ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413720

MULTILINGUAL PHONETIC DATASET FOR LOW RESOURCE SPEECH RECOGNITION

Xinjian Li; David R. Mortensen; Florian Metze; Alan W Black

Carnegie Mellon University;
xinjianl@cs.cmu.edu

ABSTRACT

Phone Recognition is one of the most important tasks in the field of multilingual speech recognition, especially for low-resource languages whose orthographies are not available. However, most speech recognition datasets so far only focus on high-resource languages, there are very few datasets available for low-resource languages, especially datasets with detailed phone annotation. In this work, we present a large multilingual phonetic dataset, which is preprocessed and aligned from the UCLA phonetic dataset. The dataset contains around 100 low-resource languages and 7000 utterances in total. This dataset would provide an ideal training/evaluation set for universal phone recognition.

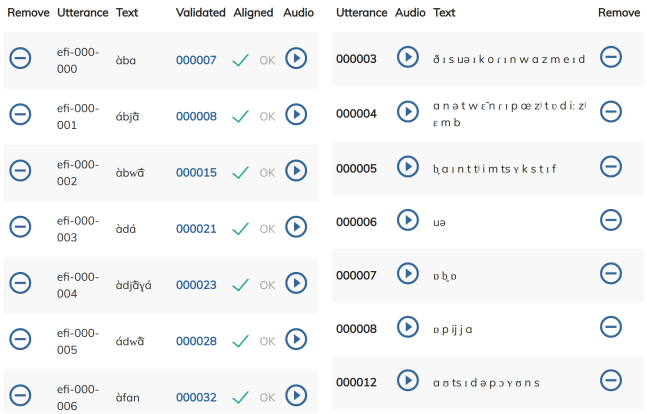
Index Terms— Multilingual Phonetic Dataset, Multilingual Speech Alignment, Low-Resource Speech recognition

1. INTRODUCTION

Recently, speech recognition communities have made significant progress towards building deep neural networks for speech recognition by taking advantage of huge volumes of training data and high-quality test sets [1, 2]. While high-resource languages such as English and Mandarin have been able to benefit from the newly developed technology [3, 4], most of the languages in the world are low-resource languages lacking large sets of training data or even small test sets. More importantly, many languages do not have standardized orthographies; speech datasets fully annotated with phonetic transcriptions are the only means of building speech technologies for them. Unfortunately, phonetically-annotated data sets are also largely limited to high-resource languages [5]. Additionally, the annotated data is usually monolingual corpus with a limited phone inventory. Ideally, a well-annotated dataset should contain a large number of languages and have a rich phone inventory. This would be useful not only to train the recognition system for the target language but also benefit to build/evaluate any language-independent universal phone recognizers [6].

In this paper, we introduce a large multilingual phonetic dataset, which is derived from the online UCLA phonetics archive [7][1]. The online phonetics archive contains a large amount of speech data collected by field linguists. For each language, there are typically a variety of materials available including the audio recordings in WAV format, transcribed word lists, information about the native speakers, etc. The total number of languages is around 300, and most of them are low-resource languages (most of which have less than 1 million native speakers). However, for each subset of the data, the archive only includes a large audio file and a table of transcriptions, along with some other information. No alignments are provided, which poses a challenging problem to create the aligned phonetic dataset.

In this work, we tackle this problem by using two-step alignment: in the first step, we segment the transcriptions and audio files into small utterances. All audio files are transcribed into phones with a recently proposed recognizer [8]. Every transcribed utterance is aligned automatically with the recognized utterances by measuring the phone feature distance. Next, all aligned utterances are manually validated and corrected by human experts. Additionally, during the second step, we implement several simple but effective strategies to speed up alignment correction. The prototype dataset contains around 100 languages and 7000 utterances, it will be



