Title	Synthetic Call Centre Data
Creator / Responsible Partner	TID
Dataset Identifier	01_Synthetic CallCentresTID_I-BiDaaS_WP2_D2.1_v0.1.tar.gz
DOI	[10.5281/zenodo.4274454](https://doi.org/10.5281/zenodo.4274454)[1]
Dataset Description	This is a simulated dataset based on both real phone interactions and conversations usually performed by real call centres. It is comprised of several simulated customer interactions with an agent representative, both roles performed by actors. Phone call recordings are performed using different mobile and landline devices. The scripting, both from customer and agent, aims to develop typical scenarios by telco-oriented call centre operations. Both raw waveform recordings and speech transcription are provided. The latter as obtained by an automatic speech recognition (ASR) prototype developed by TID. The word segmentation timestamps are also provided for those recognized. Additionally, a confidence score is also provided per token basis.
Work Package/Deliverable	WP2/D2.1
Source	TID determined the format for data by using speech processing and audio standards. Business units from Telefónica oriented on how to simulate typical use cases and scenarios.
Processing	Original recorded audio files were processed by an Automatic Speech Recognizer developed by TID to obtain word transcripts. The ASR lexicon and language model were both tuned and augmented for the specific call centre domain.
Repository	Zenodo
Language	Spanish
Code list	Each transcript file, with extension "ctm", contains word segmentation timestamps. Each line in ctm file stands for: • filename: ScriptXX.ctm, where each X is an integer number within [0-9] • start time (seconds): 12.34 • duration (seconds): 0.54 • token: please • confidence (0-1): 0.92 With the corresponding raw audio file ScriptXX.wav
Type	Numerical, Categorical
Format	RIFF (little-endian) data, WAVE audio, Microsoft PCM, 16 bit/sample, mono 8000 Hz CTM format transcripts as NIST standard
Expected Size	37.1MB
Keywords	speech, call centre, wav, transcripts
Version	V0.1

Date of Repository Submission	16/11/2020
Necessary software	N/A
Rights	TID
Access Information	Open