DICTIONARY-BASED DISCRIMINATIVE HMM PARAMETER ESTIMATION FOR CONTINUOUS SPEECH RECOGNITION SYSTEMS

Daniel Willett, Christoph Neukirchen, Jörg Rottland

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
e-mail: {willett,chn,rottland}@fb9-ti.uni-duisburg.de

ABSTRACT

The estimation of the HMM parameters has always been a major issue in the design of speech recognition systems. Discriminative objectives like Maximum Mutual Information (MMI) or Minimum Classification Error (MCE) have proved to be superior over the common Maximum Likelihood Estimation (MLE) in cases where a robust estimation of the probabilistic density functions (pdfs) is not possible. The determination of the overall likelihood of an acoustic observation is the most crucial point of the MMI-parameter estimation when applied to continuous speech systems. Contrary to the common approaches that estimate the overall likelihood of the training observations by evaluating the most confusing sentences or by applying global state frequencies, this paper suggests to perform a dictionary analysis in order to get estimates for the dictionary-based risk of mixing up each two HMM states. These estimates are used to estimate the observations' likelihood and to control the discriminative MMI training procedure.
Results on a monophone SCHMM speech recognition system are presented that prove the practicability of the new approach.

1. INTRODUCTION

The most popular approach for the estimation of the parameters of an HMM speech recognition system is the Maximum Likelihood Estimation (MLE). It aims to maximize the likelihood of the training observations X as output of a given correct transcription W, while neglecting the output's overall likelihood. It can be written as

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} p_\lambda(X|W) \tag{1}$$

The better the assumed family of distributions covers the true distribution of the data and the more training data is available, the closer the pdfs estimated according to Maximum Likelihood converge to the true distributions [1]. The more these assumptions are not met, however, discriminative training procedures like the Maximum Mutual Information Estimation (MMIE) can provide more useful parameter estimates [2].

The objective of the MMIE differs from the MLE by relating the observation's likelihood given the correct transcription to its overall likelihood:

$$\hat{\lambda}_{MMI} = \arg\max_{\lambda} I_\lambda(X,W)$$

$$= \arg\max_{\lambda}(H_\lambda(X) - H_\lambda(X|W)) = \arg\max_{\lambda} \frac{p_\lambda(X|W)}{p_\lambda(X)} \tag{2}$$

This relation takes into account that the success of the recognition procedure not only depends on the likelihood of the correct HMM-sequence but on the likelihood of the possible incorrect sequences as well. Several publications reported the superiority of the MMIE over the MLE for tasks like isolated word recognition [2] and phonetic recognition [3] where the overall likelihood of an observation can be determined exactly according to

$$p_\lambda(X) = \sum_{all\ \widehat{W}} p_\lambda(X|\widehat{W})p(\widehat{W}) \tag{3}$$

as the number of possible transcriptions is limited.
In continuous speech recognition systems though, the exact computation of $p_\lambda(X)$, that has to take into account all possible word sequences, is considered to be too complex or it is simply impossible in large vocabulary systems. Hence methods for the approximation of $p_\lambda(X)$ have to be found. Common approaches estimate $p_\lambda(X)$ by considering only the most confusing ("best") sentences which are those sentences W with high values for $p(X|W)p(W)$.

$$p_\lambda(X) = \sum_{all\ \widehat{W}} p_\lambda(X|\widehat{W})p(\widehat{W}) \approx \sum_{best\ \widehat{W}} p_\lambda(X|\widehat{W})p(\widehat{W}) \tag{4}$$

N-best lists [4] and word-lattices [5, 6] have been successfully applied for such an estimation.
These approaches have two major disadvantages. The one is the computational complexity caused by the need for several n-best or word-lattice construction procedures on the training data. The other one is the dependence on the confusion measured on the training data that only contains a small fraction of all the possible confusion and that might be specifically distributed misrepresenting the distribution of confusion on unseen test data.

A second possible approach for a discriminative parameter estimation of continuous speech recognition systems is based on the usage of global estimates for the HMM states' overall frequency of appearance $p(w_j)$. For the sake of simplicity we now provide that each of the P HMMs consists

of S emitting states. With T as the number of training samples, $X = (x(1) \dots x(T))$ now representing the training samples' feature vectors, $W = (w(1) \dots w(T))$ representing the phone- or HMM-transcription of these observations [1] and $S^\lambda = (s^\lambda(1) \dots s^\lambda(T))$ representing the state alignment with respect to the parameter set λ, such that $w(n)$ is in state $s^\lambda(n)$ at a time-step n, $p_\lambda(X)$ is estimated by

$$p_\lambda(X) \approx \prod_{i=1}^{T} \sum_{j=1}^{P} \sum_{k=1}^{S} p_\lambda(x(i)|w_j,k)p(w_j,k) \qquad (5)$$

$p_\lambda(w_j,k)$ resembles the overall frequency of HMM w_j being in state k and $p_\lambda(x(i)|w_j,k)$ is the output probability of HMM w_j for $x(i)$ when being in state k. (The parameter set λ is supposed to exclude the HMMs' transition probabilities, as they are not to be estimated discriminatively. Therefore they have been left aside in equation (5).)
In this approach each HMM state's output probability is weighted by its overall frequency of appearance $p(w_j,k)$. With the $p(w_j,k)$ usually measured on the training data this equation is to some extent certainly correlated with the language model, namely as much as the training data is. Nevertheless, it will be referred to as language model independent, as it neglects the impact of the language model on the possible HMM sequences.

