Hierarchical Multi-Stream Posterior Based Speech Recognition System

Hamed Ketabdar[1,2], Hervé Bourlard[1,2] and Samy Bengio[1]

[1] IDIAP Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract. In this paper, we present initial results towards boosting posterior based speech recognition systems by estimating more informative posteriors using multiple streams of features and taking into account acoustic context (e.g., as available in the whole utterance), as well as possible prior information (such as topological constraints). These posteriors are estimated based on "state gamma posterior" definition (typically used in standard HMMs training) extended to the case of multi-stream HMMs.This approach provides a new, principled, theoretical framework for hierarchical estimation/use of posteriors, multi-stream feature combination, and integrating appropriate context and prior knowledge in posterior estimates. In the present work, we used the resulting gamma posteriors as features for a standard HMM/GMM layer. On the OGI Digits database and on a reduced vocabulary version (1000 words) of the DARPA Conversational Telephone Speech-to-text (CTS) task, this resulted in significant performance improvement, compared to the state-of-the-art Tandem systems.

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 features or as local scores (measures) 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. Hybrid HMM/ANN method allows for discriminant training, as well as for the possibility of using short 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 only local frame or limited number of local frames as context) are used.

In [6], a method was presented to estimate more informative posterior probabilities based on "state gamma posterior" definition (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 their approach, posterior probabilities are estimated based on state gamma posterior definition in a HMM configuration, which, after some transformations, 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. In [7], these posteriors are used as local scores for a Viterbi decoder. It also showed improvement over hybrid HMM/ANN approach which uses local posteriors as local scores.

Building upon the idea of multi-stream HMMs [8, 9], in this paper we present initial investigations towards extending the mentioned posterior estimation method to multi-stream case. We show that the posterior probabilities can be estimated through a multi-stream HMM configuration based on multi-stream state gamma definition, thus giving the estimate of posteriors by combining multiple streams of input features and also taking into account whole context in each stream as well as prior knowledge encoded in the model. Our hierarchical approach is as follows: The input feature streams are PLP cepstral [10] and TRAP temporal [11] features which are known to have some complementary information. We estimate the posteriors based on state gamma posterior definition through a multi-stream HMM configuration. These posteriors are used after some transformations as features for a standard HMM/GMM layer. This hierarchical approach provides a new, principled, theoretical framework for combining different streams of features taking into account context and model knowledge. We show that this method gives significant performance improvement over baseline PLP-TANDEM [5] and TRAP-TANDEM [11] techniques and also entropy based combination method [12] on OGI digits [13] and a reduced vocabulary version (1000 words) of CTS [6] databases.

In the present paper, Section 2 reviews single stream gamma posterior estimation method. The extension of this method to multi-stream case is explained in Section 3. Section 4 explains the configuration of our hierarchical multi-stream posterior based ASR system. Experiments and results are presented in Section 5. Conclusions and future work plans are discussed in Section 6.

2 Single stream "gamma posterior" estimation

In this section, we show how posterior probabilities can be estimated taking into account whole context in a stream and prior knowledge (e.g. topological constraints) encoded in the model. These posteriors are estimated based on "state gamma posterior" definition (as it is referred to in HMM formalism) through an HMM configuration.

In phone based speech recognition systems, phones are usually modeled by a few number of states. The posteriors are first estimated for each state (called "state gamma posteriors" as referred to in HMM formalism and used in HMM training), which then can be integrated to phone or higher level posteriors.

According to standard HMM formalism, the state gamma posterior $\gamma(i,t|M)$ 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):

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

where x_t is a feature vector at time t, $x_{1:T} = \{x_1, \ldots, x_T\}$ is an acoustic observation sequence of size T and q_t is HMM state at time t, which value can range from 1 to N_q (total number of possible HMM states). In the following, we will drop all the dependencies on M, always 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) [14] using local emission likelihoods $p(x_t|q_i = t)$ (e.g., modeled by GMMs):

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

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

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

As mentioned above, we recall that the α and β recursions are processed through a specific HMM, which is used to represent prior knowledge.

