

# Robustness Aspects of Active Learning for Acoustic Modeling

Teresa M. Kamm and Gerard G. L. Meyer

Center for Language and Speech Processing  
Department of Electrical and Computer Engineering  
The Johns Hopkins University, Baltimore, Maryland, USA  
tkamm@acm.org, gglmeyer@jhu.edu

## Abstract

We previously proposed [1] an iterative word-selective training method to cost-effectively utilize data preparation resources without compromising system performance. We continue this work and investigate the robustness of our active learning approach with respect to the starting conditions and further propose a stopping criterion that supports our objective to make effective use of transcription effort while minimizing system error. In particular, we demonstrate robustness to seven initial conditions, showing that we can select around 20 hours of training data and achieve a range of error rates between 8.6% and 9.0%, compared to an error rate of 10% when using all 50 hours of the training set. Additionally, we give empirical evidence that our proposed stopping criterion is in general a good predictor of when the minimum error rate is achieved, demonstrated for each of the initial conditions.

## 1. Introduction

We are working to find a better way to select acoustic training data for speech recognition modeling, with the goal of making more effective use of transcription resources without compromising system performance. We assume that our system configuration is fixed and therefore our only opportunity to achieve a low error system is by selective use of training data. We also assume that the data is not already transcribed, so we must design a data selection criterion that does not rely on knowing the true transcription.

Conventional wisdom tells us that in general, "more data is better", as depicted in Figure 1a, and if sufficient resources are available to acquire more data, it is a clear path to lower error rate. However, if resource constraints only allow for transcription of data of size  $|S|$ , for instance, it is clearly preferred to choose data set "o" rather than data set "x" as shown in Figure 1a.

It is possible that an alternate scenario exists, that being "more data can increase the error", as depicted in Figure 1b. In this case, there is a subset "o" of size  $|T|$  of the training data that gives rise to a system with a lower error rate than using all the data. While this scenario presents a clear opportunity to reduce the error rate by using less data, it's not clear that the presence of this scenario could even be detected, as suggested by the black line in Figure 1b that shows a hypothetical progression of acquiring data.

For either scenario, the goal is clear: we want to find the best set of training data for any given size, thereby tracking the minimum error rate curve (the bottom of the gray area) of the hypothetical sampling scenarios. Clearly, we can never know the true sampling scenario or the optimal subset "o" for a given training size simply because exhaustive search is impractical. As an alternative, we propose an iterative

approach to find a reasonable subset of training data with both minimization of data usage and error in mind.

Our approach, as reported in [1], is to start with some small initial set of randomly chosen training data and then selectively incorporate additional training in small steps, with the goal of making effective use of transcription resources while minimizing system error. We showed that we could select 30% of a given training set of an alphadigit corpus and reduce the error rate from 10.3% to 9.3%, compared to using all of the given training data. Additionally, we showed a similar ability to select and utilize a 60% subset of training data from the Wall Street Journal corpus without compromising error rate. Both of these results were achieved with either no or minimal knowledge of the true transcription. Complete details are given in [2].

As with any iterative algorithm, it is important to understand the degree to which the outcome depends on the initial condition. Clearly, it is most desirable to obtain a result that does not depend on the starting point at all. To that end, we present a detailed study of the dependency of the selection algorithm on the initial conditions for a simple alphadigit corpus, concluding that the algorithm is indeed robust to initial conditions.

It is also desirable to define a stopping rule that supports our objective to make effective use of transcription effort while minimizing system error. We propose as a stopping criterion the minimization of the training likelihood, computed as an average across all non-silent frames, and demonstrate a strong correlation between the minimum likelihood point and the stopping point that minimizes error for the given test set. This stopping criterion has the distinct advantage that it is not a function of any test set. Additionally,

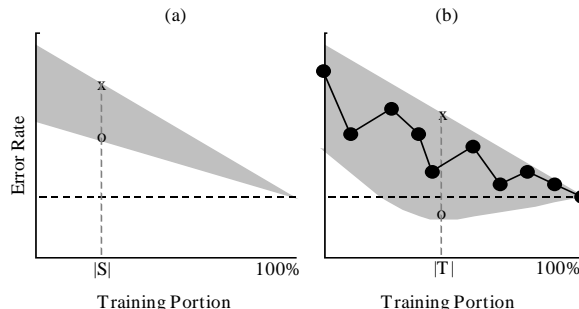


Figure 1: Sampling scenarios (gray areas) showing the possible relationship between error rate and all subsets of a given training size. In each scenario a particular instance is called out and depicted as "x" and "o". In (b), a hypothetical data acquisition curve is shown in black, suggesting one possible path.

observation of the minimum likelihood criterion offers a way to assess the potential of any *future candidate training data*, requiring only a minimum of transcription effort to make the assessment.

