CMU WILDERNESS MULTILINGUAL SPEECH DATASET

Alan W Black

Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA, USA. awb@cs.cmu.edu

ABSTRACT

This paper describes the CMU Wilderness Multilingual Speech Dataset. A dataset of over 700 different languages providing audio, aligned text and word pronunciations. On average each language provides around 20 hours of sentence-lengthed transcriptions. We describe our multi-pass alignment techniques and evaluate the results by building speech synthesizers on the aligned data. Most of the resulting synthesizers are good enough for deployment and use. The tools to do this work are released as open source, and instructions on how to apply such alignment for novel languages are given.

Index Terms— found speech data, multilingual, speech synthesis, speech recognition.

1. INTRODUCTION

Speech processing, recognition and synthesis, have reached the stage that given good training data, we can build reasonable models that are sufficient to be used in tasks such as speech translation and dialog systems. There has been a number of efforts to make the design and collection of data efficient and such techniques have helped expand ASR and TTS to more languages. But there are around 7000 languages identified, many of which may not have many speakers and hence it is not always easy to find speakers to record. Thus especially for less commercially viable languages, these languages may be left behind. One potential source of data for such other languages is **found** data where recordings collected for some other use might be used to build recognition and synthesis models.

Audiobooks, broadcast news, political speeches are potential sources and have been used when available, but even sites like Librivox.org and news outlets still only have a limited number of languages. Some multilingual datasets aready exist (e.g. TUNDRA [1] and Babel [2]) but they are still limited in the number of languages. To find a much larger number of languages it seems necessary to find places where people are collecting examples for some other, usually non-commerical, purpose. Religious texts, such as the Christian Bible and the Quran are good examples. These are often translated and recorded in order to spread the word. Recordings are often collected for low resource languages as these languages might also be languages of low literacy. We have been building tools over time to help utilize such recorded examples, often mined from Youtube, but the data itself can never be used as is. All speech training systems require well-aligned, relatively short examples of audio and text. Whole chapters or books cannot be used directly. If we have an existing speech acoustic model it is possible to segment such longer recordings into "sentence"-sized chunks, but as we are addressing low resource languages that do not yet have any resources that initial segmentation is not immediately possible. Even once data is found, selecting the best examples for models is a difficult but worthwhile task [3].

This paper introduces the CMU Wilderness Speech Dataset which offers on average 20 hours per language of aligned sentence-lengthed text and audio, derived from around 700 different languages. The dataset is derived from the read New Testaments available from the www.bible.is website. The languages come from all over the world though most are relatively low-resource languages. The dataset quality was evaluated by building a speech synthesizer using standard techniques, that can be deployed on any Android platform.

2. FOUND DATA AND TRANSCRIPTIONS

Processing of found speech data offers the potential for getting lots of speech data without having to actively record it. However such data also brings other issues, it may have large channel variation, and may not be suitably recorded (far field microphones, background music, other background noises). Broadcast news quality recordings may sometimes be available, but even they may have music in the background.

Found data can have textual transcriptions, this is typically the most useful form of found data. Such descriptions can be pre or post recording. Audio books are a good example of pre-recording transcriptions. The reader is reading a particular text. But the reader may sometimes modify the text in simple ways. Post-recording transcriptions, best typified by subtitling, too have their problems. The person doing subtitling may simplify the speech, e.g. omitting hestitations and false starts. Thus although there may be transcriptions attached to an audio recording, they are not of the quality we would expect from transcribed audio that we have conventionally required in ASR, or TTS databases.

There are potentially two ways to find alignments. You can run an ASR system and find the words then try match those to the text you want to align, or you can use the text to generate the phones and try to find those phones in the audio. The first technique is more general but you need a good ASR

to be successful. In this work we use the second technique, which for us so far, has been sufficient.

3. BIBLE.IS DATA

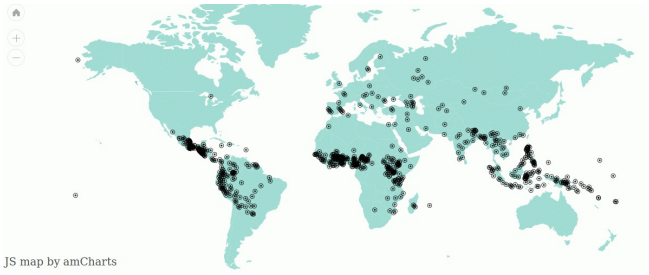
Although on-line forums like Youtube.com often have recordings of less studied languages, it is rare to find large quantities of recordings in a similar format that can be automatically processed. After processing a number of religious recordings from Youtube.com, we noted that the site bible.is has around 1250 different bibles already available in different languages. Of those, around 900 have audio recordings (though not all of those have textual forms). Given the size of the data, the general uniformity of the collections and the diversity of the languages it was felt worthwhile to try to automatically process as many of these as possible in order to provide one of the largest language-diversity speech databases we have ever seen.

