HIERARCHICAL INTEGRATION OF PHONETIC AND LEXICAL KNOWLEDGE IN PHONE POSTERIOR ESTIMATION

Hamed Ketabdar and Hervé Bourlard

IDIAP Research Institute, Martigny, Switzerland
Swiss Federal Institute of Technology at Lausanne (EPFL)

ABSTRACT

Phone posteriors has recently quite often used (as additional features or as local scores) to improve state-of-the-art automatic speech recognition (ASR) systems. Usually, better phone posterior estimates yield better ASR performance. In the present paper we present some initial, yet promising, work towards hierarchically improving these phone posteriors, by implicitly integrating phonetic and lexical knowledge. In the approach investigated here, phone posteriors estimated with a multilayer perceptron (MLP) and short (9 frames) temporal context, are used as input to a second MLP, spanning a longer temporal context (e.g. 19 frames of posteriors) and trained to refine the phone posterior estimates. The rationale behind this is that at the output of every MLP, the information stream is getting simpler (converging to a sequence of binary posterior vectors), and can thus be further processed (using a simpler classifier) by looking at a larger temporal window. Longer term dependencies can be interpreted as phonetic, sub-lexical and lexical knowledge. The resulting enhanced posteriors can then be used for phone and word recognition, in the same way as regular phone posteriors, in hybrid HMM/ANN or Tandem systems. The proposed method has been tested on TIMIT, OGI Numbers and Conversational Telephone Speech (CTS) databases, always resulting in consistent and significant improvements in both phone and word recognition rates.

Index Terms— Phone posterior estimation, Neural Networks, Temporal posterior context, Phonetic and lexical knowledge, Enhanced phone posteriors.

1. INTRODUCTION

Recently the estimation and usage of posterior probabilities has become popular for boosting the performance of speech recognition 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 Multilayer 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.

In both hybrid HMM/ANN and Tandem approaches, phone posteriors are often simply estimated from the acoustic information available in a local frame or a limited number of local frames using MLPs. The information in a limited window of spectral features in not the only source of knowledge available about phonemes. Information about phonemes are spread over time and there is no sharp boundaries between phonemes. Phonemes have specific duration constraints (phonetic knowledge), follow specific sub-lexical and lexical rules (lexical knowledge), etc. These extra sources of knowledge are usually not taken into account in (local) phone posterior estimation.

In this paper, we propose a simple yet very effective approach for integrating phonetic and lexical knowledge in the phone posterior estimation. We use a secondary MLP to learn long term dependencies between phone evidences (posteriors) estimated initially by a first MLP. These long term dependencies can be interpreted as phonetic and sub-lexical or lexical knowledge. In the present work, the second MLP has been trained on the same database, but using a longer temporal context of initial phone posteriors as input, and the same phone labels as targets. In the future tough, we expect further improvements by training the two MLPs on different databases. Another alternative will also be to use the second MLP only for (task) adaptation purposes. The second net is able to integrate the learned phonetic and lexical knowledge in the phone posterior estimation. This leads to enhanced phone posteriors as compared to initial phone posteriors. Since the posterior vector sequences (sometimes referred to as "posteriogram") resulting of the first MLP is much simpler (converging to a sequence of binary vectors), the second MLP can extract longer-term information and act as a filter, smoothing out evidences which are not matching the learned phonetic and lexical knowledge.

We have compared the performance of the enhanced phone posteriors and initial posteriors for phone and word recognition, in Tandem and hybrid HMM/ANN configurations. We have used TIMIT [6], OGI Numbers [7], and CTS [8] databases for the experiments. It is shown that the enhanced posteriors perform significantly and consistently better than the initial phone posteriors for phone and word recognition over different databases, and different configurations.

The paper is organized as follows: Section 2 briefly discusses the MLP-based phone posterior estimation. Section 3 presents the method for integrating longer-term temporal span (implicit phonetic and lexical knowledge), and enhancing posterior estimation. Section 4 describes phone and word evaluation experiments on TIMIT, OGI Numbers, and CTS, all resulting in improved performance with respect to MLP-based phone posteriors currently used in many systems.

