

台灣口音英語語料庫說明
English Across Taiwan
(EAT)

台灣口音英語(English Across Taiwan, or EAT)語料庫說明

一. EAT 錄音計畫說明

EAT 錄音計畫共發出 600 份錄音提示卡,每份提示卡皆含 80 個錄音句,其中包含英文長句,英文短句,英文單詞及中英夾雜句等.

600 份提示卡由五個單位合力完成錄音,每個單位負責 120 份提示卡,而每一份提示卡需分別由英語系及非英語系學生各錄製一份,每一學生需錄製麥克風語料及電話語料各一份,麥克風語料錄製 16khz 取樣頻率 16bits 的取樣點音檔,電話語料錄製 8khz 取樣頻率 16bits 的取樣點音檔.其中電話語料又可細分為 600 份(英語系+非英語系)固定式電話(PSTN)語料 及 600 份(英語系+非英語系)行動電話(GSM)語料,歸納如下表所列,

600份提示卡

- 600個英語系學生(提示卡編號100000-100599)
 - 麥克風語料
 - 600份(發給學生自行錄製或集中錄音)
 - 電話語料
 - PSTN語料300份(各校架站收集)
 - GSM語料300份(統一撥至0800351151收集)
- 600個非英語系學生(提示卡編號101000-101599)
 - 麥克風語料
 - 600份(發給學生自行錄製或集中錄音)
 - 電話語料
 - PSTN語料300份(各校架站收集)
 - GSM語料300份(統一撥至0800351151收集)

各單位將負責收集 120 份 PSTN 及 120 份 GSM 的語料,其中各單位的 PSTN 語料將由各單位自行架錄音站收集,而 GSM 語料則統一由工研院的語料錄音站收集.每個單位錄音完成後,計含 240 份麥克風語料,120 份 PSTN 及 120 份 GSM 語料.各提示卡號的分配如下:

提示卡分配

- 師大: (100000-100119, 101000-101119)
 - 陳柏琳老師
- 交大: (100120-100239, 101120-101239)
 - 陳信宏老師, 王逸如老師
- 清大: (100240-100359, 101240-101359)
 - 張俊盛老師, 張智星老師
- 成大: (100360-100479, 101360-101479)
 - 簡仁宗老師
- 台大: (100480-100599, 101480-101599)
 - 李琳山老師

二. 錄音設備及環境

EAT 語料分為電話及麥克風語料,電話語料部份是透過 Dialogic 電話語音介面卡,以所錄得的 8KHz,8Bits,Mulaw 格式的取樣點,經程式轉成 8khz, 16bits,pcm 格式的取樣點,然後將所有取樣點存放一.wav 格式的音檔,麥克風語料則是由各錄音單位所準備的個人電腦及麥克風,直接從 pc 的音效卡錄製 16khz,16bits 的聲音訊號,然後將所有取樣點存成一.wav 格式的音檔.

注意: 所有音檔內容皆屬於 raw 格式,也就是沒有經過 dc-offset 及 silence removal 的處理.

三. EAT 語料統計

EAT 語料從 2004 年 5 月開始收集,至 2005 年 1 月初步完成收集,從各單位回收之語料經由工研院電通所匯整並請專人做語料庫整理,整理後之語料依音檔之品質及所唸內容之正確性分為可用(usable)及不可用(unusable)兩大類,可用之語料再依英語系及非英語系細分,然後再依性別做最後的分類,綜合所得的語料,依 PSTN,MIC 及 GSM 分類,得到如下的統計結果:

MIC16K語料				
	可用			
	英語系		非英語系	
	男性	女性	男性	女性
句數	11977	30094	25432	15540
人數	166	406	368	224

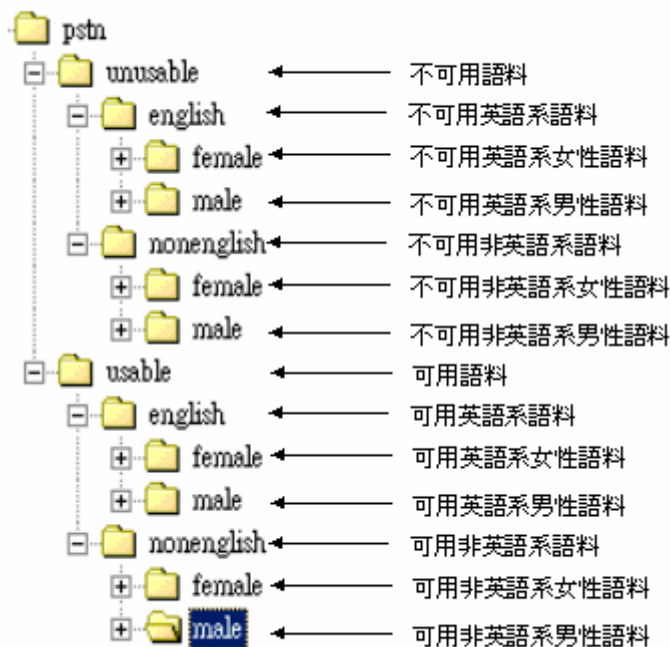
GSM語料				
	可用			

	英語系		非英語系	
	男性	女性	男性	女性
句數	6168	15681	12721	8048
人數	85	216	192	122

PSTN語料	可用			
	英語系		非英語系	
	男性	女性	男性	女性
句數	5582	14244	10584	6685
人數	82	206	160	103

四. EAT 語料庫光碟說明

EAT 語料,依 PSTN,MIC16K 及 GSM 三組不同的 CHANNEL 共存放在三張 DVD 光碟中,其中 PSTN 及 GSM 語料放在同一張光碟中並且標示為 PSTN+GSM, MIC16K 語料因 SAMPLING RATE 較高,語料量較大,故依英語系及非英語系分開存放在兩張不同的光碟上,分別標示 Mic16K English 及 Mic16K NonEnglish,以 PSTN+GSM 光碟為例,下面是該光碟目錄結構中有關 PSTN 部份的說明,



而在各性別目錄下的語料,則依提示卡號,每一個提示卡號皆有一個存放目錄,在提示卡號目錄下則存放聲音檔(.wav)及聲音內容標示檔(.lab),其中.wav 為標準的

windows wave 檔格式,其檔頭大小為 56 bytes,而 sampling rate 依不同 channel 有不同的 sampling rate,取樣點則皆為 16Bits 解析度

PSTN: 8KHz, 16Bits















GSM: 8KHz, 16Bits

MIC16K: 16KHz, 16Bits

而每一個聲音內容標示檔內皆有三列,其格式說明如下:

1	←	音檔可用否, 0: 不可用, 1: 可用
0	←	性別, 0: 男性, 1: 女性
derive from	←	音檔文字內容

爲了方便取所有的音檔,在每張光碟根目錄下,存放所有音檔的列表,以 PSTN 光碟片爲例有如下的列表檔(.lst 檔)

 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

以下是一個.lst 檔的範例,

```
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 :