Similar recursions, also yielding "state gamma posteriors" and using the same assumptions as the case of likelihood based recursions, can be developed for local posterior based systems such as hybrid HMM/ANN systems using MLPs to estimate HMM emission probabilities [6]. In standard HMM/ANN systems, these local posteriors are usually turned into "scaled likelihoods" by dividing MLP outputs $p(q_t = i|x_t)$ 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" α_s and "scaled beta" β_s recursions to yield gamma posterior estimates [6]. These recursions are similar to the previous recursions except that the likelihood term is replaced by the scaled likelihood:

$$\begin{aligned}
\alpha_s(i,t) &\triangleq \frac{p(x_{1:t}, q_t = i)}{\prod_{\tau=1}^{t} p(x_\tau)} \\
&= \frac{p(x_t|q_t = i)}{p(x_t)} \sum_j p(q_t = i|q_{t-1} = j)\alpha_s(j, t-1) \\
&= \frac{p(q_t = i|x_t)}{p(q_t = i)} \sum_j p(q_t = i|q_{t-1} = j)\alpha_s(j, t-1) \qquad (5) \\
\beta_s(i,t) &\triangleq \frac{p(x_{t+1:T}|q_t = i)}{\prod_{\tau=t+1}^{T} p(x_\tau)} \\
&= \sum_j \frac{p(x_{t+1} = j|q_{t+1} = j)}{p(x_{t+1} = j)} p(q_{t+1} = j|q_t = i)\beta_s(j, t+1) \\
&= \sum_j \frac{p(q_{t+1} = j|x_{t+1})}{p(q_{t+1} = j)} p(q_{t+1} = j|q_t = i)\beta_s(j, t+1) \qquad (6) \\
\gamma(i,t) &\triangleq p(q_t = i|x_{1:T}) = \frac{\alpha_s(i,t)\beta_s(i,t)}{\sum_j \alpha_s(j,t)\beta_s(j,t)} \qquad (7)
\end{aligned}$$

subscript s indicates that the recursion is based on scaled likelihood. In the above equations, exactly the same independence assumptions as standard HMMs are used, beside the fact that the local correlation may be better captured if the ANN is presented with acoustic context.

3 Multi-stream "gamma posterior" estimation

In multi-stream HMM configuration, the definition of the state gamma posterior is extended to the probability of being in specific state i at specific time t, given the whole observation sequences for *all streams*, and model M encoding specific prior knowledge:

$$\gamma(i,t) \triangleq p(q_t = i|x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}, M) \qquad (8)$$

where superscript n indicates the stream number. We call the state gamma posterior estimated using multiple streams of features as "multi-stream state gamma". As we show in this section, multi-stream state gammas can be estimated using multi-stream forward α and backward β recursions. The multi-stream α and β recursions can also be written based on individual stream α^n and β^n recursions. In this work, we focus on the posterior based systems, therefore all the recursions are written using scaled likelihoods. The same multi-stream recursions but for likelihood based systems has been explained in [15].

We start with individual stream forward α^n and backward β^n recursions:

$$\begin{aligned}
\alpha_s^n(i,t) &\triangleq \frac{p(x_{1:t}^n, q_t = i)}{\prod_{\tau=1}^t p(x_\tau^n)} \\
&= \frac{p(q_t = i|x_t^n)}{p(q_t = i)} \sum_j p(q_t = i|q_{t-1} = j)\alpha_s^n(j, t-1) \qquad (9) \\
\beta_s^n(i,t) &\triangleq \frac{p(x_{t+1:T}^n|q_t = i)}{\prod_{\tau=t+1}^T p(x_\tau^n)} \\
&= \sum_j \frac{p(q_{t+1} = j|x_{t+1}^n)}{p(q_{t+1} = j)} p(q_{t+1} = j|q_t = i)\beta_s^n(j, t+1) \qquad (10)
\end{aligned}$$

where $\alpha_s^n(i,t)$ and $\beta_s^n(i,t)$ show the forward and backward recursions for stream n. Subscript s indicates that the recursion is written using scaled likelihoods.

