PAPER

Minimum Bayes Risk Estimation and Decoding in Large Vocabulary Continuous Speech Recognition

William BYRNE[†], ***Nonmember***

SUMMARY Minimum Bayes risk estimation and decoding strategies based on lattice segmentation techniques can be used to refine large vocabulary continuous speech recognition systems through the estimation of the parameters of the underlying hidden Markov models and through the identification of smaller recognition tasks which provides the opportunity to incorporate novel modeling and decoding procedures in LVCSR. These techniques are discussed in the context of going 'beyond HMMs', showing in particular that this process of subproblem identification makes it possible to train and apply small-domain binary pattern classifiers, such as Support Vector Machines, to large vocabulary continuous speech recognition.
key words:

1. Introduction

The use of statistical methods now dominates the theory and practice of automatic speech processing and recognition. These methods are predominantly based on Hidden Markov Models and their estimation and decoding algorithms. The underlying statistical model is a joint distribution $P(O, W; \theta)$ defined over an acoustic observation sequence O and a word sequence W [1], [2]. If the likelihood under this distribution can be computed, decision theory provides the form of the recognizer, typically : $\hat{W} = \text{argmax}_W P(W|O; \theta)$.

The goal of modeling is to support decoding, and assumptions about statistical independence are made so that the search can be carried out. The now-standard generative modeling assumption is that the joint probability distribution can be factored into acoustic and language models:

$$P(O, W; \theta) = P(O|W; \theta)P(W; \theta)$$

where both components are modeled directly through parametric distributions. In addition to leading to efficient search procedures, the parameters of these generative models can be estimated over the available training data, which typically consists of a transcribed acoustic training set $\{O, \bar{W}\}$. The conditional independence assumptions that separate the acoustic model from the language model make it possible to perform maximum likelihood (ML) estimation of the parameters of each; $\text{argmax}_\theta P(O, \bar{W}; \theta)$ decomposes into the two separate modeling problems $\text{argmax}_{\theta_A} P(O|\bar{W}; \theta_A)$ and $\text{argmax}_{\theta_L} P(\bar{W}; \theta_L)$.

In large vocabulary continuous speech recognition tasks, much is expected of both the acoustic and language models. Speakers are allowed to speak freely and say whatever they wish in whatever manner they chose. Consequently the language model must be able to assign likelihood to any word sequence that might be uttered by a speaker, and the acoustic model must be designed so that it can provide an acoustic score to any acoustic observation for any hypothesis allowed under the language model. In practice, the scope of the problem is defined by domain-specific collections of acoustic and language model training data. This leads to the all-important issue of generalization. Both the acoustic and language model components are trained with as much in-domain data as can be obtained, and this is done with the goal of building models that can generalize from the training data to unseen test data. Two sets of procedures play a crucial role in this. Baum Welch reestimation, and its Viterbi variants, make it possible to train on the large amounts of data needed to ensure generalization. Techniques such as triphone state clustering [3], Gaussian mixture splitting, and backoff strategies in n-gram language models, control the growth of model complexity during training [4]. Training using large amounts of data and controlling model complexity are crucial to achieve generalization.

The choice of model architecture is also driven by the need for generalization. The goal of speech recognition is to generate transcriptions in the writing system used by speakers of a language, and the natural units of transcription are orthographic. However the need for generalization usually demands the use of subword acoustic models. These allow the speech sounds of frequently occurring words and word combinations to be used to construct models that can be used for less frequently observed words and acoustic contexts. If it were possible to model word sequences directly, we would; it is arguably the need to construct hierarchical models based on sub-word components that leads to the inclusion of phonetics, syntax, and morphology in ASR.

Summarizing the discussion thus far, the modeling approach is to build a single system via maximum

Manuscript received January 1, 2003.
Manuscript revised January 1, 2003.
Final manuscript received January 1, 2003.
[†]The author is with the Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ U.K.

likelihood parameter estimation algorithms and perform recognition via MAP decoding procedures. Algorithms and architectures are favored that are consistent with the need to achieve generalization through ever larger training sets and more complex hierarchical models. This is why HMMs are pervasive. HMMs are often described by their topology and details, e.g. a triphone model with left-to-right transitions and Gaussian mixture observation distributions, and from this view HMMs are easy to criticize since their shortcomings as descriptive models are plainly obvious. But a more holistic (albeit still simplistic) view is that HMMs are also the conditional independence assumptions needed to implement efficient estimation and decoding procedures using dynamic programming; these are the assumptions that make it possible to perform estimation and decoding over large amounts of data. Abandoning these efficient algorithms to move 'beyond HMMs' therefore has risks. Research in new models and algorithms is typically done on small problems for various practical reasons. Within the statistical modeling framework, unless there is the promise of eventually performing estimation over large amounts of data, and thus achieve generalization, there is little hope of supplanting HMMs completely.