The organization of this paper is as follows. Section 2 reviews the active learning paradigm and introduces a new selection criterion that does not require normalization as well as a stopping criterion. Section 3 discusses the OGI alphas digit corpus and baseline system error rates. Section 4 presents the new results achieved that address the robustness issues and demonstrate viability of the stopping criterion. Finally, a summary and discussion are presented in section 5.

## 2. Active Learning Paradigm

This section reviews the iterative selection training procedure and selection criterion that facilitates active learning for speech recognition and introduces the stopping criterion.

### 2.1. Selective Training

We start with the well-known maximum likelihood criterion to estimate a model  $\lambda$ :

$$\lambda = \arg \max_{\lambda'} P(\mathbf{O}|\mathbf{W}; \lambda') \quad (1)$$

where  $\mathbf{O}$  is one long string of observations and  $\mathbf{W}$  is the associated class labels (transcription). We partition the observations  $\mathbf{O}$  into utterances  $\mathbf{U} = (u_1 u_2 \dots u_N)$  and likewise partition the class labels  $\mathbf{W}$  such that  $W_u$  is the associated class labels for each utterance  $u \in \mathbf{U}$ . This partitioning allows us to rewrite the training criterion in terms of a product:

$$\lambda = \arg \max_{\lambda'} \prod_{u \in \mathbf{U}} P(u|W_u; \lambda') \quad (2)$$

thereby facilitating selective training simply by specifying the members of the set  $\mathbf{U}$ . This is demonstrated in Figure 2 where we present the generalized iterative selection training procedure.

- Given a set of candidate training data  $\mathbf{U}$ ,
1. Select the initial training set  $\mathbf{U}_0$  from  $\mathbf{U}$  and let  $i = 0$ .
  2. Obtain labels for the training set  $\mathbf{U}_i$ .
  3. Train a model,  $\lambda_i = \arg \max_{\lambda} \prod_{u \in \mathbf{U}_i} P(u|W_u; \lambda)$ .
  4. Accumulate additional training:  $\mathbf{U}_{i+1} = \mathbf{U}_i \cup \mathbf{U}'$ , where  $\mathbf{U}' = \{u_i \in (\mathbf{U} - \mathbf{U}_i) \text{ and } u_i \text{ satisfies some constraint}\}$ .
  5. Evaluate the stopping criterion: Stop or go to step 2.

Figure 2: Generalized Iterative Selective Training Procedure

### 2.2. Active Selection Criterion for Acoustic Training

Clearly, how we choose the additional training  $\mathbf{U}'$  is critical to achieve our goal of reducing data preparation burden without compromising system performance. Our most successful criterion to date, which is a word-selective criterion, was first presented in [2] and is as follows.

We apply the model,  $\lambda_i$ , to hypothesize the most likely words  $\hat{W}$  for a given candidate training utterance  $u$ :

$$\hat{W} = \arg \max_w P(W)P(u|W; \lambda_i). \quad (3)$$

As side information, we also obtain the hypothesized word segmentation  $S_{\hat{W}}$  by observing the alignment of the hypothesized words  $\hat{W}$  onto the utterance  $u$ . Then, for each hypothesized word  $\hat{w}$  in  $\hat{W}$  and the associated word segment  $S_{\hat{w}}$ , we find the two best matching words and generate the *boundary distance score*, which we define as:

$$P(\hat{w}_1|S_{\hat{w}}; \lambda_i) = \frac{P(S_{\hat{w}}|\hat{w}_1; \lambda_i)}{P(S_{\hat{w}}|\hat{w}_1; \lambda_i) + P(S_{\hat{w}}|\hat{w}_2; \lambda_i)} \quad (4)$$

where  $\hat{w}_1$  is the best matching word for segment  $S_{\hat{w}}$ , and  $\hat{w}_2$  is the next best matching word. The boundary distance score is related to the word posterior probability that was previously used as a selection criterion [1], but requires no normalization. The value ranges from 0.5 to 1.0 and a low value indicates closer to the decision boundary and a high value indicate far from the boundary.