We note here that we need to estimate $p(q_t = i)$, which can be done recursively as follows:

$$p(q_t = i) = \sum_j p(q_t = i|q_{t-1} = j)p(q_{t-1} = j) \qquad (11)$$

Using individual stream forward recursions α_s^n and applying the usual HMM assumptions, we can write multi-stream forward α_s recursion as follows:

$$\begin{aligned}
\alpha_s(i,t) &\triangleq \frac{p(x_{1:t}^1, x_{1:t}^2, ..., x_{1:t}^N, q_t = i)}{\prod_{\tau=1}^t p(x_\tau^1) \prod_{\tau=1}^t p(x_\tau^2) ... \prod_{\tau=1}^t p(x_\tau^N)} \qquad (12) \\
&= \frac{p(x_{1:t}^1, x_{1:t}^2, ..., x_{1:t}^N|q_t = i)p(q_t = i)}{\prod_{\tau=1}^t p(x_\tau^1) \prod_{\tau=1}^t p(x_\tau^2) ... \prod_{\tau=1}^t p(x_\tau^N)} \qquad (13) \\
&= \frac{p(x_{1:t}^1|q_t = i)p(x_{1:t}^2|q_t = i)...p(x_{1:t}^N|q_t = i)p(q_t = i)}{\prod_{\tau=1}^t p(x_\tau^1) \prod_{\tau=1}^t p(x_\tau^2) ... \prod_{\tau=1}^t p(x_\tau^N)} \qquad (14) \\
&= \frac{\frac{p(x_{1:t}^1, q_t=i)}{\prod_{\tau=1}^t p(x_\tau^1)}}{p(q_t = i)} \frac{\frac{p(x_{1:t}^2, q_t=i)}{\prod_{\tau=1}^t p(x_\tau^2)}}{p(q_t = i)} ... \frac{\frac{p(x_{1:t}^N, q_t=i)}{\prod_{\tau=1}^t p(x_\tau^N)}}{p(q_t = i)} p(q_t = i) \qquad (15) \\
&= \frac{\alpha_s^1(i,t)}{p(q_t = i)} \frac{\alpha_s^2(i,t)}{p(q_t = i)} ... \frac{\alpha_s^N(i,t)}{p(q_t = i)} p(q_t = i) \qquad (16) \\
&= \frac{\prod_{n=1}^N \alpha_s^n(i,t)}{p(q_t = i)^{N-1}} \qquad (17)
\end{aligned}$$

when going form (13) to (14), we add the following reasonable assumption:

$$p(x_{1:t}^1, x_{1:t}^2, ..., x_{1:t}^N|q_t = i) = p(x_{1:t}^1|q_t = i)p(x_{1:t}^2|q_t = i)...p(x_{1:t}^N|q_t = i) \qquad (18)$$

while (14) is rewritten as (15) simply by applying Bayes rule.

The multi-stream β_s recursion can also be written using individual stream β_s^n recursions:

$$\beta_s(i,t) \triangleq \frac{p(x^1_{t+1:T}, x^2_{t+1:T}, ..., x^N_{t+1:T}|q_t=i)}{\prod_{\tau=t+1}^T p(x^1_\tau) \prod_{\tau=t+1}^T p(x^2_\tau) ... \prod_{\tau=t+1}^T p(x^N_\tau)} \tag{19}$$

$$= \frac{p(x^1_{t+1:T}|q_t=i)p(x^2_{t+1:T}|q_t=i)...p(x^N_{t+1:T}|q_t=i)}{\prod_{\tau=t+1}^T p(x^1_\tau) \prod_{\tau=t+1}^T p(x^2_\tau) ... \prod_{\tau=t+1}^T p(x^N_\tau)} \tag{20}$$

$$= \frac{p(x^1_{t+1:T}|q_t=i)}{\prod_{\tau=t+1}^T p(x^1_\tau)} \frac{p(x^2_{t+1:T}|q_t=i)}{\prod_{\tau=t+1}^T p(x^2_\tau)} ... \frac{p(x^N_{t+1:T}|q_t=i)}{\prod_{\tau=t+1}^T p(x^N_\tau)} \tag{21}$$

$$= \beta_s^1(i,t)\beta_s^2(i,t)...\beta_s^N(i,t) \tag{22}$$

$$= \prod_{n=1}^N \beta_s^n(i,t) \tag{23}$$

Note that (19) is rewritten as (20) assuming

$$p(x^1_{t+1:T}, x^2_{t+1:T}, ..., x^N_{t+1:T}| \ \ q_t=i) = \tag{24}$$
$$p(x^1_{t+1:T}|q_t=i)p(x^2_{t+1:T}|q_t=i)...p(x^N_{t+1:T}|q_t=i)$$

