Computational Linguistics and Chinese Language Processing
Vol. 10, No. 2, June 2005, pp. 237-250

TAICAR - The Collection and Annotation of an In-Car Speech Database Created in Taiwan

Hsien-Chang Wang*, Chung-Hsien Yang+, Jhing-Fa Wang+,

Chung-Hsien Wu and Jen-Tzung Chien****

Abstract

This paper describes a project that aims to create a Mandarin speech database for the automobile setting (TAICAR). A group of researchers from several universities and research institutes in Taiwan have participated in the project. The goal is to generate a corpus for the development and testing of various speech-processing techniques. There are six recording sites in this project. Various words, sentences, and spontaneously queries uttered in the vehicular navigation setting have been collected in this project. A preliminary corpus of utterances from 192 speakers was created from utterances generated in different vehicles. The database contains more than 163,000 files, occupying 16.8 gigabytes of disk space.

Keywords: TAICAR, in-car speech, speech database, multi-channel recording, corpus collection and annotation

1. Introduction

1.1 In-car speech corpora review

Driver information systems are becoming increasingly complex as more and more functions are integrated into modern cars. Speech-enabled functions will enhance the safety and convenience of operating for future vehicles. To realize such functions, in-car speech processing techniques need to be built and tested first. Thus, it is necessary to collect an in-car speech database. Although many speech corpora [Tapisa *et al.* 1994], [Roach *et al.* 1996], [Kudo *et al.* 1994], [Bernstein *et al.* 1994] have been created to improve speech-processing effectiveness, few in-car speech databases have been reported.

* Department of Information Management, Chang Jung Christian University,
396 Chang Jung Road, Sec.1, Kway Jen, Tainan, Taiwan, R.O.C.
Tel: 886-6-2785123 ext. 2071; Fax: 886-6-2785657
E-Mail: wangbb@mail.cju.edu.tw
+ Department of Electrical Engineering, National Cheng Kung University
** Department of Computer Science and Information Engineering, National Cheng Kung University

[Received November 30, 2004; Revised April 19, 2005; Accepted April 20, 2005]

Researches on speech processing in the vehicular environment, including works on speech recognition, noise reduction and speaker adaptation, have been published at numerous conferences, for example, the *International Workshop on Hand-Free Speech Communication*, which was held in 2001 in Kyoto, Japan; the biannual *European Conference on Speech Communication and Technology (EuroSpeech)*; and the *International Conference on Spoken Language Processing (ICSLP)*. To our knowledge, several research organizations have carried out in-car speech database collection. In Japan, professor Itakura at CIAIR collected multimedia data, such as audio, video, and auxiliary vehicle information, from dialogues spoken in moving cars [Itakura 2001]. The system was built in a Data Collection Vehicle (DCV) supporting the synchronous recording of multi-channel audio and video data through microphones and cameras. In Europe, researchers in countries such as France, Germany, Britain, and Spain joined in a cooperative project, SpeechDat [Heuvel *et al.* 1999] to collect an in-car speech database for multi-lingual speech processing purposes. The resulting SpeechDat-Car database contains speech data recorded from three microphones and one cellular phone. A similar project has also been reported by Langmann and his colleagues. [Langmann *et al.* 1998]. Researchers at the University of Illinois in Champaign-Urbana designed a project whose purpose was to collect multi-channel database consisting of both speech and video data. One hundred speakers participated in the project, and a total of 59,000 utterances were collected [Lee *et al.* 2004]. Table 1 shows a brief comparison of some existing in-car speech corpora and the TAICAR corpus.

Table 1. Survey of several in-car speech corpora

Corpus name (year)	CSDC-MoTiV (1998)	SpeechDat-Car (1999)	CU-Move (2000)	CIAIR-HCC (2001)	CMU (2001)	AVICAR (2004)	TAICAR (2004)
Country	Germany	Europe	USA	Japan	USA	USA	Taiwan
# of People	641	N/A	N/A	ongoing	43	100	192
Microphone	Array	Array	Array	Mesh	Array	Array	Array
Content	Digits; Commands	Multi-lingual	Digits; Commands	Digits; Words; Sentences	Short words	Digits; Letters; Sentences	Digits; Words; FAQs
Need Specific Car	No	No	No	Yes	No	No	No

1.2 Motivation and Setup

A group of researchers in the field of speech processing in Taiwan initiated an in-car speech collection project called TAICAR (Taiwan in-CAR speech database). The goal is to generate an in-car speech database to be applied to various noisy speech processing researches. In order to generate the corpus rapidly and usefully, some considerations with regard to setting up the data collection procedure were deemed important. These considerations are described below.

