TOWARDS USING HIERARCHICAL POSTERIORS FOR FLEXIBLE AUTOMATIC SPEECH RECOGNITION SYSTEMS

Hervé Bourlard [a] Samy Bengio [a]
Mathew Magimai Doss [a] Qifeng Zhu [b]
Bertrand Mesot [a] Nelson Morgan [b]

IDIAP–RR 04-58

NOVEMBER 2004

SUBMITTED FOR PUBLICATION

[a] IDIAP Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland
[b] International Computer Science Institute, Berkeley, CA 94704, USA

IDIAP Research Report 04-58

TOWARDS USING HIERARCHICAL POSTERIORS FOR FLEXIBLE AUTOMATIC SPEECH RECOGNITION SYSTEMS

Hervé Bourlard   Samy Bengio   Mathew Magimai Doss   Qifeng Zhu
Bertrand Mesot   Nelson Morgan

November 2004

submitted for publication

Abstract. Local state (or phone) posterior probabilities are often investigated as local classifiers (e.g., hybrid HMM/ANN systems) or as transformed acoustic features (e.g., "Tandem") towards improved speech recognition systems. In this paper, we present initial results towards boosting these approaches by improving the local state, phone, or word posterior estimates, using all possible acoustic information (as available in the whole utterance), as well as possible prior information (such as topological constraints). Furthermore, this approach results in a family of new HMM based systems, where only (local and global) posterior probabilities are used, while also providing a new, principled, approach towards a hierarchical use/integration of these posteriors, from the frame level up to the sentence level. Initial results on several speech (as well as other multimodal) tasks resulted in significant improvements. In this paper, we present recognition results on Numbers'95 and on a reduced vocabulary version (1000 words) of the DARPA Conversational Telephone Speech-to-text (CTS) task.

1 INTRODUCTION

Over the last 10-15 years, posterior probabilities have been increasingly explored as a possible way to improve automatic speech recognition (ASR) systems, initially with the goal of providing more discriminant training and local HMM probabilities, and more recently as compact features (possibly resulting of the merging of several features).

Both approaches are certainly valid and have shown some success, e.g., in the case of hybrid HMM/ANN system (where posteriors are used as local classifiers) or in the case of "Tandem" systems (where posteriors are used as features fed into standard HMMs). However, their efficacy strongly depends on the quality of these posterior estimates, usually based on statistical tools such as multi-layer perceptrons (MLP) or normalized Gaussian mixture models (GMM), possibly exploiting some contextual acoustic input.

In this paper, we present the results of some initial investigation towards new ways to improve the estimation of local posteriors (hence the resulting performance) by using the so called "gamma" recursion (as usually referred to in the HMM formalism) to generate local posteriors taking into account all the acoustic information available in each utterance, possibly complemented by additional prior information. Interestingly, using these "state/phone gammas" not only yields improved recognition performance, as shown here on Numbers'95 and CTS, but also opens up several innovative and principled approaches towards the hierarchical use of posterior probabilities from the frame level up to the sentence level.

Finally, we believe that what is presented in the present paper provides a general framework for a theory of using posteriors as local measures (classifiers) or features in hierarchical, with the additional advantage of being able to accommodate, and possibly take advantage of, larger acoustic context, as well as specifically designed prior knowledge such as topological constraints.

The notation used in this paper will be the following:

- $X = x_1^T = \{x_1, \ldots, x_T\}$ an acoustic observation sequence
- q_t be an HMM state at time t, which value can range from 1 to N_q (total number of possible HMM states)
- p_t be a phoneme at time t, which value ranges from 1 to N_p (total number of phones)
- w_t be a word at time t, which value ranges from 1 to N_w (total number of words)
- Events "$q_t = i$", "$p_t = i$" and "$w_t = i$" will, respectively, often be written as q_t^i, p_t^i, and w_t^i.

2 POSTERIORS AS LOCAL CLASSIFIERS OR AS FEATURES

2.1 Posteriors as local classifiers

Hybrid HMM/ANN approaches were probably among the first ones to make extensive use of a posteriori probabilities in speech recognition. In these approaches, Artificial Neural Networks (ANN), and more specifically MLPs are used to compute the emission probabilities required in HMM systems [6]. It has indeed been shown that if each output unit of an MLP is associated with a particular state k of the set of possible HMM states $Q = \{1, 2 \ldots, k, \ldots, N_q\}$, it is possible to train the MLP to generate a posteriori probabilities of the output classes conditioned on the input, i.e., $p(q_t^i|x_t)$. While allowing for discriminant training, such an approach also has the advantage of possibly accommodating acoustic context by providing several frames at the MLP input, thus estimating $p(q_t^i|x_{t-c}^{t+c})$, where c is typically equal to 4. However, context of up to $c = 50$ has also been successfully used [15].

More recently, a posteriori probabilities started to be used as local measures for different ASR purposes, such as (1) estimating confidence measures [5, 12, 27], (2) beam search pruning [1], or (3) word lattice rescoring [20].

In all of these cases, posterior probability estimates appear to have been quite useful.

