Interspeech 2018
2-6 September 2018, Hyderabad

GlobalTIMIT: Acoustic-Phonetic Datasets for the World's Languages

Nattanun Chanchaochai[1], Christopher Cieri[1], Japhet Debrah[1], Hongwei Ding[2], Yue Jiang[3], Sishi Liao[2], Mark Liberman[1], Jonathan Wright[1], Jiahong Yuan[1], Juhong Zhan[3], Yuqing Zhan[2]

[1]Linguistic Data Consortium, University of Pennsylvania, U.S.A.
[2]School of Foreign Languages, Shanghai Jiao Tong University, P.R.C.
[3]School of Foreign Studies, Xi'an Jiaotong University, P.R.C.

nattanun@sas.upenn.edu, ccieri@ldc.upenn.edu, debrahj@seas.upenn.edu, hwding@sjtu.edu.cn, jiang8@mail.xjtu.edu.cn, liao_ssh@sjtu.edu.cn, myl@cis.upenn.edu, jdwright@ldc.upenn.edu, jiahong@ldc.upenn.edu, juhongzhan@foxmail.com, lancaster_zhan@sjtu.edu.cn

Abstract

Although the TIMIT acoustic-phonetic dataset ([1], [2]) was created three decades ago, it remains in wide use, with more than 20000 Google Scholar references, and more than 1000 since 2017. Despite TIMIT's antiquity and relatively small size, inspection of these references shows that it is still used in many research areas: speech recognition, speaker recognition, speech synthesis, speech coding, speech enhancement, voice activity detection, speech perception, overlap detection and source separation, diagnosis of speech and language disorders, and linguistic phonetics, among others.

Nevertheless, comparable datasets are not available even for other widely-studied languages, much less for under-documented languages and varieties. Therefore, we have developed a method for creating TIMIT-like datasets in new languages with modest effort and cost, and we have applied this method in standard Thai, standard Mandarin Chinese, English from Chinese L2 learners, the Guanzhong dialect of Mandarin Chinese, and the Ga language of West Africa. Other collections are planned or underway.

The resulting datasets will be published through the LDC, along with instructions and open-source tools for replicating this method in other languages, covering the steps of sentence selection and assignment to speakers, speaker recruiting and recording, proof-listening, and forced alignment.

Index Terms: speech datasets, acoustic phonetics

1. Introduction

Recently, a researcher who specializes in advanced neurophysiologic brain mapping methods, including awake speech and motor mapping, wrote to one of this paper's authors to ask if there is "something equivalent to TIMIT in Mandarin," hoping for "something well annotated with tone as well as phonemes." His response to getting a pre-publication copy of Chinese TIMIT [3] was "This sounds perfect! And Global TIMIT is such a great idea."

By "equivalent to TIMIT" he meant a dataset with:

- Multiple (anonymously) identified speakers
- Wide range of phonetically-representative inputs
- Wideband recordings with good acoustic quality
- Time-aligned lexical and phonemic transcripts
- Easily availability to anyone

Although Mandarin Chinese is one of the largest and best-documented languages in the world, he was not able to find any available resources meeting his needs. Many experiences of this kind over the past few years have motivated us to find a simple and inexpensive approach to designing, implementing, and distributing TIMIT-like resources that could plausibly be generalized across all of the world's languages.

This paper describes our exploration of this method in half a dozen test cases.

2. The original TIMIT design

The name "TIMIT" is a blend of *Texas Instruments* ("TI"), where the dataset was recorded, and Massachusetts Institute of Technology ("MIT"), where transcription, alignment, and other processing were done.

The original TIMIT dataset contains 6300 recorded utterances; 10 spoken by each of 630 speakers, representing about 30.8 seconds of speech per speaker on average, and a total of 5:23:59.7 of audio in all. The texts of these utterances comprise 2342 sentences, including 2 "dialect shibboleth" sentences designed at SRI, which were read by all speakers; 450 "phonetically-compact" sentences designed at MIT, which were read by 7 speakers each; and 1890 "phonetically-diverse" sentences selected at TI, read by 1 speaker each. These 2342 sentences contain 6099 distinct words.

Each speaker reads 2 "dialect" sentences (denominated SA1 and SA2), 5 "compact" sentences (SX1 to SX450), and 3 "diverse" sentences (SI1 to SI1890).

The speakers were primarily TI employees, often enrolled as part of the initial process of joining the company's offices in Dallas. They were recorded in a sound booth at TI, using a Sennheiser headset-mounted microphone, with 53 dB SPL of background noise played through headphones "to eliminate unusual voice quality produced by the dead room effect". The recordings were made in 1987.

