MAT-2000 – DESIGN, COLLECTION, AND VALIDATION OF A MANDARIN 2000-SPEAKER TELEPHONE SPEECH DATABASE

Hsiao-Chuan Wang, Frank Seide, Chiu-Yu Tseng, Lin-Shan Lee*

Association for Computational Linguistics and Chinese Language Processing, Taipei
*Philips Research East-Asia, Taipei

ABSTRACT

Mandarin speech data Across Taiwan (MAT) is a project initiated by members of the Association for Computational Linguistics and Chinese Language Processing (ACLCLP) to collect speech data through public telephone networks in Taiwan. Totally over 7000 Taiwanese individuals have provided speech data. The results were released as a series of MAT speech databases to the research community in Taiwan. Two databases, MAT-160 and MAT-400, have been used for the first and second Assessment of Speech Recognition Technique in Taiwan. Now, release preparation of a larger database of over 2000 speakers, called MAT-2000, has been completed. In this joint project conducted by ACLCLP and Philips Research East-Asia, considerable effort has been spent on validating the database to ensure its quality. MAT-2000 consists of over 80 hours of recordings and contains about 640,000 Mandarin syllables in over 140,000 speech files. These speech files are grouped into five sub-databases for different application purposes.

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. Researchers in many countries have produced speech databases of their own languages [Godfrey 1994]. Some of them are designed as multilingual databases [LDC 1996]. Cooperative projects have worked well in European countries [ELRA 1996].

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 telephone speech database of Mandarin Chinese spoken in Taiwan. This project was sponsored by the National Science Council of Taiwan during 1995-1998. Nine universities and research institutes were involved in this three-year project. In total, over 7000 individuals in Taiwan had provided speech data to generate the speech database. A number of subsets of MAT databases have been released for academic research purpose, such as MAT-160, MAT-400, and MAT-2400. In 1999, a joint validation project conducted by Association of Computational Linguistics and Chinese Language Processing (ACLCLP) and Philips Research East-Asia produced a database called MAT-2000, based on MAT-2400. The purpose of the validation was to ensure the quality of the speech database.

2. SPEECH DATA COLLECTION SYSTEM

In this project, we set up nine speech data collection systems in universities and research institutes. Each speech data collection system consisted of a personal computer equipped with a telephone interface card. It allowed speakers to input their voices by using any telephone handset around Taiwan through the public switching telephone network. The input speech signal was sampled in 8kHz with 16-bit linear PCM.

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

(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 speakers, the prompt for speech input, extraction of speech signals, detection of endpoints of utterances, initiation of the file header, and the 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 (telephone line or microphone), and the encoding mode.

(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.

3. SPEECH FILE FORMAT

Each utterance is 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 header carries the entire annotation information, including parameters and transcription. The length of the header is 256 bytes. The file format allows the header to be extended to 512 bytes by attaching an additional 256-byte block to allow extended transcriptions. However, to simplify data processing, we only use 256-byte headers for MAT-2000.

The audio data are stored after the header as an uncompressed sequence of integer values, such that just skipping the header yields access to the audio data without further processing.

3.1. File Header: Parameter Section

The file header was 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]. There are 28 parameters and 3 blocks defined in the file header. The parameters in the 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.

3.2. File Header: Transcription Section

Two more blocks are used to store the graphemic transliteration (Chinese characters) and the phonetic transcription. The maximum number of transliterated Chinese characters is 27. This is enough to pronounce an ordinary sentence.

Chinese characters are represented in Big-5 code. In MAT-2000, all graphemic symbols are represented as double-byte Big-5 codes, including the spelled roman letters and digits. ASCII characters are not allowed. To ensure a one-to-one correspondence with the phonetic transcription, the transliteration contains only actually spoken items; punctuation marks, spaces, or other unpronounced tokens are not allowed. With this, every Big-5 character corresponds to a single phonetic symbol (Chinese syllable or a English letter).

The phonetic symbols are denoted in a tonal pinyin notation (plain ASCII). In MAT-2000, the pinyin set was extended by:

1. the syllables "be1", "pe1", and "fe1" (bopomofo spelling)
2. (spelled) English letters (e.g. "A")
3. an optional two-digit tone syntax.

