IDIAP Research Report

Enhanced Phone Posteriors for Improving Speech Recognition Systems

Hamed Ketabdar [a] Herve Bourlard [a]

IDIAP–RR 08-39

March 2008

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**

Av. des Prés−Beudin 20 P.O. Box 592 1920 Martigny − Switzerland

Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-39

Enhanced Phone Posteriors for Improving Speech Recognition Systems

Hamed Ketabdar Herve Bourlard

March 2008

1 INTRODUCTION

The use of posterior probabilities for improving Automatic Speech Recognition (ASR) systems has become popular and frequently investigated in the past decade. Posterior probabilities have mainly been used either as local acoustic scores (measures) or as acoustic features in ASR 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 HMMs. Hybrid HMM/ANN method allows for discriminant training, as well as for the possibility of using small acoustic context by presenting few 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 (usually logarithm and Karhunen-Loeve transform), are used as acoustic feature inputs to a HMM/GMM module. Tandem takes the advantage of discriminative acoustic model training, as well as being able to use the techniques developed for standard HMM/GMM systems.

In both hybrid HMM/ANN and Tandem approaches, posteriors are estimated using ANNs (more specifically MLPs), based only on the acoustic information in a local frame or a limited number of local frames. In this paper, we call these posteriors "MLP posteriors" or "regular posteriors". However, a limited window of spectral features is not the only source of knowledge available about phones. Information about phones are spread over time, and there are no sharp boundaries between phones [6, 7]. Phonemes have specific duration constraints (phonetic knowledge), follow specific sub-lexical and lexical rules (lexical knowledge), etc. These long contextual and prior sources of knowledge can help in providing better phone posterior estimates, however they are not usually taken into account in the MLP based phone posterior estimation. There have been few recent studies with the goal of integrating context and prior knowledge in the posterior estimation [8, 9, 10]. In these studies, different methods for estimating posterior probability of a word hypothesis, given all acoustic observations of the utterance is proposed. These posteriors are estimated on HMMs or word graphs by the forward-backward (Baum-Welch) algorithm [11], and used for word confidence measurement. These studies are mainly focused on estimating word posteriors for the purpose of hypothesis confidence measurement.

In this paper, we present a principled framework for enhancing the estimation of posteriors (particularly phone posteriors) by integrating long acoustic context, as well as prior phonetic and lexical knowledge. However, as opposed to the above approaches, the goal here is to provide enhanced posteriors which can be used in frame synchronous posterior based ASR applications. The input in our approaches is regular phone posteriors estimated by an MLP, and the outcome is the "enhanced posteriors" of *phones*[1] at the *frame* level. Many posterior based ASR algorithms are based on phone evidences at the frame level. Therefore, the resulting frame based enhanced posteriors can be used in a wide range of posterior based ASR systems (e.g. Tandem and hybrid HMM/ANN), as replacement or in combination with the regular MLP posteriors in a straightforward manner.

We propose two approaches for estimating these posteriors. The first approach uses a HMM to integrate the prior phonetic and lexical knowledge. The phonetic and lexical knowledge is encoded in the topology of the HMM. The integration is realized by using the regular MLP posteriors as emission probabilities in the HMM forward and backward recursions (Baum-Welch approach) [11]. This yields new enhanced posterior estimates taking into account the encoded knowledge in the topology of the HMM. The second approach uses a secondary neural network (MLP) to post-process a temporal context of regular phone posteriors, and learn long term intra and inter dependencies between regular phone evidences (posteriors) estimated initially by the first MLP. These long term dependencies can be interpreted as prior phonetic knowledge. The learned prior phonetic knowledge is integrated in the phone posterior estimation, during the inference (forward pass) of the second MLP, resulting in enhanced posteriors.

The proposed methods provide a general framework for integrating acoustic context and different

[1] Although as it is shown in Section 6, we can also use our approach for word posterior estimation.

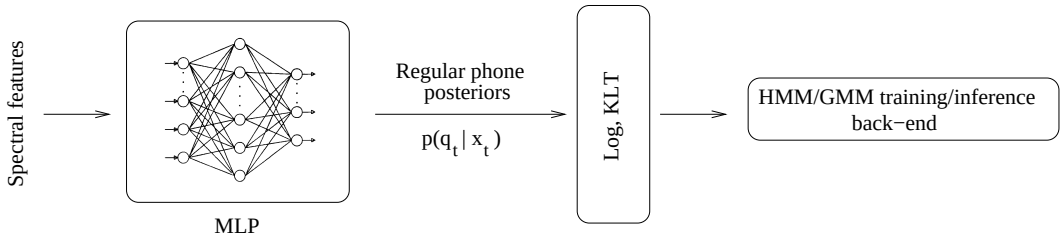

Figure 1: Standard approach for deriving and using Tandem features. The phone posterior vectors $p(q_t|x_t)$ are estimated using MLP. $p(q_t|x_t)$ is a vector of phone posterior probabilities at time t. These posteriors are gaussianized and decorrelated using log and KLT transforms. The result of the transformation is used as acoustic features for training and inference in a standard HMM/GMM back-end.

