Developing and Enhancing Posterior Based Speech Recognition Systems

*Hamed Ketabdar**, *Jithendra Vepa, Samy Bengio, Hervé Bourlard**

IDIAP Research Institute, Martigny, Switzerland

{ketabdar,vepa,bengio,bourlard}@idiap.ch

Abstract

Local state or phone posterior probabilities are often investigated as local scores (e.g., hybrid HMM/ANN systems) or as transformed acoustic features (e.g., "Tandem") to improve speech recognition systems. In this paper, we present initial results towards boosting these approaches by improving posterior estimates, using acoustic context (e.g., as available in the whole utterance), as well as possible prior information (such as topological constraints). In the present work, the enhanced posterior distribution is associated with the "gamma" distribution typically used in standard HMMs training, and estimated from local likelihoods (GMM) or local posteriors (ANN). This approach results in a family of new HMM based systems, where only posterior probabilities are used, while also providing a new, principled, approach towards a hierarchical use/integration of these posteriors, from the frame level up to the phone and word levels, and integrating the appropriate context and prior knowledge in each level. In the present work, we used the resulting posteriors as local scores in a Viterbi decoder. On the OGI Numbers'95 database, this resulted in improved recognition performance, compared to a state-of-the-art hybrid HMM/ANN system.

1. Introduction

Using posterior probabilities for Automatic Speech Recognition (ASR) has become popular and frequently investigated in the past decade. Posterior probabilities have been mainly used either as local scores (measures) or as features in speech recognition systems. Hybrid Hidden Markov Model / Artificial Neural Network (HMM/ANN) approaches [1] were among the first ones to make use of posterior probabilities as local scores. In these approaches, ANNs and more specifically Multi-Layer Perceptrons (MLPs) are used to estimate the emission probabilities required in HMM system. Hybrid HMM/ANN method allows for discriminant training, as well as for the possibility of using acoustic context by presenting several frames at MLP input. Posterior probabilities have also been used as local scores for word lattice rescoring [2], beam search pruning [3] and confidence measures estimation [4]. Regarding the use of posterior probabilities as features, one successful approach is Tandem [5]. In Tandem, a trained MLP is used for estimating local phone posteriors. These posteriors, after some transformations, can be used alone or appended to standard features (such as MFCC or PLP) as input features to HMMs. Tandem technique takes the advantage of discriminative acoustic model training, as well as being able to use the techniques developed for standard HMM systems. In both hybrid HMM/ANN and Tandem approaches, local posteriors (i.e., posteriors estimated using local frame or limited number of local frames) are used.

*Also with Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland.

In [6], a method was presented to enhance the estimation of posterior probabilities by using the so called "gamma" recursion (as usually referred to in HMM formalism) to generate posteriors taking into account all acoustic information available in each utterance, as well as prior knowledge, possibly formulated in terms of HMM topological constraints. In this paper, the authors investigated the estimation and usage of these posteriors as features for a standard HMM layer. In their approach, posterior probabilities are estimated using gamma recursion in a HMM configuration, which, after some transformation, are fed as features into a second layer consisting of standard HMM/Gaussian Mixture Models (HMM/GMM). Such an approach was shown to yield significant performance improvement over Tandem approach on Numbers'95 and on a reduced vocabulary version (1'000 words) of the DARPA Conversational Telephone Speech-to-text (CTS) task.

In the present paper, we follow the same direction, i.e., investigating the estimation of posteriors taking into account the whole context and prior knowledge. However, instead of being used as new HMM/GMM features, these posteriors are now used as local scores in a Viterbi decoder. Our approach is as follows: First, the posterior probabilities are estimated using gamma recursion through a HMM layer, then these posteriors are used as local scores for a Viterbi decoder properly integrating these posteriors with phone transition and word transition probabilities. We have achieved performance improvement comparing with hybrid HMM/ANN approach which uses local posteriors also as local scores. This approach provides a new, principled, framework for the hierarchical estimation, integration and use of these posteriors, from the frame level up to the phone and word levels. In the framework of the experiment described below, the resulting system was also shown to be less sensitive to tuning factors (such as word insertion penalty), which are usually required in standard likelihood-based HMM for numerical compensation. This is due to the use of posteriors (having small numerical dynamic range) instead of likelihoods for decoding. This implies that there is less need for tuning in our system to get the best possible performance.

In the present paper, Section 2 shows how posterior probabilities can be estimated to capture the whole context and prior knowledge. Section 3 explains decoding method using these posteriors. Experiments and results are presented in Section 4. Conclusions and future work plans are discussed in Section 5.

