IDIAP RESEARCH REPORT

Using KL-based Acoustic Models in a Large Vocabulary Recognition Task

Guillermo Aradilla [a] Hervé Bourlard [a]
Mathew Magimai Doss [a]

IDIAP–RR 08-14

April 2008

[a] IDIAP Research Institute and Ecole Polytechnique Fédérale de Lausanne (EPFL)

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-14

Using KL-based Acoustic Models in a Large Vocabulary Recognition Task

Guillermo Aradilla Hervé Bourlard Mathew Magimai Doss

April 2008

Abstract. Posterior probabilities of sub-word units have been shown to be an effective front-end for ASR. However, attempts to model this type of features either do not benefit from modeling context-dependent phonemes, or use an inefficient distribution to estimate the state likelihood. This paper presents a novel acoustic model for posterior features that overcomes these limitations. The proposed model can be seen as a HMM where the score associated with each state is the KL divergence between a distribution characterizing the state and the posterior features from the test utterance. This KL-based acoustic model establishes a framework where other models for posterior features such as hybrid HMM/MLP and discrete HMM can be seen as particular cases. Experiments on the WSJ database show that the KL-based acoustic model can significantly outperform these latter approaches. Moreover, the proposed model can obtain comparable results to complex systems, such as HMM/GMM, using significantly fewer parameters.

1 Introduction

Posterior probabilities have been successfully applied in the automatic speech recognition (ASR) field. For example, they have been used for word lattice re-scoring [1], beam search pruning [2] or estimating word confidence measures [3]. Posterior probabilities of sub-word units, such as phonemes, have been shown to be an effective front-end for ASR [4]. The most common method for estimating posterior features is through a multi-layer perceptron (MLP) [5].

Posterior probabilities can play three different roles when used as inputs for acoustic models. They can be used as *scores*, *labels* or *features*. Hybrid HMM/MLP (multi-layer perceptron) systems [5] use posteriors as scores. In this case, Bayes' rule is applied to posteriors for estimating the state scaled likelihoods. Their main limitation is that each state is associated to an output node of the MLP. This can present difficulties when states represent a large number of models, e.g. context-dependent phonemes, while keeping a reasonable size of the MLP. Discrete HMMs can use posteriors to obtain the labels (codewords) [6]. Each label is defined as the class with the highest probability of the posterior feature. This acoustic model is mainly used when fast decoding is required because state likelihoods can be quickly obtained from a look-up table. It has been shown that they can provide good performance because they can model the feature variability within each state through a discrete emission probability [7]. However, they make a simplification on the posterior features because only the most probable class is considered. Posteriors are used as features in the approach known as "tandem" [8]. They are post-processed and then modeled by a HMM where state likelihoods are described by Gaussian mixture models (GMMs). The tandem system thus benefits from both the discriminative behavior of posterior features and the good modeling properties of GMMs. This approach, though, requires a large number of parameters to properly characterize each state likelihood distribution.

In this work, we applied the acoustic model presented in [9] to a large vocabulary recognition task. This model has a topology equivalent to typical HMMs for ASR. The parameters that describe each state belong to the same space as the posteriors, i.e., each state is characterized by a multinomial distribution that lies on the same simplex-space as the posterior features. These state distributions can be estimated from a training dataset. A score is defined for each state based on the Kullback-Leibler (KL) divergence [10] between the posterior features and the state distributions. Since posterior features can be seen as probability distributions over the space of classes, the KL divergence appears to be a reasonable choice because it is a natural dissimilarity measure between distributions.

The presented model offers some advantages with respect to the acoustic models described above. Since each state is only characterized by a multinomial distribution, context-dependent models can be easily used without changing the structure of the posterior estimator as it is the case of hybrid HMM/MLP. Also, unlike discrete HMMs, all the components of the posterior features are taken into account without making any simplification. Moreover, only a few parameters are necessary for properly characterizing each state, in contrast with HMM/GMM-based approaches. In this paper, we also show that the proposed model establishes a general framework where hybrid HMM/MLP and discrete HMM can be seen as particular cases.

This paper is structured as follows: Section 2 briefly describes posterior features. Section 3 presents the KL-based acoustic model. Section 4 shows the links between the KL-based model and state-of-the-art acoustic models for posterior features. Section 5 describes the recognition experiments on the Wall Street Journal database [11] and discusses the results. Finally, Section 6 concludes this paper.