prior knowledge for improving posterior estimation in ASR, from phone up to the word units. In this paper, we mainly focus on phone posteriors. We present different aspects and applications of these enhanced posterior estimates for improving ASR systems. We show that they can be used as features, or as complementary features to regular phone posteriors in Tandem configuration. We have achieved consistent word recognition improvement with the new Tandem configuration on Numbers'95 [12] and Conversational Telephone Speech (CTS) [13] databases. The enhanced posteriors are also used as local scores for decoding in hybrid HMM/MLP configuration. We have again observed improved recognition performance on these databases (plus TIMIT database [14]), and also interesting results on the robustness of the performance with respect to ad-hoc tuning parameters (e.g. phone and word insertion penalties). Simply stated, we propose to replace or complement the use of regular MLP posteriors by the new enhanced estimates of these posteriors, and we show some important practical cases. One can think of other frame synchronous posterior based ASR systems, and simply use the enhanced posteriors as replacement or in combination with the regular MLP posteriors.

The paper is organized as follows: Section 2 reviews different approaches for estimating and using phone posterior probabilities in ASR. In Section 3, we present two approaches for integrating context, phonetic and lexical knowledge in the posterior estimation. In Section 4, we discuss the usage of the enhanced posteriors as features in Tandem configuration. Section 5 studies the usage of the enhanced posteriors as local scores in a hybrid HMM/MLP decoder. Section 6 discusses about other possible usages of the enhanced posteriors in ASR. Summary and conclusions will appear in Section 7.

2 POSTERIORS IN SPEECH RECOGNITION SYSTEMS

Sub-word (phone) posterior probabilities have been mainly estimated using Artificial Neural Networks (ANNs), and particularly Multi-Layer Perceptrons (MLPs). In these approaches, a limited number of spectral feature frames is presented at the input of the MLP. Each output of the MLP is associated with a particular phone. The MLP is discriminatively trained to find a mapping between the spectral features at the input, and the phone targets at the output. The MLP estimates $p(q_t^i|x_t)$, where x_t is a spectral feature frame at time t, $q^i \in Q = \{q^1, ..., q^i, ..., q^{N_q}\}$ (N_q total number of MLP outputs) is the i^{th} MLP output associated with phone i, and q_t^i represents the event of having phone i at time t. Although ANNs have been the most dominant tool for estimating phone posteriors in ASR, some other principled approaches have also been studied [15, 16]. In [15], a method based on using Gaussian Mixture Models (GMMs) for estimating posteriors has been proposed. In this method, a large number of Gaussians are pooled from an acoustic model trained with maximum likelihood (ML) criterion. The likelihoods estimated using these Gaussians are normalized (assuming equal priors) to obtain a sparse set of posteriors. The dimensionality of this set is reduced by a transformation learned along with Minimum Phone Error (MPE) training [17]. In [16], likelihoods estimated by GMMs (trained on acoustic data) are turned into posteriors through conditional random fields. In

this work, we are mainly concerned about the approaches using ANNs for posterior estimation.

2.1 Posteriors As Local Classifiers (scores)

Hybrid HMM/ANN approaches were probably among the first ones to make extensive use of posterior probabilities in speech recognition. In these approaches, ANNs and more specifically MLPs are used to estimate the emission probabilities required in HMM systems [1]. It has been shown that if each output unit $q^i \in Q = \{q^1, ..., q^i, ..., q^{N_q}\}$ of an MLP is associated with a particular state of the set of possible HMM states, it is possible to train the MLP in a discriminative way, to generate posterior probabilities of the output classes conditioned on the input, i.e., $p(q_t^i|x_t)$, where x_t is a spectral feature frame at time t, and q_t^i represents the event of having phone i at time t. Usually more than one frame of acoustic features (small context) is presented at the input of the MLP, thus it estimates $p(q_t^i|x_{t-c}^{t+c})$, where c is typically equal to 4. x_{t-c}^{t+c} represents a short temporal context obtained by concatenating acoustic feature vectors in $\{x_{t-c}, ..., x_t, ..., x_{t+c}\}$. This is in fact very limited context[2].

