Computational Linguistics and Chinese Language Processing
vol.2, no.1, February 1997, pp. 73-90



# MAT -- A Project to Collect Mandarin Speech Data Through Telephone Networks in Taiwan

**Hsiao-Chuan Wang***

**Abstract**

A cooperative project, called ***Polyphone***, was initiated by the Coordinating Committee on Speech Databases and Speech I/O Systems Assessment (***COCOSDA***) in 1992. Accordingly, a project to collect ***M***andarin speech data across ***T***aiwan (***MAT***) was conducted by a group of researchers from several universities and research organizations in Taiwan. The purpose was to generate a speech corpus for the development of Mandarin-based speech technology and products. The speech data were collected at eight recording stations through telephone networks. The speakers were chosen so as to reflect the population of the gender, the dialect, the educational level, and the residence in Taiwan. A preliminary Mandarin speech database of 800 speakers has been produced. The final goal is to generate a speech database of at least 5000 speakers.

**Key words: Mandarin speech, Speech database, Speech I/O systems assessment, Telephone network**

## 1. Introduction

The "Polyphone" project was initiated by the Coordinating Committee on Speech Databases and Speech I/O Systems Assessment (COCOSDA) during the International Conference on Spoken Language Processing (ICSLP-92) held in Banff, Canada, in October, 1992 [Jones and Mariani 1992]. The purpose was to coordinate the speech data collection of major languages in the world. Representatives from Europe, the USA, Canada, Australia, Japan, China, and Taiwan showed great interest in joining this project. Each language group had to seek funding from its own country. The status and progress of the "Polyphone" project of each language group were reported in the following COCOSDA meetings held in Berlin, Germany (September 1993), Yokohama, Japan (September 1994), Madrid, Spain (September 1995) and Philadelphia, the USA (September 1996). Furthermore, a Linguistic Data Consortium (LDC) was formed in the USA, in 1992. This consortium, started with a $5 million grant, was hosted by the University of Pennsylvania, USA. The mission of LDC was to share experiences and to

* Department of Electrical Engineering, National Tsing Hua University, Taiwan. E-mail: hcwang@ee.nthu.edu.tw

coordinate the release of corpora developed by its members. By 1995, 29 speech corpora had been released by LDC [LDC 1996]. An additional 31 speech corpora were scheduled to be released in 1996 and 1997. Most of the speech corpora are English collected in the United States. Mandarin Chinese only appears as a subset of the ORI Multi-language telephone corpus. Another group coordinating the speech databases is the European Language Resources Association (ELRA). Thirty-four speech databases were listed in the catalogue of ELRA by December 1996 [ELRA 1996]. Appendix A shows some databases reported by LDC and ELRA. Some speech data collection projects reported in recent years are briefly described as follows;

TED (Translanguage English Database) --
224 oral presentations recorded at EUROSPEECH'93 in Berlin [Lamel *et al.* 1994].

VAJ (Voice across Japan) --
a telephone database of 10,000 speakers collected in Japan [Kudo *et al.* 1994].

Dutch Polyphone Corpus --
a telephone database of 5,000 speakers [Damhuis *et al.* 1994].

RAFAEL.0 (Scandinavian) --
telephone speech data of 3000 speakers from Denmark, Norway and Sweden [Rosenbeck *et al.* 1994].

VESTEL (Spanish) --
a telephone speech corpus collected in Spanish [Tapias *et al.* 1994].

VAHA (Voice across Hispanic America) --
telephone speech of 5000 speakers in the US [Godfrey 1994]

CEUDEX (Spanish) --
400 phonetically balanced sentences from 300 speakers [Torre *et al.* 1995].

Bulgarian speech database (Bulgarian) [Misheva *et al.* 1995].

PhonDat (German Verbmobil project) [Hess *et al.* 1995].

EUROM -(Danish, Dutch, English, French, Italian) --
a spoken language resource for the EU [Chan *et al.* 1995].

BABEL (Bulgarian, Estonian, Hungarian, Romanian, and Polish) --
an eastern European multi-language database [Roach *et al.* 1995].

USTC-95 (Putonghua) --
Mandarin Chinese speech database collected from 280 speakers in Mainland China [Wang *et al.* 1996].

SIVA (Italian) --
speech database for speaker verification [Falcone and Gallo 1996].

FRESCO (French) --
a telephone speech collection project to generate a database of 1000 speakers [Langmann *et al.* 1996].

Korean speech database --
a multi-level speech database for spontaneous speech processing [Hahn *et al.* 1996].

Romanian speech database --
a project to collect Romanian speech data for speech recognition and synthesis [Boldea *et al.* 1996].

It is obvious that researchers in many countries have attempted to produce speech databases of their own languages. Some of them are designed as multilingual databases. Cooperative projects have worked well in European countries. Mandarin Chinese is a spoken language used by about a quarter of the population in the world. However, no Mandarin speech database of more than one thousand speakers has been reported.