1-4244-1484-9/08/$25.00 ©2008 IEEE

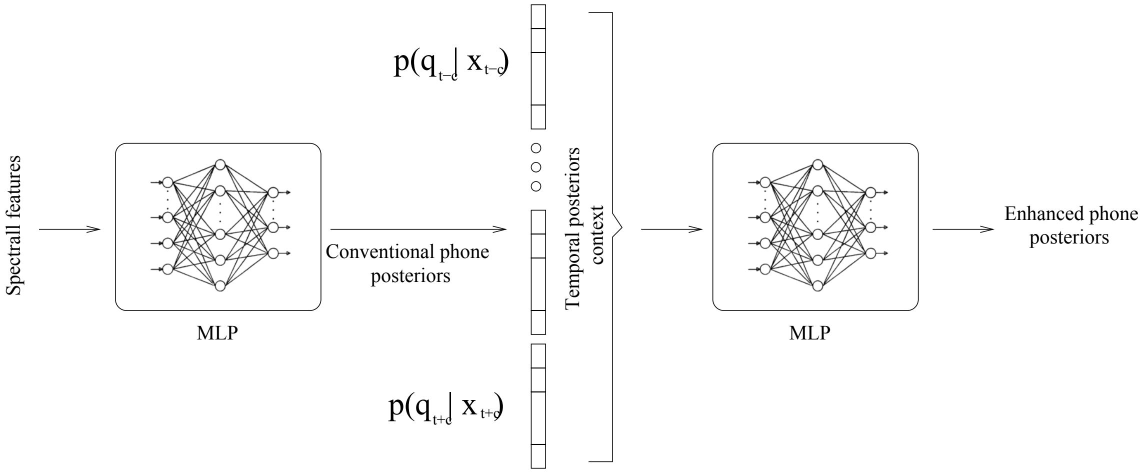

Spectrall features
MLP
Conventional phone posteriors
p(q_{t−c}| x_{t−c})
p(q_{t+c}| x_{t+c})
Temporal posteriors context
MLP
Enhanced phone posteriors

Fig. 1. Our approach for enhancing phone posterior estimation: The first MLP is transforming acoustic (cepstral) features to phone evidences in terms of posterior probabilities. The second MLP processes a temporal context of the phone evidences estimated by the first MLP, and learns the longer term dependencies between phone evidences. These dependencies can be interpreted as phonetic, sub-lexical and lexical knowledge.

2. PHONE POSTERIOR ESTIMATION

As already mentioned, the phone posteriors are usually estimated only based on information in a local or limited span of spectral feature frames. Among different approaches for estimating phone posteriors, 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, q_i is the event $q_t = i$, and x_t is the acoustic features at time t. 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] or as scores in hybrid HMM/ANN configuration [1]. The time limited spectral information is not the only source of knowledge available for a specific phoneme. Usually other sources of knowledge, such as phonetic and lexical knowledge, and long context can provide additional information about phonemes, and possibly help to enhance the estimation of phone posteriors.

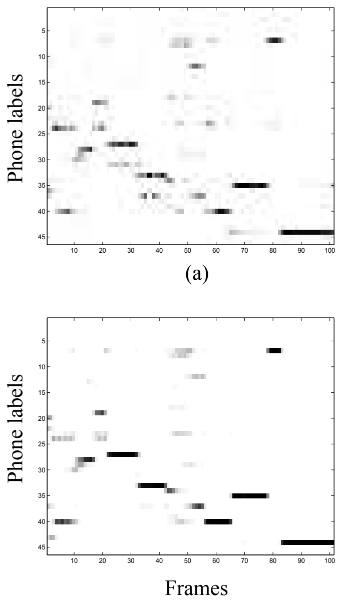
3. INTEGRATING PHONETIC AND LEXICAL KNOWLEDGE IN THE POSTERIOR ESTIMATION

Based on our discussion in the previous section, there are some extra sources of knowledge which are usually not taken into account in MLP-based phone posterior estimation. In this section, we propose a new configuration for integrating these extra sources of knowledge in the phone posterior estimation. The configuration is shown in Figure 1. We have two MLPs in this configuration. The first MLP performs the initial phone posterior probability estimation by transforming a small context of acoustic features (cepstral features) to phone posteriors. The second MLP takes a long temporal context of phone posteriors estimated by the first MLP as its input, and outputs phone posteriors for the same set of phonemes as the first MLP. The first MLP is typically trained with the cepstral features as input and phone targets as output, while the second MLP is trained with a long context of phone posteriors as input and the same phone targets as output. The first MLP learns the transformation form acoustic features to phone evidences, while the second MLP gets the phone evidences as the input and learns long term dependencies between phone evidences. This long term phone dependencies can be interpreted as sub-lexical or lexical knowledge, phoneme trajectory shape and phone duration information (phonetic knowledge). Therefore, the second MLP implicitly learns phonetic and lexical information, and is able to introduce these extra knowledge in the phone posterior estimation during the forward pass (inference). This leads to enhancement of phone posteriors. The phonetic and lexical knowledge is learned from data, and not provided by our prior assumptions on lexical rules or phone duration constraints. In the current work the two MLPs are trained using the same database. In future tough, we study further improvements by training the two MLPs on different databases. Another alternative will also be to use the second MLP for task adaptation purposes.

