2018年1月　第97卷 第1期
January 2018　Volume 97 Number 1

基於言語產生的粵語語音多模態資料庫建設

侯興泉、楊　鋒

暨南大學

提要

隨著多模態的研究方法越來越多地被運用到現代語音學的研究當中，語料庫語言學亦已開始進入多模態語料庫語言學時代。設計並構建粵語言語語音多模態資料庫是適應新時代粵語語音研究的必然要求。同步採錄的多模態粵語語音資料有益於粵語的語音生理研究、言語交際、言語工程、口傳文化、病理治療與康復等薄弱領域的發展，同時也有利於粵語研究與相關學科的交流與融合。

關鍵詞

粵語，多模態，語音資料庫

1. 前言

1.1. 粵語語音資料庫建設的現狀

粵語語音資料庫的建設是粵語語音學研究、言語科學與言語工程以及跟粵語語音相關的各個領域研究和應用的基礎。目前已建好的粵語語音資料庫主要有兩大類型：（1）面向語言學的粵語音檔庫；（2）面向言語工程的粵語語音資料庫。下面我們分別對這兩類語料庫進行一個簡單的介紹。

1.1.1. 面向語言學的粵語音檔庫

粵語音檔庫主要是為粵語的教學與研究服務的。最早的粵語音檔是李新魁、陳慧英、麥耘、方舟（1995）編著的《廣州話音檔》，該音檔包括了一本書和一盒磁帶，收錄了廣州話音系、764 個常用字、180 個詞目、55 條語法例句以及若干長篇語料的文字及錄音。邵慧君、秦綠葉（2008）主持了九十多個粵方言點的單字音語料庫的建設工作，該庫主要對粵語的 3600 個左右的字音進行錄音和國際音標標注。該庫據悉已完成預設的建設任務，但尚未開放給公眾使用。相關的粵語音檔還有劉新中（2014）的專著《廣州話單音節語圖冊》（附光碟），該圖冊除了包括廣州話各個音節的語圖和相應的輔音、元音和聲調的相關聲學參數外，還配有所有音節的錄音，因此可看作是一個廣州話的單音節音檔庫。

此外，粵語學界還建立了許多根據口語語音材料轉寫而成的口語語料庫，由於這些語料庫給公眾檢索使用的主要是轉寫後的文本材料而非原始的語音素材，故還不能算是嚴格意義上的語音資料庫，這裡不作具體的介紹。

1.1.2. 面向言語工程的粵語語音資料庫

目前已建成的面向言語工程的粵語語音資料庫基本都是單一的聲學資料庫。除了帶粵語口音的漢語言語資料庫（CACSC）是由清華大學電腦科學與技術系開發外，其他的粵語語音資料庫基本上都是由香港中文大學電子工程系開發的，包括CUCorpora、CUCall、CU2C、CUMIX 等。Li, Zheng, Xu, Song and Fang（1999）、Lo, Chow, Lee and Ching（1998）、Lo, Ching, Lee and Meng（2001）、Zheng, Qin, Lee and Ching（2010）、Chan, Ching and Lee（2005）對以上幾個資料庫皆有較為詳細的介紹。粵語的語音辨識、合成乃至自然言語處理之所以取得較為顯著的成績，跟這些資料庫的建設及相關研究有著密不可分的聯繫。

1.2. 國內外多模態語料庫的建設和研究概況

隨著現代科學技術的不斷發展，近十多年來在國際語料庫學界誕生了一個新的研究方向——多模態語料庫語言學。基於不同感覺系統的材料、重點面向言語交際的多模態語料庫的建設受到越來越多領域專家的關注。Knight（2011）較為詳細地介紹了國外目前已經建成的十多個大小規模不一的多模態語料庫。據筆者所知，國內目前已建好的多模態語音庫有中國社會科學院的“現代漢語現場即席話語多模態語料庫”和北京大學中文系語言實驗室的“普通話語音多模態資料庫”，兩者都尚未開放給公眾使用。

