Using Synthesized Speech to Improve Speech Recognition for Low-Resource Languages

Luise Valentin Rygaard
University of California, Berkeley
Spoken Language Group, Columbia University

Abstract

The task of building good automatic speech recognizers (ASR) for low-resource languages (LRL) such as Zulu, Tok Pisin, Amharic, and Tagalog is and ongoing challenge due to the limited amount of available computational tools and training data for these languages. In particular, the lack of orthographically transcribed speech data makes it difficult to create useful acoustic models for ASR to learn to associate sounds in the language with written words. We are exploring the possible gains from data augmentation approaches, by synthesizing web data (which occurs in large volume for LRLs) using Text-to-Speech (TTS) synthesis and using this synthesized audio to train acoustic models for ASR. Our results from pilot studies of American English show that even small amounts of synthetic speech from easily obtained text data improve ASR's word error rate (WER) with up to 0.93 absolute points for English. We are exploring differences in performance of ASR engines trained by adding various amounts of synthetic speech created from different genres of web data and using different strategies for creating the TTS systems that synthesize them.

Introduction

Automatic speech recognition becomes still more widely used. Most smartphones come with Google Now, Apple Siri, or similar technologies, and the need for keyword search (KWS) in online audio is increasingly sought for, as more and more speech data is on the web [1] [2]. While the necessary ASR technologies both exist and steadily improve for high resource languages (HRLs) like English and Mandarin Chinese, most LRLs lag behind in the development of ASR systems. In the attempt to overcome this gap, a significant hurdle is the lack of transcribed speech and computational tools for LRLs. Another issue is out of vocabulary words (OOV), which we won't focus on in this paper. The interested reader can find more about research in handling OOVs for ASR in [3] [4] [5].

To overcome the problem of too little orthographically transcribed speech data for LRLs, much interest has been shown in the possibilities of data augmentation. Data augmentation commonly refers to the strategy of introducing unobserved data to expand the dataset for training the ASR. As explained in [6], some methods of data augmentation include semi-supervised training, multilingual processing, acoustic data perturbation [7], and speech synthesis. While the potential of these techniques is widely acknowledged, there is to our knowledge only little research demonstrating successful use of speech synthesis for ASR development [8].

Large amounts of web data are easily accessible for many LRLs, so it seems appealing to try using this for data augmentation. In [9] it is shown how scraped web data can improve the construction of the language model (LM), but also the acoustic model can benefit from the found data. In this paper we use the little explored method of synthesizing found text and including it in the training set to improve the acoustic model that describes the mapping between sounds and written words during ASR training.

We take part in the research motivated by IARPA Babel [10] to build robust ASR systems for Keyword Search (KWS) for LRLs. For this project, IARPA Babel has provided varying amounts of collected telephone conversations in LRLs, and we therefore focus on training our ASR systems on conversational speech. For the data augmentation we use blog text and movie subtitles because [9] showed that those genres lead to relatively better ASR performance than other tested web data when added to the LM in a conversational speech ASR system.

Our text-to-speech synthesizer is trained on small amounts of radio news read by professional speakers in controlled settings. We explore how synthesized blog text and movie subtitles affect our ASR system differently, and how this effect varies as we change the amount of "real speech data" – that is, recorded and transcribed phone conversations -- available. We evaluate performance by comparing the ASR systems' word error rate (WER), which is given as the number of insertions, deletions and substitutions divided by the number of words in the reference text (this might be greater that 100%).

Our pilot studies are all made on American English, because this makes both debugging and informal evaluation easier.

Data

To imitate the situation of creating an ASR system with the data provided by IARPA Babel, we mainly train our ASR on corpora of phone conversations. We used the following three corpora:

CALLHOME American English (CE) consists of 18.3 hours of spontaneous, conversational American English telephone speech, whereof approximately 12 hours are transcribed. The training, development and test sets have 80, 20 and 20 conversations of ~30 minutes each, with ~10 minutes transcribed. Most speakers only take part in one conversation, so there are ~240 different speakers.

Fisher English (FE) is made up of two parts. We used subsets of Part 2 for our recognizers. The full dataset contains 5849 phone conversations lasting up to 10 minutes each. Each speaker participates in a few conversations each, and in most conversations, the speakers do not know each other. The speakers are given a topic to talk about to ensure that a wide variety of words are covered.

Boston University Radio Speech Corpus (BURNC) is a collection of professionally read radio news. The corpus has around 7 hours of transcribed speech read by seven 3 female and 4 male speakers.

All three corpora are collected and distributed by the Linguistic Data Consortium (LDC) [11].

Data Augmentation

Data augmentation is a strategy for creating unobserved data to be used during training of the ASR system. One form of data augmentation is speech synthesis, which we will explore further in this paper. [6] describes several additional types of data augmentation that may improve ASR systems for LRLs.