Figure 2 is showing an example of initial and corresponding enhanced posteriors. The enhanced (second MLP) posteriors are less noisy than the initial (first MLP) posteriors. The second MLP acts as a filter which smooth out evidences not matching the learned phonetic and lexical knowledge. These new posteriors can be used in the same way as the initial posteriors in speech recognition systems. They can be used as features in Tandem configuration, or as local scores in hybrid HMM/ANN configuration. Ideally, this approach can be used for post-processing the output of any posterior estimator to integrate higher level knowledge (phonetic, lexical knowledge).

4. EXPERIMENTS AND RESULTS

We have set up phone and word recognition experiments to evaluate and compare the performance of the enhanced phone posteriors with the initial posteriors. The enhanced posteriors are obtained by post processing initial posteriors using a secondary Neural Network (MLP). The comparison is done in terms of frame, phone and word error rates using hybrid HMM/ANN and Tandem configurations for the recognition.



(a)

(b)

Fig. 2. (a) Initial posteriors estimated by the first MLP, and (b) enhanced phone posteriors estimated by the second MLP integrating phonetic and lexical knowledge. The new enhanced posteriors are less noisy.

4.1. Phone recognition experiments

For the phone recognition experiments, TIMIT database [6] is used. The training data consists of 3000 utterances from 375 speakers, cross validation data set consists of 696 utterances from 87 speakers and the test data set consists of 1344 utterances from 168 speakers. There are 39 context independent phonemes. The acoustic features are PLP, delta and double delta features. For estimating initial posteriors we have used an MLP with 351 input nodes (9 frames of PLPs), 1000 hidden nodes and 39 (corresponding to the number of phonemes) output nodes.

In order to estimate enhanced posteriors, a 19 frames temporal context of the initial posteriors are post processed by a secondary MLP (as explained in Section 3). This MLP has 741 (39x19) input nodes, 1000 hidden nodes and 39 output nodes (corresponding to the number of phonemes). The size of the temporal posterior context, and the structure of the second MLP is obtained empirically for all the experiments. For the phone recognition, we have used NOWAY [9] which is a hybrid HMM/ANN decoder. In this decoder, each phoneme is modeled with 5 states, and a bi-gram phoneme level language model is used. Frame and phone recognition results are shown in Table 1. The second column is showing the performance of the initial posteriors in terms of frame and phone error rates. The third column is showing the performance of corresponding enhanced posteriors obtained by post processing the initial posteriors. The enhanced posteriors perform significantly better than the initial posteriors for frame and phone recognition.

Error rates	Initial posteriors	Enhanced posteriors
FER	29.9%	**27.4%**
PER	31.2%	**28.5%**

Table 1. Frame error rates (FER) and phone error rates (PER) for initial and enhanced phone posteriors, on TIMIT database.

4.2. Word recognition experiments

The Conversational Telephone Speech (CTS) [8] and OGI Numbers [7] databases are used for word recognition experiments. In Numbers database, there are 31 word classes and 27 context independent phonemes. The training set is about 1.5 hours and the test set is about 0.9 hour. The acoustic features are PLP, delta and double delta features. For estimating initial posteriors, we have used an MLP with 351 input (9 frames of acoustic features), 1800 hidden, and 27 output (corresponding to the number of phones) nodes.

In CTS database, there are 46 phone and 1000 word classes. The acoustic features are PLP, delta and double delta features. For initial posterior estimation, we have used an MLP with 351 input, 2000 hidden, and 46 output nodes (corresponding to the number of phones). The training set is about 15 hours and the test set is about 0.6 hour.

In order to enhance phone posterior estimates for the Numbers database, a second MLP post-processing 19 frames of initial posteriors is used (as explained in Section 3). It has 513 (19x27) input nodes, 1000 hidden nodes and 27 output nodes. For the CTS database, a second MLP with 598 (13x46) input nodes, 2000 hidden nodes and 46 output nodes is used to post-process 13 frames of the initial posteriors. Table 2 is showing frame error rates for the initial and enhanced posteriors for Numbers and CTS databases. Again, lower error rates can be observed for the enhanced posteriors in both databases. For word recognition, we have used Tandem and hybrid HMM/ANN configurations.

In Tandem approach, the posteriors are first gaussianized and decorrelated using log and Karhunen Loeve (KLT) transforms. The result of the transformation is used as features for training and inference in a HMM/GMM layer. The HMM/GMM layer is implemented in HTK [10], and trained with the posterior features. In case of Numbers database, context dependent models with 12 mixtures per state is used. In case of CTS database, context independent models with 32 mixtures per state, and a bi-gram language model is used. Table 3 is showing the word recognition performances for initial and enhanced posteriors. It can be observed that the enhanced posteriors are consistently performing better than the initial posteriors for the two databases.

