IDIAP RESEARCH REPORT

In-Context Phone Posteriors as Complementary Features for Tandem ASR

IDIAP–RR 08-43

June 2008

PUBLISHED IN
Proc. ICSLP'07

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 08-43

In-Context Phone Posteriors as Complementary Features for Tandem ASR

June 2008

published in
Proc. ICSLP'07

Abstract. In this paper, we present a method for integrating possible prior knowledge (such as phonetic and lexical knowledge), as well as acoustic context (e.g., the whole utterance) in the phone posterior estimation, and we propose to use the obtained posteriors as complementary posterior features in Tandem ASR configuration. These posteriors are estimated based on HMM state posterior probability definition (typically used in standard HMMs training). In this way, by integrating the appropriate prior knowledge and context, we enhance the estimation of phone posteriors. These new posteriors are called 'in-context' or HMM posteriors. We combine these posteriors as complementary evidences with the posteriors estimated from a Multi Layer Perceptron (MLP), and use the combined evidence as features for training and inference in Tandem configuration. This approach has improved the performance, as compared to using only MLP estimated posteriors as features in Tandem, on OGI Numbers , Conversational Telephone speech (CTS), and Wall Street Journal (WSJ) databases.

1 Introduction

Over the past 10 to 15 years, posterior probability based approaches have become popular for boosting speech processing systems. The posterior based systems can be categorized mainly to either the approaches which use posteriors as local scores (measures), or the approaches which use posteriors as features. Hybrid Hidden Markov Model / Artificial Neural Network (HMM/ANN) method [1] is one of the the first methods to use posterior probabilities as local scores. In this method, ANNs (more specifically Multi-Layer Perceptrons, MLPs) are used to estimate the emission probabilities required in HMM. Hybrid HMM/ANN method provides the possibility of discriminant training, as well as 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]. Considering the use of posterior probabilities as features, the most successful approach is Tandem [5]. In Tandem, MLP estimated phone posteriors are used as input features for training/inference in 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.

Conventionally, as in both hybrid HMM/ANN and Tandem approaches, posteriors are estimated based only on the acoustic information in a local frame or a limited number of local frames. We call these posteriors as MLP estimated posteriors. In [6, 7], we have presented a method to estimate more informative posteriors by integrating prior knowledge (such as phonetic and lexical constraints), and contextual knowledge in the posterior estimation. These posteriors are estimated based on HMM state posterior probability definition (usually used in HMMs training). The prior knowledge is formulated in terms of HMM topological constraints. This approach provides a principled framework for estimating more informative posteriors taking into account prior and contextual knowledge. We call these posteriors as 'in-context' or HMM posteriors. In [6], we have used these posteriors as local scores for decoding. There we could show that the usage of these posteriors improves the performance for decoding as compared to the MLP posteriors. We also showed that the decoder using these posteriors is less sensitive to tuning parameters such as insertion penalties, scaling the language model, etc.

In the present paper, compared to our previous work, we take a new direction for using these more informative 'in-context' posterior estimates, and we investigate how these posteriors can help to provide better evidences, this time as features for a training and inference layer (e.g. Tandem configuration). In the other words, we study how we can use these posteriors to boost the performance of a system such as Tandem which conventionally uses MLP estimated phone posteriors as features. We propose to use the 'in-context' posteriors as complementary features along with the MLP posteriors in Tandem configuration. We combine these two types of posteriors and use the combined evidence for training and inference in Tandem HMM/GMM layer. It is shown that when we use the 'in-context' posterior estimates as complementary features, we improve the performance of the recognizer as compared to the use of only MLP posteriors as features, on different small and large vocabulary tasks. This shows that the prior knowledge and context encoded in the posterior estimates can help for enhancing features in Tandem configuration.

Section 2 studies the MLP based posterior estimation, and the method for integrating prior and contextual knowledge to estimate 'in-context' posteriors. Section 3 talks about the usage of more informative 'in-context' posteriors as complementary features/evidences. Experiments, results and comparisons are presented in Section 4. Finally, the conclusions appear in Section 5.

