LATTICE-BASED UNSUPERVISED ACOUSTIC MODEL TRAINING

Thiago Fraga-Silva, Jean-Luc Gauvain, Lori Lamel

Spoken Language Processing Group
LIMSI - CNRS B.P. 133 91403 Orsay cedex FRANCE
{thfraga,gauvain,lamel}@limsi.fr

ABSTRACT

Unsupervised acoustic model training has been successfully used to improve the performance of automatic speech recognition systems when only a small amount of manually transcribed data is available for the target domain. The most common approach is use automatic transcriptions to guide acoustic model estimation. However, since the best recognition hypotheses are known to contain errors, we propose to consider multiple transcription hypotheses during training. The idea is that the EM process can benefit from the estimated posterior probabilities of the hypotheses to converge to a better solution. The proposed unsupervised training method is based on lattices. Lattice-based training gives a relative improvement of 2.2% over 1-best training on a Broadcast News transcription task and converges faster with the iterative incremental training.

Index Terms— Unsupervised training, Acoustic Modeling, Lattice-based training, Speech recognition

1. INTRODUCTION

Acoustic model development for Automatic System Recognition (ASR) relies on a large amount of transcribed audio data to achieve suitable performance levels. However, obtaining manual transcriptions is both expensive and time-consuming. A well known technique that has been gaining popularity as a means to reduce the human effort for this task is to train an acoustic model on a small amount of manually annotated data and, then, use this model to recognize several hours of training data. Once automatic transcriptions are available, they can be used to train a new acoustic model. This method, known as *unsupervised training*, has been used with success for Broadcast News and Broadcast Conversation data [1] and has been applied to different languages [1, 2, 3, 4, 5].

The main problem of the unsupervised training method is that the automatic transcriptions contain more errors than the manual ones. The lightly supervised method [3] provides an efficient way to remove incorrectly recognized words, but only if approximate transcriptions, such as closed captions, are available. Since this is not always the case, other filtering techniques should be applied to improve the performance. The most common method to errors filter out is based on lattice confidence measures at the word [2, 6, 7] or state level [8]. In this case, a segment is used for training only if its confidence measure is greater than a given threshold. Nevertheless, this data selection method is not exempt of errors. For instance, if a small threshold is chosen, some incorrectly recognized words remain in the training data, misleading the parameter estimation. On the other hand, if a high threshold is set, not enough information is added to the model, i.e., the system only learns what it already knows.

Unsupervised training is evidently an incomplete data problem since the transcription is unknown, in addition to the HMM state and Gaussian level alignments. The Expectation-Maximization (EM) algorithm can be used to solve this problem by iteratively estimating the word hypothesis with the current acoustic model. Here we propose to consider a set of possible hypotheses to obtain a better estimate (in the maximum likelihood sense) of the model parameters. In our case, the different hypotheses for each utterance are limited to a word lattice generated with the current best acoustic model and appropriate pruning thresholds. The hypothesis posterior probabilities are derived from the corresponding lattice.

Lattices have been successfully used in many areas of speech recognition. For instance, in [9] the authors compare lattice and 1-best based MLLR for speaker adaptation. However, to the best of our knowledge, these are the first experiments reported on the use of lattices on Maximum Likelihood (ML) based unsupervised acoustic training.

The unsupervised training method has been used with both ML estimation [2, 3, 8] or discriminative training approaches [1]. This study is restricted to the ML framework, but the work can be extended to discriminative training.

The remainder of this paper is organized as follows. After a brief presentation of the 1-best unsupervised training method, the next section describes lattice-based training, focusing on the influence of the posterior probability estimation and the size of the lattice on model estimation. Section 3 reports the results obtained comparing the 1-best and the lattice-based methods and the impact of some parameters on lattice-based training.

This work has been partially supported by OSEO, the French State agency for innovation, under the Quaero program.

2. LATTICE-BASED TRAINING

Usually 1-best unsupervised training is an iterative and incremental procedure which is briefly described as follows. First, a bootstrap acoustic model is trained on a very small amount of manually transcribed data. Alternatively, one can use an existing model trained for another target domain or language. This model is used to generate transcriptions of a larger amount of training data. These transcriptions can be optionally filtered to remove incorrectly recognized segments. The automatic transcriptions are then used to estimate a new acoustic model. The process is reiterated using a larger amount of data at each time, usually doubling the amount of training data at each iteration.

