DATA SELECTION FOR SPEECH RECOGNITION

Yi Wu, Rong Zhang, Alexander Rudnicky

Language Technology Institute, Carnegie Mellon University

ABSTRACT

This paper presents a strategy for efficiently selecting informative data from large corpora of transcribed speech. We propose to choose data uniformly according to the distribution of some target speech unit (phoneme, word, character, etc). In our experiment, in contrast to the common belief that "there is no data like more data", we found it possible to select a highly informative subset of data that produces recognition performance comparable to a system that makes use of a much larger amount of data. At the same time, our selection process is efficient and fast.

Index Terms— data selection, maximum entropy, speech recognition, acoustic modeling

1. INTRODUCTION

In the speech recognition community, there is a long-standing belief that "there is no data like more data". Following this idea, a number of efforts have been undertaken to label large amounts of data in order to improve recognition performance, as a result significant amounts of transcribed data has become available for training use. For example, the first year of the GALE (Global Autonomous Language Exploitation) project has produced more than 800 hours of well transcribed (Mandarin) data. Additional efforts have been directed to using untranscribed data as a source of training material. However, it is commonly observed that as we continue adding data to model training, the improvement in performance becomes smaller and smaller. This suggests a large degree of redundancy within large speech corpora. A natural question to ask is how to decide which data is more important than the other and this question led us to examine data selection.

The main advantage of successful data selection is reducing training time. As we increase the amount of training data, the time needed for training also increases. Moreover, modern systems use additional steps such as LDA, MLLT and VTLN that require multi-pass training, further increasing training time. Acoustic training becomes extremely time consuming and perhaps limits progress in acoustic modeling. Another motivation is research [9] in the rapid deployment of ASR systems for new Languages where it's been found important to have criteria for choosing the most informative data given the expense of acquiring data.

Facing above problems and challenges, we feel it is important to look at the problem of data selection in speech training. In this paper, we will address the problem of selecting a small amount of highly informative data from a much larger training corpus. We would expect a useful selection technique to meet two criteria:

- Good Performance: The performance of the system trained on the selected data should be comparable with the performance using a much larger corpus.
- Cheap and Rapid Selection: Since one of our motivations for doing data selection is to save training time, the selection method needs to itself be efficient.

We address the specific issues that come up in the training of acoustic models for Mandarin recognition. The GALE program has provided access to over 800 hours of speech data (broadcast news and broadcast conversation). But to make better use of these data for research we would like to reduce the data to a manageable amount. For example, training a baseline system using the entire corpus will take several days on our servers, impeding progress in research. Motivated by this practical problem as well as curiosity about redundancies in training material, we developed a data selection strategy that is guided by the maximum entropy principle in choosing those speech utterances that contribute to a uniform distribution across speech units such as words, characters, and phonemes. We might also consider this as a criterion that minimizes, for a given size of training corpus, under-sampling for units of interest. Using this method, we found ourselves able to use a small portion (150 hours) of the training corpus to get performance comparable to a system using the entire corpus.

In the following sections, we will first describe related work in data selection for speech recognition. In section 3, we will describe our algorithm. Section 4 describes our experiments and results.

2. RELATED WORK

There is already an extensive literature related to the problem of speech data selection, although from a perspective different from the one taken in this paper.

In [1] and [2], the authors study the problem of transcribed data selection for digit data. The authors primarily investigate techniques such as Principal Component Analysis and clustering. The authors also suggest the importance of utterance length in selection. [3, 4, 5, 6], describe interesting research concerning the use of active learning related to data selection. These studies focus on how to select the most informative untranscribed data for a human to label in further training. The main method is to select those data with low confidence. There is also some research on unsupervised data selection [14]. The main approach is to decode the data first and then select those data with high-confidence decodings for inclusion in training. Some researchers [7] have also studied the problem of "lightly supervised learning". The challenge there is to avoid those data that are poorly transcribed. The main suggestion has been to choose those data that have good match with the corresponding decoding result.

Nevertheless, none of the above approaches take the distribution of data into direct consideration. Adding new data that is redundant with the existing training material actually does not appear to help. Therefore we propose to choose data uniformly using relevant dimensions of speech. We believe that this allows us to incrementally add the most informative data on each cycle and provides a robust estimation of the parameters for modeling those speech units.

3. CURRENT APPROACH

Our selection algorithm is currently applicable to corpora of transcribed speech data. Current selection only uses the transcription which will make our selection fast in practice. As we mentioned, we propose to select the subset of data which have an uniform distribution across different speech unit like word and phoneme. Practically, it is not always possible to get the exact uniform distribution. We will choose the sample set that is close to uniform distribution.

