Developing a Corpus-based Online Pronunciation Learning System for Cantonese Learners of Mandarin

CHEN Hsueh Chu, Rebecca
WANG Li Xun
KEUNG Shuet
HAN Qian Wen, Veronica
veronicahan923@gmail.com
The Project - Aim

• To develop a corpus-based online pronunciation learning system for Mandarin teachers, learners and researchers

• To help users better understand the major problems in learning Mandarin pronunciation encountered by Hong Kong Cantonese learners
The Project - Motivation

• Hong Kong is a metropolitan city that serves as an international financial centre, which has given rise to multi-lingual characteristics in recent years.

• In addition to Cantonese and English which serve mostly as first and second languages, Hong Kong residents have increasingly started to develop a third or even a fourth language.

• Mandarin is encouraged by the Biliteracy and Trilingualism (两文三语) language policy.
Mandarin Education in Hong Kong

- The education of Mandarin started in 1984 in Hong Kong school curriculum as an **elective subject** and became a **core subject** in primary and secondary school curricula in 1998 (Leung & Hui, 2011).
- The Hong Kong Examinations Authority (HKEA) included the Test of Proficiency in Putonghua as **public examination** (Tse et al., 1995).
- Exposure to Putonghua does **not** lead to a ‘satisfactory advance in the levels of Putonghua proficiency’ due to the fact that media and political debates and education are not supportive enough to the learning of Putonghua (Leung and Hui, 2011, p. 18).
Spoken Corpora

• It is believed that language in spoken form can be an effective guide to the basis of language learning than in written form (Chui & Lai, 2008).

• As information technology progresses, there is a growing interest in developing spoken corpora. A number of spoken corpora have been constructed for various research and pedagogical purposes.
Chinese Spoken Corpora

• The National Chengchi University (NCCU) Corpus of Spoken Chinese which established an online access to spoken Mandarin, spoken Hakka, and spoken Southern Min (Chui & Lai, 2008)
• The Mandarin Conversational Dialogue Corpus (MCDC) contains 30 sets of spontaneous dialogues in Mandarin
• The Mandarin Map Task Corpus (MTCC) involves 26 sets of dialogues that are orthographically transcribed
• The Hong Kong Cantonese Adult Language Corpus (HKCAC) contains both the orthographic and phonetic transcriptions of spoken Cantonese from the radio in late 1990s
• The PolyU Corpus of Spoken Chinese comprises of conversational exchanges in spoken Cantonese with 28 transcripts in early 2010s
Research Gap

1. Most of the Chinese spoken corpora simply compile speech data without annotation or analysis.

2. Researchers can use data from the corpora for research purpose, but it is difficult for learners of Chinese to learn from the corpora.

3. Teaching and learning resources of Chinese can be developed based on a corpus, to be useful for researchers, teachers, and learners.
Corpus-based Online Mandarin Pronunciation Learning System

http://ec-concord.ied.edu.hk/Mandarin_Pronunciation/
Corpus-based Online Mandarin Pronunciation Learning System

http://ec-concord.ied.edu.hk/Mandarin_Pronunciation/
The learning system has been developed for learners to:

• practice segmental and suprasegmental aspects of Mandarin pronunciation through a variety of perception and production exercises

• discover the possible causes of common pronunciation features found in the corpus
Corpus-based Online Mandarin Pronunciation Learning System

A Spoken Corpus of Hong Kong Learners

Useful Resources for Mandarin Pronunciation Learning

A Corpus-based Online Mandarin Pronunciation Learning System

Pronunciation Practice for Mandarin Learners

Praat Beginners’ Manual
Our Spoken Corpus

Visit our spoken corpus

Our spoken corpus contains audio data with phonological annotations that focus on
Our Spoken Corpus - Characteristics

• It provides high-quality recordings that are ideally suited for phonetic and acoustic analysis by researchers around the world.

• It produces recordings and phonological annotations that are easily accessible and immediately available to all learners, teachers and researchers, both in and outside EdUHK.

• It provides a platform for learners to access and rate the corpus data in order to discover the linguistic features on their own and to enhance their active engagement in their own learning.