2 MLP based posteriors and 'in-context' (HMM) posteriors

2.1 MLP based phone posteriors

As already mentioned, the phone posteriors are conventionally estimated only based on information in a local or limited span of spectral feature frames. Among different approaches for estimating phone posteriors, ANNs and more specifically Multi Layer Perceptrons (MLPs) provide a discriminative way

of estimating phoneme posteriors. The MLP, trained on the training part of the database, estimates the posterior probabilities of phoneme classes at each frame $p(q_t^i|x_t)$, where q_t is a phoneme at time t, and q_t^i is the event $q_t = i$. We call these posteriors as MLP estimated posteriors in this paper. These posteriors can be possibly used as features for training and inference in a HMM/GMM layer, as they are used in Tandem configuration [5].

2.2 'In-context' (HMM) phone posteriors

The time limited spectral information is not the only source of knowledge available for a specific phoneme. Usually other sources of knowledge, such as prior phonetic and lexical knowledge, and long context can provide additional information about phonemes, and possibly help to enhance the estimation of phone posteriors. Information about phones are spread over time in the speech signal and there is no sharp boundaries between phonemes, therefor taking into account long contextual information can be useful. Moreover, some prior knowledge such as duration of phonemes and the lexical usage of phonemes in a word can also help to have better evidence estimates for phonemes.

In [6], we have studied how such prior phonetic and lexical knowledge, as well as long acoustic context information can be integrated in the phone posterior estimation, in order to enhance the estimates. The basic idea 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 posterior".

The state posterior $\gamma(i,t)$ can be estimated by using forward α and backward β recursions (as referred to in HMM formalism) [8] using local emission probabilities $p(x_t|q_i^t)$ (e.g., modeled by GMMs or MLPs) [8]:

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

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

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 can be developed for local posterior based systems such as hybrid HMM/ANN systems using MLPs to estimate HMM emission probabilities [1].

If we assume that a phoneme is represented by one state (q) in our HMM configuration, then $\gamma(i,t) = p(q_t = i|x_{1:T}, M)$ is the 'in-context' or HMM phone posterior for phone i at time t. Otherwise if a phoneme is modeled with more than one HMM state, the in context phoneme posterior can be simply estimated by adding up posteriors of all states composing the phone in the HMM (for more details refer to [6]).

The type of prior knowledge which is integrated is specified by the HMM topological constraints. Here, as the simplest case, we model each phone with a minimum number of states (i.e. minimum

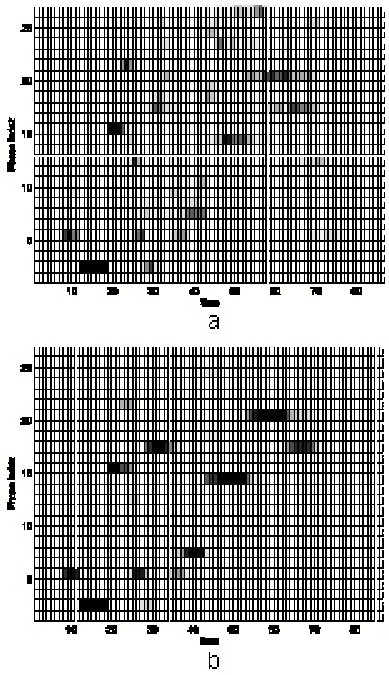

Figure 1: *MLP estimated phone posteriors (a) and corresponding 'in-context' phone posteriors (b). The y-axis is showing phone labels and x-axis is showing frames. Intensity of each block shows posterior value.*

