IDIAP RESEARCH REPORT

AN ACOUSTIC MODEL BASED ON KULLBACK-LEIBLER DIVERGENCE FOR POSTERIOR FEATURES

Guillermo Aradilla [a] [b] Jithendra Vepa [b]
Hervé Bourlard [a] [b]

IDIAP–RR 06-60

JANUARY 2007

[a] IDIAP Research Institute, Martigny, Switzerland

[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 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 06-60

An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features

Guillermo Aradilla Jithendra Vepa Hervé Bourlard

January 2007

Abstract. This paper investigates the use of features based on posterior probabilities of subword units such as phonemes. These features are typically transformed when used as inputs for a hidden Markov model with mixture of Gaussians as emission distribution (HMM/GMM). In this work, we introduce a novel acoustic model that avoids the Gaussian assumption and directly uses posterior features without any transformation. This model is described by a finite state machine where each state is characterized by a target distribution and the cost function associated to each state is given by the Kullback-Leibler (KL) divergence between its target distribution and the posterior features. Furthermore, hybrid HMM/ANN system can be seen as a particular case of this KL-based model where state target distributions are predefined. A training method is also presented that minimizes the KL-divergence between the state target distributions and the posteriors features.

1 Introduction

Posterior probabilities have recently gained importance in the automatic speech recognition (ASR) field. They have been applied for confidence measures [1], beam search pruning [2] or word lattice re-scoring [3]. Posteriors of subword units, such as phonemes, can also be used in ASR. A multi-layer perceptron (MLP) with outputs representing these subword units can be trained to estimate these probabilities. Experiments have shown that this speech representation yields better results because of the long acoustic context used as input for the MLP and its discriminative training procedure [4].

Phoneme posterior probabilities have mainly played two roles in ASR. In Tandem [5], they are used as observation vectors in a state-of-the-art HMM/GMM system. In this case, posteriors are post-processed with a log-transformation and a KLT-based decorrelation to make these features more Gaussian-like and hence, easier to be modeled by a GMM. However, a large amount of training data is required for estimating the parameters of the GMM and the actual meaning of the posterior features is lost after the post-processing.

In hybrid HMM/ANN system [6], posterior features are used to directly estimate the state emission probabilities of a HMM by using Bayes' rule. In spite its well-founded mathematical formulation (given some independence assumptions, HMM/ANN is able to estimate the global posterior probability of a model given a sequence of acoustic data [7]), this approach does not obtain as good performance as Tandem. This can be explained by the fact that each state of the hybrid HMM/ANN model must correspond explicitly to an MLP output. This rigidness does not allow to easily use strategies that are often applied in conventional HMM/GMM approach such as context-dependent phonemes as subword units or state-tying for parameter estimation. Context-dependent phonemes can be used as target labels for the MLP. In this case, states of the hybrid HMM/ANN would represent these subword units but this approach becomes impractical for a large vocabulary task because it increases the size of the MLP enormously.

In this paper, we present a novel acoustic model that alleviates the rigidness of the hybrid HMM/ANN system. This model is described by a finite state machine where each state is parameterized by a target posterior distribution. The cost function associated to every state is given by the KL-divergence between its target posterior distribution and the posterior features. In this way, states are not tied to a particular MLP output and hence, they have more flexibility to represent other types of subword units without changing the structure of the MLP. We also show that hybrid HMM/ANN system can be interpreted as a particular case of this model where state target distributions are fixed and equal to a delta distribution. Furthermore, this system naturally extends our previous work where we successfully applied posterior features and KL-divergence to the template matching approach for ASR [8].

This paper is organized as follows: Section 2 describes the hybrid system approach for speech recognition, Section 3 presents the KL-based model, which can be seen as a generalization of the hybrid system, Section 4 explains the experiments and their results and finally Section 5 gives conclusions and some ideas for future work.

2 Hybrid HMM/ANN system

A hybrid HMM/ANN system is a HMM-based model where state emission distributions are derived from posterior probabilities obtained through an MLP.

A sequence of spectral-based features $\{x_t\}_{t=1}^{T}$ of length T is used as input vectors for the hybrid HMM/ANN system. For each input vector x_t, a vector of conditional posterior probabilities $z_t = \{p(q_k|x_t)\}_{k=1}^{K}$ is computed using an MLP[1] with K outputs. Since each HMM state is related to a particular MLP output, K also denotes the number of HMM states.

[1] We are using this notation for the sake of simplicity, but in fact an acoustic context is used as input of the MLP, hence, the rigorous notation should be $p(q_k|x_{t-\Delta}^{t+\Delta})$.

