USING MORE INFORMATIVE POSTERIOR PROBABILITIES FOR SPEECH RECOGNITION

Hamed Ketabdar, Jithendra Vepa, Samy Bengio and Hervé Bourlard

IDIAP Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{ketabdar, vepa, bengio, bourlard}@idiap.ch

ABSTRACT

In this paper, we present initial investigations towards boosting posterior probability based speech recognition systems by estimating more informative posteriors taking into account acoustic context (e.g., the whole utterance), as well as possible prior information (such as phonetic and lexical knowledge). These posteriors are estimated based on HMM state posterior probability definition (typically used in standard HMMs training). This approach provides a new, principled, theoretical framework for hierarchical estimation/use of more informative posteriors integrating appropriate context and prior knowledge. In the present work, we used the resulting posteriors as local scores for decoding. On the OGI numbers database, this resulted in significant performance improvement, compared to using MLP estimated posteriors for decoding (hybrid HMM/ANN approach) for clean and more specially for noisy speech. The system is also shown to be much less sensitive to tuning factors (such as phone deletion penalty, language model scaling) compared to the standard HMM/ANN and HMM/GMM systems, thus practically it does not need to be tuned to achieve the best possible performance.

1. INTRODUCTION

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 use 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 systems. Hybrid HMM/ANN method allows for discriminant training, as well as the possibility of using small acoustic context by presenting a few number of 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, the most successful approach is Tandem [5]. In Tandem, MLP estimated posteriors are used as input features for a standard HMM/GMM configuration. Tandem 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, posteriors are estimated based only on the information in local frame or a limited number of local frames. In [6, 7], a method was presented to estimate more informative posteriors based on HMM state posterior probability definition (usually used in HMMs training) to estimate posteriors taking into account all acoustic information available in each utterance, as well as prior knowledge, possibly formulated in terms of HMM topological constraints. This approach provides a new, principled, theoretical framework for hierarchical estimation, integration and use of more informative posteriors, from the frame level up to the phone and word levels. They investigated the estimation and usage of these posteriors as features for a standard HMM/GMM layer. Such an approach was shown to yield significant performance improvement over Tandem approach on Numbers'95 and on a reduced vocabulary version of the DARPA Conversational Telephone Speech-to-text (CTS) task. In [8], these new posteriors were used as local scores for decoding and the resulting system was favorably compared with a standard HMM/GMM system.

In the present paper, we continue investigating the estimation and use of these more informative posteriors as scores for decoding. However, compared to the previous work [8], here we compare the new posteriors with MLP estimated posteriors, and explore some additional new aspects of the system such as sensitivity and stability to tuning, as well as the behavior and more efficiency of the method when there is a lack of clear acoustic information (noisy speech). In our system, the new more informative posteriors are estimated from MLP estimated posteriors by introducing prior and contextual knowledge. We then use these more informative posteriors for decoding. Therefore, comparing with hybrid HMM/ANN approach which uses MLP estimated posteriors for decoding, we use more informative posteriors for decoding. We have shown that these posteriors perform significantly better than MLP estimated posteriors for decoding (hybrid HMM/ANN approach) for clean and noisy speech. We also show that the relative improvement is higher for more noisy speech. Since some acoustic information are lost in noisy speech, the role of integrating prior knowledge in getting more informative posteriors is more evident. It confirms that integration of prior knowledge can compensate the lack of clear acoustic information. The resulting system is also much less sensitive to tuning factors (such as phone deletion penalty, language model scaling), which are usually required in standard HMM/ANN or HMM/GMM systems for numerical compensation during decoding. Therefore, practically it does not need to be tuned to reach 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 and the complete recognition system using these posteriors. Experiments and results are presented in Section 4. Conclusions and future work plans are discussed in Section 5.

2. INTEGRATING PRIOR AND CONTEXTUAL INFORMATION IN POSTERIOR ESTIMATION

In this section, we study how more informative posteriors can be estimated by integrating possible prior knowledge, as well as acoustic

context information (e.g., using the whole utterance). The basic idea as studied in [6, 7, 8] is to estimate posteriors based on HMM state posterior probability definition (as usually used in HMMs training). According to the standard HMM formalism, this posterior 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 the HMM state at time t, which value can range from 1 to N_q (total number of 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. We call $\gamma(i,t)$ as "state gamma posterior" or simply "state gamma".

