Open Source Multi-Language Audio Database for Spoken Language Processing Applications

Stephen A. Zahorian, Jiang Wu, Montri Karnjanadecha
Chandra SekharVootkuri, Brian Wong, Andrew Hwang, Eldar Tokhtamyshev
Department of Electrical and Computer Engineering, Binghamton University, USA
{zahorian,jwu1,kmontri,cvootku1,bwong5,ahwang1,etokhta1}@binghamton.edu

Abstract

Over the past few decades, research in automatic speech recognition and automatic speaker recognition has been greatly facilitated by the sharing of large annotated speech databases such as those distributed by the Linguistic Data Consortium (LDC). Open sources, particularly web sites such as YouTube, contain vast and varied speech recordings in a variety of languages. However, these "open sources" for speech data are largely untapped as resources for speech research. In this paper, a project to collect, organize, and annotate a large group of this speech data is described. The data consists of approximately 30 hours of speech in each of three languages, English, Mandarin Chinese, and Russian. Each of 900 recordings has been orthographically transcribed at the sentence/phrase level by human listeners. Some of the issues related to working with this low quality, varied, noisy speech data in three languages are described.

Index Terms: open source speech database; forced alignment; transcribe; speech recognition

1. Introduction

The need for large well-labeled databases for spoken language processing is well known. "There is no data like more data." is a comment made by MIT speech researcher Victor Zue at a speech recognition workshop in the 1980s. Despite the large number and vast sizes of speech databases developed since the 1980s, the comment by Victor Zue still rings true [1].

One of the first large databases developed for speech research was the TIMIT acoustic-phonetic continuous speech corpus. With joint efforts from Texas Instruments, SR International and MIT, TIMIT was published by the LDC in 1993. This database contains recordings of 630 speakers, each reading 10 sentences, in "studio" conditions. The data was then manually labeled with starting and ending points for each phone in each sentence. Even today, TIMIT is one of the most widely used speech corpora for phonetic level speech research.

However, the 5,040 sentences in TIMIT (the SA sentences are often removed), with a typical sentence duration of 5 seconds, only provide about 7 hours of total speech, which is insufficient for many recognition tasks. Since TIMIT, many other speech databases have been collected, transcribed, catalogued, and distributed by the LDC. For example, the English Broadcast News Transcripts (HUB4) [2] database was launched by the LDC in 1996 and now contains approximately 100 hours of broadcast news and has all speech manually transcribed at the phrasal level, using the Rich Transcription (RT-04S) Evaluation Data guidelines developed by the National Institute of Standard and Technology (NIST) in 2004 [3]. In recent years, LDC announced their plan for developing a new speech database. The DARPA GALE program [4] is a very large speech database project with the goal of collecting speech in multiple languages from global broadcast news. In 2009 the LDC [5] reported that 4,000 hours of Arabic broadcast has been collected and 2,400 hours were selected for transcribing.

Currently, public video sharing websites such as YouTube are booming because sharing homemade videos has become very easy and more popular. About 65,000 videos have been uploaded daily since 2006, and this number continuously increases [6]. This seemingly "infinite" number of videos found on the web can provide a vast collection of speech data for speech research, and the topics and speaking styles corresponding to these collections are much more varied than those found in broadcast news. To tap into this large resource, an "open source multi-language speech database" project was developed and is described in this paper.

2. Structure of the database

2.1. The goals

The database was developed with collections from three different languages: English, Mandarin Chinese, and Russian. The intent was to collect about 30 hours of speech in each language, consisting of 300 videos per language, with videos averaging about 7 minutes in duration. The intent was also to collect three videos from each of 100 speakers per language, with the three recordings from each speaker originally spoken on different days and under different recording conditions. Another goal was to have approximately an equal number of male and female speakers. As is discussed later, most but not all, of these guidelines were met. The only firm guidelines were that the audio portion of each video be of sufficient quality to be "reasonably" intelligible by a "typical" native speaker of the language, that there not be constant background noise (i.e., not have background music throughout the entire passage), and that no single passage be shorter than 1.5 minutes or longer than 16 minutes in duration. Since many videos were in fact longer than 16 minutes, a "stand-alone" primarily speech portion of the video was extracted (using Xilisoft Video converter) for the database.