HMM training requires alignment between audio and phone models. On traditional unsupervised training, the recognizer provides a unique hypothesis of alignment for the EM-based training. Assuming that we use the alignment as non-observable variable, the HMM parameters are estimated by the following maximization step:

$$\hat{\lambda} = \arg\max_{\lambda} f(X, W|\lambda) \tag{1}$$

where λ represents the model parameters, X the audio observed and W is the word hypothesis.

The 1-best approach assumes that the given hypothesis, filtered or not, is a good estimation of the true transcription. However, as discussed in [2], the training data will still contain errors if a weak filter is applied. Using several relevant hypotheses with their probabilities should be a better representation of the truth of the recognized data.

Considering this new representation, some adjustments to the training procedure need to be made. The various hypotheses considered can be associated with different words, phones and alignments. The hypothesis probabilities also have to be taken into account during parameter estimation.

The likelihood maximization is now performed considering the different lattice hypotheses as follows:

$$\hat{\lambda} = \arg\max_{\lambda} \sum\nolimits_{W \in \mathcal{L}} f(X, W|\lambda) \tag{2}$$

where the summation is taken over all the word hypotheses W in the lattice $\mathcal{L}$.

The success of the process described above relies on a good estimation of the hypothesis posterior probabilities used to solve equation 2. Let us consider that the posterior probability is given by the following equation:

$$f(W|X) = \frac{(f(X|W)^{1/\alpha} P(W))^{\gamma}}{\sum_{\hat{W} \in \mathcal{L}} (f(X|\hat{W})^{1/\alpha} \cdot P(\hat{W}))^{\gamma}} \tag{3}$$

where the parameter α represents the well-known acoustic model scaling factor and γ controls the likelihood ratio among the hypotheses. As γ increases the posterior probability approaches a delta distribution, that is approaching the 1-best solution. Alternatively, as γ decreases, the posterior distribution becomes uniform. Both α and γ can be used to adjust the ratio probabilities among the many hypothesis. However, since α also changes the 1-best solution (and γ does not), this work has focused only on the impact of γ.

Besides the posterior probabilities, other parameters may affect the performance of the lattice based models, in particular the number of relevant hypotheses considered. Usually, the hypotheses with lower probabilities do not carry any significant improvement to the system and, not rarely, may be harmful. Moreover, increasing the size of the lattices also increases the time spent on training.

To choose the most relevant hypotheses in a lattice, a well known technique that can be used is simply reject the segments for which the posterior probabilities are inferior to a given threshold. The lattice can also be reduced by a likelihood pruning. Basically, in this method, if the probability of any path leading to a node is lower than a certain threshold, the path is removed from the lattice. In this case, this threshold is calculated by the probability of the best path leading to this same node multiplied by a given parameter B. The impact of both methods are reported in this work.

3. EXPERIMENTS

The experiments were carried out using the LIMSI speech recognition toolkit, with acoustic, lexical and language models developed for the Portuguese language and tested for a Broadcast News (BN) task.

3.1. Corpora

The manually transcribed training data used in these experiments were collected from the Portuguese channel RTP from different shows broadcast on April 2000. The amount of usable data is about 3h. The untranscribed corpus used in this work consists of over 173h of usable data collected from different shows and epochs from three sources: RTP from 2001 (56h) and 2010 (78h), Voice of America (VOA) from 2009 (21h) and Euronews (Euro) from 2008 to 2010 (18h). These sources have significant differences in terms of quality. VOA and Euronews are composed by clean speech data with good background conditions and a small number of speakers. On the other hand, the RTP shows contain many different speakers, have rapid turns between speakers and varied background conditions. Results are reported on four evaluation sets, each set containing data from the same training source. The RTP00 and RTP09 evaluation sets have about 76 and 71 minutes respectively, while Voa09 and Euro10 have about 2 hours each. All the acoustic data available for training and evaluation are in European Portuguese.

The language model training data include 483M words of newspaper, newswire and blog texts, closed captions and

fine transcriptions from different sources from 1991 to 2010. They are comprised of both European Portuguese (79%) and Brazilian Portuguese (21%) texts. The vast majority of the training texts are from written sources with about only 5.9M words of closed caption data. The transcriptions used for language modeling are the same ones that are used for supervised acoustic model training, containing about 32k words.

3.2. System description

The system used in these experiments is quite similar to other BN transcription systems used at LIMSI [10].

