GlobalTIMIT: Progress and Prospects



Mark Liberman

University of Pennsylvania

Outline:

- 1. What?
- 2. Why?
- 3. How?
- 4. Examples
- 5. Plans

WHAT is TIMIT?

TIMIT is an acoustic-phonetic dataset, collected and annotated between 1987 and 1990 at Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT), with help from SRI and NIST.

630 speakers read 10 sentences each – thus 6300 sentences in all – comprising 54,217 word tokens and 5:23:59.7 of audio.

There were 2 sentences that all speakers read, 450 sentences read by 7 speakers each, 1890 sentences read by just one speaker —

Overall, TIMIT contains 2342 distinct sentences and 6099 distinct words.

TIMIT is probably the single most widely-used speech database:

≡	Google Scholar	TIMIT
•	Articles	About 21,100 results (0.08 sec)
	Any time Since 2018 Since 2017 Since 2014 Custom range Sort by relevance	DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1 JS Garofolo, LF Lamel, WM Fisher NASA STI/Recon, 1993 - adsabs.harvard.edu Abstract The Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech ★ 99 Cited by 2378 Related articles All 4 versions ≫
	Sort by date Include patents Include citations Create alert	Speech database development at MIT: TIMIT and beyond V Zue, S Seneff, J Glass - Speech Communication, 1990 - Elsevier Automatic speech recognition by computers can provide the most natural and efficient method of communication between humans and computers. While in recent years high performance speech recognition systems are beginning to emerge from research ☆ ワワ Cited by 452 Related articles All 7 versions Web of Science: 178 ≫
		[োমনাওম] Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database JS Garofolo - National Institute of Standards and Technology (NIST), 1988 র্ন গুয় Cited by 401 Related articles

And TIMIT usage continues –

although by modern standards it's a small dataset:

=	Google Scholar	ТІМІТ
٠	Articles	About 1,400 results (0.03 sec)
	Any time Since 2018 Since 2017 Since 2014 Custom range	Speaker identification evaluation based on the speech biometric and i-vector model using the TIMIT and NTIMIT databases <u>MTS Al-Kaltakchi, WL Woo</u> , SS Dlay and Forensics (IWBF), 2017 - ieeexplore.ieee.org Physiological and behavioural human characteristics are exploited in biometrics and performance metrics are used to measure some characteristic of an individual. The measure might lead to a one-to-one match, which is called authentication or one-from-N, and a match
	Sort by relevance Sort by date ✓ include patents ✓ include citations ✓ Create alert	☆ 99 Related articles ≫ Variation of voice disorders among speakers in the database TIMIT and TIMIT
		I Daly, <u>Z Hajaiej</u> , A Gharsallah - Advanced Systems and …, 2017 - ieeexplore.ieee.org In this article we present the variation of the speech signal of the database TIMIT and NTIMIT with its morphological characteristics of the speakers as well as the extrinsic variability related to the conditions of transmission and acquisition of the signal. For this study we …
		Comparison of I-vector and GMM-UBM approaches to speaker identification with TIMIT and NIST 2008 databases in challenging environments MTS Al-Kaltakchi, WL Woo, SS Dlay EUSIPCO), 2017 25th, 2017 - iseexplore.isee.org In this paper, two models, the I-vector and the Gaussian Mixture Model-Universal Background Model (GMM-UBM), are compared for the speaker identification task. Four feature combinations of I-vectors with seven fusion techniques are considered: maximum ☆ 99 Cited by 1 Related articles All 2 versions
		[PDF] NTCD-TIMIT: A New Database and Baseline for Noise-robust Audio-visual Speech Recognition <u>AH Abdelaziz</u> - Proc. Interspeech 2017, 2017 - pdfs.semanticscholar.org Although audio-visual speech is well known to improve the robustness properties of automatic speech recognition (ASR) systems against noise, the realm of audio-visual ASR (AV-ASR) has not gathered the research momentum it deserves. This is mainly due to the ☆ 99 Cited by 1 Related articles All 3 versions 👀
		[PDF] Automatic Speech Recognition: Introduction S Renals, H Shimodaira, ASRASR Lecture - 2018 - pdfs.semanticscholar.org ASR Lecture 1 Automatic Speech Recognition: Introduction 13 Page 19. Example: TIMIT Corpus Page 20. Basic speech recognition on TIMIT Train a classifier of some sort to associate each feature vector with its corresponding label ☆ ワワ All 2 versions
		Learning Hard Alignments with Variational Inference

<u>D Lawson, CC Chiu, G Tucker, C Raffel</u>... - arXiv preprint arXiv ..., 2017 - arxiv.org ... Finally we show experimentally that our approach improves performance and substantially improves training time for speech recognition on the **TIMIT** dataset as well as a challenging noisy, multi-speaker version of **TIMIT** that we call Multi-**TIMIT**. -Equal contribution. x1 x2 x2 ...