Phone-level transcription and alignment were done at MIT, where a post-hoc division into "training" and "test" sets was also performed. These steps were carried out between 1987 and 1990, when the standard TIMIT CD-ROM came out. A provisional version of the dataset was released in 1988.

The cost of the original TIMIT dataset creation, during the period 1987-1990, was about $1.5 million (personal communication from a former DARPA program manager), which corresponds to about $3.3 million in 2018 money. This

10.21437/Interspeech.2018-1185

is because it involved substantial portions of the work of several senior researchers over a period of several years, as well as even more work by lower-level staff and student research assistants.

3. Our Approach

We want to design a method for creating an "acoustic-phonetic continuous speech corpus" in an arbitrary language, with as many as possible of the properties that have made TIMIT so widely useful, while limiting the effort and cost involved to a few weeks of expert labor, or what might be accomplished as a student term project or a summer internship.

We retain key features of the original TIMIT dataset:

- a large number of fluently-read sentences, containing a representative sample of phonetic, lexical, syntactic, semantic, and pragmatic patterns;
- a relatively large number of speakers;
- time-aligned lexical and phonetic transcription of all utterances;
- Some sentences read by all speakers, others read by a few speakers, and others read by just one speaker.

But in order to keep the required effort and cost to a reasonable level, we modify some other features, which also seem less essential to us.

3.1. Speakers and sessions

In today's landscape, the overhead involved in recruiting and recording 630 speakers seems both problematic and unnecessary. There are many useful speaker-recognition datasets with even larger numbers of speakers, recorded in more realistic settings (see e.g. [4]).

In choosing a target speaker count, we reason roughly as follows. A plausible session duration for an individual speaker is 20 minutes of actual reading. Allowing an average of 10 seconds of elapsed time per 3-second sentence, we get 120 sentences per speaker. If we want about 6000 utterances in total, this gives us something like 6000/120 = 50 speakers.

Rounding each speaker's recording session up to half an hour, the total recording time required becomes something like 50/2 = 25 hours – which can plausibly be accomplished over the course of a week or two, if the speakers are accessible and scheduling can be arranged.

3.2. Assigning sentences to speakers and speaker groups

TIMIT's idea of "dialect shibboleth" sentences has not turned out to be useful, at least as originally implemented. Instead, most users of the dataset have treated each speaker's productions of the SA sentences as "calibration" utterances. We replicate and extend this idea by selecting 20 calibration sentences (for each dataset) that all subjects read.

And the idea of a larger number of sentences read by more than one speaker also seems worthwhile. Thus, we divide the set of 50 subjects into 5 groups of 10 speakers each, and we create 5 sets of 40 sentences, with all the members of each group reading all of the 40 sentences assigned to their group. This requires 5*40 = 200 distinct sentences, each of which will be read by 10 speakers.

Finally, we increase the sample size for each dataset by adding 60 unique sentences to the list to be read by each speaker, so that each speaker's 120 sentences are divided into 20 calibration sentences, 40 group sentences, and 60 unique sentences. The overall number of distinct sentences required for a given instance of this dataset design is 20 + 40*5 + 60*50 = 3220, compared to 2342 for original TIMIT.

This design also makes diverse train/test (or cross-validation) divisions easy, since we merely need to keep all the members of each of the five speaker groups together in order to divide the dataset up by speakers, and to use only the 50*60 = 3000 "unique" sentences in order to guarantee that sentences will not be duplicated across groups.

There are two ways to add speakers beyond the designed set of 50, while still retaining the structure and the advantages of the overall design: (1) We can add each additional speaker to one of the 5 groups, adding 60 new "unique" sentences for each added speaker; (2) We can create additional 10-speaker groups, adding 40 new "group" sentences for each added group, and 60 new "unique" sentences for each added speaker.

It is obviously also possible to modify the design in other ways, such as making each recording session longer or shorter by adding or subtracting from the set of sentences to be read.

3.3. Sentence selection

To create the set of 3220 sentences for each dataset, we rely on some variant of the following process:

1. Choose a large set of texts – a Wikipedia snapshot, newswire or newspaper text, etc.
2. Automatically divide the texts into "sentences" (perhaps with some errors).
3. Eliminate "sentences" that are too short or too long.
4. Optionally eliminate sentences by other automatic criteria, such as inappropriate characters, too-rare words, etc.
5. Make a random selection of ~10000 candidate sentences.
6. Manually screen the selected subset for suitability until 3220 are found.