For acoustic modeling, context-dependent triphone models are used. The acoustic feature vector has 39 components, compounded of 12 cepstrum coefficients and the log energy, with their first and second derivatives. The phone set contains 35 phones, as well as special units for silence, breath and hesitation markers.

Language models were obtained by interpolation of component backoff n-gram models trained on the different text sources. Three language models were used, one for each target data type (RTP, VOA and Euro). For each one, a development set containing the same type of data, but having no overlap with the training or evaluation sets, was used to optimize the interpolation coefficients.

A large lexicon containing 158k words was used so as to limit out-of-vocabulary (OOV) rate of the development data. With this vocabulary the OOV rate of all three LM development text sets is under 1%. A pronunciation dictionary was obtained for this word list via a ruled-based grapheme to phoneme (G2P) converter that has four main steps: 1) pre-syllabification; 2) stress syllable marking; 3) post-syllabification; and 4) G2P conversion. The G2P system used in these experiments has about 530 rules and generates (with few exceptions) a unique pronunciation for each word. Some pronunciation variants were manually added for a few frequent words. Alternative pronunciations for acronyms were automatically generated.

3.3. Results

3.3.1. Impact of γ and lattice size reduction

The baseline acoustic model, trained on 3 hours of manually transcribed data, gives an average Word Error Rate (WER) of 27.0% on the combined evaluation sets, as shown in the top entry of Table 1. This acoustic model set was used to decode the first 11 hours of training data, with which a 1-best and a lattice-based models were trained. During decoding, the same α was used for both models, the α that maximizes the performance of the 1-best model. For the lattice model, this first decoding was performed with $\gamma = 1$, $B = 10^{-8}$ and a probability threshold of 0.01. The performance of the unsupervised 1-best and lattice trained models is given in the lower part of Table 1. Even with only 11 hours of audio data, both methods of unsupervised acoustic training outperform supervised training with only 3 hours of data.

Table 1. WER on individual evaluation sets and average WER for the first iteration of unsupervised acoustic training with 1-best hypotheses and lattices. Baseline corresponds to supervised AM training on a 3 hour corpus.

Model	RTP00	RTP09	VOA09	Euro10	Avg
Baseline	36.1	36.2	21.4	22.5	27.0
1-best	32.7	33.1	16.0	18.1	22.7
Lattice	32.2	32.8	15.7	17.6	22.3

Table 2. Impact of γ on lattice-based models.

γ	0.1	0.5	1.0	1.5	3.0	10
WER [%]	22.7	22.4	22.3	22.2	22.3	22.6

Since the baseline model was trained only on RTP data, the relative improvement obtained on the Voa09 and Euro10 test sets are much greater ($> 19\%$) than the improvement on the RTP test data (9-11%). On average, the model trained with the 1-best hypotheses has a relative improvement of 15.9% over the baseline model, while, for the model trained with lattices, the improvement is about 17.8%. In a direct comparison of the two unsupervised methods, there is an absolute improvement of 0.4% (1.7% relative) with lattices.

Before carrying out incremental unsupervised training, the impact of γ when estimating the lattice-based models was assessed. Table 2 shows the WER of different lattice-trained models with values of γ varying from 0.1 to 10.

It can be seen that with the extreme values of γ, 0.1 and 10, the performance of the lattice-based model is affected. Particularly, we observe that for $\gamma = 10$ the performance is quite similar to the 1-best model, as expected, since the lattice approaches the 1-best solution for high values of γ. For γ in the range of 0.5 to 3.0, there is only a slight impact on performance with the WER varying less than 0.2% absolute. Since the best performance is for the model estimated with $\gamma = 1.5$, this value was used in the remaining experiments.

In addition to the influence of γ, we also evaluated the impact of lattice size on training. First, a likelihood based pruning was used with values of B varying from 10^{-4} to 10^{-10}. The best performance was obtained with $B = 10^{-8}$, with less than 0.2% absolute loss in performance for the other values tested. Second, a posterior probability threshold pruning was applied. Table 3 shows the results obtained with models trained on the reduced lattices for threshold values varying from 0.001 to 0.3. Applying a threshold of 0.01 gives the best performance for the lattice-based models, with, again, about 0.2% absolute loss in performance for the extreme values.

The results show that training is quite robust to the lattice size, with a small loss in performance for the range of values tested. However, the pruning parameter affects the resulting acoustic model quality. If a weak pruning is applied, low

Table 3. Impact of the posterior probability threshold on lattice-based models.