### 2.3. Stopping Criterion

One obvious stopping criterion is to observe the error rate of the models on a held out test set and decide to stop if the error rate is not improving. Alternately, if the active selection criterion is used to acquire new data rather than to refine an available data set, then a pragmatic stopping criterion is to stop either when the transcription resource is exhausted or when error rates are not improving, whichever comes first.

While these are certainly task-focused criterion, it is our desire to find a stopping rule that is more a reflection of the model's ability to incorporate additional data. To that end, we observe the likelihood of the training data:

$$\max_{\lambda} \prod_{u \in \mathbf{U}_i} P(u|W_u; \lambda) \quad (5)$$

in the form of the average per frame likelihood *with silence removed* as observed from a forced alignment of the selected training data  $\mathbf{U}_i$  with the model  $\lambda_i$ . We conjecture that when observing the training likelihood at each selection step, we initially will see the likelihood decrease, reflecting the difficulty represented in the additional training data. Eventually, we will reach a point of minimum likelihood and incorporating any additional training data will cause the likelihood to increase. We hypothesize that this increase in training likelihood is an indication that the model has already characterized all the variability possible in its current configuration *for this given training set*.

We further speculate that a future candidate training set can be quickly assessed for selection potential simply by carrying out a single selection step on that data, starting with the minimum likelihood model already obtained from previous data, and observing the training likelihood. If the training likelihood goes down with the introduction of new data, then there is the potential to further improve the model with this new candidate data. Alternately, if the likelihood goes up, then there is no reason to continue study of this data.

### 3. The OGI Alphadigit Corpus

The OGI alphadigit corpus [3] is comprised of over 3000 subjects speaking strings of 6 alphadigits over the telephone, for a total of nearly 75 hours of speech. The alphadigits are the English letters "A" through "Z" and digits "0" through "9". Each speaker was prompted to speak either 19 or 29 sets of 6 alphadigits.

Starting from the ISIP defined training and test partitions [4], we selected 46730 (50 hours) of the 51545 training sentences and 3112 (3.26 hours) of the 3329 evaluation test sentences such that the transcription of the selected sentences matches one of the possible prompts given in the prompt list [3]. This was an inexpensive way to remove possible transcription errors from the training and test sets.

#### 3.1. System Description and Baseline Error Rates

Our system consists of 36 word models, one for each alphadigit, plus a silence and short pause model. The models are trained according to the procedures given in the HTK documentation [5].

Each word model uses a standard left-right topology with a re-entrant transition allowed in each state. The number of states for each model is duration dependent and is equal to one-half of the mean number of frames of each word. The word durations are determined by first training a system with each word model having 10 states. Then the average duration for each word is observed from the durations discovered via a forced alignment of the 10-state models with the training data. The resulting system has a total of 825 states.

The features are 12 mel-frequency cepstral coefficients plus energy, deltas, and double deltas to make a feature vector of length 39. From these features, 12 gaussians are estimated per feature per state.

An equal probability network is used to drive recognition, which is defined to be: optional silence, followed by one or more alphadigits with optional inter-alphadigit silence, followed by optional silence.

Several systems were built to demonstrate baseline performance for varying amounts of training data. Each baseline subset of the training data has a fair representation of

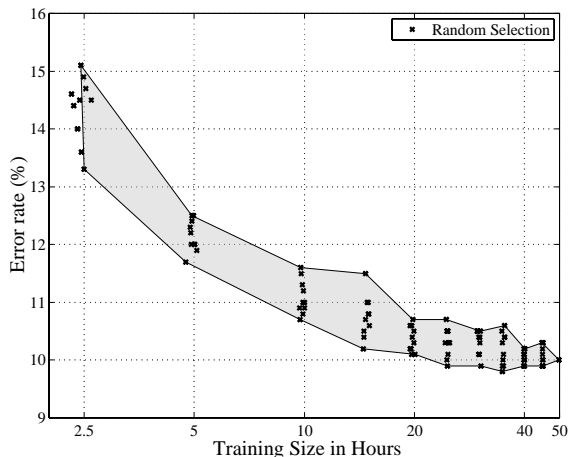


Figure 3: Relationship between error rate and training size for various random subsets of the available training set for the OGI alphadigit corpus.