Posterior probabilities have also been used as local measures for different ASR purposes, such as (1) estimating confidence measures [4, 10, 18], (2) beam search pruning [3], or (3) word lattice rescoring [2].

2.2 Posteriors As Features

The properties described above were also extended by using the MLP-generated posterior probabilities as acoustic features, which (after some transformations) 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 nonlinear discriminant analysis). One of the earlier and most successful approaches based on using posteriors as features is Tandem [5]. 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 sub-word 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 [1], is used for estimating posterior probabilities of context-independent phones. Alternatives such as GMM-derived posteriors were also investigated [19]. Hierarchical classification schemes in Tandem estimator were also investigated [20].

As illustrated in Fig. 1, the MLP phone posterior estimates $p(q_t|x_t)$ are gaussianized by a static nonlinearity (usually logarithm) and whitened by the Karhunen-Loeve transform (KLT) derived from training data. $p(q_t|x_t)$ is a vector of phone posterior probabilities at time t with the components $p(q_t^i|x_t)$ for $i \in \{1, ..., i, ..., N_q\}$. Such gaussianized and whitened posterior probabilities form the feature vector for the subsequent HMM/GMM training/inference back-end. Thus, the conventional features derived from a spectral density vector representing the spectral envelope are replaced by the transformed posteriors of acoustic events (context-independent phones). 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 [21, 22].

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 super frame of typical speech features such as 9 frames of concatenated PLP static and dynamic features [23]. Often, Tandem inputs are concatenated outputs from other sub-band classifiers (TRAP [24] or HATS [25]). TRAP has been reported to be efficient in alleviating irrelevant information [26] [27].

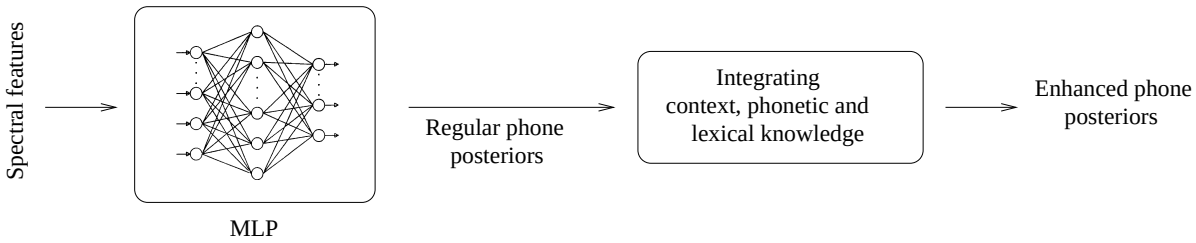

Figure 2: General idea: First, regular phone posteriors are estimated using an MLP, then these posteriors are post-processed in a secondary module to integrate context, phonetic and lexical knowledge. This results in enhanced phone posterior estimates.

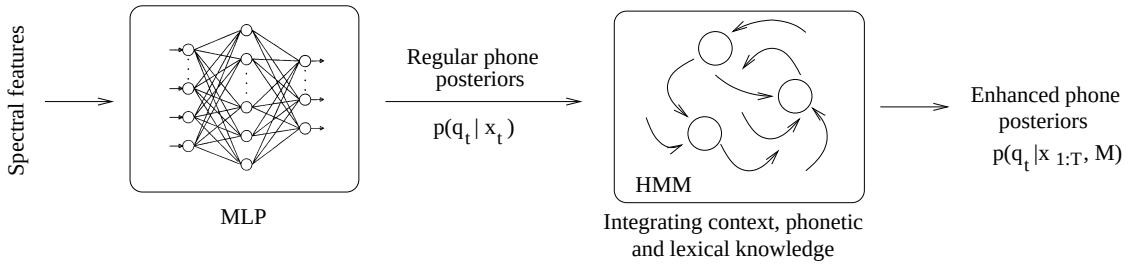

Figure 3: HMM-based enhanced posterior estimation: First, regular phone posterior vectors $p(q_t|x_t)$ are estimated using an MLP. These posteriors are used as emission probabilities in HMM recursions to estimate state posteriors. The HMM state posteriors are then integrated into enhanced phone posterior vectors $p(q_t|x_{1:T}, M)$.

3 ENHANCING POSTERIOR PROBABILITY ESTIMATION

In the previous section, we have studied the estimation and usage of posterior probabilities in speech recognition systems. Typically, the estimation of posteriors is based only on a local or limited number of spectral feature frames. In this paper, we call these posteriors as "MLP posteriors" or "regular posteriors". However, the time limited spectral information is not the only source of knowledge available about phones. There are other sources of knowledge which can help to provide more informative estimates of phone posteriors. Information about phones are spread over time in the speech signal and there are no sharp boundaries between phones [6, 7], therefore taking into account long contextual information can be useful. Moreover, some prior knowledge such as duration of phones (phonetic knowledge) and the lexical usage of phones in a word can be useful for improving posterior estimates.

