*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 3451–3456
Marseille, 20-25 June 2022


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

**Karen Jones, Kevin Walker, Christopher Caruso, Jonathan Wright, Stephanie Strassel**
Linguistic Data Consortium
University of Pennsylvania, Philadelphia, USA
{karj, walkerk, carusocr, jdwright, strassel}@ldc.upenn.edu

**Abstract**

The WeCanTalk (WCT) Corpus is a new multi-language, multi-modal resource for speaker recognition. The corpus contains Cantonese, Mandarin and English telephony and video speech data from over 200 multilingual speakers located in Hong Kong. Each speaker contributed at least 10 telephone conversations of 8-10 minutes' duration collected via a custom telephone platform based in Hong Kong. Speakers also provided at least 3 videos in which they were both speaking and visible, along with one selfie image. At least half of the calls and videos for each speaker were in Cantonese, while their remaining recordings featured one or more different languages. Both calls and videos were made in a variety of noise conditions. All speech and video recordings were manually audited for quality including presence of the expected language and for speaker identity. The WeCanTalk Corpus has been used to support the NIST 2021 Speaker Recognition Evaluation and will be published in the LDC catalog.

**Keywords:** speaker recognition, speech database, video, telephony, Cantonese, Mandarin, Chinese, English

## 1. Introduction

The WeCanTalk (WCT) Corpus is a multi-modal, multi-language corpus comprising conversational telephone speech (CTS) and audio from video (AfV) recordings from native Cantonese speakers recruited in Hong Kong who were also fluent in at least one other language. The corpus was designed to support the development and evaluation of speaker recognition technologies. The collection consists of data from 202 distinct speakers, with at least 10 telephone calls and 3 videos plus an accompanying selfie image provided by each speaker.

All speakers made at least 5 of their 10 required telephone calls and at least 2 of their 3 required videos in Cantonese. The remaining non-Cantonese recordings, whether video or call, were required to be in a different language that the participant spoke fluently. In most cases the speaker's secondary language was either English or Mandarin. Telephone calls were made and recorded in Hong Kong via a custom-built telephone collection platform designed by Linguistic Data Consortium (LDC) and built and operated by Hong Kong Polytechnic University (PolyU) under LDC's direction. Video recordings and selfie images produced by each speaker were uploaded to a custom website developed and hosted by LDC.

WCT data collection took place in June-October 2020, a time characterized by significant political and social turmoil in Hong Kong including widespread political protests and increasing impact of the COVID-19 pandemic, both which affected the collection approach, particularly for video data, as well as the overall project timeline. Despite these challenges, the collection was successfully completed and the resulting corpus was used in the NIST 2021 Speaker Recognition Evaluation (NIST, 2021; Sadjadi et al, 2022). The sections that follow describe the WCT corpus design and implementation in detail. After a general introduction we discuss speaker recruitment efforts, requirements for speakers and data, the collection platforms, data validation and auditing procedures, corpus preparation, and finally a summary of the corpus properties.

## 2. Corpus Design and Comparison with Prior Work

To support SRE evaluation goals, the WCT corpus was required to contain a number of specific features. First, all data had to be collected from speakers outside of North America, from a location whose telephone channel characteristics would introduce a new technology challenge. It was also necessary to recruit speakers who were fluent in at least two languages and could contribute data, including conversations with other people, in those languages; this meant focusing collection on a location with a large multilingual speaker population. Finally, it was important to recruit speakers who would be able to contribute video as well as telephony recordings, which pointed to the need for collection in a country where usage of video and smartphone technology was common practice. On the pragmatic side, it was also necessary to identify a location with a capable local collection partner who could manage recruitment and build and operate the collection platform.

After evaluating a number of possible options, the decision was made to conduct the collection in Hong Kong. While Cantonese is the most commonly spoken language in Hong Kong, with 96.7% of the population speaking the language (Census & Statistics Department, Hong Kong Special Administrative Region, 2016), over 50% of the population speaks English, and over 50% speaks Mandarin (Lee et al, 2012). Given this multilingual setting it is not surprising that code-switching is common in everyday speech in Hong Kong (Chan, 2019). This reality provided an opportunity to collect recordings in which participants speak a mixture of languages, providing a further challenge to SRE technology. This had to be balanced against the need for some calls and videos in which speakers spoke a single language, which led to the need for detailed instructions and explanations to study participants, as well as careful auditing of collected data for language features. In terms of smartphone usage and internet connectivity, over 98% of households under the age of 65, and 68.1% of households 65 and over, use smartphones, while over 93% of households have internet connectivity (Census and Statistics Department, 2021). The degree of smartphone

and internet usage in Hong Kong suggested that recruitment would not be hampered by lack of access to required technology.