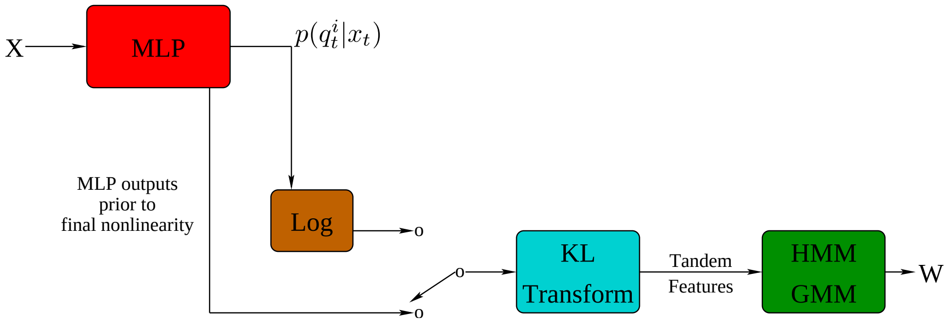

Figure 1: Standard approach for deriving Tandem features.

2.2 Posteriors as features

2.2.1 General idea

More recently, the properties described above were also extended by using the MLP-generated posterior probabilities as acoustic features, which (after some transformation) can be used alone or appended to other sets of (more traditional) features as inputs to HMMs. In this case, the MLP is considered as performing some kind of "optimal" feature extraction (using acoustic context and nonlinear discriminant analysis). In the case of multiple features (e.g., multi-band and multi-stream speech recognition), this MLP can also be used as a convenient way to integrate multiple features and generate the most compact and most discriminant representation to be used in standard HMMs.

The purpose of feature extraction in ASR is often to reduce dimensionality of data while preserving (or enhancing) the discriminant information of the data. In the process the irrelevant variability should be alleviated and the relevant variability preserved. To ensure Bayes error of the classification, at least $L-1$ features are required for discrimination among L classes (see [11], p. 444). Techniques that can satisfy this requirement based on optimal rotation of feature space such as linear discriminant analysis (LDA) has been used in feature extraction in ASR for quite some time, e.g., [17].

2.2.2 Tandem technique for deriving features for HMM-based ASR

Nonlinear alternatives for such data-guided feature extraction start to emerge. One of the earlier and most successful approach is Tandem [16]. For every speech instant (i.e. about every 10 ms in a typical ASR system), the Tandem technique derives a vector of posterior probabilities of subword speech events from any relevant evidence presented to its input. Posteriors of classes form a particularly convenient smallest set of features since the highest posterior determines the class assignment. Typically, a properly trained MLP, trained in one-hot encoding paradigm [6], is used for estimating posterior probabilities of context-independent phonemes. Alternatives such as GMM-derived posteriors were also investigated [24]. Hierarchical classification schemes in Tandem estimator were also investigated [26].

As illustrated in Figure 1, the MLP posterior probability estimates are gaussianized by a static nonlinearity and whitened by the Karhunen-Loeve transform (KLT) derived from training data. Three different techniques were investigated for the gaussianization: 1) computing logarithm of the MLP output, 2) taking MLP outputs prior to its final nonlinearities, and 3) deriving the non-linear function that properly gaussianized the histogram of posteriors from the training data. Observed differences in performance were typically minor. In the case of the experiments reported here, the logarithm was used to roughly gaussianize the features.

Such gaussianized and whitened posterior probabilities form the feature vector for the subsequent HMM recognizer. Thus, the conventional features derived from a spectral density vector representing the spectral envelope are replaced by the transformed posteriors of acoustic events (in the original concept the events were context-independent phonemes). If the targeted events are independent, the output of the trained Tandem MLP could represent an estimate of the efficient low-entropy statistically-independent code, hypothesized in perceptual processing [3, 19].

Input to Tandem can be any data that are believed to provide a relevant evidence for the classification. In its simplest form, Tandem takes as an input a superframe of speech typical conventional speech features such as 9 frames of concatenated PLP static and dynamic features. Often, Tandem inputs are concatenated outputs from other sub-band classifiers (TRAP [15] or HATS [8]). TRAP has been also reported to be efficient in combining of different features and for alleviating irrelevant information [30] [18].

In several aspects, Tandem represents a significant conceptual departure from the current practice in feature extraction for ASR:

1. The knowledge used for feature extraction is not only coming from beliefs and convictions of the designer but is mostly derived from development data. While LDA could be viewed as doing a similar function, Tandem provides a more general transformation.

2. In conjunction with sub-band classifiers (TRAP or HATS), the frequency-localized evidence is in the early stages of the feature extraction converted to frequency-localized estimates of likelihoods of speech events. In that way, many vulnerabilities of the short-term spectral envelope of speech are alleviated. Evidence used for deriving the features does not all come from the relatively short segment of speech representing a short part of the underlying sub-word class (phoneme) but the employed time span covers at least the typical coarticulation span of the phoneme. In that way, each feature vector could carry most of the available information about the underlying phoneme. While HATS/TRAP are not used in the experiments described here, they are an integral part of the larger EARS system developed by our team.

The Tandem-based ASR has been so far found most useful in combination with the conventional spectrum-based (PLP, Mel Cepstrum, etc) ASR. Thus, e.g. the system with Tandem module was shown to perform the best among all presented feature extraction techniques (including the officially accepted ETSI standard) on the small vocabulary Aurora task [2]. More recently, the Tandem-derived features were successfully used in DARPA EARS program, where they brought more than 10% relative improvement in error rate on a smaller (1000 word task) and scaled successfully on a full vocabulary task [22, 8, 30].

