Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104-2653 USA
T: +1.215.898.0464
F: +1.215.573.2175
W: www.ldc.upenn.edu

LDC Spoken Language Sampler – 6th Release, LDC2023S07
Corpus Descriptions

2017 NIST OpenSAT Pilot – SSSF, LDC2022S01

Developed by NIST (National Institute of Standards and Technology), 2017 NIST OpenSAT Pilot - SSSF contains approximately one hour of operational speech data, transcripts and annotation files used in the speech activity detection, automatic speech recognition, and keyword search tasks of the 2017 OpenSAT Pilot evaluation. The source audio consists of radio and telephone dispatches during the Sofa Super Store fire (SSSF) in Charleston, South Carolina which claimed the lives of nine firefighters in June 2007.

The NIST Open Speech Analytic Technologies (OpenSAT) Evaluation Series was designed to bring together researchers developing different types of technologies to address speech analytic challenges present in some of the most difficult acoustic conditions with the end goal of improving the state-of-the-art through objective, large-scale common evaluations. The 2017 pilot focused on the public safety communications domain. The SSSF audio represents real-world, fire response, operational data with multiple challenges for system analytics, such as land-mobile-radio transmission effects, significant background noise, speech under stress and variable decibel levels.

2017 NIST Language Recognition Evaluation Training and Development Sets, LDC2022S10

2017 NIST Language Recognition Evaluation Training and Development Sets contains approximately 2,100 hours of conversational telephone speech, broadcast conversation, broadcast narrow band speech, and speech from video in 14 languages, dialects, and varieties including Arabic (Iraqi, Levantine, Maghrebi, Egyptian), English (British, American), Polish, Russian, Portuguese (Brazilian), Spanish (Caribbean, European, Latin American Continental), and Chinese (Mandarin, Min Nan).

The goal of the NIST (National Institute of Standards and Technology) Language Recognition Evaluation (LRE) is to establish the baseline of current performance capability for language recognition of conversational telephone speech and to lay the groundwork for further research efforts in the field.

2019 NIST Speaker Recognition Evaluation Test Set -- CTS Challenge, LDC2023S03

Developed by LDC and NIST, 2019 NIST Speaker Recognition Evaluation Test Set -- CTS Challenge contains approximately 635 hours of Tunisian Arabic telephone recordings for development and testing, answer keys, enrollment, trial files and documentation from the CTS Challenge portion of the NIST-sponsored 2019 Speaker Recognition Evaluation (SRE).

The SRE evaluations are intended to be of interest to researchers working on the general problem of text independent speaker recognition. The 2019 evaluation task was speaker detection, conducted in two parts: (1) a leaderboard-style challenge based on conversational telephone speech from LDC's Call

My Net 2 (CMN2) corpus; and (2) a separate evaluation using audio-visual material collected by LDC for the VAST (Video Annotation for Speech Technology) project.

AIDA Ukrainian Broadcast and Telephone Speech Audio and Transcripts, LDC2023S01

Developed by LDC, AIDA Ukrainian Broadcast and Telephone Speech Audio and Transcripts is comprised of approximately 156 hours of Ukrainian conversational telephone speech and broadcast news audio with 1.2 million words of corresponding orthographic transcripts.

The broadcast recordings and transcripts were produced to support the DARPA AIDA (Active Interpretation of Disparate Alternatives) program which aimed to develop a multi-hypothesis semantic engine to generate explicit alternative interpretations of events, situations and trends from a variety of unstructured sources. LDC supported AIDA by collecting, creating and annotating multimodal linguistic resources in multiple languages.

The telephone speech audio recordings were collected to support the NIST 2011 LRE which focused on pair discrimination for 24 languages/dialects. These recording are also contained in Multi-Language Conversational Telephone Speech 2011 – Slavic Group LDC2016S11. The goal of NIST's LRE series is to establish the baseline of current performance capability for language recognition of conversational telephone speech and to lay the groundwork for further research efforts in the field.

Althingi Parliamentary Speech, LDC2021S01

Althingi Parliamentary Speech consists of approximately 542 hours of recorded speech from 197 speakers from Althingi, the Icelandic Parliament, along with corresponding transcripts, a pronunciation dictionary and two language models.

