Multi-linguistic Considerations for Chinese Speech Database

Chiu-yu Tseng
Institute of Linguistics (Preparatory Office), Academia Sinica, Taiwan

Introduction

Research in Chinese speech synthesis and speech recognition began nearly two decades ago in Taiwan. Initial efforts mostly came from faculty at electrical engineering departments at universities as well as telecommunication laboratories and industries. From the earlier stages researchers realized that the quality of synthetic speech as well as the correct recognition depended crucially on collected speech database. However, in those years there was no carefully designed speech database. Labs at various universities and research institutions made use of what they could get hold of on hand. Often times speech data with almost no linguistic considerations were solicited from and recorded by graduate students right in each respective labs. These data could hardly be called speech database. The need of a Mandarin speech database was finally recognized when funding became available. .Two major projects were funded separately while overlapping partially in time. One was a 5-year Theme Project (July 1994 till June 1999) funded by Academia Sinica. Another was a 3-year group research project (August 1995 till July 1998) funded by the National Science Council, Taiwan.

Under the Academia Sinica Theme Project, where Mandarin speech database was one of its sub-project, The former is an interdisciplinary Chinese information project aimed at better integrating research in information science and linguistics in order to facilitate better Chinese information research environment in general.. The National Science Council Project MAT (Mandarin across Taiwan) was aimed strictly at speech data collection across Taiwan.

Recent efforts in designing and constructing speech database in Taiwan have begun with Mandarin Chinese and subsequently gone beyond. The present paper addresses the following questions: 1.) the design of a monolingual Mandarin speech database, 2.) the linguistic and phonetic issues considered, 3.) the language engineering concerned, 4.) the machine-readable transcription system designed, and 5.) the expansion from Mandarin to Southern Min (Taiwanese) and Hakka.

1. Design of a monolingual Mandarin speech database,

The monolingual Mandarin speech database includes the following: A phonetically balanced corpus, a phonetically rich corpus, a prosody oriented corpus, and a dialogue corpus. The phonetically balanced corpus consists of 42 hours of digitized speech from 6 speakers (3 males and 3 females),

each producing 7 hours of all possible Mandarin syllables, most 1500 frequently used disyllabic-, tri-syllabic-, and quadra-syllabic- lexical items plus 599 short paragraphs. The phonetically rich corpus consists of 100 hours of digitized speech from 100 speakers (50 males and 50 females) of digitized speech, each a reduced portion randomized from the phonetically balanced set. The prosody oriented corpus consists of digitized speech from 6 speakers (3 males and 3 females), each producing 7 hours of a corpus that includes equal amount of examples from a wide variety of Mandarin sentence types. The dialogue corpus consists of around 500 pieces of digitized telephone conversation of inquiries and reservation of round-trip train ticketing processes.

2. Linguistic and phonetic issues considered

The linguistic issues considered included information at both the phonetic levels and the suprasegmental levels. The solicited information was then utilized to design desired tool software. These various levels of information was then reflected in the tools development. The speech database was tagged and labeled at different layers. For phonetic information, we considered all possible phonetic and tonal combination in order to perform somewhat detailed analysis. The aim was to extract phonetic characteristics so that software could be designed for automatic transcription. Needless to say, trained human transcribers were needed for this stage of work. With regard to suprasengmental information, we considered the majority of prosody types. Detailed analyses were also performed to characterize the prosodic features of Mandarin. Through these studies, we propose that intonation plays a minor role in the prosodic structure whereas prosodic group is the major operating unit of Mandarin Speech.

3. Language engineering concerned

Tools were necessary as the speech data were collected. These tools include software that would automatically transcribe, though somewhat crudely, the actual speech data collected. Using workstations SUN SPARC 10 and a SUN Ultra SPARC, we develop software tools under commercialized software ESPS (Entropic Speech Processing System). Electrical engineers were essential at this part of the project. While we perform analysis in speech science, we also came into a phase of language engineering that require close collaboration between linguists and engineers.

4. Machine-readable transcription system designed

An ASCII encoding of transcription system for speech database of Chinese was designed on the basis of the International Phonetic Alphabet (IPA) to include notations at segmental tonal levels which are phonemic, plus notations at he prosodic levels. The proposed system also aims to transcribe three major Chinese dialects spoken in Taiwan, namely, Mandarin, Taiwanese and

Hakka. Since existing ASCII versions of phonetic transcription systems such as SAMPA appear to aim at transcribing European languages only, they prove to be insufficient to accommodate syllable based tonal languages such as Chinese.

5. Expansion from Mandarin to Southern Min (Taiwanese) and Hakka.

While designing a machine-readable broad phonetic transcription system, we also included elaboration of the proposed system so that it would accommodate Taiwanese and Hakka. The idea was that once the set-up and construction of a Mandarin speech database are completed, we should also be able to use the framework for other Chinese dialects spoken in Taiwan. Expanding it to include Taiwanese and Hakka would be the most feasible next step.

Reference

(1) Chiu-yu Tseng, "A Phonetically Oriented Speech Database for Mandarin Chinese" Proceedings of The XIIIth International Congress of Phonetic Sciences (ICPhS95) (
) (August 13-19, 1995), Stockholm.

(2)   ,   (April 1, 1996)

(3) Chiu-yu Tseng, "Design and techniques to develop a speech database of Mandarin Chinese" The Fifth International Conference on Chinese Linguistics (ICCL-5)(
)(June 27-29, 1996),

(4) Chiu-yu Tseng, "A Database of Mandarin Speech and Its Application" Workshop on Computational Linguistics:Computational Resources for Research in Chinese Linguistic, (June 29, 1996), Academia Sinica, Taipei.

(5) Chiu-yu Tseng, "Prosodic Group: Suprasegmental Characteristics of Mandarin Connected Speech from a Speech Database". The Sixth International Conference on Chinese Linguistics(ICCL-6)(   ) (June 18-21, 1997), Leiden, the Netherlands.

(6) Chiu-yu Tseng, Fu-chiang Chou, "Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan". The First Oriental COCOSDA Workshop, (May 18-19, 1998), Tsukuba, Japan.

(7) Chiu-yu Tseng, "Networking in Mandarin Chinese   Another phase of a Mandarin speech database".   , (May 28-30, 1998),   .

(8)   , <   >,
              , (May 22, 1998),   ,   .