• It describes the distinctive linguistic features of Mandarin production of university students in Hong Kong.
Our Spoken Corpus

The spoken corpus contains:

• 8 hours recording data of Mandarin spoken tasks (reading of monosyllabic words, reading of multisyllabic words, reading of passage, and free speech) from university students in Hong Kong

• detailed phonological annotations of the recordings mainly focused on segmental features (vowels & consonants) and suprasegmental features (tone & retroflex finals)
Our Spoken Corpus

Users can:

• browse the recordings by speakers’ biographic information (e.g. gender, age, major)

• search the recordings by the pronunciation features they are interested in
Our Spoken Corpus - Browse

Online Mandarin Pronunciation Learning
A Spoken Corpus of the Mandarin of Hong Kong learners
Our Spoken Corpus - Search

Online Mandarin Pronunciation Learning
A Spoken Corpus of the Mandarin of Hong Kong learners
Our Spoken Corpus – Spoken Tasks

• Task 1: Reading of monosyllabic words
• Task 2: Reading of multisyllabic words
• Task 3: Reading of passage
• Task 4: Free speech
Our Spoken Corpus – Spoken Tasks

Rationale

Task 1 - Reading of monosyllabic words

• Cover all tones
• Even distribution of the 4 tones
• Avoid polyphones
• Avoid overlapping with words in other tasks
Our Spoken Corpus – Spoken Tasks

Rationale

Task 2 - Reading of multisyllabic words

• Include confusing clusters (e.g. Tone1+Tone4, Tone2+Tone3 etc.)

• Include tone sandhi (e.g. Tone sandhi of Tone3, ‘yi’, ‘bu’, and neutral tone)

• Include confusing vowels (e.g. nasal vs back nasal)
Our Spoken Corpus – Spoken Tasks

Rationale

Task 2 - Reading of multisyllabic words

• Word final ‘r’ vs ‘er’ syllable
• Include difficult consonants (e.g. retroflex consonants ‘zh, ch, sh’, nasal consonant ‘n’ etc.)
• Avoid polyphones
• Avoid overlapping with words in other tasks
Our Spoken Corpus – Spoken Tasks

Rationale

Task 3 - Reading of passage

• Collect comprehensive speech data that reveals speakers’ pronunciation of vowels, consonants, tones, tone sandhi etc.
Our Spoken Corpus – Spoken Tasks

Rationale

Task 4 – Free Speech

• Speakers selected one topic from ‘My Studying Life’ and ‘An Unforgettable Trip’

• Collect authentic speech data of Mandarin for research concerning discourse, phonology, grammar, vocabulary etc.
Our Spoken Corpus – Spoken Tasks

• Task 1: Reading of monosyllabic words

Tasks 1&2 Word Reading

Speaker: 23
Gender: Female
Age: 21
Education Level: Bachelor
Major: English
First Language: Cantonese
Second Language: English
Third Language: Mandarin
Year of Mandarin Learning: 10

Remarks: Please read the below marking note for details.

Speaker's Record: 0:57
Beijing Sample: 1:33
Taiwan Sample: 0:57

Marking Scheme of Reading Task 1

佳 丰 域 努 稀 恒

奉 興 吼 愛 智 氣

帥 徐 遙 罕 隱 卓

軒 喬 阻 劃 粉 杞

漬 胚 恩 凶 屏 慘

寺 紐 粒 嶽 翟 繪

淹 倍 殊 喪
Our Spoken Corpus – Spoken Tasks

• Task 1: Reading of monosyllabic words
Our Spoken Corpus – Spoken Tasks

• Task 2: Reading of multisyllabic words

Marking Scheme of Reading Task 2

<table>
<thead>
<tr>
<th>不用</th>
<th>包括</th>
<th>徐個</th>
<th>狹窄</th>
<th>狹窄</th>
<th>反省</th>
</tr>
</thead>
<tbody>
<tr>
<td>践踏</td>
<td>汛期</td>
<td>轉臉</td>
<td>瓦房</td>
<td>兒歌</td>
<td>耕耘</td>
</tr>
<tr>
<td>炎熱</td>
<td>沉著</td>
<td>提防</td>
<td>提防</td>
<td>圓圈兒</td>
<td>講講</td>
</tr>
<tr>
<td>遵守</td>
<td>遵守</td>
<td>挑戰</td>
<td>舒服</td>
<td>電影兒</td>
<td>性質</td>
</tr>
<tr>
<td>研究</td>
<td>等於</td>
<td>廣村</td>
<td>概念</td>
<td>分水嶺</td>
<td>視網膜</td>
</tr>
<tr>
<td>海市蜃樓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our Spoken Corpus – Spoken Tasks

• Task 2: Reading of multisyllabic words

Marking Scheme of Reading Task 2

- 不用
- 包括
- 排個
- 狹窄
- 狹窄
- 反省
  - xià→xié
  - zhāi→zāi

- 蹉跎
- 汛期
- 時時
- 瓦房
- 兒歌
- 耕耘

- 熱
- 熱
- 逐著
- 撚訪
- 撚訪
- 圓圓
- 謹慎
  - rè→rì
  - zhuó→zúó
  - dǐ→lǐ
  - fāng→fāng
  - ràng→yàng