2.3 Layered Approaches Based on Posteriors

Recently, several researchers have proposed the use of a layered approach to simplify the encoding of relations between observations and high-level events (such as sentences) [23, 29]. In this approach, an HMM is built for each layer of semantic complexity; for instance, in speech recognition, one could think of a layer encoding phonemes, followed by a layer encoding words and a final layer encoding sentences. One advantage of such an approach is that by decomposing the problem hierarchically, learning is performed on lower-dimensional observation spaces and, prior information related to each level of semantic can more easily be included. While in the literature [23, 29] the observations of each layer was encoded using some form of event likelihoods, we believe that they would probably be best encoded using local posteriors. Section 5 explains how these posteriors could be computed from layer to layer and how these layers could incorporate different prior information.

3 IMPROVING STATE POSTERIOR ESTIMATION

In this section, we show how the estimation of local posterior probabilities can be improved by using all the acoustic information available, as well as possible prior knowledge. We recall and discuss here possible ways to estimate and use "state gammas" (state posteriors taking into account all available acoustic, as well as possible prior information and/or topological constraints M) to improve state-of-the-art speech recognition systems. As further discussed later, the resulting approach is illustrated in Figure 2.

In the following, we show that these state gammas can be estimated in two different ways, depending on whether the local estimators being used are likelihoods or posteriors (MLP).

3.1 Standard likelihood based systems

In standard (likelihood based) HMMs, using local emission likelihoods $p(x_t|q_t^i)$ (usually modeled by GMMs), we know that we can compute the "state gamma" $\gamma(i,t)$, defined as $p(q_t^i|x_1^T, M)$, thus estimating state posterior probabilities, but actually taking into account the whole acoustic sequence x_1^T, as well as some possible prior information encapsulated through some HMM underlying topology M. In the following, we will often drop the M, keeping in mind that all recursions are processed through some prior (Markov) model M.

These "state gammas" are typically estimated by using the following α recursion:

$$\begin{aligned} \alpha(i,t) &= p(x_1^t, q_t^i) \\ &= p(x_t|q_t^i)\sum_j p(q_t^i|q_{t-1}^j)\alpha(j,t-1) \end{aligned} \tag{1}$$

and β recursion:

$$\begin{aligned} \beta(i,t) &= p(x_{t+1}^T|q_t^i) \\ &= \sum_j p(x_{t+1}|q_{t+1}^j)p(q_{t+1}^j|q_t^i)\beta(j,t+1) \end{aligned} \tag{2}$$

thus yielding the estimate of $p(q_t^i|x_1^T)$:

$$\begin{aligned} \gamma(i,t) &= p(q_t^i|x_1^T) \\ &= \frac{\alpha(i,t)\beta(i,t)}{\sum_j \alpha(j,T)} \end{aligned} \tag{3}$$

3.2 Posterior based systems

Similar recursions, also yielding to "state gammas", can also be developed for systems based on local posterior probabilities, such as hybrid HMM/ANN systems using MLPs to estimate HMM emission probabilities [6]. In this case, typically, an MLP is fed with the local acoustic vector x_t, possibly complemented by its acoustic context, i.e., x_{t-c}^{t+c} (with c typically equal to 4), and is trained to estimate $p(q_t^i|x_t)$ or $p(q_t^i|x_{t-c}^{t+c})$ at its output. In standard HMM/ANN systems, these local posteriors are usually turned into "scaled likelihood" by dividing MLP outputs by their respective a priori probability $p(q_t^i)$, as estimated on the training data, i.e.:

$$\frac{p(x_t|q_t^i)}{p(x_t)} = \frac{p(q_t^i|x_t)}{p(q_t^i)} \tag{4}$$

where the left hand side[1] of the equality is referred to as "scaled likelihood" and can be used in standard HMMs since, during recognition, $1/p(x_t)$ is simply a normalization factor independent of the state q_t^i.

In [13], it was shown that these scaled likelihood models can be used in "scaled alpha" $\alpha^s(i,t)$ and "scaled beta" $\beta^s(i,t)$ recursions to yield other gamma estimates. A different way of using scaled

[1] In the sequel of this paper, and for simplicity sake, we will often write MLP posterior outputs as $p(q_t^i|x_t)$, keeping in mind though that there are often estimating $p(q_t^i|x_{t-c}^{t+c})$ if acoustic context is provided at the input.

likelihoods was also given in [10], where the goal was initially to alleviate the scaling (underflow) problems inherent in Baum's original formulation of the forward-backward algorithm [4]. Of course, all these gammas (computed from local likelihoods or local posteriors) have the same theoretical definition (i.e., local posteriors integrating all available acoustic information, as well as possible topological constraints) and thus result in the same theoretical value. However, their estimate may be better, or have different properties, depending on the properties of the local estimators being used.