2 Posterior Features

Given a sequence of cepstral-based features $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T\}$ extracted from an spoken utterance, a sequence of posterior features $Z = \{\mathbf{z}_1, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_T\}$ can be estimated from a MLP. Each posterior feature $\mathbf{z}_t = [p(c_1|\mathbf{x}_t), \cdots, p(c_k|\mathbf{x}_t), \cdots, p(c_K|\mathbf{x}_t)]^\mathsf{T}$ is the set of MLP output values when the acoustic feature $\mathbf{x}_t$ is used as input. In this work, the MLP has K output classes corresponding to phonemes.

3 KL-based Acoustic Model

The KL-based acoustic model can be described as a finite state machine of Q states, where each state $i \in \{1, \dots, Q\}$ is parameterized by a multinomial distribution $\mathbf{y}^i$. As in HMMs used for ASR, states have a left-to-right topology. A score $S(\mathbf{y}^i, \mathbf{z})$ is defined for each state as the KL divergence between the state distribution $\mathbf{y}^i$ and the posterior features $\mathbf{z}$. In a similar way as standard ASR systems, the transition cost from state i to the state j is the negative logarithm of the state transition probability $a_{ij} = P(q_t = j | q_{t-1} = i)$. The structure of this model is illustrated in Figure 1.

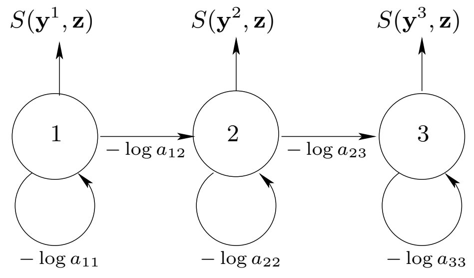

Figure 1: Scheme of the KL-based acoustic model formed by three states.

The KL divergence between two discrete distributions $\mathbf{a}$ and $\mathbf{b}$ is defined as [10]:

$$KL(\mathbf{a}||\mathbf{b}) = \sum_{k=1}^{K} a(k) \log \frac{a(k)}{b(k)} \tag{1}$$

where K denotes the dimension of the distributions. The distributions $\mathbf{a}$ and $\mathbf{b}$ play different and meaningful roles: $\mathbf{a}$ is the reference distribution and $\mathbf{b}$ is the distribution to be tested. In this work, we consider three possible cases:

HMM/KL state distributions are the reference distribution: $S(\mathbf{y}^i, \mathbf{z}) = KL(\mathbf{y}^i, \mathbf{z})$

HMM/RKL (reverse) posterior features are the reference: $S(\mathbf{y}^i, \mathbf{z}) = KL(\mathbf{z}, \mathbf{y}^i)$

HMM/SKL (symmetric) KL and RKL are combined: $S(\mathbf{y}^i, \mathbf{z}) = \frac{1}{2}[KL(\mathbf{y}^i, \mathbf{z}) + KL(\mathbf{z}, \mathbf{y}^i)]$

Given a linguistic unit m and a sequence of posterior features $Z = \mathbf{z}_1, \cdots, \mathbf{z}_T$, the global score $J_m(Z)$ is expressed as

$$J_m(Z) = \min_{\mathcal{Q}(m)} \left[\sum_{t=1}^{T} S(\mathbf{y}^{q_t}, \mathbf{z}_t) - \log a_{q_{t-1} q_t} \right] \tag{2}$$

where the state score $S(\mathbf{y}^i, \mathbf{z})$ varies depending on the configuration: HMM/KL, HMM/RKL or HMM/SKL. The set $\mathcal{Q}(m)$ denotes all possible state sequences allowed by the linguistic unit m.

Training and decoding procedures are based on the minimization of this score $J_m(Z)$. Algorithms based on the Viterbi approximation are explained in detail in [9].

4 Links with other Acoustic Models

In this section, we present the links between the KL-based acoustic and hybrid HMM/MLP and discrete HMM.

4.1 Derivation of hybrid HMM/MLP

As it has already mentioned in the introduction, hybrid HMM/MLP estimates the state scaled likelihood from the MLP posteriors via Bayes' rule [5]. Hence, if we assume that the phoneme priors are uniform, the score of a sequence of posterior features Z given the hybrid HMM/MLP model is defined as

$$J_m^H(Z) = \max_{\mathcal{Q}(m)} \left[\sum_{t=1}^{T} \log P(q_t|\mathbf{x}_t) + \log a_{q_{t-1}q_t} \right] \tag{3}$$

where $\mathcal{Q}(m)$ denotes the set of all possible state sequences allowed by the unit m.