This dataset was collected in 2016 by the ASR for Althingi project at Reykjavik University in collaboration with the Althingi speech department. The purpose of the project was to develop an ASR (automatic speech recognition) system for parliamentary speech to replace the procedure of manually transcribing performed speeches.

CALLFRIEND American English-Southern Dialect Second Edition, LDC2020S08

Developed by LDC in support of language identification technology development, CALLFRIEND American English-Southern Dialect Second Edition consists of approximately 26 hours of unscripted telephone conversations between native speakers of Southern dialects of American English. This second edition updates the audio files to wav format, simplifies the directory structure and adds documentation and metadata. Participants could speak with a person of their choice on any topic for up to 30 minutes.

Columbia Games Corpus, LDC2021S02

Columbia Games Corpus was developed by the Spoken Language Group, Columbia University and the Department of Linguistics, Northwestern University. It consists of approximately 10 hours of spontaneous English conversation along with corresponding orthographic transcripts and annotation. Speech recordings are comprised of two subjects playing a series of computer games requiring verbal communication to achieve joint goals of identifying and moving images on the screen to reach a combined number of points.

Each player used a separate laptop computer and could not see the screen of the other player. Participants played two games: the Cards Game and the Objects Game. In the Cards Game, one participant described a card and depending on the task in the game, the second participant searched for the described card or tried to match it from cards shown on their screen. In the Objects Game, each player's screen displayed 5-7 objects, one of which was the target object. One player described the target object's location on their screen, and the other player tried to move that object to the same position on their screen. 13 subjects participated in the collection.

Ethnobotanical Research and Language Documentation of Nahuatl, LDC2021S06

Ethnobotanical Research and Language Documentation of Nahuatl consists of approximately 190 hours of field recordings collected in the Sierra Nororiental and Sierra Norte regions of Puebla, Mexico. The corpus contains audio and video recordings of native Nahuatl speakers during the collection of particular plants; partial transcripts (Nahuatl and Spanish); a Highland Puebla Nahuat dictionary; botanical and ethnobotanical data; and speaker metadata.

Nahuatl is one of the most widely spoken indigenous languages in the Americas with approximately 1.5 million speakers in Mexico. Many distinct and sometimes mutually intelligible varieties have been recognized. The recordings in this release were collected between 2008 and 2019 in two different municipalities: Cuetzalan del Progreso and Tepetzintla. Speech from Cuetzalan represents Highland Puebla Nahuat, and speech from Tepetzintla represents Zacatlán-Ahuacatlám-Tepetzintla Nahuatl.

Global TIMIT Mandarin Chinese-Guanzhong Dialect, LDC2020S12

Developed by LDC and Xi'an Jiaotong University, Global TIMIT Mandarin Chinese-Guanzhong Dialect consists of approximately five hours of read speech and transcripts in the Guanzhong dialect of Mandarin Chinese as spoken in Shannxi province. 50 speakers read 120 sentences selected from Chinese Gigaword Fifth Edition (LDC2011T13). 20 sentences were read by all speakers, 40 sentences were read by 10 speakers, and 60 sentences were read by one speaker, for a total of 3220 sentence types.

The Global TIMIT project aimed to create a series of corpora in a variety of languages with a similar set of key features as in the original TIMIT Acoustic-Phonetic Continuous Speech Corpus (LDC93S1) which was designed for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. These features included a large number of fluently read sentences containing a representative sample of phonetic, lexical, syntactic, semantic, and pragmatic patterns.

Global TIMIT Thai, LDC2022S13

Developed by LDC, Global TIMIT Thai consists of approximately 12 hours of read speech and time-aligned transcripts in Standard Thai. 50 speakers (33 female, 17 male) read 120 sentences selected from the Thai National Corpus, the Thai Junior Encyclopedia, and Thai Wikipedia, for a total of 6,000 utterances. Among the 120 sentences, 24 sentences were read by all speakers, 300 sentences were read by 10 speakers, and 1,800 sentences were read by one speaker, for a total of 2,124 sentence types.