The discussion thus far ignores the extent to which mainstream ASR research has already moved beyond the original HMM formalism. Speaker adaptation [5] and discriminative training [6] are now ubiquitous, and system combination techniques such as ROVER [7] are applied whenever possible. These approaches are usually based on ML-trained HMMs, and now that they have entered the research mainstream, the extent to which they themselves go beyond the basic HMM framework is taken for granted. Adaptation and normalization will not be discussed in this paper other than to comment that they work to refine a carefully trained speaker independent ASR system and in the process reduce its ability to generalize to new speakers, that had been so carefully built into the system in the first place. MMI by definition departs somewhat from the generative ML framework. One view of the algorithm is that discriminative training sharpens up the basic HMMs, by using them to define a distribution $P(O|\bar{W};\theta)$ which is optimized over the training set. Of course, the underlying HMMs are retained and used to compute the posterior distribution using Bayes rule. A different view of the process is one in which the generative model is completely undone and a new model is constructed from its components which are estimated under a maximum likelihood criterion [8], [9]. MMI also departs from the HMM framework in that it implicitly discards training data: training utterances do not contribute to MMI when the most likely hypothesis is both correct but strongly dominant. This occurs implicitly and over whole utterances: the entire sentence hypothesis has to be correct and dominant before this effect is observed. However despite these departures, the HMM model architecture remains and standard MAP decoding can still be performed with the 'proper' integration of the acoustic and language models. But the point remains that the MMI formulation requires abandoning either the ML estimation or the generative modeling framework. A similar point can be made with respect to system combination techniques which consider replacing the most likely hypotheses generated by one model with a synthesis of less likely hypotheses from multiple models. This essentially abandons the original idea that there is a single optimum model that can be trained and used in ASR.

In summary, in one way or another, speaker adaptation, MMI, and system combination depart from the maximum likelihood generative modeling framework based on a single general model estimated with all available training data. What remains from the underlying modeling framework is that the entirety of the training data continues to define the scope of the recognition problem. The overall goal is still to train sets of models capable of tackling the entire recognition task. The issue that will be discussed in this paper is whether the need to rely on models capable of this extreme generality works against the ability to discriminate. Models and algorithms selected based on how well they work well 'in general' may fail in particular instances; conversely, modeling approaches which may be well-suited for making certain specific distinctions are difficult to incorporate. Furthermore, it can be difficult even to identify the instances in which the general purpose models are failing and may need to be helped.

The remainder of the paper will discuss modeling techniques meant to identify small vocabulary recognition problems within LVCSR tasks and to provide a theoretical and practical framework for incorporating novel modeling techniques into LVCSR. This approach merges the techniques of system combination and discriminative training. One of the key goals of this work is to make it possible to retain the good performance of state of the art HMM recognition systems when trying out new ideas. The approach allows the modeler to specify the size of the problems to be tackled. These can then be solved with one set of models or with specialized models, and these models can be of any complexity - even very simple binary classifiers can be used. The point is that the complexity of the models being developed need not be driven by the entire LVCSR problem. Of course, there is a trade-off. As will be discussed, the gains with respect to the LVCSR baseline inevitably decrease as smaller and smaller subproblems are singled out. There are several practical benefits. The first is that it is possible to assess the value of a particular novel modeling scheme. Suppose a novel classifier is proposed to make specific types of distinctions, say between certain classes of words. This procedure can determine how many times the baseline system itself

makes errors over those classes, and can also provide an indication of how much gain over the baseline the novel classifier can optimistically be expected to yield. A second consequence is that since solving small problems tends only to yield small gains, statistical significance becomes an issue. Unusually large test sets may be required to ensure that small gains can be trusted as an indication that improvements are real.

Another benefit of this approach relative to developing new techniques is that by focusing on small, independent problems the size of the training data relevant to each problem can be greatly reduced. This has multiple benefits. Training data is selected that is particularly relevant to the classification problem: the specialized classifiers can be estimated over regions of the training set where the baseline HMM itself fails to make the correct distinction. This leads to more specialized models. As a practical consideration, this also makes it easier to train novel architectures using algorithms that otherwise might be unable to make use of all the entire training set. The next section discusses how lattice segmentation techniques developed originally Minimum Bayes Risk decoding procedures can be used to support the development of novel acoustic modeling and recognition within LVCSR tasks.