The multi-stream state gamma $\gamma(i,t)$ can then be obtained using multi-stream α_s and β_s recursions:

$$\gamma(i,t) \triangleq p(q_t=i|x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T})$$

$$= \frac{p(x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}, q_t=i)}{p(x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T})}$$

$$= \frac{p(x^1_{1:t}, x^1_{t+1:T}, x^2_{1:t}, x^2_{t+1:T}, ..., x^N_{1:t}, x^N_{t+1:T}, q_t=i)}{\sum_j p(x^1_{1:t}, x^1_{t+1:T}, x^2_{1:t}, x^2_{t+1:T}, ..., x^N_{1:t}, x^N_{t+1:T}, q_t=j)}$$

$$= \frac{p(x^1_{1:t}, x^2_{1:t}, ..., x^N_{1:t}, q_t=i)p(x^1_{t+1:T}, x^2_{t+1:T}, ..., x^N_{t+1:T}|q_t=i)}{\sum_j p(x^1_{1:t}, x^2_{1:t}, ..., x^N_{1:t}, q_t=j)p(x^1_{t+1:T}, x^2_{t+1:T}, ..., x^N_{t+1:T}|q_t=j)}$$

$$= \frac{\frac{p(x^1_{1:t}, x^2_{1:t}, ..., x^N_{1:t}, q_t=i)p(x^1_{t+1:T}, x^2_{t+1:T}, ..., x^N_{t+1:T}|q_t=i)}{\prod_{n=1}^N \prod_{\tau=1}^t p(x^n_\tau) \prod_{n=1}^N \prod_{\tau=t+1}^T p(x^n_\tau)}}{\frac{\sum_j p(x^1_{1:t}, x^2_{1:t}, ..., x^N_{1:t}, q_t=j)p(x^1_{t+1:T}, x^2_{t+1:T}, ..., x^N_{t+1:T}|q_t=j)}{\prod_{n=1}^N \prod_{\tau=1}^t p(x^n_\tau) \prod_{n=1}^N \prod_{\tau=t+1}^T p(x^n_\tau)}}$$

$$= \frac{\alpha_s(i,t)\beta_s(i,t)}{\sum_j \alpha_s(j,t)\beta_s(j,t)} \tag{25}$$

We remind that all multi-stream recursions are processed through a (Markov) model M encoding some prior knowledge (e.g. topological constraints).

3.1 Ergodic HMM with uniform transition probabilities

As already mentioned above, all single stream, as well as multi-stream α and β recursions are applied through a given HMM topology representing some prior

knowledge. When no specific prior knowledge is available, the simplest solution consists in using ergodic HMM with uniform transition probabilities, i.e. $p(q_t = i|q_{t-1} = j) = K$. In this case, the multi-stream gamma estimation equation (25) can be rewritten as follows:

$$\begin{aligned}
\gamma(i,t) &= p(q_t = i|x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}) \\
&= \frac{\alpha_s(i,t)\beta_s(i,t)}{\sum_j \alpha_s(j,t)\beta_s(j,t)} \\
&= \frac{\frac{\prod_{n=1}^N \alpha_s^n(i,t)}{p(q_t=i)^{N-1}}\beta_s(i,t)}{\sum_j \frac{\prod_{n=1}^N \alpha_s^n(j,t)}{p(q_t=j)^{N-1}}\beta_s(j,t)} \\
&= \frac{\frac{\prod_{n=1}^N \frac{p(q_t=i|x_t^n)}{p(q_t=i)} \sum_k p(q_t=i|q_{t-1}=k)\alpha_s^n(k,t-1)}{p(q_t=i)^{N-1}}\beta_s(i,t)}{\sum_j \frac{\prod_{n=1}^N \frac{p(q_t=j|x_t^n)}{p(q_t=j)} \sum_k p(q_t=j|q_{t-1}=k)\alpha_s^n(k,t-1)}{p(q_t=j)^{N-1}}\beta_s(j,t)}
\end{aligned} \tag{26}$$

Assuming ergodic uniform transition probabilities, the sum over k factors in above numerator and denominator and also β_s factors are identical and can thus be droped. Moreover, the state prior $p(q_t = i)$ is constant, thus yielding:

$$\gamma(i,t) = \frac{\prod_{n=1}^N p(q_t = i|x_t^n)}{\sum_j \prod_{n=1}^N p(q_t = j|x_t^n)} \tag{27}$$

Therefore, the multi-stream state gamma is the normalized product of posteriors (MLP outputs) and gammas do not capture context and specific prior knowledge. In this case, the multi-stream gamma estimation method can be interpreted as a principled way to combine two streams of features.