The Global TIMIT project aimed to create a series of corpora in a variety of languages with a similar set of key features as in the original TIMIT Acoustic-Phonetic Continuous Speech Corpus (LDC93S1) which was designed for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. These features included a large number of fluently read sentences containing a representative sample of phonetic, lexical, syntactic, semantic, and pragmatic patterns.

MASRI Synthetic, LDC2022S08

MASRI (Maltese Automatic Speech Recognition I) Synthetic was developed by the MASRI team at the University of Malta and consists of approximately 99 hours of synthesized Maltese speech. Source sentences were extracted from the Maltese Language Resource Server (MLRS) corpus, comprised of written or transcribed Maltese covering various genres, including parliamentary debates, news, law, opinion, sports, culture, academic, literature and religious texts.

Mixer 3 Speech, LDC2023S02

Developed by LDC, Mixer 3 Speech contains 3,200 hours of audio recordings of conversational telephone speech involving 3,875 speakers and 26 distinct languages. This material was collected by LDC from 2005-2007 as part of the Mixer project, and recordings in this corpus were used in NIST Speaker Recognition Evaluation (SRE) and NIST Language Recognition Evaluation (LRE) corpora, including 2006 SRE and 2007 LRE.

The audio recordings were generated using LDC's computer telephony system capable of collecting speech from the telephone network. Recruited speakers were connected through a robot operator to carry on casual conversations lasting up to 10 minutes. Subjects fluent in languages other than English were asked to complete at least one non-English call.

Mixer 7 Spanish Speech, LDC2023S04

Mixer 7 Spanish Speech was developed by LDC and contains 9,600 hours of audio recordings of interviews, transcript readings and conversational telephone speech involving 191 distinct native Spanish speakers. This material was collected by LDC in 2011 and 2012 as part of the Mixer project. The recordings in this corpus were used in the 2012 NIST Speaker Recognition Evaluation test set.

The telephone collection protocol was similar to other LDC Mixer collections: recruited speakers were connected through a robot operator to carry on casual conversations lasting up to 10 minutes, usually about a daily topic announced by the robot operator at the start of the call. The raw digital audio content for each call side was captured as a separate channel, and each full conversation is presented as a 2-channel interleaved audio file, with 8000 samples/second and u-law sample encoding. Each speaker was asked to complete 15 calls.

The multi-microphone portion of the collection utilized 14 distinct microphones installed identically in two multi-channel audio recording rooms at LDC. Each session was guided by collection staff using prompting and recording software to conduct the following activities: (1) repeat questions (less than one minute); (2) informal conversation (typically 15 minutes); (3) transcript reading (15 minutes); and (4) up to three telephone calls under varying conditions (10 minutes). The 14 channels were recorded synchronously into separate single-channel files, using 16-bit PCM sample encoding at 16000 samples/second.

MyST Children's Conversational Speech, LDC2021S05

MyST (My Science Tutor) Children's Conversational Speech was developed by Boulder Learning Inc. It is comprised of approximately 470 hours of English speech from 1371 students in grades 3-5 conversing with a virtual science tutor in eight areas of science instruction, along with transcripts and a pronunciation dictionary.

Data was collected in two phases between 2008 and 2017. In both phases, spoken dialogs with the virtual tutor were aligned to classroom instruction using the Full Option Science System (FOSS), a research-based science curriculum for grades K-8. The eight FOSS science modules represented in this data set consisted of an average of 16 small-group classroom science investigations. Following the investigations, students conversed with the virtual science tutor for 15-20 minutes. The tutor asked open-ended questions about media presented on-screen, and students produced spoken answers.

Speech data was collected in 10,496 sessions for a total of 227,567 utterances. Approximately 45% of those utterances (102,433) were transcribed. All data collected in Phase I was transcribed using rich transcription guidelines; data collected in Phase II was partially transcribed using a reduced version of those guidelines. The transcription guidelines are included.

RATS Speaker Identification, LDC2021S08

Developed by LDC, RATS Speaker Identification contains approximately 1,900 hours of Levantine Arabic, Farsi, Dari, Pashto and Urdu conversational telephone speech with annotations of speech segments. The audio was retransmitted over eight channels, making 17,000 hours of total audio. The corpus was created to provide training and development sets for the Speaker Identification (SID) task in the DARPA RATS (Robust Automatic Transcription of Speech) program.

