Release Date: 17.01.22

Welcome to *Common Phone 1.0*

Legal Information

Common Phone is a subset of the *Common Voice* corpus collected by *Mozilla Corporation*. By using *Common Phone*, you agree to the Common Voice Legal Terms. *Common Phone* is maintained and distributed by speech researchers from Pattern Recognition Lab of Friedrich-Alexander-University Erlangen-Nuremberg (FAU) under a CCo license.
Like for *Common Voice*, you must not make any attempt to identify speakers that contributed to *Common Phone*.

About *Common Phone*

This corpus aims to provide a basis for Machine Learning (ML) researchers and enthusiasts to train and test their models against a wide variety of speakers, hardware/software ecosystems and acoustic conditions to improve generalization and availability of ML in real-world speech applications.
The current version of *Common Phone* comprises 116,5 hours of speech samples, collected from 11.246 speakers in 6 languages:

Language	**Speakers**	**Hours**
	`train` / `dev` / `test`	`train` / `dev` / `test`
English	4716 / 771 / 774	14.1 / 2.3 / 2.3
French	796 / 138 / 135	13.6 / 2.3 / 2.2
German	1176 / 202 / 206	14.5 / 2.5 / 2.6
Italian	1031 / 176 / 178	14.6 / 2.5 / 2.5
Spanish	508 / 88 / 91	16.5 / 3.0 / 3.1
Russian	190 / 34 / 36	12.7 / 2.6 / 2.8
Total	8417 / 1409 / 1420	85.8 / 15.2 / 15.5

Presented `train`, `dev` and `test` splits are **not identical** to those shipped with *Common Voice*. Speaker separation among splits was realized by only using those speakers that had provided age

and gender information. This information can only be provided as a registered user on the website. When logged in, the session ID of contributed recordings is always linked to your user, thus we could easily link recordings to individual speakers. Keep in mind this would not be possible for unregistered users, as their session ID changes if they decide to contribute more than once.
During speaker selection, we considered that some speakers had contributed to more than one of the six *Common Voice* datasets (one for each language). In *Common Phone*, a speaker will only appear in one language.
The dataset is structured as follows:

- Six top-level directories, one for each language.
- Each language folder contains:
    - [train|dev|test].csv files listing audio files, respective speaker ID and plain text transcript.
    - meta.csv provides speaker information: age group, gender, language, accent (if available) and which of the three splits this speaker was assigned to. File names match corresponding audio file names except their extension.
    - /grids/ contains phonetic transcription for every audio file in Praat TextGrid format.
    - /mp3/ contains audio files in mp3, identical to those of *Common Voice*, e.g., sampling rates have been preserved and may vary for different files.
    - /wav/ contains raw audio files in 16 bits/sample, 16 kHz single channel. They had been created from the original mp3 audios. We provide them for convenience, keep in mind that their source had undergone MP3-compression.

Where does the phonetic annotation come from?

Phonetic annotation was computed via BAS Web Services. We used the regular Pipeline (G2P-MAUS) without ASR to create an alignment of text transcripts with audio signals. We chose International Phonetic Alphabet (IPA) output symbols as they work well even in a multi-lingual setup. *Common Phone* annotation comprises 101 phonetic symbols, including silence.

Why *Common Phone*?

- Large number of speakers and varying acoustic conditions to improve robustness of ML models
- Time-aligned IPA phonetic transcription for every audio sample
- Gender-balanced across age-groups
- Support for six different languages to leverage multi-lingual approaches
- Original MP3 files as shipped with CV, plus standard WAVE files

Is there any publication available?

Yes, a paper describing Common Phone in detail is currently under revision for LREC 2022. You can access a pre-print version on arXiv (scheduled for publication on 19.01.22) entitled “Common Phone: A Multilingual Dataset for Robust Acoustic Modelling”.