(k)|, \qquad (1)$$

where $\bar{w}_i(k)$ are the real numbers that give the minimum regression error, i.e.,

$$\bar{w}_i(k) = \arg\min_{w_i(k)} E\left[d^2\right], \qquad (2)$$

where

$$d^2 = \sum_{k=1}^{K} \left\{ \log |X_0(k)| - \sum_{i=1}^{N} w_i(k) \log |X_i(k)| \right\}^2. \qquad (3)$$

Here, the expectation, $E[]$, is calculated over the training utterances.

Note that the regression error $\min E\left[d^2\right]$ is equal to the cepstral distance between the approximated and the target spectrum because of the orthogonality of the discrete time cosine transform (DCT) matrix. Therefore, the method can be considered as an extension of *feature average* in the cepstrum domain by replacing the average value with the weighted sum. Furthermore, applying the regression analysis in the log spectrum domain has the following two merits: (1) the spectrum flooring due to the oversubtraction can be avoided, and (2) the target spectrum for a wider range of in

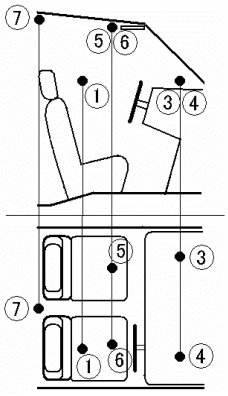

Figure 4: Microphone arrangement for data collection.

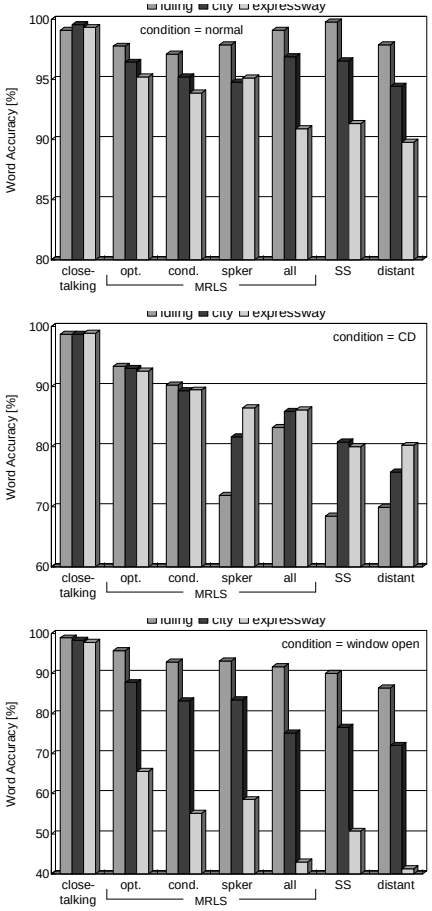

Figure 5: Recognition results under various driving conditions: normal, CD on and window open (from top to bottom.)

In order to evaluate the effectiveness of the MRLS method, recognition experiments have been performed. Throughout this section, the same structure is used for the set of triphone HMMs, i.e., 1) they share 1000 states; 2) each state has 16 component mixture Gaussian distributions; and 3) the feature vector is a 25 (12 MFCC + 12 Δ MFCC + Δ logpower) dimensional vector.

For comparison, the following three different sets of HMM are trained: 1) close-talking model is trained using the close-talking microphone speech, 2) distant microphone model is trained using the speech at the nearest microphone (mic. 6 in Figure 4), and 3) MRLS model is trained using the spectra obtained from the MRLS method. For training the MRLS model, the regression weights are optimally determined for each training sentence. The total

number of training sentences is about 8,000; 2,000 of them are uttered while driving and 6,000 of them are uttered in the idling car.

The test data includes a isolated word utterances of a 50 word set. Each of 18 speakers uttered the word set under 18 different car conditions. For each utterance, six different versions of the speech data are recognized. They are 1) speech recorded using the close-talking microphone, 2) speech recorded at the nearest microphone, 3) MRLS output with the optimally determined weights for each utterance, 4) MRLS output with the optimally determined weights for each speaker, 5) MRLS output with the optimally determined weights for each driving condition, and 6) MRLS output with the optimally decided weights for all of the training data.

Note that case 3) is an unrealistic case in that the close-talking speech itself can be used for recognition. The results for this case indicate the upper bound of the MRLS. On the other hand, cases 4) and 5) assume that the optimal weights for MRLS are constant for each speaker or driving condition.