- 遵守
- 遵守
- 挑戰
- 舒服
- 電影
- 性質
  - zūn→zhūn
  - shǒu→xiú

- 研究
- 等於
- 農村
- 概念
- 分水嶺
- 視線
- 視線
  - děng→děn
  - fēn

- 海市蜃樓
- 事願
- shèn→zhūn
Our Spoken Corpus – Spoken Tasks

• Task 3: Reading of passage

Marking Scheme of Reading Task 3

生活對於任何人都非易事，我們必須相信，我們對每一件事情都具有天賦的才能，並且無論付出任何代價，都要把這件事完成。當事情結束的時候，你要能問心無愧地說："我已經盡我所能了。"

Marking Scheme of Reading Task 3

有一年的春天，我因病被迫在家裏休息數周。我注視著我的女兒們所養的蠶正在結繭，這使我很感興趣。望着這些蠶執著地、勤奮地工作，我感到我和它們非常相似。像它們一樣，我總是耐性地把自己的努力集中在一個目標上。我之以如此，或許是因為有某種力量在鞭策著我——正如蠶被鞭策著去結繭一般。
Our Spoken Corpus – Spoken Tasks

• Task 3: Reading of passage

生活對於(yú→yǔ)任何人都非易事。我們必須相信，我們對每一件事情都具有天賦(fù→fǔ)的才(cái→chái)能，並且Incomplete Tone 3無論付出任何代價(jià→jiè)，都要把這件事情完成。當事情結束(shù→sù)的時候(hou→hòu)，你要能問心無愧地說：“我已經盡我所能(suǒ→shuō)了。”

有一年的春天，我因病被迫在家休息(xi→xī)數周。我注視(shì→shì)著我的女兒們所養的蠶(cán→chán)正在結繭Incomplete Tone 3，這(zhe→zhe'r)使我感興趣(qu→qù)。望著這些蠶(cán→jiān)執著地、勤奮地工作，我感到我和它們(mén→mēn)非常相似。像它們一樣，我總是耐心(xīn→xīng)地把自己的努力集中(zhōng→zōng)在一個目標上(shāng→shàng)。我之所以(zhǐ→zǐ)如此，或許因為有某種力量在鞭策著(zhě→zé)我——正如蠶(cán→chán)被鞭策著去結繭一般。
Our Spoken Corpus – Spoken Tasks

• Task 4: Free Speech

Task 4 Free Speech

Speaker: 23
Gender: Female
Age: 21

Education Level: Bachelor
Major: English

First Language: Cantonese
Second Language: English
Third Language: Mandarin
Year of Mandarin Learning: 10

Speaker’s Record: [Audio Player]
Download

Topic: 我的學習生活

我的學習生活在小學的時候是非常的煩悶，因為小學的生活很簡單，老師都會為我們安排了一切，所以小時候的生活主要就是唸書，然後應付考試、測驗，等等。父母對我的要求也並不是很

高，不像是現在那樣很多的怪獸父母都會要求自己的孩子去上什麼什麼補習班。我當時是非常自

由，總是隨自己意願去做的。所以當然煩悶，但是那些測驗、考試都是能應付得了的。最後就很

成功地上了一所挺不錯的第一等級的中學，在中學上面，中學的學習生活比小學的步伐快很多，

測驗的次數的...還有那個頻率都比小學來的高好幾倍。所以我在中學的學習生活都是在非常繁忙

的考驗和測試中度過，學習生活之中我有遇到過幾個好師，他們對我的...學科上面，讓他們學科

上面的知識增長了不少。尤其是高中最後的一年，我遇上...發現在我身邊的師傅其實都是...個寶

貝。他們的學歷...他們的能力非常的高，教學的...質量非常的高，在他們的帶領下，我對...我

從...我在那個...高中那個公開考試裡面就更好地掌握了。
### High Frequency Feature List

<table>
<thead>
<tr>
<th>category</th>
<th>feature</th>
<th>example</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>segmental</td>
<td>嘭 (sa→sai)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>嘭 (sa→saɪ)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>跳 (na→nia)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>嘭 (tia→tia)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>嘭 (mii→mii)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>跳心動 (por→pail)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vowel</td>
<td>言 (ma→ma)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>言 (me→me)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>言 (tai→tai)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>漲 (yi→ye)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>嘭 (tui→tai)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>嘭 (o→u)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>嘭 (rao→rang)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pronunciation Practice

1. Tongue Twisters 總口令
2. Songs 歌曲
3. Jokes 笑話
4. Quiz 測試
Pronunciation Practice – Tongue Twisters

Short Vowels

1. 坡上立着一只鸡 (jī), 坡下就是一条河 (hé). 寬闊的河, 滾肥的豬, 鴨要過河, 鴨要渡河, 不知是誰過河, 還是何處渡河？

Sample Reading

Slow:

Standard:

Search ...

