Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 763–768
Marseille, 20-25 June 2022

Common Phone: A Multilingual Dataset for Robust Acoustic Modelling

Philipp Klumpp[1], Tomás Arias-Vergara[1,2], Paula-Andrea Pérez-Toro[1,2]
Elmar Nöth[1], Juan Rafael Orozco-Arroyave[1,2]
[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, [2]Universidad de Antioquia
[1]Lehrstuhl für Mustererkennung, Martensstrasse 3, 91058 Erlangen, Germany
[2]Calle 67 # 53-108, UdeA Campus Principal, Medellín, Colombia
philipp.klumpp@fau.de

Abstract

Current state of the art acoustic models can easily comprise more than 100 million parameters. This growing complexity demands larger training datasets to maintain a decent generalization of the final decision function. An ideal dataset is not necessarily large in size, but large with respect to the amount of unique speakers, utilized hardware and varying recording conditions. This enables a machine learning model to explore as much of the domain-specific input space as possible during parameter estimation. This work introduces *Common Phone*, a gender-balanced, multilingual corpus recorded from more than 11,000 contributors of *Mozilla's Common Voice* project. It comprises around 116 hours of speech enriched with automatically generated phonetic segmentation. A Wav2Vec 2.0 acoustic model was trained with the *Common Phone* to perform phonetic symbol recognition and validate the quality of the generated phonetic annotation. The architecture achieved a PER of 18.1 % on the entire test set, computed with all 101 unique phonetic symbols, showing slight differences between the individual languages. We conclude that *Common Phone* provides sufficient variability and reliable phonetic annotation to help bridging the gap between research and application of acoustic models.

Keywords: Speech Dataset, Acoustic Modelling, Multilingual Corpus

1. Introduction

In the past two years, Wav2Vec (Schneider et al., 2019) and Wav2Vec 2.0 (Baevski et al., 2020) have leveraged the state-of-the-art of acoustic models to a new level, with the latter being able to achieve a Phoneme Error Rate (PER) of 8.3 % on the TIMIT (Garofolo et al., 1993) test set. TIMIT is among the most important corpora for acoustic model evaluation, specifically because of the precise, manually annotated, phonetic reference. All speakers had been recorded in equal acoustic conditions using the same microphone. For acoustic model validation, these conditions can be considered perfect. However, the constricted acoustic environment also implies limited robustness to altered recording conditions. The corpus is also not gender-balanced, with 439 (70 %) of the 630 speakers being male. For any application-driven project, the total number of speakers is quite small as well for modern standards.

A large variability of training samples is key to enable a deep architecture to explore as much of the input space as possible. Any perturbation in said space could result in the model performance to collapse (Wang et al., 2018). Deviations from the *known input-space* could be so small that it would be impossible for a human rater to perceive them, thus they might be exploited in scenarios like adversarial attacks (Schönherr et al., 2018; Hu et al., 2019). Robustness could be improved during training through techniques such as weight regularization, drop-out or batch-normalization (Kukačka et al., 2017). Another way would be to employ a dataset that provides a large variability with respect to acoustic conditions, recording hardware, contributing speakers, dialects and other parameters.

One such corpus could be *Common Voice* (Ardila et al., 2020) (CV). It is an ongoing initiative maintained by *Mozilla Foundation* that aims to collect spoken text samples from contributors of many different languages. Everyone could donate their speech to enrich the corpus via the project website [1]. The most recent release *7.0* from July 2021 comprised datasets in 76 different languages. Contributions could be made anonymously, additional information such as age and gender could be provided after registration on the website. Volunteers are not limited to donate their speech to help improve the quality of the corpus. They could also help to validate new speech donations, e. g. verify that the spoken text in an audio sample matches with the prompted text transcript. This crowd-based approach of automated donation and validation procedures enabled CV to collect data from a large amount of speakers. For example, the English *7.0* corpus comprises more than 75,000 different speakers.