☆ 99 Cited by 1 Related articles All 6 versions >>>

WHY has TIMIT been so popular?

lt's

- 1. easily available,
- 2.compact,
- 3. phonetically, lexically, and syntactically representative,
- 4. phonetically transcribed and aligned.

Then WHY aren't there TIMIT-like datasets for other languages?

- Creating TIMIT was a lot of work (and expense): At least 15 person-years of work at four institutions, budget estimated at \$1.5 million.
- 2. Today's focus is (properly) on conversational speech, meeting speech, broadcast news, etc.

BUT . . .

TIMIT's continued popularity suggests that TIMIT versions in other languages would be still be useful

- for phonetics research
- for teaching and student projects in speech technology
- for documentation of languages and varieties

AND...

With today's technology,

a redesigned TIMIT-equivalent dataset for a new language can be planned and created with two or three person-months of work

... most of which can be done by an interested student.

(With planned toolkit design, this may become even easier.)

HOW can this be done?

- 1. Streamlined dataset design
- 2. Modern software and/or data sources for
 - a. Selecting sentences
 - b. Recording speakers
 - c. Creating pronouncing dictionary and grapheme-to-phoneme rules d. Implementing forced alignment

(... And note that the forced aligner can be distributed with the dataset!)

Dataset design:

630 speakers is

- A hard recruiting and scheduling problem (Reading 10 sentences takes only about 90 seconds)
- 2. Not necessary
- 3. Not optimal for most purposes

Instead, suppose we ask each speaker to read for 20 minutes -

TIMIT sentences are about 3 seconds long, so if the inter-sentence interval is 10 seconds, 20 minutes gives us 20*60/10 = 120 sentences per speaker.

Then 50 speakers will give us 6000 sentences, a round number comparable to TIMIT's 6300.

And if we can schedule 5 speakers per day, recording can be finished in 10 days.

(Or a much shorter time using distributed recording methods...)

How about repeating sentences across speakers?

Remember that each TIMIT speaker reads

- 2 sentences that everyone reads
- 5 sentences that 6 other speakers read
- 3 sentences that only they read

We adjusted the proportions a bit, so that each speaker reads

- 20 sentences that everyone reads
- 40 sentences that 9 other speakers read
- 60 sentences that only they read

So we need 20 + 40*5 +60*50 = 3220 distinct sentences

And we end up with 5 groups of 10 speakers each who share 100 sentences, making for easy train/test divisions if desired.

(Note that we can add additional groups of 10 speakers given 40 + 60*10 = 640 additional sentences per group.)

How can we select sentences efficiently?

1. Start with a text collection, such as

- a. a Wikipedia snapshot,
- b. a Wikiquote snapshot,
- c. newswire or magazine text,
- d. e-novels, web forums, bible translations, whatever.
- a. Divide it into sentences
- 1. Reject sentences that are too short, too long, have proper names or etc.
- 2. Pick ~10000 at random
- 3. Have a human judge select 3220 from this set

For example, the LDC's Spanish Gigaword newswire collection contains

- 35,822,282 sentences
- 8,188,987 sentences between 8 and 20 words long
- 2,637,151 sentences between 8 and 20 words long with no non-initial capital letters

Selecting 10 at random from the list of 2,637,151 we get:

Yo no propongo, propugno, ni acepto brotes de violencia.

La brutalidad contra las mujeres es mala sin excepción.

Afirmó que ella trotó a través del apartamento con el niño sobre su hombro.

La ley antimaras tuvo ayer su primera victoria en tribunales.

Láminas de roble francés cuelgan dentro de algunos de los tanques.

Este sábado había varios miles de personas presentes para aclamar a los vencedores.