The likelihood $p(x_t|q_k)$ of a feature vector x_t given the kth state can be estimated using Bayes' rule. Posteriors are divided by their prior probabilities to obtain scaled likelihoods, which are used as emission distributions.

Transition and prior probabilities can be ignored without affecting significantly the final performance of the system since these terms can be assumed to be uniformly distributed. Therefore, by taking logarithm and changing the sign, the cost function given by the model corresponding to the word W can be expressed by:

$$J_W(x_1^T) = \min_{\phi^W} \sum_{t=1}^{T} -\log p(q_{\phi_t}|x_t) \tag{1}$$

where ϕ^W represents the set of all possible state sequences of length T allowed by the word W. This minimization can be efficiently obtained by applying Viterbi algorithm.

The main limitation of the hybrid HMM/ANN system is that MLP outputs are tied to HMM states. This system usually considers context-independent phonemes as subword units since MLP outputs tipically represent phonemes. Choosing another type of subword unit, such as context-dependent phonemes, implies changing the structure of the MLP and, if the number of possible subword units is too large, it can be a problem for training since there must be an MLP output for each unit.

3 Proposed KL-based model

3.1 Model Description

The description of this new model is based on the structure of the hybrid HMM/ANN system. From the previous section, we have seen that a sequence of posterior features $\{z_t\}_{t=1}^T$ is obtained from the spectral-based features $\{x_t\}_{t=1}^T$ through an MLP. These posterior features are discrete probability distributions on the state space. A similarity measure between two posterior distributions y and z can then be based on the KL-divergence [9]:

$$KL(y \,||\, z) = \sum_{k=1}^{K} y(k) \log \frac{y(k)}{z(k)} \tag{2}$$

Using the above KL-divergence definition, we can rewrite Equation 1 as

$$J_W(x_1^T) = \min_{\phi^W} \sum_{t=1}^{T} KL(\delta_{\phi_t} \,||\, z_t) \tag{3}$$

where δ_k is defined as the discrete delta distribution centered on the kth state.

Equation 3 allows us to extend the hybrid system to a more general model by substituting the delta distribution δ_k by a posterior vector y_k that characterizes the kth state. This posterior vector is denoted as *state target posterior distribution*. Then, the cost function of this model becomes

$$J_W(x_1^T) = \min_{\phi^W} \sum_{t=1}^{T} KL(y_{\phi_t} \,||\, z_t) \tag{4}$$

The model described by the above equation can be defined as a finite state machine where each state has an associated cost given by the KL-divergence between its corresponding target distribution and the posterior vector obtained from the input acoustic feature. This idea is illustrated in Figure 1.

In this new model, states are not tied to particular outputs of the MLP but they are only characterized by their target posterior distributions. States can thus represent other types of subword units that are not related with MLP outputs. For instance, we can easily use context-dependent phonemes without changing the structure of the MLP.

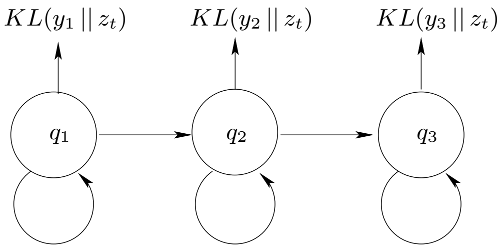

Figure 1: Scheme of KL-based model for a word formed by three phonemes. The cost function of the state q_k is given by the KL-divergence between its associated state target distribution y_k and the posterior feature z_t given at time t.

Furthermore, each target posterior distribution describes the pronunciation of the subword unit represented by its corresponding state. In this way, pronunciation variants can be modeled from the training data, avoiding thus to add manually multiple phonetic transcriptions to the dictionary.

3.2 Training

In this work, we consider that the MLP for obtaining the phoneme posteriors is fixed and already trained. Hence, the parameters of the KL-based model that must be estimated are the state target posterior distributions $\{y_k\}_{k=1}^{K}$. The criterion for estimating these parameters is based on minimizing the total cost function described in Equation 4 over a given training set. An iterative algorithm for training is presented here based on Viterbi segmentation. Through global convergence theorem [10], this algorithm ensures that the total cost function is minimized at each iteration and converges to a local minimum because KL-divergence is a convex function.