In this section, we study how these extra sources of knowledge (acoustic context and prior phonetic and lexical knowledge) can be integrated in the posterior estimation to improve the posterior estimates. The general idea is illustrated in Fig. 2. The regular phone posteriors estimated by a neural network (MLP) are post-processed by a secondary module to integrate context, phonetic, and lexical knowledge. We propose two different approaches for integrating these higher level knowledge in the posteriors estimation. The first approach is based on estimating posteriors through a HMM, to integrate the phonetic and lexical knowledge encoded in the HMM topology in the posterior estimation. The second one is based on using a secondary neural network (MLP) to post-process a long temporal context of regular phone posteriors. In the following, we study these approaches.

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

3.1 HMM-based Integration of Prior and Contextual Knowledge

Topological constraints in a HMM encode specific prior phonetic and lexical knowledge. This knowledge can be integrated in the regular MLP posteriors to get an enhanced version of these posterior estimates. This objective can be formulated as turning the regular estimate of phone posteriors $p(q_t^i|x_t)$ obtained by MLP, to a more informative posterior $p(q_t^i|x_{1:T}, M)$, where q_t^i is the event of having phone i at time t, $x_{1:T} = \{x_1, ..., x_t, ..., x_T\}$ is the acoustic context as available possibly in the whole utterance, and M is HMM model encoding specific prior knowledge. We have used HMM/ANN formalism for integrating HMM topological constraints in the MLP posterior estimates. The integration is done by using phone posteriors $p(q_t^i|x_t)$ as state emission probabilities in the HMM. Each state s^k of the set of HMM states $S = \{s^1, ..., s^k, ..., S^{N_s}\}$ (N_s total number of HMM states) is associated with one of MLP outputs representing a phone posterior probability. The state emission probabilities are used in HMM forward-backward recursions [11] to integrate HMM topological constraints (encoding specific prior knowledge). This gives the estimates of HMM state posteriors $p(s_t^k|x_{1:T}, M)$, where s_t^k is the event of having state k at time t. The state posteriors will then integrated to enhanced phone posteriors $p(q_t^i|x_{1:T}, M)$ by accumulating posteriors of all the states modeling phone i in the HMM. In the forward-backward recursions and state posterior estimation, we have the contribution of the HMM topological constraints (prior knowledge) in addition to the MLP posteriors (emission probabilities). Therefore, the state posteriors (and consequently phone posteriors) can be interpreted as the integration of topological constraints (prior knowledge) in the MLP posteriors. Here we first review the forward-backward recursions for conventional likelihood based HMM systems, then we study forward-backward recursions for the case of modeling state probability distributions with MLP outputs.

According to the standard HMM formalism, the state posterior is defined as the probability of being in state k at time t, s_t^k, given the whole observation sequence $x_{1:T}$ and the HMM model M encoding specific prior knowledge (topological/temporal constraints):

$$\gamma(k,t) = p(s_t^k|x_{1:T}, M) \tag{1}$$

where x_t is a feature vector at time t, $x_{1:T} = \{x_1, \ldots, x_T\}$ is an acoustic observation sequence, s_t is the HMM state at time t, which value can range from 1 to N_s (total number of HMM states), and s_t^k shows the event "$s_t = k$". In the following, we will often drop the M, keeping in mind that all recursions are processed through some prior (Markov) model M. We call $\gamma(k,t)$ as "state posterior".

The state posteriors $\gamma(i,t)$ can be estimated using forward α and backward β recursions (as referred to in HMM formalism) [11] using local emission likelihoods $p(x_t|s_t^k)$:

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

thus yielding the estimate of $p(s_t^k|x_{1:T}, M)$:

$$\gamma(k,t) = p(s_t^k|x_{1:T}, M) = \frac{\alpha(k,t)\beta(k,t)}{\sum_j \alpha(j,T)} \tag{4}$$

Similar recursions, also yielding to "state posteriors", can also be developed for systems based on local posterior probabilities, such as hybrid HMM/ANN systems using MLPs to estimate HMM emission probabilities [1]. Each HMM state k is associated with one MLP output $p(q_t^i|x_t)$ representing posterior probability for phone i at time t. 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. $\frac{p(q_t^i|x_t)}{p(q_t^i)}$. The scaled likelihoods are used as state emission probabilities in HMM/ANN ASR. For HMM state k at time t associated with the phone i we have:

$$\frac{p(x_t|s_t^k)}{p(x_t)} = \frac{p(s_t^k|x_t)}{p(s_t^k)} \tag{5}$$

The scaled likelihood at the left hand side of (5) is used in standard HMMs since, during recognition, $1/p(x_t)$ is simply a normalization factor independent of the state s_t^k.

