IMPORTANCE SAMPLING OF DELTA-AUC: A BASIS FOR ACTIVE LEARNING FOR IMPROVED KEYWORD SEARCH

Kerri Barnes *Matthew Snover* *Man-Hung Siu* *Herbert Gish*

Raytheon BBN Technologies Cambridge, MA 02138, USA
{kbarnes,msnover,msiu,hgish}@bbn.com

ABSTRACT

We present an importance sampling based approach to the active learning problem of selecting additional training data to supplement a seed model. Our proposed Δ-AUC selection optimizes AUC improvement in keyword search and is evaluated on the Spanish Fisher corpus. We show that over different training data sizes, Δ-AUC selection consistently outperforms random sampling by 1.05% to 2.69% absolute AUC and requires no more than 60% of the transcriptions needed by random sampling to achieve the same AUC. On terms not seen in the original seed model training, the proposed algorithm achieves a 3.47% better AUC and 4.66% reduction in word error rate. We also introduce a regression analysis model that can refine our Δ-AUC strategy in the future.

Index Terms— Active learning, CTS, STT, KWS

1. INTRODUCTION

In this paper we consider improving the training efficiency of a speech recognizer through the process of active learning, in which algorithmically selected utterances are transcribed for incorporation into a training set such that overall transcription effort can be significantly reduced.

We consider the case where we have at most 40 hours of transcribed conversational telephone speech (CTS) available. With this maximum amount of training available we can expect word error rates in the vicinity of 40-60% depending on language and other variables. Transcripts at such word error rates are often not in the range that is considered usefully readable. However, such systems when used for keyword search (KWS) can return accurate enough hits. Thus we focus on the AUC (area under the word precision recall curves) metric, sometimes referred to as MAP (mean average precision) which measures our ability to spot a set of words. For this metric, each query term produces a ranked list of hits, and the area under the precision-recall curve is averaged with all other query terms to produce the AUC. Note that AUC weights all terms equally, regardless of their frequency. Word error rate (WER), on the other hand, weights frequently occurring words most heavily, so large improvements to rare terms will not have a large impact on improving overall WER.

We assume that we have an initial out-of-domain seed recognizer available that has been trained with a nominal amount of transcribed speech. The problem we address is the selection of additional utterances to improve performance. Our approach is to automatically transcribe the untranscribed speech with the seed recognizer and then rank each automatically transcribed utterance with respect to its importance to the Δ-AUC (the improvement in AUC after adding additional training). We follow the use of importance sampling in Monte Carlo methods, in which one chooses a sampling function to emphasize the selection of samples that are most important to evaluate the expectation of a function [1]. This typically leads to an accurate estimate with a low variance using many fewer samples than would be used with random sampling. With this motivation we have chosen a reward function that ranks the importance of a sample, i.e., an utterance, according to its potential of increasing the Δ-AUC rather than improving WER. The highest ranked utterances are employed for providing a desired amount of new training to be transcribed.

In Section 2 we describe our Δ-AUC selection criterion and the details of our reward ranking function. Section 3 describes the experimental setup, and Section 4 gives our results where we show that the selected reward function significantly outperforms the use of word confidence as a selection criterion and can reduce the required amount of utterances to be transcribed by over 50% when compared to random sampling of utterances. Section 5 provides a regression analysis of measurable terms to determine their importance to increasing Δ-AUC. This leads us to Section 6 in which we conclude and discuss future work. While there have been other papers [2–6] that employ active learning techniques, they have focused on improving WER, although [7] reports KWS results in addition to WER. None, however, have considered importance sampling of the Δ-AUC as a basis for active learning. Also, unlike other approaches, our active learning strategy focuses directly on our scoring function - the AUC. In effect, we are estimating the gradient of the AUC function.