除了建設多模態語料庫外，國內外學者在多模態語料庫建設的理論和方法依據，材料的收集、轉寫、標注和分析等方面也湧現了不少的成果，比較突出的如 Bernsen and Dybkjær（2007）、Allwood（2008）、孔江平（2008）、Knight（2009, 2011）、Kipp, Martin, Paggio and Heylen（2009）、Thompson（2010）、黃立鶴（2015）等人的研究成果。儘管成果不少，但是迄今為止國際上還沒能形成一套大家都能接受的多模態語料庫建設的理論和方法。

正如 Gu（2006）所說，目前國內外主要有兩類學者對多模態語料庫感興趣：一類是對多模態話語分析感興趣的學者，另一類主要是對提高人機交互效率感興趣的學者。前者（主要來自語用學界，也有部分言語工程學者）主要關注跟言語聲音有關的肢體或手勢，以及跟言語使用有關的各種社會文化背景。後者（主要來自現代語音學和言語工程學界）主要對言語產生過程各種不同模態或類型的信號感興趣，借此弄清言語產生的機制。

言語產生的過程本身就是多模態的。隨著現代語音技術發展的日臻成熟，當前國內外對語音學的研究已呈現多元化和多模態化的趨勢，粵語語音的研究也應迎頭趕上，向多模態研究領域邁進。本文正是在這樣的研究背景下，探討粵語語音多模態資料庫的設計、建設及應用等問題。

2. 粵語語音多模態資料庫的設計

2.1. 資料類型

本資料庫需要採集的資料包括粵語權威方言點及粵語下屬主要方言點的語音材料。語料風格類型分為朗讀語音、口語語音、藝術語音三大類，其中朗讀語音和口語語音為必錄資料，藝術語音為選錄資料。

朗讀語音部分選取各個粵方言點的所有單音節，覆蓋每個方言點聲韻調的所有組合。雙音節包括每個方言點的各種聲調組合（如廣州話有 81 組雙音節聲調組合）。句子根據句式和焦點分佈情況每個點選取 50 個左右。短文文體分為新聞、小說、散文、詩詞四類，每個方言點約錄 30 篇，詩詞每篇字數 20-100 左右；其餘文體每篇字數 500-1000 左右。

口語語音的自由交談由兩位發音人共同完成，交流的話題至少有兩個，每個話題錄製 15 分鐘左右。話題自述部分讓發音人從 12 個話題中選擇三個，每個話題錄製 10 分鐘左右。

藝術語音部分為選錄資料，發音人可自由選擇講故事（說書）、相聲、吟誦或歌曲中的一項來進行，我們會給每種類型的藝術語言各提供三個樣本，發音人也可以不根據樣本而選擇自己熟悉的項目來進行錄製。每類資料的錄音時長控制在 10-20 分鐘之間。

2.2. 採錄儀器及信號類型

資料庫使用 AD Instruments PowerLab PL3516 十六通道高速記錄儀同步採錄 4 個通道的信號：第 1 通道為通過麥克風和調音台採集的語音信號，第 2 通道是通過電子聲門儀（EGG）採集的嗓音信號，第 3 通道為通過呼吸帶感測器採集的胸呼吸信號，第 4 通道為腹呼吸信號。採樣頻率均為 40kHz。錄音使用的話筒是 Sony ECM-44B，調音台是 Behringer XENYX502。電子聲門儀是 KAY 公司的 6103。胸腹兩根呼吸帶是 AD Instrument 公司生產的 MLT1132。同時使用一至三台高清數位影像錄影機（機型為 Sony NEX-VG30）同步錄製這三類語料的視頻圖像資訊。

2.3. 發音人要求

每個粵方言點至少錄製 6 位發音人的語音資料，要求所有的發音人都在本地出生並以當地粵語為第一語言。其中男女至少各 3 人。老中青每個年齡層至少各 2 人：青年的年齡段為 16-30 歲；中年的年齡段為 31-60 歲；老年為 60 歲以上。對發音人的受教育程度、職業、出生地等背景資訊不做嚴格的控制，但儘可能兼顧，不讓某一方面的發音人過於集中。