1.3 Setup of the TaiCar project

The philosophy behind the TaiCar corpus collection procedure is to use convenient and readily available equipment to collect speech and environmental noise in various vehicles. The following are the ten considerations deemed important.

1. The platform for in-car speech collection should be a notebook PC.
2. The resulting speech database should follow the Microsoft file format for audio waveforms.
3. Multiple channels of microphone signals should be recorded.
4. A channel of clean speech signal should be recorded simultaneously for reference purposes.
5. The recording devices (microphones, recording card, etc.) should be readily available.
6. The speakers should reflect Taiwan's demographics in terms of gender, dialect, education, age, and population.
7. The database should cover all the phonetic properties of Mandarin.
8. In addition to the speech data, the corpus should also include environmental noises.
9. The database should reflect two real-world road conditions of the real world. The vehicle should be routed through a downtown area and along a highway during a recording session.
10. The database should contain some spontaneous sentences to facilitate research on mobile dialogue systems.

This paper is organized as follows. Section 2 describes the recording procedure. Section 3 presents the annotating procedure. Preliminary results of the TaiCar project are given in Section 4, and Section 5 gives a conclusion.

2. Recording Procedures

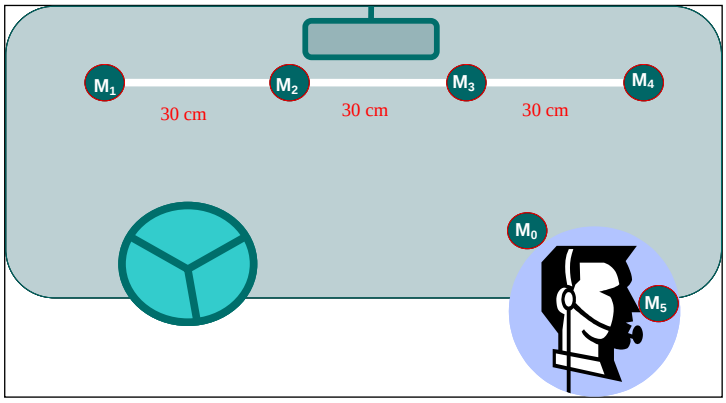
2.1 Data collection system

In this project, six recording sites at universities and research institutions have been set up so far across Taiwan. Each site uses a notebook PC equipped with a PCMCIA multi-channel signal-recording card as the recording platform. A pre-amplification circuit amplifies the input signals, which go to the recording card from the microphones. Six microphones are placed in the vehicle. A microphone array with four omni-microphones is placed on the sun visors. The distance between the microphones is 30 cm. Another microphone is bound above the notebook PC placed on the lap of the speaker. Due to safety considerations, the speaker should be the navigator instead of the driver. The last microphone, a unidirectional anti-noise one, is worn

on the head of the speaker. The reason for using such a good microphone is to provide nearly clean speech for reference purposes. The hardware elements are described in detail below:

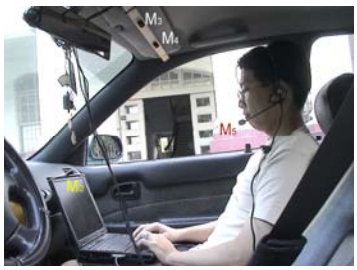
1. A DAQP PCMCIA multi-channel signal recording card capable of recording up to 16 channels of signal is plugged into the notebook PC as the recording interface.
2. Four omni-directional microphones form a linear microphone array (channels 0-3).
3. One omni-directional microphone is placed in front of the speaker (channel 4).
4. One unidirectional microphone is worn on the head of the speaker (channel 5).
5. A pre-amplification circuit is utilized before the speech signal is fed to the PCMCIA card.

Figure 1 shows the configuration and the positioning of the microphone array, the navigator, and the pre-amplification circuit.



