Call My Net Corpus: A Multilingual Corpus for Evaluation of Speaker Recognition Technology

Karen Jones[1], *Stephanie Strassel*[1], *Kevin Walker*[1], *David Graff*[1], *Jonathan Wright*[1],

[1]Linguistic Data Consortium, University of Pennsylvania, USA

karj@ldc.upenn.edu, strassel@ldc.upenn.edu, walkerk@ldc.upenn.edu, graff@ldc.upenn.edu, jdwright@ldc.upenn.edu

Abstract

The Call My Net 2015 (CMN15) corpus presents a new resource for Speaker Recognition Evaluation and related technologies. The corpus includes conversational telephone speech recordings for a total of 220 speakers spanning 4 languages: Tagalog, Cantonese, Mandarin and Cebuano. The corpus includes 10 calls per speaker made under a variety of noise conditions. Calls were manually audited for language, speaker identity and overall quality. The resulting data has been used in the NIST 2016 SRE Evaluation [1] and will eventually be published in the Linguistic Data Consortium catalog. We describe the goals of the CMN15 corpus, including details of the collection protocol and auditing procedure and discussion of the unique properties of this corpus compared to prior NIST SRE evaluation corpora.

Index Terms: speaker recognition, speech, telephone

1. Introduction

The Call My Net 2015 Corpus is a collection of telephone conversations in four Asian languages that was created to support the 2016 Speaker Recognition Evaluation (SRE). The main objective of the SRE evaluation series is to support technological development in the field of text independent speaker recognition, so it is essential that participants providing speech data are labeled with unique and persistent IDs. The Call My Net corpus consists of telephone recordings from a total of 220 unique speakers who each took part in at least 10 calls. One requirement for this corpus that differs from previous collection efforts was that all the telephone conversations were routed entirely outside of the North American telephone network, which posed challenges in terms of participant recruitment, collection management and resolution of technical issues.

2. Major and Minor Languages

Unlike LDC's previous SRE collection effort (the REMIX collection), which consisted of phone calls made by speakers of US English, the CMN15 collection consisted of telephone recordings in four Southeast Asian languages. The selection of languages was based on careful consideration of the following factors:

- A requirement to ensure that all calls were conducted entirely off the North American phone network
- Availability of call collection infrastructure in a non-North American locale
- Ease of participant recruitment at an overseas location
- Ability to recruit auditors for selected languages

Ultimately, four languages were collected for the CMN15 corpus as follows:

Major Languages (100 speakers per language)

- Tagalog
- Cantonese

Minor Languages (10 speakers per language)

- Cebuano
- Mandarin

3. Telephone Platforms and Speaker Locations

Despite prior successful experience in establishing and maintaining remote collection platforms, the tight collection time frame made setting up a new non-American platform for the Call My Net collection unfeasible. An experienced vendor with prior experience of partnering with LDC on overseas CTS collection and with active collection platforms was selected to assist with the collection effort. The configuration of the telephone platform is outlined in Table 1.

Table 1: *SRE16 Telephone platform configuration.*

Component	Details
Codec	a-law
Telephone System	E-1; ISDN-PRI signalling (no VOIP)
Hardware	1U Intel server form factor; Digium TE220 PCI_Express x1
Software	Custom Asterix 1.6 core application
Call Flow	Claque (caller) schedules call; platform dials out to claque and callees
Location	UK; Australia

A significant feature of the CTS collection set up was that the speakers participating in the collection were based in a different location from the recording platform. Specifically, while the Cantonese and Mandarin speakers were based in Guangzhou, China, the recording platform for their calls was located in London; and while the Tagalog and Cebuano speakers were based in the Philippines (in Manila or Davao) the

recording platform for these telephone conversations was in Sydney.