## 2. Speech Databases in Taiwan

Researches on speech processing have been conducted by many universities and organizations in Taiwan during the past decade. Some researches have produced remarkable results, such as the Mandarin dictation machine developed at National Taiwan University and Academia Sinica [Lee *et al.* 1995], the Chinese text-to-speech system developed at National Chiao Tung University [Hwang, Chen and Wang 1996], and the Venus Dictate (a large vocabulary speech recognition system) developed at National Cheng Kung University [Huang and Wang 1994]. Many speech databases have been generated for these research purposes. Telecommunication Laboratories (TL) of the Chung Hua Telecommunication Co. is a research organization which has carried out many researches on speech recognition and synthesis for telephone network services. This research organization has also generated some speech databases and distributed them to universities for research purposes. Some of speech the databases produced in Taiwan are shown in the table in Appendix B.

In this survey, we find that most of these speech databases were collected using microphones in laboratories. The speech data were mostly contributed by college students. For each database, the number of speakers was around 100. Since close-talk microphones were used, the speech quality was well controlled. These databases can satisfy some basic researches, but may not be practical for developing a telephone speech recognition technique where the speech signal is contaminated by noises and channel

distortion. Furthermore, the databases were produced individually in different laboratories so that there was no common format or description standard. Because of the lack of a standard speech database, it is not possible to perform a fair assessment for these developed prototypes.

## 3. MAT Project

In response to the "Polyphone" project, a group of researchers in the area of speech processing in Taiwan also initiated a speech data collection project called MAT (***M*** andarin speech data ***a***cross ***T***aiwan). The objective of the MAT project was to produce a speech database of Mandarin Chinese spoken in Taiwan. Some targets were set for this project.

(1) The PC will be the platform for speech data collection and speech file editing.

(2) The speech data can be collected through a microphone or the telephone system.

(3) A popularly used sound card will be chosen for speech signal input and output.

(4) A generalized format must be designed for the speech data files.

(5) The database should cover all the phonetic properties of Mandarin speech.

(6) The speakers should reflect the population of the genders, the dialects, the educational level, and the residence in Taiwan.

Since August, 1995, this project has been sponsored by the National Science Council, the ROC. In its first year, the scheduled tasks were

(1) to set up standard hardware and software for speech data collection,

(2) to design the contents of the database, and

(3) to generate a preliminary database of at least 500 speakers.

The result of the first year of work is a preliminary database of 800 speakers. This preliminary database will be tested by potential users to verify its availability and quality. The final goal is to produce a speech database of at least 5000 speakers.

## 4. Data Collection System

In this project, we set up eight speech data collection stations in universities and research organizations. Each station consisted of a personal computer equipped with a telephone interface card to allow speakers to input their voices by using any telephone handset around Taiwan through the public switching telephone network. A detailed description of the hardware organization is giving below:

(1) A PC with an Intel 486 or Pentium CPU is the kernel of the speech data collection station.

(2) A Dialogic D/41D card is added to the PC to provide a telephone interface.

(3) A Sound Blaster card with 16-bit ADC and DAC channels is used for speech input. It is also used to play back speech during speech file editing.

(4) The Analog Extension Bus (AEB) on the Dialogic card is connected to the Line-In port of the Sound Blaster card to get 16-bit linear PCM signal input. This allows users to choose input from a telephone line or from a microphone.

(5) An external hard disk and optical disk are necessary for mass storage.

The software is designed in two parts: a speech recording program and a speech file editing program.

### 4.1 Speech Recording Program (VCORDER)

This program runs in a DOS environment and provides the functions of an I/O driver, an interface with the speaker, prompt for speech input, extraction of speech signals, detection of endpoints of utterances, initiation of the file header, and generation of speech files. A menu-type user interface is designed to allow users to specify the default file header parameters, the recording environment, the signal input channel (through a telephone line or a microphone), and the encoding mode. For MAT data collection, the sampling rate is set to 8 kHz, and the data mode is 16-bit linear PCM.

### 4.2 Speech File Editing Program (VEDITOR)

This program works in a Windows environment. It provides a tool for users to edit speech files. The file header parameters, as well as the waveform, can be displayed on the screen. The user can edit the file header, modify the waveform, and playback the edited voice in an interactive mode.

Furthermore, the prompting voices and prompting sheet can be easily changed by replacing the prompting voice files and the prompting text files. That means that the system can be easily converted to fit any specific purpose. This speech data collection system has been designed not only for the MAT project, but also for other speech data collection projects in the future. It will produce speech data files in a standard format.

