Automatic Data Selection for MLP-based Feature Extraction for ASR

Carmen Peláez-Moreno[1], *Qifeng Zhu*[2], *Barry Chen*[2], *Nelson Morgan*[2]

[1]Department of Signal Theory and Communications, University Carlos III, Madrid, Spain
carmen@tsc.uc3m.es
[2]International Computer Science Institute (ICSI), Berkeley, California, USA
qifeng, byc, morgan@icsi.berkeley.edu

Abstract

The use of huge databases in ASR has become an important source of ASR system improvements in the last years. However, their use demands an increase of the computational resources necessary to train the recognizers. Several techniques have been proposed in the literature with the purpose of making a better use of these enormous databases by selecting the most 'informative' portions and thus reducing the computational burden. In this paper, we present a technique to select samples from a database that allows us to obtain similar results in MLP-based feature extraction stages by using around 60% of the data.

1. Introduction

In the last years, databases for speech recognition are becoming bigger and bigger because the usage of these huge databases has become an important source of accuracy improvement. However the inclusion of these enormous resources into the recognition engines does not come without disadvantages: the computational demands have increased considerably.

The use of the outputs of MLP neural networks as features that allow us to incorporate long term information into the feature vectors have proven to successfully increase the recognition performance [1, 2, 3]. Nevertheless, again this inclusion demands significantly more computational effort. Not only do these demands pose a problem to the final recognition systems, but also take an important role in the research stages that has motivated the use of intermediate tasks and several strategies to ease and make the process more efficient [4, 5].

In this context, we propose to look into the data provided in those databases and study its 'usefulness', i.e., to look for and eliminate both the redundancies and the potentially harmful data such as outliers or mislabelled data. Besides we put forward that there are some types of data which are more easily learned than others, and therefore the amount of data pertaining those easy groups could be safely removed from the training without paying any price in the accuracy of the recognition.

Here we report some experiments that have been made in order to show the validity of the previous assertions.

2. Data Selection

Several data selection techniques have been proposed in the literature and variations of them can be found under names such as novelty detection, selective sampling or active learning. However the goals of those techniques can be different [6]:

- **Generative methods** aim at selecting the best samples from unlabeled data to maximize data labeling investments.
- **Selective methods** try to select an adequate subset from labeled data to maximize performance or reduce the computational effort while maintaining a similar performance. Here we can further distinguish between *wrapper* and *filter* approaches. The former employs a statistical re-sampling technique (such as cross validation) and uses the actual target learning algorithm to estimate the accuracy of the subsets. Its disadvantage is its high cost because the learning algorithm has to be called repeatedly. The *filter* techniques are based on selectors that operate independently of the learning algorithm, i.e., undesirable samples are filtered out of the data before the induction commences [7].

Though generative methods have also important applications in speech recognition, here we are primarily concerned with the selection of already labeled data. As we will further explain, the technique we are proposing is based on the *filter* approach but does not employ the actual labels making it suitable for its use as a generative method. However, if used in the later fashion a labeling stage performed after the selection must be undertaken. Besides, it shares the inspiration from the *wrapper* approach of using the target learning algorithm to perform the selection. This makes the selection and learning criteria match, using, however, for the selection, a reduced version of that algorithm to avoid the high computational cost.

To gain a better understanding of selective methods it is worth mentioning that these methods obtain their benefits from two facts:

- Reducing the redundancy existing in the database can help to reduce the costs of learning achieving the same performance with less effort. Redundancy, however, should not be measured in terms of the number of examples present for each class due to two reasons: in the first place, not all the classes are equally separable making certain classes easier to learn and therefore its samples more redundant for the learning machine and second, the most common samples in the training set are usually the most common samples in the test set as well making it wise to model some classes better than others.
- However, over-represented examples in the database can harm the generalization capabilities of a given learning machine biasing its modeling toward those classes. This can be negative if the distribution of testing samples among classes is not the same as seen in training.

2.1. Evaluation methods and sampling criteria

For the selection of data based on the *filter* approach we need an *evaluation method* that allows us to sort the data according

to some *sampling criteria* or definition of *usefulness* of the data.

The evaluation method should meet the following,

- it should be easy to compute so that the whole process of the selection plus the training with the selected samples results in training time reductions without sacrificing much performance compared to training with the complete database,
- it should complement the learning algorithm so that the selection matches the needs of the classification machine.

Therefore our choice for the evaluation method is a classifier of the same family of the one we will use for the final classification but with a reduced size.

As for the sampling criterion we will make the following considerations:

- According to the MCE (Minimum Classification Error) criterion the most useful data is the one that best helps to define decision boundaries. Therefore, it would be useful to select those examples that are in the region of uncertainty because in that region the hypothesis made by the model is very likely to be an error,
- however, we are also concerned with the correct estimation of the posterior distributions and therefore are also interested in modeling the non boundary regions,
- while lying in the region of uncertainty the outliers or miss-labeled samples are not desirable to learn.