In [28], it was shown that these scaled likelihoods can be used in "scaled alpha" $\alpha^{scale}(k,t)$ and "scaled beta" $\beta^{scale}(k,t)$ recursions to yield state posterior estimates:

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

$$\alpha^{scale}(k,t) = \frac{p(x_1^t, s_t^k)}{\prod_{\tau=1}^t p(x_\tau)} \tag{6}$$

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, will never make any additional temporal independence assumption than the usual state conditional independence assumption.

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

$$\begin{aligned}
\alpha^{scale}(k,t) &= \frac{p(x_t|s_t^k)}{p(x_t)} \sum_j p(s_t^k|s_{t-1}^j) \frac{p(x_{1:t-1}, s_{t-1}^j)}{\prod_{\tau=1}^{t-1} p(x_\tau)} \\
&= \frac{p(x_t|s_t^k)}{p(x_t)} \sum_j p(s_t^k|s_{t-1}^j) \alpha^{scale}(j,t-1) \\
\alpha^{scale}(k,t) &= \frac{p(s_t^k|x_t)}{p(s_t^k)} \sum_j p(s_t^k|s_{t-1}^j) \alpha^{scale}(j,t-1)
\end{aligned} \tag{7}$$

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

$$\begin{aligned}
\beta^{scale}(k,t) &= \frac{p(x_{t+1:T}|s_t^k)}{\prod_{\tau=t+1}^T p(x_\tau)} \\
&= \sum_j \frac{p(s_{t+1}^j|x_{t+1})}{p(s_{t+1}^j)} p(s_{t+1}^j|s_t^k) \beta^{scale}(j,t+1)
\end{aligned} \tag{8}$$

Given that all values required in (7) and (8) are available from the MLP output, another estimate of the state posteriors $p(s_t^k|x_{1:T}, M)$, denoted here as $\gamma^{scale}(k,t)$, can thus be obtained as:

$$\begin{aligned}
\gamma^{scale}(k,t) &= p(s_t^k|x_{1:T}, M) \\
&= \frac{p(s_t^k, x_{1:T})}{p(x_{1:T})} \\
&= \frac{p(x_{t+1:T}|s_t^k) p(s_t^k, x_{1:t})}{p(x_{1:T})} \\
&= \frac{p(x_{t+1:T}|s_t^k) p(s_t^k, x_{1:t}) \prod_{\tau=1}^T p(x_\tau)}{p(x_{1:T}) \prod_{\tau=1}^T p(x_\tau)} \\
&= \frac{p(x_{t+1:T}|s_t^k) p(s_t^k, x_{1:t}) \prod_{\tau=1}^T p(x_\tau)}{p(x_{1:T}) \prod_{\tau=1}^t p(x_\tau) \prod_{\tau=t+1}^T p(x_\tau)}
\end{aligned}$$

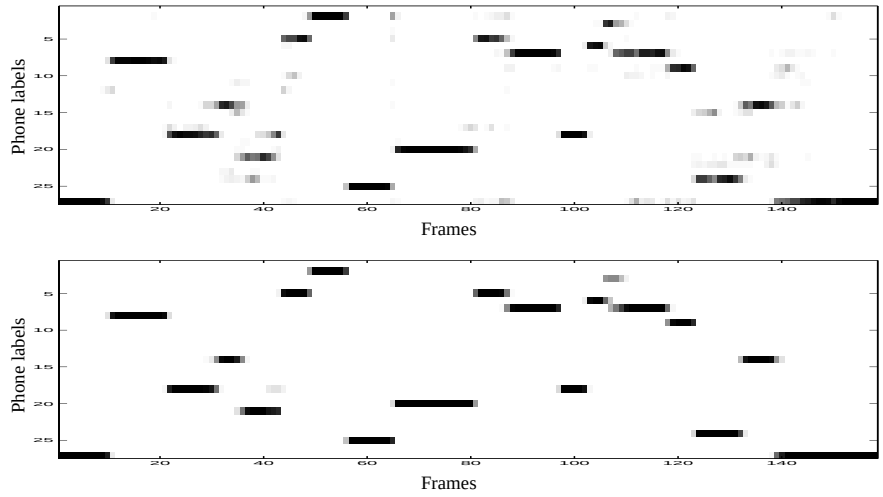

Figure 4: (top) MLP estimated phone posteriors, and (bottom) corresponding enhanced phone posteriors. The y-axis is showing phone labels and x-axis is showing frames. Intensity of each block shows the posterior value. The enhanced posteriors look more confident.