3.1. Sample Uniformly

We will now define our uniformly sampling proposal formally. Suppose we have a utterance set U. We have the transcription $X_1...X_n$ for each utterance u_i in U. X_i can be either the word, character or phoneme. Then our uniform sampling proposal is that we should sample a subset of U which have an approximately uniform empirical distribution of X_i.

To understand our uniform sampling idea, imagine following simple example: suppose we have k different classes and each class is generated by some distribution, say gaussian distribution with density function $f_k(\mu_k, \sigma_k)$ with prior π_k. And we (somehow) already have prior π_k for each class beforehand. In training, we will use the MLE estimator for the model parameter, namely use sample mean and sample variance to estimate the mean and variance for each gaussian. And we are given access to a total of N samples from the k classes while we have the choice to determine how many sample we want from each class. When a new example comes according to the generative model, we will make a prediction on which class it is from by our trained model. The question is: how much should we sample from each class to train a model with minimum classification error?

Our proposal is to sample from each class equally. Our claim is based on following statement. Ideally, if we have the true parameter u_i and σ_i of f, we would use the optimal Bayes classifier[11]: for a new example x, we will choose the class label i that maximize the posterior as follows:

$$i = arg\max_i \pi_i f_i(x) = arg\max_i(\log(\pi_i) + \log(f_i(x)))$$

Notice $\log(\pi_i)$ is some fixed number we already know. In order to achieve the optimal Bayes error, we need to make our MLE estimation of $\hat{f}_i(x)$ accurate. Suppose we have error bound $e_i(x)$ in estimating $f_i(x)$ using $\hat{f}_i(x)$:

$$|f_i(x) - \hat{f}_i(x)| \leq e_i(x).$$

Then our true error will be bounded by be $\max e_i(x)$ plus the optimal Bayes error. Since we have totally no idea of what $f_i(x)$ is before selection, sampling uniformly would give a small error bound $e_i(x)$ for each class.

Notice that the above process is similar to the true speech recognition system. In the classic formula for speech recognition, we are using the following Bayes formula [12]

$$arg\max_w P(W|X) = arg\max P(W) * P(X|W).$$

Here the prior $P(W)$ is estimated by language model beforehand and $P(X|W)$ is given by acoustic model. Although the mechanism of a real ASR system is more complex, the essential thing here is that we have a independent estimation of the prior P(W) by the language model, so the optimal sampling for acoustic model training is to sample each class W uniformly to minimize the error of estimating $P(X|W)$. Notice that the language model defined at word level can also be viewed as prior on phoneme or character because it contain information which word (hence which phoneme, character) will happen more often. Therefore, we will explore the idea of a sample uniform distribution on class defined at phoneme, character and word.

3.2. Maximum Entropy Principle

Usually we cannot get a truly uniform distribution. To measure how close a distribution is to the uniform distribution, we can use the notion of entropy. Entropy is defined as the uncertainty of random variables. For discrete random variable X, if it can take the possible value from $\{X_1, X_2, ..X_n\}$, then its entropy is defined as

$$\begin{aligned} H(x) &= \sum p(X_i) \log_2 1/p(X_i) \\ &= -\sum p(X_i) \log_2 p(X_i) \end{aligned}$$

Also, for probability distributions X and Y of a discrete random variable, the Kullback-Leibler divergence of X from Y is defined to be

$$\sum \log p(X_i) \frac{p(X_i)}{p(Y_i)}$$

It is a natural distance measure of two distributions. The maximum entropy principle states that the entropy is maximized when the distribution on X is uniform. Essentially, notice that the entropy of X is actually the Kullback-Leibler divergence between X and the uniform distribution on $X_1, X_2...X_n$ with some additive constant. So it is a good measurement of the "uniformness" of a distribution. Overall, we formulate our entropy based selection principle as follows: given the size of the data we want to select, we would prefer to choose the subset that has the largest entropy.

3.3. A greedy algorithm in search

One technical difficulty is the computational intractability of finding the global optimal subset that maximize the entropy. To make the computation efficient, we use a greedy search algorithm as follows:

Algorithm 1 Greedy Search

```
for all utterance u_i do
  if Adding u_i increase entropy by some threshold e then
    add u_i
  end if
end for
```

We can use the threshold e to control the amount of data we want to select. Notice this algorithm can be executed very fast since its only needs to check the transcription of each utterance once.

3.4. Choice of Classes and their combination

As we mentioned, we will use word, character and phoneme as classes for the distributions we will consider. A natural question to ask is how to combine those distributions because we have multiple sources of entropy to maximize. One direct way is to maximize their weighted sum. However, to determine weights is not easy.

We do the combination in a quite practical way. We notice that in our experiments, the entropy of all of the above distributions except word saturate very early. We can only select roughly 10 hour of

the data by maximizing entropy of those distributions. So in our selection strategy, we use the word distribution as our basic selection criterion. Then we maximize the entropy of the other distributions by adding some more data.