Unsuitable candidates would be non-sentences, sentences that don't make sense out of context, sentences containing words that speakers are likely not to be able to pronounce or to understand, etc. In our experience, between a third and a half of the automatically-selected candidates are judged to be suitable. With a simple computer interface, suitability judgments can be accomplished at a rate of about 15 per minute, so that the selection process takes something like 8 to 10 hours of human labor.

Special sentence sets such as collections of proverbs may be incorporated as a whole, if desired.

There are several approaches to dividing the selected sentences into the calibration, group, and unique sets. The easiest method is simply to make random selections (without replacement) of the needed numbers. A second approach is to select from the candidate pool so as to optimize some desired criterion, for example selecting the calibration sentences so as to cover the maximum number of syllable types or phone n-grams, using a greedy algorithm [5].

3.4. Recording procedures

From the overall set of 3220 sentences, we create 50 ordered lists of 120 sentences, one for each of the 50 planned speakers. A given list will include the 20 calibration sentences, one of the 5 sets of 40 group sentences, and a random selection (without

replacement) of 60 sentences from the 60*50 = 3000 unique sentences.

Each such list of 120 sentences is presented to its assigned speaker in a randomly-permuted order. We recommend using the SpeechRecorder software [6] for presentation of prompts and recording of responses. For speakers who are not fluently literate in the language being recorded, audio prompts could be used, although we have not tried that approach yet.

Use of a sound booth or formal recording studio is possible but by no means necessary. We have gotten good results by recording in a quiet environment using an inexpensive head-mounted noise-cancelling microphone with integrated A-to-D conversion and USB connection, such as the Logitech H390.

Techniques for recruiting speakers differed across the collections that we have done so far, and will be sketched in the section of this paper describing the individual collections.

3.5. Transcriptions and alignment methods

There are three steps in automatic phonetic alignment for a project of this type:

1. Creation of candidate phone sequences for each sentence, using a combination of a pronouncing dictionary and grapheme-to-phoneme rules.
2. Training acoustic models on the whole corpus.
3. Using the results to accomplish forced alignment.

In languages like Chinese and Thai, where word boundaries are not marked in the orthography, an initial (automatic or manual) 'word' division also will be required.

In some languages, accomplishing the overall orthography-to-word-divided-phone-sequence mapping may be the hardest part of the project. In other cases, the orthographic system may be so phonologically transparent that the mapping is nearly or exactly the identity function.

In the worst case, once word divisions are accomplished, pronunciations for all the distinct words in the dataset might need to be added by hand. But even if the language's orthographic system is phonologically opaque, it may be possible to get pronunciation fields from a dictionary in digital form, and expand that mapping if needed using something like Phonetisaurus [7].

In our experiments so far, we have used an HTK-derived system for training acoustic models and accomplishing the final forced alignment, as described in [3] and [8]. The results have in general been excellent. Thus, in the case of the Chinese collection, 50 randomly selected sentences were manually segmented, and we found that in 93.2% of the well-defined phonetic boundaries, the forced-alignment time points were within 20ms of the manual segmentation, which compares well with state-of-the-art results as in [9].

Note that there should also be two stages of quality control, one at the start of this process and one at the end:

1. An initial pass of "proof-listening" to be sure that each recorded utterance is actually a performance of the associated orthographic form.
2. A final check that the phone sequence created for each utterance corresponds adequately to the way it was actually pronounced, and that the forced-alignment output is close enough.

If there are no problems in the basic collection, each of these steps should take only about two person-days of work per dataset. And given a method of publication and distribution that allows for version control, additional quality checking will be provided by end users of each dataset.

4. GlobalTIMIT experiments

We have completed five collections, with several more planned or in progress. Individual collections will be documented in separate papers – here we will simply sketch the process and results for each case.

4.1. Completed collections

For each of these five datasets, we have been through the process of selecting a sentence set, dividing the set into "calibration", "group", and "unique" subsets, creating the randomized sentence lists for each speaker, recruiting and recording the speakers, proof-listening the results, creating grapheme-to-phoneme mapping methods and acoustic models, implementing and applying an HTK-based forced alignment system, and checking the resulting alignments.

4.1.1. Standard Thai: "THAIMIT"

Designed and collected by Nattanun Chanchaochai in 2016, this was the first experiment in the GlobalTIMIT set, and used a slightly different design for the division of sentences among the 50 speakers. For THAIMIT, we projected the TIMIT proportions of 2-5-3 onto 120 sentences as 24-60-36, whereas for later collections we adopted the proportions of 20-40-60. In all collections, each speaker read 120 sentences, so that we recorded 6000 total utterances in each collection.