After establishing Hong Kong as a suitable collection locale we identified Hong Kong Polytechnic University, led by staff in the Department of Chinese & Bilingual Studies, as a collection partner who had the expertise, technical infrastructure and data collection experience required to handle collection operations in Hong Kong.

Although prior SRE corpora produced by LDC include Cantonese and Mandarin data, most notably the Call My Net Corpus (Jones et al, 2017) built to support the NIST 2016 SRE Evaluation (NIST, 2016; Sadjadi et al, 2017), the WCT corpus includes a number of unique features. First, WCT doubles the number of Cantonese speakers relative to Call My Net, from 100 to 200. WCT also is the first SRE corpus developed by LDC to include multilingual Cantonese speakers; Call My Net used separate speakers for Mandarin versus. Cantonese. Additionally, WCT is the first corpus developed in support of NIST SRE evaluations that includes both video and telephony data from every speaker. Finally, in the WCT corpus, all data was collected from participants in Hong Kong using a telephony platform based in the same locale, whereas in Call My Net speakers were in Guangzhou, China while their calls were collected via a call collection platform in London (a factor that may have introduced distinctive channel effects into the audio recordings).

## 3. Speaker Recruitment and Enrollment

The WCT data collection protocol was subject to Institutional Review Board (IRB) approval at the University of Pennsylvania (LDC's host institution) as well as at PolyU. All speakers provided informed consent prior to contributing data, and were compensated for each recording produced plus a bonus upon completion of all requirements. Participant recruitment was managed by PolyU in Hong Kong and was primarily based on word of mouth. While extensive on-campus recruitment was also planned, this was severely hampered by two major events. First, during the 2019-2020 Hong Kong protests, the PolyU campus became the focal point of social unrest. Consequently, the campus was closed to staff and students for several months. Second, the increased presence of COVID-19 in Hong Kong in summer 2020 necessitated further campus closures. Plans to post fliers on campus and to work on technical updates to the telephone collection platform hosted on campus had to be deferred until the University became accessible.

Given the resulting delays, LDC began making contingency plans for collection in other locations. These included organizing a satellite recruitment and collection site in Macau, and adding a secondary collection site in Philadelphia. Ultimately it was determined that the overall collection timeline could be extended such that there was no need to execute these alternatives in the end.

Participants registered for the WCT collection via an enrollment website hosted by LDC, which provided detailed information about the requirements for participation, privacy protections for subjects, and compensation details. After enrollment subjects could log into the account to update their contact information, check their collection progress and contact recruitment staff at PolyU who had responsibility for all aspects of participant care and recruitment progress tracking. Study coordinators and recruiters could also query the website's database backend to view real-time information about enrollment and collection for the study as a whole or for individual participants.

Despite the challenges of recruitment presented by the ongoing protests and the global pandemic, a total of 315 participants were recruited for the collection. Although the collection required only 200 speakers, it is well understood that some recruited participants never produce any data, while others begin but drop out before completion. For that reason, over-recruitment is necessary to ensure that at least 200 complete speakers could be collected before the end of the project.

## 4. Speaker Requirements

As in other recent LDC data collections to support NIST SRE evaluation, WCT used a model in which a recruited speaker makes calls to multiple callees in their social network, as well as multiple videos that may or may not involve other people. Each recruited speaker was required to complete 11 calls and 4 videos to be considered "complete", with calls and videos meeting the additional requirements described in the sections that follow.

In addition to being native Cantonese speakers with fluency in at least one other language, all speakers were required to be adults located in Hong Kong. Recruited speakers self-reported basic demographic information including sex, year of birth and native language upon enrollment. Once enrolled, each recruited speaker was assigned a unique, persistent anonymous ID (PIN) which was then associated with all of their data. The callees contacted by the enrolled speaker, and other speakers present in video recordings, were entirely anonymous and no demographic information was collected for them, nor were they assigned a speaker ID. Both enrolled speakers and non-enrolled callees provided consent to be recorded prior to each telephone call.

