INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

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

Karen Jones, Stephanie Strassel, Kevin Walker, David Graff, Jonathan Wright

Linguistic Data Consortium, University of Pennsylvania, USA

{karj, strassel, walkerk, graff, 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 and will 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: linguistic resources, speaker recognition, multilingual, conversational telephone speech

1. Introduction

The Call My Net 2015 (CMN15) Corpus is a new collection of conversational telephone speech recordings from Linguistic Data Consortium (LDC), covering four languages and comprising multiple calls made by 220 unique speakers. The corpus is designed to support the development of robust speaker recognition technology, providing carefully collected and audited speech from a large pool of speakers recorded in different communicative situations. Native speakers of Tagalog, Cantonese, Cebuano or Mandarin made a total of 10 calls each, talking to people within their existing social networks. Speakers were encouraged to use different telephone instruments in a variety of acoustic settings, and were instructed to talk for 8-10 minutes on a topic of their own choosing. All conversations were collected outside of North America. Collected data was encoded as a-law sampled at 8kHz in SPHERE formatted files. CMN15 has been used as training, development and test data in the 2016 NIST Speaker Recognition Evaluation (SRE16) [1].

2. Collection Protocol

The CMN15 corpus collection protocol relies on the notion of "claques". Claques are speakers recruited to make calls to multiple individuals within their established social networks, including friends, family members co-workers and other acquaintances. The primary advantage of a claque-based methodology is that it encourages natural, realistic conversation since speakers are talking with people they already know. Recruitment strategies for SRE collections typically involve hiring almost double the required number of claques to counter the occurrence of people enrolling but failing to complete the study. Claques are directly compensated for their participation in the collection, earning a base payment for each successful call and a bonus once all calls are completed. Callees are not directly compensated though some claques may choose to share their compensation with their call partners. Both the claque and non-claque (callee) side are recorded and included in the corpus, but the claque side is the focus for SRE evaluation and is the only side that counts toward the goal of 220 total speakers.

Claques were required to call at least 3 different callees over the course of 10 calls, but were strongly encouraged to call a different person each time in order to maximize speaker pair variety. However, repeated speaker pairings do occur. Additionally, because claques may have overlapping social networks, multiple claques may call the same callee. Similarly, a claque may also be a callee for a different claque.

All claques were assigned a unique, persistent ID number that was used throughout the collection. Because only the claque call side counts toward the "deliverable speaker" goal it was not strictly necessary to assign a unique persistent ID to callees, although this was done. 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 ID number was identical across both speaker roles.

Although collection of demographic information was a strict requirement for claques only, basic information including sex, year of birth and language was collected for both claques and callees. No effort was made to balance demographics along any dimension.

As with all LDC collections involving human subjects, the CMN15 collection was conducted with oversight from the University of Pennsylvania's Institutional Review Board. Both claques and callees provide informed consent to be recorded and to have their calls and non-identifying call metadata published for research purposes. Care is taken to ensure that any personal identifying information (e.g. name and address collected for compensation purposes) is handled appropriately and is never divulged in the published corpus.

3. Collection Languages

Multilingual data has been a feature of some prior SRE corpora developed by LDC including some of the Mixer corpora [2], but CMN15 is the first large speaker collection containing no English speakers. Instead, the corpus comprises two major and two minor languages. Major languages include 100 speakers per language, while minor languages include only 10 speakers. The SRE16 development set contained data from both the major and minor languages, while the test set contained only major language speakers.

http://dx.doi.org/10.21437/Interspeech.2017-1521

Major Languages (100 speakers per language)

- Tagalog
- Cantonese

Minor Languages (10 speakers per language)

- Cebuano
- Mandarin

The CMN15 languages were chosen in consultation with SRE evaluation coordinators and in consideration of logistical constraints on the collection itself.

4. Recording Platforms

To support the goal of developing channel-robust speaker recognition technology, a primary requirement of the CMN15 corpus was that all calls must be collected outside of North America, utilizing entirely non-North American telephony networks. The collection budget and timeline dictated that we make use of recording platforms and related infrastructure already in place in the collection locales, rather than developing and deploying a new collection system, and so LDC selected an experienced vendor with telephone speech collection capabilities and existing recording platforms in suitable locations to handle the data collection. The vendor was also responsible for speaker recruitment.

LDC provided detailed technical and speaker specifications to the vendor to ensure that the resulting data met the corpus requirements and was generally compatible with prior SRE collections. Basic features of the recording platforms are summarized in Table 1.