While CV could be considered a decent corpus for any end-to-end automated speech recognition (ASR) task, there are several important drawbacks. First of all, CV provides only a text transcript as ground truth reference. For training and testing acoustic models, CV does not provide any phonetic transcript or segmentation. Furthermore, the distribution of speakers and speech samples is not ideal in many cases. The English CV 7.0 corpus for example comprised 45 % male but only 15 % female speakers. For the remaining contributions, gen-

[1] `https://commonvoice.mozilla.org/`

	Age < 19	19 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70 - 79	80 - 89
English	276	867	417	270	287	170	57	5
	276	870	421	274	290	173	58	5
German	24	127	94	46	71	28	5	0
	24	127	95	50	72	28	5	0
Spanish	72	263	110	76	55	12	0	0
	72	263	110	76	55	12	0	0
French	38	186	90	78	76	41	6	0
	38	186	90	78	77	41	6	0
Italian	9	82	53	42	42	22	4	0
	9	82	53	42	42	22	4	0
Russian	18	48	17	9	3	0	0	0
	18	48	17	9	3	0	0	0
Total	**437**	**1573**	**781**	**521**	**534**	**273**	**72**	**5**
	437	**1576**	**786**	**529**	**539**	**276**	**73**	**5**

Table 1: Speaker distribution among different age groups in the training set for female (top) and male (bottom).

	Age < 19	19 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70 - 79	80 - 89
English	45	135	65	47	47	29	10	0
	45	144	71	45	49	29	9	1
German	4	22	16	9	12	5	1	0
	4	21	17	8	13	5	1	0
Spanish	12	44	19	13	10	3	0	0
	12	44	19	13	10	3	0	0
French	7	31	15	13	13	7	2	0
	7	31	15	13	13	7	2	0
Italian	2	14	9	7	7	4	1	0
	2	14	9	7	7	4	1	0
Russian	3	8	3	2	1	0	0	0
	3	8	3	2	1	0	0	0
Total	**73**	**254**	**127**	**91**	**90**	**48**	**14**	**0**
	73	**262**	**134**	**88**	**93**	**48**	**13**	**1**

Table 2: Speaker distribution among different age groups in the development set for female (top) and male (bottom).

der information was unavailable. Another major problem is the number of samples certain speakers were able to contribute. In the official training split shipped with the previously mentioned English CV dataset, the most overrepresented speaker (according to ID) contributed more than 35,000 samples, which equaled 4.7 % of all training samples.

This paper introduces *Common Phone* (CP), a refined version of CV, which alleviates the aforementioned drawbacks to provide a corpus that meets modern machine learning (ML) requirements for acoustic modelling in a multi-lingual setup. After a brief summary of the structure of CP, the speaker selection process is explained, including an overview of speaker distributions in the entire dataset. Afterwards, the automated phonetic annotation procedure is described, as well as the utilized phonetic inventory. To validate the quality of the phonetic labels, we fine-tuned a Wav2Vec 2.0 model and tested on CP's test split. The resulting PERs showed that after training with CP, the model was able to reliably predict sequences of phonetic symbols across different languages.

2. Materials and Methods

2.1. Corpus structure

The structure of CP is very similar to that of CV. The directory of each language (English, German, Spanish, French, Italian and Russian) contains CSV-files for the respective train, development and test splits, and an additional one summarizing meta information of all speakers. Directory *mp3* contains the original recordings from CV. These recordings have not been altered in any way. It is important to notice that audio files in CV do not share a common sampling rate (we found 32, 44.1 and 48 kHz), thus varying values should be expected when working with the original recordings.

In an additional folder *wav*, raw PCM files were provided through simple decompression of their respective *mp3* counterparts. This was done for two main reasons: Firstly, all files could be converted to a format common in speech signal processing. We chose a sampling rate of 16 kHz, 16 bits depth and mono-channel configuration. Additionally, most existing ML environments and projects expect (or at least support) raw *wav* files as input. As the waveform had been reconstructed from a lossy compression (Brandenburg, 1999), it is not to be

	Age < 19	19 - 29	30 - 39	40 - 49	50 - 59	60 - 69	70 - 79	80 - 89
English	43 47	143 145	65 70	44 46	46 47	27 29	10 10	1 1
German	5 5	20 19	15 15	9 9	13 11	5 5	1 1	1 1
Spanish	12 12	44 44	19 19	13 13	10 10	3 3	1 1	1 1
French	7 7	31 31	16 16	13 13	13 13	7 7	2 2	0 0
Italian	2 2	13 14	9 9	8 8	8 8	4 4	1 1	0 0
Russian	4 4	8 8	3 3	2 2	1 1	0 0	0 0	0 0
Total	**73** **77**	**259** **261**	**127** **132**	**89** **91**	**91** **90**	**46** **48**	**15** **15**	**3** **3**

Table 3: Speaker distribution among different age groups in the test set for female (top) and male (bottom).

mistaken for a lossless version of CV recordings.