To use scaled likelihoods, we start by defining "scaled" α as:

$$\alpha^s(i,t) = \frac{p(x_1^t, q_t^i)}{\prod_{\tau=1}^{t} p(x_\tau)} \tag{5}$$

We note here that this is simply a *definition.* Thus, the product in the denominator does not imply that we have made any explicit temporal independence assumption. In fact, all the recursions used below, as well as in Section 3.1, will never make any additional temporal independence assumption than the usual *state conditional independence assumption.*[2]

Starting from (5), we can the express the scaled α recursion as follows:

$$\begin{aligned}
\alpha^s(i,t) &= \frac{p(x_t|q_t^i)}{p(x_t)} \sum_j p(q_t^i|q_{t-1}^j) \frac{p(x_1^{t-1}, q_{t-1}^j)}{\prod_{\tau=1}^{t-1} p(x_\tau)} \\
&= \frac{p(x_t|q_t^i)}{p(x_t)} \sum_j p(q_t^i|q_{t-1}^j) \alpha^s(j,t-1) \\
\alpha^s(i,t) &= \frac{p(q_t^i|x_t)}{p(q_t^i)} \sum_j p(q_t^i|q_{t-1}^j) \alpha^s(j,t-1) \qquad (6)
\end{aligned}$$

Similarly, we can define the "scaled" β and β recursion as follows:

$$\begin{aligned}
\beta^s(i,t) &= \frac{p(x_{t+1}^T|q_t^i)}{\prod_{\tau=t+1}^{T} p(x_\tau)} \qquad (7) \\
&= \sum_j \frac{p(q_{t+1}^j|x_{t+1})}{p(q_{t+1}^j)} p(q_{t+1}^j|q_t^i) \beta^s(j,t+1) \\
& \qquad (8)
\end{aligned}$$

Given that all values required in (6) and (8) are available from the MLP output, another estimate

[2] Which, in fact, may even be relaxed a bit in the case of hybrid HMM/ANN systems where we used acoustic context to estimate local posteriors.

of the state gammas $p(q_t^i|x_1^T)$ (3), denoted here as $\gamma^s(i,t)$, can thus be obtained as:

$$\begin{aligned}
\gamma^s(i,t) &= p(q_t^i|x_1^T) && (9)\\
&= \frac{p(q_t^i, x_1^T)}{p(x_1^T)} \\
&= \frac{p(x_{t+1}^T|q_t^i)p(q_t^i, x_1^t)}{p(x_1^T)} \\
&= \frac{p(x_{t+1}^T|q_t^i)p(q_t^i, x_1^t)\prod_{\tau=1}^T p(x_\tau)}{p(x_1^T)\prod_{\tau=1}^T p(x_\tau)} \\
&= \frac{p(x_{t+1}^T|q_t^i)p(q_t^i, x_1^t)\prod_{\tau=1}^T p(x_\tau)}{p(x_1^T)\prod_{\tau=1}^t p(x_\tau)\prod_{\tau=t+1}^T p(x_\tau)} \\
&= \frac{\alpha^s(i,t)\beta^s(i,t)\prod_{\tau=1}^T p(x_\tau)}{p(x_1^T)} \\
&= \frac{\alpha^s(i,t)\beta^s(i,t)\prod_{\tau=1}^T p(x_\tau)}{\sum_j p(x_1^T, q_T^j)} && (10)\\
&= \frac{\alpha^s(i,t)\beta^s(i,t)}{\sum_j \alpha^s(j,T)} && (11)
\end{aligned}$$

Again, in theory, we have:

$$\gamma(i,t) = \gamma^s(i,t) = P(q_t^i|x_1^T) \qquad (12)$$

although their estimated values will be different since different local estimators, possibly with different properties, will usually be used.

3.3 Special Case: Ergodic HMM with uniform transition probabilities

As already mentioned above, and further illustrated later, the advantages in using $\gamma(i,t)$'s, defined as $p(q_t^i|x_1^T, M)$, as "local" posterior probabilities are numerous, including:

1. Possibility of generating better posterior estimates:

   (a) By making use of a large acoustic context, which can easily extend to the whole utterance x_1^T.

   (b) Integrating specific prior knowledge, e.g., using specific HMM topologies M.

2. Possibility of using these posterior probabilities in hierarchical structures, focusing on different blocks of the recognition system, and possibly using more "optimal" (better suited) prior knowledge (e.g., HMM topologies) to the recognition level being considered (e.f., phones, words, sentences). When used as features, these γ's could also be complemented with level-specific additional features.

For our initial investigations though, we often experimented with a special case where we do not make use of any specific prior information, thus equivalent to using an ergodic HMM with uniform transition probabilities for M. In this case, we show below that the resulting γ estimates are well known values, which can be estimated locally (without requiring to run the α and β recursions).

Starting from the following equality (still requiring no additional assumption than the usual temporal state conditional independence assumption):

$$\begin{aligned}
p(x_1^T) &= \sum_k p(x_1^T, q_t^k), \quad \forall t \\
&= \sum_k p(x_{t+1}^T | q_t^k) p(x_1^t, q_t^k) \\
&= \sum_k \frac{p(x_{t+1}^T | q_t^k) \prod_{\tau=t+1}^T p(x_\tau)}{\prod_{\tau=t+1}^T p(x_\tau)} \\
&\quad \cdot \frac{p(x_1^t, q_t^k) \prod_{\tau=1}^t p(x_\tau)}{\prod_{\tau=1}^t p(x_\tau)} \\
p(x_1^T) &= \sum_k \alpha^s(t,k) \beta^s(t,k) \prod_{\tau=1}^T p(x_\tau) \qquad (13)
\end{aligned}$$