## 5. Speech File Format

Each utterance is collected and stored as a speech file. The speech file is identified by the file name extension ".vat". It is a binary file composed of two parts; the file header and the sampled data. The file format is designed to contain as many parameters as in the

"Macrophone" project initiated by SRI International [Bernstein, Taussig, and Godfrey 1994] and in the VAJ project initiated by Texas Instruments [Kudo *et al.* 1994].

### 5.1 File Header

The length of a file header is 256 bytes. It can be extended to 512 bytes by attaching an additional 256-byte block. This extended block can be used for additional transcripts of speech data. There are 28 parameters and 3 blocks defined in the file header. The parameters in the file header can be grouped into several categories.

(1) Basic data - including the header length, the sampled data length, the recording date, the recording time, the recording site, and the database name.

(2) Data type - including the encoding type, the sampling rate, and the number of bits per sample.

(3) Content description- including the prompting sheet number, the item number, and the number of transcribed characters of the recorded utterance.

(4) Speaker's personal data - including the speaker's gender, age, accent, education level, mother tongue, daily language, and residence.

(5) Speaking style and quality- including the speaking rate, articulation, effort, mode, and quality.

(6) Signal conditions - including the signal condition and signal quality.

Three blocks are used to store the transcribed Chinese characters, the phonetic symbols of the Chinese characters, and the user defined information. The maximum number of transcribed Chinese characters is 27. This is enough to pronounce an ordinary sentence. The Chinese characters are represented in Big-5 code. Their corresponding phonetic symbols are denoted using Pin-Yin. A detailed description of the file header is given in Appendix C.

### 5.2 Sampled Date

The sampled data of speech signal are in binary format. This sequence of sampled data retains the waveform of the recorded utterance as well as its preceding and succeeding silent portions. After a speech file is edited, the silence portion is set to about 0.5 seconds before and after the speech signal. This allows the user to get the background noise information from the retained silent portions.

## 6. Material Design

### 6.1 Spoken Materials

The spoken materials are designed for generation of speech models and evaluation of the telephone-based speech recognition systems developed for Mandarin speakers. The framework of the material design was created by Dr. Chiu-Yu Tseng of Academia Sinica [Tseng 1995]. The materials were extracted from two text corpora of 77,324 lexical entries and 5,353 sentences. Forty sets of speech materials were produced to generate the prompting sheets. A brief description of the speech materials is given as follows.

(1) They cover 407 base-syllables without concerning the tones in Mandarin Chinese.

(2) They contain 1062 words with two to four syllables in each word. These words cover 338 tone combinations and 1351 voiced vs. voiced/unvoiced combinations.

(3) They contain 400 sentences with at most 27 Chinese characters in each sentence. These sentences cover 399 base-syllables, 289 tone combinations, and 1434 voiced vs. voiced/unvoiced combinations.

This database also contains 200 numbers pronounced in five different ways. A set of examples is shown below. Their pronunciation is transcribed in Pin-Yin shown in parentheses:

(1) Digit sequence - 118 2720 (yi1 yi1 ba1 er4 qi1 er4 ling2).

(2) Date - 2nd of October (shi2 yue4 er4 ri4)

(3) Time - 10:33 am (shang4 wu3 shi2 shi2 san1 shi2 san3 fen1)

(4) Price - 1341 dollars (ti1 qian1 san1 bai3 si4 shi2 yi1 yuan2)

(5) Car plate - WB 4522 ( W - B - si4 wu3 er4 er4)

## 6.2 Prompting Sheet

The prompting sheets are designed to guide speakers as they input speech data. The necessary information for speakers is given on the first page. This page also has nine questions used to gather information about the speaker, such as the speaker's gender, age, language background, education level, and residence. These data are used to set some parameters in the file header. The speaker's responses to these questions are collected as spontaneous speech data. The second page contains 57 items. The speaker is asked to read these items following instructions given by the system. These items are grouped into four parts:

(1) 5 numbers spoken in different ways (Prompting Item No. 10 - 14),

(2) 12 isolated Mandarin syllables (Prompting Item No. 15 - 26),

(3) 30 isolated words of 2 ~ 4 characters (Prompting Item No. 27 - 56), and

(4) 10 phonetically balanced sentences (Prompting Item No. 57 - 66).

These materials are arranged into 40 phonetically rich sets so that 40 prompting sheets are accordingly generated. That means that a speaker can provide utterances with as many syllables and phonetic combinations as possible. A sample of the prompting sheet is shown in Appendix D.

## 7. A Preliminary Database

A preliminary database was produced in June 1996. This first version of the MAT database (MAT Database version 1.0) contains 52671 speech data files. Each file is an utterance collected from a speaker. There are 424 male and 376 female speakers who provided the speech data. For convenience, the speech data are arranged into five subsets, which are shown in Table 1.