Remove	Utterance	Text	Validated	Aligned	Audio
⊖	efi-000-000	àba	000007	✓ OK	▷
⊖	efi-000-001	ábjã̀	000008	✓ OK	▷
⊖	efi-000-002	àbwã̀	000015	✓ OK	▷
⊖	efi-000-003	àdá	000021	✓ OK	▷
⊖	efi-000-004	àdjðɣá	000023	✓ OK	▷
⊖	efi-000-005	ádwã̀	000028	✓ OK	▷
⊖	efi-000-006	àfan	000032	✓ OK	▷

Utterance	Audio	Text	Remove
000003	▷	ð ɪ s u ə ɪ k o ɾ ɪ n w a z m e ɪ d	⊖
000004	▷	ɑ n ə t w ɛ̃ n ɾ ɪ p œ z ʲ t ɒ d iː z ʲ ɛ m b	⊖
000005	▷	b̥ ɑ ɪ n t tʲ i m t s ʏ k s t ɪ f	⊖
000006	▷	u ə	⊖
000007	▷	ɒ b̥ ɒ	⊖
000008	▷	ɒ p i j j ɑ	⊖
000012	▷	ɑ ɑ t s ɪ d ə p ɔ ʏ ɑ n s	⊖

Fig. 1. A alignment sample from the dataset where the left Table shows the annotated phones/utterances extracted from the website, the table on the right side is the segmented audio chunks and the recognized phones. Two tables are first aligned automatically with phonetic features distances and then fixed manually.

[1] http://archive.phonetics.ucla.edu/

distributed to benefit future work[2]. Note that the number of utterances and languages might change in the final version.

2. RELATED WORK

Previously, many multilingual datasets have been created from different sources such as audiobooks, broadcast news, and online recordings. These include the Babel database [9], TUNDRA corpus [10], Voxforge collections [11] and common voice dataset [12]. While those datasets sometimes cover more than 10 languages, the target languages are typically high-resource languages, whereas low-resource languages are rarely included. More recently, a dataset has been prepared for a much broader group of languages [13]. However, the dataset is automatically aligned and alignment quality differs among languages. Additionally, the transcription is in the orthographic form where the phonetic transcriptions are not available. In this paper, we prepare a low-resource language dataset with fully annotated and validated phonetic transcriptions.

To develop a good alignment tool is essential for this work, as a fully manually alignment would be a poor use of valuable expert time. The automatic alignment problem, arising when curating speech corpora or synchronizing audiobooks, has been addressed in prior works [14, 15, 16, 13]. There are typically two approaches to finding alignments between audio and transcriptions. The first is to utilize a speech recognizer to transcribe audio into text or phones, and then estimate the alignment between the outputs with the provided transcriptions [14, 15]. The second, on the other hand, obtains the audio signals by synthesizing transcriptions, then aligns the original audio with the generated audio [16, 13]. While both groups have achieved some success in obtaining usable alignments, they typically require some prior knowledge of the target language, and the aligned pairs are usually not systematically validated. In this work, we establish the alignment for around 100 languages while assuming little prior information. Additionally, human feedback is used efficiently to validate and correct alignments.

3. APPROACH

In this section, we introduce the methods used to develop the dataset. We obtain the raw dataset by crawling the archive pages. We then automatically align the recognized phones and annotated phones for the utterances. Finally, experts employ an online tool to manually but efficiently validate and correct the alignments.

3.1. Preprocessing

The crawler first downloads the top page and extracts all available languages. It then recursively parses the individual links to the pages for each language and extracts all annotated word or utterance lists, together with the corresponding audio files. The utterance lists are typically contained in tables whose headers document the content type of each column. As the headers do not always follow identical naming conventions, several regular expressions are used to determine which column contains the phone annotations. From each utterance list, we typically extract 10 to 100 annotated words/utterances.