The latter means that every syllable has a one or two-digit tone marker. In the standard cases, a single digit (1-5) is used (5 stands for the neutral tone). Double digits are used to denote tone sandhi by appending the default tone as well as the actually spoken tone to the base syllable (e.g. "wu32 bai3"). In addition, this notation is used when the tone deviates from the citation form while still being valid (e.g. "gou3 gou31") and in clear cases of neutralization ("ba4 ba45").

3.3. Sampled Data

The sampled data of the speech signals 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.

For the MAT data collection, the sampling rate is 8 kHz, and the audio data is encoded as 16-bit linear PCM in little-endian convention (the lower-order byte is stored first).

4. MATERIAL DESIGN

4.1 Spoken Material

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

1. By design, they cover 407 base-syllables without concerning the tones in Mandarin Chinese. In addition, "be1", "pe1", and "fe1" are contained since quite a few speakers "spelled out" the bopomofo symbols of some syllables.
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.

The prompt sheets also contained 200 different numbers of five different types. A set of examples is shown below. Their pronunciation is transcribed in pinyin shown in parentheses:

1. Digit strings – 118 2720 (yi1 yi1 ba1 er4 qi1 er4 ling2).
2. Dates – 2nd of October (shi2 yue4 er4 ri4)
3. Times – 10:33 am (shang4 wu3 shi2 dian3 san1 shi2 san3 fen1)
4. Prices – 1341 dollars (yi1 qian1 san1 bai3 si4 shi2 yi1 yuan2)
5. Car plates (English spelled letters and digit string) – WB 4522 ("W" "B" si4 wu3 er4 er4)

4.2. 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 has nine questions used to gather information about the speaker, such as the speaker's gender, age, language background, education level, and residence. 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 items no. 10 – 14),
2. 12 isolated Mandarin syllables (items no. 15 – 26),
3. 30 isolated words of 2 ~ 4 characters (item no. 27 – 56),
4. 10 phonetically balanced sentences (items 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.

5. DATABASE DESIGN

Each utterance provided by a speaker is stored as a separate speech file. Sixty-six files correspond to 66 prompting items that are collected from each speaker. As shown in Table 1, the speech data is arranged into five subsets according to the five parts in the prompting sheet described in the previous section.

Table 1. Database subsets.

Subset	Type	Items
MATDB-1	spontaneous, prompting items	1 – 9
MATDB-2	read, numbers	10 – 14
MATDB-3	read, isolated Mandarin syllables	15 – 26
MATDB-4	read, isolated words	27 – 56
MATDB-5	read, sentences	57 – 66

The speech-file editing program VEDITOR and the conversion program VATWAV are provided. The latter is for converting the speech file (.vat) to a standard wave file (.wav).

6. DATABASE VALIDATION

The MAT-2000 release was prepared in a joint database-validation project by ACLCLP and Philips Research East-Asia. Database validation often means checking a database against certain acceptance criteria. In our case, it meant ensuring the quality of the MAT-2000 release, as a combination of checking and correcting. The specific goals were to ensure correctness of annotations (transcriptions and parameters), enhanced consistency (the data was collected at nine different sites), and formal fulfillment of the specification (e.g. pinyin syntax).

MAT-2000 should contain speech data from 1000 speakers per gender. We started with data from 2444 speakers that had already been partly processed: The 9 spontaneous items had been manually transcribed; for the remaining items, initial transcriptions had been inserted from the prompt sheets; and corrections and quality annotations had been made partially.

During our work, we encountered various special cases that required us to refine the specification, either because these cases had not been anticipated, or details were left open. These cases are described in the sections below.

Utterances were specified to be unusable if a syllable was cut in the middle, or the recording was otherwise inappropriate. The latter criterion was also more precisely defined during validation (see below). We also decided that a speaker with over 50% of unusable files was to be marked entirely unusable.

6.1. The Process

The total number of files to process was 163,215, so efficiency was key – saving one second per file would yield an overall saving of 45 working hours. Work was split into two phases:

1. Audit all files and identify case that required correction – as efficiently as possible, using efficiency-optimized tools.
2. Correct these cases and verify corrections.

In phase 1, 15% of all files were identified as problem cases.

6.1.1. Phase 1: Initial verification

First, a speech recognizer did a forced alignment of the speech data to verify leading/trailing silence; silence over 0.5s was cut.