## 5. Language Requirements

Speakers recruited for the WCT collection were instructed to produce calls and videos with a specific language distribution. The primary collection language was Cantonese, with 5 Cantonese calls and 2 Cantonese videos required from each speaker. Speakers were also required to make 5 additional monolingual calls and 1 additional monolingual video in a different language of their choosing, nearly always English or Mandarin. The same speaker was permitted to contribute data in both English and Mandarin as long as each recording contained only a single language. Both the primary speaker and their callee were required to speak the same language for the Cantonese and non-Cantonese monolingual calls, and for all monolingual videos. Finally, each speaker was required to produce one call and one video designated as "Freestyle", in which they and their call partners could use any language or mix of languages, including code-switching. The WCT language requirements are summarized in Table 1.

| Language | Calls per Speaker | Videos per Speaker |
|---|---|---|
| Cantonese monolingual | 5 | 2 |
| Non-Cantonese monolingual | 5 | 1 |
| Freestyle (monolingual or mixed) | 1 | 1 |
| **Total** | **11** | **4** |

Table 1: Call and video language requirements

Note that while the corpus design required only 10 calls and 3 videos per speaker, participants were instructed to make an extra recording in each modality. This strategy helped keep collection goals on track even if individual recordings turned out to be unacceptable for some reason.

## 6. Call Requirements

Telephony data collected for WCT had the following requirements. The call duration was 8-10 minutes, designed to yield at least 3-5 minutes of speech per call side, though enrolled participants were encouraged to do at least half of the talking. Speakers could discuss any topic of their choosing that would result in a natural conversation, with care taken to avoid sensitive topics and personal identifying information such as full names. At least 25% of the speaker's calls were required to be made in a noisy environment, and all calls were to be made from a landline or cellphone. There were no requirements to make VOIP calls.

Speakers were instructed to make no more than 1 call per day. Calling the same person more than once was allowed, but unique pairings were encouraged with at least 3 distinct pairings per speaker required. While there was no strict handset variety requirement, speakers were strongly encouraged to use at least 2 distinct handsets across their 11 calls; distinct handsets could include using the same device in speakerphone mode or with a headset.

## 7. Video and Selfie Requirements

Because WCT was the first LDC speaker recognition collection effort targeting both video and telephony data from every enrolled speaker, it was not known ahead of time how difficult it might be for speakers to reach the video collection goals. Therefore, video requirements were fairly lightweight in order to maximize the likelihood of meeting the overall collection goals. Each enrolled speaker was required to contribute at least 4 video recordings where they were both visible and audible, with a strong preference for the speaker's face to be visible in at least part of the recording. Video duration was required to be between 3-10 minutes, and speakers generally preferred to contribute shorter videos. Videos could be newly recorded for WCT or they could be existing videos that met requirements; if the existing video was already available online (e.g. on YouTube) speakers could provide the URL rather than uploading the video directly. Videos could be recorded in any setting, with speech on any topic, and could involve the enrolled speaker without any other individuals present or could include other people visible and/or audible in the recording.

Speakers were encouraged but not strictly required to contribute videos that varied across a range of dimensions including setting, noise condition, topic and number of speakers. Although the original collection design included plans for group video recordings at PolyU to increase variety, these plans had to be abandoned due to the closure of campus in light of the political protests and COVID-19 concerns; this resulted in fewer multi-party video conversations than originally anticipated. In addition to their 4 videos, each enrolled speaker was required to upload one selfie image clearly showing their face.

## 8. Collection

### 8.1 CTS Collection Platform

The telephone collection platform was designed by LDC and built and operated by PolyU, following detailed specifications for hardware and software selection, installation and configuration. Platform design enabled LDC staff in Philadelphia to conduct remote testing, monitoring and control of the system although the platform was physically located in Hong Kong. Major components of the telephone platform include:

- Control computer for handling both the recorded messages participants hear when they interact with the collection platform (prompts) and all recording functions
- Custom Interactive Voice Recording software with Cantonese prompts designed by LDC and recorded and installed by PolyU
- Asterisk dialplan for routing calls programmatically
- Database servers at PolyU and LDC for call logging and capture of speaker metadata
- VPNs for storing collected calls and videos securely at the Hong Kong site before transfer to LDC servers

All database interactions between the enrollment website, the telephone platform and the collection servers at both LDC and PolyU were configured, tested and implemented by LDC. File transfer protocols and network security involved careful selection and deployment of suitable VPNs and firewalls.