2. Δ-AUC SELECTION CRITERION

To improve the AUC of a given term, one of the most important factors, after adding it to the decoding dictionary, is

to find instances of that word for training. As shown in Table 1, a small number of training examples can give a large boost in AUC. Δ-AUC is shown for a set of terms that had no training in an out-of-domain seed model. An additional 10 hours of in-domain data[1] was added to the model and we examined the Δ-AUC for terms where N in-domain examples were added to training. The Δ-AUC as measured on a 168 hour test set shows that terms with no additional training examples ($N = 0$) gained modestly due to general acoustic and language model improvement. Much larger gains were found for adding a single training example of a term, but adding a second instance gave diminishing improvements.

N	#Terms	Seed AUC	Seed + 10 hr AUC	Δ-AUC
+0	4903	0.3912	0.4168	0.0256
+1	1027	0.4576	0.5178	0.0601
+2	249	0.4607	0.5367	0.0760

Table 1. Δ-AUC on rare terms when N additional instances are added to training. All terms had 0 training instances in the seed model. AUC is reported only on terms in each row.

We capture the effect of diminishing returns from additional training for word w with an importance weighting function: $I(w) = 2^{-T(w)}$, where $T(w)$ is the number of instances of w that have been transcribed, either in training or in previous iterations of active learning. This weighting function will be used in the Δ-AUC criterion to capture the importance of including an additional instance of word w.

Because AUC weights all words equally, improving terms that are Out-Of-Training (OOT) or Rare-In-Training (RIT) can have as much impact on AUC as improving more frequent terms. For the experiments in this paper, we defined RIT as terms that are seen in training, but occur less than 5 times. The set of words that are in training and seen 5 or more times will be called Frequent-In-Training (FIT). During active learning selection, since we believe that finding more instances of the FIT terms will have diminishing returns, we instead target the rare terms (including both RIT and OOT terms) by assigning a score to each utterance based upon the importance function $I(w)$ and the following factors:

- $\Pr(w_i)$: the probability of word instance w_i in the consensus network output of the recognizer. This factor helps the selection to avoid hallucinated instances of a term, so that when an annotator transcribes the utterance, the term is more likely to be present.
- $C(\Pr(w_i))$: how close the probability of word instance w_i is to an ideal confidence. This factor serves to balance the $\Pr(w_i)$ term, encouraging the selection of terms that might be incorrect, possibly due to substitution with Out-Of-Vocabulary (OOV) terms. We define this factor as $\exp(\frac{-(c-\Pr(w_i))^2}{\beta})$. The ideal confidence, c, used for these experiments was 0.7, which causes us to choose instances of rare words that are likely but are still uncertain and thus valuable for training. β is a scaling factor that we set to $\frac{1}{15}$.
- $\rho(w)$: the prior probability of the word w in the candidate set of the utterances, estimated from the consensus output. This is not the prior probability of the term in the seed training data, but rather the expected prior in the in-domain data. This factor causes selection of terms that are more likely to occur in the in-domain data and are therefore of more interest to users, even though these terms were rare in the out-of-domain seed training.

The score $s(u)$ for each utterance u in Equation 1 is a linear combination of the factors, weighted by the importance and summed over all of the instances of rare terms in the utterance[2].

$$\mathrm{s}(u) = \sum_{w_i \in \mathrm{rare}(u)} \frac{I(w)}{D(u)} (\alpha_1 \Pr(w_i) + \alpha_2 C(\Pr(w_i)) + \alpha_3 \rho(w)) \quad (1)$$

The score for the utterance is normalized by the duration of the utterance $D(u)$, so that the score reflects the expected benefit per second of audio transcribed. The utterances with the highest scores are selected for transcription using a greedy search. After each utterance is selected, $T(w)$ is updated with the new counts from the transcribed utterance so that an updated $I(w)$ can be computed[3]. Because $I(w)$ is biased towards unseen words, this causes the selection strategy to diversify its lexicon rather than focusing on a small number of rare words.