Thus, using the fact that

$$\sum_k p(x_1^T, q_t^k) = \sum_k \alpha^s(t,k) \beta^s(t,k) \prod_{\tau=1}^T p(x_\tau), \quad \forall t$$

in (10), we get the following expression for the ergodic gamma estimate $\gamma_e^s(i,t)$

$$\gamma_e^s(i,t) = \frac{\alpha^s(i,t)\beta^s(i,t)}{\sum_k \alpha^s(k,t)\beta^s(k,t)} \qquad (14)$$

Integrating the definition of $\alpha^s(i,t)$ in (14), we obtain:

$$\gamma_e^s(i,t) = \frac{\frac{p(q_t^i|x_t)}{p(q_t^i)} \sum_j p(q_t^i|q_{t-1}^j) \alpha^s(j,t-1) \beta^s(i,t)}{\sum_k \frac{p(q_t^k|x_t)}{p(q_t^k)} \sum_j p(q_t^k|q_{t-1}^j) \alpha^s(j,t-1) \beta^s(k,t)}$$

In the case of ergodic HMM with uniform transition probabilities, the sum factors in the above numerator and denominator are constant and identical and can thus be dropped, yielding:

$$\gamma_e^s(i,t) = \frac{\frac{p(q_t^i|x_t)}{p(q_t^i)}}{\sum_k \frac{p(q_t^k|x_t)}{p(q_t^k)}} \qquad (15)$$

where (15) will be referred to as "**normalized scaled likelihood**". Thus, estimating "state gammas" from local posteriors through an ergodic, uniform transition probability model, doesn't require to run α and β recursions, and it is enough to compute the above local normalized scaled likelihoods.

Furthermore, in the case of likelihood-based systems where we only have access to local likelihoods $p(x_t|q_t^i)$, equation (15) can, of course, be rewritten as:

$$\gamma_e(i,t) = \frac{p(x_t|q_t^i)}{\sum_k p(x_t|q_t^k)} \qquad (16)$$

which is then the usual "**normalized likelihood**" typically used to get posterior estimates from likelihoods (as often used, e.g., to estimate confidence measures).

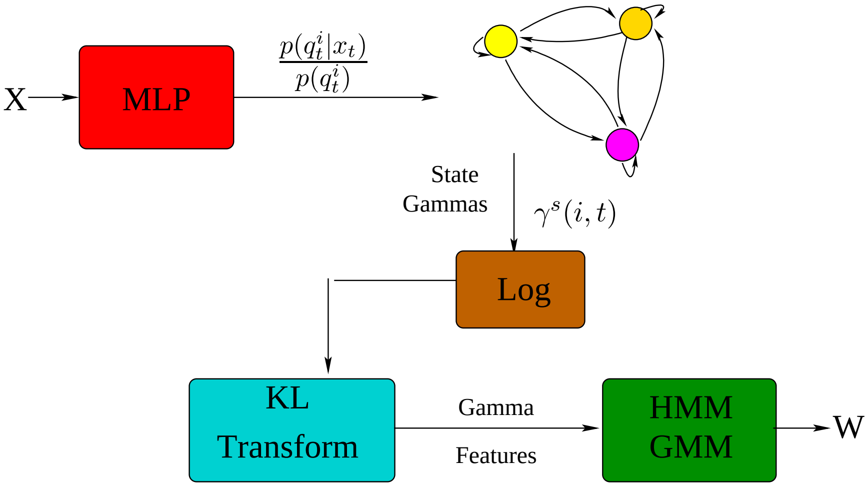

Figure 2: Approach for deriving state/phone gamma features, used as HMM-GMM features after gaussianization through log and KL transformations.

4 USING STATE GAMMAS AS FEATURES

In the next sections, and building upon the success of the Tandem approach, we investigate how our new "local" posteriors (state/phone gammas) can be used to improve state-of-the-art ASR performance. In this case, the resulting (MLP, posterior-based) system is illustrated in Figure 2. Compared with Tandem (Figure 1), we now divide the MLP output by the state priors, resulting in scaled likelihoods which are used in some HMM structure M to estimate the state gammas $\gamma^s(i,t)$, which are then used as HMM-GMM features after gaussianization through log and KL transformations.

Results are presented on Numbers'95 and a reduced vocabulary version of the DARPA Conversational Telephone Speech-to-text (CTS) task (1'000 words).

It is also possible to use gammas to recurrently compute and integrate phone gammas, word gammas, and sentence gammas, providing a convenient approach towards hierarchical HMM structures. Although we haven't yet experimented with these approaches, we discussing these ideas towards the end of the paper, in Section 5.

4.1 Numbers'95

We used the OGI Numbers database for connected word recognition task [9]. The definition of the training set, validation set and test set is similar to the one defined in [21]. The training set contains 3'233 utterances spoken by different speakers (approximately 1.5 hours) and the validation set consists of 357 utterances (used during MLP training). The test set contains 1'206 utterances. The vocabulary consists of 31 words (including silence) with a single pronunciation for each word.

The acoustic vector x_t is the PLP cepstral coefficients [14] extracted from the speech signal using a window of 25 ms with a shift of 12.5 ms, followed by cepstral mean subtraction. At each time frame t, 13 PLP cepstral coefficients, their first-order and second-order derivatives were extracted, resulting in 39 dimensional acoustic vector. There were 24 context-independent phonemes including silence.