In hybrid HMM/ANN approach, the posteriors are used as local scores for decoding. We have used NOWAY as the hybrid decoder for Numbers database, and JUICER [11] for CTS database. In case of Numbers database, phonemes are modeled with 5 states in the decoder. In case of CTS database, phonemes are modeled with 5 states, and a bi-gram language model is used. Table 4 is showing the word recognition performances for initial and enhanced posteriors. It can be again observed that the enhanced posteriors are performing better than the initial posteriors.

Overall, the experiments show that enhancing posterior estimates, consistently improves the performance in terms of frame, phone and word recognition.

In addition to the recognition studies, we study the entropy of the enhanced and regular posteriors. The entropy provides a measure of consistency/confusion in the posteriors. The entropy is measured for each frame (posteriors vector), and averaged over all the frames. Table 5 shows the average entropies for the enhanced and initial posteriors. Enhanced posteriors have smaller entropy than the initial posteriors. This indicates that there is more consistency in the enhanced posteriors, or in th other words they are less noisy.

Database	Initial posteriors	Enhanced posteriors
Numbers	19.5%	**16.7%**
CTS	28.7%	**25.9%**

Table 2. Frame error rates (FER) on Numbers'95 and CTS tasks, for initial and enhanced phone posteriors.

Database	Initial posteriors	Enhanced posteriors
Numbers	4.7%	**4.2%**
CTS	47.3%	**45.1%**

Table 3. Word error rates (WER) on Numbers'95 and CTS tasks, for initial and enhanced phone posteriors. The phone posteriors are used in Tandem configuration for the recognition. Enhanced phone posteriors perform significantly and consistently better than the initial posteriors.

Database	Initial posteriors	Enhanced posteriors
Numbers	9.2%	**8.2%**
CTS	53.6%	**49.2%**

Table 4. Word error rates (WER) on Numbers'95 and CTS tasks, for initial and enhanced phone posteriors. The phone posteriors are used in hybrid HMM/ANN configuration for the recognition. Enhanced phone posteriors perform better than the initial posteriors.

Database	Initial posteriors	Enhanced posteriors
TIMIT	1.16	**0.84**
CTS	1.64	**1.29**
Numbers	0.67	**0.40**

Table 5. Average entropy of enhanced and initial phone posteriors for different databases.

5. CONCLUSIONS

In this paper, we have proposed a simple yet very effective approach for enhancing the estimation of phone posteriors by hierarchically processing longer time spans of simpler information streams. In this approach, we use a secondary MLP to post-process the posteriors, and learn long term intra- and inter-phone posterior dependencies. The long term dependencies can be interpreted as phonetic, sub-lexical and lexical knowledge. This learned knowledge is integrated in the phone posterior estimation during the forward pass (inference), resulting in an enhanced version of the phone posterior estimates. The resulting enhanced posteriors are shown to consistently perform better than the initial posteriors for frame, phone and word recognition, over different small and large vocabulary tasks. Ideally, this approach can be used at the end of any posterior estimator to learn and integrate higher level of prior knowledge in the posterior estimation.

In this work, the second MLP has been trained on the same database. In the future, we study possible further improvements by training the two MLPs on different databases, or splitting the existing database between the two MLPs (with a percentage of overlap). Another alternative will also be to use the second MLP only for (task) adaptation purposes. For instance, the first MLP can be trained on a general English database, while the second MLP is trained on a second database of specific accent or dialect. In this case, the first MLP acts as a general phone posterior estimator, and the second MLP adapts the posterior estimation for a specific task.

We will also further study the strategies and possibilities of optimizing and using a simpler structure for the second MLP. This will provide the possibility of processing longer temporal context. The initial phone posteriors have simpler and possibly linearly separable patterns, as compared to the acoustic features. Therefore, it is potentially possible to use a relatively simpler MLP for post processing the phone posteriors. Additional MLPs in the hierarchy can even have simpler structure, since the posteriors become smoother.

6. ACKNOWLEDGMENTS

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), and Augmented Multi-party Interaction with Distance Access (AMIDA) projects. We also thank Hynek Hermansky (IDIAP) and Mathew Magimai Doss (IDIAP) 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] Fisher, W.M, Doddingtion, G.R, and Goudie-Marshall, K.M. "The DARPA Speech Recognition Research Database: specifications and Status," Proc. of DARPA Workshop on Speech Recognition, pp. 93-99, Feb. 1986.

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

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

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

[10] 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.

[11] Moore, D., Dines, J., Doss, M., Vepa, J., Cheng, O., Hain, T., "Juicer: A Weighted Finite-State Transducer speech decoder", MLMI'06, 2006.