2. Enhancing posterior probability estimation

In this section, we show how posterior probability estimation can be enhanced by using acoustic context information (in our case, using the whole utterance), as well as possible prior knowledge. Therefore, enhancing posterior estimates refers to estimating new and more informative posteriors. These posteriors can be estimated for different levels, ranging from phones to words, or even sentences [6]. In the present paper though, we focus on phone level posteriors only.

2.1. "State gamma" estimation

In phone-based HMM speech recognition systems, phones are typically modeled by specific HMMs M with a few number of states. The posteriors are first estimated for each state (called "state gammas" as referred to in HMM formalism and used in HMMs training), which are then integrated to phone or higher level posteriors. According to standard HMM formalism, the state gamma $\gamma(i, t|M)$ is defined as the probability of being in state i at time t, given the whole observation sequence $x_{1:T}$ and model M encoding specific prior knowledge (topological/temporal constraints):

$$\gamma(i, t|M) = p(q_t^i|x_{1:T}, M) \tag{1}$$

where, x_t is a feature vector at time t, $x_{1:T} = \{x_1, \ldots, x_T\}$ is an acoustic observation sequence, q_t is HMM state at time t, which value can range from 1 to N_q (total number of possible HMM states), and q_t^i shows the event "$q_t = i$". In the following, we will drop the M, keeping in mind that all recursions are processed through some prior (Markov) model M.

In standard likelihood-based HMMs, the state gammas $\gamma(i, t)$ can be estimated by using forward α and backward β recursions (as referred to in HMM formalism) [7] using local emission likelihoods $p(x_t|q_i^t)$ (e.g., modeled by GMMs):

$$\begin{aligned} \alpha(i, t) &= p(x_{1:t}, q_t^i) \\ &= p(x_t|q_t^i) \sum_j p(q_t^i|q_{t-1}^j)\alpha(j, t-1) \end{aligned} \tag{2}$$

$$\begin{aligned} \beta(i, t) &= p(x_{t+1:T}|q_t^i) \\ &= \sum_j p(x_{t+1}|q_{t+1}^j)p(q_{t+1}^j|q_t^i)\beta(j, t+1) \end{aligned} \tag{3}$$

thus yielding the estimate of $p(q_t^i|x_{1:T})$:

$$\gamma(i, t) = p(q_t^i|x_{1:T}) = \frac{\alpha(i, t)\beta(i, t)}{\sum_j \alpha(j, T)} \tag{4}$$

Similar recursions, also yielding "state gammas", can be developed for local posterior based systems such as hybrid HMM/ANN systems using MLPs to estimate HMM emission probabilities. In standard HMM/ANN systems, these local posteriors are usually turned into "scaled likelihoods" by dividing MLP outputs by their respective prior probabilities $p(q_t^i)$, i.e.: $\frac{p(x_t|q_t^i)}{p(x_t)} = \frac{p(q_t^i|x_t)}{p(q_t^i)}$. These scaled likelihoods can be used in "scaled alpha" $\alpha^s(i, t)$ and "scaled beta" $\beta^s(i, t)$ recursions to yield gamma estimates [6]. These recursions are similar to the previous recursions except that the likelihood term is replaced by the scaled likelihood.

All these gammas, either computed from local likelihoods or local posteriors, have the same theoretical definition (i.e., posteriors integrating all available acoustic information, as well as possible topological constraints) and thus result in the same theoretical value. However, their estimated values will be different since different local estimators, possibly with different properties, have been used.

2.2. Phoneme posterior (gamma) estimation

The estimated state gammas can then be used to estimate phone posteriors (phone gammas), which in turn can be used to estimate sentence posteriors. In the following, we call these phone posteriors as "phone gammas" $\gamma_p(i, t)$, which can be expressed in terms of state gammas $\gamma(i, t)$ as follows:

$$\begin{aligned} \gamma_p(i, t) &= p(p_t^i|x_{1:T}) = \sum_{j=1}^{N_q} p(p_t^i, q_t^j|x_{1:T}) \\ &= \sum_{j=1}^{N_q} p(p_t^i|q_t^j, x_{1:T})p(q_t^j|x_{1:T}) \\ &= \sum_{j=1}^{N_q} p(p_t^i|q_t^j, x_{1:T})\gamma(j, t) \end{aligned} \tag{5}$$

where p_t is a phone at time t and p_t^i represents the event "$p_t = i$". Probability $p(p_t^i|q_t^j, x_{1:T})$ represents the probability of being in a given phone i at time t knowing to be in the state j at time t. If there is no parameter sharing between phones, this is deterministic and equal to 1 or 0. Otherwise, this can be estimated from the training data. In this work, we assume that there is no parameter sharing between the phones, thus a phone gamma is estimated by adding up all state gammas associated with the phone in the whole model. The posterior estimation for different kinds of phones (context-dependent or context-independent) is basically the same, the difference is only in modeling the phone in the HMM layer used for posterior estimation.