Synthesized data can both refer to perturbed existing data, and to synthetically created new data. We use the latter strategy and synthesize new TTS speech from text data. We use HTS [12], a widely used HHM-based speech synthesis system, to for both training and synthesis. HTS uses the technique of statistical parametric speech synthesis (SPSS) as opposed to the currently prevalent method of unit selection. The main difference in the two approaches is that unit selection synthesizes speech by piecing together actual segments from the original training data, whereas parametric synthesis creates and uses an *average* from similar sounds in the training data. The main advantage of parametric synthesis over unit-selection is that parametric synthesis systems can synthesize words it has not seen before, because it can create the necessary sequence of phones. Conversely, unit-selection uses original sounds and thus requires a large database of speech to ensure good coverage. For more details on SPSS,

unit-selection and the main differences between the two, see [13].

Speech recognition

We use the online open-source toolkit Kaldi for building our speech recognizers. Kaldi's ASR system is based on finite state transducers (FST), which are commonly used in ASR systems to represent the hidden Markov model (HMM), lexicon, language model and more [14].

We use mel frequency cepstral coefficient (MFCC) for feature extraction – that is, for identifying the linguistic content of the audio and filtering out noise. The acoustic model for mapping sounds to text is represented by a Gaussian mixture model (GMM), and the language model (LM) is an FST. More details about Kaldi, graph creation, decoding and more can be found in [15].

Because the training data determines the content of the language model, different kinds of training data lead to different levels of ASR performance. Namely, the more similar the training data is to the data we decode, the better the results we get, roughly. We therefore focus our ASR system on a specific category of data, namely, conversational speech. For that reason we also synthesize "conversational text" (blogs and movie subtitles) for the data augmentation.

Experiments and Results

To measure the effect of different types of synthesized and the different amounts of transcribed recorded speech, we trained ASR systems on 10 different combinations of real and synthesized speech, and 5 with only recorded speech. Subsequently, we used each recognizer to decode the CE test and development sets, and then compared the WERs.

We used the BURNC corpus to train our TTS system and create a synthetic voice for each of the seven speakers. We then synthesized 10,000 utterances scraped from English blogs as described in [9] and 10,000 English movie subtitles. For the ASR training data, we divided the 10,000 utterances evenly between the seven speakers. Thus, we only used one speaker's version of each utterance so that all speakers had the same number of utterances, and each utterance only appeared in the training set once. This way, we avoid overtraining the language models to the 10,000 sentences. The 10,000 synthesized blog utterances make up 11.3 hours of speech, and the subtitles add up to 4.4 hours.

For the recorded speech, we made subsets with 3, 10 and 13 hours of CE speech (CE#), and 40 and 80 hours of FE data (FE#). We trained recognizers on each of these subsets without any TTS utterances, with the 10,000 blog utterances (B10) and with 10,000 subtitle utterances (S10). For the recognizers trained on the CE subsets, we used the PRONLEX lexicon [16] for the grammar FST, and for the FE based recognizers, we used CMU dictionary [17].

WER	**CE3**	**CE3, B10**	**CE3, S10**
Test	**80.68**	84.26	85.01
Dev	**61.66**	65.74	66.04

Table 1

WER	**CE10**	**CE10, B10**	**CE10, S10**
Test	74.24	***73.33***	*73.43*
Dev	53.1	53.22	***53.03***

Table 2

WER	**CE13**	**CE13, B10**	**CE13, S10**
Test	72.26	*72.12*	***71.92***
Dev	51.79	51.98	***51.08***

Table 3

WER	**FE40**	**FE40, B10**	**FE40, S10**
Test	74.95	75.63	***74.94***
Dev	**54.27**	56.24	54.94

Table 4

WER	FE80	FE80, B10	FE80, S10
Test	**71.9**	73.47	72.35
Dev	**51.61**	52.76	***51.61***

Table 5

In tables 1-5, we see the WER of the different ASR systems. The best performances are in bold, and all recognizers where TTS data improves performance are in italics. As the tables show, the TTS utterances improve the performance of ASR systems trained on the CE subsets of 10 and 13 hours. For the remaining subsets, TTS data either makes no difference or hurts performance. In general, subtitles lead to better performance than blog utterances (probably due to the more conversational style of subtitles), but blog utterances lead to the single biggest improvement of 0.93 absolute points when added to the 10 CE subset. It is interesting to note, however, that only 4.4 hours of synthesized subtitles improve WER of the CE10 ASR by 0.81 absolute points. Overall, subtitles lead to improvements for CE10 and CE13 subsets, and do relatively better than blog utterances for the FE sets, where all TTS experiments hurt performance. Both blog data and subtitles significantly hurt performance of the ASR trained on the CE3 subset.

There are a few things to note about the FE results. First of all, despite the noticeably larger volume of training data, the FE40 and FE80 recognizers do worse than the recognizer trained on the smaller CE13 ASR. One likely explanation is that the CMU dictionary used for the FE recognizers is very large (~140,000 words), and significantly larger than PRONLEX (~93,000 words) used for the CE ASR systems. The large lexicon might cause a too large language model, which will likely hurt performance of the ASR. For this reason, it is not clear how much to conclude from the results from FE + TTS recognizers.