Thus, for THAIMIT there were 24 "calibration" sentences read by all speakers, 300 "group" sentences read by 10 speakers each (60 per speaker), and 1800 "unique" sentences read by just one speaker, for a total of 2124 distinct sentence types.

The sentences were selected from three text sources: the Thai National Corpus II [10] (75%), the Thai Junior Encyclopedia [11] (13%), and Thai Wikipedia (12%). In the case of the Thai National Corpus, selection was based on searches using the most frequent words in the corpus documentation, selecting examples from each of the six corpus genres: fiction, newspaper, non-academic, academic, law, and miscellaneous.

The Standard Thai dialect is natively spoken only by people in a region centered around the Bangkok Metropolitan Area. People in other regions of Thailand acquire the standard variety through education and media exposure. For this initial collection, we did not require subjects to be native speakers of Standard Thai, but only that they be literate people born and raised in Thailand.

All speakers were recruited in the Bangkok Metropolitan area, and were fluent in Standard Thai. Demographic details were collected, including gender, age, geographical history, height, education, etc., and will be published with the dataset.

In the case of Thai, developing a forced aligner required some extra steps, since Thai is written without spaces between words. To divide the text, the Smart Word Analysis for Thai (SWATH) tool [12] was used, with the divided text manually checked and corrected for accuracy. Creation of a pronouncing dictionary for this collection began with data from the Mary R. Haas Thai Dictionary Project [13]; about one thousand words in the selected material were not found in that dictionary, and

pronunciations for those words were added by hand, using the same system of transcription.

A forced aligner for Thai was then developed to fully annotate the corpus at the levels of phones, tones, and words, using the methods described in [8] and [9]. Details are provided in the corpus documentation and in a separate paper.

4.1.2. Standard Mandarin Chinese: "CHIMIT"

This dataset has been fully documented in [3]. It was designed by Jiahong Yuan, and collected at Shanghai Jiao Tong University by Hongwei Ding, Sishi Liao, and Yuqing Zhan.

The sentences for this collection were selected from the Chinese Gigaword Fifth Edition [14], a large archive of text data from Chinese news sources. The steps in the selection process were: 1. Extract sentences 10-20 characters long, excluding any containing characters not among the 3500 most frequently used; 2. Inspect the large resulting set of sentences in randomized order, removing any with uncommon words or inappropriate meanings, and dividing the character sequences into words, to produce a set of 5000 candidate sentences (containing about 6600 unique words and 2200 unique characters); 3. Use a computer program implementing greedy search to choose (a) 20 "calibration" sentences to cover the maximum number of (tone-independent) syllable types, and (b) 200 "group" sentences to cover the maximum number of tones and (within-word) tonal combinations; 4. Select 3000 "unique" sentences at random from the remainder of the list.

The speakers in this dataset were 50 students at Shanghai Jiao Tong University, 25 males and 25 females, who scored Class 2 Level 1 or better on the *Putonghua Shuiping Ceshi* proficiency test. The recordings were done in a sound-treated booth at Shanghai Jiao Tong University.

4.1.3. Chinese learners' L2 English

The sentences for this dataset were selected from the original English TIMIT, to make the two datasets more comparable. Two graduate students at Shanghai Jiao Tong University (Sishi Liao and Yuqing Zhan) examined the TIMIT sentences and selected approximately 1000 of those that are not difficult to understand and to read aloud for college students in China. Because the number of these "simple TIMIT sentences" is too few, we adopted a modified 20-40-60 design that requires only 820 sentences: 20 calibration sentences for all 50 speakers (20*1 = 20); 40 sentences for every 10-speaker group (40*5 = 200); and 60 sentences for every 5-speaker group (60*10=600). In this new setting, every speaker still reads 120 sentences, but every sentence is read by at least 5 speakers.

This dataset was collected at Shanghai Jiao Tong University with the same 50 speakers as in the Standard Mandarin Chinese TIMIT.

4.1.4. The Guanzhong variety of Mandarin Chinese

Besides Standard Mandarin, we have also made an effort to create TIMIT-like datasets for other dialects in China. The first attempt was for the Guanzhong dialect, which is a variety of Mandarin Chinese, spoken in the Guanzhong region in Shaanxi province, including the city of Xi'an. The sentences for this collection were the same as for Standard Mandarin.

This dataset was collected by Yue Jiang and Juhong Zhan (with their students) at a local high school in Chengcheng, Weinan. The speakers were 50 high school students, 25 males and 25 females, who speak the Guanzhong dialect as their native language.

