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 – 5th Release, LDC2019S17
Corpus Descriptions

2011 NIST Language Recognition Evaluation Test Set, LDC2018S06

2011 NIST Language Recognition Evaluation Test Set contains selected training data and the evaluation test set for the 2011 NIST Language Recognition Evaluation. It consists of approximately 204 hours of conversational telephone speech and broadcast audio collected by the Linguistic Data Consortium (LDC) in the following 24 languages and dialects: Arabic (Iraqi), Arabic (Levantine), Arabic (Maghrebi), Arabic (Standard), Bengali, Czech, Dari, English (American), English (Indian), Farsi, Hindi, Lao, Mandarin, Punjabi, Pashto, Polish, Russian, Slovak, Spanish, Tamil, Thai, Turkish, Ukrainian and Urdu.

The 2011 evaluation emphasized the language pair condition and involved both conversational telephone speech (CTS) and broadcast narrow-band speech (BNBS).

This release includes training data for nine language varieties that had not been represented in prior LRE cycles -- Arabic (Iraqi), Arabic (Levantine), Arabic (Maghrebi), Arabic (Standard), Czech, Lao, Panjabi, Polish and Slovak -- contained in 893 audited segments of roughly 30 seconds duration and in 400 full-length CTS recordings. The evaluation test set comprises a total of 29,511 audio files, all manually audited at LDC for language and divided equally into three different test conditions according to the nominal amount of speech content per segment.

AISHELL-1, LDC2018S14

Developed by Beijing Shell Shell Technology Co., Ltd, AISHELL-1 contains approximately 520 hours of Chinese Mandarin speech from 400 speakers recorded simultaneously on three different devices with associated transcripts. The goal of the collection was to support speech recognition system development in domains such as smart homes, autonomous driving, entertainment, finance, and science and technology.

Participants read 500 sentences chosen for their speech and phonetic characteristics and were recorded in a quiet indoor environment on a high-fidelity microphone and two mobile phones (Android and IOS).

Speakers were recruited from different accent areas across China, including North, South and Yue-Gui-Min regions. There were 214 female speakers and 186 male speakers. Additional demographic information about the participants is included in this release.

Avatar Education Portuguese, LDC2018S15

Developed by the University of Pernambuco, Avatar Education Portuguese contains approximately 80 minutes of Brazilian Portuguese microphone speech with phonetic and orthographic transcriptions. The data was developed for Avatar Education, an animated virtual assistant designed to enhance communication and interaction in educational contexts, such as online learning.

This release contains 1,400 utterances (700 male and 700 female) of read and spontaneous speech spoken by two professional speakers. Utterances were transcribed at the word level (without time alignments) and at the phoneme level (with time alignment labels). The acoustic environment was controlled to allow for background conditions that occur in application environments.

CALLFRIEND Vietnamese, LDC96S60

CALLFRIEND Vietnamese was developed by LDC and consists of approximately 60 unscripted telephone conversations between native speakers of Vietnamese. The duration of each conversation was between 5-30 minutes. The corpus also includes documentation describing speaker information (sex, age, education, callee telephone number) and call information (channel quality, number of speakers).

The CALLFRIEND project covered fifteen languages and supported the development of language identification technology. For each language collection, participants were native speakers of the target language. All calls were made inside the continental United States and Canada.

CIEMPIESS Experimentation, LDC2019S07

CIEMPIESS (Corpus de Investigación en Español de México del Posgrado de Ingeniería Eléctrica y Servicio Social) Experimentation was developed at the National Autonomous University of Mexico (UNAM) and consists of approximately 22 hours of Mexican Spanish broadcast and read speech with associated transcripts. The goal of this work was to create acoustic models for automatic speech recognition.

This corpus is comprised of three different data sets, specifically Complementary, Fem and Test. Complementary is a phonetically-balanced corpus of isolated Spanish words spoken in Central Mexico. Fem contains broadcast speech from 21 female speakers, collected to balance by gender the number of recordings from male speakers in other CIEMPIESS collections. Test consists of 10 hours of broadcast speech and transcripts and is intended for use as a standard test data set alongside other CIEMPIESS corpora.