These considerations will lead us to balance the amount of 'difficult' examples we consider with that of the 'common' or usual examples as will become apparent in the next section.

3. Data selection for MLP-based Feature Extraction in ASR

The inclusion of probabilities a posteriori from different time spans as features to train the HMM acoustic models has been proven to give enhanced performance [2, 3, 1] in the final speech recognition system.

Nevertheless, this does not come without a very important computational effort given that enormous MLP nets must be trained to profit from the increasingly bigger databases. Several 'ad hoc' tricks must be used to reduce the time and computational burden of the networks training. The automatic selection of data and its corresponding reduction in the computational time needed can help to reduce the time necessary without having to rely on personal experiences and intuitions.

3.1. Evaluation method

According to the requirements for a desired evaluation method discussed in section 2.1, we have chosen the entropy of the outputs of a smaller classifier in the *family* of learning machines that we want to train. We have chosen as an indication of the *usefulness* of the data the entropy of the posterior probabilities coming from a smaller MLP.

Therefore, as a first step, we have to train an MLP selector, *s*, using a small subset of the data that will result in a set of parameters, $\vec{\theta}$.

Afterwards, given those parameters $\vec{\theta}$ we can then obtain the probabilities a posteriori for the rest of the data,

$$p_s\left(q_k|\vec{x}[n],\vec{\theta}\right) = s\left(\vec{x}[n]\right) \quad k = 0,..,K-1 \tag{1}$$

for every feature frame $\vec{x}[n]$ and phoneme, q_k. We can now compute the entropy value for each feature frame as:

$$h[n] = \sum_{k=0}^{K-1} p_s\left(q_k|\vec{x}[n],\vec{\theta}\right) log_2 p_s\left(q_k|\vec{x}[n],\vec{\theta}\right) \tag{2}$$

of we could even use the perplexity ($p[n] = 2^{h[n]}$) as we have found that the probability distribution of the latter is smoother and can ease the search for the appropriate thresholds that separate the selected from the unselected samples. Both, $h[n]$ and $p[n]$ are now measurements of the *difficulty* of the decision that the classifier does given that the frame error rate will be now computed as,

$$FER = \frac{1}{N}\sum_{n=0}^{N-1} step\left\{\left|l\left(\vec{x}[n]\right) - \hat{l}\left(\vec{x}[n]\right)\right|\right\} \tag{3}$$

where $l\left(\vec{x}[n]\right)$ and $\hat{l}\left(\vec{x}[n]\right) \in \{0,..,K-1\}$ are the true and estimated labels for the frame feature vector $\vec{x}[n]$, respectively, and $\hat{l}\left(\vec{x}[n]\right)$ is computed as follows:

$$\hat{l}\left(\vec{x}[n]\right) = \underset{k}{\text{argmax}}\left\{p_s\left(q_k|\vec{x}[n],\vec{\theta}\right)\right\} \tag{4}$$

3.2. Sampling criteria

The sampling criterion we have to chose has to take into account the observations we have made in section 2.1. Using our entropy measure we can now say that:

- High entropy values indicate that taking a decision is going to be difficult
- Low entropy values indicate that the decision is easy to make (not necessarily implying it will be the right one)

but

- Very high entropy values may account for outliers or mislabeled examples: non-separable data. We usually avoid learning these examples by limiting the capacity of the classifier or using procedures such as early stopping in the MLP.
- Very low entropy values can account for overrepresented or easily learnt examples. This overrepresentation can harm the classifier abilities by forcing too much detail in the corresponding class. Besides, these examples usually correspond to the mean (or central) part of the probability distributions, and decision boundaries are not placed there.

Therefore, we are interested in selecting high entropy examples (avoiding the extremely high or low entropy ones) and thus using the capacity of the MLP to model the tails or the probability distributions, that is, the place where the decision boundaries are usually situated.

4. Experiments

4.1. Task definition

In our experiments we have used a large vocabulary continuous speech recognition task with conversational telephone speech data. The training set contains 23 hours of speech mainly from the Switchboard Corpus, randomly sub-sampled. The test set contains a 1.4 hour from the NIST 2001 Hub5 evaluation set.