表一 粵語語音多模態資料庫的語料及信號類型

<table>
<tr><th>錄製要求</th><th>語料種類</th><th>內容</th><th>信號</th><th>用途</th><th>人數</th></tr>
<tr><td rowspan="5">必錄</td><td rowspan="3">朗讀語音</td><td>單、雙音節</td><td>語音、嗓音、圖像</td><td>元音、輔音、聲調、變調、唇形和發聲分析</td><td rowspan="9">每個方言點至少 6 位發音人：男女各 3 人；老中青各 2 人</td></tr>
<tr><td>句子</td><td rowspan="4">語音、嗓音、圖像
胸呼吸、腹呼吸</td><td rowspan="4">語調、語篇、韻律節奏
輔助語言、發聲分析</td></tr>
<tr><td>不同文體短文</td></tr>
<tr><td rowspan="2">口語語音</td><td>話題自述</td></tr>
<tr><td>自由交談</td></tr>
<tr><td rowspan="4">選錄</td><td rowspan="4">藝術語音</td><td>講古（說書）</td></tr>
<tr><td>相聲</td></tr>
<tr><td>吟誦</td></tr>
<tr><td>歌曲</td></tr>
</table>

3. 粵語語音多模態資料庫的構建

跟單一模態的文本語料庫或聲學資料庫比較起來，多模態資料庫的構建無論是在資料獲取、資料處理抑或是在資料檢索等方面都要比單模態資料庫複雜得多。首先，要利用多種不同的設備同步採集多種不同類型的資料，採集資料的軟硬體要求要比單模態資料高得多。其次，資料的處理更加複雜，目前國際上還沒有通用或標準化的多模態處理軟體可供使用，各家多根據自身的研究需求來處理資料。最後，不同模態的語料檢索也要比單模態資料庫困難。有鑑於此，我們對多模態粵語語音資料庫的構建分三個階段來開展：第一階段，按照前面第 2 節的設計採集資料，資料經過初步降噪、剪輯處理後存放好，建成一個多模態的粵語語音資源庫或毛語料庫；第二階段，對採集來的視頻圖像、聲學及生理資料分別進行轉寫、標注，並提取相應的參數，建成多模態的粵語語音參數資料庫；第三階段，開發相應的檢索軟體，開放給研究機構或社會公眾使用。下面我們分別對這三個階段的工作進行詳細說明。

3.1. 資料獲取

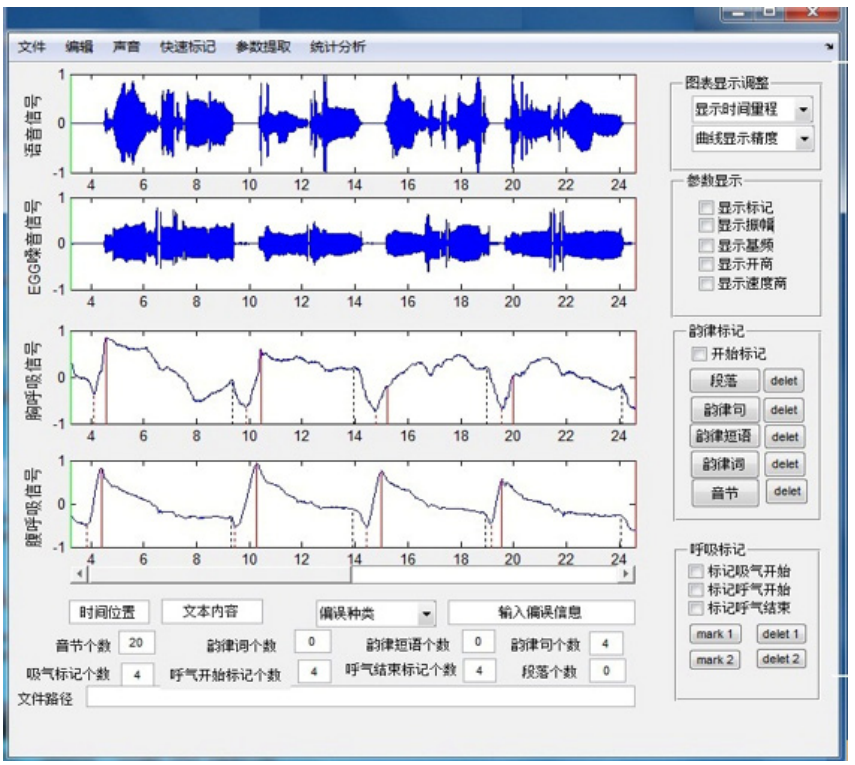
利用胸腹呼吸帶、喉頭儀（EGG）、麥克風、高清數位影像錄影機來同步錄製粵語主要方言點的語音、視頻、嗓音、胸腹呼吸信號，並記錄發音人的相關資訊。錄製完成之後首先使用 PowerLab 記錄儀和 Chart 軟體將信號保存為 adicht 格式檔，然後轉換為 wav 格式的四通道語音檔，並進行濾波、降噪、編輯等預處理。視頻資料跟通道資料通過時間同步對齊後統一存放在每一個發音人對應的文檔目錄下。