***Table 1.*** *Subsets of the MAT speech database*

| Subject | Number of Files | Prompting Item Numbers | Speaking Style | Description |
|---|---|---|---|---|
| MATDB-1 | 7191 | 1 - 9 | spontaneous | short answering |
| MATDB-2 | 3991 | 10 - 14 | read | statements numbers pronounced in five different ways |
| MATDB-3 | 9576 | 15 - 26 | read | Mandarin syllables |
| MATDB-4 | 23939 | 27 - 56 | read | words of 2 to 4 syllables |
| MATDB-5 | 7974 | 57 - 66 | read | phonetically balanced sentences |

No screening mechanism is used to discard files with poor signal quality. The signal quality and condition are specified by the parameters in the file header. The statistics show that 901 speech files were inappropriately recorded. Futhermore, about 18% of the speech files were found to be contaminated by noises or distortion. Meanwhile, we allowed the database to be tested by potential users. A method for selecting good speech files will be decided upon after the testing period.

Since the number of speakers is still small in this preliminary database, the distributions of the speaker age, educational level, and residence can not reflect the distributions of the population in Taiwan. This situation will be improved when the number of speakers increases to 5000 in the subsequent recording process.

The language background of the speakers is also interesting. The statistics for speaker accent, mother tongue, and daily language are shown in Table 2.

***Table 2.*** *Statistics for Speaker Language background*

| | Accent | Mother Tongue | Daily Language |
|---|---|---|---|
| Taiwanese Hokkian | 120 | 499 | 108 |
| Mandarin Chinese | 378 | 202 | 661 |
| Kakka | 10 | 53 | 9 |
| Aboriginal | 0 | 0 | 0 |
| non-Chinese | 0 | 0 | 0 |
| Unidentified | 282 | 46 | 22 |

The table shows that more than 80% of the speakers use Mandarin as their daily spoken language even though many of their mothers speak Taiwanese Hokkian. This may come from the fact that most of the speakers are students. The table also shows that a speaker's accent is hard to identify during the editing process. The reason is that listeners in the editing process can not distinguish the accent. In fact, most of the people in Taiwan speak Mandarin with a Taiwanese accent.

## 8. Conclusion

The MAT is a project designed not only for collecting Mandarin speech data through a telephone network, but also for setting up a standard format for speech data files in Taiwan. Using this standard file format, speech databases produced in Taiwan can be easily shared by researchers and industry. The preliminary database is not perfect but can be used as test data for evaluating speech recognition systems. These telephone speech data can also be used to develop techniques for noise compensation, channel equalization, and robust speech recognition. The final goal of the MAT project is to produce a speech database of a large number of speakers so that we will have a good tool for developing Mandarin-based speech technologies in Taiwan.

## Acknowledgments

This work has been sponsored by the National Science Council, the ROC, under contract no. NSC-85-2213-E-007-043. The author would like to thank his colleagues for their support and contributions in data collection. They are (in alphabetical order) Dr. Chao-Huang Chang, Prof. Sin-Horng Chen, Prof. Ching-Tang Hsieh, Dr. Eng-Fong Huang, Prof. Yau-Tarng Juang, Prof. Lin-Shan Lee, Prof. Keh-Yih Su, Prof. Chiu-Yu Tseng, and Prof. Jhing-Fa Wang.

## References

Bernstein, J. and K. Taussig, J. Godfrey, "MACROPHONE: An American English Telephone Speech Corpus for Polyphone Project," *ICASSP'94*, Adelaide, Australia, 1994, I-81-I-84.

Boldea, M. and A. Doroga, T. Dumitrescu, M. Pescaru, "Preliminaries of a Romanian Speech Database," *ICSLP'96*, Philadelphia, PA, 1996, pp.1934-1937.

Chan, D. *et al.*, "EUROM -- a Spoken Language Resource for the EU," *EUROSPEECH'95*, Madrid, Spain, 1995, pp.867-870.

Damhuis, M. *et al.*, "Creation and Analysis of Dutch Polyphone Corpus," *ICSLP'94*, Yokohama, Japan, 1994, pp.1803-1806.

ELRA, ELRA Catalogue release 1.4, December 1996.

Falcone, M. and A. Gallo, "The SIVA Speech Database for Speaker Verification: Description and Evaluation," *ICSLP'96*, Philadelphia, PA, 1996, pp.1902-1904.

Godfrey, J. "Polyphone: Second Anniversary Report," Notes form the COCOSDA Workshop 94,Yokohama, Japan, 1994

Hahn, M. and S. Kim, J. C. Lee, Y. J. Lee, "Constructing Multi-level Speech Database for Spontaneous Speech Processing," *ICSLP'96*, Philadelphia, PA, 1996, pp.1930-1933.

Hess, W. and K.J. Kohler, H.G. Tillmann, "The PhonDat-verbmobil Speech Corpus," *EUROSPEECH'95*, Madrid, Spain, 1995, pp.863-866.