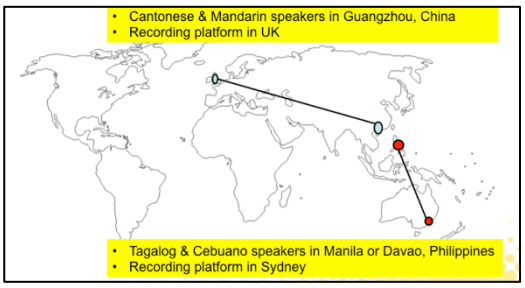

Figure 1: *Speaker locations and telephone platform locations.*

The distal separation of speaker from platform may have contributed to some anomalous call properties discussed in section 7.

4. Speaker Recruitment

The primary speaker recruitment model for the Call My Net collection was "claque-based". This model relies on recruited speakers (claques) to make calls to multiple individuals within their established social networks (friends, family members, acquaintances). The main advantage of this approach is that it yields natural, realistic conversation. On this model the following scenarios were permissible:

- Different claques may have overlapping networks i.e. multiple claques might call the same non-claque
- A non-claque may be called several times by the same claque (but it was a requirement that each claque have at least three different call partners)
- A claque could be a callee in another claque's network

Because of the risk of ID confusion, each of these scenarios was carefully monitored.

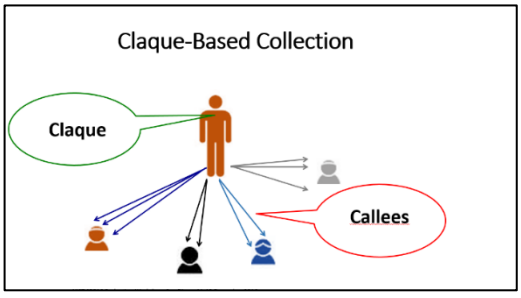

Figure 2: *Claque makes calls to multiple individuals.*

5. Speaker Requirements

5.1. Speaker Fluency

The recruitment strategy required that only native speakers of the collection languages were hired in order to satisfy the condition that claques be at least highly fluent if not a native speaker of the language in question.

5.2 Unique and Persistent Speaker ID

For the CMN15 collection, the claque side of the call (side A) was of primary interest and each claque was assigned a unique, persistent pin number that was used throughout the collection. Though not a hard requirement, each callee was also assigned a unique pin number. Great care was taken to ensure that in cases where a speaker participated both as a claque and as a callee (i.e. the call partner for a different claque) that the pin number was identical across both speaker roles.

5.3 Demographic Information

Obtaining demographic information including sex, year of birth and language was a hard requirement for claques. In fact, LDC succeeded in obtaining this demographic information for both claques and callees.

As with all LDC speech studies involving human subjects, managing the collection of demographic data requires a careful balancing act; on the one hand the final corpus should contain sufficient metadata to support speaker recognition research, while on the other any fears and suspicions subjects may have about providing personal data needed to be anticipated and addressed. In line with standard practice, LDC ensured that

- LDC project coordinators and technical staff had undertaken training in working with human subjects under the Collaborative Institutional Training Initiative
- the collection was approved by the University of Pennsylvania's Institutional Review Board
- the vendor's collection activity was constrained to ensure that personal identifying information was not divulged.

6. Call and Handset Requirements

In line with collection requirements, the CMN15 calls all meet the following conditions:

- The telephone network used was outside of North America
- Claques and callees were located outside of North America
- 10 telephone conversations per speaker (from each of at least 200 speakers)
- 3-5 minutes of speech per conversation
- Conversations are natural (claques were instructed to talk about topics of their own choosing, and to avoid using people's full names, telephone numbers or other personal identifying information during their phone conversations)

- Calls were in the specified language
- All phone numbers were uniquely identified by means of distinct strings that allowed for tracking of phone-set re-use without exposing the actual phone numbers.

An additional requirement that claques make no more than one call per day was complicated by a high incidence of connection problems in the Philippines. This issue is discussed in more detail in section 7.1.