3.2. 資料處理

錄製的資料主要有聲學信號、生理信號和視頻圖像信號三大類型，處理的時候把視頻圖像信號單獨進行處理，聲學和生理信號則使用團隊自行開發的軟體來同步進行處理。因為資料獲取結束，在存檔的時候已經利用時間進行了同步，因此最後這三大類型的信號都可以通過時間來進行關聯。

3.2.1. 聲學和生理資料的處理

圖一　粵語語音標注和分析程序的界面



由於目前國際上還沒有見到可以同步分析嗓音、呼吸以及語音聲學信號的軟體，因此我們利用自主開發的程序對四個通道不同的資料進行標記和分析，[1] 程序界面如圖一所示。具體流程如下：

首先，對不同類型的信號進行標注。由於聲學信號和嗓音信號有整齊的對應關係，而呼吸信號跟聲音信號的對應規律則比較複雜，因此需要分別以聲學信號和呼吸信號為參照來進行標注。第一步是以聲學信號為參照，按照語音韻律層級（話語－語調短語－韻律短句－韻律詞－音步－音節）對錄製的資料進行標注，除了單音節的語音信號需要在音節之下進一步標注出聲母和韻母外，其他語音信號一律只標注到音節。第二步以胸腹呼吸信號為參照來進行標注，分別標記出胸腹吸氣開始時間點、胸腹呼氣開始時間點和胸腹呼氣結束時間點。

第二，對標注過的語音信號進行拼音和文本轉寫。轉寫以聲學信號為參照，完成跟音節對應的拼音和文本資料登錄後，在第一欄波形的上下位置顯示拼音和文本內容。

第三，提取相應的語音聲學參數。提取跟音節輔音、母音、聲調相關的聲學參數，諸如嗓音起始時間（VOT）、共振峰、基頻、時長等信息。

第四，提取相應的嗓音發聲參數。根據嗓音 EGG 信號提取不同韻律層級語音單位的開商、速度商和基頻參數，對於特殊的發聲方式進行標記，如緊嗓音、氣嗓音等。

第五，提取呼吸參數。根據前面標記的吸氣開始時間點、呼氣開始時間點、呼氣結束時間點，計算出呼吸重置幅度、吸氣段時長、呼氣段時長、吸氣段斜率、呼氣段斜率、吸氣段面積、呼氣段面積等參數。