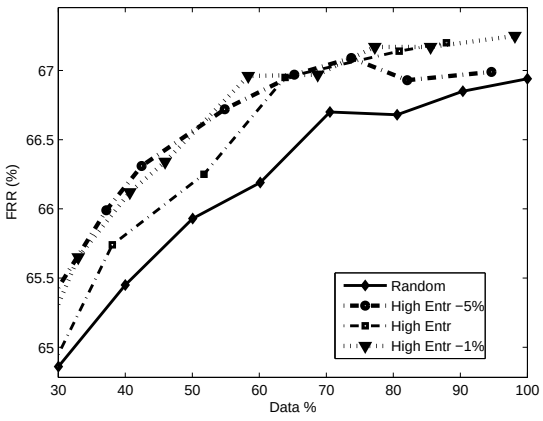

Figure 1: *Sample selection vs Random selection evaluated on Eval2001 for males*

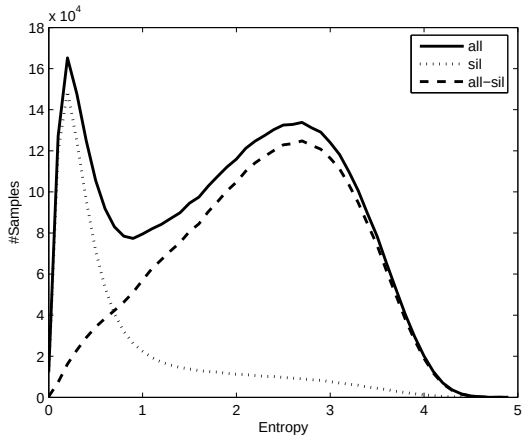

Figure 2: *Entropy distribution of the outputs of the selector.*

First the most frequent 500 words are identified, and the selection of the set set is based on the criterion that each utterance has lower than 10% out-of-vocabulary (OOV) rate. More details about this task can be found in [5].

4.2. Frame error level results

As a baseline for the frame error level recognition we have used an MLP network with:

- 351 inputs (a context window of 9 frames and a feature vector dimension of 39 containing 12 PLP + 1 Energy + 12 ΔPLP + 1 ΔE + 12 $\Delta\Delta$PLP + 1 $\Delta\Delta$E),
- 1300 hidden units
- 46 outputs corresponding to 46 phonemes

For the training of the selector we have used a smaller MLP network (20% of the parameters which makes it fast to train) that uses a randomly selected 20% of the database. Finally, we have validated our results using FRR (Frame Recognition Rate) measures over an independent test set (Eval2001).

Figure 1 compares the results obtained for different training amounts of data (horizontal axis) performing,

1. a random selection, i.e., different amounts of data are selected randomly in each experiment
2. a high entropy selection, i.e., the samples are selected in order from the ones with the highest entropy toward the ones with lowest entropy plus the data used to train the selector (that had been selected randomly),
3. same as previous but discarding the 1% with the highest entropy plus the data used to train the selector,
4. same as in 2. but discarding the 5% with the highest entropy plus the data used to train the selector.

We can draw the following conclusions from this figure:

- The entropy selection always performs better than the random selection for the same amount of data.
- In the range of 50-60 % the entropy selection achieves the same FRR results that the random selection with the 100 % of the data.
- Avoiding the highest entropies is better in the ranges of interest (50-65%) but it becomes less important when most of the data is back in the MLP.

4.3. Word error level results

Automatic speech recognition experiments are conducted to verify the data selection scheme. The SRI Decipher system is used to conduct the experiments. Gender dependent HMMs are trained with the maximum likelihood criterion. A bigram language model was used in the decoding. To verify the data selection scheme in this paper, we only used one gender (males) in ASR experiments.

Table 1 shows the Frame Error Rate (FER) results obtained on Eval2001 and Word Error Rate (WER) for: a) the tandem baseline (PLP + PLP/MLP) (i.e. using all the available data), b) a selection of the 58% of data (avoiding the 1% top highest entropy and including the data used for the training of the selector) and c) a random selection using 60% of the data.

Table 1: *Comparison of results using entropy selected and unselected data.*

Selection	Data (%)	FER (%)	WER (%)
Baseline	100	33.06	42.1
Random	59.68	36.28	42.8
Selected	58.35	33.04	42.2

As it can be observed the entropy-based selection process is clearly better than a random selection, both in FER and WER and what is more important: by selecting around 60 % of the data it is possible to obtain almost the same results than with the whole database.

5. Discussion

As we are interested in knowing what effects on the MLP training result from the data selection, we have further analyzed the results of the baseline and the entropy based selected sets by obtaining their corresponding entropy distributions.

In figure 2, we can observe the entropy distribution of the outputs of the selector for all of the phonemes, the silence phoneme and the rest (non-silence phonemes).