### 8.2 CTS Collection Protocol

As with the other recent LDC speech collections, the collection protocol for WCT involved recruiting participants (called "claques") to make calls to their own friends, relatives and acquaintances. The advantage of this model is that the resulting speech, since it is a conversation between people who know each other, is natural and realistic. Participants could talk about anything they wished, though they were reminded to avoid revealing any personal identifying information and discussing sensitive subject matters they did not want recorded. On this claque model, the following scenarios were permissible:

- Different claques could call the same person in cases where their networks overlapped
- A claque could be a callee in another claque's network
- Claques could call the same person more than once (claques were instructed to call at least 3 different people)

Although callees (i.e., non-claque call sides) were entirely anonymous and were not assigned a unique, persistent speaker ID, claques did indicate, for each call, whether they had called this person before as part of the

WCT collection and this information was used to understand the presence of unique vs. duplicate speaker pairings in the collection.

### 8.3 Participant Experience - CTS

The sequence of steps involved in making a call was as follows:

- Claque dials local Hong Kong call platform number and listens to questions
- Claque uses phone keypad to provide:
  - Consent to be recorded
  - PIN
  - Language to be used in the call (Cantonese, English, Mandarin, Other)
  - Phone type (mobile, landline)
  - Microphone type (internal mic, speaker, headset)
  - Noise level (noisy, not noisy)
  - Repeat callee (yes, no)
- Claque enters call partner's telephone number and the platform automatically dials that number
- Callee hears a greeting and provides consent to be recorded
- Platform connects both speakers and starts recording
- After 10 minutes, the platform terminates the recording.

### 8.4 Video/Selfie Collection Platform and Participant Experience

Participants utilized a simple user interface incorporated into the WCT enrollment website in order to submit videos and selfies. The site included a reminder of the requirements for each video recording (as described in Section 7), and data could be submitted in two ways :

- upload a file directly from the claque's phone or computer
- provide a URL pointing to pre-existing videos on a video hosting site

Videos with durations of between 3-10 minutes in any format were permitted, although claques generally gravitated towards shorter videos in part because they were faster to upload. Claques were required to answer a handful of questions about each video:

- Who is in the video (me, me plus one other, me plus multiple others)
- Whose voice can be heard in the video (me, me plus one other, me plus multiple others)
- What languages are spoken in the video (only Cantonese, Only Mandarin, Mix, Other)

When uploading a video/selfie or specifying a URL for existing videos, claques were also asked to indicate their consent for the data to be collected and used in the corpus, by clicking a button in a dialog box.

Once the claque submitted a video or selfie, the site would report back on the success or failure of the upload with information about filename, file size and duration.

## 9. Data Validation and Manual Auditing

### 9.1 Automatic Validation Checks

Automatic checks were performed on all incoming videos and calls as they were collected. These checks included confirmation that file duration meets requirements; confirmation that the audio recording contains sufficient amounts of speech as measured by the LDC HMM Speech Activity Detector v1.0.5 (Ryant 2013), and confirmation that md5 checksums on submitted videos were unique. If all of these conditions were true, then the recording was provisionally deemed successful and was then subject to manual review for language, speaker and overall quality. Calls or videos that failed the automatic quality check were flagged for review by study personnel who could then follow up with claques needing further guidance or help.

### 9.2 Manual Auditing Overview

In addition to automatic checks, a manual audit was completed for all calls and videos. Manual auditing has two primary goals. First, auditing ensures that each individual call or video meets basic requirement in terms of recording quality, language and amount of speech, and that basic information about language, number of speakers present and noise conditions in the recording is accurate. Second, auditing ensures that the set of videos and calls assigned to the same speaker ID really do contain speech from the same individual.

Auditing of individual video and call recordings was conducted as soon as possible after the item was collected, while auditing for speaker consistency was conducted after all data for a given speaker was collected. Custom web-based user interfaces were designed and implemented by LDC for each stage of auditing, and auditing was conducted by trained annotators who were native Cantonese speakers also fluent in both Mandarin and English. Auditors reviewed the full video during auditing, but only examined the claque call side during call auditing since the callee side was not used as SRE evaluation data.

### 9.3 Call Quality Audit

Prior to manual auditing of individual calls, LDC extracted individual call segments from the claque call side as follows:

- The initial 15-30 seconds of each call were earmarked to use as a "reference segment" for speaker-specific greetings and other characteristics
- The remainder of the call was divided into thirds (beginning, middle and end) and a 60-second segment containing the most speech was selected from each third

This resulted in a total of around 3 minutes of claque call side speech to audit per call.