The data typically consists of the New Testament recorded as individual chapters (typically 260 chapters). There is a page for each chapter that contains the text of that chapter in the language's orthography, and a link to an mp3 recording. On average, each chapter is about 13 minutes long (varying from 5 to 30 minutes). The recordings are almost always single speaker and mostly male. The recordings seem to be well recorded, probably in a studio, and we suspect collected over a number of different years. Unfortunately, from the speech processing point of view, there is added background music, which is relatively quiet: a flute or string instrament.

There are a few languages where there are multiple recordings from different speakers (and sometimes different versions of the New Testament). Some of the languages have multiple speakers (e.g. Bahasa Indonesian has 3 recordings) or colonial languages (Spanish has 4 recordings, Portuguese has 3) or varying dialects (e.g. there four languages labeled as Arabic and although the text is MSA the speakers are clearly from different parts of the Arabic speaking world). For languages where we have access to speakers we can confirm that the speakers of say Spanish, Portuguese, Russian are not European dialects but South American/Central Asian versions. There are two English recordings both standard southern UK English.

The language distribution tends towards Central and South America, West and East Africa, and South East Asia. Although there are some Chinese dialects (Mandarin, Cantonese, Hakka and Hokkien) and some Central Asian languages. It easy to name languages that are missing from the collection. Also there are only two US/Canadian indigenous language (Objibwe and St Lawrence Yupik) and no Australian indigenous languages.

The orthography used is typically the native orthography, but there there are obvious exceptions: Hakka and Hokkien use a romanized pinyin. We also suspect in some cases the the orthographic choice is partly a legacy one, as in it was the translation that was available at the time. (e.g. Objibwe is available in both a romanized form and in CAS (Canadian Aboriginal Syllabics)). There is a wide range of non-roman scripts, including, Arabic, Hebrew, Cyrillic, various Brahmic scripts, Thai, and Ge'ez. Although the majority of the languages use a roman-based script many use additional characters beyond extended ASCII.



699 Languages Successfully Aligned

4. OVERALL ALIGNMENT PROCESS

We will describe the alignment process in this section. Our intial process was not as complex, but after our experience over several languages we refined the process to what is described here. Each recording on bible.is is identified by a six letter/number code. The first three letters identifies the language (which is often, but not always, the iso 639-2 language code). The second three letters, perhaps, identifies the organization that recorded the data.

Our basic process involves the following steps

- Download html and mp3 versions of the target language from bible.is
- Convert mp3 to wav, and extract the text from the html. This is aligned per chapter.
- Prune silence from waveforms. This is done using an F0 labeling tool, which we have often used in processing speech synthesis databases. We remove sections longer than 250ms that contain no detactable F0. Without this removal our initial alignment sometimes fails, or is measurably worse.
- Initial cross-lingual alignment (Pass 0): using a cross-lingual version of Interslice [4], we find matching initial subsegments of the waveforms that sufficiently match the phone strings generated from the text.
- Build a synthesizer from the successfully aligned data, and resynthesize each utterance and measure each utterance's accuracy.
- Build a target-language acoustic model from the best 85% of the initially aligned data
- target-language alignment (Pass 1): using a model trained on Pass 0 alignment, this usually gives both a larger number of aligned utterances and a better alignment
- Build a new synthesizer and again resynthesize and score the utterances.

5. PRONUNCIATION MODELING

In order for this to work at all, we need a method to generate pronunciations for all of the scripts into a common phone system. Although end-to-end synthesis and recognition systems are currently popular (and can often be quite successful), an end-to-end system requires initial sentence level alignment before it will work. That is exactly what we do not have from this initial data. The work here is trying to create data that would make end-to-end system training feasible. Thus we must find a pronunciation model to allow cross-linguistic grounding of phones from whatever the orthography of the language is.

We use a modification of UniTran [5] that has been substantially modified over the last few years to provide an XSampa phone string for any orthography (that is supported). UniTran's history involves the pronunciation options that appear in the unicode standard. However although we have been using Unitran for much of our grapheme-based synthesis research, we know it is far from optimal. It is basically language independent, thus it ignores orthography distributions in any particular language thus misses digraph/trigraph and contextualized grapheme-to-phone mappings that could be better. Also UniTran is fundamentally segmental. Each orthographic character goes to zero, one, two of more phones. An orthographic character can never modify a contextual phone choice, or change their order (both of which can happen in real othographies). When we train grapheme based speech models (i.e. end-to-end systems) this restriction is usually addressed by the machine learning method. But in our initial alignment, we are treating the orthography as purely segmental and probably losing accuracy because of it.