1. Training data is segmented uniformly according to the phonetic transcriptions.
2. Target distributions are computed to minimize the distance according to the segmentation. Given a set of posterior vectors $\{z_i\}_{i=1}^{N}$ corresponding to the same state q_k, the target posterior distribution y_k that minimizes the total distance can be computed as[2]

$$y_k(l) = \frac{\tilde{y}_k(l)}{\sum_{d=1}^{K} \tilde{y}_k(d)} \quad \text{and} \quad \tilde{y}_k(l) = \sqrt[N]{\prod_{i=1}^{N} z_i(l)} \tag{5}$$

   where the $y_k(l)$ denotes the lth component of the target distribution corresponding to the kth state.
3. Training data is segmented to minimize the global cost function using Viterbi algorithm.
4. Steps 2 and 3 are repeated until convergence of the cost function.

One advantage of this algorithm is that it starts from uniform segmentation and hence, the training set does not need to be labeled.

3.3 Derivations of the KL-based model

We can further extend this new model by substituting the KL-divergence used for computing the cost function by other metrics between distributions. In this work, we have explored two derivations of the KL-divergence:

[2] The proof is presented in the appendix.

- Since KL-divergence is not symmetric, we can switch the term of Equation 2 to obtain the reverse KL (RKL): $RKL(y\,||\,z) = KL(z\,||\,y)$. In this case, instead of Equation 5, the minimum distance is given by

$$y_k(l) = \frac{1}{N}\sum_{i=1}^{N} z_i(l) \tag{6}$$

- Also, the symmetric version of KL (SKL) can also be studied: $SKL(y\,||\,z) = KL(y\,||\,z) + KL(z\,||\,y)$. Since no closed solution exists for estimating the state target distributions in this case, we have used a convergent iterative procedure presented in [11].

4 Experiments and Results

Experiments have been carried out to explore the properties of the KL-based model described in the previous section. The task chosen in this work consists in continuous speech recognition on the numbers database [12]. The lexicon is formed by 30 words using 27 different phonemes and the MLP for generating the phoneme posteriors has been trained on a smaller version of the same database and is fixed for all experiments. Each subword unit (context-dependent or -independent phonemes) has 3 states and equivalently, a minimum duration constraint of 3 frames has been imposed to they hybrid HMM/ANN system. Also, word insertion penalty is used to equalize deletion and insertion errors and a single phonetic transcription is used per word. The test set is formed by 3576 utterances.

A first comparison was carried out between the KL-based model and the hybrid HMM/ANN system. States of the KL-based model represent context-independent phonemes for a fair comparison with the hybrid HMM/ANN system. A different number of training utterances was used for training the KL-based model to investigate the effect of training set size.

Model	# utterances	Accuracy
Hybrid HMM/ANN	-	89.8
KL	100	89.2
KL	200	91.0
KL	500	91.2
KL	1000	91.2
KL	2000	91.2
KL	5000	91.3

Table 1: System accuracy for both models. Different number of training utterances have been used for the KL-based model. Since in this work we are not considering the transition probabilities, no training files are required for the hybrid HMM/ANN model.

The increase in performance of the KL-based model when compared to hybrid HMM/ANN can be explained because of the higher flexibility for the state representation. As explained in Section 3, state target distributions are fixed for the hybrid case, whereas they are estimated for the KL-based model. Interestingly, significant improvement is already achieved with a reduced number of training data (200 utterances) and the further increase of the training set does not affect significantly the performance because of the reduced number of parameters.

We describe an example to illustrate how state target distributions can automatically model pronunciation variants using this KL-based model: words “four” and “oh” are phonetically transcribed as */f ao r/* and */ow/* respectively; “four” is sometimes misrecognized as “oh” because “four” can also be pronounced as */f ow r/*. In the hybrid approach, a new transcription should be added manually to the system dictionary. Nevertheless, KL-based model is able to deal with this variation by selecting more appropriate target distributions for the states associated to the phonemes */ao/*.

Phoneme	1st state	2nd state	3rd state
r	0.0	0.0	0.4
ao	0.3	0.1	0.2
ow	0.7	0.9	0.4

Table 2: Target distributions for the states corresponding to the phoneme */ao/* for the KL-based model.

In Table 2 we can see that the target distributions for the states corresponding to the phoneme */ao/* have been split into the phonemes */ao/* and */ow/*. Also, */r/*, which is the following phoneme for "four", has already been represented in the last state. In this way, KL-based model can better model the variability and therefore, the mismatch between "four" and "ow" have thus been alleviated without modifying the dictionary.

In the next experiment, we use context-dependent phones as subword units. We compare the performance of our model with Tandem system. Also, derivations of the KL-based model are evaluated.

# utterances	KL	RKL	SKL	Tandem
100	91.8	91.4	91.8	88.1 (5)
200	92.6	92.6	93.0	90.1 (7)
500	93.0	92.8	93.2	91.9 (7)
1000	93.3	93.0	93.6	93.2 (7)
2000	93.4	93.1	93.8	94.0 (10)
5000	93.5	93.4	94.0	95.0 (14)

Table 3: System accuracy for context-dependent KL-based model. Different sets of training utterances have been used. For the Tandem system, the number of Gaussians per state has also been indicated between parentheses.