3. EXPERIMENTAL SETUP

3.1. Corpora

We have evaluated our active learning selection algorithm on Spanish CTS. The seed recognizer was trained on 20 hours of transcribed speech selected from the Spanish CallHome, CallFriend, and Ricardo corpora. The target domain for this task was Fisher Spanish; accordingly, we reserved 6 hours of the Fisher corpus for an evaluation set, and the remaining 178

[1] See Section 3 for more details on corpora used in Table 1. 10 hours were randomly selected from an active learning candidate set and remaining 168 hours were used for testing.

[2] $\alpha_1 = 0.3, \alpha_2 = 0.3, \alpha_3 = 0.4$. A slightly higher weight is given to α_3 since the prior is a very low value for rare terms.

[3] The selection can also be run without immediate transcription feedback by using the expected count of the words instances in the selected utterance, $\Pr(w_i)$, in place of the true counts to update $T(w)$. Using only the expected counts without ever adjusting for the actual counts reduces the effectiveness of the active learning strategy.

hours were used as the untranscribed candidate set for active learning simulation.

3.2. LVCSR and KWS

Acoustic models for all systems were built using perceptual linear predictive cepstral features for every 10ms frames, followed by concatenation of 9 feature frames and then projection to 46 dimensional features. Vocal tract length normalization, speaker adaptive training and speaker adaptation were all applied in a similar manner as in the system described in [8]. GMM-HMM triphone and quinphone models were discriminatively trained with the minimum phone error (MPE) criterion.

KWS was performed by computing the posterior probability that any word ends at a particular time in the output lattice for an utterance [9]. A ranked list for a given query is returned, with hypothesized instances sorted by posterior scores. Unless otherwise stated, AUC was measured on all words in the evaluation set after discarding stopwords.

3.3. Active learning

Since the motivation behind the proposed Δ-AUC active learning strategy is to find new instances of rare and unseen terms, we began by expanding the initial decoding dictionary of approximately 13,000 terms to approximately 250,000 terms. This was achieved by gathering term lists from the Spanish Gigaword text corpus as well as scraping the Spanish Voice of America news website for new terms [10]. The experiments reported here do not use these external sources for additional language model training, but we observed similar results when doing so.

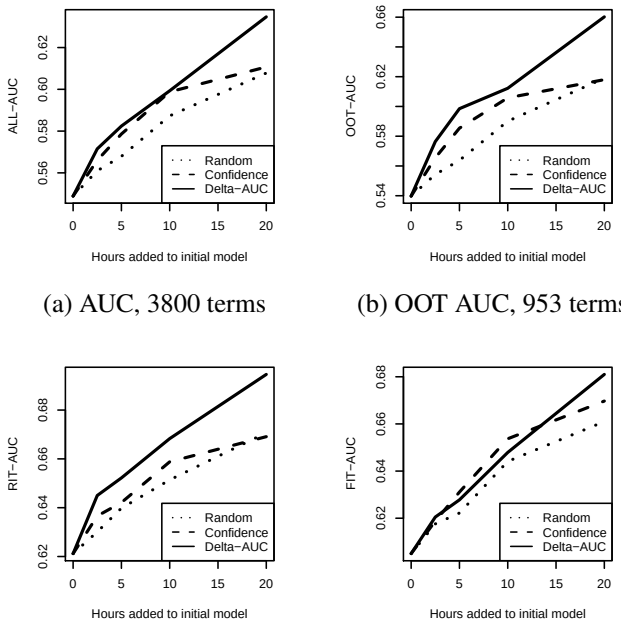
We compared our active learning algorithm to two baseline approaches: random and confidence-based selection. There are several variations on confidence based selection, e.g. [2, 3]. For these experiments, we sorted the utterances by the average posterior probability of the words in the one best using the expanded dictionary, and then selected the least confident among them. Utterances at the top of this confidence-sorted list can be noisy and not helpful for training, so we found that we got better results by excluding 40%[4] of the top of the list, thus removing the most noisy utterances.