Given a research preference for handset variety, each claque was required to and succeeded in using at least 3 different handsets (or at least 3 different configurations of phone/microphone/headset) to ensure device variety in the collection. The handsets used were self-reported by the claques.

Likewise, given a research preference for calls to be made in varied noise conditions, claques were required to and succeeded in making calls in at least three distinct acoustic settings. At least two of each claque's calls were made in noisy environments. Claques were given examples of what constituted a quiet background (quiet room at home, library setting, quiet office etc.) and a noisy background (busy street, busy café, busy shopping mall etc.).

7. Anomalous Call Properties

7.1. Multi-part Calls

Claques in the Philippines experienced a relatively high incidence of connection failure, which in turn caused a higher than expected number of dropped calls. This impacted on speaker behavior with claques occasionally needing to redial their call partner to continue their conversation.

If a call was cut off due to connection problems, it was allowable for a claque to re-dial within a short time span to continue the conversation. This resulted in several cases of "multi-part" calls. A "call_group_label" was used in the metadata to identify cases where two or three distinct calls were made within a short span of time, involving the same subjects, the same phone numbers and the same environmental conditions. Calls that share a given call_group_label were not to be considered as independent samples; rather, they are cases where the intended 10-minute conversation was broken up by one or two network outages, and the two subjects reconnected within the next few minutes in order to complete the remainder of the 10-minute conversation.

7.2. Unexpected Regions of Silence

Another anomaly observed in the collection came in the form of sporadic cases of unusually abrupt voice-onset and voice-offset. Durations of mid-syllable dropouts range between approximately 15 and 30 milliseconds.

This property appeared to affect both Tagalog and Cantonese, both A and B channels. It is possible that this anomaly is a characteristic of the mobile network rather than the ISDN line; the fact that the length of the dropout is the same as the length of a GSM frame i.e. 20ms seems to support this idea.

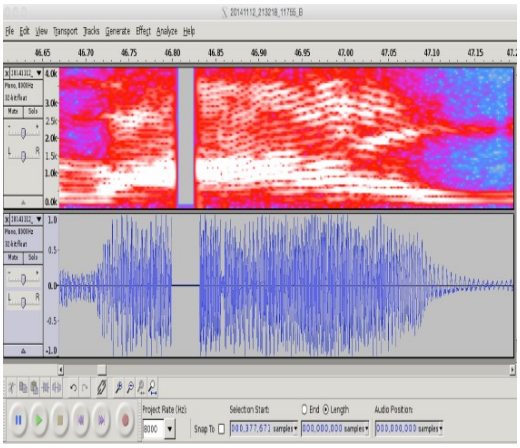

Figure 3: *Signal drop out.*

7.3. Misalignment of Call Sides

Throughout collection LDC liaised with Doug Reynolds and Eliot Singer at MIT-LL, who provided guidance on expected characteristics of collected data. One joint observation involved a subset of calls in which the channel A and B sides were found to be misaligned. Unlike LDC's typical platform settings, the vendor platform used for the CMN2 collection resulted in different recording start times for claques and callees. The behavior of the call platform was as follows:

- Claques provided consent and entered phone numbers of their call partners via a website
- Platform then simultaneously dialed out to claque and callee
- Callee recording began after pressing the key to give consent to be recorded
- Claque recording began after callee recording started

While most calls exhibited negligible lag between the recording start times on both channels, 56 calls (all Tagalog) had lag times of over 1 second.

Following consultation with MIT-LL and NIST, LDC obtained fine grained recording start times for each channel then carried out the following steps to remedy the problem:

- Elided samples from the beginning of each B-channel recording in order to align it with the beginning of the corresponding A-channel
- Elided or added a suitable number of samples from the end of the B channel if misalignment persisted after the previous step

Auditors who took part in a blind, randomized listening experiment to see which version of alignment was better confirmed the steps taken to correct alignment produced considerable improvement.