male and female speakers as well as examples of each alphadigit. Selection was done on a per speaker basis, meaning that once a speaker was selected all of the data contributed by that speaker was included in the set. This implies that the smaller sets include fewer speakers than the larger sets. Additionally, the larger sets are made up of different combinations of the smaller sets. This is meant to mimic the way that corpora are typically collected, where we acquire a large amount of speech per speaker from a small set of speakers, in contrast to acquiring a small amount of speech per speaker from a large set. The baseline error rates for various training sizes are shown in Figure 3. The minimum and maximum error rates across the training sizes are plotted with a line and serve to highlight the wide range of error rates possible with a random sample of training data. An error rate of 10% is achieved when using all of the training data.

### 4. Robustness Observations

In this section, we describe experiments and associated results that explore the robustness issues presented in Section 1.

#### 4.1. Initial Conditions

To investigate the sensitivity of the iterative selection algorithm to initial conditions, we selected 7 mutually exclusive, 2.5 hour, subsets of the training data. Each subset was made up of approximately 110 speakers, roughly half male and half female. Additionally, each set of alphadigits within each subset contained examples from both genders.

We ran seven separate selection experiments, each starting with one of the seven mutually exclusive subsets. The results of these experiments are summarized in Figure 4 and are depicted by the darker gray area that shows the range of minimum and maximum error rates achieved across the seven experiments. In particular, after selecting an additional 3.5 or more hours of data (at least 6 hours of training), all seven selection experiments outperform random selection of the

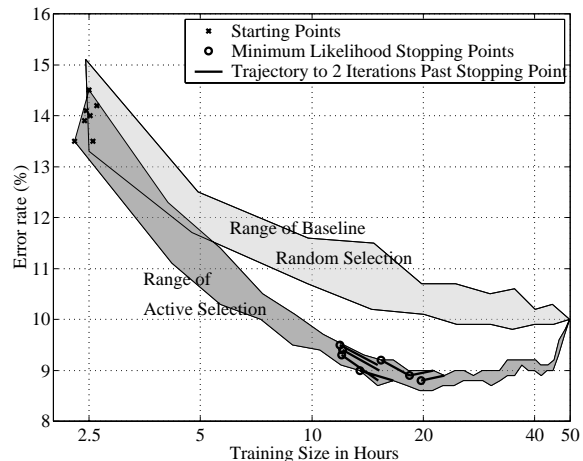


Figure 4: Range of error rates obtained by active selection from 7 different initial conditions compared to baseline random selection. Also shown is the minimum likelihood stopping point generated for each of the 7 initial conditions as well as an alternate stopping point two iterations beyond.

same size. Additionally, the range of error rates become more focused as additional data is selected, and in general is minimized after selecting an additional 17.5 hours of data (totaling 20 hours of training), with the error rates ranging from 8.6% to 9%. This is a significant system improvement over the 10% error rate achieved when using all of the training data and represents a substantial 60% savings in data preparation if the selection algorithm were in fact used to actively acquire data.

## 4.2. Stopping Criterion

We investigated the viability of using the training likelihood as a stopping criterion by observing this likelihood for each sampling experiment discussed in the previous section. One particular example that is representative of the seven selection experiments is given in Figure 5. This example clearly gives indication that the first minimum likelihood point (achieved after selecting 15 hours of data), or one or two iterations after, is a reasonable stopping criterion. The behavior of the training likelihood after selecting around 40 hours of data is ignored, as we attribute its shape to finite data effects.

Results across the seven experiments are summarized in Figure 4, showing the training size and error rate (o's) achieved at the minimum likelihood point as well as the trajectory indicating an alternate stopping point two iterations later. In general, the minimum likelihood point is a good predictor of the "knee" of the error rate curve and clearly only one or two more iterations are needed to get to the "best" error rate for each selection experiment on this test set.