The corresponding WAV file is usually a long audio recording containing the entire contents of the utterance list. Besides, it usually contains many unrelated contents such as the introduction of the native speaker, instructions regarding what to read next, and some incidental conversation. Since most of our annotated utterances typically contain a single word in each utterance, voice activity detection is applied to segment the audio into small chunks. For each annotated utterance, one particular chunk is expected to contain its speech. It is important to note that the acoustic environment varies significantly across different languages' recordings; some are clean enough for the voice activity detection to work efficiently, but others contain so much noise and overlapping speech that voice activity detection cannot consistently distinguish silence and speaking intervals.

3.2. First Pass Alignment

Next, all audio chunks are fed into a recently proposed multilingual phone recognizer [8], by which each chunk is transformed into an appropriate sequence of phones. The first-pass alignment is done by matching the golden annotated phone labels and the recognized phone labels. Typically, the phone-level alignment is done using standard string edit distance and greedy search. i.e, for each annotated utterance, we compute the edit distance with all recognized phone sequences and select the utterance with the lowest cost. However, this baseline alignment fails to produce a good first-pass alignment in this case, due to two challenges: First, the gold phone transcriptions are partial transcriptions. Many speaking parts in the recordings are not transcribed as they are not related to annotation (e.g: instructions to native speakers of what to read next). Second, the recognizer has not seen most of the languages (and, understandably, performs worse on languages it has not seen). However, by taking advantage of several properties in the dataset and the recognizer, we arrive at alignments that are much better than those produced by this baseline. Three approaches are introduced in this section.

3.2.1. Monotonic Alignment

First, the annotated utterances are not listed in random order. The relationship between the annotated word list and the associated recording is typically monotonic. While other material may intervene in the recording, the utterances of interest are in the same order in the recordings as in the annotations. We note there are several cases in which this order

[2]https://github.com/xinjli/ucla-phonetic-corpus

fails to be monotonic, for example, the native speaker occasionally forgets reading some utterances and returns to those utterances later. However, by imposing this constraint, the available matching pairs are greatly reduced. Coupled with dynamic programming, this makes alignment much more efficient.

3.2.2. Phonological Distance

Next, we use phonological features to measure the distance between annotated phones and recognized phones (instead of using the exact phone match). The phonological distance enables us to quantify similarity more precisely. In particular, we use the PanPhon tool to compute the phonological distance between two utterances where 22 phonological features are taken into account [17]. For example, [syllabic], [sonorant], [consonantal], etc. Instead of penalizing phone mismatch with 1 cost, it imposes a penalty based on partial feature mismatch.

3.2.3. Consecutive Segment Merger

Another improvement in the alignment can be made by merging consecutive vowels or consonants in the recognized phones. During the experiment, we found that the recognizer tends to generate more than one vowel or consonant for a single phone when that specific phone context is rare in the training set. This issue tends to increase the distance even when the recognized phones are close to the annotation. Table 1 shows such an example in which merging multiple vowels and consonants could lead to a more accurate distance. We note that it is not always correct to merge vowels and consonants since sequences of multiple vowels or multiple consonants do occur in many languages; however, we find this approach helps to reduce many misalignments in practice.

Annotated Phones	Recognized Phones	Distance
[tʰɑɪbɒ]	[m a z]	1.45
[tʰɑɪbɒ]	[t ɕe i: b̪ uə ə]	3.04
[tʰɑɪbɒ]	[t e b̪ uə]	1.18

Table 1. An actual example from the experiment to merge consecutive vowels and consonants into one phone. The annotated phones [tʰɑɪb] should be aligned with the [t ɕe i: b̪ uə ə], but was originally misaligned with [m a z] as it has less distance, after merging vowels and consonants in the 3rd row, it has less distance and could be aligned correctly.

3.3. Second Pass Alignment: Real-Time Feedback

During the second phase, we use our online tool to update the first pass alignment in real time based on feedback (validation or correction) from annotators. In particular, we exploit two types of feedback to improve the alignment. Both types are fast enough to update alignments in real time.