We trained an MLP with 351 input nodes (9x39 vector), 1200 hidden units and 24 output units corresponding to the 24 context-independent phonemes. After training, the phoneme posteriors for the training set and test set were estimated and, were scaled by their respective priors (estimated from the training segmentation) to obtain scaled-likelihoods. The scaled-likelihoods were then used to estimate the state gammas according to (15). Finally, the state gammas at each frame t were transformed to 24 dimensional gamma feature vector by performing log operation and KLT[3], as done in case of tandem feature extraction [16].

For comparison purpose, we also extracted the regular tandem features by taking the value of output units prior to the nonlinearity and performing KLT, resulting in a 24 dimensional tandem feature vector at every time frame t.

For each type of feature, we trained a HMM-GMM system with 80 context-dependent phonemes, 3 emitting states per phoneme and 12 mixtures per state using the HTK-toolkit [28]. The results of the recognition studies are given in Table 1. The system using "state gamma" features performs better than both conventional ASR system using PLP features and standard Tandem system.

Features	WER
PLP	6.9%
Tandem phone posteriors (alone)	4.9%
Gamma phone posteriors (alone)	4.6%

Table 1: Word error rate (WER) on the Numbers'95 task: 31 lexicon words, 9x39(PLP)-1200-24 MLP (resulting in 25 features after KLT), 80 CD HMM/GMM phone models , 24 dimensional Tandem features, 1206 test utterances.

4.2 DARPA CTS task

The use of "state gamma" as features was further evaluated on a conversational telephone speech (CTS) recognition task. The training set for this task contained 32 hours of gender balanced CTS speech randomly selected from the Fisher Corpus and the Switchboard Corpus. The tuning/test set was a subset selected from the the NIST 2003 evaluation set. Only those utterances that covered the top most frequent 500 words with lower than 10% OOV rate were selected, resulting in 2.5 hours of data, which was further divided into a 1.2 hour tuning set and a 1.3 hour test set. The tuning and test set contained similar ratio of the number of utterances from Fisher corpus to the number of utterances from the Switchboard corpus.

Gender dependent MLPs (with 9 frames of PLP acoustic features) were trained using the training set. Each MLP was trained with 14.6 hours of speech with the remaining 1.4 hours of speech used as a cross-validation set to prevent over-training. The input layer of the MLP had 351 nodes containing 9 frames of PLP features, together with their first and second order derivatives. The hidden layer had 1300 nodes and the output layer had 46 outputs, associated with the 46 phones defined in the 2003 SRI Decipher system. This resulted in an MLP with about 500K parameters. After training, the phone posteriors for the training set and the test set were computed, and used to estimate the "state gammas", which were then sent as features to a HMM-GMM system for recognition.

The baseline system used PLP with the first two derivatives as the feature. It was computed with vocal tract normalization and mean and variance normalization. The SRI decipher system was used for the recognition (HMM-GMM) experiments. Gender dependent models were trained through 7 iterations: 2 on context independent models, 2 on context dependent models, and 3 on Genone clustered models [25] using an ML criterion. The trained model had 1498 Gaussian clusters for female and 1725 clusters for male. Each Gaussian cluster contained 64 Gaussians. A 1000 word dictionary with multi-words and multi-pronunciations was used in decoding, using a bi-gram language model (LM) based on the 1000 words and extracted from the LM in the 2004 SRI system.

State posteriors (as typically used in Tandem systems) or "state gammas" (as proposed here) were combined with the baseline PLP feature in the same way as described in [30] through log, KLT and truncation. The resulting 25 dimension feature vector is then appended to the baseline PLP feature vector. For such a long feature, the Gaussian weight [30] is set to 0.3 compared to 0.8 for the baseline

[3] PCA statistics obtained from the training data.

system. The resulting augmented feature then go through the same HMM-GMM decoding processing as the baseline feature.

For the initial experiments we only tested on one gender (male, which for these data is more difficult to recognize than the female test set), and only used an ergodic HMM for the estimation of the state gamma, which is thus equivalent to using the normalized scaled likelihoods (15) estimated from the MLP output as features.

The results reported in Table 2 show that the "state gamma" features (i.e., PLP features augmented with Gamma phone posteriors) give significant improvement compared with the PLP baseline, and the PLP feature augmented with phone posteriors (Tandem).

Features	WER
PLP	44.3%
PLP+Tandem phone posteriors	42.5%
PLP+Gamma phone posteriors	41.7%

Table 2: Word error rate (WER) on the male part of a reduced vocabulary version of the DARPA Conversational Telephone Speech-to-text (CTS) task: 1'000 lexicon words, with multi-words and multi-pronunciations, 9x39(PLP)-1300-46 MLP (resulting in 25 features after KLT).

5 HIERARCHICAL POSTERIOR BASED ASR

5.1 Hierarchical gamma estimation

Although not evaluated here, another extension to the approach discussed in the present paper, is to integrate the state gammas γ (i.e., improved local posterior estimates, taking all acoustic data, as well as possible prior information or topological constraints, into account), estimated according to (3) or (11), into new recursions, yielding estimates of phone gammas γ_p and word gammas γ_w.