4. EXPERIMENTS

4.1. System Setup

We evaluate our new selection method using the broadcast subset of the RT-04 Mandarin test set. The training system is SPHINX III. The corpus we do our selection on is the transcribed part of the Mandarin release from GALE Q1-Q4 as well as the Mandarin HUB4 1997 training data. These comprise a total of 840 hours in all. We use a 39-dimension MFCC as our features and a 64k word dictionary. The system is speaker independent and uses between 32 $\sim$ 128 gaussians per mixture depending on the amount of data available.

4.2. Results

4.2.1. Selection by Word Distribution Alone

In our first experiment, we use our maximum entropy method to select 30, 50 and 100 hours of data from the 840 hours data according to the word level distribution. We stop at 100 hours since the entropy of words ceases to increase at around 100 hours of data. We compare the resulting model with one prepared using a random sampling of the same amount of data (using multiple samplings to minimize error). To see how well our selection actually works, we also trained with the complete 840 hours. Table 1 shows the main results. More detailed results, broken down by the different components of RT-04 are shown in Table 2˜4.

Table 1. *Overall Result Using Word Distribution*

	30 hour	50 hour	100 hour	840 hour
random	27.6	27.1	26.1	24.3
max-entropy	27.1	26.2	24.8	

Table 2. *30 Hour Result*

	CCTV	NTDTV	RFA	All
random (30 hr)	17.0	24.7	42.9	27.6
max-entropy(30 hr)	15.0	23.0	45.8	27.0

Table 3. *50 Hour Result*

	CCTV	NTDTV	RFA	All
random(50 hr)	15.7	24.2	43.6	27.1
max-entropy(50 hr)	14.0	22.3	44.8	26.2

Table 4. *100 Hour Result*

	CCTV	NTDTV	RFA	All
random (100 hr)	14.3	23.0	42.0	25.8
max-entropy(100hr)	13.0	21.1	42.7	24.8

4.2.2. Combination with Character and Phoneme Distribution

In the second experiment, we add additional data that maximizes distribution on the phoneme and character levels. We keep the 100 hours of data from the word level distribution and add 50 hours of data that maximizes the phoneme distribution and character entropy respectively. We can see the result in Table 5:

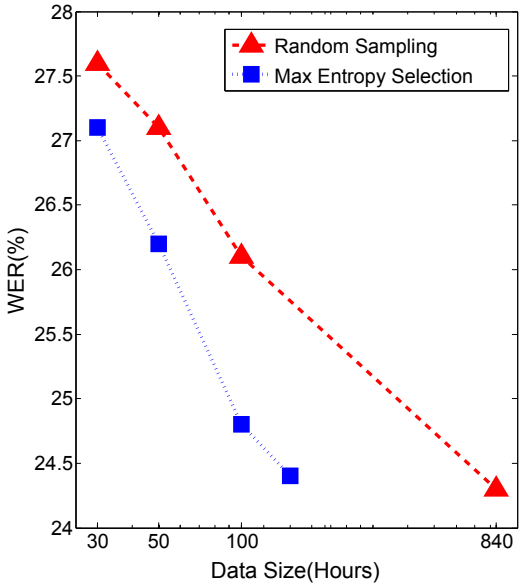

Fig. 1. *Recognition performance for selected and random training sets on RT04 Mandarin test set. Random results for 30h and 50h based on three samples; 100h on 2 samples.*

Table 5. *150 Hour Result*

	CCTV	NTDTV	RFA	All
random (150hr)	13.6	22.2	41.1	25.0
max-entropy (word+char)	12.2	21.8	42.3	24.7
max-entropy (word+phone)	13.1	20.5	41.8	24.4
All data (840hr)	12.9	21.0	41.0	24.3

4.3. Additional Result with VTLN

As our method selects a small subset of a corpus using a limited number of criteria, there is a question of whether it undermines other sources of variability, specifically ones that support subsequent stages of training. We examined whether data selection is additive with speaker adaptive methods such as VTLN. After applying VTLN to the 150 hour selected data, we observe a improvement from 24.4% to 22.5% for character error rate roughly comparable to what we might expect to observe for non-selected data. Notice that the basic principle of our proposed selection could easily use speaker information; we therefore believe our approach will not greatly impair those speaker adaptive methods.

4.4. Analysis of Results

Figure 1 shows overall performance. Note that the Random performance at 30 hours is based on 3 samples, while the performance at 150 hours is based on two separate samples. We can see that the proposed selection procedure is quite effective. By selecting just 17% (150 hours) of the entire corpus, we achieve almost the same per-

formance as by using the entire corpus. In addition, our selection performs better than random sampling at every sample size.