2.5 Results of MRLS

For the evaluation, six recognition experiments are performed: 1) recognize close-talking speech by the close-talking model (close-talking), 2) recognize nearest microphone speech by the distant microphone model (distant), 3) recognize optimal MRLS output by the MRLS model (MRLS opt.), 4) recognize MRLS output optimized for each speaker by the MRLS model (MRLS spker), 5) recognize MRLS output optimized for each driving condition using the MRLS model (MRLS cond.), and 6) recognize MRLS output optimized for all training data using the MRLS model (MRLS all). In addition, the results of spectrum subtraction (SS) are also compared where the training and the test speech at the nearest distant microphone are enhanced by the spectrum subtraction.

The results are shown in Figure 5 for each car condition. It is found that MRLS outperforms the nearest distant microphone result even in the MRLS:all case. This result suggests the robustness of the method to the change of the location of the noise sources, because the primary noise locations are different between 'open window' and 'cd' cases. It is also found that the improvement is larger when the performance of the distant microphone is lower. Furthermore, by optimizing the regression weights for each speaker or driving condition, recognition accuracy can be further improved, but the performance is still not as high as the result of the upper bound.

3 LAVALIER MICROPHONE SPEECH CORPUS

For the realworld speech applications, *hands-free* is an important issue for making use of the merit of speech interface, i.e., remote input. The speech captured at the distant microphone, however, is distorted by the addition of the background or interfering noise and/or convolution of the acoustic channel; and is more difficult to recognize.

Using lavailer microphone is a compromising between the distant microphone and the close-talking microphone because lavalier microphones can be attached to any part of the body. They are also lightweight and the SNR difference between them and close-talking microphones is not significant if they are placed near the mouth area. Therefore, a large corpus of lavailer microphone speech is collected and the recognition experiments using the lavailer microphone speech are performed.

3.1 Recording Environments

Speech was recorded by using Sony ECM77B lavalier microphones. This is an ultra mini omnidirectional electret condenser lavalier suitable for many different applications, ranging from recording of news and interviews to recording in theaters and for instrument pick-up. Its frequency response is 40-20 kHz with upper range lift for extra presence. Directivity is optimized to ensure uniform output, regardless of direction of the sound source. The metal mesh windshield effectively eliminates both outdoors wind noise and "popping" in close microphone situations. It is 5.6mm in diameter and 12.55mm in length.

Each subject was equipped with two lavalier microphones. One was attached to the frame of the provided spectacles and the other was attached to the subject's shirt around the chest area. The recording scene is as shown in Figure 6. Input speech was quantized to 16 bits, and sampled at 48 kHz.

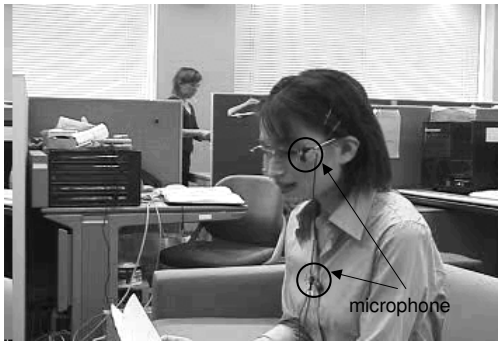

Figure 6: Recording scene - A lavalier microphone is placed on the subjects shirt around the chest area and the other is on the spectacles frame.

This database was constructed to carry out recognition experiments for real world applications. Speech was recorded in four different environments which included recording in a sound-proof room, in an office space, on a street, and inside a car. The noise level of the sound-proof room was about 22 dB(A), and the reverberation time was approximately 150 ms. Three types of cars were used: a Sedan, a station wagon, and a one-box type car. A driver was instructed to drive these cars in the Nagoya city suburbs. The subjects sat in the passenger seat and were instructed to read a list of phonetically balanced sentences. The traffic on the street was relatively heavy.

3.2 Sentence Composition

For constructing this speech corpus, we used 10 sets of ATR phonetically balanced Japanese sentences (Sagisaka et al., 1990)), consisting of a total of 503 sentences. The Acoustical Society of Japan (ASJ) continuous speech corpus (Japanese Newspaper Article Sentences:JNAS) (Itou et al., 1999), consisting of 100 sentences in total was also used.

Each subject read one set from the ATR phonetically balanced sentences and 5 sentences from 100 JNAS sentences. Phonetically balanced sentences were used for to build the training data, and the JNAS sentences were used for the test data. Based on this recording method, each speaker read 60 sentences in each environment. In total, 53 speakers (26 males and 27 females) participated in building this database.