以上所有的 wav 檔與該檔的所有標記資訊和參數共同存放在一個 mat 格式的檔中，同時按類別分別輸出語音、嗓音、呼吸參數資訊，存放在 Excel 表中。

3.2.2. 視頻影像處理

視頻圖像資料不是本資料庫的處理重點，大多情況下只作為聲學和呼吸信號的輔助觀察信號來使用。但是我們會重點處理跟粵語語音密切相關的唇形資料。由於唇形資料標注和提取參數比其他資料更為複雜，也更消耗時間，一般情況下只選取其中最有代表的一名男性和一名女性發單音節時的正面視頻資料為研究物件，進行標注並提取其二維唇形的參數。其餘資料留待研究需要時再做相應的處理。

二維唇資料的處理按以下流程開展：（1）切分音節；（2）對錄影畫面有效資料

[1] 團隊成員楊鋒博士共開發了兩個軟體標注及分析系統，分別是：（1）言語呼吸分析系統（軟著登記第 0393007 號）；（2）韻律標記系統（軟著登記第 0409637）。

進行裁剪；（3）對人臉大小資料進行歸一化處理；（4）建立二維唇模型；（5）提取二維唇參數；（6）資料存放。具體提取和存放的二維唇參數包括外唇寬度、內唇寬度、上唇外輪廓開口度、下唇外輪廓開口度、上唇內輪廓開口度、下唇內輪廓開口度、唇角閉合曲線開口度、人中凹陷程度、下唇圓弧度、頭部傾斜程度和歪嘴程度等。詳細介紹參看潘曉聲（2011）對普通話單音節二維唇參數的提取和分析。

3.3. 資料的管理和檢索

所有提取好的參數都存放在我們專門開發的多模態粵語語音資料庫當中。為了方便檢索，後期還將開發專門的檢索軟體。檢索軟體主要設置的檢索條件有：性別（男女）、年齡（老中青）、參數類型等，可以按要求輸出符合條件的 wav 語音檔、聲學參數、嗓音參數、呼吸參數、文本、圖像等資訊。

4. 基於粵語語音多模態資料庫的相關研究

4.1. 粵語語音的生理研究

從 1.1. 的介紹我們不難看出，目前已建好的粵語語音資料庫基本都是聲學資料庫，大部分的研究成果也主要集中在語音聲學領域。粵語語音生理方面的研究成果則相對較少。粵語語音多模態資料庫同步採集了多種類型的生理資料，包括用呼吸帶採集的胸呼吸和腹呼吸資料，用喉頭儀採集的嗓音資料，以及用高清數位影像錄影機採錄的唇形資料。這些生理資料對於研究粵語語音的呼吸韻律機制、嗓音發聲機制以及唇形搭配機制有著重要的作用。由於所有的生理資料都是同步採集的，我們還可以利用這些資料研究呼吸與嗓音、唇形乃至聲學語音的搭配及相互作用，這對我們弄清粵語的發音原理無疑會起到很大的促進作用。目前我們已利用該資料庫開展過多項粵語語音的生理研究，如李煥哲、曹慶松、宋秀豹、吳南開、侯興泉（2015）基於 EGG 的粵語母語者病理嗓音與正常嗓音比較研究，Hou, Deng and Yang（2017）對粵語吟誦七言詩的呼吸韻律研究，鄧德崇（2017）對粵語吟誦和朗讀格律詩詞的語音和呼吸韻律所展開的對比研究。

4.2. 粵語口傳文化的保護與研究

近十多年來，如何保護和傳承以語言或方言為載體的口傳文化引起了國內外眾多研究機構、團體和個人的關注。但是如何在技術層面實現全面保存這些口傳文化資料，學界尚未能形成統一的共識。運用多模態的理念和方法來保護和傳承口傳文化無疑是未來的一個主要方向。本資料庫有意識地對粵語的各種口傳文化資源進行採集、保存和研究。對跟粵語語音相關的呼吸信號、嗓音信號以及唇位、伴隨姿勢等信號的收集

和提取，這比傳統單一的錄音錄影更有利於全面地保存粵語講古、相聲、民歌、粵曲等口傳文化。目前我們正在有序地開展粵語吟誦、粵語民歌等口傳文化的多模態保護及研究等相關工作，已採錄幾十個小時的粵語吟誦、東莞木魚歌、中山咸水歌、粵西山歌等一手素材。