Table 1: *CMN15 recording platform features*

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

Because it was necessary to rely entirely on existing recording platforms, some aspects of the collection were sub-optimal. One noteworthy feature is that the speakers participating in the collection were based in a different location from the recording platforms themselves. 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 respectively) the recording platform for these telephone conversations was in Sydney. The distal separation of speaker from platform may have contributed to some anomalous call properties discussed in Section 6. Further research is needed to thoroughly evaluate and quantify how calls made in this manner compare to those made with local connections only. While a possible association of channel characteristics with language would have a serious impact on a collection supporting a language recognition evaluation, the presence of specific channel features is less significant for a speaker recognition program.

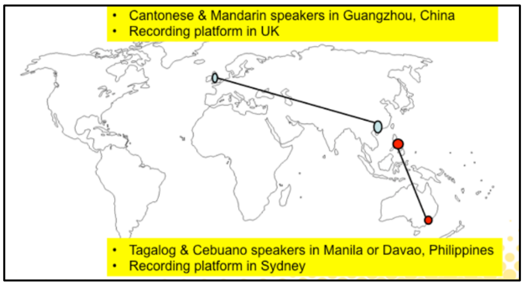
• Cantonese & Mandarin speakers in Guangzhou, China
• Recording platform in UK
• Tagalog & Cebuano speakers in Manila or Davao, Philippines
• Recording platform in Sydney

Figure 1: *CMN15 speaker and recording platform locations*

5. Call Requirements

To support the goals of the CMN15 collection and SRE16 evaluation, the following requirements were established for the collected data.

All claques were required to make at least 10 calls. Calls should last 8-10 minutes in order to yield 3-5 minutes of claque speech.

Conversation should be natural, on a topic of the claque's own choosing. Claques and callees were also instructed to avoid discussing sensitive topics, personal identifying information (like full names) or other information that they did not wish to have recorded. All conversations were required be in the specified language, with minimal code-switching.

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.

Given a research preference for handset variety, each claque was required to use 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 make calls in at least three distinct acoustic settings, with at least two of each claque's calls made in noisy environments. Claques were given examples of what constituted a quiet background (quiet room at home, library setting, quiet outdoor space etc.) and a noisy background (traffic noise in background, busy café, busy shopping mall etc.).

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 6.1.

6. Anomalous Call Properties

6.1. 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 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. All multipart calls contained at least one part with sufficient claque speech (3-5 mins), so there was no impact on the usability of calls for evaluation.

6.2. 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 2 below illustrates the signal dropout issue for one call.

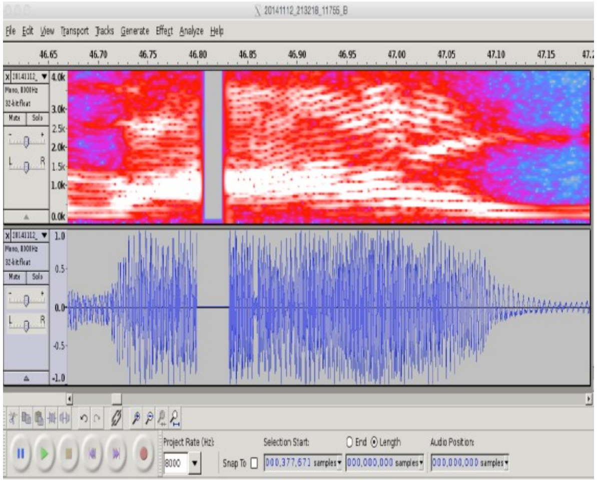

Figure 2: *Signal drop out.*

6.3. Misalignment of Call Sides

Although the technical requirements for the collection specified that call sides be time-aligned, it was noted in early data deliveries that the channel A and B sides were noticeably misaligned for a subset of the calls. Unlike LDC's typical platform settings, the vendor platform used for the CMN15 collection resulted in different recording onset 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 simultaneously dialed 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. Because the “speaker deliverable” in CMN15 was the claque call side and not the full call, misaligned call sides did not represent a problem for producing a corpus that could fully support the SRE16 evaluation. However, LDC decided to take steps to remedy the situation in order to add value to the corpus and enable it to be used for other kinds of speech research.

To this end, the vendor provided LDC with fine-grained recording start times for each channel. LDC 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 that the steps taken to correct alignment produced considerable improvement.

