Crowdsourced high-quality Colombian Spanish [es-co] multi-speaker speech dataset

research.google/tools/datasets/colombian-spanish-tts/

This dataset was created for speech research purposes and contains about 4,900 recordings of participants reading a script in Spanish as spoken in Colombia, one sentence at a time. Each example contains the audio files and the associated text. The audio is high-quality (16-bit, 48kHz) recorded in a quiet environment using cardioid condenser microphone. The dataset is multi-speaker, containing recordings from 33 volunteers (male and female), where each volunteer contributed up to 150 recordings. The recordings took place in Bogota, Colombia in 2018.

PUBLISHER(S)

Google LLC

INDUSTRY TYPE

Corporate - Tech

KEY APPLICATIONS

Machine Learning, Speech Technology

INTENDED USE CASE(S)

Multi-speaker and multi-lingual model speech synthesis models building
Evaluating dialects affects on speech recognition models
Linguistic research

PRIMARY DATA TYPE

Speech data

DATASET FUNCTION(S)

Training, Testing

DATASET CHARACTERISTICS

Number of recorded lines	4,903 lines
Avg. number of lines per participant	148.6 lines
Avg. number of words in script	9.4 words
Number of participants	33 participants
Total length of recordings	7h 34m 46s
Avg. length of recordings	5.6 s
Avg. recording file size	526 kB
Human verified instances	all
Recording format	WAVE, PCM 16-bit mono at 48 kHz

NATURE OF CONTENT

The dataset contains recordings of Spanish as spoken in Colombia in 2018. The participants read a script, approximately one sentence per file. The data is delivered in audio files and the associated transcription of the audio. All the script lines are listed with the corresponding audio files in a file named line_index.tsv, which has two columns. The first column contains the FileID of the file, and second the column contains the text read in the corresponding audio file. The columns are tab separated.

EXAMPLE COMPONENTS

The file line_index.tsv gives a transcription of each audio file in the following format:

Audio: FileID	cof_12345_1234
Script: Text	Me gusta la idea

DESCRIPTIONS OF EXAMPLE COMPONENTS

cof_12345_1234 is the FileID of the file containing the Text in the line. The FileID is composed of three parts, delimited by an underscore "_". The first part is unique for the dataset and gender, the second part is a unique identification of the user, and the third is a unique number for the file.

LICENSE TYPE(S)

CC-BY-4.0-SA

FIRST RELEASED	August 2019
CURRENT VERSION	Version 1
MAINTENANCE STATUS	Limited maintenance
CHANGE LOG	N/A
ACCESS LINK	research.google/tools/datasets/colombian-spanish-tts/

ATTRIBUTION

Crowdsourced high-quality Colombian Spanish [es-co] multi-speaker speech dataset, by Google LLC. Available at https://research.google/tools/datasets/colombian-spanish-tts/ under Creative Commons Commons Attribution 4.0 Share Alike.

CC BY-SA 4.0 license

DATA COLLECTION METHOD(S)

Scripts: Compensated Workers

DATA SOURCE(S)

Scripts Generated by dataset publishers.

DATA SOURCE(S) DESCRIPTION

Compensated workers, native Spanish speakers located in USA and Mexico. No further demographic information can be reported on the workers as the sample size is limited.

DATA COLLECTION PROCEDURE

The initial set was created based on
• Internally collected conversational recordings.
• About 30 sentences which were generated by hand to contrast phenomena in different dialects in Spanish as spoken in Latin America.

DATA SELECTION

Lines were randomized and assigned to each user for the recordings. Each script contained a subset of the 30 contrasting lines.

Audio: Crowdsourced

DATA SOURCE(S)

Recorded audio from volunteers.

DATA SOURCE DESCRIPTION(S)

Volunteers which were reached with the help of Google employees and Google Developer Groups in Bogota, Colombia.

DATA SOURCE DISTRIBUTION: GEOGRAPHIC

Volunteers in Bogota, Colombia. Only self reported gender information was collected. All volunteers were older than 21 when the data collection was performed.

DATA SOURCE DISTRIBUTION: GENDER

Female	48.3%
Male	51.7%

DATA COLLECTION PROCEDURE

The recordings were performed in a quiet environment using a Neumann KM-184 microphone, Blue Icicle USB XLR A/D converter and an Asus Fanless laptop using proprietary software.

DATA SELECTION

Other than the age limits on the participants, no further limitations were in place.

FILTERING CRITERIA

No filtering was done on the audio during the data collection.

SAMPLING METHOD(S)

Scripts: Unsampled
Audio: Unsampled

SAMPLING TASK(S)	N/A
SAMPLING DESCRIPTION(S)	N/A
SAMPLING POLICY SUMMARY	N/A

VALIDATION METHOD(S):

Scripts: Not validated

VALIDATION TASK(S)	N/A
VALIDATION DESCRIPTION(S)	N/A
VALIDATOR CHARACTERISTICS	N/A
VALIDATION POLICY SUMMARY	N/A

Audio: Human Verified

VALIDATION TASK(S)

Validate audio quality
Validate the text matches the audio.

VALIDATION DESCRIPTION(S)

Validate that the audio files, and double check that the script represent the recorded audio.

VALIDATOR CHARACTERISTICS

The same workers were used for the validation as for the script generation.

EXCLUDED DATA

Any collected data that did not pass validation procedures has been excluded.

VALIDATION POLICY SUMMARY

Validate that the audio is audible, that no audio flaws such as very loud background noises, and major disfluencies were not present such as coughing and sneezing. The workers also validated that the audio recorded and the text matched. When the mismatches could be fixed, the scripts were updated to reflect the audio.

Each line was validated by one worker.

VALIDATOR TRAINING SUMMARY

Validators did not get any training other than for using the tool to perform the validation. The validators were native Spanish speakers.

Data Card Version	1.0
Data Card Template Version	1.0

Submit Feedback