(a)

(b)



(c)

(d)

Figure 1. (a) The configuration of TAICAR recording system. The distance between the microphones in the array is 30 cm. (b) The microphone array attached to the sun visor above; (c) the positions of the speaker and recording notebook PC; (d) the amplification circuit board for multi-channel recording.

During the recording process, the notebook PC is placed on the lap of the navigator. The material to be uttered is shown on the screen in prompts so that the speaker can follow. A sample screenshot captured during the recording procedure is shown in Figure 2.

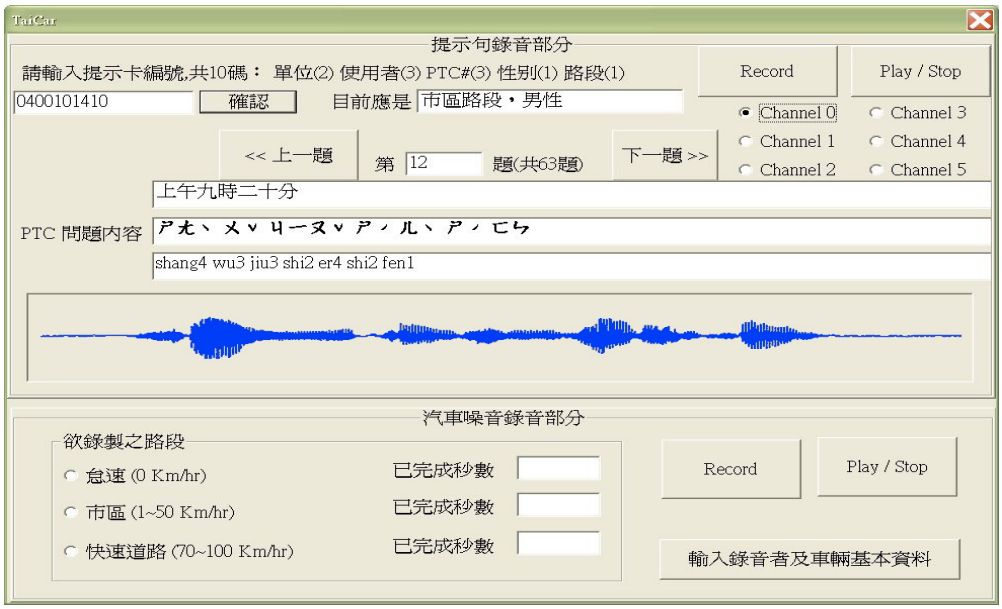


Figure 2. A screenshot from the TAICAR database recording procedure

2.2 Speech Files Format

For each utterance, six speech files are recorded. The files, saved in the MS-Windows file format for audio waveforms, are composed of two parts: a file header and sampled data. The file header contains the following information about the speech: 1) the number of channels, which indicates whether the speech was recorded in mono or stereo; 2) the number of samples recorded per second; 3) the number of bits per sample; and 4) the size of the speech data. The sampled data of speech signals are in the binary format. The files retain the waveforms of the recorded utterances as well as the preceding and following silence.

Unlike several existing speech databases, for example, MAT [Wang 1997], the transcribed Chinese characters are stored in separate files using Big-5 code. This makes it convenient to preview these files using common text processing programs under most operating systems.

2.3 Corpus Design

The TAICAR database material contains two parts. The first part is used to collect the reading speech of the speakers. It is generated by following the philosophy of the creation of

MAT-2400 database material [Wang 1997]. The framework for this material was created by Dr. Tseng of Academia Sinica [Tseng 1995]. The materials were extracted from two text corpora consisting of 77,324 lexical entries and 5,353 sentences. The material contains 407 base-syllables in Mandarin Chinese without tones; 1,062 words with two to four syllables; and 200 numbers in five different contexts, including digital sequences, dates, time, prices, and car license plate numbers.

The second part consists of spontaneous FAQ's (Frequently Asked Questions) collected from the general public in Taiwan. This material was generated by asking them several questions. The scenario questions were given to ordinary citizens, and their answers were transcribed and used as the material for the spontaneous FAQ's. The scenario questions include a description of seven query domains containing questions which are usually asked while driving a car. The seven query domains and some collected FAQ's are listed in Table 2.