It should highlighted that the first mode of the bimodal distribution obtained is due mostly to the silence phonemes (that account for the 25.12 % of the samples). However, it should be pointed out that this selection method does not perform a mere

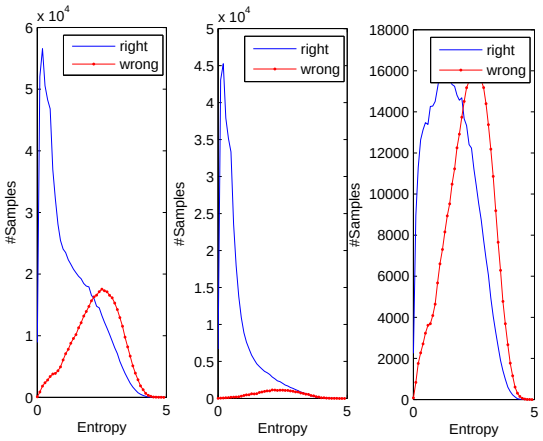

Figure 3: *Same as Figure 2 showing the distributions of the right and wrong decisions for the evaluation set obtained by the MLP net trained using no selection (baseline results) for (a) the whole database, (b) the silence samples and (c) all except the silence samples*

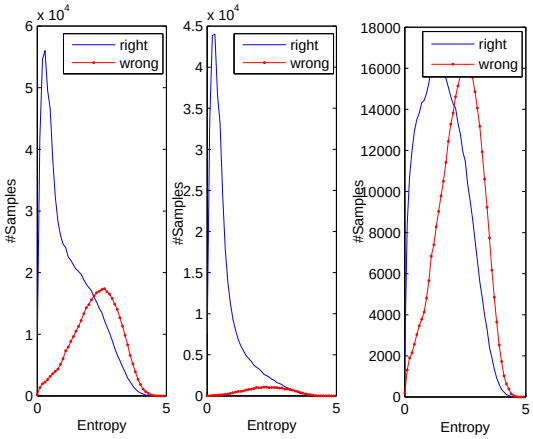

Figure 4: *Same as Figure 3 for the MLP net trained using the selection of 58% of the data*

elimination of the silence phonemes (some of them have middle entropy values) but also eliminates other low entropy samples.

Also, for each of these distributions we are interested in a separate analysis of the entropy distributions of the data that finally results in right and wrong decisions. We depict these distributions in Figure 4 for the selected data set, and in Figure 3 for the whole unselected database. As we can observe the two figures, they are almost identical showing that the MLP nets trained using the selected data set and the whole database have a similar performance. It is also worth noting that, as we expected, wrong decisions happen mostly for high entropy samples. What is especially striking is the distribution for wrong decisions in silence samples: almost every silence sample is correctly classified.

6. Conclusions

A selective sampling method based on the entropy values of an MLP selector has been employed to reduce the size of the training set of an MLP network used to obtain probabilities a posteriori for its use as features in an ASR system. With this method we have obtained almost the same ASR results than with the full database by only using around 60% of the data. This can be of great help given the present trend of increasing the size of the databases to obtain better recognition performances.

7. Future Directions

The theory behind the sample selection presented here can be applied to full database selection for the whole recognition system. Word or sentence level approaches similar to [5, 8, 9, 10] should be selected in that case. However several considerations about how to adapt the evaluations measures must be taken into account.

8. Acknowledgements

The authors would like to thank the Spanish Education and Science Ministry for supporting this research under the ICSI fellowship program for Spanish technologists.

9. References

[1] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, " On using MLP features in LVCSR," in *Proceedings of Intl. Conf. Spoken Language Processing*, Jeju, Korea, October 2004.

[2] B. Chen, Q. Zhu, and N. Morgan, "Learning Long-Term Temporal Features in LVCSR Using Neural Networks," in *Proceedings of Intl. Conf. Spoken Language Processing*, Jeju, Korea, October 2004.

[3] N. Morgan, B. Y. Chen, Q. Zhu, and A. Stolcke, " TRAPping Conversational Speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition," in *Proceedings of Intl. Conf. on Audio, Speech and Signal Processing*, Montreal, Canada, May 2004.

[4] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke, "Scaling up: learning large-scale recognition methods from small-scale recognition tasks," in *Special Workshop in Maui (SWIM) paper 218*, Maui, USA, 2004.

[5] B. Chen and et al., "A CTS task for meaningful fast-turnaround experiments," in *EARS Workshop*, New York, USA, November 2004.

[6] S. Vijayajumar and H. Ogawa, "Improving generalization ability through active learning," *IEICE Trans. Info. and Syst.*, vol. W82-D, no. 2, pp. 480–487, February 1999.

[7] M. Hall and L. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper," in *Proc. of the Florida Artificial Intelligence Symposium (FLAIRS-99)*, Florida, USA, 1999.

[8] T. Kamm and G. Meyer, "Automatic Selection of Transcribed Training Material," in *Proceedings of ASRU*, Trento, Italy, December 2001.

[9] ——, "Selective sampling of training data for speech recognition," in *Proceedings of Human Language Technology*, San Diego, USA, 2002.

[10] T. M. Kamm, *Active Learning for acoustic speech recognition modeling*. PhD Dissertation.