2. Lattice Segmentation and the Selection of Recognition Subproblems

Minimum Bayes Risk decoders [10]–[12] find a sentence hypothesis with the least expected error under a loss function as

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmin}} \sum_{W' \in \mathcal{W}} l(W, W') P(W'|O;\theta). \qquad (1)$$

The motivation for the approach is that $l(\cdot,\cdot)$ describes a task performance metric, such as the Levenshtein distance associated with Word Error Rate, so that models estimated under some other criterion can be 'tuned' to a task by modifying the decoding procedure [13]. Although this formulation may appear overly formal, it can be shown that ASR rescoring and system combination techniques such as ROVER [7] can be formulated as MBR decoding procedures [14].

This is a search problem in which $\mathcal{W}$ are N-Best lists or lattices that incorporate $P(W'|O)$ as a posterior distribution on word strings. For each hypothesis $W \in \mathcal{W}$, the risk is computed

$$R(W,\ \mathcal{W};\ \theta) = \sum_{W' \in \mathcal{W}} l(W, W')\ P(W'|O;\theta) \qquad (2)$$

and the hypothesis with the least risk is selected

$$\hat{W} = \underset{W \in \mathcal{W}}{\operatorname{argmin}}\, R(W,\ \mathcal{W};\ \theta)\ . \qquad (3)$$

Efficient algorithms have been developed to compute

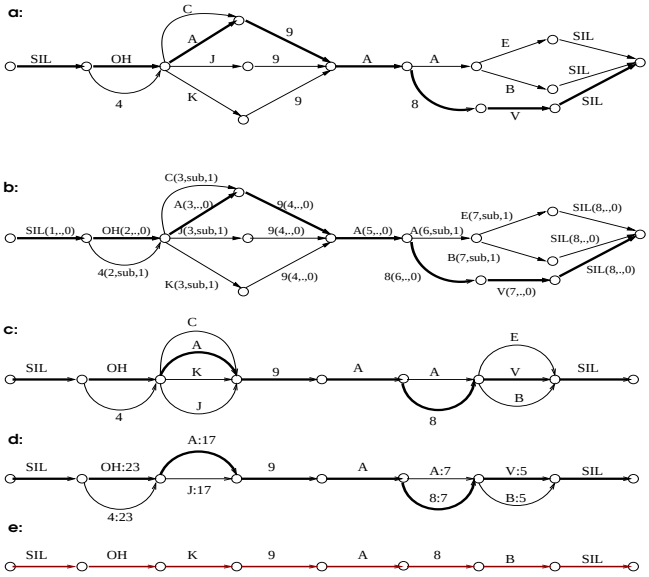

Fig. 1 Lattice Segmentation. *a*: First-pass lattice with MAP hypothesis in bold; *b*: Lattice to string alignment under the Levenshtein distance to the MAP path; *c*: Segment sets that arise from the alignment; *d*: Refined Search Space consisting of binary segment sets found by pruning segment sets produced in Step *c*. Word hypotheses are tagged so specialized models can be used in lattice rescoring. *e:* Correct transcription used to measured the quality of the selected pairs.

the risk of a hypothesis W under the Levenshtein loss function [15]. Since it is straightforward to compute likelihoods such as $P(W'|O;\theta)$ over ASR lattices, the key is an efficient *lattice-to-string alignment* algorithm to find and represent the costs $l(W, W')$ for all W' in any lattice $\mathcal{W}$. In addition to making the MBR search problem feasible under certain loss functions, we use these algorithms to segment the first-pass ASR lattices. The alignment algorithm is reviewed briefly in the following section within the context of identifying recognition subproblems.

2.1 Recognition Subproblem Selection and Analysis

We now summarize recent studies of lattice segmentation in discriminative training [16] and applications of Support Vector Machines in LVCSR rescoring [17], [18]. Lattice segmentation converts a first-pass lattice into a sequence of smaller sub-lattices through a Levenshtein alignment of the lattice to a reference path [15]. Here, test set lattices (Fig. 1, a) are aligned to the primary hypothesis so that word sequences from the lattice are aligned with words in the primary hypothesis (Fig. 1, b). This produces *segment sets*, which are groups of substrings from the lattice identified as alternatives to words in the primary hypothesis (Fig. 1, c).

These segment sets define the LVCSR subproblems that can be considered in subsequent decoding passes. However, not all segment sets are of equal value in improving the baseline. In some segment sets, the refer-

ence (MAP) hypothesis is likely to be correct - ideally these should be left alone. In others, the segment set does not contain the truth as one of the alternatives - these should be ignored. The ideal problems to attack are those defined by segment sets within which the MAP path is wrong and the correct path is available as an alternative. An additional concern is that, if specialized models are to be trained, the selected subproblems should also appear frequently enough in training data so that models can be reliably estimated. Ideally, this should all be carried out with a minimum of supervision.