The state gammas $\gamma(i,t)$ can be estimated by using forward α and backward β recursions (as referred to in HMM formalism) [9] using local emission likelihoods $p(x_t|q_i^t)$ (e.g., modeled by GMMs):

$$\begin{aligned}\alpha(i,t) &= p(x_{1:t}, q_t^i) \\ &= p(x_t|q_t^i)\sum_j p(q_t^i|q_{t-1}^j)\alpha(j,t-1) \qquad (2)\\ \beta(i,t) &= p(x_{t+1:T}|q_t^i) \\ &= \sum_j p(x_{t+1}|q_{t+1}^j)p(q_{t+1}^j|q_t^i)\beta(j,t+1) \qquad (3)\end{aligned}$$

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 [1].

The estimated state gammas can then be used to estimate phone posteriors or higher level posteriors. 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:

$$\begin{aligned}\gamma_p(i,t) &= p(p_t^i|x_{1:T}) = \sum_{j=1}^{N_q} p(p_t^i, q_t^j|x_{1:T}) \qquad (5)\\ &= \sum_{j=1}^{N_q} p(p_t^i|q_t^j, x_{1:T})p(q_t^j|x_{1:T}) \qquad (6)\\ &= \sum_{j=1}^{N_q} p(p_t^i|q_t^j, x_{1:T})\gamma(j,t) \qquad (7)\end{aligned}$$

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 phones, thus a phone gamma is estimated by adding up all state gammas associated with the phone in the whole model.

Although in this paper we only study phone level posteriors, this posterior estimation/integration 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. The ultimate goal is to build a hierarchical processing system, in which each layer enhances the estimation of posteriors coming from the previous layer by introducing appropriate prior knowledge, context or even auxiliary information.

The HMM layer used for gamma posterior estimation can have different topologies, thus encoding different types of prior knowledge. As the simplest case, we can model each phone with a minimum number of states and connect phone models with ergodic uniform transition probabilities. In this case, only the prior knowledge about minimum duration of phones is introduced in the posterior estimation. We can do one more step and use real estimated phone transitions instead of ergodic transitions between phone models. In this case, we can also introduce some phonetic prior knowledge. Finally, we can have a fully constrained model composed of connected word models made by phone models, and each phone modeled by a minimum number of states. The parameters of this model are estimated from the training set. This topology can integrate phonetic and lexical knowledge in the posterior estimation.

3. DECODING AND RECOGNITION

Decoding is performed by a Viterbi decoder (NOWAY decoder [10]) using phone gammas as local scores. For each phone, 3 states are reserved in the decoder structure. Phone models belonging to each word are connected to make words. Words are also connected based on the language model. The local scores in the decoder are phone gammas and the transition penalties between states are state, phone or word transition probabilities.

The whole recognition system is composed of three layers which are shown in Figure 1. The first layer is an MLP or GMM layer which estimates initial evidences for phones in the form of posteriors or likelihoods. The second Layer is a HMM layer which integrates prior and contextual knowledge by using the initial evidences in forward and backward HMM recursions (Eq. 2, 3) to get the estimate of gamma state posteriors (Eq. 4). These state gamma posteriors are integrated into phone gammas using Eq. 7, then they are used as local scores in the last layer which is a decoding layer. Conceptually, the second layer gets phone initial evidences as input and acts as a corrective filter by introducing some context and prior knowledge. The prior knowledge has been encoded in the topology of HMM in this layer. The corrective filter suppresses the effect of evidences not matching with prior knowledge or contextual information, and magnifies the effect of evidences matching them. The output of this corrective filter is more informative evidences in the form of posteriors. The decoder makes decision about the word sequence based on this more informative posteriors.

4. EXPERIMENTS AND RESULTS

In this section, we compare the gamma posteriors with MLP posteriors (for clean and noisy speech) to investigate the role of integrating prior and contextual information in estimating more informative posteriors. We also compare and discuss the sensitivity of gamma posterior based system and MLP posterior based system to tuning factors (e.g. phone deletion penalty, scaling of the language model).

