WeCanTalk: A New Multi-language, Multi-modal Resource for Speaker Recognition

Karen Jones, Kevin Walker, Christopher Caruso, Jonathan Wright, Stephanie Strassel

Linguistic Data Consortium

Introduction

- The **We Can Talk (WCT) Corpus** was created to support NIST Speaker Recognition Evaluation **(SRE)** Campaign
- **Cantonese, Mandarin and English** recordings from **200 multilingual speakers** in Hong Kong
- **Conversational telephone speech** (CTS) and **audio from video** (AfV) from each speaker
- Used in NIST **2021 Speaker Recognition Evaluation**
- Data will be **published in LDC Catalog** after use in evaluations

Language Requirements per Speaker	Calls/ Speaker	Videos/ Speaker
Cantonese monolingual	5	2
Non-Cantonese monolingual	5	1
Freestyle (any mono or mixed language)	1	1
Total	**11**	**4**

Plus one selfie image

Speakers

Sex (self-reported)	Count
Female	154
Male	45

Year of Birth (self-reported)	Count
1960-1969	1
1970-1979	4
1980-1989	4
1990-1999	131
2000-2009	61

- Collection at Hong Kong Polytechnic University Jun-Oct 2020
- Coincided with COVID-19 lockdowns, Hong Kong protests
- Speaker recruitment and management at PolyU
- IRB oversight from Penn and PolyU
- Adult speakers in Hong Kong
- Native Cantonese speakers, fluent in 1+ additional language
- Speakers compensated per recording plus completion bonus
- No strict demographic requirements
- 315 enrolled speakers, 202 completed all requirements

CTS Collection

Collection System

- LDC-designed, implemented by PolyU with LDC remote access
- Control computer for handling prompts and recording
- Asterisk dialplan for call routing
- Custom Interactive Voice Response (IVR) software
- Database servers at both sites; VPNs and firewalls for file transfer

Protocol

- Enrolled speakers call friends and family
- Use dialpad to enter call details (e.g., language, device, noise)
- Callees remain anonymous
- Both enrolled speakers and callees provide consent for recording

Call Requirements

- ☑ 11 calls
- ☑ specified langs
- ☑ 8-10 mins/call
- ☑ 3-5 mins speech
- ☑ 3+ unique callees
- ☑ flag repeat callees
- ☑ familiar callees
- ☑ unrestricted topics
- ☑ 25%+ noisy calls
- ☑ landline, mobile only
- ☑ 2+ handsets
- ☑ 1 call/day max

AfV and Selfie Collection

Collection System

- Web form integrated into participant enrollment app, hosted by LDC

Protocol

- Enrolled speakers log into website and either upload video/image, or specify existing URL with their data to be downloaded by LDC
- Use webform to enter video details (e.g., language, number of speakers visible and audio)
- Provide consent prior to upload
- Webform reports upload results, file size and duration
- Limited requirements to create low barrier to participate

Video Requirements

Required

- ☑ 4 videos
- ☑ specified langs
- ☑ 3-10 mins/video
- ☑ speaker audible & visible

Desired

- ☑ mix of monolog and multi-party
- ☑ varied setting, speaker appearance
- ☑ varied noise conditions
- ☑ varied recording devices

Collection Auditing and Quality Control

Automated Checks

- File duration, speech duration (SAD), duplicate video (md5sum)

CTS Manual Auditing Shortly After Recording

- Enrolled speaker side only (callees are not SRE target speakers)
- Listen to first 15 seconds of each call - characteristic greetings etc.
- Plus extracted 60-sec segments from start, mid, end of call
- Verify quality, language, speaker demographics, noise conditions

AfV Manual Auditing Shortly After Upload

- Review full recording, sampling as much as needed
- Verify quality, language, speaker demographics, count of other speakers present, selfie match with video speaker, noise conditions

Speaker Auditing After Speaker Has Completed All Data

- Verify that all recordings (and selfie) are the same speaker

Results

Corpus Package

- CTS: 8 KHz alaw single channel files with sph headers
- AfV, selfie: original format (and video codec) as provided by speaker
- Speaker, call and audit metadata tables

Language	Calls (2359)	Videos (840)
Cantonese	52.01%	59.88%
Mandarin	29.67%	11.67%
English	15.56%	19.29%
Mixed	2.37%	8.21%
Other	0.38%	0.95%

Noise Condition	Calls (1 n/a)	Videos
Noisy	803	77
Not Noisy	1555	763

Unique CTS...	Count	Speakers
Devices	1	2
	2	62
	3+	138
Phone Numbers	1	162
	2	35
	3	5

CTS Device Type	Calls (2359)
Internal mic	39.6%
Speakerphone	23.7%
Headset	36.8%