OUR PAGES

Our Spoken Corpus
General Learning Resources
For Cantonese Learners
Our Praat Manual
Recommended Learning Websites
English Pronunciation Learning Website

Sample Reading

Slow:

Standard:
## Pronunciation Practice – Songs

<table>
<thead>
<tr>
<th>Song</th>
<th>Artist(s)</th>
<th>Focused Phoneme</th>
</tr>
</thead>
<tbody>
<tr>
<td>土里查</td>
<td>周定番</td>
<td>(n, l)</td>
</tr>
<tr>
<td>Lydia</td>
<td>F.L.R</td>
<td>(zh, s)</td>
</tr>
<tr>
<td>不再不愛</td>
<td>楊健樺, 梁靜儀</td>
<td>(sh, z) (zh, z)</td>
</tr>
<tr>
<td>千里之外</td>
<td>周華健, 周潤發</td>
<td>(y, j)</td>
</tr>
<tr>
<td>霸花台</td>
<td>周杰倫</td>
<td>(w, r)</td>
</tr>
<tr>
<td>Super star</td>
<td>S.H.E</td>
<td>(w, m)</td>
</tr>
<tr>
<td>說愛你</td>
<td>羅志祥</td>
<td>(g, k)</td>
</tr>
<tr>
<td>不要規定我去</td>
<td>周傑倫, 周潤發</td>
<td>(w, m)</td>
</tr>
<tr>
<td>謝幕 謝幕</td>
<td>韓磊</td>
<td>(n, l)</td>
</tr>
<tr>
<td>謝幕 謝幕</td>
<td>梁靜茹</td>
<td>(zh, x)</td>
</tr>
</tbody>
</table>
Pronunciation Practice – Songs

Artist: 曹宝清，周杰伦
Focused Phonemes: (ə) (ə̌) (ɪ)

HK&Overseas: https://www.youtube.com/watch?v=ocOq3ySyHSI
Mainland: http://www.iqiyi.com/viavue/20130327/1b675fasf4f74d90.html

千里之外
1. 有一天，小明在家中看电视，妈妈说：“等一下你去冲个凉吧，用[fa]洗。小明不明白，问：“什么是[fa]？”妈妈想了想说：“哦，我搞错了，应该是[luà]洗！

2. 小明约了朋友见面，迟到了。他和朋友说：“不好意思，刚和妈妈吃了饭！”朋友不解。问：“吃花[ban]（[ban]）？”小明羞愧到自己说错了，“哦不是，是吃饭（[fan]）！不要笑话我了哈哈！
## Pronunciation Practice – Quiz

### Online Exercise

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Vowel</th>
<th>Tone</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$100</td>
<td>$100</td>
<td>$100</td>
<td>$100</td>
</tr>
<tr>
<td>$200</td>
<td>$200</td>
<td>$200</td>
<td>$200</td>
</tr>
<tr>
<td>$300</td>
<td>$300</td>
<td>$300</td>
<td>$300</td>
</tr>
<tr>
<td>$400</td>
<td>$400</td>
<td>$400</td>
<td>$400</td>
</tr>
<tr>
<td>$500</td>
<td>$500</td>
<td>$500</td>
<td>$500</td>
</tr>
</tbody>
</table>

---

The Education University of Hong Kong Library

For private study or research only. Not for publication or further reproduction.
Pronunciation Practice – Quiz

**CONSONANT $100**

**Question:** 請選擇你聽到的聲母（呼讀音）:

- j
- q
- x
- s

*Check Answer*
1. General Introduction 概述

Mandarin Chinese, also called Standard Chinese, has been the official language of China for a few decades. It is commonly used in Chinese schools, communities, institutions and media. The Chinese language is actually a collection of various dialects that can be categorized into eight groups: Northern Chinese (Mandarin), Yueh (Cantonese), Wu, Kan, Northern Min, Southern Min, Hsiang, and Hakka (Chang, 2001). Though 70% of the Chinese population are native speakers of Northern Chinese, which is the basis of modern Standard Chinese, most of the Chinese people do not speak Mandarin accurately (Chang, 2001). According to a survey (Chinese Ministry of Education, 2004), only 10% of the Chinese population can speak Mandarin fluently and accurately.