Approach	Acc. Mean	Acc. Std
First Pass (baseline)	4.88%	6.37%
First Pass (+ monotonic)	27.3%	21.6%
First Pass (+ distance)	6.62%	12.5%
First Pass (+ merge)	5.64%	8.32%
First Pass (+ all)	38.0%	26.4%
Second Pass	**56.0%**	24.3%

Table 2. Alignment accuracy of different approaches. The first pass on the first row is the baseline alignment, in which there are no constraints in the alignment. Additionally, we add three different First Pass approaches and measure the performance separately and jointly. The Second-Pass shows the improved alignment accuracy by using real-time feedback.

First, we use the anchor point to improve the alignments. When a new validation is confirmed or a new alignment is fixed, the aligned utterance index and audio index are sent to the server, notifying it of the new anchor point. The remaining unverified alignments are updated, subject to this new anchor point. In the first pass alignment, the alignment errors tend to propagate through the last utterance whenever there is a large mismatch. Fixing the anchor point could bring the alignment back to the correct starting point.

Next, we use the index interval information to improve the alignment. During the experiment, we noticed that the aligned audio index has a typical index interval in each dataset. For example, the aligned audio index might be 10, 12, 14, etc: the first utterance is aligned to the 10-th audio, the second utterance is aligned to the 12-th audio. This is because the native speaker and the linguist are talking in turns: one reads the utterance, then the other instructs what to read next. Each dataset has a different pattern, but the index interval is usually consistent in each dataset. During the second pass, we use the validated utterances to estimate the typical index interval by taking the mean of validated/fixed utterances intervals. The interval is then taken into account as a new distance factor when updating the alignments. By combining those two types of real-time feedback, the validation and fixing process requires much less manual works.

4. EXPERIMENTS

In this section, we evaluate our alignment approach and provide statistics for the collected dataset. In the first version of our dataset, we provide alignments for 106 languages. For each language's dataset, the alignment is first automatically aligned and then validated/fixed by an expert.

4.1. Alignment Evaluation Results

We first evaluate the alignment performance across all 106 languages. The metric is the *alignment accuracy*: whether each annotated utterance is correctly aligned with the target audio chunk or not. As we handle a large number of languages in the experiment, instead of showing the alignment accuracy for each language, we show the mean and standard deviation of accuracy across all languages. The results are shown in Table.2, in which we compare several approaches we mentioned in the last section. First, we consider the naive first pass alignment in which we greedily match each utterance with all audio candidates. The results are around 5% accuracy, which is hardly useful as the first pass alignment. Next, we try the three approaches mentioned above: imposing the monotonic order, using phonetic distance instead of the naive edit distance, merging consecutive vowels and consonants. The monotonic constraint improves the alignment significantly by about 20% accuracy. The other two only increase the metric marginally when used separately, however, when all three approaches combined, it improves the accuracy by more than 30%.

During the first pass alignment, we notice there is a huge accuracy variance across different languages as shown by the standard deviation: some datasets are aligned very successfully with almost 100% accuracy. On the other hand, some corpora fail with near 0% accuracy. The variance can be explained by several factors: first, the audio quality varies significantly across different languages: some recordings were made in a clean environment, while others were done in relatively noisy rooms. The audio quality affects the recognition accuracy and therefore makes a huge difference during the first pass. Second, the recordings are segmented by voice activity detection. Some speakers wait for 1–2 seconds between every utterance while others continue to speak several utterances without any interruption. As there is no silence between the utterances, the single audio chunk contains several utterances and could not get aligned with any of the target annotations. Finally, imposing the monotonic order might propagate the alignment error to the last utterance. While the first pass alignment could align 40% correctly, it still requires a huge amount of effort to fix the remaining 60% utterances. In the second pass, we apply the real-time feedback to the system and automatically fix many alignment errors with the new anchor point and interval information. The table suggests that alignment accuracy is further improved to around 60% in the second pass. Finally, the remaining 40% misaligned utterances are fixed manually.

