English Across Taiwan
(EAT-ALL)

English Across Taiwan (EAT-ALL) Corpus Description

I. Statement of EAT Project

EAT project prepared 600 recording sheets. Each sheet had 80 reading sentences, including English long sentences, English short sentences, English words and mixed Chinese-English sentences etc. Five academic affiliations joined this project. Each affiliation finished 120 reading sheets. Each sheet was used for speech recording individually for English Department people and non-English Department people. The recording using hand-held microphone and wire/wireless telephone was fulfilled. Microphone corpus was recorded as sound files with 16 kHz sample rate and 16 bit sample resolution. Telephone corpus was recorded as sound file with 8 kHz sample rate and 16 bit sample resolution. Telephone corpus was divided into 600 copies (English Department + non-English Department) of PSTN corpora and 600 copies (English Department + non-English Department) of GSM corpora. We summarize the recording sheets as follows:

600 copies of recording sheets:

- 600 English Department people (number of recording sheets: 101000-101599)
  - Microphone Corpus
    - 600 copies (recorded by student or by affiliation)
  - Telephone Corpus
    - 300 copies of PSTN corpus (recorded by affiliation)
    - 300 copies of GSM corpus (recorded by toll free phone recording system, its phone number is 0800351151)
- 600 non-English Department people (number of recording sheets: 100000-100599)
  - Microphone Corpus
    - 600 copies (recorded by student or by affiliation)
  - Telephone Corpus
    - 300 copies of PSTN corpus (recorded by affiliation)
    - 300 copies of GSM corpus (recorded by toll free phone recording system, its phone number is 0800351151)

Each affiliation had 120 copies of PSTN recording sheets and 120 copies of GSM recording sheets. PSTN speech data were collected using the recording station setup at each affiliation. GSM speech data were collected by the recording station setup at Industrial Technology Research Institute (ITRI). Totally, there were 240 copies of microphone recording sheets, 120 copies of PSTN recording sheets and 120 copies of GSM recording sheets. Numbers of recording sheets were distributed as follows:

Distribution of Recording Sheets

- National Taiwan Normal University: (100000-100119, 101000-101119)
  - Berlin Chen
- National Chiao Tung University: (100120-100239, 101120-101239)
  - Sin-Horng Chen and Yih-Ru Wang
- National Tsing Hua University: (100240-100359, 101240-101359)
  - Jason S. Chang and Jyh-Shing Roger Jang
- National Cheng Kung University: (100360-100479, 101360-101479)
  - Jen-Tzung Chien
- National Taiwan University: (100480-100599, 101480-101599)
  - Lin-shan Lee

II. Recording Equipments and Environments

EAT corpus was divided into telephone speech and microphone speech. Telephone speech were recorded by DIALOGIC card with the specifications of Mu law decoding, 8 KHz sampling rate and 8 bit sample resolution. Speech data were converted to the PCM format with 8 KHz sample rate and 16 bit sample resolution. A sound file with .wav format contained all sampling points. Each affiliation prepared microphone and personal computer to acquire hand-held microphone speech. Microphone speech in .wav format was recorded by the sound card in personal computer with 16 KHz sampling rate and 16 bit sample resolution. All sound files are all in raw format. No dc-offset and silence removal were performed.

III. EAT Corpus Statistics

EAT corpus has been collected since May 2004. The recording was finished by January 2005. Corpus was checked and annotated by CCL/ITRI. After checking, this corpus was divided into two classes, *usable* and *unusable*, in accordance with the sound quality and the correctness of reading contents. Usable corpus was further divided into two sub-groups, English Department and non-English Department. We also divided the corpus into male and female groups. Corpus was classified into PSTN, MIC and GSM groups. The statistics was given as follow:

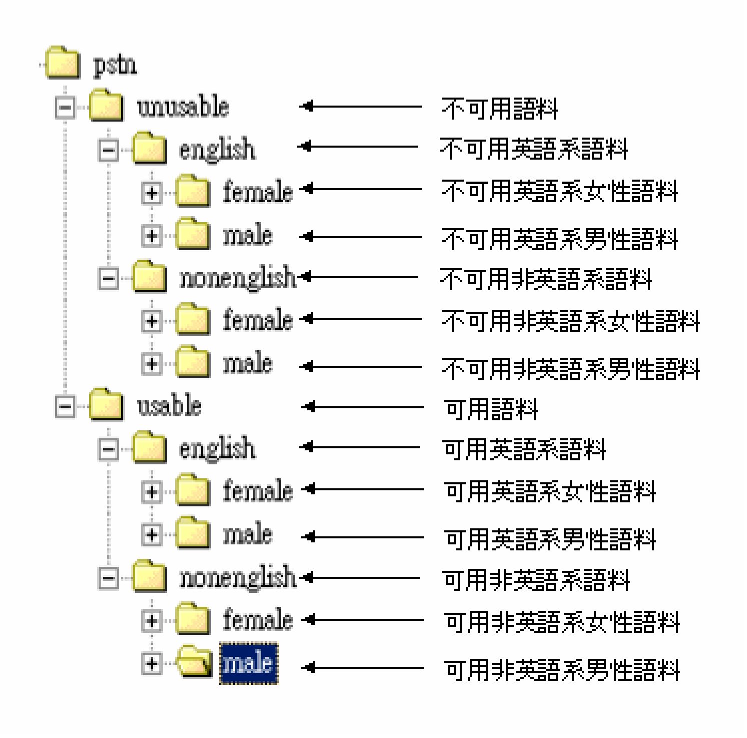
MIC16K corpus				
	Usable			
	English Department		non-English Department	
	Male	female	male	female
number of sentences	11977	30094	25432	15540
number of persons	166	406	368	224

	GSM corpus			
	Usable			
	English Department		non-English Department	
	Male	female	male	female
number of sentences	6168	15681	12721	8048
number of persons	85	216	192	122

PSTN corpus				
	Usable			
	English Department		non-English Department	
	Male	female	male	female
number of sentence	5582	14244	10584	6685
number of person	82	206	160	103

IV. Description of EAT-ALL Corpus DVD

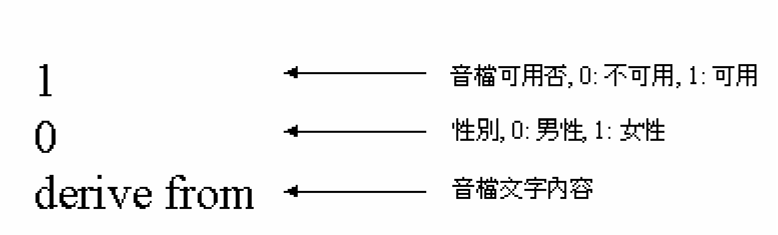
EAT corpus containing three groups of channels: PSTN, MIC16K and GSM was stored in three DVD discs. PSTN and GSM corpora were stored in the same DVD disc which is label as “PSTN +GSM”. Because the sampling rate of MIC16K speech data was high, the resulting storage requirement was huge. We stored MIC16K speech in two DVD discs labeled by “Mic16K English” and “Mic16K NonEnglish” for English Department and non-English Department, respectively. For example, the “PSTN +GSM” DVD disc had the file structure as follows:



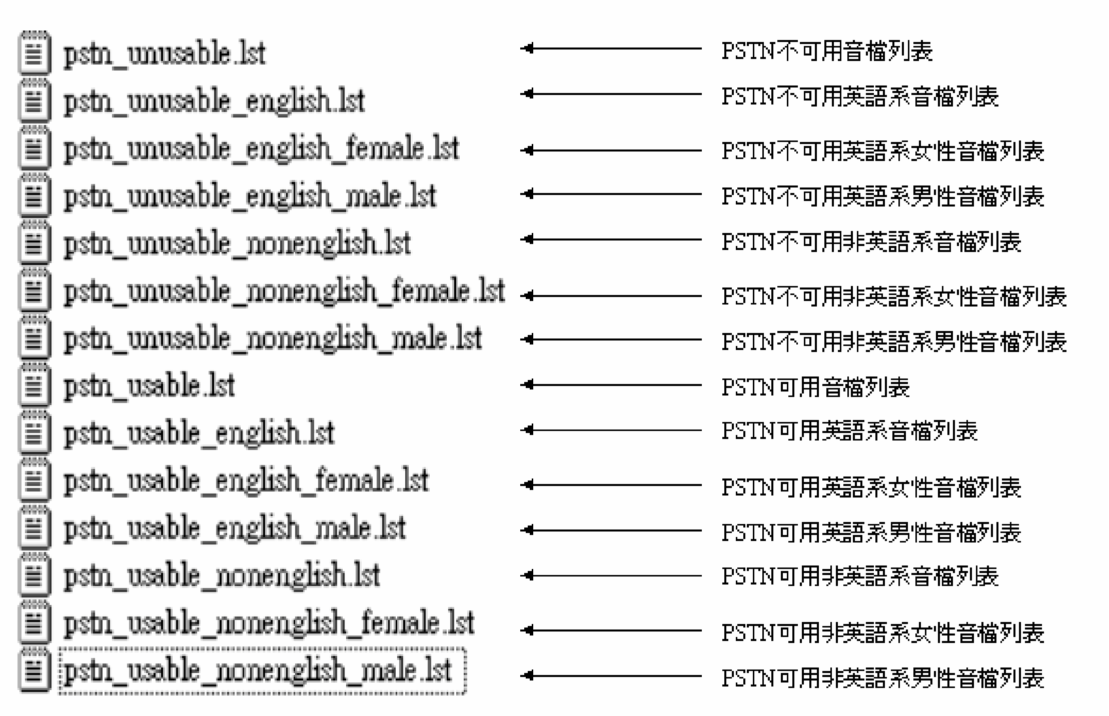
For male and female groups, we put sound files corresponding to one recording card into one file folder. Every recording card had its own file folder. The sound file (.wav) and the sound content label file (.lab) were stored into its own file folder. Sound files (.wav) were in standard format of Windows wave file. Size of file header was 56 bytes. Sampling rates were different for different recording channels. Sample resolution 16 bit was the same for all files.

PSTN: 8 KHz, 16 bits
GSM: 8 KHz, 16 bits
MIC16K: 16 KHz, 16 bits

There were three description labels in every sound file. Their formats are described as follows:



In order to conveniently fetch all sound files, we stored the file list of all sound files under the root directory of each DVD disc. We take the PSTN DVD disc as an example. It had the file list (.lst file) as follows:



pstn_unusable.lst	PSTN不可用音檔列表
pstn_unusable_english.lst	PSTN不可用英語系音檔列表
pstn_unusable_english_female.lst	PSTN不可用英語系女性音檔列表
pstn_unusable_english_male.lst	PSTN不可用英語系男性音檔列表
pstn_unusable_nonenglish.lst	PSTN不可用非英語系音檔列表
pstn_unusable_nonenglish_female.lst	PSTN不可用非英語系女性音檔列表
pstn_unusable_nonenglish_male.lst	PSTN不可用非英語系男性音檔列表
pstn_usable.lst	PSTN可用音檔列表
pstn_usable_english.lst	PSTN可用英語系音檔列表
pstn_usable_english_female.lst	PSTN可用英語系女性音檔列表
pstn_usable_english_male.lst	PSTN可用英語系男性音檔列表
pstn_usable_nonenglish.lst	PSTN可用非英語系音檔列表
pstn_usable_nonenglish_female.lst	PSTN可用非英語系女性音檔列表
pstn_usable_nonenglish_male.lst	PSTN可用非英語系男性音檔列表

The following was one example of the content in the .lst file

```
pstn/usable/english/male/100060/10006001.wav
pstn/usable/english/male/100060/10006002.wav
pstn/usable/english/male/100060/10006003.wav
pstn/usable/english/male/100060/10006004.wav
pstn/usable/english/male/100060/10006005.wav
pstn/usable/english/male/100060/10006006.wav
pstn/usable/english/male/100060/10006007.wav
pstn/usable/english/male/100060/10006008.wav
pstn/usable/english/male/100060/10006009.wav
pstn/usable/english/male/100060/10006010.wav
pstn/usable/english/male/100060/10006011.wav
pstn/usable/english/male/100060/10006012.wav
```