Table 2. Some example of FAQ's

Domain	Scenario	Collected FAQ
Food	♦ You are hungry and looking for a restaurant.	– Where is the nearest fast-food restaurant? – Guide me to the nearest restaurant.
Clothes	♦ You want to buy some clothes.	– Where is the nearest Hang Ten? – I want to go to Far Eastern Department Store.
Lodging	♦ You are looking for a place to stay.	– I would like to know the location of the Hilton Hotel. – Show me the nearby hotels.
Navigation	♦ You want to know how to get to a destination.	– How do I go to CKS airport? – Where is City Hall?
Entertainment	♦ You want to have fun.	– How do I get to the nearest theater?
Others	♦ Weather conditions. ♦ Other information one wants to know while driving.	– What's the temperature in Taipei? – Is there any museum nearby? – Turn the CD player on.

The collected FAQ's were randomly chosen to be included in the TAICAR prompt sheets. Each prompt sheet contains 10 FAQ's that the speaker utters spontaneously.

2.4 Prompt Sheet

The prompt sheets are designed to serve as guides for the speaker to follow while uttering speech. The prompt sheet contains two parts of the aforementioned materials--the spontaneous speech and FAQ's. A total of 72 items are listed on a prompt sheet. The items are:

- 5 numbers spoken in different ways (No's. 10-14);
- 12 isolated Mandarin syllables (No's. 15-26);
- 45 isolated words (No's. 27-56, 67-82);
- 10 FAQ sentences (No's. 57-66).

The prompt sheet is designed to contain as many syllable and phonetic combinations as possible. The FAQ's are also included on the prompt sheet since they are useful for research of vehicular dialogue systems. An example of a prompt sheet is shown in Appendix A.

3. The Annotating Process

For a speech corpus to be useful, various phenomena of speaker behaviour and the deficiencies of the speech files should be annotated correctly. Since the annotation of a speech database is a labour consuming task, the tagging procedure for the TAICAR database was designed to be as convenient as possible. In the annotation phase, the annotators check whether the speech files are intelligible and whether the auto-transcribed syllables match the speaker's utterances, and they mark the starting and ending points of the speech. Figure 3 shows a screenshot of the annotation process.

Figure 3. A screenshot of TAICAR database tagging

(1) Prompt sheet number. (2) & (3) Text and phonetic transcript of the current speech. (4) & (5) Back/Proceed to other speech.(6) Mini-keyboard for modifying the phonetic syllables. (7) Click this area to select one of the six channels. The speech waveform shown will change correspondingly. (8) Play/Stop-playing this speech. (9) & (10) Left/Right Click on the waveform to mark the starting/ending point of the speech. (11) Update the database when tagging is finished.

If the starting or ending point of an utterance does not match the syllables, the annotator should mark another boundary of the utterance and correct both the text content and phonetic syllables in the database.

4. Preliminary Data Collection Result

The TAICAR project was carried out between 2002 and 2003. According to the initial plan, researchers at each recording site would record the speech of 40 speakers and annotate the utterances. However, for technical and financial reasons, researchers at some sites did not complete these tasks. In all, 192 speakers at the six recording sites participated in this project. The result was an in-car speech database consisting of utterances recorded in both downtown and highway environments. Since it is hard to accurately read long sentences on a screen while driving, utterances consisting of FAQ sentences were collected at only one site. Some statistics for the resulting database are shown in Table 3. Note that the number of files or contents is for 192 speakers driving along two different routes with six recording-channels.

Table 3. Some brief statistics for the preliminary result of TAICAR database collection

Items	Description	
Speakers	Total:	192
	Male:	115 (59.8%)
	Female:	77 (40.2%)
	Age:	from 19 to 58, mostly 20~30 (71.3%)
	Education:	most has BS degrees (89.6%)
	Daily Language:	Taiwanese (64.6%)
Car	Type:	mostly sedans (71.3%)
	Engine capacity:	below 2.0L: (57.8%); 2.0~3.0L (37.5%)
Speech data amount	6 DVDs 163,890 files 16.8 gigabytes 145.8 hours	
Database content	16,128 digits 4,608 English letters 27,648 Isolated syllables 101,376 Words with two-four characters 960 FAQ's	

Figure 4 shows the waveforms of the utterance “EQ7673” from channel 0 to channel 5.