The Call Quality Audit user interface presented annotators with the four extracted segments and a set of questions to answer for each call. Auditors would play each segment in its entirety and then answer the following questions about the call:

- Is there speech throughout most of the call?
- How clear is the phone line?
- Is this a noisy call?
- Is all of the speech from a single speaker?
- What is the speaker's sex?
- What language does the speaker use?
- Specify other language if known (optional)

### 9.4 Video Quality Audit

During video auditing, the Video Quality Audit user interface presented auditors with the entire video recording

along with the selfie submitted by the speaker. Auditors were instructed to watch and listen to as much of the video as required to accurately answer questions.

- Who is visible in the video? (speaker alone, one other, multiple others)
- Whose voice can be heard in the video? (speaker alone, one other, multiple others)
- Confirm presence of languages indicated by speaker during video upload
- Is speaker talking throughout the video?
- What is the video image quality?
- What is the video audio quality?
- Is there significant background noise in the video?
- What is the speaker sex?

### 9.5 Speaker Audit

Once all of the individual calls and videos for one speaker ID had been audited for quality, a comprehensive speaker audit was performed to ensure that the same speaker appeared in all recordings. The Speaker Quality Audit user interface presented auditors with the selfie image, which served as a kind of ground truth for the speaker's appearance. The first call recorded by the speaker served as an initial point of reference for the speaker's voice. All other calls and videos were judged in relation to these reference points.

Auditors were instructed to sample portions of each call and each video and to answer the question: "Is this same speaker as in the reference call?" For each video, the auditor was also asked to confirm that the speaker was the same as the person in the selfie.

## 10. Data Observations

### 10.1 Noise Conditions

To meet the noise requirement of at least 25% noisy calls in the collection as a whole, claques were instructed to make five of their 10 calls from noisy environments. There were no requirements to meet a specific percentage of noisy videos but auditors were still asked to make a judgement about noise level. Noisy conditions included such environments as busy cafes, shopping malls, transit stations, construction sites, sporting events, concerts, parties, rallies, or a room with a loud radio or TV playing. Quiet environments included such places as a quiet office, a park or room at home.

| Noise Condition Calls | Number of Calls |
|---|---|
| Noisy | 803 |
| Not noisy | 1555 |
| No response | 1 |

Table 2: Call Noise Conditions as Judged by Auditors

| Noise Condition Videos | Number of Videos |
|---|---|
| Noisy | 77 |
| Not noisy | 763 |

Table 3: Video Noise Conditions as Judged by Auditors

### 10.2 Call Devices, Microphone & Unique Phone Numbers

Unsurprisingly given the prevalence of smartphone usage in Hong Kong, most recordings were made from a mobile phone.

| Device Type | Number of Calls |
|---|---|
| Mobile | 2241 |
| Landline | 118 |

Table 4: Device Types in CMN2

Along with providing information about the kind of device they used to make a specific call, claques also gave details about the mode in which they used the device e.g. with or without a headset, with or without a speakerphone, wired or wireless.

| Type | Number of Calls |
|---|---|
| Internal mic | 934 |
| Speakerphone | 558 |
| Headset | 867 |

Table 5: Device Modes in CMN2

Since it was not always feasible for claques to use multiple handsets or devices to make their calls, the same device used in different modes was counted as two instances of a handset. For example, the same cellphone used with a headphone, then without a headphone was counted as two unique devices.

The number of unique devices per claque is presented in Table 6, and the number of unique phone numbers per claque is shown in Table 7.

| Unique Devices | Claques |
|---|---|
| 1 | 2 |
| 2 | 62 |
| 3 | 115 |
| 4 | 16 |
| 5 | 6 |
| 6 | 1 |

Table 6: Number of Unique Devices in CMN2

| Unique Phone Numbers | Claques |
|---|---|
| 1 | 162 |
| 2 | 35 |
| 3 | 5 |

Table 7: Unique Phone Numbers per Claque

### 10.3 Language Observations

The number of calls and videos made in specific languages was calculated on the basis of an analysis of audit judgements and any associated comments on language that auditors made.

| Language | Number of Calls |
|---|---|
| Cantonese | 1227 |
| Mandarin | 700 |
| English | 367 |
| Cantonse/English | 48 |
| Cantonese/Mandarin | 7 |
| Cantonse/Mandarin/English | 1 |
| Other | 9 |

Table 8: Call Language

| Language | Number of Videos |
|---|---|
| Cantonese | 503 |
| Mandarin | 98 |
| English | 162 |
| Cantonese/English | 45 |
| Cantonese/Mandarin | 7 |
| Cantonese/Other | 9 |
| English/Other | 5 |
| Mandarin/English | 1 |
| Mandarin/Other | 2 |
| Other | 8 |

Table 9: Video Language

Approximately 2% of calls consisted of a mix of languages compared to 8% of videos containing a mixture. It is not clear what accounts for this slight difference; it may reflect patterns of language usage in Hong Kong in different settings and modalities, or it may relate to minor differences in the data collection and auditing procedures (e.g. call auditing asked "What language does the speaker use?" whereas video auditing asked about "languages" with the instruction to "check all that apply").