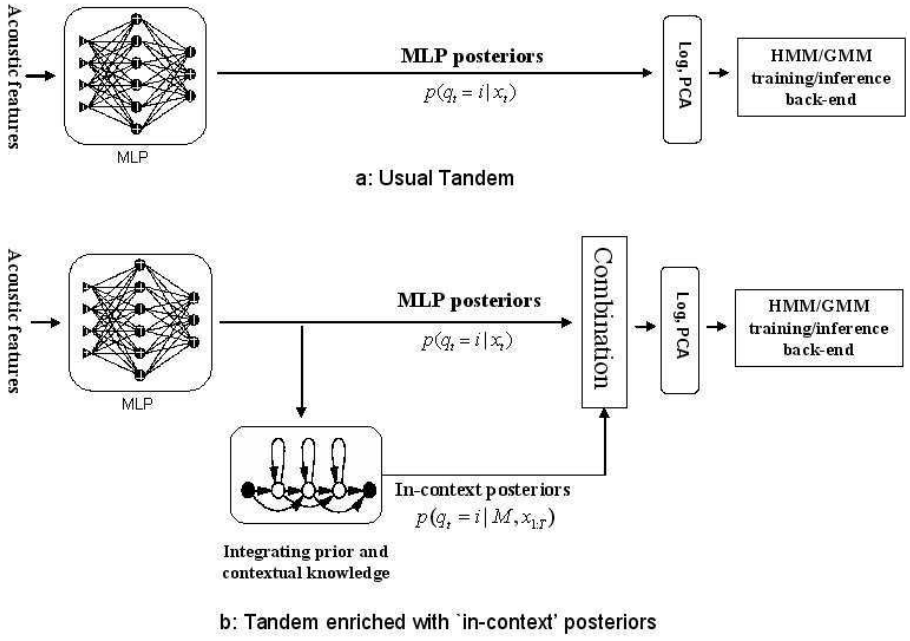

Figure 2: *Usual Tandem, and Tandem system using 'in-context' posteriors as complementary features.*

duration) and connect phone models with ergodic uniform transition probabilities. In this case, the knowledge about minimum duration of phones is introduced in the posterior estimation. Phone models can be also connected based on word lexical constraints. This topology integrates lexical knowledge in the posterior estimation.

As already mentioned, the MLP posteriors can be used as the HMM emission probabilities in the above recursions (2-4) to estimate the 'in-context' posteriors [1]. In this case, the 'in-context' posteriors can be considered as an enriched version of the MLP posteriors. Conceptually, the HMM layer gets phone initial evidences (MLP posteriors) as input and acts as a corrective filter by introducing context and prior knowledge. 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 the corrective filter is referred to as 'in-context' posteriors. Fig. 1 shows an example of enhancing and integrating extra knowledge in the posterior estimates. The upper plot is showing posteriors for different phone classes over time for an utterance, estimated using MLP. The lower plot shows the posteriors for the same utterance, estimated as explained in Section 2.2 by integrating prior and contextual knowledge. The prior knowledge imposed in this case is a minimum duration of 3 frames for each phoneme. The local estimators (emission probabilities) for the HMM estimating 'in-context' posteriors are the MLP posteriors. Therefore, the 'in-context' posteriors in the lower plot can be considered as enhanced estimates of the MLP posteriors in the upper plot. The 'in-context' posteriors look less noisy and more smooth.

3 In-context posteriors as complementary features

In the previous section, we have shown that we can enhance the posterior estimates by integrating prior and contextual knowledge. Besides the advantages of integrating extra knowledge for enhancing the estimation of posteriors, it should be noticed how and to what extent the extra knowledge is reliable. Although the prior knowledge is assumed to be usually correct, but as the name "prior" suggests, there can be few cases in which the true data is not matching the prior knowledge. For example, the assumed lexical knowledge may not include some rare but truly existing pronunciation variants for a word, while such a cases may appear in data. In these cases, the enhanced posteriors start deviating from the MLP posteriors and they may not represent the data correctly. Therefor, although prior knowledge helps in many cases to improve the estimation of posteriors, there can be some cases in which the resulting posteriors are not matching the data. This means there is a trade off between the smoothness obtained by integrating prior knowledge, and deviation from the real data. Considering the possible risk of deviation form the real data due to the partially incorrect prior knowledge, a safe compromise is using the 'in-context' posteriors as complementary features along with the original MLP posteriors. In the other words, the 'in-context' posteriors should be combined with the MLP posteriors. Considering a configuration similar to Tandem, the combined evidences are then used as features for training and inference. This way, the raw evidences (MLP posteriors) representing the data are preserved, while there is also the access to the posteriors enriched by the prior knowledge.

Fig. 2 is showing a diagram of the normal Tandem using MLP posteriors as features, and Tandem system using 'in-context' posteriors as complementary features along with the MLP posteriors. The emission probabilities in the HMM layer integrating prior knowledge are provided by the MLP. The 'in-context' posteriors are obtained by processing MLP posteriors in the HMM layer to integrate prior and contextual knowledge. The conventional Tandem configuration uses only the MLP estimated posteriors as features for training/inference in the HMM/GMM layer, while in our method, the 'in-context' posteriors are combined as complementary features with the MLP posteriors. The combined evidence is then used as features for the HMM/GMM layer.

We have studied addition (average), multiplication, and concatenation as the combination rules. In case of addition (average), the combined evidence is written as:

$$C_t^i = \frac{p(q_t = i|x_t) + p(q_t = i|M, x_{1:T})}{2} \quad (5)$$

and in case of multiplication:

$$C_t^i = p(q_t = i|x_t)p(q_t = i|M, x_{1:T}) \quad (6)$$

where C_t^i shows the combined evidence for phone i at frame t. In case of concatenation rule, the MLP and 'in-context' posterior vectors at frame t are concatenated. The dimension of the resulting vector is reduced by applying Principal Component Analysis (PCA).