Phone gammas $\gamma_p(i,t)$ can indeed be expressed in terms of state gammas $\gamma(i,t)$ as follows:

$$\begin{aligned}
\gamma_p(i,t) &= p(p_t^i|x_1^T) && (17)\\
&= \sum_{j=1}^{N_q} p(p_t^i, q_t^j|x_1^T) \\
&= \sum_{j=1}^{N_q} p(p_t^i|q_t^j, x_1^T)p(q_t^j|x_1^T) \\
&= \sum_{j=1}^{N_q} p(p_t^i|q_t^j, x_1^T)\gamma(j,t) && (18)
\end{aligned}$$

where probability $p(p_t^i|q_t^j, x_1^T)$ represents the probability of being in a given phoneme p_t^i at time t knowing the state q_t^i at time t. If there is no parameter sharing between phonemes, this is deterministic and equal to 1 or 0. Otherwise, this can also easily be estimated from the training data.

Similarly, at a next level, we can also integrate phone gammas $\gamma_p(i,t)$ into word gammas $\gamma_w(i,t)$

as follows:

$$\begin{aligned}\gamma_w(i,t) &= p(w_t^i|x_1^T) \qquad (19)\\ &= \sum_{j=1}^{N_p}\sum_{k=1}^{N_p} p(w_t^i, p_t^j, p_{t-1}^k|x_1^T)\\ &= \sum_{j=1}^{N_p}\sum_{k=1}^{N_p} p(w_t^i|p_t^j, p_{t-1}^k, x_1^T)p(p_t^j, p_{t-1}^k|x_1^T)\\ &= \sum_{j=1}^{N_p}\sum_{k=1}^{N_p} p(w_t^i|p_t^j, p_{t-1}^k, x_1^T)p(p_t^j|p_{t-1}^k, x_1^T)\\ &\qquad\qquad \cdot p(p_{t-1}^k|x_1^T)\\ &= \sum_{j=1}^{N_p}\sum_{k=1}^{N_p} p(w_t^i|p_t^j, p_{t-1}^k, x_1^T)p(p_t^j|p_{t-1}^k, x_1^T)\\ &\qquad\qquad \cdot \sum_{r=1}^{N_q} p(p_{t-1}^k|q_{t-1}^r, x_1^T)\gamma(r,t-1)\end{aligned} \qquad (20)$$

where the last sum factor is nothing else but $\gamma_p(k, t-1)$. Probability $p(w_t^i|p_t^j, p_{t-1}^k, x_1^T)$ now represents the probability of being in a given word knowing the phoneme sequence $\{p_{t-1}^k, p_t^j\}$, which thus encodes some form of lexical information. Obviously, this probability is also independent of x_1^T, and can easily be estimated during training from the results of forced alignment. Probability $p(p_t^j|p_{t-1}^k, x_1^T)$ represents the phoneme transition probability. Assuming this probability to be independent of x_1^T, it can be estimated from the training data.[4]

5.2 Decoding

Decoding can then be performed in different ways using gammas. In the case of Viterbi decoding, we define

$$V(i,t) = \max_{w_1^{t-1}} p(w_t^i, w_1^{t-1}|x_1^T) \qquad (21)$$

which can be derived recursively as follows:

$$\begin{aligned}V(i,t) &= \max_{j, w_1^{t-2}} p(w_t^i, w_{t-1}^j, w_1^{t-2}|x_1^T)\\ &= \max_j [p(w_t^i|w_{t-1}^j, w_1^{t-2}, x_1^T)\cdot\\ &\qquad \max_{w_1^{t-2}} p(w_{t-1}^j, w_1^{t-2}|x_1^T)]\\ &= \max_j p(w_t^i|w_{t-1}^j, w_1^{t-2}, x_1^T)V(j,t-1)\end{aligned} \qquad (22)$$

We can now for instance reasonably decide to estimate $p(w_t^i|w_{t-1}^j, w_1^{t-2}, x_1^T)$ as follows:

$$\begin{aligned}&p(w_t^i|w_{t-1}^j, w_1^{t-2}, x_1^T) \qquad (23)\\ &= \begin{cases} p(w_t^i|x_1^T) & \text{if } i = j\\ p(w_t^i|x_1^T)p(w_t^i|w_{t-1}^j) & \text{otherwise.}\end{cases}\end{aligned}$$

[4] We note here that the value of those transition probabilities will now have a bigger impact than in standard likelihood-based systems since they are combined only with a posteriori probabilities.

hence, the gamma term on the words reappear, as well as a language model over words.

6 CONCLUSIONS

In this paper, we first briefly discussed the approaches currently using local posterior probabilities as local measures or as features in ASR systems. Indeed, several approaches in that direction have recently been shown to have a significant potential to improve state-of-the-art ASR systems. However, we also believe that further progress in that direction will critically depend on several factors such as (1) improving the quality of these posterior estimates (using, e.g., GMMs or MLPs), while (2) preserving a strong theoretical framework, permitting the hierarchical integration of posteriors corresponding to processing level, while also integrating at each of these levels all possible information present in the data, as well as all possible appropriate a priori information (e.g., represented as specific HMM topological constraints).

We believe that the present paper introduced a general framework in this direction. Simply stated, we proposed here to replace the use of local posterior probabilities (used as local measures or as features) by new estimates of those local posteriors, usually referred to as "gamma posteriors", using different versions of the α and β (likelihood or posterior-based) recursions.