Table 3: Recognition Model Parameters

Sampling rate	48 kHz
Window	Hamming window
Frame length	25 ms
Frame shift	10 ms
Feature Vector	12 MFCC + 12 Δ MFCC + Δ-log-POWER
HMM	32-mixture triphone
Number of states	500
Training data	2150 sentences

3.3 Recognition Experiments

HMM based continuous speech recognition experiments were performed using the lavalier microphone speech corpus with a vocabulary size of 20,000. Four HMMs were trained for each environment. For training each model 2150 sentences were used from 43 speakers (21 males and 22 females). The other remaining conditions are summarized in Table 3. The test data were 50 JNAS sentences spoken by 10 speakers (5 males and females each) different from those who recorded the training data. The decoder used was julius 3.1 (Kawahara et al., 1999). Recognition performance is evaluated by the word correct rate. For comparison, we considered the IPA[1] standard Japanese HMM. This model has 2,000 states, with 16 mixtures, and is of gender independent triphone type.

3.4 Recognition Results

The results obtained from the HMM continuous speech recognition tests are shown in Figures 7 and 8. In these figures, 'IPA' refers to the IPA standard Japanese HMM. 'Train' refers to the HMM trained using 2,150 sentences in the same environment as sentences in the test set. 'Train ALL' refers to the HMM trained by 8,600 sentences in all environment conditions. Compared with the results of IPA standard HMM, it is evident that the recognition rates using environment-dependent HMMs are better than those using IPA HMMs in all environments except for the sound-proof room. Especially, at such noisy environments as the street and inside the car, the recognition rates are over 10% better than those using IPA and only 2,150 sentences were used. The recognition rate of speech utterances recorded in the sound-proof

[1]Information-technology Promotion Agency, Japan

room is about 6% higher if the IPA standard HMM is used rather than the HMM trained by lavalier microphone speech utterances. This is because 1) In the sound-proof room, there is little difference between lavalier microphone speech and close-talking microphone speech; and 2) IPA standard HMM is trained with about 20 times more training data than the HMMs trained here.

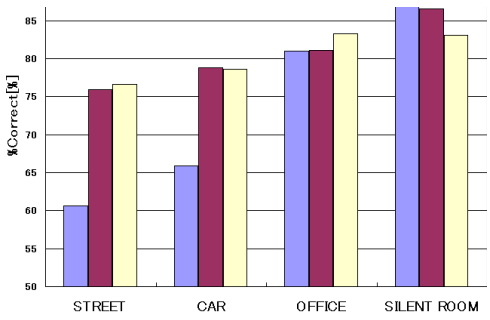

Figure 7: Recognition results for speech in different environment recorded by a lavalier microphone attached to the spectacles.

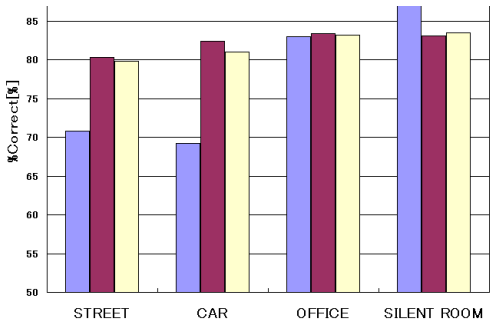

Figure 8: Recognition results for speech different environments recorded by a lavalier microphone attached to the shirt around the chest area.

4 WHISPERED SPEECH CORPUS

Whispered speech is produced by speaking without vibration of the vocal cords. Since exhalation is the source of sound in whispered speech, its acoustic characteristics differ from those of normal speech. In particular, the magnitude (power) in the low-frequency region of whispered speech is weaker than that in normal speech. Therefore, the signal-to-noise ratio (SNR) of whispered speech in a real environment where the background noise is present is low. Accordingly, whispered speech recognition is considered to be more difficult.

We have built a speech corpus consisting of whispered speech and normal speech of more than 6,000 sentences from 123 speakers. One hundred and twenty three speakers (68 males and 55 females) participated in speech recording. They produced both normal speech and whispered speech. Each speaker read one set (50 sentences) from sets A to I in ATR phonetically balanced Japanese sentences. For the test data, 50 sentences from newspaper articles were used.