$$
\begin{aligned}
&= \frac{\alpha^{scale}(k,t)\beta^{scale}(k,t)\prod_{\tau=1}^{T} p(x_\tau)}{p(x_{1:T})} \\
&= \frac{\alpha^{scale}(k,t)\beta^{scale}(k,t)\prod_{\tau=1}^{T} p(x_\tau)}{\sum_j p(x_{1:T}, s_t^j)} \\
&= \frac{\alpha^{scale}(k,t)\beta^{scale}(k,t)}{\sum_j \alpha^{scale}(j,T)}
\end{aligned} \tag{9}
$$

Again, in theory, we have:

$$\gamma(k,t) = \gamma^{scale}(k,t) = p(s_t^k|x_{1:T}, M) \tag{10}$$

In this work, we always use hybrid HMM/ANN configuration for the estimation of HMM state posterior probabilities. This means that the MLP posteriors (after turning to scaled likelihoods), are used as emission probabilities in the forward-backward recursions.

The estimated state posteriors are then used to estimate phone posteriors. The enhanced phone posteriors $p(q_t^i|x_{1:T,M})$ can be expressed in terms of state posteriors $\gamma(k,t)$ as follows:

$$
\begin{aligned}
p(q_t^i|x_{1:T,M}) &= \sum_{k=1}^{N_s} p(q_t^i, s_t^k|x_{1:T}) \\
&= \sum_{k=1}^{N_s} p(q_t^i|s_t^k, x_{1:T}) p(s_t^k|x_{1:T}) \\
&= \sum_{k=1}^{N_s} p(q_t^i|s_t^k, x_{1:T}) \gamma(k,t)
\end{aligned} \tag{11}
$$

where $p(q_t^i|x_{1:T}, M)$ is the enhanced phone posterior for phone i at time t. Probability $p(q_t^i|s_t^k, x_{1:T})$ represents the probability of being in a given phone i at time t knowing to be in the state k at time t. If there is no parameter sharing between phones, this is deterministic and equal to 1 or 0. Otherwise, this can be estimated from the training data. In this work, we assume that there is no parameter sharing between phones, thus a phone posterior is estimated by adding up all state posteriors associated with

the phone in the whole model. This way, the new enhanced phone posterior estimates $p(q_t^i|x_{1:T}, M)$ integrating context and prior knowledge is obtained. In the reminder of the paper, we call them as "HMM-based enhanced posteriors".

Figure 3 is showing the configuration for the HMM-based integration of prior and contextual knowledge. As it is shown, the regular phone posterior vectors $p(q_t|x_t)$ are initially estimated using an MLP. $p(q_t|x_t)$ is a vector of phone posteriors at time t with the components $p(q_t^i|x_t)$ for $i \in \{1, ..., i, ..., N_q\}$. These phone posteriors are turned into scaled likelihoods (by dividing them by the corresponding priors), and used as emission likelihoods in the HMM. The HMM state posteriors are estimated using HMM forward-backward recursions. The state posteriors are then integrated to enhanced phone posteriors $p(q_t|x_{1:T}, M)$. $p(q_t|x_{1:T}, M)$ is a vector of enhanced phone posteriors at time t. The obtained phone posteriors are more informative (enhanced) than regular MLP posteriors, since the prior knowledge (encoded in the topology of the HMM), and long acoustic context (as available in the whole utterance) is additionally taken into account to estimate them. In fact, the second module (the HMM) gets phone initial evidences (MLP posteriors) as input, and acts as a corrective filter by introducing context and prior knowledge. The corrective filter suppresses the effect of evidences not matching with prior knowledge or contextual information, and magnifies the effect of evidences matching them. The output of this corrective filter is enhanced evidences in the form of posteriors.

Figure 4 is showing a sample of regular MLP posteriors and corresponding enhanced posteriors obtained by integrating phone duration information. The enhanced posterior estimates look more confident. The MLP posteriors at the top are used as local estimators (emission probabilities) in the HMM estimating enhanced posteriors (bottom).

The HMM module used for enhanced posterior estimation can have different topologies, thus encoding different types of prior knowledge. As the simplest case, phones can be modeled with a minimum number of states, and be connected using ergodic uniform transition probabilities. In this case, only the prior phonetic knowledge about minimum duration of phones is introduced in the posterior estimation. Next step is using non ergodic phone transitions estimated from a labeled data, instead of ergodic transitions. Finally, we can have a fully constrained model composed of connected word models and phone models. The parameters of this model are estimated from the training set. This topology integrates full phonetic and lexical knowledge in the posterior estimation.

