APPROACHES TO AUTOMATIC LEXICON LEARNING WITH LIMITED TRAINING EXAMPLES

Nagendra Goel[1], Samuel Thomas[2],
Mohit Agarwal[3], Pinar Akyazi[4], Lukáš Burget[5], Kai Feng[6], Arnab Ghoshal[7], Ondřej Glembek[5],
Martin Karafiát[5], Daniel Povey[8], Ariya Rastrow[2], Richard C. Rose[9], Petr Schwarz[5]

[1] Go-Vivace Inc., Virginia, USA, nagendra.goel@gmail.com;
[2] Johns Hopkins University, MD, samuel@jhu.edu; [3] IIIT Allahabad, India;
[4] Boğaziçi University, Turkey; [5] Brno University of Technology, Czech Republic;
[6] Hong Kong UST; [7] Saarland University, Germany;
[8] Microsoft Research, Redmond, WA;[9] McGill University, Canada

ABSTRACT

Preparation of a lexicon for speech recognition systems can be a significant effort in languages where the written form is not exactly phonetic. On the other hand, in languages where the written form is quite phonetic, some common words are often mispronounced. In this paper, we use a combination of lexicon learning techniques to explore whether a lexicon can be learned when only a small lexicon is available for boot-strapping. We discover that for a phonetic language such as Spanish, it is possible to do that better than what is possible from generic rules or hand-crafted pronunciations. For a more complex language such as English, we find that it is still possible but with some loss of accuracy.

Index Terms— Lexicon Learning, LVCSR

1. INTRODUCTION

This paper describes work done during the Johns Hopkins University 2009 summer workshop by the group titled "Low Development Cost, High Quality Speech Recognition for New Languages and Domains". For other work also done by the same team also see [1] which describes work on UBM models, [2] which describes in more detail our work as it relates to cross-language acoustic model training, and [3] which provides more details on issues of speaker adaptation in this framework.

Traditionally pronunciation dictionaries or lexicons are hand-crafted using a predefined phone-set. For building ASR systems in a new language, having a hand-crafted dictionary covering the entire vocabulary of the recognizer can be an expensive option. Linguistically trained human resources may be scarce and prone to errors. Therefore it is desirable to have automated methods that can leverage on a limited amount of acoustic training data and a small pronunciation dictionary, to generate a much larger lexicon for the recognizer.

This work was conducted at the Johns Hopkins University Summer Workshop which was supported by National Science Foundation Grant Number IIS-0833652, with supplemental funding from Google Research, DARPA's GALE program and the Johns Hopkins University Human Language Technology Center of Excellence. BUT researchers were partially supported by Czech MPO project No. FR-TI1/034. Thanks to CLSP staff and faculty, to Tomas Kašpárek for system support, to Patrick Nguyen for introducing the participants, to Mark Gales for advice and HTK help, and to Jan Černocký for proofreading and useful comments.

In this paper, we explore some approaches for automatically generating pronunciations for words using limited hand-crafted training examples. To address the issues of using these dictionaries in different acoustic conditions, or to determine a phone-set inventory, other approaches have been proposed [4, 5]. Use of multiple pronunciations when a much larger amount of acoustic data is available for those words is explored in [6].

In order to cover the words that are not seen in the acoustic training data, it is necessary to have a grapheme-to-phoneme (G2P) system that uses the word orthography to guess the pronunciation of the word. Our main approach is to iteratively refine this G2P system by adding more pronunciations to the training pool if they can be reliably estimated from the acoustics.

We find that for a language like English, the G2P models trained on a small startup lexicon can be very inaccurate. It is necessary to iteratively refine the pronunciations generated by the G2P for each word, while constraining the pronunciation search space to the top N pronunciations. On the other hand, if the language is very graphemic in pronunciation, such as Spanish, G2P models may be very accurate, but miss a number of common alternate pronunciations. Therefore to add more alternates, it helps to use free phonetic speech recognition and align it with the transcripts.

The rest of the paper is organized as follows. In Section 2 we describe the approaches we use to estimate pronunciations. We discuss how we use these approaches for experiments using two languages - English and Spanish - in Section 3. Section 4 talks about the results using the proposed approaches. We conclude with a discussion of the results in Section 5.

2. PRONUNCIATION ESTIMATION

Theoretically, the problem of lexicon estimation of words can be defined as

$$\hat{\text{Prn}} = \arg\max_{\text{Prn}} P(\text{Prn}|W, X), \quad (1)$$

where $P(\text{Prn}|W, X)$ is the likelihood of the pronunciation given the word sequence and acoustic data. If optimized in a un-constrained manner (for the words for which acoustic data is available), each instance of a word could potentially have a different optimal pronunciation. It has been found in practice that doing such an optimization without additional constraints does not improve the system's performance. Also, this approach is not applicable to words that have not

been seen in the acoustic training data. For these words it is necessary to have a well trained G2P system.

2.1. Deriving pronunciations from graphemes

We use the joint-multigram approach for grapheme-to-phoneme conversion proposed in [7, 8] to learn these pronunciation rules in a data driven fashion. Using a G2P engine gives us one additional advantage. Due to the statistical nature of the engine that we use, it is now possible to estimate not only the most likely pronunciation of a word but also to get a list of other less likely pronunciations. Now we can split the pronunciation search into two parts. In the first part, we find a set of N possible pronunciations for each word W by training a G2P with a bootstrap lexicon. We then use the acoustic data X, to choose the pronunciation $\hat{\text{Prn}}$ that maximizes the likelihood of the data.

Using a set of graphoneme (pair of grapheme and phoneme sequence) probabilities, the pronunciation generation models learn the best rules to align graphemes to phonemes. The trained models are used to derive the most probable pronunciation $\hat{\text{Prn}}$ for each word W, such that

$$\hat{\text{Prn}} = \arg\max_{\text{Prn}} P(W, \text{Prn}), \quad (2)$$

where $P(W, \text{Prn})$ is the joint probability of the word and its possible pronunciations. Trained acoustic models are then used to derive the most probable pronunciation $\hat{\text{Prn}}$ for each word W in the acoustic data. Using the acoustic data X, we approximate Eqn 1. as

$$\hat{\text{Prn}} = \arg\max_{\text{Prn}} P(X|\text{Prn})P(\text{Prn}|W) \quad (3)$$

Limiting the number of alternate pronunciations for each word to the top N pronunciations of the word and assuming $P(\text{Prn}|W)$ to be a constant for each word, Eqn 3. reduces to

$$\hat{\text{Prn}} = \arg\max_{\text{Prn} \in \text{ Top N pron. of W}} P(X|\text{Prn}), \quad (4)$$

The trained G2P models are used to generate pronunciations for the remaining words in the training corpus and the recognition language model of the ASR system, not present in the initial pronunciation dictionary.

2.2. Refining pronunciations

We start the iterative process of building a lexicon, using a initial pronunciation dictionary containing a few hand-crafted pronunciations. We use this dictionary as a bootstrap lexicon for training G2P models as described in the previous section. Since we do not have any trained acoustic models yet, we use the G2P models to generate pronunciations for all the remaining words in the recognizer's vocabulary. Our first acoustic models are now trained using this dictionary.

We now use this initial acoustic model to search for the best pronunciations of words as described earlier. In Eq.4, which is essentially a forced alignment step involving a Viterbi search through the word lattices, pronunciations that increase the likelihood of the training data are picked up. We use the set of pronunciations derived from this process to create a new pronunciation dictionary. This new pronunciation dictionary, along with the initial pronunciation dictionary with hand-crafted pronunciations, is used to re-train the G2P models and subsequently new acoustic models. Using these acoustic models to force align the data, we recreate the pronunciation dictionary with the new pronunciations to build new G2P models. This procedure is repeated iteratively until the best performing acoustic models are obtained. We do not retain multiple pronunciations in the dictionary for each word as we did not find this to be helpful. Instead we pick the pronunciation with the maximum number of aligned instances for the word. Before using the resulting dictionary to train the G2P we also discard words where the chosen pronunciation had only one aligned instance in the data.