An unsupervised approach based on the posterior scores over lattice segment sets has been developed [18]. Segment sets are identified as described by lattice-to-string alignment under the Levenshtein distance, but the joint acoustic and language model scores from the lattice are retained and can be used to define posterior distributions over the hypotheses in the segment sets, and by extension over the portions of the segment sets themselves. This allows us to prune the segment sets to finally obtain confusion pairs (Fig. 1, d).

The selection process summarized above and depicted in Figure 1 is referred to as *lattice pinching*. It consists of consecutive alignment, segmentation, and pruning steps to identify segment sets. As mentioned, the process is effective only to the extent that it identifies weaknesses in the primary hypothesis and offers useful alternative hypotheses.

We have evaluated our approach in the MALACH spontaneous Czech conversational domain [19]. The system consists of speaker independent continuous mixture density, tied state, cross-word, gender-independent, triphone HMMs trained with HTK using 65 hours of transcribed speech (24065 utterances). The speech was parameterized into 39-dimensional, MFCC coefficients, with delta and acceleration coefficients. The AT&T Large Vocabulary Decoder was used to generate lattices over the training and test sets with a bigram language model based on a 83000 word vocabulary. Lattice-based MMI [6], [20] was performed. The test set studied in this section consisted of approx. 8400 utterances spoken by ten held-out speakers (approx. 25 hours of speech). Unsupervised MLLR transforms for each of the test-set speakers were estimated on a 1000 utterance subset of the test set. The baseline system produced a test set lattices with WER of 45.6% and 22.3% Lattice Error Rate (LER).

We now analyze the performance of the lattice pinching procedure. Referring to Table 1, we can see that the oracle Lattice Error Rate increases due to pinching and pruning the test set lattices. This is the inevitable limitation of this approach: focusing on small decoding problems with the larger ASR problem inevitably limits the gains available over the baseline system. We first prune the original lattices so that lattice-to-string alignment is more easily done; this increases the LER from 22.3% to 27.3%. Subsequent pruning as reported in Table 1 is performed relative to the link posterior scores of the MAP alternatives within each segment set; high pruning thresholds discard low confidence alternative word hypotheses. Note that for pruning thresholds above 1.0, there tends to be only two hypotheses per segment set. LER continues to increase as pruning increases, since many segment sets are completely pruned away, leaving only the MAP hypothesis.

Pruning Threshold	LER	Avg. # Hyps. / Segment Set	Segment Sets	
			Types	Tokens
0.00	27.3	11.65	94029	1393099
0.05	35.3	2.82	49837	212852
0.10	37.9	2.35	35278	134252
0.20	41.1	2.06	17132	63267
0.30	43.2	2.00	7288	26913
0.40	44.7	2.00	2249	7930
0.50	45.6	-	0	0

Table 1 Segment Set Analysis Over A 25 Hour Test Set. The average number of hypotheses per segment set, number of distinct segment sets, and total number of segment sets after posterior-based pruning as described in Section 2.

Based on these results we selected a pruning theshold of 0.3, since we are interested in finding binary classification problems. All non-binary segment sets are pruned back to the MAP hypothesis. We also restricted our attention to those confusion pairs observed in the test data at least 100 times. The reason for doing this is that we wished to be able to measure the performance of the individual binary classifiers we trained; for each classifier we have a test set of at least 100 instances. This is not a necessary limitation, however, and if we were more interested in overall WER reduction and less interested in assessing the quality of the novel classifiers, we would have chosen more pairs. Referring to Figure 1 d, only these frequently occurring confusion pairs are retained, and all others are pruned back to the baseline hypothesis.

The process so far is unsupervised. To further analyze the confusion pairs, we Levenshtein-align the pinched lattices (Fig 1 d) to the truth (Fig 1 e). We first count the number of Confusion Pair Errors (CPERR), defined as confusion pairs that do not contain the truth. For example, in Fig. 1 d, (A:17, J:17) is classified as CPERR since it does not contain the true word 'K'; the other sets are classified as Confusion Pair Oracle Correct (CPOC). Within the CPOC segments we can distinguish those in which the MAP path agrees with the oracle path (MAPCOR) and those in which the MAP path is in error (MAPERR). In Fig. 1, d the pair (V:5, B:5) is classified as MAPERR, and the pairs (OH:23, 4:23) and (A:7, 8:7) are MAPCOR; both these sets are CPOC.