Looking into more detail, we can see that phoneme entropy performs better than character entropy when combined with word level distribution. In our understanding, the main reason is that phoneme distribution is further removed from word distribution (while character distribution is less so). So it would likely contain add useful information in the combination. Another interesting phenomenon is that the max-entropy selection model shows much better gain for the broadcast news part (CCTV, NTDTV), compared with the random model. But it shows some degradation for the broadcast conversation set (RFA). We do not fully understand why this should happen, although we note that the selection algorithm initially favors the broadcast news data, apparently because it contains a more varied vocabulary. Within the proposed selection framework this naturally suggests that some additional distributional criteria based on the properties of the conversational speech would be a natural thing to try (or perhaps this simply points out the benefit of modeling these two types of speech separately).

5. FUTURE WORK

Although our results could be considered preliminary (for example we do not yet know how this selection process interacts with various subsequent modeling steps), we believe that these results suggest a number of interesting research directions in data selection that could be pursued.

5.1. Combine Additional Information

As mentioned earlier, one interesting problem is how to combine distribution at different levels of representation. There are other natural speech unit distributions, such as triphone and senone that we can try to maximize the distribution. Along these lines, we can even ask if there might be other class information available such as speaker, gender, source channel for each speech utterance. Notice that these distributions do not have their prior information captured by the language model and so could benefit from distribution-based selection at the acoustic level. And these classes are usually not directly modeled by the ASR system (unless by separate parallel models). Being able to take these other classes into consideration for selection might prove to be useful. To our best understanding, maximizing the cross entropy between these distributions with the prior (if there is any) appears to be a good strategy.

5.2. Unsupervised Selection and Lightly Supervised Selection

We believe that the idea of sampling a uniform distribution over speech units would also apply, without substantial modification, to unsupervised and lightly supervised data selection. We are interested in combining our approach with traditional confidence-based method to explore the selection of un-transcribed data. Some preliminary results [8] suggests that this may be practical.

5.3. Identifying Training Data for a New Language

Acquiring training data for "less resourced" languages can be an expensive undertaking. Entropy-based data selection method can also be used to guide the acquisition of data for rapid training of acoustic models for a new language [9].

6. CONCLUSION

We have described a framework for efficient data selection for supervised acoustic model training. Our strategy is based on identifying that subset of data that yields the maximum entropy over key properties of the data. We show that this produces recognition performance equivalent to that obtained with much larger, randomly selected, training sets. We find that this algorithm gives good performance with high efficiency. Since the technique is not specific to the data properties that we examined in this study, we believe the approach can be applied in a variety of circumstances.

7. ACKNOWLEDGMENT

This work has been sponsored by a contract from the United States Government under the DARPA GALE program. All views expressed in this paper are those of the authors and not those of the sponsor.

8. REFERENCES

[1] A. Nagórski, L. Boves, and H. Steeneken, "Optimal Selection of Speech Data for Automatic Speech Recognition Systems", Proc. ICSLP, Denver, vol. 4, pp. 2473-2476, 2002.

[2] A. Nagórski , L. Boves , H. Steeneken, "In Search of optimal data selection for training of automatic speech recognition systems" Proc. of ASRU, St. Thomas, 2003, pp. 67-72.

[3] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in Proc. of ICASSP, Orlando, FL, 2002, pp. 3904-3907.

[4] G. Riccardi and D. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," in Proc. of EUROSPEECH, 2003.

[5] G. Tur, R. E. Schapire, and D. Hakkani-Tür, "Active learning for spoken language understanding," in Proc. ICASSP, Hong Kong, May 2003

[6] T. M. Kamm and G. G. L. Meyer, "Selective sampling of training data for speech recognition," in Proc. Human Language Technology Conf., San Diego, CA, 2002.

[7] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, "Lightly supervised and unsupervised acoustic model training", Computer Speech and Language, vol. 16, pp. 115.129, 2002.

[8] R. Zhang and A. Rudnicky, "A New Data Selection Approach for Semi-Supervised Acoustic Modeling", Proc. of ICASSP 2006.

[9] T. Schultz and A. W Black, "Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs", in Proc. of ICASSP, 2006.

[10] T. Hastie, R. Tibshirani and J. Friedman. "Elements of Statistical Learning". Springer, 2001.

[11] R. Duda, P. Hart and D. Stork "Pattern Classification (2nd Edition)", Wiley, 2000.

[12] Xuedong Huang, Alex Acero and Hsiao-wuen Hon, "Spoken Language Processing", Prentice Hall PTR, NJ, 2001

[13] K. Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System". Ph.D. Thesis, Carnegie Mellon University, April 1988.

[14] F. Wessel, H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition". In Proc. of ASRU, Trento, 2001.