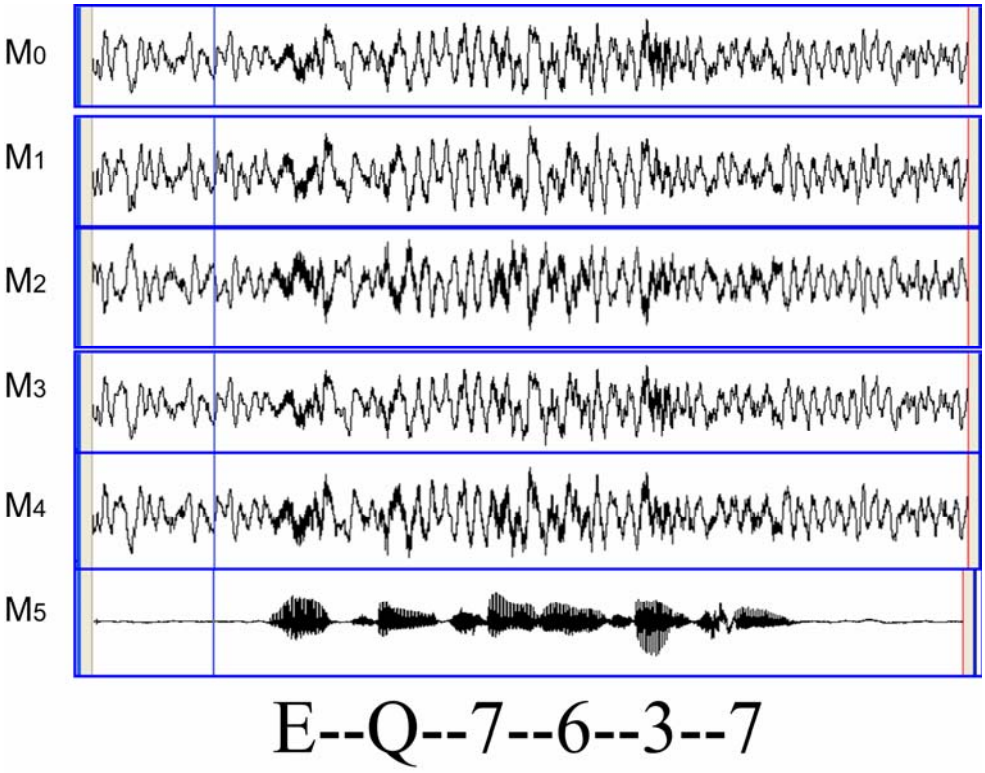


Figure 4. Speech waveforms of the utterance “EQ7637” (in Mandarin), from channel 0 to channel 5.

As mentioned in Section 2, the microphone for channel 5 is unidirectional and anti-noise. It is adopted to record the reference signal for calculating the signal-to-noise ratio (SNR) and the time shift for other channels. The SNR can be computed as follows:

$$SNR = 10 \cdot \log_{10} \frac{E[speech]}{E[noise]}, \qquad \text{(in dB)}$$

where *E[x]* stands for the energy of signal *x*.

To estimate the SNR for M_5, the speech region is detected first. Then the noise can be estimated from the non-speech part. Based on the estimated noise level, the average SNR in the speech region can be determined. By aligning the signal of M_5 with the signals of M_0~M_4, one can locate the speech regions in M_0~M_4. Then the noise level and SNRs for M_0~M_4 can be computed. The SNRs for different routes measured in the downtown and highway environments are reported in Table 4.

Table 4. SNRs (in dB) for microphones M_0~M_5 measured in the highway and downtown environments

	Channel 0	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5
Highway	-2.8550	-2.4770	-2.7171	-2.5318	-2.6763	11.4040
Downtown	-2.6187	-2.2637	-2.0714	-2.4655	-2.5261	11.2245

To calculate the time shift between channel k $(0 \leq k \leq 4)$ and channel 5, the configuration of all six microphones should be considered first, as shown in Figure 5. The microphone for channel k is M_k. The distance between M_i and M_j is $D_{i,j.}$ The distances for $D_{3,5}$ and $D_{0,5}$ are predefined as 40 cm and 60 cm, respectively. The distance between the microphones in the microphone array is 30 cm, i.e. $D_{1,2} = D_{2,3} = D_{3,4} = 30$ cm. Applying the Pythagorean Theorem, we can calculate the distances $D_{1,5}$, $D_{2,5}$, and $D_{4,5}$ obtaining 72, 50, and 50 cm, respectively. Because the speed of sound is 32,000 cm/sec, the time shift between M_i and M_j $(0 \leq i,j \leq 4)$, denoted as $T_{i,j}$, can be determined. Since M_5 is placed in front of the mouth of the speaker, it can be regarded as the original source of the utterance. The time shift for each channel can be determined as $T_{0,5}$=0.00187, $T_{1,5}$=0.00156, $T_{2,5}$=0.00125, $T_{3,5}$=0.00156, and $T_{4,5}$=0.00221.

