# Call My Net 2: A New Resource for Speaker Recognition

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

**Abstract**

We introduce the Call My Net 2 (CMN2) Corpus, a new resource for speaker recognition featuring Tunisian Arabic conversations between friends and family, incorporating both traditional telephony and VoIP data. The corpus contains data from over 400 Tunisian Arabic speakers collected via a custom-built platform deployed in Tunis, with each speaker making 10 or more calls each lasting up to 10 minutes. Calls include speech in various realistic and natural acoustic settings, both noisy and non-noisy. Speakers used a variety of handsets, including landline and mobile devices, and made VoIP calls from tablets or computers. All calls were subject to a series of manual and automatic quality checks, including speech duration, audio quality, language identity and speaker identity. The CMN2 corpus has been used in two NIST Speaker Recognition Evaluations (SRE18 and SRE19), and the SRE test sets as well as the full CMN2 corpus will be published in the Linguistic Data Consortium Catalog. We describe CMN2 corpus requirements, the telephone collection platform, and procedures for call collection. We review properties of the CMN2 dataset and discuss features of the corpus that distinguish it from prior SRE collection efforts, including some of the technical challenges encountered with collecting VoIP data.

**Keywords:** speaker recognition, speech database, telephony, VoIP, data centers, Tunisian Arabic

## 1. Introduction

The Call My Net 2 (CMN2) Corpus is a large collection of Tunisian Arabic conversational telephone speech (CTS) recordings that was created to support the advancement of technology in the field of text independent speaker recognition. CMN2 adds a new language to LDC's SRE collections since Tunisian Arabic has not specifically been collected in the past. Also, in addition to traditional telephony data the CMN2 corpus contains VoIP data for the first time. All call recordings were collected in Tunis and the creation of this corpus involved developing several new infrastructural solutions to hosting and managing a remote collection. In total, 4,562 calls were made by 408 Tunisian Arabic speakers who made at least 10 calls each from December 2016 to January 2018. 2306 recordings from 213 speakers were used in the NIST 2018 Speaker Recognition Evaluation (NIST, 2018; Sadjadi et al., 2019). Calls from the remaining speakers were reserved for the NIST 2019 Speaker Recognition Evaluation (NIST, 2019).

## 2. Language

Together, the varieties of Arabic spoken throughout the Maghreb comprise a dialect continuum that encompasses the everyday Arabic dialects spoken in Libya, Tunisia, Algeria, Morocco, Western Sahara and Mauritania. While Maghrebi Arabic has featured in prior LDC collections such as LRE11, the regional dialects were not specifically distinguished. In contrast, one of the requirements for the CMN2 collection was that all study participants be native or highly fluent speakers of Tunisian Arabic.

The decision to specifically collect Tunisian Arabic was based on a number of factors:

- A general requirement that calls were made entirely outside of North America
- The linguistic expertise of the selected vendor's collection team; in addition to being native Tunisian Arabic speakers many team members possess doctorates in linguistics and have experience of annotation, quality control and linguistic analysis of previous Arabic language collections
- An opportunity to create a large archive of naturally-occurring, everyday speech in a dialect of Arabic for which there are limited spoken data resources.

Since French is commonly spoken in Tunisia, and Modern Standard Arabic is the country's official language, code-switching is a common characteristic of everyday speech, presenting challenges for a corpus required to be unambiguously Tunisian Arabic. For this reason, careful management of recruited speakers was a necessity, as was careful auditing of speech segments to confirm language and dialect.

The CMN2 collection differs from prior Tunisian speech datasets such as the Spoken Tunisian Arabic Corpus (Zribi et al., 2015) and the Tunisian Arabic Railway Interactive Corpus (Masmoudi et al., 2014) in that it consists of spontaneous conversational telephone speech on open topics between speakers who know each other.