Then, validators checked all files for annotation errors using a special tool based on GNU Emacs (a free text editor popular in the UNIX / Linux world that has a powerful built-in programming language). It displays filenames and annotations (transliteration, transcription, audio-assessment parameters) of all 66 utterances of one speaker in a text-editor window. With a single keystroke, the validator could mark the annotation of an utterance as correct (verified), advance the cursor to the next utterance, and also play this next file back, all in one step.

Thus, processing of most of the initially correct 85% recordings merely required the playback time of the file plus a single key stroke. For the remaining 15% erroneous files, the type of the error (transliteration, transcription, signal-condition annotation, noise annotation, etc.) was marked for further processing using other hot keys. To ensure the correctness of the initial verification stage, 10% was checked again by another person.

6.1.2. Phase 2: Correction

For efficient correction, data was grouped by error type. The most important annotation errors (transliteration) were corrected first. All modifications were verified by a second person. When a second less important error was detected, it was marked, to be fixed and again verified in the next round, and so on. In the most complicated cases, 7 iterations were required.

6.1.3. Automatic specification verification

Automated tools were used to verify pinyin syntax (typing errors were a frequent source of errors) and character/pinyin correspondence. A semi-automatic process was used to check and fix speaker statistics in the headers (from the 9 spontaneous items). However, manual verification was needed since, e.g., some speakers gave obviously wrong gender information.

6.2. Transcriptions

The transcription must represent what was actually spoken. When speakers deviated from the prompt sheet, existing transcriptions (mostly derived from the prompt sheets) were often inaccurate and needed correction. Some cases even fell out of the specification. We encountered the following cases:

1. Sentence read incompletely or with modifications; or transcription was even completely different (e.g. due to mixed-up prompt sheets).
2. For a character with multiple valid pronunciations (e.g. er4 / liang3), the transcribed syllable was not the actually spoken one; or a character's pinyin syllable had been corrected already but in a wrong way or with a typing error.
3. A character is spoken as a valid syllable that is no correct pronunciation of that character (mis-reading; metathesis).
4. Extraneous speech (hesitations, false starts, restarts), but consisting solely of valid Mandarin syllables (see also 7.).
5. Character spelled out in bopomofo ("bang – be ang bang", frequently found in the isolated syllable sub-database, and occasionally for an entire sentence). The bopomofo symbols "be", "pe", and "fe" are not valid Mandarin syllables.
6. Wrong tone transcription due to tone sandhi, neutralization, or some unsystematic but acceptable tone deviation from the citation form.
7. Malformed syllables (e.g. stutter, restart without completing the syllable, or any other abnormal speaking like laughter, speech impediments, whispering etc.)
8. Non-Mandarin speech (e.g. English words; two speakers read the entire prompt sheet in Taiwanese dialect).

Cases 1.–2. are the trivial ones. For case 6., we extended the tone notation as described before. Cases 7. and 8. fall out of the scope of MAT-2000, so such sentences were marked unusable.

Cases 3.–5. were not fully covered by the specification, and existing transcriptions were inconsistent. A refinement of the specification was necessary. For that, our underlying assumption was that MAT-2000 is intended mainly for acoustic model training. All utterances containing valid Mandarin syllables, English letters, or bopomofo spelling, spoken in a correct and fluent way, should be used.

However, whereas a phonetic transcription is possible, graphemic transliteration is problematic in view of other potential uses of the database such as deriving test sets for LVCSR, generating pronunciation lexica, or studying pronunciation variation. For these three cases, an accurate transliteration in Chinese characters is either impossible or would lead to meaningless text and/or incorrect character/pinyin correspondence.

We decided to still keep these sentences, but to replace the corresponding characters by a special character (a star). These sentences can easily be detected and left out where undesirable.

6.3. Signal Condition / Quality Annotation

Recordings with acoustic distortions are useful and valuable for model training, if these effects are likely to occur in the deployment situation. The file header defines parameters for signal condition and quality. However, only one type of distortion can be annotated per file. When two types were present, we had to choose the more predominant one. The original specification also did not cover speaker noise (cough, lip smack, loud breath, breathing on the microphone, clearing throat). We reassigned the broad class “abnormal-noise” for these. For future databases, we suggest to extend this part of the specification.