We did two sets of experiments to investigate different aspects of our gamma posterior based system. In the first set of experiments, we compare our system with the state-of-the-art hybrid HMM/ANN method in which MLP estimated posteriors are used as scores for decoding. The configuration of our system is the same as explained

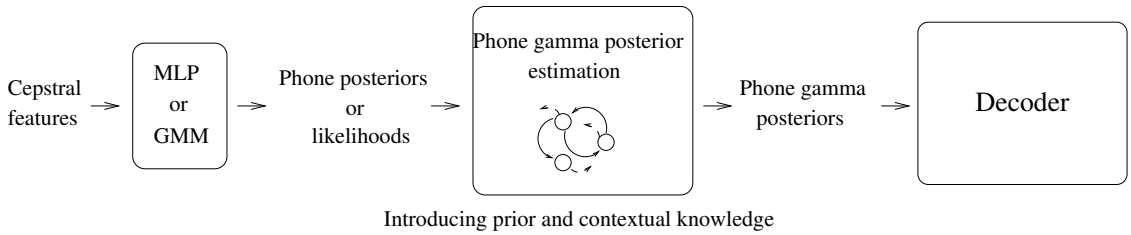

Fig. 1. *The whole recognition system. First, initial phone evidences are estimated using GMMs or MLPs, then these evidences are used to estimate gamma state posteriors through a HMM, which are then integrated into phone gammas. Finally, phone gammas are used as local scores for decoding.*

in Section 3. In this system, the MLP estimated initial posteriors are used in HMM forward and backward recursions to get gamma state posteriors. These more informative posteriors are then used as scores, instead of MLP estimated posteriors for decoding. Therefore, the difference between our system and the hybrid HMM/ANN system is in the posteriors used for decoding. The former uses more informative posteriors estimated from MLP posteriors by integrating prior and contextual knowledge, while the latter uses directly MLP estimated posteriors for decoding. For the experiments in this paper, we used a fully constrained model (as explained in Section 2) to get estimates of gamma posteriors. This means we integrate lexical and phonetic knowledge in the posterior estimation. The decoder structure was explained in Section 3 and it is the same for both systems.

We used OGI Numbers'95 database for connected word recognition task [11]. The training set contains 3330 utterances spoken by different speakers. The test set contains 2250 utterances (8688 words). The vocabulary consists of 31 words with a single pronunciation for each word. There are 27 context-independent phones (monophones). The acoustic vector is the PLP cepstral coefficients extracted from the speech signal using a window of 32 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. An MLP with 351 input nodes (9x39 vector, corresponding to the concatenation of 9 frames of 39 dimensional acoustic vector) and 27 output units corresponding to the 27 monophones were used to estimate initial posteriors.

Table 1 compares the performance of the two systems (gamma based system and hybrid HMM/ANN system) for clean speech as well as different levels of factory noise (the numbers appearing in the second column inside brackets will be explained in the next paragraph). It is clear that the decoder which uses gamma posteriors performs significantly better than the one which uses MLP estimated posteriors (hybrid method) [1]. It is also interesting to observe that the relative improvement increases by increasing the noise level. This implies that integrating prior and contextual knowledge can be even more useful when there is no clear acoustic information, because it provides extra knowledge which can compensate the lack of acoustic information.

The second set of experiments compares the sensitivity of the two mentioned systems to tuning factors (e.g. phone deletion penalty). Phone deletion penalty (or word deletion penalty which comes from the same idea) is a tuning factor and an engineering trick which is used for numerical compensation of scores for different paths during decoding [12]. It can significantly affect the recognition performance of standard HMM/ANN and HMM/GMM systems[2]. In order to compare the sensitivity of the systems, we vary the phone deletion penalty value in the decoder and observe the change of performance for two systems. Figure 2 shows the results. Comparing the two curves, we can conclude that the gamma based system is much less sensitive to tuning than the standard hybrid HMM/ANN system. It can be explained by the fact that gamma posteriors tend to have very close to binary values (like a decision) because they are estimated by integrating some extra knowledge, while the MLP posteriors can change more smoothly between 0 and 1, thus the accumulated scores obtained by gamma posteriors during decoding tend to be discrete while it is continuous for the case of MLP posteriors. Tuning which slightly changes the scores can affect the decision made based on continuous scores more than the one made based on discrete scores. This is another advantage of the gamma based approach which means it needs much less tuning to achieve the best performance. Moreover, the numbers inside brackets in the second column of Table 1 show the recognition rates of the MLP posterior based system when it is tuned to reach the best performance. Again, you can see how the performance of MLP posterior based system can be sensitive and rely on tuning to reach the best, which is not the case for gamma based system. The sensitivity of the gamma based system to tuning is also much less than standard HMM/GMM systems using likelihoods for decoding. The same less sensitivity properties was also observed to scaling of language model (another tuning factor) for gamma based system comparing with standard HMM/ANN and HMM/GMM systems.