## 5. Summary

In this paper we introduced an improved selection criterion that it does not require normalization. We demonstrated its effectiveness as a selector by observing the outcome of several selection experiments, each starting with a different initial condition.

We presented evidence that our iterative selection

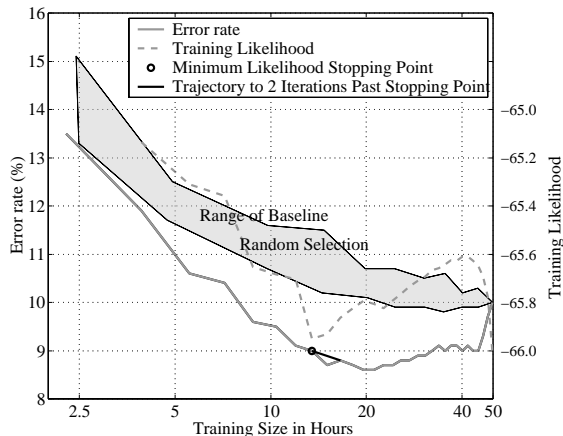


Figure 5: Observation of training likelihood for one selection experiment. For the purposes of selecting a minimum likelihood point, the first minimizing point is selected, ignoring the behavior of the training likelihood after selecting around 40 hours of data.

algorithm is robust to the initial conditions, by investigating several initial training sets that have a fair representation of both male and female speakers and class examples. We showed that after two iterations all selection experiments, regardless of the starting point, are able to continually beat the best error rate predicated by random selection of a similar amount of training data. Additionally, we showed that as selection progresses, the range of error rates achieved from difference starting points becomes more focused, culminating to a minimum error rate in the range of 8.6% to 9.0% after selecting 20 hours of training. Future work will investigate how sensitive the selection algorithm is to the careful split between male and females speakers enforced for these experiments, with the goal of demonstrating that the algorithm can adjust to any initial training set so long as it contains at least a few examples of each class.

We also introduced a stopping criterion, that being the point when training likelihood is minimized. We present empirical evidence that this minimum point is a good predictor of the "knee" of the error rate curve and observe for multiple selection experiments that the "best" error rate on a particular test set is achieved within 1 or 2 iterations past the minimum training likelihood point.

A desirable side benefit of using the minimum training likelihood as a stopping criterion is that it is not tied to any particular test set. Future work will address the robustness of this stopping criterion across test sets.

Observation of the path to minimum training likelihood appears to track the model's ability to make use of the information available in the training data, as reflected in the also improving error rate. Once the minimum training likelihood is achieved, the error rate curve starts to flatten, and we observe minimal or no improvement thereafter by introducing additional data. This flatness is maintained over a wide range of training sizes, until the addition of the last 5% of training data resulting in an error rate that is suddenly severely degraded. This phenomenon is certainly worthy of future investigation. Additionally, future work will investigate using the training likelihood to assess new candidate data, with the goal of making a quick assessment of the potential offered by this new data.

## 6. References

- [1] T. M. Kamm and G. G. L. Meyer, "Word-selective training for speech recognition," In *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, 2003.
- [2] T. M. Kamm, "Active Learning for Acoustic Speech Recognition Modeling," Ph.D., The Johns Hopkins University, Baltimore, 2004.
- [3] M. Noel, (1997), "Alphadigits," Center for Spoken Lang. Understand., Oregon Graduate Inst. Sci. Technol., Portland, OR, [Online] Available: <http://www.cse.ogi.edu/CSLU/corpora/alphadigit>
- [4] J. Hamaker, A. Ganapathiraju, and J. Picone, (1997), "A proposal for a standard partitioning of the OGI AlphaDigit corpus," Inst. Signal Inform. Process., Mississippi State Univ., [Online] Available: [http://www.isip.msstate.edu/projects/speech/software/asr/research/syllable/alphadigits/d ata/ogi\\_alphadigits/eval\\_trans.txt](http://www.isip.msstate.edu/projects/speech/software/asr/research/syllable/alphadigits/d ata/ogi_alphadigits/eval_trans.txt)
- [5] S. Young, et al, *The HTK Book, Version 3.2*: Cambridge University Engineering Department, 2002.