We further process the pinched lattices constructed from the frequently occurring confusion pairs. We renormalize these lattices to define the posterior distri-

Pruning Threshold	#CPOC/ #CPERR	#MAPERR/ #MAPCOR	Segment Sets	
			Types	Tokens
0.00	14.30	0.24	22	7324
0.05	4.7	0.64	26	8022
0.10	3.3	0.92	26	6860
0.20	3.2	1.17	17	3831
0.30	4.2	1.15	6	1405
0.40	11.0	1.04	2	337
0.50	-	-	0	0

Table 2 Ratio of #CPOC/#CPERR segments and #MAPERR/#MAPCOR segments for the confusion pairs observed at least 100 times in the 25 hour test set.

bution over these binary confusion pairs, and again apply a posterior-based pruning to these instances of the confusion pairs. The results are as reported in Table 2. At a pruning threshold of 0.4, the surviving confusion pairs are high quality: the CPERR pairs occur far less frequently than CPOC pairs; and within these the the MAPERR count is about equal to the MAPCOR count, so about half the MAP hypotheses are incorrect. Unfortunately, there are only two distinct confusion pairs and pruning eliminates all but 337 instances of them. In the subsequent experiments, we prune at a threshold of 0.1. At this level, we still have three times as many CPOC pairs as CPOERR, the system is still making errors roughly half the time (MAPERR $\approx$ MAPCOR), and we have a diverse test set of 6860 observations of 26 distinct confusion pairs. Since we are specifically interested in acoustic modeling, we discard confusion pairs consisting of homonyms only; this leaves us with 2991 instances of 21 confusion pairs.

2.2 Support Vector Machines for LVCSR Subproblems

We now review our approach to building SVMs for these confusion pairs identified by the LVCSR system. We begin by training special purpose, whole-word HMMs for the words in the confusion pairs; these will complement the cross-word triphone acoustic HMMs used in the baseline LVCSR system. We next clone these whole-word models for the confusion pairs, e.g. the model for the word 'A' is replicated so that A:17 and A:7 are two different whole-word HMMs. For example, to train the models for the confusion pair (A:7, 8:7), an acoustic training subset is created by extracting all the acoustic segments for 'A' and '8' from the training data. MMI is then used to further train the models A:7 and 8:7 over this training subset. This allows to accumulate statistics over different recognition problems and thus create specialized decoders from specialized training sets.

To train SVMs for the binary confusion pairs, we use the score-space approach developed by Smith and Gales [17], [21], [22]. Statistics derived from the HMM likelihoods are used to transform a variable-length sequence into a static fixed-dimensional representation which can be used in SVM training and classification; the dimension of the features to be classified is derived from the number of parameters in each whole-word HMM and not from the length of the speech segment.

Our choice of this approach was driven mainly by expedience; using MMI-trained whole word models to extract statistics may indeed confer modeling advantages in that the statistics are generated by models tuned to the specific binary classification problem. However, alternative approaches based on monophone models, rather than word models, can also be used (M. Gales, personal communication), which may be particularly useful if data sparsity is an issue. This approach avoids the need to generate lattices over the training set; MMI over confusion pairs can be performed simply by using two version of the transcriptions that differ by the word in question. All training statistics over confusion pairs can be obtained using Baum Welch. Following the selection criteria as explained, we chose 21 confusion pairs to study. On average, 0.58 hours of speech was selected as a training set for each confusion pair. The *Gini*SVM toolkit [23] was used to train classifiers based on mean and likelihood-ratio scores derived from the MMI trained word HMMs; details are provided in [18].

The objective is to apply these specialized SVM classifiers in an unsupervised manner to resolve binary word confusions found in lattices generated by the baseline LVCSR system. As discussed, selecting the test set to include at least 100 instances of each confusion pair allowed us to make meaningful comparisons of the performance of the SVM trained for each pair to the MAP baseline performance. We found that performance relative to the MAP baseline is mixed; there are not consistent improvements due to using the SVM alone. However the lattice-to-string alignment procedure is carefully designed so that the complete original paths and their likelihoods are retained throughout pinching and pruning. We can thus derive reliable posteriors over the remaining baseline hypotheses and perform hypothesis combination. To combine the SVM and MAP hypotheses, a posterior distribution over the SVM decisions was estimated by logistic regression [23].

This associates a confidence (estimated likelihood of being correct) with each SVM choice. For a particular instance of a confusion pair with words (w_1, w_2), let $p_h(w)$ be the MAP posterior over the pinched lattices, and $p_s(w)$ be the SVM confidence in each decision. A simple linear interpolation with weighting λ gives a combined likelihood over the word pair. With $\lambda = 0.5$, the error count decreases in 18 of the 21 pairs. The influence of these reductions on the overall WER over the complete test set is necessarily limited, as already discussed. Under the MAP-SVM combination system, the baseline MAP WER is reduced from 45.6% to 45.5%. However small, these gains are statistically significant

and stable with respect to λ: we obtained this performance improvement for $\lambda = 0.4, 0.5, 0.6$, and 0.7, and in all instances the significance test p-values [24] were less than 0.001, so we can claim improvement with great confidence. These gains are small from the point of view of improving an LVCSR system. However, from the point of view of validating our proposed SVM modeling approach, we can simultaneously claim significant improvements on both a small vocabulary recognition problem and an LVCSR system incorporating MMI, MLLR speaker adaptation and other state of the art techniques. We also offer this example to show that, if the underlying experiments are carefully constructed, even novel techniques can be applied to LVCSR problems even in the early stages of their development.

3. Minimum Bayes Risk Parameter Estimation

Risk-based parameter estimation procedures attempt to minimize the expected risk over the training set. Given a transcribed database $\{\bar{W},O\}$, the estimation objective is

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, R(\bar{W}, \mathcal{W}; \theta) \tag{4}$$

where

$$R(\bar{W}, \mathcal{W}; \theta) = \sum_{W' \in \mathcal{W}} l(\bar{W}, W') P(W'|O;\theta)\,.$$

$\mathcal{W}$ is taken to be a set of hypotheses being considered as alternatives to the truth $\bar{W}$, and we assume that their distance to the correct transcription $\bar{W}$ is measured by the string edit or Levenstein distance $l(\bar{W}, W)$ associated with Word Error Rate (WER).

The estimation problem hinges on determining the contribution to the overall risk of each hypothesis W' in $\mathcal{W}$. If a relatively likely hypothesis W' differs significantly from $\bar{W}$ as measured by $l(\bar{W}, W')$, it will add substantially to the overall risk. Thus a successful estimation strategy is one that moves probability mass towards those hypotheses that are close to the reference while reducing the likelihood of those hypotheses that are far away.

While the loss function $l(\bar{W}, W')$ and the likelihood under the current model parameters dominate the overall risk, $\mathcal{W}$ also plays an important role in that, since the risk is measured over $\mathcal{W}$, it must provide a representative sample of hypotheses that are both likely and error-full. If $\mathcal{W}$ is not chosen well, the risk measurements will be biased. In particular there is a danger of underestimating the risk.

Kaiser *et al.* [25] have shown how the Extended Baum Welch [26] algorithm can be applied to obtain a risk-minimizing variant of the MMI re-estimation procedure for the parameters of state-dependent Gaussian observation distributions, as shown for the Gaussian means

$$\bar{\mu}_s = \frac{\sum\limits_{W' \in \mathcal{W}} K(W', \mathcal{W}; \theta) \sum_\tau \gamma_s(\tau; W') o(\tau) + D_s \mu_s}{\sum\limits_{W' \in \mathcal{W}} K(W', \mathcal{W}; \theta) \sum_\tau \gamma_s(\tau; W') + D_s} \tag{5}$$

where $K(W', \mathcal{W}; \theta)$ is computed as

$$[\sum_{W'' \in \mathcal{W}} P(W''|O;\theta) l(\bar{W}, W'') - l(\bar{W}, W')]\, P(W'|O;\theta).$$

Here, the $\gamma_s(t; W')$ are the occupancy statistics associated with the s^{th} Gaussian as found by the Forward-Backward algorithm with respect to the hypothesis W'. The D_s are constants that arise directly from the formulation of the Extended Baum Welch algorithm. As discussed by Gopalakrishnan *et al.* [26] these constants are set to some sufficiently large value to ensure increase in the training objective function. We follow the approach described by Woodland and Povey [6] in using Gaussian-specific constants to ensure that convergence is not overly slow.

The quantity $K(W', \mathcal{W}; \theta)$ determines the sensitivity of the overall risk to the contribution of each hypothesis W'. The relationship to the MMI procedure is readily apparent. For the 0/1 loss function, Equation 5 reduces to the usual MMI update relationship as derived as an Extended Baum Welch procedure. This approach is also reminiscent of Minimum Classification Error training [27] in that the aim is to improve the modeling of correct hypotheses relative to erroneous hypotheses. The difference is that, rather than select a single most likely incorrect hypothesis (for example) as the training alternative, a very large collection of hypotheses are considered under a weighting provided by the $K(W', \mathcal{W}; \theta)$ statistics. This form of risk minimization also differs from MCE in the use of the Extended Baum Welch algorithm which provides a closed form iterative parameter estimation procedure as an alternative to gradient based searches.

As was done with MMI estimation in small vocabulary recognition tasks by Normandin [28], Kaiser *et al.* [25] demonstrated that the statistics needed to perform minimum Bayes risk estimation of HMM parameters can be found by the Forward-Backward procedure over N-Best lists of competing hypotheses. However, Equation 5 is not easily implemented over lattices, which limits its usefulness in LVCSR tasks, where the N-Best lists would have to be exceedingly deep to contain a significant portion of the likely, erroneous hypothesis. The difficulty is that the term $l(\bar{W}, W')$ in $K(W', \mathcal{W}; \theta)$ must be found for all $W' \in \mathcal{W}$. If $l(\bar{W}, W')$ was a likelihood based quantity, computation would be straightforward, but since the loss is based on the Levenshtein distance, the computation must be done for each complete path through the lattice. Fortuitously, this is exactly the computation that is performed by lattice-to-string alignment. That procedure

produces a new lattice in which each path is marked with the information needed to align it to the reference string $\bar{W}$. This information is preserved through lattice pinching and pruning, and results in an *induced loss function* $l_I(\bar{W}, W')$. We refer to it as induced by the lattice pinching because the optimum alignment between $\bar{W}$ and W' may be not allowed within the pinched lattice $\tilde{\mathcal{W}}$, e.g. Fig 1a allows more diverse alignments than Fig 1d. Hence l_I only approximates l. But it allows us to perform estimation under the following criterion

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, R_I(\bar{W}, \mathcal{W}; \theta) \tag{6}$$

where

$$R_I(\bar{W}, \mathcal{W}; \theta) = \sum_{W' \in \tilde{\mathcal{W}}} l_I(\bar{W}, W') P(W'|O; \theta)\,.$$

By controlling the amount of pruning performed after lattice alignment, the size of the pinched lattice $\tilde{\mathcal{W}}$ can be restricted. This allows us to expand the pinched lattice into a reasonable-sized N-Best list, with alignment costs $l_I(\bar{W}, W')$, so that the update procedure of Kasier *et al.* can be performed in LVCSR. We refer to this procedure as *Pinched Lattice Minimum Bayes Risk Discriminative Training* (PLMBRDT). Over the MALACH Czech ASR training set described earlier, lattices were generated using the baseline MMI system and lattice-to-string alignment with respect to the reference transcription was performed. Lattice segmentation was done, focusing again on the (word level) error pairs that occurred 100 times or more; this yielded 117 confusion pairs observed a total of 48,302 times in training.

For every word in the reference hypotheses there was an average of 0.13 confusion pairs. As a result, after pinching, not all lattices contained confusion pairs; put another way, the induced risk over these lattices was zero. These utterances were therefore discarded from training, reducing the training set from 62.4 hours to 52.4 hours of speech. The remaining pinched lattices were expanded into N-Best lists, with an average depth of 36.5 hypotheses. On a 2 hour subset of the full 25 hour test set, the ML baseline performance of the system was 44.3% WER. This was reduced to 41.5% by five MMI iterations, and was further reduced to 41.1% by five PLMBRDT iterations (with p-value 0.013 relative to the MMI hypotheses).

4. Discussion

Lattice segmentation and Pinched Lattice Minimum Bayes Risk Discriminative Training have been discussed as two procedures that are based on HMMs and at the same time depart significantly from the original HMM framework. These techniques evolved from techniques intended to minimize risk, rather than maximize likelihood, in ASR decoding. As modeling procedures, estimation and decoding are obviously distinct tasks. However within this formulation the two problems are linked through the calculation of empirical risk with respect to a set of underlying HMMs. Since there is this common need to evaluate empirical risk, the distinct tasks of MBR decoding and estimation can be carried out efficiently using shared techniques of risk calculation.

Lattice segmentation can be used to define recognition subproblems within LVCSR tasks. 'Defining a subproblem' implies more than just selecting a small recognition task such as a particular binary word choice: we suggest a procedure to identify which particular confusion sets in the test set should be selected as candidates for correction; we suggest how training data might be collected to train models to solve these problems; and we show how the final simple classifier can be reincorporated into the overall large vocabulary ASR problem. The overall framework is still under development, but our studies of Support Vector Machines demonstrate that novel techniques can be applied to LVCSR problems even in the early stages of their development. There are also interesting issues in training and test set sizes. For reasons of statistical significance, large test sets seem inevitable in working on small LVCSR subproblems. On the other hand, the training set is reduced in these LVCSR subproblems, and in PLMBRDT, as subsets of the training set are selected to solve specific problems. Selecting small training sets for LVCSR subproblems offers practical advantages in developing new techniques, and it may in addition provide modeling advantage in that the sets contain exactly those training instances over which the baseline HMM is weak. The determination of training set size can be made more rigorous, in that it follows from the training objective function, through the use of lattice pinching and pruning to define an induced loss function over the training set lattices. Training set utterances which incur no risk under the induced loss function are simply not considered by PLMBRDT, with a resulting reduction in the original training set. Apart from these somewhat abstract considerations, the overall approach offers a route for the the development of novel modeling and decoding procedures to improve HMM-based ASR systems.

Acknowledgments Many thanks to M. Gales for comments and suggestions. My colleagues V. Goel, S. Kumar, V. Doumpiotis, S. Tsakalidis, S. Chakrabartty, and V. Venkataramani deserve the credit for the work and results discussed here.

References

[1] F. Jelinek, L. Bahl, and R. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," IEEE Transactions Information Theory, vol.21, pp.250–256, 1975.

[2] F. Jelinek, "Continuous speech recognition by statistical methods," Proceedings of the IEEE, vol.64, pp.32–556, 1976.

[3] J. Odell, THE USE OF CONTEXT IN LARGE VOCABULARY SPEECH RECOGNITION, Ph.D. thesis, University of Cambridge, 1995.
[4] S. Young *et al.*, The HTK Book Version 3.0, Cambridge University, 2000.
[5] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol.9, pp.171–185, April 1995.
[6] P. Woodland and D. Povey, "Large scale discriminative training for speech recognition," Proc. ITW ASR, ISCA, 2000.
[7] J. Fiscus, "A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," IEEE ASRU Workshop, pp.347–352, 1997.
[8] Y. Ephraim and L.R. Rabiner, "On the relations between modeling approaches for information sources," IEEE Transactions Information Theory, vol.36, no.2, pp.372–380, March 1990.
[9] W. Byrne, "Information geometry and maximum likelihood criteria," Conference on Information Sciences and Systems, Princeton, NJ, 1996.
[10] A. Stolcke, Y. Konig, and M. Weintraub, "Explicit word error minimization in N-Best list rescoring," Proceedings of the European Conference on Speech Technology, pp.163–166, Rhodes, Greece, 1997.
[11] V. Goel, S. Kumar, and W. Byrne, "Confidence based lattice segmentation and minimum Bayes-risk decoding of lattice segments," Proceedings of the European Conference on Speech Technology, Aalborg, Denmark, 2001.
[12] S. Kumar and W. Byrne, "Risk based lattice cutting for segmental minimum Bayes-risk decoding," ICSLP 2002, Denver, CO, USA, pp.373–376, 2002.
[13] V. Goel and W. Byrne, "Minimum Bayes-Risk automatic speech recognition," Computer Speech and Language, vol.14(2), pp.115–135, 2000.
[14] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," in Pattern Recognition in Speech and Language Processing, ed. W. Chou and B.H. Juang, CRC Press, 2003.
[15] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum Bayes-risk decoding for automatic speech recognition," IEEE Transactions on Speech and Audio Processing, pp.234–249, May 2004.
[16] V. Doumpiotis and W. Byrne, "Lattice segmentation and minimum Bayes risk discriminative training for large vocabulary continuous speech recognition," Speech Communication. Accepted. To appear.
[17] V. Venkataramani, S. Chakrabartty, and W. Byrne, "*Gini*support vector machines for segmental minimum Bayes risk decoding of continuous speech," Computer Speech and Language. Accepted. To appear.
[18] V. Venkataramani and W. Byrne, "Lattice segmentation and support vector machines for large vocabulary continuous speech recognition," Proc. ICASSP, 2005.
[19] W. Byrne *et al.*, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," IEEE Transactions Speech and Audio Proc., pp.420 – 435, July, 2004.
[20] V. Doumpiotis, S. Tsakalidis, and W. Byrne, "Lattice segmentation and minimum Bayes risk discriminative training," Proceedings of the European Conference on Speech Technology, 2003.
[21] N. Smith and M. Gales, "Using SVMs and discriminative models for speech recognition," Proc. ICASSP, 2002.
[22] V. Venkataramani and W. Byrne, "Support vector machines for segmental minimum Bayes risk decoding of continuous speech," ASRU, 2003.
[23] S. Chakrabartty and G. Cauwenberghs, "Forward decoding kernel machines: A hybrid HMM/SVM approach to sequence recognition," Proc. SVM 2002, Lecture Notes in Computer Science, MIT Press, 2002. Toolkit: http://bach.ece.jhu.edu/svm/ginisvm/.
[24] D. Pallett *et al.*, "Tools for the analysis of benchmark speech recognition tests.," Proc. ICASSP, pp.97–100, 1990.
[25] J. Kaiser, B. Horvat, and Z. Kacic, "Overall risk criterion estimation of hidden Markov model parameters," Speech Communication, vol.38, no.3–4, pp.383–398, 2002.
[26] P.S. Gopalakrishnan, D. Kanevsky, A. Nádas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," IEEE Transactions Information Theory, pp.107–113, Jan. 1991.
[27] B.H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," IEEE Transactions on Speech and Audio Processing, vol.40, no.12, pp.3043–3054, Dec. 1992.
[28] Y. Normandin, "Maximum mutual information estimation of hidden Markov models," in Automatic Speech and Speaker Recognition, ed. C.H. Lee, F.K. Soong, and K.K. Paliwal, pp.57–82, Kluwer Academic Publishers, 1996.