From the above table, we can observe that our approach outperforms Tandem when a limited amount of training data is available. Indeed, one of the main drawbacks of Tandem is that a large amount of data is required to estimate all the parameters of the GMM. For example, in the case of using 5000 training utterances, the best performance of Tandem is achieved when using 14 Gaussians per state, since diagonal covariance matrices are used, a total of (14 components x [27 mean coefficients + 27 covariance coefficients] + 13 weights) 769 parameters are needed per state, while in our approach we only use 27 parameters. Hence, our approach is faster and consumes less computational resources than conventional HMM/GMM.

5 Conclusions and Future Work

In this paper, we have presented an extension of the hybrid system that alleviates some constraints of the hybrid HMM/ANN system while maintaining the actual interpretation of posterior features. Based on the experiments carried out on a continuous speech recognition task, the following properties of this new acoustic model can be drawn:

- Using a minimum amount of training data (only 200 utterances), KL-based model can significantly improve the accuracy of the hybrid system.

- This system can easily use to model context-dependent phonemes as subword units. In this case, results are comparable to more sophisticated systems, like Tandem.

- When reduced amount of training data, this model is particularly suitable because of its limited set of parameters (a single target posterior distribution characterizes each state).

A promising research direction is based on fitting this model into a probabilistic framework by finding an appropritate emission distribution for posterior features. Hence, all the advantages of the HMM theory could be exploit, such as Expectation-Maximization (EM) method for parameter estimation or discriminative training. Furthermore, similarly to Tandem, where a mixture of Gaussians is used as emission distribution, a mixture of posterior-based distributions could also be used to increase the capacity of the system and then performance of the system could improve when increasing the amount of training data.

6 Appendix

In this section, the proof for obtaining the state target posteriors distributions that minimizes Equation 4, once the segmentation has been fixed, is given for the KL and RKL metrics:

KL: The problem is defined as: Given a set of posterior vectors $\{z_i\}_{i=1}^{N}$, find y_k (the target distribution of state q_k) such that minimizes

$$f(y_k) = \sum_{i=1}^{N} KL(y_k \,||\, z_i) \tag{7}$$

Since y_k is also a posterior vector, the constraint $\sum_{n=1}^{K} y_k^n = 1$ must be satisfied. We use the method of Lagrange's multipliers to find the solution:

$$\frac{\partial}{\partial y_k(l)} \left[\sum_{i=1}^{N} KL(y_k \,||\, z_i) + \lambda \left(\sum_{n=1}^{K} y_k(n) - 1 \right) \right] = 0 \tag{8}$$

After some computations, we find that

$$y_k(l) = \lambda' \sqrt[N]{\prod_{i=1}^{N} z_i(l)} \tag{9}$$

where λ' is the normalization factor that satisfies the constraint.

RKL: Similarly to the KL case, the function to be minimized is

$$f(y_k) = \sum_{i=1}^{N} KL(z_k \,||\, y_i) \tag{10}$$

and after similar computations as the previous case, we obtain

$$y_k(l) = \frac{1}{N} \sum_{i=1}^{N} z_i(l) \tag{11}$$

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, “On Using MLP features in LVCSR,” *Proceedings of ICSLP*, 2004.

[5] H. Hermansky, D. Ellis, and S. Sharma, “Tandem Connectionist Feature Extraction for Conventional HMM Systems,” *Proceedings of ICASSP*, 2000.

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

[7] J. Hennebert, C. Ris, H. Bourlard, S. Renals, and N. Morgan, “Estimation of global posteriors and forward-backward training of hybrid hmm/ann systems,” *Proceedings of Eurospeech*, pp. 1951–1954, 1997.

[8] G. Aradilla, J. Vepa, and H. Bourlard, “Using Posterior-Based Features in Template Matching for Speech Recognition,” *Proceedings of ICSLP*, 2006.

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

[10] B. H. Juang and L. R. Rabiner, “The Segmental k-Means Algorithm for Estimating Parameters of Hidden Markov Models,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1639–1641, 1990.

[11] R. Veldhuis, “The Centroid of the Symmetrical Kullback-Leibler Distance,” *IEEE Signal Processing Letters*, vol. 9, pp. 96–99, 2002.

[12] R. Cole, M. Fanty, Noel M., and T. Lander, “New Telephone Speech Corpora at CSLU,” *Proceedings of Eurospeech*, pp. 821–824, 1995.