2.3. HMM topologies for posterior probability estimation

The HMM layer used for posterior estimation can have different topologies. The topology of this layer affects the state gamma posteriors since they capture some prior knowledge (such as HMM topology). We have studied two different general topologies:

1. **Ergodic topology:** In this HMM topology, a phone is modeled by one state. Phone models are connected to each other with uniform transition probabilities. Due to the ergodic uniform transition probabilities, we do not make use of specific prior information. Moreover, in this case the state gammas are equal to local normalized (scaled) likelihoods[1] [6]. Therefore, they do not capture contextual information contained in the whole utterance.
2. **Non ergodic topology:** In this topology, a phone is modeled by a left to right, self loop 3-state HMM. Instead of ergodic connection, phone models belonging to each word are connected to make word model and the word models are connected together based on the language model. This is similar to the standard HMM topology normally used for speech recognition. This topology allows capturing prior information encoded in the model topology as well as context information in the whole utterance. Parameters of this model are estimated using the training set.

3. Decoding and recognition

Decoding is performed by a Viterbi decoder using phone gammas as local scores. For each phone, a state is reserved in the decoder structure. Phone states belonging to each word are connected based on phone transition probabilities to make words. Words are also connected based on language model. The local scores in the decoder are phone gammas and the transition

[1] Normalized (scaled) likelihood is the (scaled) likelihood divided by the sum of all classes (scaled) likelihoods.

penalties between states are phone transition probabilities or transition probabilities between words.
We define:

$$V(i,t) = \max_{p_{1:t-1}} p(p_t^i, p_{1:t-1} | x_{1:T}) \quad (6)$$

which can be derived recursively as follows:

$$\begin{aligned} V(i,t) &= \max_{j, p_{1:t-2}} p(p_t^i, p_{t-1}^j, p_{1:t-2} | x_{1:T}) \\ &= \max_j [p(p_t^i | p_{t-1}^j, p_{1:t-2}, x_{1:T}) \cdot \\ &\qquad \max_{p_{1:t-2}} p(p_{t-1}^j, p_{1:t-2} | x_{1:T})] \\ &= \max_j p(p_t^i | p_{t-1}^j, p_{1:t-2}, x_{1:T}) V(j, t-1) \quad (7) \end{aligned}$$

where $p_{1:t} = \{p_1, \ldots, p_t\}$ is the phone sequence. Assuming a left to right model, the term $p(p_t^i | p_{t-1}^j, p_{1:t-2}, x_{1:T})$ can be estimated as:

$$p(p_t^i | p_{t-1}^j, p_{1:t-2}, x_{1:T}) = p(p_t^i | x_{1:T}) p(p_t^i | p_{t-1}^j) \quad (8)$$

The term $p(p_t^i | x_{1:T})$ is the phone gamma and the term $p(p_t^i | p_{t-1}^j)$ is the frame level phone transition probability inside a word (or transition probability between words while going from the last phone in a word to the first phone in the next word)[2]. Simply stated, the decoding process finds the phone sequence (consequently word sequence) having maximum posterior probability. It is generally similar to the decoding in standard HMMs except that the local scores are phone gammas instead of likelihoods.

The whole recognition system is composed of two layers: the posterior probability estimator and the decoder. Figure 1 shows a diagram of the whole system. The first layer is a HMM layer estimating state gamma posteriors using (4), having the whole utterance acoustic features. This layer can have different topologies, such as ergodic or non ergodic as explained in Section 2.3. The state gammas are then integrated to phone gammas using (5). The second layer is a Viterbi decoder which uses phone gammas as local scores. Conceptually, the first layer gets some features as input and acts as a corrective filter by introducing some context and prior knowledge. The prior knowledge has been encoded in the HMM topology. This corrective filter suppresses the effect of features or local posteriors not matching with prior knowledge about the problem or contextual information in the utterance, and magnifies the effect of features or posteriors matching the prior and contextual information. The output of this corrective filter which is in the form of posteriors is fed as local scores to the decoder. The role of the decoder is to make the decision based on these enhanced posterior estimates. The decoder also uses phone and word transition probabilities as transition penalties to force legal phone and word sequences. Note that the value of these transition probabilities will now have a bigger impact than in standard likelihood or scaled likelihood based systems since they are combined with posterior probabilities (ranging between 0 and 1) instead of likelihoods or scaled likelihoods.