When comparing the expression (3) with the model score from HMM/KL (2), it can be noted that they are equivalent if we define the state distribution $\mathbf{y}^{q_t}$ as a delta distribution centered at the phoneme represented by the state q_t. Thus,

$$J_m^H(Z) = \max_{\mathcal{Q}(m)} \left[\sum_{t=1}^{T} -KL(\delta_{\rho(q_t)}||\mathbf{z}_t) + \log a_{q_{t-1}q_t} \right] \tag{4}$$

where $\rho(q)$ is the mapping from each state q to its corresponding class (MLP output). This represents the major limitation of hybrid HMM/MLP because each state must correspond to an output class from the MLP. HMM/KL removes this constraint because each state is represented by a multinomial distribution whose values are estimated from the training data.

Therefore, hybrid HMM/MLP can be seen as a particular case of HMM/KL where state distributions are delta distributions and phoneme priors are assumed uniform.

4.2 Derivation of discrete HMM

In discrete HMM, the sequence of speech features is first quantized into clusters. The most common methods to quantize are based on K-means [12] and MLP [6]. This latter approach yields better accuracy because of its capability of modeling non-linear boundaries and discriminative training. For each feature vector $\mathbf{x}_t$, its corresponding label v_t is defined as the MLP output with the highest probability, $v_t = \arg\max_k p(c_k|\mathbf{x_t})$. Given a sequence of cluster indexes $V = v_1 \cdots v_t \cdots v_T$, the score of a discrete HMM is defined as

$$J_m^D(V) = \max_{\mathcal{Q}(m)} \left[\sum_{t=1}^{T} \log P(v_t|q_t) + \log a_{q_{t-1}q_t} \right] \tag{5}$$

This score function can then be represented in terms of the HMM/RKL configuration by turning the indexes into delta distributions centered at the dimension given by the index.

$$J_m^D(V) = \max_{\mathcal{Q}(m)} \left[\sum_{t=1}^{T} -KL(\delta_{v_t}||\mathbf{y}^{q_t}) + \log a_{q_{t-1}q_t} \right] \tag{6}$$

Therefore, discrete HMM can be seen as a particular case of HMM/RKL where posterior features are delta distributions centered at the component with the highest probability. It should be noted that this relation have already been observed in [13]. In that case, the discrete HMM was using a fuzzy vector quantizer.

5 Experiments and Results

In previous experiments [9], we applied the KL-based acoustic model to a small vocabulary task. In this work, the Wall Street Journal (WSJ) database [11] is used to carry out recognition experiments using the acoustic models described in the previous sections. A set of 38250 utterances ($\sim$ 80 hours) has been used for training the MLP. For each posterior feature $\mathbf{z}_t$, a context of 9 PLP acoustic vectors ($\mathbf{x}_{t-4} \cdots \mathbf{x}_t \cdots \mathbf{x}_{t+4}$) is used as inputs (39 x 9 = 351 units) for the MLP. It contains 3652 hidden units and 45 output nodes, which are also the set of phonemes $\{c_k\}_{k=1}^{K}$, hence $K = 45$. The total number of weights of the MLP corresponds to 5% of the total amount of samples in the training data. The test data used in this work is known as WSJ 5K task. A bigram language model is applied to decode 913 test utterances ($\sim$ 2 hours) using a lexicon of 5150 words. Experiments using context independent (CI) and context dependent (CD) phonemes have been carried out. Phonemes are modeled by a 3-state KL-based model. In the case of CD phonemes, the 4000 word-internal triphones appearing most often in the training data have been used (they cover 85% of the training dataset). For this case, the 45 CI phonemes are used to model unseen triphones on the test set. For the tandem system, 16 Gaussian distributions are used to describe each state likelihood. In this case, the post-processing of the posterior features is done using the standard procedure (log + PCA) described in the original paper [8].

model	CI	CD
hybrid HMM/MLP	23.9	-
HMM/KL	23.5	22.3
discrete HMM	25.5	40.0
HMM/RKL	26.6	22.4
HMM/SKL	23.3	20.9
tandem	27.7	20.0