Because of the outlined inconveniences of these common approaches, we developed the dictionary-based MMI-training, a very different approach that does not incorporate the confusion on the training data, but predicts possible errors by an analysis of the dictionary.

2. BASIC IDEA OF THE DICTIONARY-BASED APPROACH

The dictionary, the language model and the HMM topology control which states of which HMMs compete with each other in the recognition procedure and which are less relevant for a specific observation. And as a matter of fact it is most important to discriminate the pdfs of those states that compete the most. So the idea is to extract estimates for the degrees of competition between each two states straight from the dictionary. The following example dictionary illustrates the basic idea:

book	/b/	/U/	/k/	
look	/l/	/U/	/k/	
looks	/l/	/U/	/k/	/s/
brick	/b/	/r/	/I/	/k/

It is very important with respect to the recognition error to discriminate the pdfs of the HMMs `/b/` and `/l/` as the words *book* and *look* differ only by these phones. Furthermore it is important to discriminate (the ending of) HMM `/k/` from HMM `/s/` in order not to mix up *look* and *looks*.

[1] Actually, the phonetic transcriptions are most commonly not given with respect to each feature vector, but simply as a sequence of phones $p_1 \dots p_K$ (possibly derived from a sequence of words) and the precise assignments $w(i) = p_k$ are part of the parameter estimation process. This circumstance is left aside for the sake of simplicity. The $w(i)$ can be assumed to be computed by a viterbi-alignment according to the parameter set λ.

Additionally the HMM `/U/` should be discriminated from the concatenation of `/r/` and `/I/` in order to distinguish *book* and *brick*.
On the contrary, discriminating `/l/` and `/k/` is less important in this case because mixing up these phones does not lead directly to another word of the dictionary.

3. FORMALISM FOR THE DICTIONARY-BASED TRAINING

With the language model independent estimation of $p_\lambda(X)$ according to (5) the MMI-equation (2) yields

$$\hat{\lambda}_{\substack{MMI \\ indep}} = \arg\max_\lambda \prod_{i=1}^{T} \frac{p_\lambda(x(i)|w(i),s^\lambda(i))}{\sum_{j=1}^{P}\sum_{k=1}^{S} p_\lambda(x(i)|w_j,k)p_\lambda(w_j,k)} \qquad (6)$$

In this equation the term $p_\lambda(w_j,k)$ weights the influence of each states' output probability according to its frequency of appearance. From the point of view stated in the previous section, these weights can be interpreted as estimates for the importance of discrimination or as a degree of competition between $w(i).s^\lambda(i)$ [2] and $w_j.k$. They can be replaced by general estimates $c(w',s',w'',s'')$ for the degree of competition among each two states s' and s'' of each two HMMs w' and w'' which leads to an estimation of $p_\lambda(X)$ according to

$$p_\lambda(X) \approx \prod_{i=1}^{T} \sum_{j=1}^{P} \sum_{k=1}^{S} p_\lambda(x(i)|w_j,k)c(w(i),s^\lambda(i),w_j,k) \qquad (7)$$

and the general MMI training objective

$$\hat{\lambda}_{\substack{MMI \\ general}} = \arg\max_\lambda \prod_{i=1}^{T} \frac{p_\lambda(x(i)|w(i),s^\lambda(i))}{\sum_{j=1}^{P}\sum_{k=1}^{S} p_\lambda(x(i)|w_j,k)c(w(i),s^\lambda(i),w_j,k)} \qquad (8)$$

In this context the language model independent MMI-equation (6) can be seen as a special case of (8) where the $c(w(i),s^\lambda(i),w_j,k)$ are estimated as $p_\lambda(w_j,k)$. Certainly, the frequency of appearance of state k of phone w_j is a suitable measure for the importance of discrimination, but the next section and our results will show that better estimates can be gained from a careful dictionary analysis.