4.3. 多學科的交叉研究

為適應交叉學科發展需要而建設的多模態粵語語音資料庫，它所提供的資源和資訊不僅可以運用到粵語語音研究本身，還可以運用到言語工程、粵語教學、言語治療和康復、聲樂教學等相關領域。這有利於粵語研究跟中文資訊處理、病理治療與康復、聲樂、言語工程等學科的交融。我們可以利用多模態的語音資料來開拓以上領域的研究和應用。李煥哲、曹慶松、宋秀豹、吳南開、侯興泉（2015）的研究就是一項典型的由語言學者和五官科及康復科醫生合作完成的跨學科研究，這對弄清粵語人群病理嗓音和正常嗓音的區別和聯繫有著重要的啟示意義。我們現正跟音樂學院民族聲樂系的師生合作開展的粵語民歌多模態研究也是一項跨學科的研究。相信未來跨學科的粵語語音研究成果會越來越多。

5. 結語

粵語語音的多模態研究是一項基於多學科交叉的創新型研究，需綜合運用方言學、現代語音學、言語工程學、嗓音生理學、聲樂學以及電腦科學等學科的理論和方法來展開研究。建立粵語言語語音多模態資料庫是開展粵語語音多模態研究的前提和基礎。鑒於之前已經建好的粵語語音資料庫都是單一信號來源的聲學信號庫，設計並構建多模態的粵語語音資料庫顯得尤為重要。本資料庫在設計和構建的過程中充分考慮到多學科交叉的需求，在材料的設計和採集以及資料的提取等方面都儘量滿足語言學本體、中文資訊處理、言語工程等學科的需要。

粵語語音多模態資料庫所採集的胸腹呼吸資料、嗓音資料以及唇形資料將有利於推進粵語語音的生理研究，並有利於生理研究跟聲學研究的結合。針對粵語語音多模態資料庫而開發的相關技術對粵語口傳文化的保護和傳承也是一種促進。粵語語音多模態資料庫所提供的資源和資訊不僅可以運用到粵語語音研究本身，還可以運用到言語工程、粵語教學、言語治療和康復、聲樂教學等相關領域，這有利於粵語研究跟相關學科的交流與融合。

鳴謝

本文是“暨南大學創新資金”啟明星項目“多學科交叉視角下的粵語語音研究”（15JNQM024）的階段性成果。感謝匿名評審對本文提出的寶貴修改意見。

參考文獻

鄧德崇。2017。格律詩詞的粵語吟誦和朗讀的語音及呼吸韻律研究。暨南大學文學碩士論文。

黃立鶴。2015。語料庫 4.0：多模態語料庫建設及其應用。《解放軍外國語學院學報》第 3 期，頁 1-7/48。

孔江平。2008。語音多模態研究和多元化語音學研究。收錄於《中國語音學報》編委會編：《中國語音學報》（第一輯）。北京：商務印書館。

李煥哲、曹慶松、宋秀豹、吳南開、侯興泉。2015。基於 EGG 的粵語母語者病理嗓音與正常嗓音比較研究。發表於第二十屆國際粵方言研討會，香港中文大學。

李新魁、陳慧英、麥耘、方舟。1995。《廣州話音檔》。收錄於侯精一主編：《現代漢語方言音庫》。上海：上海教育出版社。

劉新中。2014。《廣州話單音節語圖冊》。廣州：世界圖書出版公司。

潘曉聲。2011。漢語普通話唇形協同發音及可視語音感知研究。北京大學博士論文。

邵慧君、秦綠葉。2008。論粵方言語音資料庫的建設。《學術研究》第 4 期，頁 147-150。

Allwood, Jens. 2008. Multimodal corpora. In *Corpus Linguistics: An International Handbook*, ed. Anke Lüdeling, and Merja Kytö, 207-225. Berlin: Mouton de Gruyter.