Table 1: Word error rate expressed in percentage

Results are shown in Table 1. The performance of hybrid HMM/MLP only corresponds to the monophone case since, in this work, the MLP estimates posteriors of context-independent phonemes. As it was shown in [9], HMM/KL performs better than hybrid HMM/MLP and significant improvement is obtained when using context-dependent models. This can be explained because state distributions $\{\mathbf{y}^i\}$ can better describe the co-articulation effects. Table 5 illustrates this effect. The highest components of the state distributions characterizing the triphone /*r*/-/*ah*/+/*m*/ are shown. It can be noted that the left context (/*r*/) is represented in the first state and that the right context (/*m*/) appears in the third state. The second state corresponds to the central phoneme (/*ah*/), which is also represented by the neutral vowel (/*ax*/).

/*r*/-/*ah*/+/*m*/		
/*ah*/ (0.5)	/*ah*/ (0.8)	/*ah*/ (0.5)
/*r*/ (0.2)	/*ax*/ (0.1)	/*m*/ (0.3)

Table 2: The two highest components of the three state state distributions characterizing hte triphone /*r*/-/*ah*/+/*m*/. The corresponding phoneme and its weight value are represented.

When comparing the performance of discrete HMM/RKL and discrete HMM, we can observe that discrete HMM fails when modeling context-dependent models. Since there is a large number of models, less training samples are assigned to each one. Thus, these models are not properly characterized because posteriors are simplified to only their highest component. This effect does not appear when using HMM/RKL because all the information from the posterior features is used.

Experiments have been carried out where delta distributions (labels from posteriors) have been used for the HMM/RKL models. These experiments have yielded comparable results to the HMM/RKL case but decoding time has been significantly reduced. Hence, we can take advantage of the good generalization properties of HMM/RKL and the fast decoding time of discrete HMMs.

HMM/SKL outperforms the rest of KL-based acoustic models as it was also observed in [9]. This can be explained by analyzing the entropy of the posterior features and the reference distributions.

Table 3 presents the average entropy of the reference distributions for all the KL-based acoustic models.

model	CI	CD
HMM/KL	0.21	0.25
HMM/RKL	0.91	0.87
HMM/SKL	0.49	0.50
posteriors	0.51	

Table 3: Average entropy of the state distributions for the KL-based acoustic models. The last row corresponds to the average entropy of the posterior features. The entropy of a posterior feature $\mathbf{z}_t$ is computed as $H(\mathbf{z}_t) = -\sum_k p(c_k|\mathbf{x}_t) \log p(c_k|\mathbf{x}_t)$.

From Table 3, we can observe that the average entropy of the reference distributions when using HMM/SKL is similar to the average entropy of the posterior features. This effect is also observed at the state level: the entropy of each state distribution is similar to the average entropy of the training samples used to estimate that state distribution[1]. Since the entropy is a measure of uncertainty [10], this means that the uncertainty of a state distribution is equivalent to the average uncertainty of the training samples used to estimate that distribution. This suggests that HMM/SKL yields better performance because the symmetric KL criterion is able to estimate more appropriate state distributions.

5.1 System Complexity

In this section, we evaluate the complexity of the different acoustic models described in this paper. We compare the performance and the number of total parameters for different training data sizes. Table 4 presents the total number of parameters of the different models and Figure 2 shows the evolution of the performance using context-dependent models. The parameters of hybrid HMM/MLP are only the weights of the MLP. The number of hidden units of the MLP is chosen so that the number of weights represents 5% of the training feature vectors.

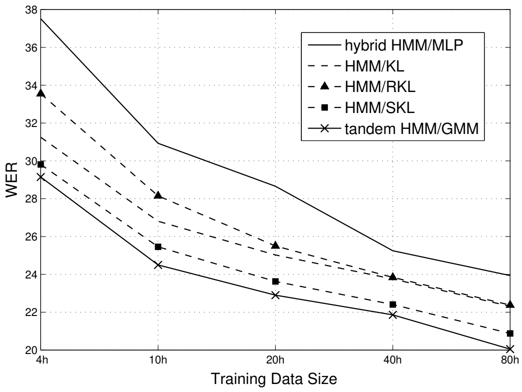

Figure 2: Word error rate depending on the training data size.