Although in this paper we only studied phone level posteriors, this approach provides a theoretical framework for hierarchical estimation, integration and use of posteriors, from the frame level up to the phone and word levels. Word gammas can be estimated basically in the same way as state gammas are integrated into phone gammas. These higher level gammas can also

[2]The transition probabilities are estimated by using forced alignment phone and word level transcriptions of the training set.

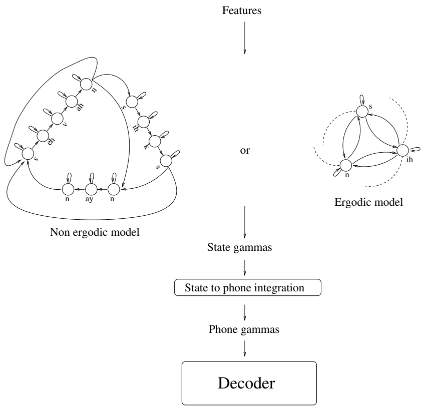

Figure 1: *The whole recognition system. State gammas are estimated through a HMM, then they are integrated into phone gammas. These phone gammas are used as local scores for decoding.*

be used as local scores for a decoder or as features for a standard HMM layer. The ultimate goal is to build a hierarchical processing system, in which each layer enhances and smooths the estimation of posteriors coming from the previous layer by introducing appropriate prior knowledge, context or even auxiliary information, and without ever taking local decisions.

4. Experiments and results

We used OGI Numbers'95 database for connected word recognition task [8]. The training set contains 3'233 utterances spoken by different speakers. The test set contains 1'206 utterances. The vocabulary consists of 31 words (including silence) with a single pronunciation for each word. There are 24 context-independent phones (monophones) and 80 context-dependent phones (triphones) including silence.

The acoustic vector is the PLP cepstral coefficients extracted from the speech signal using a window of 25 ms with a shift of 12.5 ms. At each frame t, 13 PLP coefficients, their first and second order derivatives are extracted resulting in 39 dimensional acoustic vector.

Our system consists of two layers. The first layer is a standard HMM/GMM layer used to estimate state gammas using (4) which are then integrated to triphone gammas using (5). We used the non ergodic topology for this layer to be able to estimate posteriors using context and prior knowledge. This HMM/GMM layer has 12 Gaussians per state, and 3-state left to right, self loop model for each triphone. Parameters of the model are estimated using the training set. Word transition probabilities are considered uniform for the case of numbers database. The estimated triphone gammas are used as local scores in the second layer which is a Viterbi decoder. Table 1 shows the recognition results. The first row in the table shows our system performance. The second row shows the performance of HMM/ANN baseline system and the third row shows the performance for standard HMM/GMM speech recognition system. All the tests were done without using word insertion penalties for decoding. Comparing with hybrid HMM/ANN approach which uses local posteriors (in the form of scaled likelihoods) also as local scores, our system performs better. In our system, we have replaced the local posteriors with enhanced estimates of posteriors taking into account the whole context and prior knowledge. In addition, transition probabilities have more impact in our decoder than hybrid HMM/ANN system due to the combination with the posteriors having the same numerical range, instead of scaled likelihoods. Our system also performs better than the standard HMM/GMM speech recognition system, having the same HMM/GMM model parameters as the one used for estimating posteriors in the first layer of our system. We did the same experiment using monophone models and we got similar conclusions.

Moreover, we found that the new system is less sensitive to changes in word insertion penalties than standard HMM/GMM speech recognition system. The word insertion penalty is a tuning factor to compensate for difference in words length. Since our system uses posteriors instead of likelihoods for decoding, it has less numerical dynamic range and thus, less sensitivity to this factor. This is another advantage of this approach which means it needs less tuning to achieve the best performance.