The CMU Kids Corpus, LDC97S63

The CMU Kids Corpus was developed in 1995-1996 and is a database of sentences read aloud by 76 children, totaling 5,180 utterances. This data set was designed as a training set of children's speech for the SPHINX II automatic speech recognizer in the LISTEN project at Carnegie Mellon University.

The child participants (24 male, 52 female) ranged in age from 6-11 years old and were in the first through third grades in Pittsburgh at the time of recording. They represented two separate populations: 44 readers who provided “good examples” of reading aloud, and 32 readers who supplied examples of errorful reading and dialectic variants. Transcripts are included in this release.

CSLU: Portland Cellular Telephone Speech Version 1.3, LDC2008S01

Created by the Center for Spoken Language Understanding (CSLU) at Oregon Health and Science University, CSLU: Portland Cellular Telephone Speech Version 1.3 is a collection of cellular telephone speech (7,571 utterances) and corresponding orthographic and phonetic transcriptions.

Speech was collected from 515 different speakers calling the CSLU data collection system on cellular telephones. Callers were asked to repeat certain phrases and respond to other prompts. Two prompt protocols were used: an In Vehicle Protocol for speakers calling from inside a vehicle and a Not in Vehicle Protocol for those calling from outside a vehicle. Each protocol contained distinct queries designed to probe the conditions of the caller's in vehicle/not in vehicle surroundings.

DIRHA English WSJ Audio, LDC2018S01

DIRHA English WSJ Audio is comprised of approximately 85 hours of real and simulated read speech by six native American English speakers. It was developed as part of the Distant-Speech Interaction for Robust Home Applications (DIRHA) Project, which addressed natural spontaneous speech interaction with distant microphones in a domestic environment.

This release contains signals of different characteristics in terms of noise and reverberation, making it suitable for various multi-microphone signal processing and distant speech recognition tasks.

Speech was collected in a real apartment setting with typical domestic background noise and inter/intra-room reverberation effects. A total of 32 microphones were placed in the living-room (26 microphones) and in the kitchen (6 microphones). The target utterances were taken from CSR-I (WSJ0) Complete (LDC93S6A), specifically, the 5,000 word subset of read speech from Wall Street Journal news text.

Annotations for each acoustic sequence - such as microphone positions, speaker id, speaker gender and speaker position – and additional metadata about the speakers and images of the apartment setting are provided.

The DKU-JNU-EMA Electromagnetic Articulography Database, LDC2019S14

The DKU-JNU-EMA Electromagnetic Articulography Database was developed by Duke Kunshan University and Jinan University and contains approximately 10 hours of articulography and speech data in Mandarin, Cantonese, Hakka, and Teochew Chinese from two to seven native speakers for each dialect.

Articulatory measurements were made using the NDI electromagnetic articulography wave research system to capture real-time vocal tract variable trajectories. Subjects had six sensors placed in various locations in their mouth and one reference sensor was placed on the bridge of their nose. For simultaneous recording of speech signals, subjects also wore a head-mounted close-talk microphone.

Speakers engaged in four different types of recording sessions: one in which they read complete sentences or short texts, and three sessions in which they read related words of a specific common consonant, vowel or tone.

Emotional Prosody Speech and Transcripts, LDC2002S28

Emotional Prosody Speech and Transcripts was developed by LDC and contains audio recordings and corresponding transcripts, designed to support research in emotional prosody and collected over an eight-month period in 2000-2001. The recordings consist of professional actors reading a series of semantically neutral utterances (dates and numbers) spanning 14 distinct emotional categories, selected

after Banse & Scherer's study of vocal emotional expression in German. (Banse, R. & Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology, 70, 614-636.)

The Prosody Recordings Project was interested in capturing the aspects of speech (emotion, intonation) left out of the written form of a message. In these experiments, simple phrases were expressed to reflect varied contexts, for example, to answer different questions, to address listeners at different distances from the speaker, or to express different emotional states.

The data set consists of 15 recordings in sphere format and their transcripts. The original session recordings are provided in their entirety, including informal chit-chat and discussion between each emotion category elicitation task. Time alignment is limited to utterances within the formal elicitation tasks and miscellaneous regions have been marked as such.