2.2. Video download and post-preparation

The first step in this development was to identify, download, and store audio/video clips from public video sharing websites. All videos were downloaded in the highest quality format that the sharing sites supported and then stored in a standard format. Table 1 shows the typical sites used for each language, the download tools used, and the original file formats.

For those videos in a format other than MP4, further processing was done, so that all videos were saved as MP4 files. The step was done by Xilisoft video converter, too. A copy of each video in its original format was also kept. Then all files were designated with a specific name based on the language, the gender of the speaker, the initials of the speaker, and the category of recording. The category of recording

comprised "formal presentation" and "casual conversation" which refer to the conditions under which the recording was made. Table 2 shows some criteria for making a decision on which category a video clip belonged to, although the decisions were somewhat subjective.

Table 1. *Typical video sharing websites and tools for downloading.*

Language	*Typical Webs*	*Download Tools*	*Original Format*
English	Youtube.com	www.savevid.com	MP4
Chinese	Youku.com	www.flvcd.com	FLV
Russian	Rutube.ru	www.savevid.com	MP4

Table 2. *Some criteria for separating "Formal Presentation" and "Casual Conversation"*

	Formal Presentation	***Casual Conversation***
Speech type	Speech prepared in advance	Casual talk, semi-spontaneous
Noise/Interruptions in recording environment	Quiet and not much noise	More noise and even distortion effects
Slang/Disfluencies	Very little	Usually a lot
Background music	None	Maybe a little

A standard file name consists of 5 parts, chosen to make it easier to sort and organize the videos. Figure 1 illustrates a file name given to a Chinese language video sample, which is categorized as a "casual conversation," spoken by a male speaker who has initials "WX." The detailed description and possible options for each part as well as their abbreviations in file name are given in Table 3.

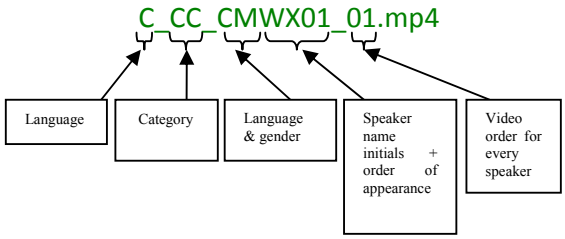

Figure 1. *Illustration of naming conventions for video files*

Table 3. *Filename notation and descriptions*

	Description	***In filename***
Language	English	E
	Mandarin Chinese	C
	Russian	R
Recording Category	Formal Presentation	FP
	Casual Conversation	CC
Gender	English/Chinese/Russian	E/C/R
	Male/Female	M/F
Speaker Name Initials	Initials of first / last name	
	Order of appearance in database	01, 02, 03…
Video Order	Indicates which video of each speaker	01, 02, 03…

3. Transcribing the database

An important aspect of this database is that all speech files were manually transcribed by human listeners. The reason for this was quite simple: given the relatively low quality, the wide variations in recording conditions, the presence of background noises, and the multiple languages, it seemed very doubtful that any automatic speech recognizer would be able to establish reliable "ground truth." By carefully listening to each sentence and reviewing by different listeners, the human transcriptions would provide the best orthographic transcriptions of this "ground truth." Also, the accurately transcribed sentences provides the best starting point for an automatic forced alignment process to create time labels for words and phonemes, thus better supporting phonetic recognition research. Therefore, properly selecting the tool and building the specifications for the transcribing work was necessary.