4.1.5. Ga

Ga (ISO 639-3, gaa) belongs to the Kwa branch of the Niger-Congo language family, and is spoken by about 750,000 people in the Greater Accra region of Ghana. It is the native language of Japhet Debrah, a first-year undergraduate at the University of Pennsylvania at the time this dataset was collected.

Lacking other sources of Ga text, we used a bible translation as the main source of sentences, with a collection of Ga proverbs added in. The recordings were made in Accra during the summer of 2017, with most speakers recruited from the congregation of a church.

Ga orthography is phonologically transparent, and so our forced alignment system used the standard spelling as an adequate proxy for the phone sequences. But since Ga orthography does not mark tonal categories, tone marking remains to be done.

4.2. Collections in progress or planned

Other GlobalTIMIT collections are in various stages of development. A collection of American learners' L2 Mandarin has been fully designed, with recording about half done; a Swedish collection has been fully designed, with recording partly done. In both of those cases, the limiting factor is the relatively small number of suitable speakers in Philadelphia, where the collections are taking place. Datasets for Italian and French have been designed. We may collect several other sets for different geographical varieties of Mandarin.

5. Conclusions and future directions

Our experiments establish that it is possible to create a TIMIT-like dataset in a new language quickly and cheaply, for a definition of "TIMIT-like" explained earlier in this paper. We have collected five such datasets in diverse languages and language varieties, with several more in progress or planned.

Our next steps will be to publish the completed datasets through the Linguistic Data Consortium, and to create a set of tutorial instructions to make it easier for others to design and collect similar datasets for additional languages, varieties, or speaker groups.

There are a number of ways that collections of this type might be modified, in general or for special purposes. Thus, it might be useful to add to each recording session a short elicitation of spontaneous speech, such as a simple picture-description task. A dataset of this general type might be extended to exemplify variations in speaking rate, vocal effort, precision of articulation, etc. And speakers might be recruited to sample variations in age, gender identity, ethnic or geographical background, etc., although we acknowledge the limitations of reading sentence lists for these purposes.

6. Acknowledgements

The authors of this paper are listed in alphabetical order. The corresponding author is Mark Liberman (myl@cis.upenn.edu). These projects have so far been carried out without external funding: student support and modest subject payments were provided by the LDC and the University of Pennsylvania's China Research and Engagement Fund.

7. References

[1] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM," NISTIR 4930, 1993. [https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir4930.pdf]

[2] J. S. Garofolo et al. "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *LDC93S1*. Philadelphia: Linguistic Data Consortium, 1993.

[3] J. Yuan et al., "Chinese TIMIT: A TIMIT-like corpus of standard Chinese", *O-COCOSDA* 2017.

[4] NIST SRE (Speaker Recognition Evaluation) Plans, 1996-2016: [https://www.nist.gov/itl/iad/mig/speaker-recognition]

[5] U. Feige, "A Threshold of ln n for Approximating Set Cover*," J. of the ACM* 45(5), p. 634-652, 1998.

[6] C. Draxler and K. Jänsch, "SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software," *Proceedings of LREC 2004*, pp. 559-562.

[7] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework", *Natural Language Engineering*, 22(6):907:38, 2016.

[8] J. Yuan, N. Ryant, and M. Liberman, "Automatic phonetic segmentation in Mandarin Chinese: Boundary Models, Glottal features and tone," *Proceedings of ICASSP 2014*, pp. 2539-2543, 2014.

[9] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models", *Proceedings of Interspeech 2013*, pp. 2306-2310, 2013.

[10] "Thai National Corpus II," Department of Linguistics, Chulalongkorn University, 2013. Online: available from [http://www.arts.chula.ac.th/ling/TNCII/]

[11] "Thai Junior Encyclopedia," Thai Junior Encyclopedia by Royal Command of His Majesty the King, 1997. Online: available from [http://saranukromthai.or.th/sub/Ebook/Ebbok.php]

[12] P. Charoenpornsawat, "Feature-based Thai word segmentation," Master's thesis, Computer Engineering, Chulalongkorn University, 1999. Software by Theppitak Karoonboonyanan, Online: available from [https://linux.thai.net/projects/swath]

[13] M. R. Haas, "The Mary R. Haas Thai Dictionary Project", 1951. Online: available from [http://sealang.net/thai/dictionary.htm]

[14] R. Parker et al., *Chinese Gigaword Fifth Edition* (LDC2011T13), Linguistic Data Consortium, 2001.

[15] N. Minematsu et al., "English Speech Database Read by Japanese Learners for CALL System Development." *Proceedings of LREC 2002*.