First DIHARD Challenge Development - Eight Sources, LDC2019S09

First DIHARD Challenge Development - Eight Sources was developed by LDC and contains approximately 17 hours of English and Chinese speech data along with corresponding annotations used in support of the First DIHARD Challenge. This release, when combined with First DIHARD Challenge Development - SEEDLingS (LDC2019S10), contains the development set audio data and annotation (diarization, segmentation) as well as the official scoring tool.

The First DIHARD Challenge was an attempt to reinvigorate work on diarization through a shared task focusing on "hard" diarization; that is, speech diarization for challenging corpora where there was an expectation that existing state-of-the-art systems would fare poorly. As such, it included speech from a wide sampling of domains representing diversity in number of speakers, speaker demographics, interaction style, recording quality, and environmental conditions, including, but not limited to: clinical interviews, extended child language acquisition recordings, YouTube recordings, and conversations collected in restaurants.

IARPA Babel Zulu Language Pack IARPA-babel206b-v0.1e, LDC2017S19

IARPA Babel Zulu Language Pack IARPA-babel206b-v0.1e was developed by Appen for the IARPA (Intelligence Advanced Research Projects Activity) Babel program. It contains approximately 211 hours of Zulu conversational and scripted telephone speech collected in 2012 and 2013 along with corresponding transcripts.

The Babel program focuses on underserved languages and seeks to develop speech recognition technology that can be rapidly applied to any human language to support keyword search performance over large amounts of recorded speech.

The Zulu speech in this release represents that spoken in the KZN (KwaZulu-Natal)-urban dialect region of South Africa. The gender distribution among speakers is approximately equal; speakers' ages range from 16 years to 70 years. Calls were made using different telephones (e.g., mobile, landline) from a variety of environments including the street, a home or office, a public place, and inside a vehicle.

ICSI Meeting Speech, LDC2004S02

ICSI Meeting Speech contains approximately 72 hours of speech from 53 unique speakers in 75 meetings collected at Berkeley's International Computer Science Institute (ICSI) in 2000-2002. The recordings were made during regular weekly meetings of various ICSI working teams, including the team working on the ICSI Meeting Project. The speech files range in length from 17 to 103 minutes, but in general are less than one hour each. Word-level orthographic transcriptions are available as ICSI Meeting Transcripts (LDC2004T04).

The meetings were simultaneously recorded using close-talking microphones for each speaker (generally head-mounted, but early meetings included lapel microphones), as well as six table-top microphones: four high-quality omnidirectional PZM microphones arrayed down the center of the conference table, and two inexpensive microphone elements mounted on a mock PDA. All meetings were recorded in the same instrumented meeting room.

Meetings involved from 3-10 participants, averaging 6 attendees per meeting. The corpus includes non-native English speakers, varying in fluency from nearly-native to challenging-to-transcribe.

Malto Speech and Transcripts, LDC2012S04

Malto Speech and Transcripts contains approximately 8 hours of Malto speech data collected between 2005 and 2009 from 27 speakers (22 males, 5 females), accompanying transcripts, English translations and glosses for 6 hours of the collection. Speakers were asked to talk about themselves, their lives, rituals and folklore; elicitation interviews were then conducted. The goal of the work was to present the current state and dialectal variation of Malto.

Malto is a Dravidian language spoken in northeastern India (principally the states of Bihar, Jharkhand and West Bengal) and Bangladesh by people called the Pahariyas. Indian census data places the number of Malto speakers in a range of between 100,000-200,000 total speakers. The transcribed data accounts for 6 hours of the collection and contains 21 speakers (17 male, 4 female). The untranscribed data accounts for 2 hours of the collection and contains 10 speakers (9 male, 1 female). Four of the male speakers are present in both groups. All audio is presented in .wav format. Each audio file name includes a subject number, village name, speaker name and the topic discussed.