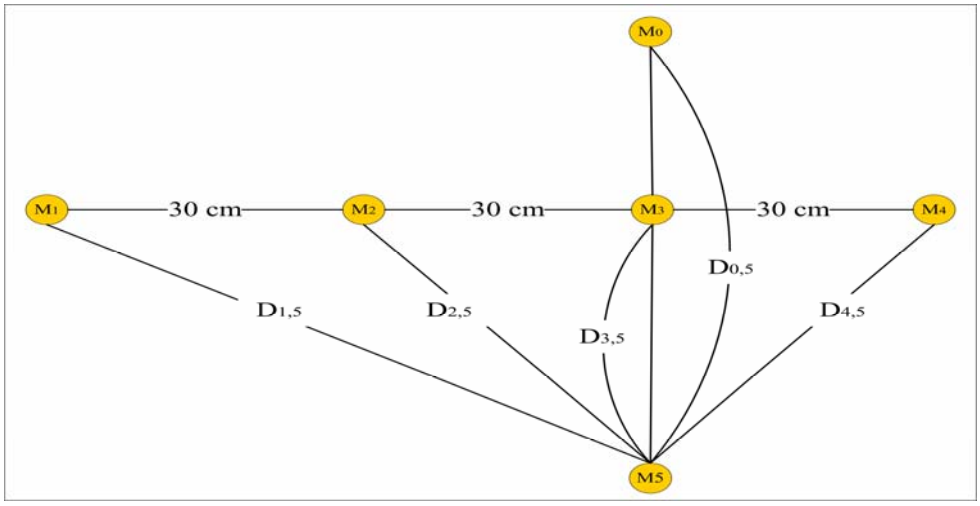


Figure 5. The detailed configuration and distances between the six microphones

5. Conclusion

This paper has described the TAICAR project that aims to create a Mandarin Chinese speech database based on the in-car environment in Taiwan. The preliminary result is a 192-speaker speech database containing 145.8 hours of utterances and environmental noises recorded in various types of automobiles. So far, two works have adopted the TaiCar corpus studies on speech enhancement in car noise environment [Yang *et al.* 2004], [Wang *et al.* 2004] and have

shown that the use of this corpus is of fundamental importance for the testing of in-car noise reduction technology. The database can also be used to develop various in-car speech processing techniques, such as speech source separation, active speech detection, channel equalization, and robust noisy speech recognition.

Acknowledgement

The authors would like to express their appreciation to the researchers from National Taiwan University, National Chiao-Tung University, National Tsing-Hua University, National Cheng-Kung University, Industrial Technology Research Institute (ITRI), and Chunghwa Telecom Laboratories (CTL). Without their help, the TaiCar project would not have been possible.

References

Bernstein, J., K. Taussig and J. Godfrey, "MACROPHONE: An American English Telephone Speech Corpus for Polyphone Project," In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 1994, Adelaide, Australia, Vol. I, pp. 81-84.

Heuvel, H., A. Bonafonte , J. Boudy, S. Dufour, P. Lockwood, A. Moreno and G. Richard, "SpeechDat-Car: Towards a collection of speech databases for automotive environments," In *Proceedings of the Nokia-COST249 Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, 1999, Tampere, Finland, pp. 135-138.

Heuvel, H., J. Boudy, R. Comeyne, S. Euler, A. Moreno and G. Richard, "The SpeechDat-Car multilingual speech databases for in-car applications: Some first validation results," In *Proceedings of 6th European Conference on Speech Communication and Technology*, 1999, Budapest, Hungary, Vol.5, pp. 2279-2282.

Itakura, F., "Multi-Media Data Collection for In-Car Speech Communication - Ongoing Data Collection and Preliminary Results," In *Proceedings of International Workshop on Hand-Free Speech communication*, 2001, Kyoto, Japan, pp. 1-5.