3.1. Transcribing tool

Transcriber, developed as a tool for assisting in creating the speech corpora in [7], was designed for manual segmentation and transcription of long duration broadcast news recordings, including annotation of speech turns, topics and acoustic conditions. With its embedded user-friendly graphical user interface, Transcriber allows listeners to perform tedious and complex operations such as modifying the time boundary of each speech segment, adding noise notations or indicate switching between speakers in a convenient way. Also, the output of Transcriber accurately records the time durations between segments, which provide support for the following automatic forced alignment process for detailed word and phonetic transcription. All these features made Transcriber extremely well-suited for the transcription task.

Other speech transcription tools considered are listed in Table 4. Although Transcriber does not have all the functions these other tools have, it does have the required features and there is no licensing fee; therefore, Transcriber 1.5.1 was used for this project.

Table 4. *Other available transcription tools*

Tool	***Main Features***
XTrans	Multi-speakers tasks (Developed by LDC)
Transana	Link the transcription place to video
SoundIndex	Directly transcribe in XML file
WaveSurfer	Waveform display/analysis

3.2. Audio preparation

Transcriber only reads WAV files as its audio input. Therefore audio-only WAV files were extracted from the video MP4 files using a software tool called AOA audio extractor. Factors contributing to overall speech quality and intelligibility include: the background noise and recording conditions when videos were made, the loss by website compression tools for uploading files from users, another round of compression by the video download tool, and the final audio extraction. In order that the only significant degradation be due to the first two factors, high quality settings were used for the last two steps. Tables 5 and 6 list the quality setting for downloads and the quality setting for audio extraction, respectively.

Table 5. *Video quality specifications*

Format	***MPEG-4***	
Video	Encoding	MPEG-4 AVC1 (H.264)
	Resolution	Original as on YouTube
Audio	Encoding	AAC
	Channels	2
	Sampling rate	44100Hz

Table 6. *Audio quality specifications*

Format	***WAV***	
Audio	Encoding	PCM
	Channels	2
	Sampling rate	22050 Hz
	Bits/sample	16

3.3. Metadata and annotation

The annotation and metadata for transcribing this database was based on the format used for the LDC project GALE [8]. Unlike studio quality recorded read speech, there is a large amount of variability in transcribing web-collected speech. Main factors which possibly diminish the transcribing quality included the background noise/music, slang, and inserted words in different languages, other than the primary language of the speaker. These additional factors were annotated to the extent feasible.

When selecting videos, background music was considered permissible if it was not "too" high level and did not overlap with speech too often. Music as well as other types of noise were labeled in the transcription using the notation in Table 7. Noise labels differed depending on when the noise occurred relative to the speech. "Burst" noise/music occurred when there was no speech. "Overlap" noise/music occurs during speech. The most commonly used notation for noise was the "Others" notation, since, from listening, it was often quite difficult to accurately determine the source or type of the noise. Many speech signals also contained static noise which was present throughout the signal. This was labeled in a master file that describes each file.

Table 7. *Notations for noise and how it is annotated*

Noise Notation	*Description/Example*	*Symbol/burst*	*Symbol/overlap*
Music		*[mu]*	*[mu-][-mu]*
Applause		*[a]*	*[a-][-a]*
Laughing		*[l]*	*[l-][-l]*
Human	coughing, inhaling,	*[h]*	*[h-] [-h]*
Nature	wind, ocean...	*[n]*	*[n-] [-n]*
Vehicle	honks, car engine…	*[v]*	*[v-][-v]*
Animal	barking, birds…	*[an]*	*[an-] [-an]*
Office	telephone…	*[o]*	*[o-] [-o]*
Machinery	fan, construction...	*[m]*	*[m-] [-m]*
Others		*[ot]*	*[ot-] [-ot]*

Additionally, special events like silence segments and language transitions are potentially as detrimental as noise or other interference. While transcribing, the listeners labeled a "pause" (Table 8) as a speech break (silence) between 0.5 and 1 second. Silences longer than 1 second were labeled differently. For some run-on sentences which commonly occur in casual conversation, great care was taken with the time labels because speech often does not end abruptly, but rather gradually fades.