The folder *grids* contains *Praat* (Boersma and Van Heuven, 2001) TextGrids with word- and phonetic-level segmentation for every recording.

2.2. Speaker Selection

The main objective during speaker selection was to distribute data evenly among languages, genders and age groups, while at the same time keeping as many speakers as possible. Selection was done on the entire set of validated data in CV, omitting the original splits for training, development and test.

In a first step, all contributions that did not include information about age and gender had been removed. Not only did this help to keep track of speaker distributions, but it also allowed to assign speakers to only one of the three splits for training, development and test. CV assigns a session ID to every speech donation. If the same speaker donated samples through multiple sessions, it would be impossible to link all contributions to that same speaker. However, if a speaker was logged in to their account on the CV website, all contributions (even over varying sessions) would be linked to a static account ID. Meta information such as age and gender could only be provided through a user account. If that information was available, the recording was made as a logged-in user and the donation could always be linked to a particular contributor.

After this pre-selection, all speakers were assigned to a slot within an age-gender grid. From each age slot, contributors were randomly selected in pairs of female and male. The first pair was assigned to the test, the second to the development set, and another five pairs were partitioned into the training split. This procedure was repeated for every age slot until no more speaker pairs were available. After repeating this procedure, speakers were distributed as shown in Tables 1, 2 and 3 for training, development and test splits, respectively.

In the following step, samples were drawn from each speaker, such that there were as many unique uttered sentences as possible. The number of samples taken from a speaker differed among languages due to their uneven amounts of speakers. We chose to draw 2 (English), 9 (Spanish), 11 (French), 13 (German), 28 (Italian) and 80 (Russian) samples per speaker to ensure that the resulting corpus was not biased towards a particular language. If it was not possible to draw at least one speech sample with a unique text transcript for a particular speaker, that speaker was omitted. For each language, *meta.csv* provides a list of all speakers by their respective ID, which is identical to the one from CV. It summarizes a contributors age group, gender, information about a possible accent if available, and what split the contributor had been assigned to.

After sample selection, CP comprised 76,307 speech samples with 73,644 unique texts, totalling 116.5 hours of recorded audio collected from 11,246 unique speakers. The data distribution between languages and splits is summarized in Table 5.

2.3. Phonetic Inventory

The entire phonetic inventory used for CP is given in International Phonetic Alphabet (Association et al., 1999) (IPA) format and comprises a total of 101 symbols. Table 4 summarizes the core set of symbols, excluding the one for silence and 26 elongated variants. The presented IPA symbols are not to be mistaken for phonemes (Moore and Skidmore, 2019), but rather resemble a set of phones that sufficed to describe the speech of all six languages.

Of course, none of the languages required the entire inventory. German (48 symbols) and Italian (47) had the largest inventories, with the former introducing numerous umlauts and the latter differentiating between normal and elongated stops (sustained closure before the burst). Russian (41 symbols) introduced the many palatalized variants of phones. By including French (39 symbols), the entire inventory was enriched by multiple nasalized vowels. English (37) and Spanish (33) were found to be the languages with the smallest inventory.

Group	Phonetic Symbols
Vowels	a ã ɐ ɑ ɒ æ Œ̃ ʌ e ẽ ɛ ɜ ə i ɨ ɪ o õ ø œ ɔ ʊ u y ʏ
Stops	b bʲ d dʲ g gʲ p pʲ t tʲ k kʲ ʔ
Fricatives	β ç ð f fʲ h ʝ s sʲ ʃ ʃʲ v vʲ x xʲ ɣ z zʲ ʒ θ
Nasals	m ɱ mʲ n ŋ ɲ nʲ
Approximants	ɥ j l lʲ w ʎ
Trills	r rʲ ʀ

Table 4: Summary of all phonetic IPA symbols used throughout the corpus. Not included in the table but part of the annotation are a symbol for silence as well as 26 elongated variants of presented phones.