Although in this paper we only study phone level posteriors, this posterior estimation/integration approach provides a theoretical framework for hierarchical estimation, integration and use of posteriors, from the state level up to the phone and word levels. Word posteriors can be estimated basically in the same way as state posteriors are integrated into phone posteriors. For more details please refer to [29].

Besides the advantages of integrating prior knowledge for enhancing posterior estimates, it should be noticed how and to what extent the knowledge is reliable. Although the prior knowledge is assumed to be correct, but as the name "prior" suggests, there can be few cases in which the true data is not matching the prior knowledge. For example, the assumed lexical knowledge may not include some rare but truly existing pronunciation variants for a word, while such cases may appear in data. In these cases, the enhanced posteriors start deviating from the MLP posteriors and they may not represent the data correctly. Therefore, although prior knowledge helps to improve the estimation of posteriors, there can be some cases that the resulting posteriors are not matching the data. This means there is a trade off between the smoothness obtained by integrating prior knowledge, and deviation from data. Considering this potential risk, as it is studied in Section 4.1, we propose to use HMM-based enhanced posteriors in combination with the original MLP posteriors. In this way, information in both posterior streams are preserved. A more detailed explanation will be given in Section 4.1.

3.2 MLP-Based Integration of Prior and Contextual Knowledge

In 3.1, we have studied the integration of phonetic and lexical knowledge (encoded in HMM topology) in the posterior estimation. The HMM topology specifies the prior knowledge based on the solid prior

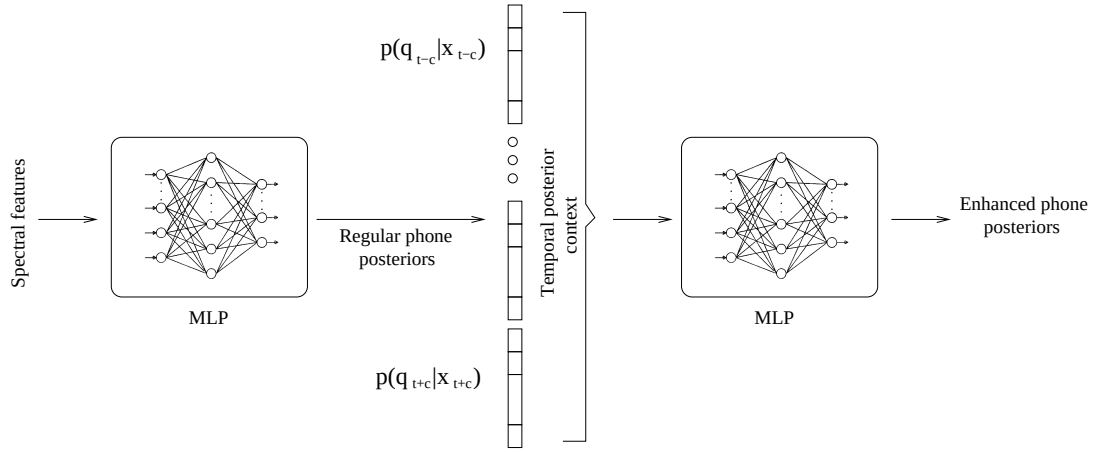

Figure 5: MLP-based enhanced phone posterior estimation: The first MLP is transforming acoustic (cepstral) features to regular phone posteriors. A temporal context of phone posteriors is made by concatenating posterior vectors in $\{p(q_{t-c}|x_{t-c}), ..., p(q_t|x_t), ..., p(q_{t+c}|x_{t+c})\}$. $p(q_{t-c}|x_{t-c})$ is a vector of phone posteriors at time $t-c$. The second MLP processes the temporal context of regular phone posteriors, and learns long term dependencies between phone evidences. These dependencies are prior phonetic knowledge. During the inference (forward pass of the second MLP), the learned knowledge is integrated in the posterior estimation, resulting in enhanced posteriors.