4. DICTIONARY ANALYSIS

According to the example in section 2 the principle procedure of the dictionary analysis is to search for all pairs of similar dictionary words and to consider these pairs as possible confusions and then to evaluate the contribution of the involved states in the assumed confusions and to set $c(w',s',w'',s'')$ proportional to the total contribution of $w'.s'$ and $w''.s''$.
The detailed realization of this method largely depends on the structure of the HMMs. Therefore the rest of this section will give a more detailed description of the way we derived dictionary-based estimates for the values $c(w',s',w'',s'')$ for a "standard" [7] recognition system of linear HMMs of three emitting states each.
We found that most of the recognition errors belong to the

[2] $w.s$ denotes state s of HMM w.

following three types that were already illustrated in the example of section 2:
1:0 confusion where one phone is not seen or one phone is added by mistake (deletion/insertion)
1:1 confusion where one phone is mistaken for another one (substitution)
1:2 confusion where one phone is mistaken for two others or two phones are mixed up by one other phone (corresponding to a substitution next to a deletion or insertion)
We estimated the contribution of each state to the confusion with each state of the confusing HMMs according to the following matrices:

1:1 confusion
In a substitution of the phones A and B the contribution of the states was weighted as

average contribution in a 1:1 confusion	HMM B State 1	HMM B State 2	HMM B State 3
HMM A State 1	0.7	0.2	0.1
HMM A State 2	0.2	0.6	0.2
HMM A State 3	0.1	0.2	0.7

The values in the matrix estimate the contribution of a state to the interference with the specific state of the other HMM. The high values in the diagonal stand for the circumstance that when mixing up A with B it is mainly the first state of A that is being mixed up with the first state of B, the second of A with the second of B and so on. Thus the dictionary-based MMI-training will have to discriminate these states the most with respect to the 1:1 confusions.

2:1 confusion
The contribution of the HMM states in a 2:1 confusion between two phones A_1,A_2 and B was estimated as

average contribution in a 2:1 confusion	HMM B State 1	HMM B State 2	HMM B State 3
HMM A_1 State 1	0.5	0.1	0
HMM A_1 State 2	0.3	0.15	0
HMM A_1 State 3	0.1	0.25	0.1
HMM A_2 State 1	0.1	0.25	0.1
HMM A_2 State 2	0	0.15	0.3
HMM A_2 State 3	0	0.1	0.5

1:0 confusion
The contribution of the states in an insertion or deletion of a phone A between the phones B_1 B_2 was weighted according to

average contribution in a 1:0 confusion	HMM A State 1	HMM A State 2	HMM A State 3
HMM B_1 State 1	0.1	0	0
HMM B_1 State 2	0.2	0.15	0
HMM B_1 State 3	0.6	0.35	0.1
HMM B_2 State 1	0.1	0.35	0.6
HMM B_2 State 2	0	0.15	0.2
HMM B_2 State 3	0	0	0.1

These matrices were obtained heuristically. More profound methods could determine individual matrices for each pair of HMMs by considering their transition probabilities and the average state durations.

We scanned the dictionary to find the possible confusions of these types and summed up the state-conditioned contributions to $n_{w',s',w'',s''}$.

$$n_{w',s',w'',s''} := \sum_{\substack{confusions\ with\ w' \\ and\ w''\ involved}} contribution\ of\ w'.s'\ and\ w''.s'' \tag{9}$$

From these values we derived the estimates $c_{w',s',w'',s''}$ according to the following normalization

$$c_{w',s',w'',s''} := \begin{cases} \frac{n_{w',s',w'',s''}}{\sum_{i=1}^{P}\sum_{k=1}^{S} n_{w',s',w_i,k}} & : w' \neq w'' \\ p_\lambda(w'',s'') & : w' = w'' \end{cases} \tag{10}$$

Certainly, the considered primitive types of confusions only represent a subset of all the possible confusion that might occur on real data. However, they seem to represent the state to state relation in all confusions very well as our results prove.

5. TRAINING ALGORITHM

For maximizing the MMI-criterion (2) several algorithms have been proposed. [8] showed that even the EM-algorithm can be used. We chose a gradient-descent procedure to optimize the means $\mathbf{m_j}$ of the Gaussian pdfs of a SCHMM-system. With diagonal covariance matrices $\boldsymbol{\sigma}_\mathbf{j}$ and mixture weights $\mathbf{d_{is}}$ for state s of HMM w_i the pdfs are defined as

$$p_\lambda(x|w_i,s) = \sum_{j=1}^{C} d_{isj} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\sigma}_\mathbf{j}|}} e^{-\frac{1}{2}\sum_{l=1}^{N} \frac{(m_{jl}-x_l)^2}{\sigma_{jl}}}$$

so that the partial derivations are given by

$$\frac{\partial(H_\lambda(X) - H_\lambda(X|W))}{\partial m_{jk}} = \frac{\partial \log p_\lambda(X|W) - \partial \log p_\lambda(X)}{\partial m_{jk}}$$

$$= \sum_{i=1}^{T} \left(\frac{\partial \log p_\lambda(x(i)|w(i), s_\lambda(i))}{\partial m_{jk}} - \frac{\partial \log \sum_{l=1}^{P}\sum_{n=1}^{S} p_\lambda(x(i)|w_l,m) c(w(i), s^\lambda(i), w_l, n)}{\partial m_{jk}} \right)$$

and

$$\frac{\partial p_\lambda(x|w_i,s)}{\partial m_{jk}} = d_{isj} \frac{(x_k - m_{jk})}{\sigma_{jk}} \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\sigma}_\mathbf{j}|}} e^{-\frac{1}{2}\sum_{l=1}^{N} \frac{(m_{jl}-x_l)^2}{\sigma_{jl}}}$$