Huang C.C. and J. F. Wang, "A Mandarin Speech Dictation System Based on Neural Network and Language Processing Model," *IEEE Trans. Consumer Electronics*, Vol. 40, No. 3, 1994, 437-445.

Hwang, S.H. and S.H. Chen, Y.R. Wang, "A Mandarin Text-to-speech System," *ICSLP'96*, Philadelphia, PA, 1996, pp. 1421-1424.

Jones, K. and J. Mariani (edited), Proceedings of the 1992 Workshop of the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment, Banff, Canada, October 1992.

Kudo, I. and T. Nakama, N. Arai, N. Fujimura, "The Database Collection of Voice Across Japan (VAJ) Project," *ICSLP'94*, Yokohama, Japan, 1994, pp.1799-1802.

Lamel, L.F. *et al.*, "The Translanguage English Database (TED)," *ICSLP'94*, Yokohama, Japan, 1994, pp.1795-1798.

Langmann, D. and R. Haeb-Umbach, L. Boves, E. den Os, "FRESCO: The French Telephone Speech Data Collection - Part of the European SPEECHDAT(M) Project," *ICSLP'96*, Philadelphia, PA, 1996, pp.1918-1921.

LDC, A note of corpora released by LDC, 1996.

Lee, L.S. *et al.*, "Golden Mandarin (III) - A User-adaptive Prosodic-segment-based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary," *ICASSP'95*, Detroit, Michigan, 1995, pp. 57-60.

Misheva, A. *et al.*, "Bulgarian speech database: a pilot study," *EUROSPEECH'95*, Madrid, Spain, 1995, pp.859-862.

Roach, P. *et al.*, "BABEL: An Eastern European Multi-language Database," *ICSLP'96*, Philadelphia, PA, 1996, pp. 1892-1893.

Rosenbeck, P. *et al.*, "The Design and Efficient Recording of a 3000 Speaker Scandinavian Telephone Speech Database: RAFAEL.0," *ICSLP'94*, Yokohama, Japan, 1994, pp.1807-1810.

Tapias, D. and A. Acero, J. Esteve, J.C. Torrecilia, "The VESTEL Telephone Speech Database," *ICSLP'94*, Yokohama, Japan, 1994, pp.1811-1814.

Torre, C. and L. Hernandez-Gomez, D. Tapias, "CEUDEX: A Database Oriented to Context-dependent Units Training in Spanish for Continuous Speech Recognition," *EUROSPEECH'95*, Madrid, Spain, 1995, pp.845-848.

Tseng, C.Y. "A Phonetically Oriented Speech Database for Mandarin Chinese," *ICPhS'95*, Stockholm, Sweden, 1995, Vol. 3, pp.326-329.

Wang, R.H. and D. Xia, J. Ni, B. Liu, "USTC95 - A Putonghua Corpus," *ICSLP'96*, Philadelphia, PA, 1996, pp.1894-1897.

## **Appendix A** - Speech Databases Reported by LDC in 1996

**Table A1.** *Speech Databases Reported by LDC in 1996*

| Database name | Sponsor | Size, CDs | Spkrs | Speaking Style | Sampling Rate, bit per sample | Contents and Descriptions |
|---|---|---|---|---|---|---|
| TMIT | ARPA | 1 | 630 | read sentences | 16 kHz, 16 bits | 10 phonetically rich sentences per speaker |
| NTIMIT, telephone version of TIMIT | NYNEX | 2 | 630 | read sentences | 8 kHz | collected by transmitting all TIMIT recordings through a telephone handset over various channels |
| RM1, Resource Management Corpus | ARPA | 4 | 160 | read sentences | 16 kHz, 16 bits | 25,000 utterances, 1000 word vocabulary |
| RM2, Extended RM1 | ARPA | 2 | 4 | read sentences | 16 kHz, 16 bits | 2652 sentences per speaker |
| ATIS0, Air Travel Information System Corpus | ARPA | 6 | 36 | spontaneous and read | 16 kHz, 16 bits | disc 1 - 912 spontaneous utterances by 36 speakers<br>disc 2 - read version of spontaneous utterances for 20 of 36 speakers<br>disc 3~6 read speech, 3171 utterances from 10 speakers |
| ATIS2, Air Travel Information System Corpus | ARPA | 4 | 450 | spontaneous | 16 kHz, 16 bits | 15,000 utterances from 450 speakers at five sits, AT&T, BBN, CMU, MIT, and SRI |
| ATIS3 - Training Data | ARPA | 3 | 137 | spontaneous | 16 kHz, 16 bits | over 774 scenarios with 7,300 utterances including the flight information for 46 cities and 52 airports |
| ATIS3 - Test Data | ARPA | 2 | 137 | spontaneous | 16 kHz, 16 bits | two 1000-utterance test sets |
| WSJ0 or CSR-I, Continuous Speech Recognition Corpus (Wall Street Journal Sentences) | ARPA | 15 | | read | 16 kHz, 16 bits | 31,000 utterances (40 hours of speech) |
| WSJ1 or CSR-II, Continuous Speech Recognition Corpus (Wall Street Journal Sentences) | ARPA | 34 | | read and spontaneous | 16 kHz, 16 bits | 78,000 utterances ( 73 hours of speech) |
| ARPA CSR III - Continuous Speech Recognition Corpus (Financial News) | ARPA | 3 | 180 | spontaneous | 16 kHz, 16 bits | 20~40 sentences per speaker |
| SWITCHBOARD Corpus (recorded Telephone Conversations) | ARPA | 28 | 543 | spontaneous | 8 kHz, $\mu$-law | two-sided telephone conversations among 543 speakers from all areas of the US |
| SWITCHBOARD Corpus Excerpts (Credit Card Conversations) | ARPA | 1 | 69 | spontaneous | 8 kHz, $\mu$-law | 35 conversations on the topic of "Credit Card Use" |