Kudo, I., T. Nakama, N. Arai and N. Fujimura, "The Database Collection of Voice Across Japan (VAJ) Project," In *Proceedings of 2nd International Conference on Spoken Language Processing*, 1994, Yokohama, Japan, pp.1799-1802.

Langmann, D., H.R. Pfitzinger, T. Schneider, R. Grudszus, A. Fischer, M. Westphal, T. Crull and U. Jekosch, "CSDC, the MoTiV Car Speech Data Collection," In *Proceedings of 1st International Conference on Language, Resources and Evaluation*, 1998, Granada, Spain, pp. 1107-1110.

Lee, B., H.-J. Mark, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. Huang, "AVICAR: Audio-Visual Speech Corpus in a Car Environment," In *Proceedings of 8th International Conference on Spoken Language Processing*, 2004, Jeju Island, Korea.

Roach, P., S. Arnfield, W. Barry, J. Baltova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum,L. Lamel, K. Marasek, A. Marchal, E. Meister and K. Vicsi, "BABEL: An Eastern European Multi-language Database," In *Proceedings of 4th International Conference on Spoken Language Processing*, 1996, Philadelphia, USA, pp. 1892-1893.

Tapias, D., A. Acero, J. Esteve and J.C. Torrecilia, "The VESTEL Telephone Speech Database," In *Proceedings of 3rd International Conference on Spoken Language Processing*, 1994, Yokohama, Japan, pp.1811-1814.

Tseng, C.Y. "A Phonetically Oriented Speech Database for Mandarin Chinese," In *Proceedings of the 13th International Congress on Phonetic Sciences*, 1995, Stockholm, Sweden, Vol. 3, pp.326-329.

Wang, H.C., "MAT - A Project to Collect Mandarin Speech Data Through Telephone Networks in Taiwan," *Computational Linguistics and Chinese Language Processing*, 2(1), 1997, pp. 73-90.

Wang, J.-F., C.-H. Yang and K.-H. Chang, "Using Perceptual Wavelet Decomposition and Subspace Tracking for Noise Removal in Car Environment," In *Proceedings of ROCLING XVI: Conference on Computational Linguistics and Speech Processing*, 2004, Taipei, Taiwan.

Yang, C.-H., J.-F. Wang and K.-H. Chang, "Subspace Tracking for Speech Enhancement in Car Noise Environments," In *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, 2004, Quebec, Canada.

Appendix A. A Sample of the Prompting Sheet

(10) 七三八 零四零八
(11) 十二月十七日
(12) 上午九時二十分
(13) 四千五百二十二元
(14) EQ 七六三七
(15) 該
(16) 案
(17) 鎮
(18) 榮
(19) 轉
(20) 卡
(21) 推
(22) 桌
(23) 怒
(24) 踐
(25) 說
(26) 的
(27) 予以
(28) 財務
(29) 喜愛
(30) 案由
(31) 給予
(32) 爲要
(33) 台銀
(34) 掃蕩
(35) 手腕
(36) 搬運
(37) 合約
(38) 應用
(39) 黨外
(40) 加速
(41) 花園
(42) 去年
(43) 佛像
(44) 尊重
(45) 狀況
(46) 內閣制
(47) 改革派
(48) 閩南語
(49) 消防車
(50) 療養院
(51) 風風雨雨
(52) 未雨綢繆
(53) 布魯塞爾
(54) 訓導主任
(55) 血本無歸
(56) 莫名其妙
(57) 我要找吃飯的地方
(58) 哪裡有餐廳
(59) 最近的加油站在哪裡
(60) 附近有沒有加油站
(61) 加油站還有多遠
(62) 最近的餐廳在哪裡
(63) 附近有沒有麥當勞速食店
(64) 我想找肯得基，哪裡有？
(65) 車子快沒油了，最近的加油站在哪裡
(66) 台南車站要怎麼走
(67) 七五九五
(68) 四五七一八七九
(69) 泰國
(70) 人事行政局
(71) 聯華電子公司
(72) 彰化商業銀行
(73) 無法
(74) 代表
(75) 不過
(76) 報導
(77) 方式
(78) 調查
(79) 工業區
(80) 台中市
(81) 戡亂時期
(82) 金融機構