The recognizer trained on the small CE3 dataset performs significantly worse when any synthesized speech is added to the training set. This might be because the synthesized data is too different from the test and development sets we evaluate the ASR system on.

Conclusion

We have presented results that show the effect of using synthesized speech for training ASR systems. Our results suggest that there is great potential in using synthesized speech to improve ASR systems for LRLs, and that movie subtitles might lead to better results than blog utterances. Since large amounts of web data are available for some LRLs, this augmentation by synthesis seems to be an interesting and promising direction to take ASR development in. By synthesizing no more than 4.4 hours of English movie subtitles, we improved performance on an American English ASR system with 0.81 absolute points. With 11.3 hours of synthesized blog utterances, we obtained 0.93 absolute points reduction of WER. It is reasonable to assume that we can get similar ASR improvements with data augmentation for LRLs.

Our experiments only cover English, so to further research with LRLs is needed. In future work it would be interesting to experiment with larger amounts of synthesized data, different genres of text, and other strategies for the speech synthesis – e.g. average voices, selected voices or data selection.

Acknowledgements

I would like to thank my mentor Professor Julia Hirschberg and PhD student Erica Cooper for their excellent supervision during my internship at Columbia University. Also CRA-W, DREU (Distributed Research Experience for Undergraduates), and Columbia University deserve special thanks for making this experience and project possible for me.

References

[1] Victor Soto, Andrew Rosenberg, Lidia Mangu, and Julia Hirschberg, "A Comparison of Multiple Methods for Rescoring Keyword Search Lists for Low Resource Languages," *Interspeech*, 2014.

[2] Victor Soto, Erica Cooper, Lidia Mangu, Andrew Rosenberg, and Julia Hirschberg, "Rescoring Confusion Networks for Keyword Search," *Int'l Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014.

[3] Dogan Can, Erica Cooper, Arnab Ghoshal, Martin Jansche, and et. al., "Web Derived Pronunciations for Spoken Term Detection," in *SIGIR '09 Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Speech Retrieval*, New York, 2009, pp. 83-90.

[4] Dogan Can et al., "Effect of Pronunciations on OOV Queries in Spoken Term Detection," *Acoustics, Speech and Signal Processing, 2009 (ICASSP)*, pp. 3957-3960, 2009.

[5] Christopher White, Abhinav Sethy, and Bhuvana Ramabhadran, "Unsupervised Pronunciation Validation," in *Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

[6] Anton Ragni, Kate Knill, Shakti Rath, and Mark Gales, "Data augmentation for low resource languages," in *Proc. Interspeech*, 2014.

[7] Navdeep Jaitly and Geoffry Hinton, "Vocal Tract Length Perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*.

[8] M. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support Vector Machines for Noise Robust ASR," in *Automatic Speech Recognition & Understanding, 2009*, 2009, pp. 205-210.

[9] Gideon Mendels, Erica Cooper, Victor Soto, Julia Hirschberg, and et. al., "Improving Speech Recognition and Keyword Search for Low Resource Languages Using Web Data," 2015.

[10] M. Harper, "IARPA Solicitation IARPA-BAA-11-02," , 2011.

[11] LDC, "https://www.ldc.upenn.edu/,".

[12] K. Tokuda et al. The HMM-based speech synthesis system (HTS). [Online]. http://hts.sp.nitech.ac.jp/

[13] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039-1064, 2009.

[14] Mehryar Mohri, Fernando Pereira, and Michael Riley, "Speech Recognition with Weighted Finite State Transducers," in *Springer Handbook of Speech Processing and Speech Communication*, Jacob Benesty, M. M. Sondhi, and Yiteng Huang, Eds. Berlin: Springer-Verlag Berlin Heidelberg, 2008.

[15] Dan Povey, Arnab Ghoshal, Gilles Boulianne, Yanmin Qian, and et. al., "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* , Big Island, Hawaii, 2011.

[16] CALLHOME American English Lexicon (PRONLEX). Linguistic Data Consortium. [Online]. https://catalog.ldc.upenn.edu/LDC97L20

[17] Carnegie Mellon University. Speech at CMU. [Online]. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[18] Victor Soto, Erica Cooper, Andrew Rosenberg, and Julia Hirschberg, "Cross-Language Phrase Boundary Detection," *Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[19] Andrew Rosenberg, Erica Cooper, Rivka Levitan, and Julia Hirschberg, "Cross-

Language Prominence Detection," *Speech Prosody*, 2012.

[20] Mark Gales and Steve Young, "The Application of Hidden Markov Models in Speech Recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195-304, 2007.

[21] John Dines, Junichi Yamagishi, and Simon King, "Measuring the Gap Between HMM-Based ASR and TTS," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, 2010.