中研院具音段標記之中文對話語音資料庫

(Sinica Phone-aligned Chinese Conversational Speech Database**)**

一 資料庫內容

聲檔大小： 1 GB (.wav)
聲檔長度： 3 小時 24 分鐘
中文字數： 5 萬 (52K)
音段總數： 12 萬(125K)
文字轉記： PRAAT 格式 (.TextGrid)

二 語料錄製與處理程序

本資料庫為 Sinica MCDC8 漢語口語對話語料庫之子集。在錄製 Sinica MCDC8 時，從錄音地點到錄音設備，盡量以語者在對談過程中感覺自然為原則。語者人選是中央研究院調查研究工作室由台北市市民中隨機抽樣選出後配對錄音。採用 SONY TCD-D10 Pro II DAT 數位錄音機與使用 Audio-Technica ATM 33a 手持式麥克風。兩位語者分錄於左右聲道。錄音地點為普通房間。

本資料庫音節邊界為人工標記結果。音段邊界由六名標記員經訓練達成一致率門檻後，分為三組各標記三分之一的語料。每組兩名標記員之標記結果，取其中間值作為最後資料庫之音段邊界。本音段標記計畫假設每個典型音段都存在。因此，在音段標記層所有典型音段都會出現。如有極度弱化或刪減音段，可由其時長判別。如遇弱化嚴重之詞語，如音節連併，則不做音段標記，在音段層會出現詞語的文字轉寫。以 V_音段轉寫者，則表示該音段發音明顯與典型有異。以[漢字]轉寫者，表示該語音內容以非 Mandarin Chinese 發音。

三 資料庫格式

檔案命名原則： PA_聲檔編號_語者性別_語者年齡_IPU 序號
檔案分割單位： 兩個停頓之間的語流單位 (IPU)
標記層名稱： 音節 (Syllable) 與音段 (Phoneme)
非漢字標記內容：

(1) 無法明確辨識內容之語音：UNCERTAIN
(2) 個人隱私之內容，語音清空：WHITE_NOISE
(3) 非語音之言語現象，例如停頓或吸氣聲：BREAK/INHALE
(4) 填充詞以大寫英文字母轉寫，例如 MHM
(5) 語氣詞，不論其是否有慣用的漢字，皆以大寫英文字母轉寫，例如 A/LA
(6) 雖可以漢字轉寫，但主要為言談功能之言談詞，亦以英文字母標示，例如 NA GE

詳細訊息請參考下列文獻 Tseng (2013)

四 音段符號對照表

本資料庫所使用之音段標記(SPCCSD)與國際音標符號(IPA)之對照表如下。

IPA	p	p^h	t	t^h	k	k^h
SPCCSD	b	p	d	t	g	k
IPA		m		n		ŋ
SPCCSD		m		n		ng
IPA		f	s	ʂ	ɕ	x
SPCCSD		f	s	s`	sj	h
IPA	ts	ts^h	tʂ	$tʂ^h$	tɕ	$tɕ^h$
SPCCSD	dz	ts	dz`	ts`	dj	tj
IPA		l	ʐ	j	w	
SPCCSD		l	Z`	j	w	
IPA	i	y	ɨ	ɯ		u
SPCCSD	i	y	U'	U`		u
IPA	e	a	ə	ɚ		o
SPCCSD	e	a	@	er		o
IPA	ai	au	ei	ye		ou
SPCCSD	ai	au	ei	ye		ou

五文獻引用

使用本資料庫所獲致之研究成果，應引用以下文獻。

Tseng, S.-C. 2013. Lexical coverage in Taiwan Mandarin conversation. *International Journal of Computational Linguistics and Chinese Language Processing* 18(1): 1-18.