### 10.4 Speaker Demographics

Claques input their year of birth and gender via the enrollment website. Aside from the requirement that all claques be at least eighteen there were no restrictions on age or sex. Note that one speaker did not provide birth year.

| Sex | Number of Claques |
|---|---|
| Female | 154 |
| Male | 48 |

Table 10: Claque Sex

| Year of Birth | Number of Claques |
|---|---|
| 1960-69 | 1 |
| 1970-79 | 4 |
| 1980-89 | 4 |
| 1990-99 | 131 |
| 2000-2009 | 61 |

Table 11: Claque Year of Birth

## 11. Conclusion

LDC delivered the complete set of WCT call recordings to NIST as full-length narrowband 1-channel 8-kHz a-law files. Both A and B channels were delivered along with all associated metadata and annotation judgements. Video recordings were also delivered to NIST in the same format that the participants submitted. NIST selected and extracted short segments from these recordings to use for the SRE21 evaluation.

Despite significant challenges to recruitment and collection in the forms of political unrest and COVID-19, the WCT corpus was a success. We collected a total of 2359 calls and 840 videos from 202 speakers who each contributed data in Cantonese and at least one other language, and who also each supplied a selfie image. The corpus was successfully used to support the NIST SRE21 evaluation, and will be published in the LDC catalog after the data is authorized for public release.

## 12. Acknowledgements

LDC would like to thank Omid Sadjadi and Craig Greenberg at NIST, and Elliot Singer at MIT Lincoln Laboratories, for their contributions to corpus planning and feedback on collected data. The authors are also very grateful to Dr. Yao Yao and her team at Hong Kong Polytechnic University for all determined efforts in overseeing the recruitment of participants and remote collection of data during challenging times.

## 13. References

### 13.1 Bibliographical References

Census and Statistics Department. 2016. "Use of Language by Hong Kong Population." Hong Kong Special Administrative Region. https://www.bycensus2016.gov.hk/data/snapshotPDF/Snapshot08.pdf [Online; accessed 10-Jan-2022]

Census and Statistics Department. 2021. "Thematic Household Survey Report No.73." Hong Kong Special Administrative Region. https://www.censtatd.gov.hk/en/data/stat_report/product/C0000031/att/B11302732021XXXXB0100.pdf [Online; accessed10-Jan-2022]

Chan, K.L.R. (2019). "Trilingual Code-switching in Hong Kong." Applied Linguistics Research Journal. https://alrjournal.com/jvi.aspx?un=ALRJ-22932 [Online; accessed 21-Dec-2021]

Jones, K., Strassel, S., Walker, K. and Wright, J. "Call My Net Corpus: A Multi-lingual Corpus for Evaluation of Speaker Recognition Technology," in INTERSPEECH 2017: 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings, 2017, pp. 2621-2624.

Lee, K.S., Leung, W.M. The status of Cantonese in the education policy of Hong Kong. Multiling.Ed. 2, 2 (2012). https://doi.org/10.1186/10.1186/2191-5059-2-2

NIST. (2016). "NIST 2016 Speaker Recognition Evaluation Plan." https://www.nist.gov/system/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf [online; accessed 17-Jan-22]

NIST. (2021). "NIST 2021 Speaker Recognition Evaluation Plan." https://www.nist.gov/system/files/documents/2021/07/12/2021_SRE_Evaluation_Plan_V5.pdf [Online; accessed 21-Dec-2021].

Sadjadi, O.S., Kheyrkhah, T., Tong, A., Greenberg, C.S., Reynolds, D.A., Singer, E., Mason, L.P., Hernandez-Cordero, J. 2017. "The 2016 NIST Speaker Recognition Evaluation." in INTERSPEECH 2017: 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden, Proceedings, 2017, pp. 1353-1357.

Sadjadi, O.S., Greenberg, C., Singer, E., Mason, L., Reynolds, D. 2022. "The 2021 NIST Speaker Recognition Evaluation." (Odyssey 2022 forthcoming)

### 13.2 Language Resource References

Ryant, N. (2013). "LDC HMM Speech Activity Detector (v1.0.5)." LDC, University of Pennsylvania