Multi-Language Conversational Telephone Speech 2011 -- Central European, LDC2018S08

Multi-Language Conversational Telephone Speech 2011 -- Central European was developed by the Linguistic Data Consortium (LDC) and is comprised of approximately 44 hours of telephone speech in two distinct language varieties of Central Europe: Czech and Slovak. The data was collected to support research and technology evaluation in automatic language identification, specifically language pair discrimination for closely related languages/dialects. Portions of these telephone calls were used in the NIST 2011 Language Recognition Evaluation.

Calls were made using LDC's telephone collection infrastructure. Participants were recruited by native speakers who contacted acquaintances in their social network. Those native speakers made one call, up to 15 minutes, to each acquaintance. Human auditors labeled calls for gender, dialect type, and noise.

N4 NATO Native and Non-Native Speech, LDC2006S13

N4 NATO Native and Non-Native Speech corpus was developed by the NATO research group on Speech and Language Technology in order to provide a military-oriented database for multilingual and non-native speech processing studies. It consists of 115 native and non-native speakers using NATO English procedure between ships and reading from a text, "The North Wind and the Sun," in both English and the speaker's native language.

Speech data was recorded in the naval transmission training centers of four countries (Germany, The Netherlands, United Kingdom, and Canada) during naval communication training sessions in 2000-2002. The duration of Navy procedure recordings ranges from 1.3 to 2.3 hours, for a total of 7.5 hours. The duration of the native language text readings ranges from 1.5 to 22.9 minutes, for a total of approximately one hour.

The corpus contains the following information about each speaker: gender, age, weight, height, possible speaking or hearing disorders, education level, living area, accent, second language, and the year English was learned (for non-native speakers).

RATS Language Identification, LDC2018S10

RATS Language Identification was developed by the Linguistic Data Consortium (LDC) and is comprised of approximately 5,400 hours of Levantine Arabic, Farsi, Dari, Pashto and Urdu conversational telephone speech with annotation of speech segments. The corpus was created to provide training, development and initial test sets for the Language Identification (LID) 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.

Conversational telephone speech recordings were audited by annotators who listened to short segments and determined whether the audio was in the target language. Annotations on the audio files include start time, end time, speech activity detection (SAD) label, SAD provenance, language ID and LID provenance.

Turkish Broadcast News Speech and Transcripts, LDC2012S06

Turkish Broadcast News Speech and Transcripts was developed by Boğaziçi University, Istanbul, Turkey and contains approximately 130 hours of Voice of America (VOA) Turkish radio broadcasts and corresponding transcripts. This is part of a larger corpus of Turkish broadcast news data collected and

transcribed with the goal to facilitate research in Turkish automatic speech recognition and its applications. The VOA material was collected between December 2006 and June 2009 using a PC and TV/radio card setup. The data collected during the period 2006-2008 was recorded from analog FM radio; the 2009 broadcasts were recorded from digital satellite transmissions. A quick manual segmentation and transcription approach was followed.

The data was recorded at 32 kHz and resampled at 16 kHz. After screening for recording quality, the files were segmented, transcribed, and verified. The segmentation occurred in two steps, an initial automatic segmentation followed by manual correction and annotation which included information such as background conditions and speaker boundaries.

Vehicle City Voices Corpus – Part I, LDC2017S17

Vehicle City Voices Corpus – Part I was developed at the University of Michigan-Flint, and is an ongoing oral history project and survey of English language variation in Flint, Michigan. It contains approximately 16 hours of speech with corresponding transcripts from 21 interviews of Flint residents conducted between 2012 and 2015. The corpus was designed to provide high-quality recordings for acoustic analysis and to examine narrative structure and discursive construction of individual and collective identity in urban spaces.

Participants (11 female, 10 male) were born between 1935 and 1991 and represented a range of ages, genders, and ethnicities. Of the interviewees, 11 were Black/African American, 8 were White/Caucasian, and 2 were biracial/mixed ethnic heritage. Questions focused on recollections of important community events, remembrances about the community, the interviewee's relationship to the auto industry and the city's physical transformation, among other topics.

Metadata (where provided by participants) includes information on gender, ethnicity, year of birth, level of education, field of employment, average income, length of time living in Flint and its surrounding areas, as well as interviewer age, gender, and ethnicity. In addition, original interview durations, edited interview durations, interview year, and transcript word counts are also provided in the metadata file.