## 3. In-country Collection

Speaker recruitment and collection was conducted entirely within Tunisia, utilizing a recording platform built and hosted by a collection partner in Tunis working under LDC direction. Hosting recording platforms at remote locations necessitates measures to minimize the risk of both unexpected platform behavior and vendor misinterpretation of collection requirements.

### 3.1 Collection Platform and Speaker Co-Location

In the first Call My Net collection (2015), the speakers and the collection platforms were located in different countries and it is likely that this distal separation contributed to a relatively high incidence of connection failures and anomalous signal properties in the calls (Jones et al., 2017). Since both the CMN2 call platform and the speakers were located in Tunis, the possibility of any connection issues arising from geographic separation was eliminated.

### 3.2 Collection Platform Specification

LDC provided detailed specifications to the collection partner in Tunis for building a portable recording platform that could support collection of both telephony and VoIP data, and the recording of up to 15 simultaneous 2-party, 4-wire calls. The specifications covered hardware and software purchase, installation and configuration.

Major components of the telephone platform include:

- Control computer for handling both the recorded messages participants hear when they interact with the collection platform and all recording functions
- LDC-designed Interactive Voice Recording software
- GSM to VoIP gateway providing access to the cellular network
- ISDN/PRI to VoIP gateway providing access to the phone traditional network
- Asterisk dialplan for routing calls programmatically
- Database servers in Tunis and LDC for call logging and capture of speaker metadata

The recorded instructions (prompts) that participants hear when they interact with the collection platform must be in compliance with the University of Pennsylvania's Institutional Review Board (IRB) approved human subjects protocol. For this reason, LDC provided the collection partner with approved prompt wording in Tunisian Arabic and a detailed set of instructions for both producing high quality prompt audio recordings and installing them onto the collection platform.

### 3.3 LDC Mirror Platform

The telephone recordings collected in Tunis were required to have no dependency on the North American telephone network, so technical staff in Tunis were tasked with ensuring that a-law codec was used for all inbound and outbound calls and recordings. However, the platform was specifically designed for LDC staff in Philadelphia to have remote deployment, testing, monitoring and control capabilities as a means for ensuring call and metadata collection progressed as required.

### 3.4 Web-based Collection Infrastructure

LDC provided a CMN2 study website that could be used by participants, recruitment and technical staff in Tunis and managers at LDC. The website provided a real-time view of recruitment and collection progress that was unavailable in prior collections. It was used for the following purposes:

- Participant enrollment – individuals interested in participating in the CMN2 collection could sign up via a secure webpage, provide contact information and the demographic information required for the collection (namely language, sex and year of birth)
- Call set up – enrolled individuals could provide information about each call they made for the collection (see Section 7)
- Progress reporting – a variety of collection progress reports, customized for different audiences, were made available to participants, the collection partner and staff at LDC.

## 4. Collection Protocol

As with the prior Call My Net (2015) collection, the primary collection model for CMN2 involved recruiting participants (called "claques") to make calls to their own friends, relatives and acquaintances. The main advantage of this model is that the resultant speech, since it is a conversation between people who know each other, is natural and realistic. Instead of presenting topics to discuss, participants are told that they can talk about anything, though care should be taken to avoid revealing any personal identifying information and discussing sensitive subject matters they do not wish to have 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 (but there was a requirement that claques must call at least 3 different people)

These various scenarios are illustrated in Figure 1.