7. Auditing

Call auditing progressed in parallel with call collection. The collection vendor supplied an initial set of metadata and audit results, and LDC conducted a complete two-stage audit prior to delivering the resulting data to NIST for use in the SRE16 evaluation. LDC's audit reviewed the claque call side only, since the non-claque side of the call was not intended to count toward the 220- speaker collection goal for SRE16. The goals of auditing are as follows:

- Verify that the speaker associated with a given ID is consistent
- Verify that each call side contains a single speaker
- Verify that the call side contains a minimum of 3 minutes of topical conversation (on any topic)
- Classify the call as noisy vs. not-noisy
- Verify that the overall audio quality of the recording (e.g. with respect to distortion, echo) is acceptable

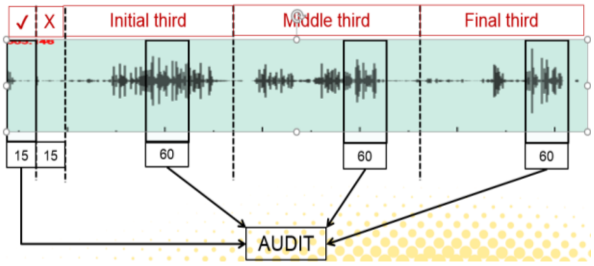

Figure 3: *Selection of segments for manual audit*

In order to make the LDC manual auditing task as efficient as possible, we pre-selected segments from each call side and packaged them into auditing “kits” for manual review by native speaker auditors. First, we ran the LDC-Ryant speech activity detector (SAD) on the claque call side [3]. The first 15 seconds of the recording was always selected as a "reference segment" for speaker-specific greetings or other characteristic patterns. The next 15 seconds was always skipped, and the remainder of the call was divided into equal thirds. From each third we selected the continuous 60-second segment that contained the most speech, as reported by SAD output. If a call was too short to yield three 60-second segments, then the full call was audited. This procedure is illustrated in Figure 3 above.

LDC auditing was conducted in two stages, using a web-based interface developed by LDC. The initial Quality Audit focused on one call at a time, and was performed by a small number of senior annotators with extensive experience in previous speech collections. Each auditing “kit” included the four audit segments selected for a given 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?

All available calls from a single speaker were audited before calls from a new speaker were presented. Note that Quality Audit judgments were made based on the recording only, without reference to the call metadata. This means that, for instance, calls self-reported as “noisy” by claques may be judged as “not noisy” by auditors.

Following the Quality Audit, native speakers of the target language performed a Speaker Audit. For this round of auditing each kit contained all calls associated with a single claque ID, including any calls that had “failed” the Quality Audit. The goal of the Speaker Audit 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 independent dual annotation to measure auditor agreement.

8. Corpus Distribution

The full CMN15 corpus, including full recordings of both call sides and complete call, speaker and audit metadata was delivered to NIST for selection of training, dev and test segments for the SRE16 evaluation. All calls were delivered, including any that “failed” auditing. Timestamps for the selected audit segments were also delivered. Table 2 summarizes the data delivered to NIST. Note that the number of calls audited includes multi-part calls.

Table 2: *Summary of delivered data*

Language	**Calls audited**	**Calls with single speaker**	**Speakers with 10+ calls**	**Speakers with < 10 calls**
Tagalog	1238	1230	100	1
Cantonese	1035	1031	100	0
Mandarin	100	100	10	0
Cebuano	100	100	10	0

All calls were delivered as 2-channel 8-bit, 8-kHz SPHERE files. Associated metadata was compiled in a series of tables that included information on:

- Subjects (subject ID and demographic information)
- Call date, time and ID
- LDC audit results
- Vendor audit results
- Noise conditions
- Handset type / phone type
- Language & country of origin
- Audit segment timestamps
- De-identified ANI/phone number

9. Conclusions

The CMN15 corpus is a new data set for the evaluation of Speaker Recognition technology. It is the first SRE collection undertaken by LDC that includes no English speakers and relies entirely on telephone networks outside of North America. The corpus includes 220 speakers from 4 distinct languages, with at least 10 calls per speaker reflecting a wide variety of handsets and acoustic environments. The data has been used by 66 sites in the SRE16 evaluation. The SRE16 Test Set and CMN15 corpus will remain sequestered until after the creation of a new blind test set, but the test set and the larger corpus will eventually be published in the LDC catalog, making the data broadly available to LDC members and non-member licensees.

10. Acknowledgements

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

11. References

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

[2] C.Cieri, L.Corson, D. Graff, K . Walker (2007). “Resources for New Research Directions in Speaker Recognition: The Mixer 3, 4 and 5 Corpora,” in *INTERSPEECH 2007 — 8th Annual Conference of the International Speech Communication Association, August 27–31, Antwerp, Belgium, Proceedings, 2007, pp. 2864–2868*

[3] N. Ryant, LDC HMM Speech Activity Detector (v.1.0.3a). LDC, University of Pennsylvania, 2013