For each active learning strategy, we selected 2.5 hours, 5 hours, 10 hours, and 20 hours from the candidate set. We combined each of these new sets with the seed training and retrained both the acoustic and language models. The decoding dictionary for each model was composed of any terms in the model's training transcripts plus the additional terms described above.

[4]Various percentages were tested, and we selected the percentage that performed the best for KWS and WER.

4. RESULTS

Figure 1a compares the AUC for the three selection approaches when 2.5, 5, 10, and 20 hours of speech are selected. The Δ-AUC selection strategy consistently outperformed random sampling at all selection points by anywhere from 1.05% to 2.69% absolute; in fact, it required as little as 40% and at most 60% of the transcriptions needed by random sampling to achieve the same AUC. In addition, the Δ-AUC method was superior to confidence-based selection, with gains of 2.4% absolute as more speech is required.



(a) AUC, 3800 terms (b) OOT AUC, 953 terms

(c) RIT AUC, 1017 terms (d) FIT AUC, 1553 terms

Fig. 1. AUC among active learning approaches for various query sets. OOV terms (total of 277) not shown.

If we examine the AUC of various word subsets for Δ-AUC, we find large improvements for the terms targeted during active learning selection (Figures 1b and 1c) over both baseline approaches. We see smaller gains for the FIT terms (Figure 1d) compared to random selection and no change compared to confidence-based selection. Of the 1.20% absolute gain in overall AUC of the 10 hour Δ-AUC selection over the 10 hour random sampling, the OOT terms were responsible for 46.19%, the RIT terms for 37.69%, and the FIT terms for 12.94%, with the remaining gain due to OOV terms that co-occurred or were confused with rare terms.

As users of audio KWS systems often search for new words, the gains for OOT terms are particularly of interest. Of the approximately 237,000 OOT terms targeted during active learning selection, 953 occurred in the test set; we observed that the Δ-AUC selection strategy resulted in large gains on

Method	Hours	AUC	OOT-WER (%)
Random	5	0.5638	79.79
Confidence	5	0.5854	79.03
Δ-AUC	5	0.5985	75.13
Random	10	0.5901	78.11

Table 2. AUC and word error rate (%) among active learning approaches on OOT terms in test set (953 terms). Selected hours are added to the 20 hour seed model transcripts.

these OOT terms (Table 2). Comparing the confidence-based method to the random baseline, we observed a 2.16% absolute improvement in AUC and a 0.76% absolute improvement in WER when selecting the same amount of speech. The Δ-AUC selection outperformed the confidence-based method, with a 3.47% absolute improvement in AUC and a 4.66% absolute improvement in WER when compared to the random baseline. Δ-AUC, using 50% less data, was almost 1% absolute better in AUC than the random sampling approach in the last row.

Since the Δ-AUC strategy focuses on improving AUC, particularly on OOT and RIT terms, the WER measured on the FIT terms was close to equal for all three selection strategies, and the overall WER (dominated by the FIT terms) was roughly equivalent between the three strategies as well. We believe that for many KWS applications, users are more likely to search for rare terms (such as the name of a person or location), so improving OOT and RIT terms is likely more beneficial than improving more common terms.

5. REGRESSION ANALYSIS

To support the choice of the three factors in the reward ranking function described in Section 2, we performed a regression analysis to determine the importance of each individual factor in relation to per-term Δ-AUC. The simple linear regressions described in this section examined the per-term differences in AUC when using two recognizers: 1) the seed recognizer, and 2) the seed recognizer plus the 10 hours of randomly selected speech from Section 3. The regressions were trained on the entire active learning candidate set minus the 10 hours used for the second model, for a total of approximately 168 hours. The per-term features used as independent variables for the regression were extracted from the 10 hour random set, as this set directly contributes to the Δ-AUC from model 1 to model 2.

