UNSUPERVISED PRONUNCIATION VALIDATION

[1]Christopher M. White, [2]Abhinav Sethy, [2]Bhuvana Ramabhadran
[3]Patrick Wolfe, [4] Erica Cooper, [5] Murat Saraclar, [1]James K. Baker

[1] HLT Center of Excellence, Johns Hopkins University, [2] IBM, [3] Harvard University
[4] Massachusetts Institute of Technology, [5] Bogazici University

ABSTRACT

This paper addresses selecting between candidate pronunciations for out-of-vocabulary words in speech processing tasks. We introduce a simple, unsupervised method that *outperforms* the conventional supervised method of forced alignment with a reference. The success of this method is independently demonstrated using three metrics from large-scale speech tasks: word error rates for large vocabulary continuous speech recognition, decision error tradeoff curves for spoken term detection, and phone error rates compared to a handcrafted pronunciation lexicon. The experiments were conducted using state-of-the-art recognition, indexing, and retrieval systems. The results were compared across many terms, hundreds of hours of speech, and well known data sets.

***Index Terms*—** Speech processing, Speech recognition, Speech synthesis

1. INTRODUCTION

Several speech processing applications including large vocabulary continuous speech recognition (LVCSR), spoken term detection (STD), and speech synthesis rely on a fixed vocabulary and a pronunciation for each word therein. This pronunciation lexicon typically contains mappings from an orthographic form of a word (e.g. QUEDA) into a phonetic form (e.g. /k aa d ax/) that can be used for decoding, indexing, retrieval, or synthesis. Although the pronunciation lexicon remains fixed, realistic use requires a constantly changing vocabulary resulting in words that are out-of-vocabulary (OOV). OOVs can be new words, rare words, foreign words, or words unknown to be important at the time the lexicon was formed. Adjusting to change in vocabulary demands the generation of pronunciations and a need to automatically select between candidates. Therefore, this work addresses the question: given a word in any language, and a set of candidate pronunciations, how can you determine the best pronunciation of that word?

Challenges within OOV modeling and pronunciation validation are not new, but issues with OOV words have traditionally been given less attention due their low impact on word error rate (WER). Recent work [1, 2] and the development of the STD task [3] have highlighted their importance as rare words, which are therefore information rich.

An initial approach to create pronunciations for OOV employs a trained linguist, but they are expensive, often produce inconsistent representations, generate few pronunciations per hour, and have limited areas of expertise [4]. Therefore effort has been made toward data-driven pronunciation modeling. Previous work [4, 5] addressing pronunciation validation comes from pronunciation modeling attempting to simultaneously generate/validate pronunciations using existing lexica [4], linguistic rules [6], speech samples [7, 8], or all of the above (see literature pronunciation modeling, grapheme-to-phoneme, letter-to-sound). Such work generally includes criteria for creating a pronunciation for an OOV that involves the modality of data used to create it (e.g. generating pronunciations from lexica tests against comparisons to held out entries in the lexica, generating pronunciations from speech forced-alignment uses accuracy or WER of speech samples). The previous work on data-driven pronunciation modeling addresses pronunciation variation [4, 8] or common words [5, 9]. In [7] they concentrate on names and places, directory services, noting that proper names can be hard where it is difficult to reuse letter-to-sound rules from common words.

For example, in [7] they learn pronunciations from audio samples along with rules from an existing lexicon and develop an iterative algorithm for pronunciation refinement; accuracy of recognition on directory assistance samples is measured. For many cases using speech samples including [6], the standard score comes from aligning the speech sample of a word against the putative pronunciation, sometimes with a filler model for likelihood ratio threshold. In [6] they augment acoustic likelihood with linguistic features and use a decision tree classifier rather than a threshold; they attempt to verify pronunciations for literacy assessment and treat the problem as estimating a confidence score over a short utterance (the word of interest).

This work departs from the standard framework of simultaneously generating and testing pronunciations. We are agnostic about where candidates come from, isolating the task to choose between them. Furthermore, we concentrate on a large number of difficult words, of which many are foreign proper names and places. Our evaluation involves large-scale speech tasks with large data sets in an effort to present results that generalize. We use two methods to select between candidate pronunciations: a conventional supervised method via forced-alignment, and unsupervised method via recognition.

We compare these two methods via three metrics: phone error rate (PER) against a reference pronunciation to analyze the difference with a handcrafted lexicon, WER for LVCSR to see impact on their recognition as well as their impact on recognizing other words in the vocabulary, and decision error tradeoff (DET) curves for STD for searching OOVs. The end goal was to identify a methodology for picking correct pronunciations. This work was conducted as part of the Johns Hopkins University summer workshop (JHUWS08) team 'Multilingual Spoken Term Detection' where pronunciations were generated via letter-to-sound models, those augmented from web data, or from transliteration models .

2. BASELINE SUPERVISED METHOD

Our baseline mechanism for choosing between two candidate pronunciations was to pick the pronunciation with higher average acoustic likelihood from a forced-alignment with a reference, with the average taken over several speech samples.

Performance is measured from approximately 500 words via three metrics: edit distance to a reference lexicon, WER on decoding 100 hrs of speech, and STD DET curves on the LVCSR lattices for the same 100 hours.

2.1. Data set, OOV terms, systems

Our goal was to address pronunciation validation using speech for OOVs in a variety of applications (recognition, retrieval, synthesis) for a variety of types of OOVs (names, places, rare/foreign words). To this end we selected speech from English broadcast news (BN) and approximately 500 OOVs. The OOVs were selected with a minimum of 5 of acoustic instances per word, and common English words were filtered out to obtain meaningful OOVs (e.g. NATALIE, PUTIN, QAEDA, HOLLOWAY). Once selected, these were removed from the recognizer's vocabulary and all speech utterances containing these words were removed from training. For each OOV, two candidate pronunciations are considered, each from a variant of a letter-to-sound system. These OOVs were taken from a larger set used to compare web-data augmented letter-to-sound systems, a subset on which two particular letter-to-sounds systems differed. For details the reader is referred to [8].

The LVCSR system was built using the IBM Speech Recognition Toolkit [10] with acoustic models trained on 300 hours of HUB4 data with utterances containing OOV words excluded. The excluded utterances (around 100 hours) were used as the test set for WER and STD experiements. The language model for the LVCSR system was trained on 400M words from various text sources. The LVCSR system's WER on a standard BN test set RT04 was 19.4%. This system was also used for lattice generation for indexing for OOV queries in the STD task along with the OpenFST based Spoken Term Detection system from Bogazici University [11].

2.2. Supervised validation

Let X denote a sequence of acoustic observation vectors; the objective of the recognizer is to find the most likely word sequence W^* given the acoustic vectors:

$$W^* = \arg\max_W p(W|X) \tag{1}$$

$$= \arg\max_W p(X|W)p(W) \tag{2}$$

where Equation 2 comes from rewriting Equation 1 using Bayes' rule and considering that $p(X)$ does not play a role in the maximization; $p(X|W)$ denotes the acoustic likelihood of the acoustic observations given a word sequence hypothesis W; $p(W)$ is the prior probability of that word sequence W as defined by a language model.

The conventional method for selecting between pronunciation candidates involves using a transcript and performing a forced alignment against it: during alignment there is a constraint in decoding path W to the reference transcript (with each word replaced by its pronunciation in the lexicon), augmented with candidate pronunciations. Speech data that contain the OOV are aligned with the acoustic models corresponding to each candidate pronunciation via Viterbi search, and the maximum likelihood acoustic score determines the 'winner' candidate [5, 4, 7, 6].

Some of the work referenced above attempts to improve the decision function or include additional information while simultaneously generating and validating pronunciations. Our work assumes that pronunciations have been provided and seeks to decide between them. Also, this work concentrates on simple and fast methods for large scale heterogeneous applications.

3. UNSUPERVISED METHOD

Using standard automatic methods (e.g. Section 2.2) for verifying pronunciations requires transcribed audio, which can cost as much as 100$/hr (common) - 400$/hr (new language) to transcribe. Transcription is time-consuming, laborious, and difficult to recruit/keep labelers for transcribing. However, in many applications meta-data can alleviate the need by pointing to speech likely to contain a word of interest, which can be used to select between candidate pronunciations for that word. For example, items in the news, television shows, etc. are a rich source of untranscribed speech for unsupervised validation.

Moreover, often we do not have access to a transcript corresponding to audio examples of an OOV, but we may have some knowledge it has occurred in an audio archive. For example, we may know from meta-data that a broadcast news episode recently aired about a conflict in Iraq, and at present it would give us high confidence to find examples of words like

word	hyp prons	ref prons	phn err%
QAEDA	k aa d ax	k ay d ax	
QAEDA	**k aa ey d ax**		
.			40
SCHIAVO	**sh ax v ow**	s k h aa v ow	
SCHIAVO	s k y ax v ow	sh iy aa v ow	

Table 1. Example pronunciations and PER

QUEDA. We may not know how many times it was spoken, or where in the audio, but we can still use the entire broadcast to help us choose between hypothesized pronunciations for QUEDA.

In the *absence* of labeled examples we use unsupervised recognition to select between candidate pronunciations. We decode data likely to contain the OOV with each candidate, calculate the average acoustic likelihood over the entire data, and choose the candidate with the highest average likelihood as the 'winner'. This corresponds to using Equation 1 to decode speech 'as is' (without the extra constraint on the decoding path to the reference as in the supervised case).

4. RESULTS

For each of the metrics below, a pronunciation lexicon was created for the set of OOVs (approximately 500). For every OOV there were two candidate pronunciations from different letter-to-sound systems, and we compare the two methods described above for choosing between the two candidates for this set (along with an 'upper-bound' and 'lower-bound'). These 500 words were removed from a handcrafted lexicon, therefore we have a set of 'true' pronunciations. The 'upper-' and 'lower-bound' take advantage of this knowledge, denoted $plex - best$ and $plex - worst$. $plex - best$ selects the candidate that is the closest (in edit distance) to a reference pronunciation that word, and $plex - worst$ selects the farthest.

For example, in Table 1 two OOVs are listed, each with two hypothesized pronunciations. Here, $plex - best$ would have as members '/k aa d ax/' and '/sh ax v ow/'.

The two methods compared are those described above, where $sup - force$ denotes the lexicon created from selecting pronunciations based on supervised forced-alignment with a reference, and $unsup - reco$ denotes the lexicon created from selection based on unsupervised decoding. For the unsupervised case approximately the time for one broadcast news show was decoded using each candidate pronunciation, making sure to include all the speech examples used for the forced-alignment somewhere in the data.

4.1. Large vocabulary continuous speech recognition

In addition to comparing methods using the performance in speech tasks, we can see which method produces pronunciations that are closest to a reference. For example in Table 1, if speech had selected the bold pronunciations, there are 4 errors out of 10 phones w.r.t. the closest reference pronunciation (e.g. QAEDA: /ay/ to /aa/, insert /ey/; SCHIAVO: insert /iy/, /ax/ to /aa/) resulting in a 40% PER.

Since the $plex - best$ was artificially selected for this metric, it becomes the upper-bound (although this isn't the case for speech tasks shown below). In Figure 1 the PER is plotted for each of the methods at 3 system configurations. The 3 configurations were created with different levels of language model pruning, and demonstrate differences based on system performance (in WER). The systems' WER on the RT04 data set at the various configurations were 29.3%, 24.5% and 19.4% corresponding to 360, 390, and 450 respectively. Note the x-axis is $\#words$, which corresponds to the number of the OOV types that were decoded via the unsupervised method, and hints at a limitation that will be discussed below. With regard to PER, the $unsup - reco$ has lower error rate at all system configurations compared to $sup - force$, which accords with the results below.

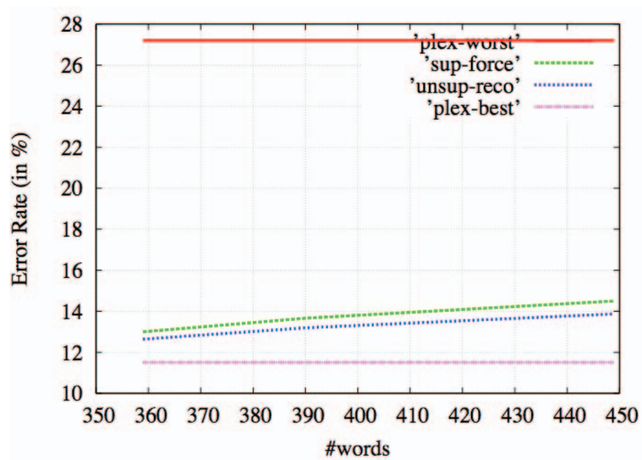

Fig. 1. Phone Error Rate w.r.t. reference lexicon

The methods for selecting between candidate pronunciations described above were used to decode 100 hours of speech that contained all of the OOVs. Standard WER was used to compare these methods in Table 2. Note that $unsup - reco$ outperforms all others. Also, note that the candidate pronunciations give about a half percent WER range (between the best and worst), and that selecting based on the phone edit distance to the reference does not directly translate to better ASR WER.

4.2. Spoken term detection

Lattices generated by the LVCSR system for the 100 hours test set were indexed and used for spoken term detection experiments in the OpenFST based architecture described in [11]. Our goal was to see whether our WER results correlated with another speech task like spoken term detection. To

Method	ASR WER%	#errors
plex-worst	17.8	193,145
sup-force	17.3	187,772
unsup-reco	**17.3**	**187,424**
plex-best	17.4	188,517

Table 2. LVCSR WER

this end, the same sets of pronunciations were used as queries to the STD system. Results from the OpenFST based indexing system are presented in a DET curve using NIST formulas and scoring functions/tools from the NIST 2006 evaluation. The DET curves in Figure 2 show that $plex - best$ and $unsup - reco$ work the best for detection at nearly all operating points.

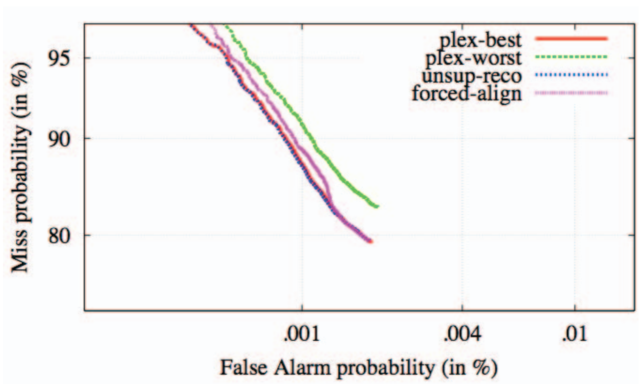

Fig. 2. STD DET Curves

5. DISCUSSION

We have presented an unsupervised method for pronunciation validation via recognition that works *better* than conventional validation via forced-alignment. This success has been demonstrated using 3 metrics for large-scale speech tasks: Phone Error Rate on a large set w.r.t. a reference lexicon, LVCSR Word Error Rate on decoding a 100 hours of speech, and STD DET Curves on the same.

The usual argument for unsupervised speech methods: they save considerable time and money over speech transcription or using a linguist, which is enticing as long as the performance degradation isn't too harmful. However, for selecting a candidate pronunciation our unsupervised method does not suffer any degradation, and actually performs better as it naturally filters out unhelpful speech samples by employing the power of comparison (search) and a language model. In all of the experiments our notion of phone errors were based on a word-to-phone pronunciation lexicon; there were no manual phonetic transcriptions used.

There are several limitations to this method. Unsupervised recognition can't always verify a word (if neither pronunciation is ever decoded), although this provides a natural check against comparing many bad candidates (alignment will always give a score). It requires having seen it or words like it in text (LM), which is not unreasonable given that a word comes into fashion somehow. It's possible that false alarms might hurt (if an OOV sounds like common word), but the 3 configuration experiments indicate that isn't a problem for these words of interest. Finally, the performance could depend on amount or type of data decoded, which is the basis of our future work.

6. REFERENCES

[1] C. White, G. Zweig, L. Burget, P. Schwarz, and H. Hermansky, "Confidence Estimation, OOV Detection, and Language ID using Phone-to-Word Transduction and Phone-Level Alignments," in *Proc. ICASSP*, 2008.

[2] L. Burget et al., "Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs," in *Proc. ICASSP*, 2008.

[3] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary Independent Spoken Term Detection," in *Proc. SIGIR*, 2007.

[4] M. Riley et al., "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, 1999.

[5] J. M. Lucassen and R. L. Mercer, "An Information Theoretic Approach to the Automatic Determination of Phonemic Baseforms," in *Proc. ICASSP*, 1984.

[6] J. Teppermann et al., "Pronunciation verification of children's speech for automatic literacy assessment," in *Proc. ICSLP*, 2006.

[7] F. Beaufays, A. Sankar, S. Williams, and M. Weintraub, "Learning Name Pronunciations in Automatic Speech Recognition Systems," in *Proc. ICTAI*, 2003.

[8] B. Ramabhadran, L. R. Bahl, P. V. deSouza, and M. Padmanabhan, "Acoustics-Only Based Automatic Phonetic Baseform Generation," in *Proc. ICASSP*, 1998.

[9] T. Vitale, "An algorithm for high accuracy name pronunciation by parametric speech synthesizer," *Computational Linguistics*, 1991.

[10] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proc. ICASSP*, 2005.

[11] Siddika Parlak and Murat Saraclar, "Spoken Term Detection for Turkish Broadcast News," in *Proc. ICASSP*, 2008.