3.2 Higher level posterior estimation

In case of having phone-based ASR system and modeling each phone with more than one state, state gamma posteriors should be integrated to phone level 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:

$$\begin{aligned}
\gamma_p(i,t) \triangleq p(p_t = i|x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}) &= \sum_{j=1}^{N_q} p(p_t = i, q_t = j|x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}) \\
&= \sum_{j=1}^{N_q} p(p_t = i|q_t = j, x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}) p(q_t = j|x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}) \\
&= \sum_{j=1}^{N_q} p(p_t = i|q_t = j, x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T}) \gamma(j,t)
\end{aligned} \tag{28}$$

where p_t is a phone at time t. Probability $p(p_t = i|q_t = j, x^1_{1:T}, x^2_{1:T}, ..., x^N_{1:T})$ 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 model each phone with one state in the multi-stream HMM, therefore in this particular case, state gammas are equal to phone gammas and we do not need to integrate state gammas to phone gammas.

4 Hierarchical multi-stream posterior-based ASR

In this section, the configuration of our hierarchical multi-stream speech recognition system is explained. The main idea is to combine N (two in our case) streams of features which have complementary information by estimating gamma posteriors through a multi-stream HMM configuration. These posteriors capture the whole context in all streams as well as prior knowledge (e.g. topological constraints) encoded in the model, thus they are expected to be more informative than individual streams of features before the combination. Figure 1 shows our hierarchical multi-stream posterior based ASR system. This hierarchical system consists of three layers: The first layer gets two streams of raw features (PLPs and TRAPs) extracted from speech signal, and estimates two streams of posteriors using MLPs. This is called "single stream posterior estimation". These streams of posteriors are used after turning to scaled likelihoods in the second layer of hierarchy, which is a multi-stream posterior based HMM to obtain the estimates of multi stream state gammas. The state gammas are then used as features after some transformations (KLT) for the third layer of hierarchy which is a standard HMM/GMM train/inference back-end. In the following, some issues related to the system is explained in more details:

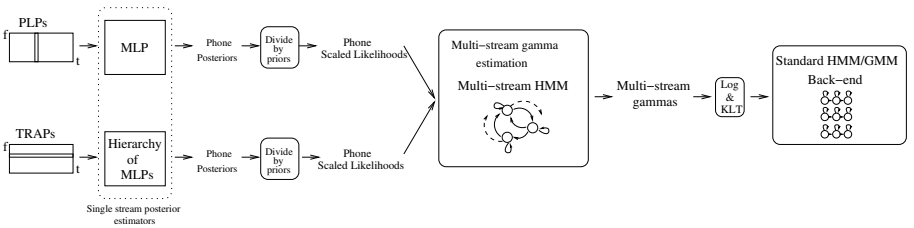

Fig. 1. Hierarchical multi-stream posterior based ASR: Two streams of posteriors are estimated form PLP and TRAP features using MLPs, then these posteriors are turned into scaled likelihoods by dividing by the priors. The resulting two streams of scaled likelihoods are fed to the multi-stream HMM. The multi-stream gammas are estimated using multi-stream forward α_s and backward β_s recursions as explained in Section 3. These multi-stream gamma posteriors are used after some transformations (KLT) as features for a standard HMM/GMM back-end system.

4.1 Input streams of features

The first step in developing the system is to choose two sets of features having complementary information. Spectral (cepstral) features and features having long temporal information are suitable candidates. We used PLP cepstral features[3] [10] and TRAPs temporal features [11] as input feature streams for the system. TRAP features represent temporal energy pattern for different bands over a long context, while PLPs represent full short-term spectrum (possibility with very short time context).