6. ALIGNMENT MODELING

We have used Interslice [4] to segment large audio recordings in our research for some time (particularly audio books) but until now we have only done this for English. The idea is given an acoustic model, a long piece of text, and a long audio recording: find an initial segment of the audio that matches the phones in the initial piece of text. We have tuned this technique to deal with a small amount of word variation. The speaker may speak the text sligthly differently (e.g. contractions, hestiations, missed/inserted words). For English, Interslice is usually quite good, and later processing allows us to find the best of the alignments, and excludes those that do not align well.

For this case though, we had to do this cross-lingually. We used our largest most varied datasets to build models. Specifically the Arctic English datasets [6] and our Indic datasets [7]. All of these are carefully read speech, with good pronunciation models, and hence we hope give sufficient variation to allow cross lingual modeling.

Initially we allowed for hand mapping of missing phones in the initial alignment. We would stop the process and an expert would look at the distribution of missing phones generated by the pronunciation model and give a mapping to a phone that existed in our cross-lingual model. However in all but a very few cases this actually was not necessary and the system is robust to a few missing phones.

As the language may contain phones that are not in our initial multilingual model, and/or phone names that are not realized in the same way, we use the results from the first multilingual alignment (Pass 0) to build a new in-language acoustic model and perform the alignment again. This gives much better results. On average we align 3305 more utterances in Pass 1 than in Pass 0.

7. MEASURING SUCCESS

Apart from a very few languages we have no access to native speakers of most of these languages. For those languages that we do have speakers (mostly European, Indian and some South East Asian languages) we can confirm that the alignments often are successful. For other languages we must rely on objective measures, though we predict the phonetic streams for all and can listen to the examples. As these are Bible verses they often contain familiar Hebrew names which can be recognized by non-natives.

For measurement of success we use a speech synthesis objective measure. Mel Cepstral Distortion [8] is a weighted mean Euclidean error metric (smaller is better) often used in voice conversion and speech synthesis. It typically ranges from 4.0 (very good) to 8.0+ (not good). Experiments over multiple languages show that a difference over 0.08 is human detectable, and that improvements of around 0.12 may be achieved by doubling the amount of data [8]. Usually this measure is used to compare improvements in modeling within a language/database as cross speaker comparisons may not hold.

We use two speech synthesis techniques to measure quality of the alignments. The first is our Clustergen Statistical Parametric Speech Synthesis System [9]. Its purpose is not to build the best speech synthesizer but to build an acoustic model from the data and give a reliable measure of how good the dataset is as a whole, and also how well a waveform synthesized from the text aligns with the audio segment that was extracted from the larger audio waveform. We use this synthesis technique after Pass 0 to help select which utterances are aligned well enough to be used for the in-language acoustic model used in Pass 1. The following table shows the mean results for the currently processed 699 languages

	Pass 0		Pass 1		
#Lang	#Utt	MCD	#Utt	MCD	Duration
699	7397	7.396	10702	6.827	19H39M

Thus on average there, is just under 20 hours of aligned utterances per language, a total of 13,725 hours of aligned speech.

We find there is an improvement on the number of utterances between Pass 0 and Pass 1 of 3305 and a decrease in MCD on 0.569.

8. BUILDING TTS MODELS

To further test our generated alignments, we build a much more elaborate TTS model. We use same basic Clustergen

model, but improve it with a segmental boundary alignment technique [10]. We also do a multi-model version using random forests [11]. Importantly not only does this produce a much better synthesis model, it also produces a runnable synthesizer for any Android device running CMU Flite [12] without any modification to the core engine.

We use only the best 2000 utterances found in Pass 1, giving around 3-5 hours of speech. The following table shows the mean MCDs for the base CG synthesizer and the Random Forest one, and the base value from Pass 1 showing that the best data in the alignment is in fact better than all the data.

Pass 1 MCD	Base CG MCD	RFS MCD
6.827	5.814	5.629

We were surprised about the improvement when we took only the best 2000 utterances, previous experiments implied more data was almost always better (or at least not worse), but these models are both smaller, and better. The improvement between the base CG model and the RFS model is consistent with earlier findings in [11]. We also confirmed that RFS improvement is greater for "worse" voices.

The following table shows the distribution of MCD scores per languages

MCD score	#Lang
>3.8 and <5.0	118
>5.0 and <6.0	418
>6.0 and <7.0	116
>7.0 and <8.0	47

Voices with MCD less than 5.0 can be considered very good, and fully understandable. Those between 5.0 and 6.0 are still good. Those between 6.0 and 7.0 can sometimes be hard to follow, and those whose MCD is greater than 7.0 and probably not usable as synthesizers.