4.1 Recording Method

Whispered and normal speech utterances were recorded in the same sound-proof room as the lavalier microphone corpus, using a DV camera and a close-talking microphone (Sennheiser HMD410). The sampling rate used was 48 kHz

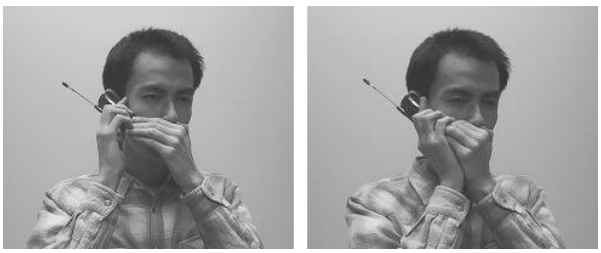

Figure 9: Whispered speech recording through cellular-phone with covering the mouth (left) and with covering both the mouth and the receiver (right).

4.2 Whispered Speech Recognition

In this section, we conduct recognition experiments using the close-talking microphone whispered speech corpus described above. The recognition model is again the hidden Markov Model (HMM).

HMMs trained by normal speech and HMM trained by whispered speech were built. For

training the whispered speech model, 4,000 sentences were used from 80 speakers (40 males and 40 females). Other remaining conditions are summarized in Table4.

Table 4: Recognition Model Parameters

Sampling rate	16 kHz
Window	Hamming window
Frame length	25 ms
Frame shift	10 ms
Feature Vector	12 MFCC + 12 Δ MFCC + Δ POWER
HMM	16-mixture monophone
Number of states	129

For comparison with the whispered HMM, a normal speech HMM was also trained using 14,000 phonetically balanced Japanese sentences of 276 speakers (138 males and 138 females) from JNAS speech database (Itou et al., 1999). This HMM has 129 states, with 16 mixtures, and is of gender independent monophone type.

Using the whispered and normal speech HMMs, continuous speech recognition experiments were performed with a vocabulary size of 20,000. The test data comprised of 200 whispered and normal speech JNAS sentences spoken by 4 speakers (2 males and females each). The decoder used was julius-3.1.

4.3 Results

The recognition rates of normal speech and whispered speech using the normal speech model and whispered speech model is shown in Table 5. The benchmark result for this experiment was the 87% recognition rate obtained for normal speech using the normal speech model. However, a 74% recognition rate was obtained for whispered speech using the whispered speech model.

We also used the normal speech model for recognizing the whispered speech. This gave a 27% recognition rate which was significantly lower than those in the cases where the same speech style was used in the acoustic model and the evaluation sentences. Also this reduction in recognition rate was found to be larger than the case where the whispered speech model was used for recognizing normal speech.

5 SUMMARY

In this paper, we presented the speech corpora collected at center for integrated acoustic information research (CIAIR) and the results of the recognition experiments using the corpora. These corpora are available for various research purposes through our WEB site `http://db.ciair.coe.nagoya-u.ac.jp/`.

Table 5: Recognition rates of normal speech and whispered speech

	Test Speech	
Models	Normal	Whisper
Normal Speech Model	87%	27%
Whisper Speech Model	62%	74%

References

J.C. Junqua and J.P. Haton: Robustness in automatic speech recognition. Kluwer Academic Publishers, 1996.

Petra Geutner, Luis Arevalo and Joerg Breuninger: VODIS - Voice-operated driver information systems: a usability study on advanced speech technologies for car environments. Proc. of International Conference on Spoken Language Processing (ICSLP2000, Beijin), pp.IV378–IV381 (2000).

Asuncion Moreno, Borge Lindberg, Christoph Draxler, Gael Richard, Khalid Choukri, Jeff Allen, Stephan Eule: SpeechDat-Car: A Large Speech Database for Automotive Environments. Pro. of 2nd Int'l Conference on Language Resources and Evaluation (LREC 2000, Athens)

Nobuo Kawaguchi, Shigeki Matsubara, Hiroyuki Iwa, Shoji Kajita, Kazuya Takeda, Fumitada Itakura and Yasuyoshi Inagaki: Construction of Speech Corpus in Moving Car Environment, Proc. of International Conference on Spoken Language Processing(ICSLP2000, Beijin), pp.362–365 (2000).

Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano and Shuichi Itahashi: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, J. Acoust. Soc. Jpn.(E),Vol. 20, No. 3, pp.199–206 (1999).

Y. Sagisaka, M. Abe, K. Takeda, S. Katagiri, T. Umeda, H. Kuwabara, "A Large-Scale Japanese Speech Database," Proc. of International Conference on Spoken Language Processing (ICSLP'90, Nov. 1990, Kobe) Vol.2, pp.1089-1092.

Tatsuya Kawahara, Tetsunori Kobayashi, Kazuya Takeda, et al. 1999. *Japanese Dictation Toolkit: Plug-and-play Framework For Speech Recognition R&D* Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'99), pp.393–396 (1999)