8. Preparing Segments for Audit

The first stage of preparing segments for audit was to run a speech activity detector (ldc_sad_hmm.v1) developed by Neville Ryant at LDC on each entire call. The first 15 seconds of the calls were always selected as a "reference segment" for speaker-specific greetings or other characteristic patterns. The next 15 seconds of the call were skipped and then the remainder of the call was divided into thirds (the beginning, middle and end), and from each third the densest 60-second segments of speech were selected for auditing. If a call was too short to yield three 60-second segments, then (roughly) the entire call would be lined up for auditing.

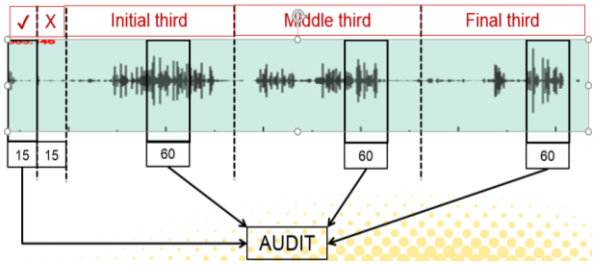

Figure 4: *Selection of speech segments from a call.*

9. Auditing

The external vendor supplied an initial set of metadata and audit results, and the LDC conducted a supplemental set of audits to confirm the validity of the vendor's information. Since the design of the collection involves using only the "A"-side subjects for speaker-recognition trials, LDC's own auditing was performed only on the A channel of each call.

Auditors used web-based software developed at LDC, and auditing consisted of two stages:

- A Quality Audit was performed by a small number of senior annotators with extensive experience in previous speech collections. Calls from speakers without multi parts were queued first. An additional prioritization built into the audit assignment logic was that all the calls for one speaker must be audited before calls for any other speaker. Annotators were given a single audit assignment presenting all the audit segments for one call along with the following set of questions:
    - Is there speech throughout most parts of this call? (yes, no)
    - How clear is the phone line? (good, acceptable, poor)
    - Is this a noisy call? (yes, no)
    - Is all the speech on the line from a single speaker? (yes, no)
    - What is the speaker's sex? (male, female, unsure)
    - Any comments?
- A Speaker Audit was performed by native speakers of the target language. For this round of auditing each assignment contained all calls associated with a single speaker ID including those that had "failed" the quality audit. The goal of this auditing task was to confirm that all calls associated with a single speaker ID are the same person and also to confirm that all calls are in the expected language. The speaker's first call was used as "reference" to compare subsequent calls against.

LDC also performed some dual annotation to measure auditor agreement.

10. Corpus Distribution

In total, LDC distributed to NIST the following numbers of calls as full 2-channel 8-bit, 8-kHz SPHERE files:

- Cebuano - 200 audio files (100 A/B pairs)
- Mandarin -200 audio files (100 A/B pairs)
- Tagalog - 2472 audio files (1236 A/B pairs)
- Cantonese - 2072 audio files (1036 A/B pairs)

Associated metadata was compiled in a series of tables which included information on:

- Subjects (subject ID and demographic information)
- Call date, time and ID
- LDC audits
- Vendor audits
- Noise conditions
- Handset type / phone type
- Language & country of origin
- Timestamps of audited judgements
- Anonymized phone numbers

Ultimately, the telephone recordings and associated annotations and metadata that make up the CMN15 corpus will be scheduled for release in LDC's general catalog.

11. Conclusions

The CMN15 collection supported the NIST SRE16 evaluation by providing telephone recordings produced with a variety languages, noise conditions and handsets. The use of non-US telephone networks and the presence of signal anomalies provided researchers who participated in the SRE16 evaluation with new and challenging data.

12. Acknowledgements

Our thanks are due to Craig Greenberg at NIST, and also to Doug Reynolds and Eliot Singer at Lincoln Laboratories, MIT for their contributions to corpus planning and verification of the collected data.

13. References

[1] NIST (2016). The 2016 NIST Speaker Recognition Evaluation Plan (SRE16). https://www.nist.gov/sites/default/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf