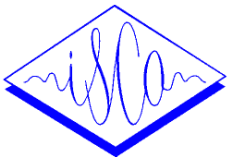

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

MINIMUM RISK ACOUSTIC CLUSTERING FOR MULTILINGUAL ACOUSTIC MODEL COMBINATION

Dimitra Vergyri *Stavros Tsakalidis* *William Byrne* *

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218-2686
{*byrne,dverg,stavros*}*@clsp.jhu.edu*

ABSTRACT

In this paper we describe procedures for combining multiple acoustic models, obtained using training corpora from different languages, in order to improve ASR performance in languages for which large amounts of training data are not available. We treat these models as multiple sources of information whose scores are combined in a log-linear model to compute the hypothesis likelihood. The model combination can either be performed in a static way, with constant combination weights, or in a dynamic way, with parameters that can vary for different segments of a hypothesis. The aim is to optimize the parameters so as to achieve minimum word error rate. In order to achieve robust parameter estimation in the dynamic combination case, the parameters are defined to be piecewise constant on different phonetic classes that form a partition of the space of hypothesis segments. The partition is defined, using phonological knowledge, on segments that correspond to hypothesized phones. We examine different ways to define such a partition, including an automatic approach that gives a binary tree structured partition which tries to achieve the minimum WER with the minimum number of classes.

1. INTRODUCTION

Multilingual acoustic modeling is motivated by the need for speech recognizers in languages and dialects for which acoustic training data is not available in large quantities. The goal of multilingual acoustic modeling is to improve ASR performance in a target language by borrowing models and data from other languages. Previous work has largely focused on the task of building a single set of target language acoustic models using data from different source languages. For example, cross-lingual phonetic mappings between the source and target languages can be created so that a pool of multilingual data can be used to train a single set of multilingual acoustic models [1]. Alternatively, well-trained acoustic models from a source language can be adapted to the target language using standard acoustic adaptation techniques [2].

An alternative methodology has recently been presented [3, 4] that extends the above procedures by using discriminative model combination techniques (DMC) [6]. Rather than merging source language acoustic data to train a single system, this technique produces a likelihood score for a recognition hypothesis by combining the scores produced for the hypothesis segments by multiple, independent, source language ASR systems. First, mappings are derived from the phones of the source languages to those of the target language. This allows mapping source language acoustic models onto target language speech. The mapped source models are then adapted by MLLR/MAP on a small amount of transcribed target language acoustic data. The best possible monolingual target language system is trained using the available data in the target language. Using this monolingual system, N-Best lists are produced for the test set. Finally, these N-Best lists are rescored by the adapted source language ASR systems and these scores are combined to produce a new likelihood for each hypothesis. The new best hypothesis is chosen based on this rescoring.

In this paper we discuss refinements of this last, crucial step, i.e. the optimal combination of the available acoustic models from different languages. We treat these source language acoustic models as independent information sources that provide separate, independent scores for each hypothesis, which are combined in a log-linear (GLM) model. The weights of this log-linear combination can either be *static* or *dynamic*. Static weights are optimally determined for each source language system (on held out data) and are held constant for all hypothesized segments. Dynamic weights allow the different language systems to contribute variably, as a function of the hypothesis. We compare the static and the dynamic combination approaches and discuss dynamic weighting algorithms in detail.

2. LOG-LINEAR COMBINATION OF MULTIPLE INFORMATION SOURCES

The DMC approach was used in previous work [3, 4] to combine the multiple monolingual ASR systems. DMC aims at an optimal integration of all possible information sources (in our case the acoustic models from multiple languages and a language model for the target language) into one log-linear posterior probability distribution. The weights of the scores for each information source can either be constant (static combination) or variable across the hypothesis (dynamic combination).

2.1. Static combination

Assume we have available m knowledge sources from which we can obtain scores for a hypothesis $\mathbf{W}$. We define the probability $P(\mathbf{W}|\mathcal{I})$, where $\mathcal{I}$ is the information available to the sources, which includes but is not limited to the acoustic vector $\mathbf{A}$, to be an

*This work was supported by the National Science Foundation under Grant No. #IIS-9810517. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

exponential combination of the scores $S_i(\mathbf{W}|\mathcal{I})$ obtained using each of the information sources. In the following we will denote these scores just with $S_i(\mathbf{W})$:

$$P(\mathbf{W}|\mathcal{I}) = \frac{\mathbf{1}}{\mathbf{Z}(\Lambda,\mathcal{I})} \prod_{i=1}^{\mathbf{m}} \mathbf{S_i}(\mathbf{W})^{\lambda_\mathbf{i}} \quad (1)$$

where $Z(\Lambda,\mathcal{I})$ is a normalization factor so that the probabilities for all $\mathbf{W} \in \mathcal{H}$ add to one.

In the experiments the scores $S_i(\mathbf{W})$ we consider are the Czech acoustic model score (A_{cz}), the English acoustic model score (A_{en}) and a score from a Czech bigram language model.

2.2. Dynamic combination

Here we combine the scores from the available information sources dynamically, within the simple form of an exponential model, by weighting each of the scores with different exponents, for different segments of a hypothesis. This allows the source language models to contribute variably depending on the hypothesis. For instance, one language may approximate a set of target language phones particularly well, while other target language phones are not modeled well at all. Thus we want the weights with which the scores from each language are combined to depend on the identity of the phones in the different hypotheses.

We denote as w_{ij} the jth segment for hypothesis W corresponding to the segmentation associated with the ith source [1]. Then we can define:

$$P(\mathbf{W}|\mathcal{I}) = \frac{\mathbf{1}}{\mathbf{Z}(\Lambda,\mathcal{I})} \prod_{\mathbf{i=1}}^{\mathbf{m}} \prod_{\mathbf{j=1}}^{\mathbf{K_i(W)}} \mathbf{S_i}(\mathbf{w_{ij}})^{\lambda_\mathbf{i}(\mathbf{w_{ij}})} \quad (2)$$

where the exponent for the score of each segment is a function of the segment. The robust optimization of the parameters $\lambda(\cdot)$ is the focus of the remainder of this paper.

Since we would like to have a small number of parameters $\lambda(\cdot)$ to optimize, we define a partition function for each source language $F_i : (\mathcal{W}) \rightarrow \{1,...,C_i\}$ that maps the space of (w_{ij}) into a small number of discrete classes. Then we can define for a hypothesis $\mathbf{W}$ the score under the ith independent source:

$$S_{ic}(\mathbf{W}) = \prod_{\mathbf{w_{ij}} \in \mathbf{W}:\mathbf{F_i}(\mathbf{w_{ij}},\mathcal{I})=\mathbf{c}} \mathbf{S_i}(\mathbf{w_{ij}}) \quad (3)$$

We can rewrite (2) as:

$$P(\mathbf{W}|\mathcal{I}) = \frac{\mathbf{1}}{\mathbf{Z}(\Lambda,\mathcal{I})} \prod_{\mathbf{i=1}}^{\mathbf{m}} \prod_{\mathbf{c=1}}^{\mathbf{C_i}} \mathbf{S_{ic}}(\mathbf{W})^{\lambda_\mathbf{i}(\mathbf{c})} \quad (4)$$

where we have grouped the scores for the segments of the hypothesis according to the class of each segment.

[1] Each source model can have a different time alignment for the same hypothesis and thus the segments might correspond to different acoustic intervals.

2.3. Optimization issues

The above defined model of equation (4) is used to rescore the N-Best lists and choose the MAP candidate. We train the parameters $\lambda(\cdot)$ in formulas (1) and (4) so that the empirical word error count induced by the model is minimized. Since the objective function is not smooth, gradient descend techniques are not appropriate for estimation. We use the simplex downhill method known as amoeba search [7] to minimize the word errors on a held out set [8]. We also consider the approach of Beyerlein [6] which minimizes a smooth function that approximates the expected error and has a closed form solution for the parameters of an exponential model.

In the case of dynamic combination, apart from the problem of finding the optimal parameters $\lambda(\cdot)$, we face another optimization issue: that of finding the optimal partition functions $F_i(\cdot)$ for each information source. Ideally we would like to jointly optimize the partition function and the parameters associated with it, i.e. find the partition of the space whose optimal parameters achieve the minimum number of word errors. In the section 3.3 we present an algorithm that approximates this search.

3. MULTILINGUAL ACOUSTIC MODEL COMBINATION

First we describe our experimental setup and give the results for some simple, knowledge based partitions. Then we describe the automatic approach that constructs a partition in an attempt to achieve the lowest WER with the minimum number of parameters $\lambda(\cdot)$.

3.1. Database description

As our source-language acoustic training data we used English Broadcast News obtained from LDC. The target language was Czech for which we used 1.0 hour of the Charles University Corpus of Financial News (CUCFN). This is read speech and was used in the '99 Summer Workshop at JHU [4] to train the baseline acoustic models for Czech. We also obtained from the same corpus an extra 1.0 hour of Czech data which we use to train the combination parameters of the log-linear model. The test set was 1.0 hour of Czech Voice of America broadcasts.

The acoustic models were trained from mel-frequency, cepstral data using HTK [10]. We used state-clustered cross-word triphone HMMs with 3 states each and 6 gaussian mixtures per state. There were a total of 6040 shared states in the system.

1000-Best hypotheses were obtained for the training and test data using the 1.0 hour baseline Czech language triphone acoustic model.

In all the experiments described in the following sections a bigram language model [5] and word insertion penalty were included. Their weights were optimized along with the acoustic model weights $\lambda_{AC}(\cdot)$.

We can see the baseline result using only the one hour trained Czech acoustic models (A_{cz}) in line (a) of Table 1. We compare this with the WER when we used only the English acoustic models adapted on the one hour of Czech data A_{en} (b). We notice

that the WER is lower with the English models. This is due to the fact that we evaluate the English model by rescoring the N-Best obtained with the original Czech models. Decoding with the English models and the Czech language model is worse than the 1.0 hour Czech system.

	$\#\lambda_{AC}$		WER (%)
		BASELINES	
(a)	1	A_{cz}	29.1
(b)	1	A_{en}	28.4
		STATIC COMB.	
(c)	2	$A_{cz} + A_{en}$	27.8
		DYNAMIC COMB. (knowledge based partition)	
(d)	6	$(\text{V+C+S})_{cz}$ + $(\text{V+C+S})_{en}$	27.5
(e)	71	$(\text{12V+27C+1S})_{cz}$ + $(\text{10V+20C+1S})_{en}$	**26.8**
(f)	28	$(\text{12V+1C+1S})_{cz}$ + $(\text{10V+1C+1S})_{en}$	26.9
(g)	51	$(\text{1V+27C+1S})_{cz}$ + $(\text{1V+20C+1S})_{en}$	27.2
		(tree partition)	
(h)	8	tree leaves in Figure 2	27.0

Table 1: Combination of English and Czech acoustic models using different acoustic classification schemes.

3.2. Knowledge based partition

The first system we evaluated utilized both the English and Czech acoustic models score in a static combination (system $A_{cz} + A_{en}$ in Table 1, line (c)). Here the two models were combined as in formula (1) using two weights: λ_{cz} and λ_{en}. Thus using only one extra weight yields a significant improvement over the monolingual systems.

Next we explored dynamic combination by examining different knowledge based partitions for the hypothesized phone segments for each information source used. We consider partitions only for the acoustic information sources; the language model receives only one weight.

The first partition simply clusters together the vowel, consonant and silence models for the English and Czech language ($(\text{V+C+S})_{cz/en}$) and assigns one weight for the phone models in each cluster. This model has 6 acoustic weights to be trained (one for each of the classes for each of the languages). The accuracy improves slightly over the static combination (line (d) in Table 1).

We find that by allowing each phone model to have its own weight we achieve a further improvement (Table 1-line (e)). In that system there are 40 weights for the Czech phone models (12V+27C+1S) and 31 weights for the English (10V+20C+1S). But we found that we could reduce the number of parameters significantly by allowing separate weights for only the vowels for each language and tying the weights for the consonants to one parameter for each model, without significantly changing WER (result (f) in table). On the other hand when only the consonants were allowed to have separate weights the result deteriorated (result (g)).

The above results suggest that we should aim to obtain the optimal partition: the minimum number of tied parameters that achieve the lowest WER.

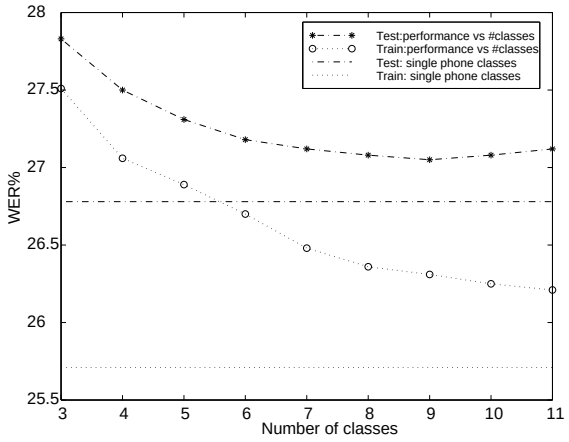

Figure 1: WER results achieved with different number of classes chosen by the automatic partition algorithm, compared with the knowledge based partition which puts one phone in each class.

3.3. Searching for optimal partition of the parameter space

The goal of this experiment is to obtain the accuracy improvements achieved in the previous section with parsimonious acoustic clustering. We use an automatic approach to find an optimal partitioning of model scores into classes. The algorithm iteratively builds a binary partition of the hypothesized phone segments. It starts with all phones (from all acoustic models) in one class-node. Phonological questions are used repeatedly to split each available class into two sub-classes. For every node, all questions are examined: weights are optimized for the resulted partitions and WER is computed. The tree grows by choosing to split the node using the question that results in classes with the best WER improvement.

We also investigated building the tree with splitting criterion the improvement in the smooth approximation of WER [6]. This objective function has a closed form solution so the tree building algorithm was much faster. However the resulting WER was higher than found with the direct minimization of WER.

The questions asked in the tree are allowed to separate classes of phones from the pool of phone-models available, and assign a separate weight to each of the classes. These questions involve either general classes of models (from both languages), such as "Is this a liquid phone?", or "Is it in the class of phone A" (this is the class of phones (A, AA, AH, AE, AX), or they can separate a specific phone, for example "Is this the 'EH' Czech phone".

In Figure 2 we see the binary tree partition chosen by the algorithm after finding 8 classes. We notice that the algorithm produces a right branching tree. That means that each question chosen identifies small classes (most of them consist of one phone), while the majority of phone models are still tied with one common weight. This suggests that only a few phones contribute most of the improvement in discrimination. Another observation we make from the figure is that the majority of questions involve vowels. This means that putting individual weights for

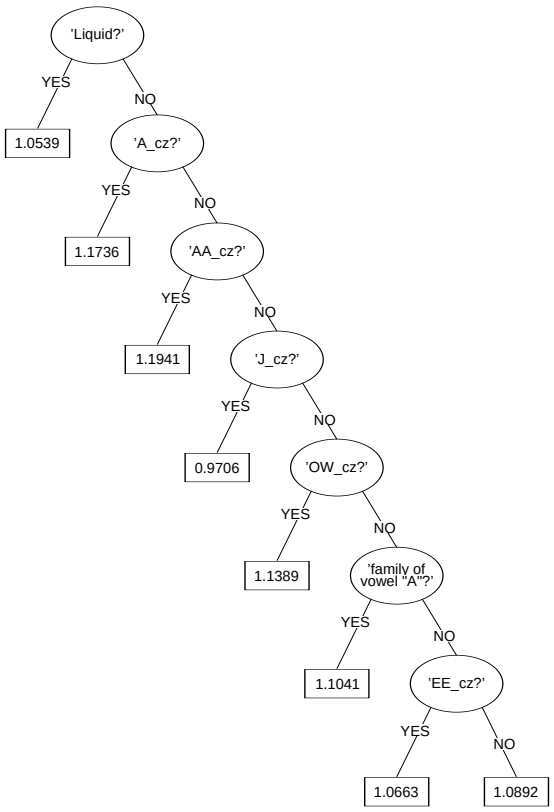

Figure 2: The binary tree partition constructed by the automatic partition algorithm. The class weights are shown in each leaf.

some vowels is more beneficial than other classes, which is consistent with the results (f)-(g) of Table 1 that suggests vowels with individual weights are more beneficial than consonants.

The results in Table 1 show that we obtain gains comparable to the knowledge based approach with far fewer parameters. We were not successful in finding a partition that achieved a higher accuracy than the sparse partition with one phone per class. However, this technique should prove robust as more information sources are incorporated into the model.

In Figure 1 we plot the accuracy achieved by the system as a function of the number of classes chosen by the automatic algorithm (on both the training and the test data); the accuracy achieved by the system with the sparse phone partition (knowledge based partition) is also plotted. We notice that after 8 classes the change in WER from adding extra classes is negligible. Investigation of more proper stopping criteria is under way.

4. DISCUSSION

We have presented a new approach for sub-word multilingual acoustic model combination. We have found that dynamic combination of multilingual acoustic phonetic classes is superior to static combination of multilingual acoustic scores.

Previous attempts at multilingual acoustic clustering have been mainly employed in maximum likelihood modeling [9]. We have shown that minimum risk acoustic clustering is effective in finding acoustic classes that directly minimize the word error rate using MAP decision rules.

Acknowledgment: We thank P. Beyerlein for helpful discussions on this topic.

5. REFERENCES

1. T. Schultz and A. Waibel: "Fast bootstrapping of LVCSR systems with multilingual phoneme sets", *Proceedings EUROSPEECH*, pp 371-374, 1997
2. P. Fung and C.Y. Ma and W.K. Liu: "MAP-based cross-language adaptation augmented by linguistic knowledge from English to Chinese", *Proceedings EUROSPEECH*, pp 871-874, 1999.
3. W. Byrne et al.: "Towards language independent acoustic modeling", *Proceedings ICASSP*, pp 1029-1032, 2000.
4. W. Byrne, et al.: "Towards language independent acoustic modeling: Final report", *www.clsp.jhu.edu/ws99/projects/asr*.
5. W. Byrne, et al., "Large vocabulary speech recognition for read and broadcast Czech," *Proceedings Wkshp. on Text Speech and Dialog, Marianske Lanze, Czech Republic*, 1999.
6. P. Beyerlein: "Discriminative model combination", *Proceedings ICASSP*, pp 481-484, 1998.
7. J. Nelder and R. Mead, "A simplex method for function minimization", *Computer Journal*, vol 7, pp. 308-313, 1965.
8. Dimitra Vergyri, "Use of word level side information to improve speech recognition", *Proceedings ICASSP*, 2000.
9. P. Cohen et al.: "Towards a universal speech recognizer for multiple languages", *Proceedings Wkshp on Automatic Speech Recognition and Understanding*, 1997.
10. S. Young and D. Kershaw and J. Odell and D. Ollason and V. Valchev and P. Woodland, *The HTK Book*, Entropic Inc., 1999.

MINIMUM RISK ACOUSTIC CLUSTERING FOR MULTILINGUAL ACOUSTIC MODEL COMBINATION

Dimitra Vergyri , Stavros Tsakalidis and William Byrne [2]
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218-2686
{*byrne,dverg,stavros*}*@clsp.jhu.edu*

In this paper we describe procedures for combining multiple acoustic models, obtained using training corpora from different languages, in order to improve ASR performance in languages for which large amounts of training data are not available. We treat these models as multiple sources of information whose scores are combined in a log-linear model to compute the hypothesis likelihood. The model combination can either be performed in a static way, with constant combination weights, or in a dynamic way, with parameters that can vary for different segments of a hypothesis. The aim is to optimize the parameters so as to achieve minimum word error rate. In order to achieve robust parameter estimation in the dynamic combination case, the parameters are defined to be piecewise constant on different phonetic classes that form a partition of the space of hypothesis segments. The partition is defined, using phonological knowledge, on segments that correspond to hypothesized phones. We examine different ways to define such a partition, including an automatic approach that gives a binary tree structured partition which tries to achieve the minimum WER with the minimum number of classes.