Table 1. Illustration of aligning phonetic and word level transcriptions

Start frame	word	Start frame	phoneme
10	w_1	8	p_1
		11	p_2
...	...	...	...
21	w_2	21	p_9
		24	p_{10}
		26	p_{11}
28	w_3	28	p_{12}
...	...	...	...
47	w_6	43	p_{25}
		44	p_{26}
		...	...
		48	p_{30}

2.2.1. Approach for phonetic languages such as Spanish

In the case of Spanish, since letter-to-sound (LTS) rules are very simple, the G2P system does not generate sufficient alternates for dictionary learning as described above. We therefore use an unsupervised approach to generate an *optimized* pronunciation dictionary using the acoustic training data. Using an ASR system built with the initial pronunciation dictionary, we decode the training data both phonetically and at the word level. We use the time stamps on these recognized outputs to pick a set of reliable phonetically annotated words. The selection procedure is illustrated with an example below. Table 1 shows an illustration of the phonetic and word level recognition of a hypothetical sentence. The sentence is transcribed at the word level into the sequence of words - "$w_1\ w_2\ w_3\ \ldots w_6$" and at the phonetic level into phonemes - "$p_1\ p_2\ p_3\ \ldots p_{30}$".

In this example, we pick the phoneme sequence "$p_9\ p_{10}\ p_{11}$" as the pronunciation of the word w_2, as their phonetic and word alignments match. In this unsupervised approach, by indirectly using the likelihoods of the acoustic data, we rely on the acoustic data to pick reliable pronunciations.

2.3. Adding more pronunciations to the dictionary using untranscribed audio data

Using the best acoustic models trained in the previous step, new pronunciations are added to the pronunciation dictionary in this step. We use the best acoustic model to decode in-domain speech from different databases. The decoded output is augmented with a confidence score representative of how reliable the recognized output is. The recognized output is also used as a reference transcript to force align the acoustic data to phonetic labels. For this forced alignment step we use a reference dictionary with the top N pronunciations (for example, N=5) from the best G2P model. Using a threshold on the confidence score, reliable words and their phonetic labels are

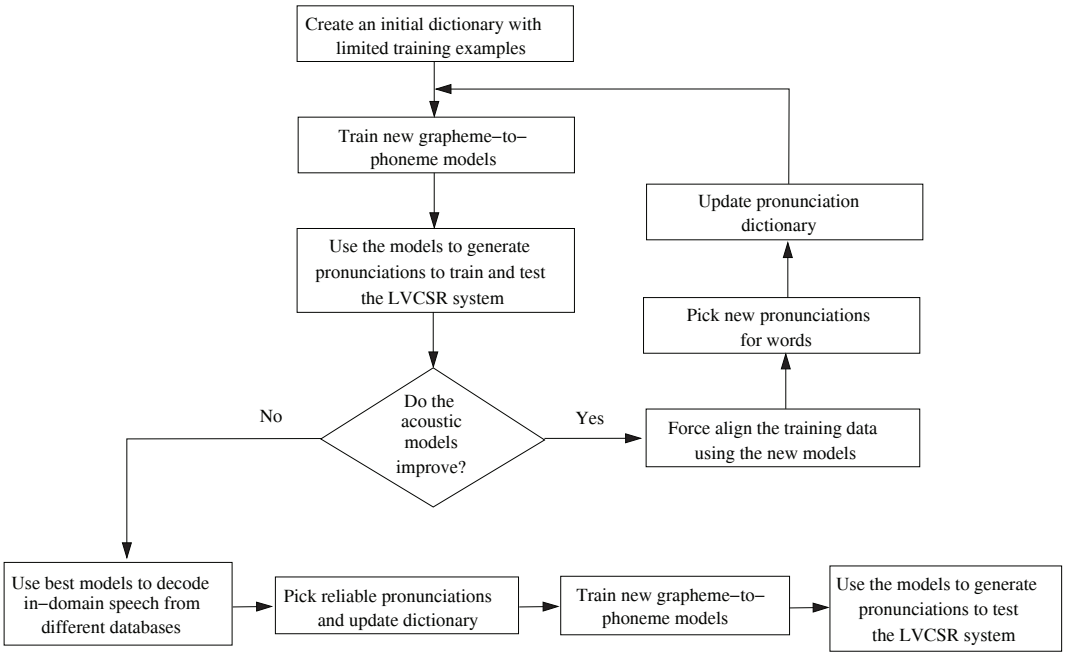

Fig. 1. Schematic of lexicon learning with limited training examples

Table 2. Illustration of a decoded sentence along with confidence scores and aligned phonetic labels