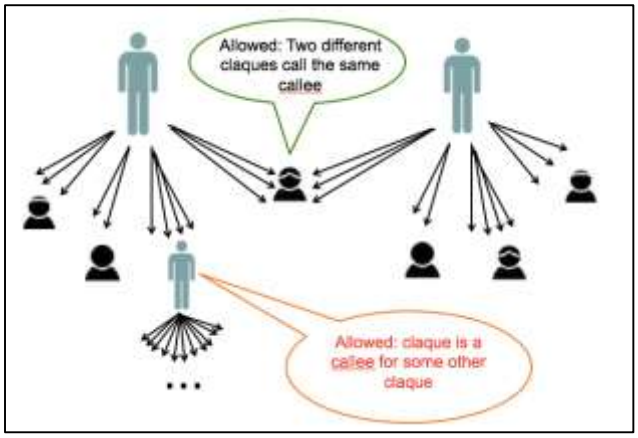


Figure 1: Claque-based collection model

## 5. Speaker Requirements

In addition to being native or highly fluent speakers of Tunisian Arabic, claques were also required to have a unique and persistent ID throughout the entire collection. They were also required to provide demographic information namely sex, year of birth and native language. This was self-reported via the web-enrollment process as discussed in Section 8.4.

Demographic information and persistent IDs were not collected for callees since they were not the primary speakers of interest for CMN2.

## 6. Call and Handset Requirements

The calls collected for CMN2 met each of the following collection requirements:

- At least 10 telephone conversations per speaker
- At least 3-5 minutes of speech per call side

- Each claque must call at least 3 people
- No more than one call per speaker per day
- Calls are entirely in Tunisian Arabic
- Conversations are on topics of speakers' own choosing, with care taken to avoid mentioning personal identifying information such as telephone numbers and full names
- At least 25% of the calls are made in noisy conditions
- Between 20-30% of the calls are initiated via VoIP
- Claque and callee phone numbers are uniquely identified, but phone numbers are anonymized
- Calls do not involve the North American telephone network

# 7. Call Flow Procedure

The sequence of steps involved in setting up and making a call was as follows. First, claques entered information about the call they were about to make via a simple web form including the following questions:

- Is this a VOIP call (yes, no)
- What kind of phone will you call from? (computer, mobile phone, landline)
- What kind of microphone will you use (internal device microphone, separate plug-in microphone, separate wireless or Bluetooth headphones)
- What is the noise level at your location (noisy, not noisy)
- Have you called this person before for this study (yes, no)

If claques selected "yes" for VOIP call, the phone type selection menu included computer but excluded landline; if they selected "no" for VOIP then the phone type menu excluded computer and included landline.

Claques were assigned a one-time call confirmation code that expired after 30 minutes. Claques dialed the collection phone number, pressed "1" to provide their consent to have the call recorded, entered the confirmation code and then entered telephone number of their call partner. The platform dialed the claque's call partner and played a message requesting that they provide consent by pressing "1". The two speakers were connected and recording began. After 10 minutes, the recording ended and the call was terminated.

After each call, a number of automatic quality checks were performed, including verification that the call had a duration of at least 7 minutes, with at least 3 minutes of speech on the claque call side as determined by automatic SAD (Ryant, 2013). If both of these conditions were true, then the call was provisionally deemed successful and was subject to manual review for language, speaker and overall quality.

# 8. Data Observations

## 8.1 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. 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. A total of 46% of claque call sides were reported as noisy.

| Device Type | Number of Calls |
|---|---|
| Noisy | 2087 |
| Not noisy | 2475 |

Table 1: Call Noise Conditions in CMN2

## 8.2 Handset

Claques were instructed to use at least two distinct handsets or devices during their participation in the study and report on the kind of devices they used e.g. cellphones, landlines or computers.

| Device Type | Number of Calls |
|---|---|
| Mobile | 4341 |
| Computer | 136 |
| Landline | 85 |

Table 2: 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 |
|---|---|
| Plugin mic | 958 |
| Internal mic | 2389 |
| Speakerphone | 1067 |
| Wireless | 148 |

Table 3: 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 4, and the number of unique phone numbers per claque is shown in Table 5. Note that phone numbers used for Skype calls were unknown since Skype calls are recorded as having an "anonymous" ID.

| Unique Devices | Claques |
|---|---|
| 1 | 3 |
| 2 | 16 |
| 3 | 60 |
| 4 | 108 |
| 5 | 93 |
| 6 | 63 |
| 7 | 35 |
| 8 | 17 |
| 9 | 10 |
| 10 | 2 |
| 11 | 0 |
| 12 | 1 |