2.4. Phonetic Annotation

To generate phonetic annotation, we used *WebMAUS* (Kisler et al., 2017), a web-service provided by the *Bavarian Archive for Speech Signals* (BAS). The Munich AUtomatic Segmentation (MAUS) toolkit provides a routine to reliably predict pronunciation from a pair of speech recording and text transcript (Schiel, 1999; Schiel, 2015). The preset pipeline *G2P_MAUS* without ASR was used with the respective language of the sample, requesting the output phonetic symbols to be encoded in IPA. The pipeline without ASR only disabled the initial ASR for transcription (which was already available), not the one for prediction of pronunciation. After running a grapheme-to-phoneme (G2P) model, MAUS estimated the true pronunciation from the ideal (G2P) and recognized (ASR) pronunciations. The default weight factors for deciding between the two options were left unchanged for all languages.

MAUS returned its segmentation result in the form of *Praat* TextGrids. Each contained word- and phonetic-level segmentation of the audio-signal. When choosing IPA as output symbol, phonetic transcription on word-level was given in IPA, but on phonetic level, MAUS yielded X-SAMPA (Wells, 1995) symbols. As this was just another coding format, translation to IPA was trivial, thus all phonetic symbols in the provided TextGrids follow IPA standards.

2.5. Acoustic Model Training

The training split from CP was used to fine-tune a Wav2Vec 2.0 (Baevski et al., 2020) base model. The model had been pre-trained on the 960 hours Librispeech (Panayotov et al., 2015) corpus of read English speech. A final linear layer was added for classification with 102 output nodes, one for each phonetic symbol and an additional one for a blank token to perform connectionist temporal classification (CTC) (Graves et al., 2006). Adam optimizer (Kingma and Ba, 2014) was used with an initial learning rate (LR) of $3 \cdot 10^{-6}$. During warm-up, the LR increased linearly to $3 \cdot 10^{-5}$ over the first ten epochs, remained constant for another 30 epochs, and would then decay exponentially by a factor of 0.96 for the remaining 120 epochs. During a single epoch, the model was shown a subset of 5,000 randomly selected samples from the training set.

Afterwards, the most probable sequence of phonetic symbols could be estimated through a beam search (beam width = 10) and CTC decoding. The beam search was not extended with a language model due to the multilingual setup. Despite the fact that our model did not predict phonemes, but phonetic symbols, PER was still considered a decent metric to evaluate the performance of the entire system, as it simply measures the amount of insertions, deletions or replacements required to transform the predicted sequence into the true sequence, relative to the true sequence' length.

Language	Train	Dev	Test
English	14.1	2.3	2.3
German	13.6	2.3	2.2
Spanish	14.5	2.5	2.6
French	14.6	2.5	2.5
Italian	16.5	3.0	3.1
Russian	12.7	2.6	2.8
Total	**85.8**	**15.2**	**15.5**

Table 5: Recorded hours of speech in the splits for all six languages.

3. Results

The model predicted phonetic symbols with average PERs of 17.8 % on the development and 18.1 % on the test set of CP. Results for the different languages are shown in Table 6. English and Spanish were found to be the easiest languages to predict phonetic symbols in our setup. On both development and test, the weakest results were observed for Russian.

Differences between predicted and true sequence had also been analyzed for frequent substitution patterns. The model sometimes struggled to correctly differentiate between short and elongated stop sounds that were common in Italian. For stop phones in general, confusions between voiced and unvoiced productions were also observed more frequently. For Russian, a frequent confusion was those of palatalized phones with their

Language	Dev	Test
English	15.5	15.6
German	19.4	19.4
Spanish	14.5	15.0
French	18.8	18.4
Italian	17.8	17.4
Russian	20.0	21.4

Table 6: PERs on development and test for the different languages.

non-palatalized counterparts. In some cases, the model would even append a palatal approximant [/j/ or /ʎ/] after the non-palatalized phone.