Recordings with obvious technical recording problems were considered unusable. We found that over 200 speakers were not usable due to, it seemed, recording-software problems like recording buffer overruns, over-amplification, and even wrong sampling rate. Recordings with unusually extreme noise or DTMF tones were also considered unusable, as were recordings containing a second speaker. Some recordings had a 60 Hz power-line hum. If it was located entirely below ~250 Hz, the recording was kept, and the noise was not annotated.

6.4. Validation Results

From the 2444 speakers (163,215 files), 212 speakers were unusable. The remaining 2232 speakers, 1227 female and 1005 male, constitute the MAT-2000 database, comprising 143,266 usable recordings with 641,936 spoken syllables, 83.7h in total.

During validation, 5703 character transliterations were changed (not counting automatic transformations), and 28168 phonetic transcriptions. Table 2 shows the number of substitutions, insertions, and deletions made. Most of the 36541 pinyin substitutions, 29196, were only a change of tone. (974 additional tone changes to double-digit notation without changing the actual tone were not counted as substitutions.) Within the 2232 speakers, 5939 recordings (4.1%) were marked unusable.

Table 2. Changes during validation.

	Subs.	**Ins.**	**Del.**	**%**
Graphemic transliteration	5268	5785	462	1.8%
Phonetic transcription	36541	6369	531	7.0%

We studied the effect of validation on the quality of acoustic models using the 2000 benchmarking task (syllable recognition). We trained our tonal baseline benchmark system [Liao et al. 2000] on (1) the original unvalidated 2444 speakers (omitting sentences originally marked invalid) and (2) the fully validated MAT-2000. Table 3 shows the syllable error rates (SER). Validation led to 1.6% relative SER reduction – small but reasonable since most of our corrections affected the tone only. We also observed a small increase of log-likelihood and a 12% search space reduction, indicating improved acoustic models.

Table 3. Recognition results.

Training set	**Spkrs**	**Sent**	**SER**
(1) Original unvalidated set	2444	126k	31.0%
(2) MAT-2000, fully validated	2232	113k	30.5%

6.5. Lessons Learned

We learned three things. First, impact of validation on system accuracy is small but notable. Second, when designing a database, it is impossible to anticipate the entire variety of acoustic conditions and speaker behaviour. Sharing experience is key. Third, it may seem that some of the technical recording problems and annotation inconsistencies could have been avoided. However, we have to acknowlegde that large multi-site data collections are complex operations with many people with varying experience involved.

We conclude that for future data collections, validation should begin as early as possible, overlapping with the recordings in order to help detect and avoid technical problems, and to make necessary specification refinements at an early stage.

7. ACKNOWLEDGEMENTS

The speech data collection was sponsored by National Science Council of Taiwan, project numbers NSC-85-2213-E-007-043, NSC-86-2213-E-007-045, NSC-87-2213-E-007-031. Funding for the MAT-2000 validation was provided by ACLCLP and Philips Research East-Asia. Authors would like to thank Professors Sin-Horng Chen, Ching-Tang Hsieh, Eng-Fong Huang, Yau-Tarng Juang, Keh-Yih Su, Jhing-Fa Wang, and Ming-Shing Yu for their contributions to the speech data collection.

8. REFERENCES

1. Bernstein, J., K. Taussig, and J. Godfrey, "MACROPHONE: An American English telephone speech corpus for polyphone project," ICASSP'94, Adelaide, Australia, 1994, I-81-I-84.
2. ELRA, ELRA Catalogue release 1.4, December 1996.
3. Godfrey, J., “Polyphone: Second anniversary report,” Notes from the COCOSDA Workshop 94, Yokohama, Japan, 1994.
4. Jones, K. and J. Mariani, (edited), Proc. Workshop of the International Coordinating Committee on Speech Databases and Speech I/O Systems Assessment, Banff, Canada, October 1992.
5. Kudo, I., T. Nakama, N. Arai, and N. Fujimura, “The database collection of Voice Across Japan (VAJ) project,” ICSLP'94, Yokohama, Japan, 1994, pp.1799-1802.
6. LDC, A note of the corpora released by LDC, 1996.
7. Tseng, C. Y., “A phonetically oriented speech database for Mandarin Chinese,” ICPhS'95, Stockholm, Sweden, 1995, vol. 3, pp.326-329.
8. Liao Y.F. et al., “Improvements of the Philips 2000 Taiwan Mandarin benchmark system,” ICSLP'2000, Beijing, China, 2000.