Table 4: Number of Unique Devices in CMN2

| Unique Phone Numbers | Claques |
|---|---|
| 1 | 208 |
| 2 | 114 |
| 3 | 59 |
| 4 | 19 |
| 5 | 6 |
| 6 | 1 |
| 7 | 1 |

Table 5: Unique Phone Numbers per Claque (excluding office phones, see section 9.2)

### 8.3 Call Order Variation

To ensure variation in the order of call types across the collection as a whole, claques were instructed to avoid making calls of all one type (for example all of their five noisy calls) in a row. Likewise, claques were also instructed to mix up their VoIP and non-VoIP calls, and also the order in which they used different handsets. To reinforce this requirement, claques were given pop-up reminders on the call set-up web page to vary the order of specific call types if their latest input about the call they were about to make revealed a run of calls of one particular type.

Overall, the CMN2 corpus succeeds in exhibiting call order variation in terms of handsets and noise conditions. However, order variation for VoIP versus non-VoIP calls was more challenging for the reasons discussed in section 9.2.

### 8.4 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.

| Sex | Number of Claques |
|---|---|
| Female | 262 |
| Male | 146 |

Table 6: Claque Sex

| Year of Birth | Number of Claques |
|---|---|
| 1940-49 | 2 |
| 1950-59 | 10 |
| 1960-69 | 22 |
| 1970-79 | 30 |
| 1980-89 | 95 |
| 1990-99 | 249 |

Table 7: Claque Year of Birth

## 9. VoIP

The CMN2 corpus is the first LDC telephone speech collection to include VoIP calls. The collection of VoIP calls presented one of the most challenging aspects of the CMN2 collection.

### 9.1 Classification of VoIP and Telephony

Determining whether a call counted as VoIP or traditional telephony was challenging because of the complex and varied interactions between:

- Call platform components relying on SIP, ISDN or GSM gateway or some combination of these
- The network connection between the two speakers also varying between SIP, ISDN or GSM gateway
- The claque's device (landline, cellphone, computer)

Our strategy was to categorize a call as VoIP if the claque self-reported that they were using a VoIP client. Since the claque side of the call was the main side of interest no consideration of the callee side, whether it involved a VoIP client or not, factored into the classification of calls as VoIP.

### 9.2 VoIP Challenges and Solutions

We utilized two clients for VoIP calls: Skype and Viber. Initial tests conducted with Skype clients on a range of devices were beset with DTMF and network connection problems. As discussed in section 7, claques must make a number of inputs into their phone device or computer keyboard when they call the study line. Inputting the six-digit call confirmation code was especially problematic – this input which should trigger an automatic validation of the code against LDC's database records frequently resulted in a failed call. Testing showed that installing the latest version of Skype helped with this problem, though network connectivity continued to be an issue. The Viber client was somewhat more reliable but callers continued to experience dropped calls and other technical problems.

Ultimately, the required ratio of VoIP to non-VoIP calls was achieved despite a number of continued call failures that were due to poor network connections. As a consequence of the delays making Skype and Viber calls, many claques made all their non-VoIP calls first, contravening the requirement that call order be varied. Another unwelcome consequence of the network connectivity issues was that several claques, in seeking to get around their own poor internet connections, resorted to using a shared handset; this phone is identified as "icg" in Table 9.

| VoIPType | Number of Calls |
|---|---|
| Skype | 322 |
| Viber | 826 |
| Non-VoIP | 3414 |

Table 8: VoIP and non-VoIP calls in CMN2

| Number of Claques | Times Using "icg" |
|---|---|
| 352 | 0 |
| 27 | 1 |
| 11 | 2 |
| 17 | 3 |
| 1 | 5 |

Table 9: Use of "icg" Phone for VoIP Calls.