4. Discussion

There were likely two main reasons for English and Spanish to yield the lowest PERs. Firstly, the entire Wav2Vec model was pre-trained on almost a thousand hours of English speech. This could potentially induce some bias towards English pronunciation. For Spanish, the fact that the language used comparatively few stop phones (12.8 % compared to the other languages [18.3 % – 21.4 %] as estimated from the ground truth) Finally, Spanish and English comprised the smallest phonetic inventories with 33 and 37 symbols, respectively.

The weaker result for Russian did not come as a surprise, because the language introduced a large number of new phones to the inventory that were not used among the other datasets.

A PER of 18.1 % may appear a lot higher than the one reported for Wav2Vec 2.0 mentioned earlier (8.3 %). This was, however, calculated on TIMIT, which includes recordings from only one microphone in a single language, collected in a clean acoustic environment. Furthermore, when computing PERs on TIMIT, it is common to collapse the annotated phonemes to 39 classes. This was done for the Wav2Vec 2.0 evaluation as well. The presented model trained with CP had to distinguish between 101 classes from six different languages. Lastly, the single-digit PER was achieved only by the large version of Wav2Vec, which comprised 317 million parameters. For this study, the base model with 95 million parameters was used, as we were not interested in showing peak PER values, but intended to validate the suitability of phonetic labels created with *WebMAUS*.

Finally, some light should be shed on the potential weaknesses of CP. During the process of speaker selection, the session ID was used as the unique identifier for a speaker. While we managed to ensure that this ID was constant throughout multiple sessions by only including users who had been logged in to their account, we could not rule out the possibility of multiple speakers contributing through the same account. However, no such case was identified during manual investigation.

It could also happen that the predicted pronunciation from *WebMaus* was not entirely correct, or that a file contained longer segments of silence, background noise, or poorly intelligible speech which confused the ASR during the alignment. Whilst these artifacts could result in a certain amount of label noise (which can be found in almost every ML dataset), they would also allow a model to explore a much larger space of acoustic conditions, speaker traits and noise patterns, ultimately resulting in a more robust decision function.

5. Conclusion

CP provides a reduced, gender-balanced version of CV comprising six different languages. All samples received phonetic annotation, which could be proven to yield reliable results when used for training a state-of-the-art acoustic model. Totalling over 100 h of recorded speech collected from more than 11,000 contributors in unsupervised recording conditions, CP can help to bridge the gap between research and application of acoustic models.

The corpus is distributed free of copyright under a *Creative Commons (CC0 1.0 Universal)* license, just like CV itself. It is distributed via `www.zenodo.org` (doi: 10.5281/zenodo.5846137). In the future, the phonetic inventory should grow further by adding more languages. CP should also receive updates along with major revisions of CV.

6. Bibliographical References

Association, I. P., Staff, I. P. A., et al. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Boersma, P. and Van Heuven, V. (2001). Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Brandenburg, K. (1999). Mp3 and aac explained. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Hu, S., Shang, X., Qin, Z., Li, M., Wang, Q., and Wang, C. (2019). Adversarial examples for automatic speech recognition: Attacks and countermeasures. *IEEE Communications Magazine*, 57(10):120–126.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, September.

Kukačka, J., Golkov, V., and Cremers, D. (2017). Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*.

Moore, R. K. and Skidmore, L. (2019). On the use/misuse of the term 'phoneme'. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September(October):2340–2344.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. of the ICPhS*, pages 607–610, San Francisco, August.

Schiel, F. (2015). A statistical model for predicting pronunciation. In *ICPhS*.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Schönherr, L., Kohls, K., Zeiler, S., Holz, T., and Kolossa, D. (2018). Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*.

Wang, T. E., Gu, Y., Mehta, D., Zhao, X., and Bernal, E. A. (2018). Towards robust deep neural networks. *arXiv preprint arXiv:1810.11726*.

Wells, J. C. (1995). Computer-coding the ipa: a proposed extension of sampa.

7. Language Resource References

Ardila, Rosana and Branson, Megan and Davis, Kelly and Henretty, Michael and Kohler, Michael and Meyer, Josh and Morais, Reuben and Saunders, Lindsay and Tyers, Francis M and Weber, Gregor. (2020). *Common voice: A massively-multilingual speech corpus*.

Garofolo, John S and Lamel, Lori F and Fisher, William M and Fiscus, Jonathan G and Pallett, David S. (1993). *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*. ISLRN 664-033-662-630-6.