6. EXPERIMENTS AND RESULTS

We applied the dictionary-based training algorithm on a monophone SCHMM speech recognition system for the Resource Management 1000 word, speaker independent, continuous speech recognition task. As features we used a 12-value mel-cepstrum, log energy and these values' first and second derivatives in four independent streams each modeled by 200 shared Gaussian mixtures. After the standard ML-training (EM-algorithm) we optimized the Gaussians' means of each stream according to the language model

Test	ML	LM indep. MMI (6)	dictioary-based MMI
February'89	9.76%	10.15%	9.25%
October'89	12.11%	11.66%	11.36%
February'91	9.66%	9.34%	9.34%
September'92	16.61%	15.98%	15.16%
Average	**12.04%**	11.78%	**11.27%**
Error reduction compared to ML		2.2%	**6.4%**

Table 1. Word error rates on the RM database

independent MMI-equation (6) and according to the new approach (8), with the values $c(w', s', w'', s'')$ estimated according to (9) and (10). Table 1 shows the word error rates on the four SI test sets that were obtained with the standard word pair grammar of perplexity 60. The ML trained system performs remarkably well considering that it does not make use of context-dependent phones. Its pdfs seem to be close to the real distributions. Regarding this, the gain of 6.4% corresponds to what has been reported for other discriminative training approaches [4, 5, 6]. It should be pointed out, that the results obtained with the ML-trained monophone SCHMM system can be considered as very good results for this database with monophone models. Therefore, the reported error reductions have to be rated as considerably strong improvements. The reduction of the error rate compared to the language model independent training proves the importance of an appropriate consideration of the language model in the MMI training procedure.

7. CONCLUSION

The paper presented a novel approach for discriminative language model dependent HMM parameter estimation for continuous speech recognition systems. It has been shown that useful estimates for the complete likelihood of the training utterances can be computed without the need for time consuming n-best or word-lattice recognition procedures. This considerably simplifies the MMIE procedure, especially for large vocabulary systems with a large amount of training data. In addition to that, the new approach is independent of the distribution of confusion on the training data so that it is capable of providing useful discriminative HMM parameter estimates in cases where the other approaches fail due to a misrepresentation of the real distribution of confusion.

Our future work will focus on the extension of the proposed HMM parameter estimation on more complex recognition systems that make use of triphones, different kinds of mixture sharing and clustered HMMs. Furthermore we will try to develop more profound procedures for the estimation of the state-to-state degrees of competition that will not only analyze the dictionary but also make use of the words' global probabilities of appearance and the language model-based probabilities of mixing up two specific words.

8. ACKNOWLEDGMENTS

This work was partly sponsored by the DFG (German Research Foundation) under contract number Ri 658/3-1. The authors thank Prof. Dr. Gerhard Rigoll from Duisburg University for his support and helpful discussions.

REFERENCES

[1] A. Nádas: "A Decision Theoretic Formulation of a Training Problem in Speech Recognition, and a Comparison of Training by Conditional versus Unconditional Maximum Likelihood" IEEE Transactions on ASSP, pages 814-817

[2] L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer: "Maximum Mutual Information of Hidden Markov Model Parameters for Speech Recognition" Proc. ICASSP'86, pages 49-52

[3] Merialdo B.: "Phonetic Recognition using Hidden Markov Models and Maximum Mutual Information Training" Proc. ICASSP'88, pages 111-114

[4] Yen-Lu Chow: "Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition using the N-Best Algorithm" Proc. ICASSP'90, pages 701-704

[5] Y. Normandin, R. Lacouture, R. Cardin: "MMIE Training for Large Vocabulary Continuous Speech Recognition" Proc. ICSLP'94, pages 1367-1370

[6] V. Valtchev, J.J. Odell, P.C. Woodland, S.J. Young "Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition" Proc. ICASSP'96, pages 605-608

[7] H. Bourlard: "Towards Increasing Speech Recognition Error Rates" Proc. EUROSPEECH'95, pages 883-894

[8] P.S. Gopalakrishnan, D. Kanevsky, A. Nádas, D. Nahamoo "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems" IEEE Transactions on Information Theory, Vol 37, pages 107-113, 1991