4.2 Single stream posterior estimation

In the first layer of hierarchy, the two input feature streams (PLPs and TRAPs) are processed by MLPs to estimate posterior probabilities of context-independent phones. For PLP cepstral features, usually 9 frames of PLP coefficients and their first and second order derivatives are concatenated as the input for a trained MLP to estimate the posterior probabilities of context-independent phones [5]. The phonetic class is defined with respect to the center of 9 frames. For the case of TRAPs, different bands temporal energy pattern over 0.5 to 1 second TRAP temporal vector are first processed by band classifier MLPs, then the outputs of these band classifiers are fed as inputs for a merger MLP [11]. The Merger MLP outputs gives the posterior estimate for context-independent phones. Again, phonetic class is defined with respect to the center of 0.5-1 second temporal vector. In the reminder of the paper, we call these single stream posterior estimates as PLP and TRAP posteriors.

4.3 Multi-stream posterior estimation

Having two stream of posteriors estimated from PLP and TRAP features using MLPs, the next step in the hierarchy is to estimate state gammas through the multi-stream HMM configuration. Posteriors are first divided by priors to obtain scaled likelihoods, i.e.: $\frac{p(x_t^n|q_t=i)}{p(x_t^n)} = \frac{p(q_t=i|x_t^n)}{p(q_t=i)}$, and then these scaled likelihoods are used in multi-stream forward α_s and backward β_s recursions according to (17, 23) to obtain estimates of state gammas. In this work, we model each phone with one state, thus state gammas are equal to phone gammas. Moreover, we assume ergodic uniform transition probabilities between phones, therefore as explained in Section 3.1, the multi-stream state gamma estimation can be interpreted as a probabilistic principled way to combine different streams of features which have complementary information.

5 Experiments and results

Results are presented on OGI digits [13] and a reduced vocabulary version of the DARPA Conversational Telephone Speech-to-text (CTS) task (1'000 words) databases [6]. We used PLP and TRAP features as input streams to our system.

[3] In the reminder of the paper, "PLP cepstral features" stands for PLP cepstral coefficients and their first and second order derivatives

The PLP cepstral coefficients [10] are extracted using 25-ms window with 10-ms shifts. At each frame, 13 PLP coefficients, their first-order and second-order derivatives are extracted and concatenated to make one feature vector.

For extracting TRAP features, the short-term critical band spectrum is computed in 25-ms windows with 10-ms shifts and the logarithm of the estimated critical band spectral densities are taken. There are 15 bands. For each band, 50 frames before and after the center of analysis is taken resulting in 101 points long temporal TRAP vector [11].

In this work, each phone is modeled by one state in the multi-stream HMM and we assume ergodic uniform transition probabilities between phones.

5.1 OGI Digits

The task is recognition of eleven words (American English Digits). The test set was derived from the subset of CSLU Speech Corpus [13], containing utterances of connected digits. There are 2169 utterances (12437 words, about 1.7 hours) in the test set. Training set contains 2547 utterances (about 1.2 hours). This set is also derived from CSLU Speech Corpus and utterances containing only connected digits are used. Standard HMM/GMM train/inference back-end system is based on HTK. There are 29 context-independent phonetic classes. The subset of OGI stories [16] plus a subset of OGI numbers [13] was used for training MLPs for single stream posterior estimation. This set has in total 3798 utterances with total length about 4.5 hours.

Two streams of posteriors (one from PLP features and the other one from TRAP features) are estimated as explained in Section 4.2 for the test and training set. They are then turned into scaled likelihoods and used in the multi-stream HMM layer to get the estimates of state (phone) gammas. These gamma posteriors are fed as features (after gaussianization and decorrelation through log and KL transform) to the standard HMM/GMM layer. For comparison purposes, we also run the standard HMM/GMM system using single stream posterior estimates as features (after log and KLT) in order to obtain the baseline performance of single stream PLP and TRAP posteriors before the combination (This corresponds to PLP-TANDEM and TRAP-TANDEM methods). Moreover, we used an inverse entropy based combination method [12] to combine PLP and TRAP posteriors, and compare the combination performance with our method. Table 1 shows the result of recognition studies.The first column shows the features (after log and KLT) which are fed to standard HMM/GMM layer. The second column shows word error rate (WER). The first row shows the baseline performance of posteriors estimated using PLP features (the first stream). The second row shows the baseline performance of posteriors estimated using TRAP features (the second stream). The third row shows the performance of features obtained by inverse entropy combination of PLP and TRAP posteriors and the fourth row shows the performance of our system which uses multi-stream gamma posteriors obtained by combining the mentioned streams of PLP and TRAP posteriors through the multi-stream HMM. The system using multi-stream gamma posteriors performs significantly better than the systems using baseline single stream posteriors before the combination and also inverse entropy based combination.