普通话，也稱標準華語，已經在過去的幾十年裏一直是中國的官方語言。普通話在中國學校、社區、研究所以及媒體中被廣泛運用着。中文其實是不同方言的合集，這些方言可以被分成八類：北方中文（普通話）、粵（粵語）、吳、閩、閩北，閩南，湘，客（楊, 2001）。雖然中國人口的70%都是以北方中文為母語的，而粵中華文又是現代標準中文的基礎，但是大部分的華人不以標準的普通話（除，2011）。據調查（中國教育部，2004），只有中國人口的10%可以流利標準的普通話。

2. Pinyin System 拼音系統

Pinyin, literally meaning “spell out the sound”, is the Romanized phonetic transcription of the Chinese language. The system is now used in China to teach Chinese to students and internationally as a second language. Pinyin system bridges the gap between writing and speech in Chinese, and enables learners to derive the meanings of characters (Sisk & Fletcher, 2001).

拼音，字面上的意思是“拼出聲音”，是用羅馬拼音系統記錄的中文。在中國，這個系統現在用來教授學生中文並且在語言上被當作第二語言。拼音系統橋樑了中文寫作和會話的關係，同時讓學習者能夠明白漢字的
Learning Resources

Recommended Learning Websites

Websites

- 拼音發音室網站 [Pinyin Learning Website]

  The website provides systematic introduction of Pinyin system, including syllable structure, correspondence to IPA system and so on. Learners can access a series of exercises whilst learning Pinyin.

- 普通話學習網 [Putonghua Learning Website]
  http://pth.lingui.org/yinyin/index.html

  This website compiles useful learning materials of Mandarin, including Pinyin system, place of articulation, segmental and suprasegmental features of Mandarin. Moreover, vocabularies, correspondences of Cantonese and Mandarin, learning exercises and useful Mandarin learning tools are introduced.

- 普通話世界 [Putonghua World]
  http://www.putonghuaworld.com/putonghua/index.htm

  Putonghua world is a platform built by an expert who has taught Mandarin for years. Mandarin teaching experience, researches, and learning resources are shared to Mandarin teachers and learners.

- 普通話自學網站 [Mandarin Self-learning Website]
Videos in Chinese

1. Articulation of tones, j, q, x, n, l

https://www.youtube.com/user/seq160/videos

These videos teach in detail how tones, and difficult sounds such as j, q, x, n, and l in Mandarin should be produced with help of model reading.

2. 50 common mistakes in speaking Mandarin

https://www.youtube.com/watch?v=uKV8tQ06P14

The video lists 50 common mistakes for Cantonese speakers in speaking Mandarin, with model reading.
Praat Beginner’s Manual

- Chapter 1. Introduction: Acoustic Phonetics and Praat
- Chapter 2. Acoustic Analysis of Vowels
- Chapter 3. Acoustic Analysis of Consonants
- Chapter 4. Prosodic Analysis with Praat
- Chapter 5. Annotating Sound Files with Praat

Suggested Answers

Online Resources
1. Other recommended Praat manuals
2. Extra exercises, tutorials, and self-tests
3. IPA resources: charts&fonts
You can learn pronunciation of Mandarin tones by analyzing your own reading with help of Praat, and compare your analyses with our sample reading.

For more practice, please download our recordings by clicking the file name, jump to a new page, right-click the audio and click 'save as'.
Tone Learning by Praat

PRAAT - SAMPLE READING ANALYSES

You can compare your audio analyses in Praat with our sample reading analyses to correct your pronunciation of Mandarin tones.

**Monosyllabic word**

Beijing Sample

Mean Pitch (Blue Line): 171.8 Hz
Mean Intensity (Yellow Line): -3.6 dB
Duration: 0.779168 s

Taiwan Sample

Mean Pitch (Blue Line): 205.8 Hz
Mean Intensity (Yellow Line): -3.6 dB
Duration: 0.779168 s
Tone Learning by Praat

Monosyllabic words

Sample Reading:
tone1 tone2 tone3 tone4

Learner Reading:
learner tone1 learner tone2 learner tone3 learner tone4

Multisyllabic words:

Sample Reading:
Sample 1 Sample 2 Sample 3 Sample 4

Learner Reading:
Learner 1 Learner 2 Learner 3 Learner 4

Search ...

OUR PAGES

Our Spoken Corpus
General Learning Resources
For Cantonese Learners
Our Praat Manual
Recommended Learning Websites
English Pronunciation Learning Website