Un calamar de 8 metros mordió el anzuelo y pudo ser filmado mientras luchaba por liberarse..

Un hombre de 51 años de edad murió al derrumbarse su casa.

Segundo, presta importancia a la "diplomacia de naciones grandes".

A continuación ofrecemos las posiciones en el inicio de la cuarta fecha.

Another obvious source would be Spanish proverbs from Wikiquote – there are 278 of them, of which 108 are 8 words or longer.

10 of these chosen at random are:

A caballo regalado no se le mira el diente.

Ara bien y hondo, cogerás pan en abando.

Mucha paya y poco grano; es por vicio del verano.

El que no oye consejo no llega a viejo.

Con el agua de la bañadera echar también al niño.

Agua blanda en piedra dura, tanto cavadura continua gotera cava la piedra.

Quien bien quiere a Pedro, no hace mal a su perro.

Quien no oye consejo, no llega a viejo.

El favo es dulce, mas pica la abeja.

El hilo siempre se rompe por lo más delgado.

Someone who knows the language and has good judgment needs to check the candidates from such sources –

but generally we can get 3,220 good sentences from 5,000-10,000 candidates.

(Obviously this is harder

for a language with little or no digital text...)

How can we recruit speakers efficiently?

We have used

1. University departments (Standard Chinese, L2 English, L2 Chinese)

- 2. High school students (Guanzhong Chinese)
- 3. Religious organizations (Thai and Ga)
- 4. Cultural groups (Swedish)

A small payment may be made directly to the speaker, or to the sponsoring organization on their behalf. How can we record speakers efficiently?

1. Make up 50 sentence lists, one for each speaker.

- 2. Feed the lists to a program like SpeechRecorder (from BAS), which
 - a. presents the items one at a time,
 - b. records the response,
 - c. Endpoints the recording and stores it in a file as instructed.
- 3. Use a noise-cancelling head-mounted USB microphone, connected to a laptop in a quiet room.

How can we create a forced alignment program efficiently?

- 1. Find or create a pronouncing dictionary (unless orthography is quasi phonetic):
 - 1. Existing digital dictionary
 - 2. Wiktionary
 - 3. e-book
 - 4. ...other more painful options
- 2. Train a grapheme-to-phoneme system (e.g. Phonetisaurus)
- 3. Use a speech recognition toolkit (e.g. HTK or Kaldi) trained on the collected speech

(We have done this for ~six GlobalTIMIT collections now, and it goes very quickly once the dictionary is in place.)

Experience so far:

Completed:

- Standard Thai -- "THAIMIT" (graduate student doing other summer research in Bangkok)
- Standard Mandarin Chinese -- "CHIMIT" (faculty and students at Shanghai Jiao Tong University)
- Chinese learners L2 English (also Shanghai Jiao Tong University)
- 4. Guanzhong dialect of Mandarin Chinese (high school students in Xi'an)
- Ga -- language of Ghana (undergraduate student during summer vacation)

In process:

- 1. Swedish -- recording has started
- 2. American learners L2 Chinese recording has started
- 3. Italian sentence selection done
- 4. French planned
- 5. Spanish?

Plans:

- 1. Do more languages and varieties
- 2. Create a toolkit and instructions

to make it easy for othersto make their own versions.

Options:

 Add brief spontaneous conversation, picture description, etc. to the recording protocol
Use audio prompts for speakers who are not
??? Some examples from collections that are complete or underway:



















Data Intensive Research on Languages of the Americas

References:

John Garofolo, Lori F. Lamel, William Fisher, Jonathan Fiscus, and David Pallett. "DARPA TIMIT acoustic-phonetic continuous speech corpus." *NIST technical report* 1993.

Jiahong Yuan, Hongwei Ding, Sishi Liao, Yuqing Zhan, and Mark Liberman. "Chinese TIMIT: A TIMIT-like Corpus of Standard Chinese", OCOCOSDA 2017.

Nattanun Chanchaochai, Christopher Cieri, Japhet Debrah, Hongwei Ding, Yue Jiang, Sishi Liao, Jonathan Wright, Jiahong Yuan, Juhong Zhan, Yuqing Zhan. "GlobalTIMIT: Acoustic-Phonetic Datasets for the World's Languages", InterSpeech 2018.