We can observe that all the models are similarly affected by the training data size. It can be noted that HMM/SKL obtains a comparable performance to tandem by using a significantly fewer number of parameters. This suggests that the KL divergence is a more convenient measure to describe the feature variability within each state than a GMM. When comparing HMM/SKL and hybrid HMM/MLP, it

[1]The same conclusions have been obtained in other databases.

training	hybrid HMM/MLP	HMM/SKL	tandem
4h	72K	612K	14856K
10h	182K	722K	48695K
20h	365K	905K	78909K
40h	730K	1270K	106008K
80h	1460K	2010K	135496K

Table 4: Parameters of the acoustic models. The number of parameters of HMM/KL, HMM/RKL and HMM/SKL is the same.

can be noted that hybrid HMM/MLP requires 80 hours of training data and a MLP of 1460K weights for obtaining the same performance as HMM/SKL, using only 20 hours of training data and 905K total parameters. Therefore, characterizing each state with a multinomial distribution is more effective than using a bigger MLP trained on more data.

6 Conclusions

In this paper we have presented a novel acoustic model for posteriors features based on the KL divergence. Since KL is not symmetric, three configurations are possible. One of them (HMM/KL) can be seen as a general case of hybrid HMM/MLP when phoneme priors are assumed uniform. The other configuration (HMM/RKL) is a general case of discrete HMM where posteriors are not simplified to delta distributions. In this work we show that the two configurations of the KL-based model outperform both hybrid HMM/MLP and discrete HMM. A symmetric combination of KL divergence can also be used as state score (HMM/SKL). Since this configuration is able to adjust the entropy of the state distributions according to the entropy of the training samples, HMM/SKL outperforms the rest of KL-based acoustic models.

In addition, a study on the complexity of the acoustic models is carried out. The proposed HMM/SKL yields comparable performance than HMM/GMM by using significantly fewer number of parameters. This suggest that KL divergence is a more appropriate measure than GMM for computing the feature variability within each state. Moreover, HMM/SKL outperforms hybrid HMM/MLP by using less amount of training data and fewer parameters. Thus, state distributions are more effective for describing the speech variability than using a more complex MLP.

Discrete HMM are very suitable acoustic models when decoding time is a relevant factor. However, they lack of generalization capabilities when using context-dependent phonemes. In this work, we also show that we can benefit from good generalization properties and fast decoding by training the system using HMM/RKL and decoding using the codewords obtained from the posteriors.

Future work should be focus on increasing the capacity of the system. This increasing of the complexity system can be done by augmenting the number of classes (in this work, classes are phonemes). A reasonable direction would be to investigate the use of more classes without the need of a large number of MLP output nodes. For example, several MLPs that estimate articulatory features could be combined [14].

7 Acknowledgements

This work was supported by the EU 6th FWP IST integrated project AMI (FP6-506811). The authors want to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on “Interactive Multimodal Information Management (IM2)”.

References

[1] G. William and S. Renals, "Confidence Measures from Local Posterior Estimate," *Computer, Speech and Language*, vol. 13, pp. 395–411, 1999.

[2] S. Abdou and M. S. Scordilis, "Beam Search Pruning in Speech Recognition Using a Posterior-based Confidence Measure," *Speech Communication*, vol. 42, pp. 409–428, 2004.

[3] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and other Applications of Confusion Networks," *Computer, Speech and Language*, vol. 14, pp. 373–400, 2000.

[4] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On Using MLP features in LVCSR," *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2004.

[5] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, vol. 247, Kluwer Academic Publishers, Boston, 1993.

[6] G. Rigoll, "Maximum Mutual Information Neural Networks for Hybrid Connectionist HMM Speech Recognition Systems," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 175–184, 1994.

[7] P. Le Cerf and D. Van Compernolle, "Using MLP's as Probability Generators vs. as Labelers: A Comparative Study," Tech. Rep., ESAT Laboratory, 1993.

[8] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.

[9] G. Aradilla, J. Vepa, and H. Bourlard, "An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[10] T. M. Cover and J. A. Thomas, *Information Theory*, John Wiley, 1991.

[11] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *DARPA Speech and Language Workshop*, 1992.

[12] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[13] E. Tsuboka and J. Nakahashi, "On the Fuzzy Vector Quantization Based Hidden Markov Model," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 637–640, 1994.

[14] S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech Using Neural Networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.