Because per-term Δ-AUCs can be very noisy and thus hard to predict – for instance, many words occur only once in a test set – we binned the data based on prior and fit a simple linear regression of the following form to the binned data points

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (2)$$

where y_i is the dependent variable for a single data point (Δ-AUC in this case), x_i is the independent variable, β_0 and β_1 are the regression coefficients, and ϵ_i is the error.

Independent Var.	$\mathbf{r^2}$	$\mathbf{r}$
$\rho(w)$	0.4747	0.6890
$\Pr(w_i)$	0.4484	0.6696
$C(\Pr(w_i), c = 0.7)$	0.4012	0.6334
$C(\Pr(w_i), c = 0.3)$	0.1115	0.3339

Table 3. Coefficient of determination (r^2) and correlation coefficient (r) for simple linear regression with Δ-AUC as dependent variable. β_1 is statistically different than 0 at a 1% significance level.

The first three rows of Table 3 show the coefficient of determination, r^2, as well as r, Pearson's correlation coefficient, for a simple linear regression of each of the three independent variables used in the reward ranking function. The coefficient of determination can be interpreted as the proportion of the variance in the dependent variable, Δ-AUC, explained by the regression [11]. For each of the factors used in the reward ranking function, we see that close to half of the variance in the data can be explained by the variable in question, so they are good candidates for predicting Δ-AUC. On the other hand, when we considered how close instances are to a confidence of 0.3 rather than 0.7 in the last row of Table 3 (a factor that seems intuitively to be less helpful), we found that only 11% of the variance in the data can be explained by this factor and the correlation coefficient is about half of the other factors.

6. CONCLUSIONS

In this work, we have introduced the framework of importance sampling as a basis for active learning. We have demonstrated that the Δ-AUC active learning methodology outperforms confidence-based selection and requires significantly less transcription (40-60%) compared to random sampling of utterances. On terms not seen in the original seed model training, Δ-AUC-based selection achieved a 3.47% better AUC and 4.66% reduction in WER.

In the future, we can explore the use of the regression analysis described in Section 5 as a basis for creating a refined reward ranking function by learning weights for the ranking function factors, effectively using the predicted Δ-AUC directly as the reward. With this regression analysis framework, we can also attempt to incorporate additional predictive features such as length and confusability.

7. REFERENCES

[1] Christian Robert and George Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics. Springer New York, 2013.

[2] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech and Language - Special Issue on Emergent Artificial Intelligence Approaches for Pattern Recognition in Speech and Language Processing*, 2009.

[3] Giuseppe Riccardi and Dilek Hakkani-Tur, "Active learning: theory and applications to automatic speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 504–511, July 2005.

[4] Teresa M. Kamm, *Active Learning for Acoustic Speech Recognition Modeling*, Ph.D. thesis, The Johns Hopkins University, 2004, AAI3130709.

[5] Teresa M. Kamm and Gerard G. L. Meyer, "Word-selective training for speech recognition," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, Nov 2003, pp. 55–60.

[6] Dilek Hakkani-Tur, Giuseppe Riccardi, and Allen Gorin, "Active learning for automatic speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, May 2002, vol. 4, pp. IV–3904–IV–3907.

[7] Thiago Fraga-Silva, Jean-Luc Gauvain, Lori Lamel, Antoine Laurent, Viet-Bac Le, and Abdel Messaoudi, "Active Learning based data selection for limited resource STT and KWS," in *Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, September 2015, pp. 3159–3163.

[8] Spyros Matsoukas and Richard Schwartz, "Improved speaker adaptation using speaker dependent feature projections," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, Nov 2003, pp. 273–278.

[9] Ivan Bulyko, Owen Kimball, Man-Hung Siu, José Herrero, and Daniel Blum, "Detection of unseen words in conversational Mandarin," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 5181–5184.

[10] "Voz de américa," `http://www.voanoticias.com`, 2015, [Online; accessed 24-September-2015].

[11] William C. Navidi, *Statistics for Engineers and Scientists*, McGraw-Hill Higher Education, 2007.