Table 1. *Comparing word recognition performance (in %) after decoding, for MLP estimated posteriors and gamma posteriors*

Noise level	MLP posterior	Gamma posterior	Relative improvement
Clean	86.6 *(90.0)*	**90.8**	4.8
SNR 12	79.0 *(82.3)*	**84.5**	7.0
SNR 6	65.5 *(70.4)*	**74.1**	13.0
SNR 0	42.8 *(49.1)*	**52.7**	23.0

[1] Better performances can even be obtained if context-dependent phone (triphone) posteriors are estimated instead of monophone posteriors [8], but training MLP for triphones is computationally expensive (particularly for larger databases) and it will not lead to new conclusions.

[2] Usually this factor is tuned using a development set to get maximum performance, which does not guarantee the same improvement on the test set, specially if the conditions (e.g. noise level, task, etc.) change. Sometimes it is even tuned over the test set which is an incorrect practice as it shows optimistically biased results! In any case, there is no strong theoretical explanation for tuning, it makes the system less robust against changes and it is time consuming.

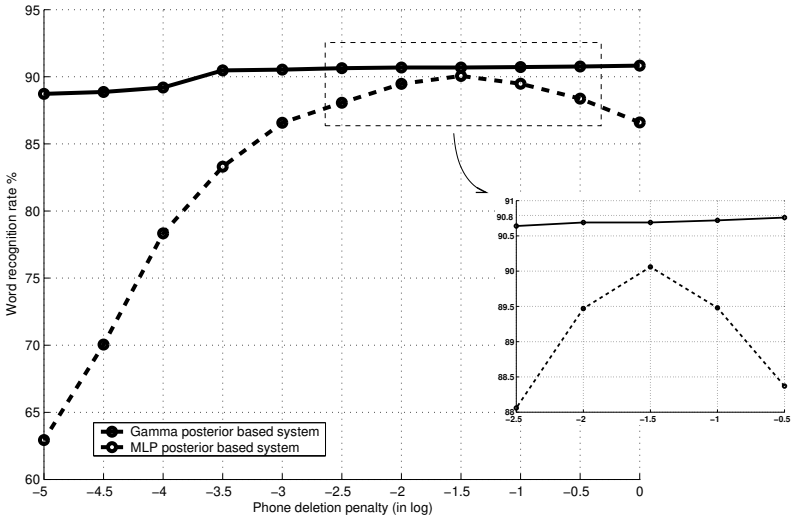

Fig. 2. *Comparing the sensitivity of gamma posterior based system and MLP posterior based system to tuning phone deletion penalty. The diagram inside is a zoom of performance curves for small values of phone deletion penalty (fine tuning).*

5. CONCLUSIONS

In this paper, we proposed a new, principled, theoretical framework for estimation, integration and use of more informative posterior probabilities in automatic speech recognition systems. It is explained how these more informative posteriors can be estimated by taking into account all possible information present in the data (whole acoustic context), as well as possible prior information (e.g. phonetic and lexical knowledge). The new posterior estimation theoretical framework also allows designing optimal hierarchical HMM structures such as proposed in [13] since it accommodate a principled way to introduce appropriate context and prior knowledge in each level of hierarchy.

We used these posteriors as local scores in a Viterbi decoder. It is shown that these posteriors perform significantly better than MLP posteriors (hybrid HMM/ANN approach) for clean and more specially for noisy speech. We observed that the relative improvement is higher for more noisy speech which confirms that integrating prior and contextual knowledge can compensate the lack of clear acoustic information. It was also shown that the proposed system is much less sensitive to tuning (e.g. phone deletion penalty) comparing to the standard HMM/ANN and HMM/GMM systems, resulting in a system which practically does not need to be tuned to reach the best possible performance.

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 like to thank Hynek Hermansky 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] Ketabdar, H., Bourlard, H., Bengio, S., "Hierarchical Multi-Stream Posterior Based Speech Recognition System", *MLMI'05 Workshop*, July 2005.

[8] Ketabdar, H., Vepa, J., Bengio, S., and Bourlard, H., "Developing and enhancing posterior based speech recognition systems", *IDIAP RR 05-23*, 2005.

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

[10] Renals, S., Hochberg, M., "Efficient search using posterior phone probability estimates", *Proc. ICASSP'95*, Detroit, USA, 1995.

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

[12] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell G., Ollason, D., Povey, D., Valtchev, V., Woodland, P., "The HTK Book", http://htk.eng.cam.ac.uk/docs/docs.shtml.

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