Start frame	word	Confidence score	phoneme
4	w_1	c_1 =0.15	p_1
			p_2
...	...	...	...
11	w_3	c_3 =0.93	p_{10}
			p_{11}
			p_{12}
15	w_4	c_4 =0.84	p_{13}
...	...	...	...
42	w_7	c_7 =0.96	p_{32}
			p_{33}
		...	...
			p_{48}

selected. Table 2 shows an illustration of a decoded sentence along with confidence scores for each word. The sentence is decoded into a sequence of words - "$w_1\ w_2\ w_3 \ldots w_8$" with confidence scores "$c_1\ c_2\ c_3 \ldots c_8$". Using the decoded sequence of words the sentence is also forced aligned into phonemes - "$p_1\ p_2\ p_3 \ldots p_{48}$".

In our case, we set a confidence score threshold of 0.9, and select words like w_3 with its phonetic transcription "$p_{10}\ p_{11}\ p_{12}$". We also remove pronunciations that are not clear winners against other competing pronunciations of the same word instance. We train G2P models after adding new words and their pronunciations derived using this unsupervised technique.

3. EXPERIMENTS AND RESULTS

For our experiments in English, we built an LVCSR system using the Callhome English corpus [9]. The conversational nature of the speech database along with high out-of-vocabulary rates, use of foreign words and telephone channel distortions make the task of speech recognition on this database challenging. The conversational telephone speech (CTS) database consists of 120 spontaneous telephone conversations between native English speakers. Eighty conversations corresponding to about 15 hours of speech, form the training set. The vocabulary size of this training set is 5K words. Instead of using a pronunciation dictionary that covers the entire 5K words, we use a dictionary that contains only the 1K most frequently occurring words. The pronunciations for these words are taken from the PRONLEX dictionary.

Two sets of 20 conversations, roughly containing 1.8 hours of speech each, form the test and development sets. With the selected set of 1K words, the OOV rate is close to 12%. We build a 62K trigram language model (LM) with an OOV rate of 0.4%. The language model is interpolated from individual models created using the English Callhome corpus, the Switchboard corpus, the Gigaword corpus and some web data. The web data is obtained by crawling the web for sentences containing high frequency bigrams and trigrams occurring in the training text of the Callhome corpus. We use the SRILM tools to build the LM. We use 39 dimensional PLP features to build a single pass HTK [10] based recognizer with 1920 tied states and 18 mixtures per state along with this LM.

In our experiments our goal is to improve the pronunciation dictionary such that it effectively covers the pronunciations of unseen words of the training and test sets. Figure 1 illustrates the iterative process we use to improve this limited pronunciation dictionary for English. We start the training process with a pronunciation dictionary of the most frequently occurring 1K words. This pronunciation dictionary is used to train G2P models which generate pronunciations for the remaining unseen words in the train and test sets of the ASR system. As describe in Section 2, we use the trained acoustic models to subsequently refine pronunciations. The forced alignment step picks pronunciations that increase the likelihood of the training

Table 3. Word Recognition Accuracies (%) using different iterations of training for English

Iteration 1	41.38
Iteration 2	42.0
Iteration 3	41.45
Iteration 4	42.93
Iteration 5	42.77
Iteration 6	42.37
Iteration 4 + new pronunciations from un-transcribed switchboard data	43.25
Full training dictionary	44.35

data from a set of 5 most likely multiple pronunciations predicted by the model. We select close to 3.5K words and their pronunciations from this forced alignment step, after throwing out singletons and words that don't have a clear preferred pronunciation. These new pronunciations along with the initial training set are then used in the next iteration. We continue this iterative process as the performance of the recognizer increases.

We start with models trained using only 1K graphonemes (word-pronunciation pairs). For each subsequent iterations, pronunciations from forced alignments are used to train new grapheme-to-phoneme models. Table 3 shows the word accuracies we obtain for different iterations of lexicon training. We obtain the best performances in Iteration 4. We use G2P models of order 4 in this experiment.