4 Experiments and results

In Section 3, we have suggested to use the enhanced 'in-context' posterior estimates as complementary posterior features for Tandem configuration. In this section, we compare the results of recognition studies for the normal Tandem, which uses only MLP posteriors as features, and the Tandem system which combines 'in-context' posteriors with the MLP posteriors, and use the combined evidence for training/inference. The two configurations are shown in Fig. 2, and details of implementation is as explained in the previous section. The prior knowledge used to obtain 'in-context' posteriors is the phonetic duration knowledge. This was achieved by considering 3 states per phoneme model in the HMM layer integrating prior knowledge. We have tried three combination rules for combining the two posterior streams: Addition (average), multiplication, and concatenation.

Results are presented on OGI Numbers [9], Conversational Telephone Speech (CTS) [10], and Wall Street Journal (WSJ) [11] Databases. For the OGI Numbers database, we have used a MLP with 351 input, 1800 hidden, and 27 output (corresponding to the number of phones) nodes. There are 31 words and 27 phonemes in the database. The training set is about 1.5 hours and the test set is about 0.9 hour. For CTS database, we have used an MLP with 351 input, 2000 hidden, and 46 output nodes (corresponding to the number of phones). There are 46 phones and 500 words in this database. The training set is about 7 hours and the test set is about 0.6 hours. For WSJ database we have used an MLP with the same size as for CTS database. There are 46 phones and 5k words in this database. The training set size is about 70 hours and the test size is about 1.1 hours. The implemented HMM/GMM layer [12] uses triphone models for Numbers and CTS databases, and monophone models for WSJ database.

Table 1. is showing the results of the recognition studies in terms of word error rate for the two mentioned configurations, using different databases. The results are presented for the best combination rules (which was addition for Numbers and WSJ, and multiplication for CTS). It shows that the combined evidence obtained from the two stream of posteriors performs better than the MLP posterior features alone, for all the databases. Using 'in-context' posteriors encoding prior and contextual knowledge in combination with MLP posteriors has helped to provide better evidences as features for Tandem. The addition rule was working the best for two out of three databases. This can be due to the more robustness of the addition rule with respect to the errors in the estimation of posteriors in the two streams [13].

Table 1: Recognition results for MLP posteriors (usual Tandem), and MLP posteriors combined with 'in-context' posteriors in Tandem configuration.

Database	MLP posterior	MLP & 'in-context' posterior
Numbers'95	4.7%	**4.3%**
CTS	54.2%	**51.6%**
WSJ	34.6%	**32.5%**

5 Conclusions

In this paper, we studied a method for boosting the performance of the Tandem ASR configuration, by providing phone posteriors enriched and enhanced by integrating prior and contextual knowledge. We saw how these new 'in-context' posteriors can be estimated, and then used in combination with the MLP posteriors as complementary features in the Tandem configuration. The combined evidence leads to better performance compared to only use of MLP posteriors as features. This shows that integrating prior and contextual knowledge can help to provide more informative evidences for phonemes.

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 for helpful discussions.

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] Ketabdar, H., Vepa, J., Bengio, S., and Bourlard, H., "Using more informative posterior probabilities for speech recognition", in "IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)", 2006.

[7] Ketabdar, H., Bourlard, H., Bengio, S., "Hierarchical Multi-Stream Posterior Based Speech Recognition System", *MLMI'05 Workshop*, July 2005.

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

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

[10] Zhu, Q., Chen, B., Morgan, N., Stolcke, A. "On Using MLP Features in LVCSR", ICSLP 2004, Korea.

[11] Fransen, J., Pye, D., Robinson, T., Woodland, P., Young, S., "WSJCAM0 corpus and recording description", Technical Report 192, Cambridge University Engineering Department, 1994.

[12] Young, S.J., Kershaw, D., Odell, J.J., Ollason, D., Valtchev, V., and Woodland, P.C., "The HTK Book (for HTK version 2.2).", Entropic Ltd., Cambridge, England, 1999.

[13] Kittler, J., Hatef, M., Duin, R., Matas, J., "On Combining Classifiers", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, Vol. 20, No. 3, March 1998.