The goal of the RATS program was to develop human language technology systems capable of performing speech detection, language identification, speaker identification and keyword spotting on the severely degraded audio signals that are typical of various radio communication channels, especially those employing various types of handheld portable transceiver systems. To support that goal, LDC assembled a system for the transmission, reception and digital capture of audio data that allowed a single source audio signal to be distributed and recorded over eight distinct transceiver configurations simultaneously. Those configurations included three frequencies -- high, very high and ultra high --

variously combined with amplitude modulation, frequency hopping spread spectrum, narrow-band frequency modulation, single-side-band or wide-band frequency modulation. Annotations on the clear source audio signal, e.g., time boundaries for the duration of speech activity, were projected onto the corresponding eight channels recorded from the radio receivers.

Samrómur Icelandic Speech 1.0, LDC2022S05

Samrómur Icelandic Speech 1.0 was developed by the Language and Voice Lab, Reykjavik University in cooperation with Almannarómur, Center for Language Technology. The corpus contains 145 hours of Icelandic prompted speech from 8,392 speakers representing 100,000 utterances. Speech data was collected between October 2019 and May 2021 using the Samrómur website which displayed prompts to participants. The prompts were mainly from The Icelandic Gigaword Corpus, which includes text from novels, news, plays, and from a list of location names in Iceland. Additional prompts were taken from the Icelandic Web of Science and others were created by combining a name followed by a question or a demand. Prompts and speaker metadata are included in the corpus.

This version 1.0 is equivalent to "Samrómur Icelandic Speech 21.05" as used by the Language Technology Programme for Icelandic 2019-2023.

Spoken Digits in Hindi and Indian English, LDC2022S03

Developed by the Birla Institute of Technology and Science Pilani, Spoken Digits in Hindi and Indian English contains approximately two hours of speech comprised of spoken digits from one to ten in Hindi and English with regional accents from across India.

The speech data was collected as follows: in person, on a mobile handset recorder app; via one-to-one online communications over social apps; and from social media sites. Each audio file represents a single spoken digit in either Hindi or Indian English. Background noise was mostly retained. Some data was recorded in a noise-free environment or cleaned after recording to avoid abrupt noises such as car horns.

A Google Colab Notebook file which can be used for basic functionalities such as removing noise or unwanted spaces is also included.

The SSNCE Database of Tamil Dysarthric Speech, LDC2021S04

The SSNCE Database of Tamil Dysarthric Speech was developed by the Speech Lab, SSN College of Engineering, India, in collaboration with the Indian National Institute of Empowerment of Persons with Multiple Disabilities (NIEPMD) and contains approximately eight hours of Tamil speech data, time-aligned transcripts and metadata collected from 30 speakers (20 dysarthric speakers and 10 non-dysarthric speakers).

Dysarthria is a speech disorder caused by muscle weakness which can result in slowed and slurred speech that is difficult to understand. Common causes of dysarthria include nervous system disorders and conditions that cause facial paralysis or tongue or throat muscle weakness.

The speech data was collected between 2015 and 2017 in two sessions at NIEPMD. In total, each speaker recorded 365 utterances consisting of single words and of sentences that included a combination of common and uncommon Tamil phrases.

UCLA Speaker Variability Database, LDC2021S09

Developed by the UCLA Speech Processing and Auditory Perception Laboratory, UCLA Speaker Variability Database is comprised of approximately 34 hours of English speech and orthographic transcripts.

This corpus was designed to sample variability in speaking within individual speakers and across a large number of speakers. 202 participants representing a variety of language backgrounds took part in six different tasks: vowel sounds, reading sentences, giving instructions, neutral conversation, happy conversation, a phone conversation, annoyed conversation, and responding to a video. Speaker metadata is included.

Wikipedia Spanish Speech and Transcripts, LDC2021S07

Wikipedia Spanish Speech and Transcripts consists of approximately 25 hours of Spanish read speech and transcripts. The read text was taken from the Spanish version of WikiProject Spoken Wikipedia, referred to as Wikipedia Grabada. The audio is comprised of short recordings from Wikipedia articles read by 193 speakers and the audio files were segmented and transcribed by native Spanish speakers.