## 10. Auditing

The procedures for preparing sections of audio files for audit as well as the protocol for performing manual audits of the data follow those that were used successfully in CMN2015. After an initial round of vendor data verifications, experienced speech annotators at LDC listened to pre-selected segments identified by automatic SAD. Audit segments were extracted from each claque call side as follows:

- Segments between 15 and 40 seconds were extracted from the first minute of the call to use as a "reference segment" for speaker-specific greetings and other characteristics
- The remainder of the call was divided into thirds and the densest 45-second speech segment was selected from each third

Auditing was conducted in two stages. First, annotators with extensive experience in audio and speech annotation listened to the three segments in each call and made judgements about the call's signal clarity, amount of speech, speaker sex, number of speakers and noise level. Second, once the complete set of calls from one speaker had been audited for quality, a speaker audit was performed in which a native Tunisian Arabic annotator sampled segments from each call in the set and judged whether the speaker was the same across all calls, as well as whether the language was Tunisian Arabic for all calls. The reference segment from the beginning of each call, often containing distinctive greetings, was especially useful in the speaker confirmation task.

## 11. Corpus Distribution

LDC delivered the complete set of CMN2 call recordings to NIST as full-length 1-channel 8-kHz a-law files. Both A and B channels were delivered along with all associated metadata and annotation judgements. The CMN2 corpus will also be released in the LDC Catalog thereby making all 4562 calls available along with metadata tables providing information on:

- Subjects (subject ID, sex, year of birth, native language)
- Calls (call ID, call date, language ID)
- Callside (side A/B, subject ID, phone ID and information about phone type, device type, noise conditions, and whether the call was Skype, Viber or non-VoIP)
- Auditor judgements on audio quality, amount of speech as well as judgements on the gender of the caller, whether there is a single speaker and whether the call contains the expected language and the voice of the expected speaker.

## 12. Conclusions

CMN2 consists of telephone conversations produced in a variety of noise conditions and with a variety of handsets, all carefully audited for speaker ID, language and audio quality. The NIST SRE18 and SRE19 evaluations successfully utilized the CMN2 corpus and presented researchers with a large set of everyday conversations between Tunisian Arabic speakers who know each other for the first time. The resources described here will be released through the LDC catalog, making them available to the general research community.

## 13. Acknowledgements

LDC would like to thank Craig Greenberg and Omid Sadjadi at NIST, and Doug Reynolds and Elliot Singer at Lincoln Laboratories, MIT for their contributions to corpus planning and feedback on collected data. The authors gratefully acknowledge the contributions of Dr. Mohamed Maamouri who provided expert input on aspects of Tunisian Arabic and oversaw the corpus collection efforts in Tunis.

## 14. References

### 14.1 Bibliographical References

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.

Masmoudi, A., Khmekhem, M.E., Yannick, E., Belguith, L.H. and Habash, N. (2014). A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 306-310, Reykjavik, Iceland, may. European Language Resource Association (ELRA).

NIST. (2018). "NIST 2018 Speaker Recognition Evaluation Plan," https://www.nist.gov/system/files/documents/2018/08/17/sre18_eval_plan_2018-05-31_v6.pdf. [Online; accessed 20-Feb-2020].

NIST. (2019). "NIST 2019 Speaker Recognition Evaluation: CTS Challenge," https://www.nist.gov/system/files/documents/2019/07/22/2019_nist_speaker_recognition_challenge_v8.pdf, 2019 [Online; accessed 20-Feb-2020]

Sadjadi, S.O., Greenberg, C., Singer, E., Reynolds, D., Mason, L. and Hernandez-Cordero, J. (2019). The 2018 NIST Speaker Recognition Evaluation. In Interspeech 2019: 20th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings, 2019, pp. 1483-1487

Zribi, I., Ellouze, M., Belguith, L.H. and Blache, P. (2015). Spoken Tunisian Arabic Corpus "STAC": Transcription and Annotation. In Research in Computing Science, pp.123-135

## 14.2 Language Resource References

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