To understand the role of HMM topology in estimation of posteriors, we did another experiment to compare posteriors estimated through the ergodic and non ergodic model. We used HMM/ANN model once with ergodic topology and the second time with non ergodic topology to estimate posteriors[3]. An MLP with 351 input nodes (9x39 vector) and 24 output units corresponding to the 24 monophones were used to estimate HMM emission probabilities. The same decoder was applied to the estimated posteriors in both cases. Table 2 shows the results of the experiment. The system which uses phone gammas estimated through the non ergodic topology performs significantly better. It shows the fact that the posteriors estimated using non ergodic model are more discriminative than the ones estimated using ergodic model. As mentioned before, ergodic configuration does not capture context or prior knowledge while non ergodic configuration takes the advantage of capturing prior knowledge and the whole utterance context enhancing the estimation of posteriors.

Table 1: Different systems performance

System Configuration	**WER**
Triphone gammas estimated by non ergodic HMM/GMM + Decoder	5.8%
Hybrid HMM/ANN baseline system	6.9%
HMM/GMM baseline system	6.8%

Table 2: Comparing ergodic and non ergodic topologies for posterior estimation

System configuration	**WER**
Monophone gammas estimated by non ergodic HMM/ANN + Decoder	9.4%
Monophone gammas estimated by ergodic HMM/ANN + Decoder	13.3%

5. Conclusions and future work

In this paper, we proposed a new, principled, theoretical framework for estimation, integration and use of posterior probabilities in automatic speech recognition systems. We explained how the posterior estimation can be enhanced taking into account all possible information present in the data (whole acoustic context), as well as possible prior information (e.g. topological constraints).

We used these posteriors as local scores in a Viterbi decoder. We showed our system performs better as compared to the hybrid HMM/ANN approach (which uses local posteriors also as local scores). Therefore, we proposed here to replace the local posteriors with new, enhanced estimates of posteriors. We also showed how the HMM topology can affect the estimation of posteriors.

The enhanced estimates of posteriors can also be used as features (e.g. for a standard HMM layer [6]). The layer used for estimating posteriors acts as a corrective filter using contextual and prior knowledge about the problem, thus resulting in more efficient features, in the form of posteriors.

This theoretical framework allows designing optimal hierarchical HMM structures [9] since it proposes a principled way to introduce appropriate context and prior knowledge in each level of hierarchy. The final goal is to make a hierarchical processing system, in which each level enhances and smooths the estimation of posteriors coming from the previous level by introducing appropriate context and prior knowledge.

6. Acknowledgments

This work was supported by the EU 6th FWP IST integrated project AMI. The authors want to thank the Swiss National Science Foundation for supporting this work through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The authors also would like to thank Hynek Hermansky, Mathew Magimai Doss and Hemant Misra for helpful discussions.

7. References

[1] Bourlard, H. and Morgan, N., "Connectionist Speech Recognition – A Hybrid Approach", Kluwer Academic Publishers, 1994.

[2] Mangu, L., Brill, E., and Stolcke, A., "Finding consensus in speech recognition: word error minimization and other applications of confusion networks", *Computer, Speech and Language*, Vol. 14, pp. 373-400, 2000.

[3] Abdou, S. and Scordilis, M.S., "Beam search pruning in speech recognition using a posterior-based confidence measure", *Speech Communication*, Vol. 42, pp. 409-428, 2004.

[4] Bernardis, G. and Bourlard, H., "Improving posterior confidence measures in hybrid HMM/ANN speech recognition system", *Proc. ICSLP*, pp. 775-778, 1998.

[5] Hermansky, H., Ellis, D.P.W., and Sharma, S., "Connectionist Feature Extraction for Conventional HMM Systems", *Proc. ICASSP*, 2000.

[6] Bourlard, H., Bengio, S., Magimai Doss, M., Zhu, Q., Mesot, B., and Morgan, N., "Towards using hierarchical posteriors for flexible automatic speech recognition systems", *DARPA RT-04 Workshop*, November 2004.

[7] Rabiner, L. R., "A tutorial on hidden Markov models and selective applications in speech recognition", *Proc. IEEE*, vol. 77, pp. 257-286, 1989.

[8] Cole, R. A., Fanty, M., Noel, M., and Lander, T., "Telephone speech corpus development at CSLU", *Proc. ICSLP*, 1994.

[9] Oliver, N., Horvitz, E., and Garg, A., "Layered representations for learning and inferring office activity from multiple sensory channels", *Proc. ICMI*, 2002.

[3] HMM/GMM configuration was not used for this experiment since *ergodic* HMM/GMM yields normalized likelihoods (estimated by GMMs) as gammas [6] which are not discriminative for decoding, while ergodic HMM/ANN takes the advantage of MLP discriminant acoustic modeling, yielding more discriminant normalized scaled likelihoods.