| | | | | | | |
|---|---|---|---|---|---|---|
| TI-46, TI 46-word Speaker-Dependent Isolated Word Corpus | TI | 1 | 16 | read | 12.5 kHz, 12 bits | 26 utterances of the 46-word vocabulary |
| TIDIGITS, TI Speaker-Independent Connected-Digit Corpus | TI | 3 | 326 | read | 20 kHz | 77 digit sequences per speaker |
| HCRC Map Task | Univ. of Glasgow | 8 | 64 | spontaneous | 20 kHz, 16 bits | 128 two-person conversations |
| ATC0, Air Traffic Control Corpus | ARPA | 8 | | spontaneous | 8 kHz, 16 bits | voice communication traffic between various controllers and pilots |
| SPIDRE Speaker Identification Corpus | ARPA | 2 | 45 | spontaneous | 8 kHz | a sub-set of SWITCHBOARD collection, 45 target speakers, 4 conversations fromeach target speaker, 100 calls of no target appearing |
| OGI Multi-Language Telephone Corpus | Oregon Graduate Institute | 2 | 200 | spontaneous | 8 kHz | responses over telephone lines by speakers of English, Farsi, French, German , Hindi, Japanese,Korean, Mandarin Chinese, Spanish,Tamil, and Vietnamese, 1927 calls, 175 calls per language |
| OGI Spelled and Spoken Word Corpus | Oregon Graduate Institute | 1 | | spontaneous | 8 kHz | 3,650 telephone calls speakers pronounced their names, where they were calling from, where they grew up, answered yes/no questions, and spelled their first names and last names. |
| MACROPHONE | SRI, LDC | 7 | 5000 | spontaneous | 8 kHz, $\mu$-law | 200,000 utterances through telephone lines from all regions of the United States |
| KING-92, Corpus for Speaker Verification Research | ITT | 1 | 51 | read | 8 kHz | one version from a high-quality microphone, one version from a telephone handset; original data was sampled at 10 kHz but has now been re-sampled at 8 kHz |
| WSJCAM0 | Cambridge Univ. | 6 | 92 | read | 16 kHz | subjects were native speakers of British English, 90 utterances per speaker |
| NYNEX PHONEBOOK Database | NYNEX | 2 | 1358 | read | 8 kHz, $\mu$-law | phonetically-rich, isolated words, 93,667 isolated-word utterances |
| LATINO-40 Spanish Read News Corpus | Entropic Research lab. | 2 | 40 | read | 16 kHz, 16 bits | sentences of shorter than 80 characters, 125 utterances per speaker |
| CTIMIT, Cellular TIMIT Speech Corpus | Lockheed-Martin Sanders | 1 | 630 | read | 8 kHz | similar to NTIMIT, TIMIT recordings were passed through cellular circuits |
| FEMTIMIT, Far Field Microphone Recording of TIMIT | ARPA | 1 | 630 | read | 16 kHz, 16 bits | including low frequency noise |