Bernsen, Niels Ole, and Laila Dybkjær. 2007. Annotation schemes for verbal and non-verbal communication: Some general issues. In *Verbal and Nonverbal Communication Behaviours*, ed. Anna Esposito, Marcos Faundez-Zanuy, Eric Keller, and Maria Marinaro, 11-22. Berlin and Heidelberg: Springer-Verlag Berlin Heidelberg.

Chan, Joyce Y. C., P.C. Ching, and Tan Lee. 2005. Development of a Cantonese-English code-mixing speech corpus. In *Proceedings of INTERSPEECH-2005*, 1533-1536.

Gu, Yue-guo. 2006. Multimodal text analysis: A corpus linguistic approach to situated discourse. *Text and Talk* 26(2): 127-167.

Hou, Xing-quan, De-chong Deng, and Feng Yang. 2017. The respiratory prosody of seven-syllable modern-style poems chanted in Cantonese. *Journal of Literature and Art Studies* 7(1): 69-76.

Kipp, Michael, Jean Claude Martin, Patrizia Paggio, and Dirk Heylen, eds. 2009. *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*. Berlin and Heidelberg: Springer-Verlag Berlin Heidelberg.

Knight, Dawn. 2009. A multi-modal corpus approach to the analysis of backchanneling behaviour. Doctoral dissertation, The University of Nottingham.

Knight, Dawn. 2011. The future of multimodal corpora. *Brazilian Journal of Applied Linguistics* 11(2): 391-415.

Li, Shu-qing, Fang Zheng, Ming-xing Xu, Zhan-jiang Song, and Di-tang Fang. 1999. A Cantonese accent Chinese speech corpus. Paper presented at the 2nd International workshop on East-Asian language resources & evaluation (Oriental COCOSDA'99), Taipei.

Lo, Wai Kit, K.F. Chow, Tan Lee, and P.C. Ching. 1998. Cantonese databases developed at CUHK for speech processing. In *Proceedings of the Conference on Phonetics of the Languages*, 77-80.

Lo, Wai Kit, P.C. Ching, Tan Lee, and Helen M. Meng. 2001. Design, compilation and processing of CUCall: A set of Cantonese spoken language corpora collected over telephone networks. In *Proceedings of Research on Computational Linguistics Conference* 14, 193-212.

Thompson, Paul. 2010. Building a specialised audio-visual corpus. In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael McCarthy, 93-103. New York: Routledge.

Zheng, Neng-heng, Chao Qin, Tan Lee, and P.C. Ching. 2010. CU2C: A dual-condition Cantonese speech database for speaker recognition. In *Computer Processing of Asian Spoken Languages*, ed. Shuichi Itahashi, and Chiu-yu Tseng, 90-93. Los Angeles: Consideration Books.

The Construction of the Multi-modal Cantonese Speech Corpus: A Speech Production Perspective

Xingquan Hou and Feng Yang

Jinan University

Abstract

The multi-modal approach has been widely applied in the modern phonetic research, and the multimodal corpus linguistics has become the major field in the current corpus linguistic study. Building a multi-modal Cantonese speech corpus is of necessity for the up-to-date Cantonese phonetic research. The multi-channel speech production signals simultaneously recorded in the corpus have diversified applications in the less-studied fields of Cantonese-related speech physiology, communication patterns, speech engineering, orally-transmitted culture, speech therapy and rehabilitation. Furthermore, it will facilitate and propel the collaboration between the Cantonese linguistics study and other related disciplines.

Keywords

Cantonese, multi-modal, speech corpus

通訊地址：廣東 廣州 天河區 暨南大學 漢語方言研究中心（侯興泉）
廣東 廣州 天河區 暨南大學 應用語言學研究院（楊　鋒）
電郵地址：thouxingquan@jnu.edu.cn（侯興泉）
yangzihuai@163.com（楊　鋒）
收稿日期：2016年1月21日
接受日期：2017年9月20日