Table 8. *Notations for important events and how it is annotated*

Event	*Description/Example*	*Symbol*
Pause	Break btw/ .5 to 1 sec	*[p]*
Language Transition	Word(s) spoken in different language	*Exp. [lang=English-] [-lang=English]*

One common issue in spoken Mandarin Chinese is that people mix English words in Chinese sentences. (In constrast, Russian speakers often use English word variants.) Thus, labeling of languages transitions became necessary. Table 8 also shows the symbol for a language transition.

The issues related to slang, truncated words, incomplete pronunciation and other informality in spoken language are clearly addressed by the GALE standard. These issues include: contractions or ambiguous words should be transcribed as closely as possible to how they actually sound; truncated words are to be ended with a dash "-"; the notation "(())" is to be used to represent an unintelligible words; and tilde "~" is to be used when each letter of an acronym is spoken individually

4. Practical issues in transcribing

4.1. Common issues

The most difficult issue in the transcribing process of all three languages stemmed from the great range of qualities in the original video recordings. For example, some videos downloaded from YouTube had High-Definition quality, which has very high resolution for both audio and video. However, most videos were recorded with much lower quality for both video and audio. Some Chinese videos were recorded with defective equipment, resulting in distortion of the speech signal thus causing difficulties even for a human listener.

4.2. English

In English, speakers often fill pauses with fillers such as "um" or "hm," briefly between words, or extensively to gain time for a next thought. Differences in pronunciation due to culture and accent promoted its own share of concerns; various accepted norms of speech (i.e. tomato), and the use of foreign language for brand names, proper nouns, descriptions, and verbs are used in conjunction with English speech. Furthermore, background noise or feedback from the medium used to record video was an additional factor.

Simplifying words and expressions through slang was very common; often times words that end in "-ing" are mispronounced and are transcribed as "-in." For example "sleepin" for "sleeping." And the use of "dope" and "hot" to express emotions or quality. The widespread use of internet acronyms such as "brb" and "lol" occurred occasionally in causal speech, implying the assimilation of today's digital jargon in verbal communication.

4.3. Mandarin Chinese

Other than embedded English words, the transcription of Mandarin Chinese was done using standard Chinese characters for transcriptions. Unlike English, there are no slang or informal language (such as truncated words) related ambiguities in the transcriptions.

However, the Chinese language has a large number of dialects with a resulting big influence on how people pronounce Mandarin [9]. Accents among Mandarin speakers differ significantly. For example, the speakers from the northeast region of China confuse "s-" [ʂ] and "sh-" [□□], and also there was no clear boundary between the nasal consonants "-ng" [ŋ] and "-n" [ɲ] for south China speakers. In the development of the database described in this paper, the listeners always transcribed according to actual pronunciation, even if matching with its neighbor characters did not make linguistic sense.

4.4. Russian

In daily life, some words of the Russian language have a different pronunciation than defined in the dictionary. For example, “тыща” ['t□□□a] is often pronounced “тысяча” ['t□sjt□a], “щас” [□□as] as “сейчас” [sej't□as], and “сёдня” ['sjdnja] as “сегодня” [se'gondnja]. Similar to Chinese and English transcriptions, all these ambiguous words were transcribed as pronounced.

Russians use English words directly sometimes. If the English word is clear with the correct pronunciation, the words are enclosed with English language tags. However, in some cases English words are misused (“Americanisms”): in such cases words are transcribed as they sound in Russian. Also, the Russian language includes some words with a Ukrainian pronunciation. Such words were marked with double parentheses.

5. Summary table of data collected

The goal of 300 videos for each language had been achieved, and roughly 30 total hours of speech for each database has been collected and transcribed, as summarized in Table 9. In the past 9 months, which included about 3 weeks for data collection and approximately 8 months for transcribing, overall, 11 students were involved in the database development.