**Table A2.** *Speech Databases reported by ELRA in 1996*

| Database Name | Description |
|---|---|
| ACCOR (7 languages) | acoustic and articulatory multilingual database, 7 languages recorded as part of the ESPRIT-ACCOR project |
| AIDA-1 (Italian) | several sets of meaningless words, phonetically dense and digits, recorded by 20 males and 20 females |
| BDLEX 23000 (French) | a phonetically transcribed French lexicon of 23000 canonical entries |
| BDLEX 50000 (French) | a phonetically transcribed French lexicon of 50000 canonical entries |
| BDSONS (French) | French and Canadian French speech database, recorded by 16 males and 16 females |
| BREF sub-corpus BREF-80 (French) | 5330 sentences read by 80 French speakers |
| BREF sub-corpus BREF-Polylot (French) | 3193 sentences read by 80 French speakers |
| COLLECT-500 (Italian) | 10 Italian digits and 5 command words, 500 speakers* calls |
| COST 232 (English) | Multi-English Speech database, 797 calls received in Italy and UK |
| Dutch Polyphone Database (Dutch) | telephone speech of 5050 Dutch speakers, 44 items per speaker, read and spontaneous |
| English Polyphone Database -DB1 (English) | 1000 speakers recorded over digital telephone lines, phonetically rich sentences, keywords, digits |
| English Polyphone Database -DB2 (English) | phonetically rich sentences sub-set of DB1 |
| Erlanger Bahnansage - ERBA (German) | 10000 utterances read by 100 German speakers, domain of train inquires |
| EUROM1 | multilingual European speech database, recorded by a microphone, 60 speakers per language, numbers, sentences, isolated words |
| EUROM1I (Italian) | Italian release of EUROM1 |
| French polyphone database FRESCO -DB1 (French) | 35000 utterances, 1000 callers over telephones in France, phonetically rich sentences, keywords, digits |
| French polyphone database FRESCO -DB2 (French) | phonetically rich sentences sub-set of DB1 |
| German polyphone database -DB1 (German) | 1000 speakers, phonetically rich sentences, keywords, digits |
| German polyphone database -DB2 (German) | phonetically rich sentences sub-set of DB1 |
| GRONINGEN (Dutch) | 20 hours of reading of short sentences and short texts, from 238 speakers |
| M2VTS (multi-language) | multiligual database designed to facilitate access control using multimodal identification of human faces |
| Onemastica Multi-language pronunciation dictionaries (11 European languages) | city names, street names, family names, first names, product names |
| PHONDAT1 PD1 (German) | 201 speakers, read 450 sentences per speakers |
| PHONDAT2 PD2 (German) | 16 speakers, 200 different sentences from train inquiry task |
| SIEMENS 100 (German) | 101 speakers, 100 sentences from German newspapers |
| SIEMENS 1000 (German) | 10 speakers, 1000 sentences from German newspapers |
| SieTill telephone speech database (German) | 36000 utterances, 730 speakers, digit sequences, dates, spelled names |
| SIVA (Italian) | speech database for verification and identification, 2000 calls in Italian language collected over telephones |
| Strange Corpus 1 -SC1(German) | "Nordwind und Sonne" story read by 72 speakers with foreign accents and 16 native German speakers |
| Swiss-French polyphone database (Swiss-French) | 5000 speakers, 10 answers to questions, 28 read items |
| Translanguage English database -TED (English) | recording of 188 oral presentations in English given at EUROSPEECH, 93 in Berlin |
| TEDPhone (multi-language) | polyphone like recording of 64 speakers in English and in their native languages |
| BDBRUIT (French) | recording of French speech corrupted with noisy, especially the Lombard |
| VERBMOBIL (German) | spontaneous speech database recorded in dialogue task |

**Appendix B** - Samples of Mandarin Speech Databases Developed in Taiwan

| Database name | Sponsor | Size | Spkrs | Speaking Style | Sampling Rate, bit per sample | Contents and Descriptions |
|---|---|---|---|---|---|---|
| TL Telephone Words and Sentences | CHTC | 380 MB | 120 | read | 8 kHz, 16 bits | 120 utterances per speaker |
| TL Telephone Isolated Words | CHTC | 100 MB | 116 | read | 8 kHz, 16 bits | 50 utterances per speaker |
| TL Digits | CHTC | 600 MB | 100 | read | 10 kHz, 16 bits | 6 sessions per speaker, 10 Mandarin digits and 30 digit sequences per session |
| TL Phonetically Balanced Words | CHTC | 84 | 101 | read | 8 kHz, 16 bits | 50 phonetically balances words per speaker |
| NTU-AS Isolated Syllables | NTU and AS | 6.15 hrs | 117 | read | 16 kHz, 16 bits | 54 sets of 1345 mandarin syllables |
| NTU-AS Isolated Words | NTU and AS | 5.44 hrs | 73 | read | 16 kHz, 16 bits | 44 sub-sets, including 3157 words |
| NTU-AS Phonetically Balanced Sentences A | NTU and AS | 5.56 hrs | 67 | read | 16 kHz, 16 bits | 30 sets of 352 phonetically balanced sentences, covering 1300 mandarin syllables |
| NTU-AS Phonetically Balanced Sentences B | NTU and AS | 2.04 hrs | 47 | read | 16 kHz, 16 bits | 11 sets of 333 phonetically balanced sentences, covering 1300 mandarin syllables |
| NTU-AS Phonetically Balanced Sentences C | NTU and AS | 3.71 hrs | 20 | read | 16 kHz, 16 bits | 20 sets of 260 phonetically balanced sentences, covering 400 mandarin base-syllables |
| AS Phonetically Balanced Speech | AS | 2.5 GB | 96 | read | 16 kHz, 16 bits | each speaker pronounced 85 words,100 sentences, and 5 unrelated paragraphs |
| CCL-200 Syllables and Words | CCL, ITRI | 12 GB | 118 | read | 16 kHz, 16 bits | each speaker pronounced 2 sets of 1266 tonal syllables and 50 words |
| NCTU Continuous Speech | NCTU | 1.5 GB | 60 | read | 16 kHz, 16 bits | each speaker read the news of about 1500 Chinese characters |