To add new words and their pronunciations to the dictionary, we decoded 300 hours of switchboard data using the best acoustic models obtained in Iteration 4. The decoded outputs were then used as labels to force align the acoustic data. Using the approach outlined in Section 2.4, we use a confidence based measure to select about 2.5K new pronunciations. These pronunciations are appended to the pronunciation dictionary used in Iteration 4. We added the pronunciations with a precedence to ensure that words in the pronunciation dictionary have the most reliable pronunciations. We used the order - limited hand-crafted pronunciations, followed by pronunciations from forced alignment with best acoustic models and finally pronunciations from unsupervised learning, while allowing only one pronunciation per word. New grapheme-to-phoneme models are trained using this dictionary. Without retraining the acoustic models, we used the new grapheme-to-phoneme models to generate a new pronunciation dictionary. This new dictionary is then used to decoding the test set. Adding additional words and pronunciations using this unsupervised technique improves the performance still further from 42.93% to 43.25%. To verify the effectiveness of our technique we use the complete PRONLEX dictionary to train the ASR system. When compared to the best performance possible with the current training set, the iterative process helps us reach within 1% WER difference with the full ASR system. We use G2P models of order 8 while training with the complete dictionary.

In the second scenario of Spanish, the written form is phonetic and simple LTS rules are usually used for creating lexicons. For our experiments, we build an LVCSR system using the Callhome Spanish corpus. We attempt to improve the pronunciation dictionary for this language by creating an *optimized* initial pronunciation dictionary using the acoustic training data. Similar to the English database, the Spanish databases consists of 120 spontaneous telephone conversation between native speakers. We use 16 hours of Spanish to train an ASR system as we described before. We use an automatically generated pronunciation dictionary from Callhome as the initial pronunciation dictionary. After training an ASR system

Table 4. Word Recognition Accuracies (%) using different initial pronunciation dictionaries for Spanish

Using automatically generated LDC pronunciations	30.45
Using optimized pronunciation dictionary	31.65

using this dictionary, we decode the training data both phonetically and at the word level. As described in Section 2.2, we derive a set of reliable pronunciations by aligning these transcripts. We use this new dictionary to train grapheme-to-phoneme models for Spanish. Similar to the English lexicon experiments, we train new acoustic models and grapheme-to-phoneme models using reliable pronunciations from a forced alignment step. Table 2 shows the results of our experiments with the Spanish data. Using an improved dictionary improves the performance of the system by over 1%.

4. CONCLUSIONS

We have proposed and explored several approaches to improve pronunciation dictionaries created with only a few hand-crafted samples. The techniques provide improvements for ASR systems in two different languages using only few training examples. However, the selection of the right techniques depends on the nature of the language. Although we explored unsupervised learning of lexicon for English, we did not combine that with unsupervised learning of acoustic models. However we plan to do that and hope that this would make a powerful learning technique for resource poor languages.

5. REFERENCES

[1] D. Povey et. al., "Subspace Gaussian mixture models for speech recognition", in submitted to: ICASSP, 2010.

[2] Lukas Burget et. al., "Multilingual acoustic modeling for speech Recognition based on subspace Gaussian mixture models", in submitted to: ICASSP, 2010.

[3] Arnab Ghoshal et. al., "A novel estimation of feature-space MLLR for full-covariance models", in submitted to: ICASSP, 2010.

[4] T. Slobada and A. Waibel, "Dictionary learning for spontaneous speech recognition", in ISCA ICSLP, 1996.

[5] R. Singh, B. Raj and R.M. Stern, "Automatic generation of phone sets and lexical transcriptions", in IEEE ICASSP, 2000, pp. 1691-1694.

[6] Chuck Wooters and Andreas Stolcke, "Multiple-Pronunciation lexical modeling in a speaker independent speech understanding system", in ICSA ICSLP, 1994

[7] Sabine Deligne and Frédéric Bimbot "Inference of variable-length linguistic and acoustic units by multigrams", in Speech Commun., vol. 23,3, 1997, pp. 223-241

[8] M. Bisani and H. Ney, "Joint sequence models for grapheme-to-phoneme conversion", Speech Communication, vol. 50, no. 5, pp. 434-451, 2008.

[9] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech," Linguistic Data Consortium, 1997.

[10] S. Young et. al., "The HTK Book," Cambridge University Engineering Department, 2009.