Threshold	0.001	0.01	0.1	0.3
WER [%]	22.4	22.2	22.3	22.4

Table 4. WER of 1-best and lattice models after each iteration of incremental unsupervised training, as well as the WER of the baseline models with supervised training.

System	Dur	RTP00	RTP09	VOA09	Euro10	Avg
Baseline	3h	36.1	36.2	21.4	22.5	27.0
1-b(1x)	11h	32.7	33.1	16.0	18.1	22.7
lat(1x)	10h	32.2	32.8	15.4	17.7	22.2
1-b(2x)	22h	31.1	31.3	14.6	16.6	21.2
lat(2x)	20h	30.5	30.9	14.2	16.2	20.7
1-b(3x)	44h	29.8	30.2	13.6	15.1	19.9
lat(3x)	41h	29.0	29.3	13.1	14.7	19.3
1-b(4x)	87h	27.6	28.4	12.9	14.3	18.7
lat(4x)	80h	27.5	28.0	12.6	13.9	18.4
1-b(5x)	173h	26.6	27.3	12.9	13.5	18.1
lat(5x)	157h	26.2	26.5	12.5	13.5	17.8
1-b(6x)	173h	26.4	27.3	12.9	13.7	18.1
lat(6x)	159h	26.0	26.8	12.6	13.3	17.7

probability segments that contain erroneous labels may mislead the parameter estimation. Even if each of these segments has only a minor influence, the joint effect is non-negligible. On the other hand, with a higher pruning level fewer hypotheses are used during training, approximating the 1-best model.

3.3.2. Incremental unsupervised training

Incremental unsupervised training was pursued with both the 1-best and lattice based methods at 6 iterations. At the first 5 iterations, the amount of training data was doubled and all data decoded with acoustic models trained for the respective configuration. At the last iteration, no training data was added. For the lattice-based models, the best parameters determined before were used during all training. Table 4 shows the results obtained. The column labeled 'Dur' specifies the amount of data used at each iteration. For the lattice-based models, the duration takes into account the probabilities of each hypothesis. Since segments with low probabilities are removed, the total amount of data is slightly smaller than that used to train the 1-best models. It can be seen in Table 4 that roughly 10% of the data are removed.

For all the evaluation sets, the lattice-based models outperform the 1-best models, with an average absolute improvement of about 0.4%. Using NIST tools, it was found that the results obtained at each iteration are statistically significant ($p < 0.01$). The relative improvement of lattice over 1-best based models increases at almost all iterations, except at the 4th, indicating that the lattice-based training converges faster. At the end of the 6th iteration, a relative improvement of 2.2% is obtained over 1-best based training.

4. CONCLUSIONS

In this paper, we have proposed a new unsupervised acoustic training method that takes into account various hypotheses of transcriptions during the HMM parameter estimation. The impact of optimizing the factor γ, which controls the posterior probabilities distribution has been experimentally explored. We also demonstrated that the lattice-based training is quite robust to lattice size reduction by likelihood pruning or probability cutoff for the range of thresholds applied. Concerning the incremental training, the present lattice-based method converges faster than the classical 1-best one, and has a relative error reduction of about 2.2%.

5. REFERENCES

[1] L. Wang, M. J. F. Gales, and P. C. Woodland, "Unsupervised training for mandarin broadcast news and conversation transcriptions," in *ICASSP*, Honolulu, Hawaii, April 2007, vol. IV, pp. 353–356.

[2] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: recent experiments," in *Eurospeech*, Budapest, Hungary, September 1999, pp. 2725–2728.

[3] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[4] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *ICASSP*, Toulouse, France, May 2006, vol. III, pp. 1056–1059.

[5] L. Lamel and B. Vieru, "Development of a speech-to-text transcription system for finnish," in *SLTU*, Penang, Malaysia, May 2010, pp. 62–67.

[6] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, February 1998, pp. 301–305.

[7] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding Workshop*, Trento, Italy, December 2001.

[8] C. Gollan, S. Hahn, R. Schlueter, and H. Ney, "An improved method for unsupervised training of LVCSR system," in *Interspeech*, Antwerp, Belgium, August 2007, pp. 2101–2104.

[9] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised MLLR for speaker adaptation," in *ISCA ITRW ASR2000*, Paris, pp. 128–131.

[10] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, pp. 89–108, 2002.