Note these are not actually full speech synthesis, they will speak any textual input, but do not address issues of numbers, symbols and other non-alphabetic words. They also do not address code-mixing which is extremely common in lower resourced languages. However the better ones are quite reasonable for being used in existing book reading apps.

9. DISTRIBUTION

The process was built to be robust, and only fails for 3% of the languages. However the current whole alignment process is still quite computationally expensive. It takes around 7 days on 12 core machine to complete. As these processes do not use GPUs finding spare machines to run this on, is relatively easy, but even at our peak we were generating only 50 languages per week.

Because the audio data from bibie.is cannot be resdistributed, we distribute indices into the data, and offer a script that will allow end users to reconstruct the aligned data. Thus they can create the aligned data without having to re-execute the whole alignment process. Rather than using up 7 days of CPU time, an aligment can be made in under 30 minutes (much of the time is for downloading and converting the mp3s to waveforms). Its is notable that bible.is is adding new language regularly and also are updating their website so we have been up grading out scripts accordingly.

The data is distributed at

```
https://github.com/festvox/
datasets-CMU_Wilderness
```

which includes a list of the all the languages, a map, example segmentations, example TTS, and standard links to Wikipedia pages describing the language.

A core script `do_found` allows all of the alignment processes to be rerun as well as a fast method to recreate the distribution from the distributed indices. The script also includes the techniques to build a full text to speech synthesizer for each language. This script is built on top the CMU FestVox voice building toolkit also distributed on that github page.

10. DISCUSSION

This is the first pass of building such a large set of languages. It is clear we should do this again with pronunciation and alignment models based on our initial attempts. We expect others will use this data to build their own multilingual (and monolingual models) which could also help improve these initial alignments.

As part of the experiment we investigated which languages do not work well in this framework. We noted that all the Arabic scripted languages (Arabic dialects, Urdu, Farsi and Dari) all had MCD scrores greater than 7.0 (and some greater than 8.0). We improved our pronunciation model for Arabic script, by not predicting a default vowel, which improved all these languages by over 1.0 in MCD score, and for some (Urdu) gave a fully practical synthesizer.

About 3% of the languages have failed to complete a build. We have not investigated all of the reasons yet. However one version of Objibwe uses Canadian Aboriginal Syllabics, which we do not yet support, one version of Hindi (Surnami) text was not Hindi at all even though the audio was. The French data is missing all accented characters (not just the accents on the letter are missing but the whole accented letters themselves are missing). But other failures have no obvious reason, but deserve further investigation, but there are only 20 or so language in this position.

11. CONCLUSION

We provide around 20 hours of aligned sentence-level text and audio for over 700 languages. We show the accuracy of the alignments through the creation of usable speech synthesizer.

We believe this the largest diversity in languages with aligned audio. Such a dataset allows the possibility of not only providing standard speech techologies to low resource languages and their speakers. Such a large dataset also allows input to computationl phonological studies.

12. REFERENCES

[1] 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.

[2] M. Harper, "The IARPA Babel multilingual speech database," `http://www.iarpa.gov/Programs/ia/Babel/babel.html`, 2011.

[3] Erica Cooper, Xinyue Wang, Alison Chang, Yocheved Levitan, and Julia Hirschberg, "Utterance selection for optimizing intelligibility of tts voices trained on asr data," *Proc. Interspeech 2017*, pp. 3971–3975, 2017.

[4] Kishore Prahallad, Arthur R Toth, and Alan W Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[5] Ting Qian, Kristy Hollingshead, Su-youn Yoon, Kyoung-young Kim, and Richard Sproat, "A Python toolkit for universal transliteration.," in *LREC Malta*, 2010.

[6] John Kominek and Alan W Black, "CMU ARCTIC databases for speech synthesis," 2003.

[7] Andrew Wilkinson, Alok Parlikar, Sunayana Sitaram, Tim White, Alan W Black, and Suresh Bazaj, "Open-source consumer-grade indic text to speech," in *9th ISCA Speech Synthesis Workshop*, pp. 190–195.

[8] John Kominek, Tanja Schultz, and Alan W Black, "Voice building from insufficient data-classroom experiences with web-based language development tools.," in *SSW6*, 2007, pp. 322–327.

[9] Alan W Black, "CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling," in *Interspeech, Pittsburgh*, 2006.

[10] Alan W Black and John Kominek, "Optimizing segment label boundaries for statistical speech synthesis," in *ICASSP, Taipei*, 2009.

[11] Alan W Black and Prasanna Kumar Muthukumar, "Random forests for statistical speech synthesis," in *Interspeech, Dresden*, 2015.

[12] Alan W Black and Kevin A Lenzo, "Flite: a small fast run-time synthesis engine," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis. cmuflite.org*, 2001.