Notes: CHTC - Chung Hua Telecommunication Co., TL - Telecommunication laboratories of CHTC, NTU - National Taiwan University, AS - Academia Sinica, ITRI - Industrial Technology Research Institute, CCL - Computer and Communication Research Laboratory of ITRI, NCTU - National Chiao Tung University

## **Appendix C** - File Header of Speech Data File

| Seq. No. | Parameter | Description |
|---|---|---|
| 1 | header_length | code indicating length of file header, 128 bytes or 512 bytes |
| 2 | data_lentgh | number of samples |
| 3 | database_name | code indicating the name of the database |
| 4 | encoding_mode | code indicating the encoding method, 16-bit linear PCM, 8-bit ADPCM, A-law PCM, $\mu$-law PCM, or others |
| 5 | bit_per_sample | number of bits per sample |
| 6 | sampling_rate | number of samples per second |
| 7 | recording_date | date of recording, year-month-day time |
| 8 | recording_time | time of recording, hour-minute |
| 9 | recording_site | code indicating the recording station |
| 10 | speaker_gender | code indicating the speaker's gender |
| 11 | speaker_age | code indicating the range of the speaker's age |
| 12 | speaker_accent | code indicating the speaker's accent |
| 13 | speaker_education | code indicating the level of the speaker's education |
| 14 | speaker_mother_tongue | code indicating the speaker's mother tongue |
| 15 | speaker_daily_language | code indicating the speaker's daily language |
| 16 | speaker_residence | code indicating the speaker's residence |
| 17 | prompting_sheet_ID | ID number of the prompting sheet, 3 digits - 4 digits |
| 18 | prompting_item | prompting item number, 2 digits |
| 19 | speaking_style | code indicating the speaking style |
| 20 | speaking_quality | code indicating the speaking quality |
| 21 | speaking_rate | code indicating the speaking rate |
| 22 | speaking_articulation | code indicating the articulation type |
| 23 | speaking_effort | code indicating the effort of speaking |
| 24 | signal_condition | code indicating the signal condition |
| 25 | signal_quality | code indicating the signal quality |
| 26 | reserved | user defined parameter |
| 27 | assessing_date | date of recently assessing, year-month-day |
| 28 | character_length | number of Chinese characters |
| 29 | Chinese_characters | up to 27 Chinese characters in Big-5 codes |
| 30 | Mandarin_syllables | Pin-Yin of transcript syllables |
| 31 | extended_block | an optional block defined by the user |

**Appendix D** - Sample of Prompting Sheet

# 在台灣之國語語音收集計畫

# MANDARIN ACROSS TAIWAN (MAT)

## 錄音提示卡 第一頁

所需時間約六分鐘

歡迎您參加在台灣之國語語音收集計畫的錄音工作，您的聲音資料，將用以建立語音資料庫，作爲科技研究之用。

請撥電話號碼 _____ ，就可以經由電話開始錄音工作。

＜電腦語音＞歡迎參加在台灣之國語語音收集計畫，您的聲音將錄音，用來建立語音資料，作爲科技研究之用，如果您不願意聲音被錄下，請立刻掛斷電話，如不掛斷，以下錄音即將開始。

＜電腦語音＞請按鍵輸入您的提示卡編號。（見第二頁上端）

（若操作不當或輸入提示卡編號不正確，錄音系統會要求您重新輸入，若三次輸入皆不正確，即終止此錄音程序。）

＜電腦語音＞ 以下請在聽到嗶聲之後回答所問問提：

1. 您日常講的是國語、閩南語、客家語或是其他那種語言？ _____
2. 您母親講的是國語、閩南語、客家語或是其他那種語言？ _____
3. 您是男性或女性？ _____
4. 您現在住在那一縣市？ _____
5. 您的教育程度是小學、國中、高中、專科、大學、或研究所程度？ _____
6. 您是在民國那一年出生？ _____
7. 請問現在時間是幾點幾分？ _____
8. 您有沒有自用汽車？ _____
9. 您是不是在學學生？ _____