Table 9. *Summary of Database*

Language	*Speech Type*	*Gender*	*Total Num. of speakers*	*Total Num. of Recordings*	*Total Time*
English	Formal	Male	88	124	14 hrs
	Formal	Female	21	26	2.6hrs
	Casual	Male	36	108	10.5 hrs
	Casual	Female	14	42	4 hrs
		Total	**159**	**300**	**31.1hrs**
Chinese	Formal	Male	59	136	11.1hrs
	Formal	Female	24	56	4.4 hrs
	Casual	Male	44	82	6.4 hrs
	Casual	Female	10	26	2 hrs
		Total	**137**	**300**	**23.9hrs**
Russian	Formal	Male	79	167	20.5hrs
	Formal	Female	26	42	4hrs
	Casual	Male	36	71	8.1hrs
	Casual	Female	17	20	2.2hrs
		Total	**158**	**300**	**34.8hrs**

6. Use of Forced Alignment for Speech Labeling

The goal of this database collection project is to make the database useful for speech processing research. In order to fulfill that goal, the database must provide accurate word-level and phone-level transcriptions, both with time marks. Although the speech data could be transcribed manually by a well-trained phonetician, the manual method is a laborious and time-consuming task, which makes it impractical for a large amount of data. An efficient way of achieving this is to apply automatically derived forced alignment for this more detailed labeling and then use humans to finalize the labeling.

By using an Automatic Speech Recognizer (ASR) “tuned” for a single passage or group of passages, that is, given phonetic models and word models in terms of phonetic lattices, the audio can be accurately time aligned with word transcriptions. The recognizer can be configured to give a best time-aligned match between the audio and transcription, using all the probabilistic constraints imposed by the phonetic and word models. Furthermore, this recognizer can be made much more accurate by configuring it for each recording. For example, based on the human transcription of a recording, the vocabulary can be restricted to only the words in that recording, and the language model derived only from the word and word-pair frequencies in that passage. We are currently experimenting with several techniques for forced alignment on a subset of the English database and the outcome looks very promising. These techniques are being tuned for this particular task and database.

7. Conclusion

In this research project, a large database of English, Mandarin, and Russian was collected, formatted, organized, annotated, and given time-aligned orthographic transcriptions at the sentence/phrase level. Due to the variability, noisiness, and low speech quality, human listeners were employed for this transcription and annotation. Automatic speech recognition techniques will be developed and used to aid in the annotation process at a more fine-grained level. This database will be useful for both automatic speech recognition research and automatic speaker recognition research. The database is derived from open source public web sites; thus it is a sampling of an “infinite,” widely accessed repository of speech.

8. Acknowledgements

This material is based on research sponsored by the Air Force Research Laboratory under agreement number FA8750-10-2-0160. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

9. Disclaimer

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government.

10. References

[1] Zue, V. and Seneff, S., “Speech database development at MIT: TIMIT and beyond,” *Speech Comm.*, 9, 351-356, 1990.

[2] Graff, D., “An overview of Broadcast News corpora,” *Speech Comm.* , 37(1-2): 15-26, 2002.

[3] Garofolo, J. S., Laprun, C. D. and Fiscus, J. G., “The rich transcription 2004 spring meeting recognition evaluation,” *NIST 2004 Spring Rich Transcription Evaluation Workshop*, Montreal, Canada, 2004.

[4] Cohen, J., “The Gale Project: A description and an update,” *ASRU IEEE Workshop on Automatic Speech Recognition & Understanding*, 237-237, Dec. 2007.

[5] Paulsson, N., Choukri, K., Mostefa, D., Dipersio, D., Glenn, M. and Strassel, S., “A large Arabic broadcast news speech data collection,” *MEDAR 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 22-23 April, 2009.

[6] http://en.wikipedia.org/wiki/YouTube

[7] Barras, C., Geoffrois, E., Wu., Z and Liberman, M., “Transcriber: Development and use of a tool for assisting speech corpora production,” *Speech Comm.*, 33(1-2): 5-22, 2001.

[8] http://projects.ldc.upenn.edu/gale/Transcription/

[9] http://mandarin.about.com/od/chineseculture/a/dialects.htm