Features	WER
PLP posteriors	3.6%
TRAP posteriors	4.8%
Inverse entropy combination	3.5%
Multi-stream gammas	**2.9%**

Table 1. Word error rates (WER) on OGI Digits task

5.2 DARPA CTS task

The use of multi-stream gamma estimation method was further evaluated on a conversational telephone speech (CTS) recognition task. The training set for this task contained 15011 utterances (about 15.9 hours) and the test set contained 951 utterances (about 0.6 hour) of male speakers CTS speech randomly selected from the Fisher Corpus and the Switchboard Corpus. There were 46 context-independent phonetic classes in this task. The layer estimating single stream posteriors were trained on the same training set using PLP and TRAP features. The standard HMM/GMM system is based on HTK. A 1000 word dictionary with multi-words and multi-pronunciations was used for decoding, using a bi-gram language model.

Similar experiments as the case of OGI Digits database was repeated. Table 2 shows the recognition results. Again, multi-stream gamma combination gives significant improvement over PLP and TRAP posteriors before the combination and also inverse entropy combination.

Features	WER
PLP posteriors	48.7%
TRAP posteriors	55.1%
Inverse entropy combination	48.7%
Multi-stream gammas	**46.8%**

Table 2. Word error rates (WER) on CTS task

6 Conclusions and future work

In this paper, we proposed a new, principled, theoretical framework for hierarchical estimation/use of posteriors and multi-stream feature combination, and we presented initial results for this theory. We explained how the posterior estimation can be enhanced by combining different streams of features and 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 features for a standard HMM/GMM system. We showed our system performs significantly better as compared to the PLP-TANDEM and TRAP-TANDEM baseline systems and inverse entropy combination method on two different ASR tasks. This theoretical framework allows designing optimal hierarchical multi-stream systems since it proposes a principled way for combining different streams of features by hierarchical posterior estimation and introducing context and prior knowledge to get better evidences in the form of posteriors.

In this work, we investigated the particular case of assuming ergodic uniform transition probabilities. We will further investigate this method by introducing

prior knowledge encoded in appropriate model to get better estimates of posteriors. The state gammas can be also used for reestimating MLP parameters in the single stream posterior estimation layer. In this case, the MLPs used for estimating single stream phone posteriors from acoustic features are retrained with multi-stream phone gamma posteriors as new labels.

7 Acknowledgments

This project was jointly funded by the European AMI and PASCAL projects and the IM2 Swiss National Center of Competence in Research. The authors want to thank Petr Fousek for providing PLP and TRAP features and 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*, 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, also IDIAP-RR 04-58.
7. Ketabdar, H., Vepa, J., Bengio, S., and Bourlard, H., "Developing and enhancing posterior based speech recognition systems", *IDIAP RR 05-23*, 2005.
8. Bourlard, H. and Dupont, S., "Sub-band-based speech recognition", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 1251-1254, 1997.
9. Dupont, S. and Luettin, J., "Audio-visual speech modeling for continuous speech recognition", *IEEE Transactions on Multimedia*, vol. 2. no. 3, pp. 141-151, 2000.
10. Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech". *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990.
11. Hermansky, H., Sharma, S., "TRAPs: classifiers of TempoRAl Patterns", *Proc. ICSLP-98*, Australia, November 98.
12. Misra, H., Bourlard, H. and Tyagi V., "New entropy based combination rules in HMM/ANN multi-stream ASR", *Proc. ICASSP*, 2003.
13. Cole, R., Fanty, M., Noel, M. and Lander T. "Telephone Speech Corpus Development at CSLU", In Proc. of ISCLP (Yokohama, Japan, 1994), pp. 1815-1818.
14. Rabiner, L. R., "A tutorial on hidden Markov models and selective applications in speech recognition", *Proc. IEEE*, vol. 77, pp. 257-286, 1989.
15. Bengio, S., "Joint training of multi-stream HMMs", *to be published as IDIAP-RR 05-22*, 2005.
16. Cole, R., Noel, M., Lander T. and Durham T. "New Telephone Speech Corpora at CSLU", In Proc. of EUROSPEECH (Madrid, Spain, 1995), pp. 821-824.