4.2. Dataset Statistics

The first version of our dataset contains 106 languages with 6,880 validated utterances. Each language contains around 60 utterances on average with 28.4 std. We find that some languages have many more utterances and speakers than others.

Language Area	Language %	Utterance %
Africa	48.5%	23.1%
America	6.15%	8.63%
Asia	26.2%	43.6%
Europe	15.3%	22.5%
Pacific	3.85%	2.17%

Table 3. Area distribution of languages and utterances

syllabic	**sonorant**	**continuant**	**delayed release**
44.3%	67.8%	68.0%	0.53%
lateral	**nasal**	**strident**	**spread glottis**
4.02%	10.8%	1.74%	1.71%
cons glottis	**anterior**	**coronal**	**distributed**
2.04%	38.2%	30.4%	4.67%
labial	**high**	**low**	**back**
16.9%	25.0%	17.4%	25.4%
round	**click**	**tense**	**long**
13.1%	0.28%	37.4%	2.61%

Table 4. Phone distribution of features

The data related to areal distribution is shown in Table.3. We show the language distribution and utterance distribution across different areas. The table suggests that nearly half of the languages in the dataset are from Africa, while only 4% of languages are from the Pacific area. The utterance distribution is relatively proportional to the language distribution. However, Asian languages dominate in the utterance count with 43.6%. African languages have fewer utterances in proportion to the number of languages. We also investigated the distributions of phones in the entire dataset. In total, we find the number of unique phones (phone types) is more than 400. 51.7% of the phones are consonants and 48.3% are vowels. The detailed feature distribution is shown in Table.4, which suggests that the phone inventory is rich in various categories. This dataset should be useful in many ways. First, it can be used to evaluate phone recognition systems for the included low-resource languages. Additionally, it might serve as a good training/evaluation set for any universal phone recognizers due to its rich inventory and large coverage of languages.

5. CONCLUSION

In this work, we introduce a new multilingual phonetic dataset for low resource languages. The dataset is prepared from an online archive by two steps alignment. The dataset contains around 100 languages and 7000 utterances, and would be released to the community to benefit speech research in low resource phone recognition.

6. REFERENCES

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[2] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[3] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 517–520.

[4] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: a resource for the next generations of speech-to-text.," .

[5] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *STIN*, vol. 93, pp. 27403, 1993.

[6] Xinjian Li, Siddharth Dalmia, David Mortensen, Juncheng Li, Alan Black, and Florian Metze, "Towards zero-shot learning for automatic phonemic transcription," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 8261–8268.

[7] Peter Ladefoged, B Barbara, and GS Russell, "Ucla phonetics lab archive," 2009.

[8] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al., "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[9] Mary Harper, "The iarpa babel multilingual speech database," 2011.

[10] Adriana Stan, Oliver Watts, Yoshitaka Mamiya, Mircea Giurgiu, Robert AJ Clark, Junichi Yamagishi, and Simon King, "Tundra: a multilingual corpus of found data for tts research created with light supervision.," in *INTERSPEECH*, 2013, pp. 2331–2335.

[11] Voxforge.org, "Free speech recognition (linux, windows and mac) - voxforge.org," http://www.voxforge.org/, accessed 06/25/2014.

[12] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[13] Alan W Black, "Cmu wilderness multilingual speech dataset," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5971–5975.

[14] Xavier Anguera, Jordi Luque, and Ciro Gracia, "Audio-to-text alignment for speech recognition with very limited resources," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[15] Germán Bordel, Mikel Penagarikano, Luis Javier Rodriguez-Fuentes, and Amparo Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[16] Fabrice Malfrère, Olivier Deroo, Thierry Dutoit, and Christophe Ris, "Phonetic alignment: speech synthesis-based vs. viterbi-based," *Speech Communication*, vol. 40, no. 4, pp. 503–515, 2003.

[17] David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin, "Panphon: A resource for mapping ipa segments to articulatory feature vectors," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 3475–3484.