When used as features, even for the simplified case that corresponds to the scaling and normalizing of local posteriors, we have shown here that gammas can yield significant performance improvements on two different speech recognition tasks.[5]

7 ACKNOWLEDGEMENTS

The authors would like to thank several of their colleagues for stimulating and fruitful discussions, including: Andreas Stolcke (SRI), Hynek Hermansky (IDIAP), Daniel Gatica (IDIAP), Jean-Marc Odobez (IDIAP), Dong Zhang (IDIAP) and Hamed Ketabdar (IDIAP).

This project was jointly funded by the DARPA EARS project, the European AMI project, as well as the IM2 Swiss National Center of Competence in Research.

References

[1] 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.

[2] Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H, , Jain, J., Kajarekar, S., Morgan, N., and Sivadas, S. "QUALCOMM-ICSI-OGI Features for ASR", *Proceedings of the Intl. Conference on Spoken Language Processing* (Denver, USA), September 2002.

[3] Atick, J.J, "Could information theory provide an ecological theory of sensory processing?", *Network: Computation in Neural Systems*, Vol. 3, pp. 213-251, 1992

[4] Baum, L.E., "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes", *Inequalities*, Vol. 3, pp. 1-8, 1972.

[5] Bernardis, G. and Bourlard, H., "Improving posterior confidence measures in hybrid HMM/ANN speech recognition system", *Proceedings of the Intl. Conference on Spoken Language Processing* (Sydney, Australia), pp. 775-778, 1998.

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

[5] Similar approaches have also shown significant improvements on complex multimodal processing tasks currently under investigation at IDIAP.

[7] Bourlard, H., Konig, Y., and Morgan, N., "A training algorithm for statistical sequence recognition with applications to transition-based speech recognition," *IEEE Signal Processing Letters*, Vol. 3, No. 7, pp. 203-205, 1996.

[8] Chen, B., Zhu, Q., and Morgan, N., "Learning long-term temporal features in LVCSR using neural networks", *Proc. Interspeech'04* (Korea), October 2004

[9] Cole, R. A., Fanty, M., Noel, M., and Lander, T., "Telephone speech corpus development at CSLU", *Proceedings of the Intl. Conference on Spoken Language Processing* (Yokohama, Japan), September 1994.

[10] Devijver, P., "Baum's forward-backward algorithm revisited," *Pattern Recognition Letters*, Vol. 3, pp. 369-373, 1985.

[11] Fukunaga, K., *Statistical Pattern Recognition*, Academic Press, San Diego, 1990.

[12] Hatch, A.O., *Word-level confidence estimation for automatic speech recognition*, M.S. Thesis, ICSI Technical Report, Berkeley, April 2002.

[13] Hennebert, J., Ris, C., Bourlard, H., Renals, S., and Morgan, N., "Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems," *Proceedings of EUROSPEECH'97* (Rhodes, Greece), pp. 1951-1954, 1997.

[14] Hermansky, H., "Perceptual linear predictive(PLP) analysis of speech", *Journal of the Acoustical Society of America*, vol. 87, number 4, pp. 1738–1752, 1990.

[15] Hermansky, H. and Sharma S., "TRAPS Classifiers of Temporal Patterns", *Proceedings of Intl. Conf. on Spoken Language Processing* (Sydney, Australia), 1998.

[16] Hermansky, H., Ellis, D.P.W., and Sharma, S., "Connectionist Feature Extraction for Conventional HMM Systems", *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing* (Istanbul, Turkey), 2000.

[17] Hunt, M.J., "A statistical approach to metrics for word and syllable recognition", *Journal Acoust. Soc. Am.*, 66(S1), S35(A), 1979

[18] Ikbal, S., Misra, H., Sivadas, S., Hermansky, H., and Bourlard, H., "Entropy Based Combination of Tandem Representations for Robust Speech Recognition", *Proc. Interspeech'04* (Korea), October 2004

[19] Lewicki, M-S., "Efficient coding of natural sounds", *Nature Neuroscience*, 5(4), pp. 356-363, 2002

[20] 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.

[21] Mirghafori, N. and Morgan, N.,"Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers", *Proceedings of the Intl. Conf. on Spoken Language Processing* (Sydney, Australia), pp. 743-746, November 1998.

[22] Morgan, N. et al., *DARPA-EARS Meeting*, Boston, MA, May 2003

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

[24] Reyes-Gomez, M.J. and Ellis, D.P.W., "Error visualization for tandem acoustic modeling on the Aurora task", *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Orlando, Florida), May 2002.

[25] Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Rao Gadde, V.R., Plauche, M., Richey, C., Shriberg, E., Sonmez, K., Weng, F., and Zheng, J., "The SRI March 2000 Hub-5 conversational speech transcription system", *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[26] Sivadas, S. and Hermansky, H., "Hierarchical Tandem Feature Extraction", *Proceedings of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing* (Orlando, Florida), May 2002.

[27] William, G and Renals, S., "Confidence measures from local posterior estimate", *Computer, Speech and Language*,Vol. 13, pp. 395-411, 1999.

[28] 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.

[29] Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., and Lathoud, G., "Modeling individual and group actions in meetings: a two-Layer HMM framework", *IEEE Workshop on Event Mining at the Conference on Computer Vision and Pattern Recognition, CVPR*, 2004.

[30] Zhu, Q., Chen, B., Morgan, N., and Stolcke, A., "On using MLP features in LVCSR", *Proc. Interspeech'04* (Korea), October 2004