assumptions about phones duration and the lexical usage of phones in the words. The alternative to this solid prior assumptions is learning the prior knowledge from data. In this section, we study a second approach for integrating prior knowledge which realizes the idea of learning priors form data. We use a secondary neural network to learn long term inter and intra dependencies between phone evidences (posteriors) in the training data. The configuration is shown in Figure 5. We have two MLPs in this configuration. The first MLP performs the regular phone posterior probability estimation by transforming a small context of acoustic features (cepstral features) to phone posteriors. The input to the second MLP is a temporal context of phone posteriors estimated by the first MLP, i.e. $\{p(q_{t-c}|x_{t-c}), ..., p(q_t|x_t), ..., p(q_{t+c}|x_{t+c})\}$, where '$c$' shows a temporal context (typically 6-9). To form this input, the posterior vectors in the mentioned temporal context are concatenated. The output of the second MLP is enhanced phone posteriors for the same set of phones as the first MLP. The phonetic class is defined with respect to the center of the temporal context. 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 same database is used for training the two MLPs. The first MLP learns the transformation form acoustic features to phone evidences, while the second MLP gets the phone evidences as input and learns long term dependencies between phone evidences. This long term phone dependencies can be interpreted as prior phonetic information, such as phone trajectory shape, co-articulation between phones, and phone duration information. Therefore, the second MLP learns prior phonetic knowledge from data, and integrates these knowledge in the phone posterior estimation during the inference (forward pass). This leads to enhancement of phone posteriors. The rational behind this is that at the output of every MLP, the information stream gets 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. In the reminder of this paper, we call the posteriors at the output of the second MLP as "MLP-based enhanced posteriors".

We have experimentally analyzed the role of the second neural network in the hierarchy. The mapping function which is learned by the MLP is nonlinear, thus the analysis of second MLP role is not straightforward. A single layer perceptron (SLP) can be a reasonable approximation for investigating

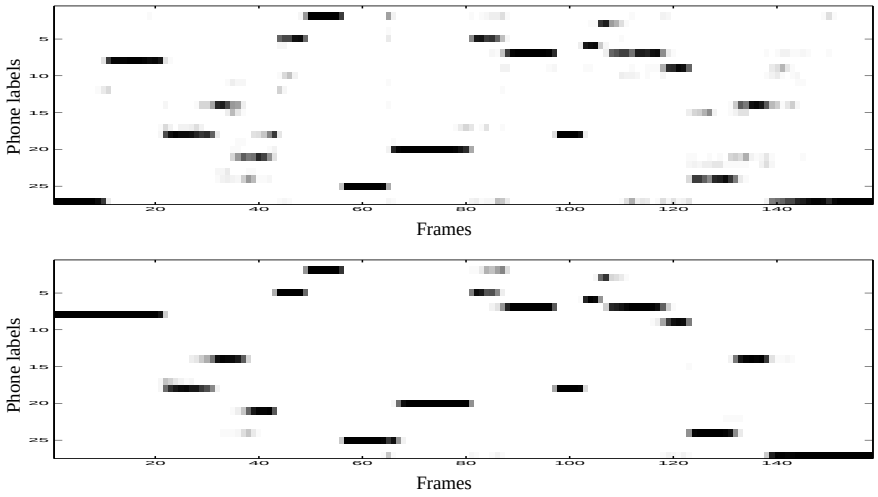

Figure 6: (top) Initial posteriors estimated by the first MLP, and (bottom) enhanced phone posteriors estimated by the second MLP, integrating phonetic prior knowledge. The y-axis is showing phone labels and x-axis is showing frames. The intensity inside each block is showing the posterior value. The new enhanced posteriors look more confident.

the role of the second MLP, and can be considered as a multi-dimensional linear matched filter [30]. Therefore, we replace the second MLP with a SLP, in order to analyze the role of the second neural network in the configuration shown in Fig. 5. The single layer perceptron can be viewed as a multi-dimensional matched filter derived jointly for all the phonemes by minimizing an error criteria. The analysis of the matched filters obtained after training the SLP shows that the matched filter for a specific phoneme (e.g. /iy/) captures the contribution of different regular phone posteriors at the input of SLP to estimate the posterior probability of the phone /iy/. These contributions are consistent with the production of this phoneme. The analysis indicates that the second neural network has learned the long term inter an intra dependencies between the regular posteriors. These dependencies are mostly prior phonetic information such as phone posterior trajectory shape, co-articulation between phones, and phone duration information.

Figure 6 is showing an example of initial and corresponding enhanced posteriors. The enhanced (second MLP) posteriors are more confident than the initial (first MLP) posteriors. The second MLP acts as a filter which smooth out evidences not matching the learned prior phonetic knowledge. Ideally, this approach can be used for post-processing the output of any posterior estimator to integrate higher level knowledge (e.g. prior phonetic knowledge).

In the MLP-based integration of the phonetic and lexical knowledge, the risk of using prior knowledge which is not matching the reality of data is less than HMM-based integration. It is due to the fact that the prior knowledge is learned from the data, instead of being obtained form solid prior assumptions. This leads to some differences in the way we use HMM-based and MLP-based enhanced posteriors for speech recognition systems. It will be studied in more detail in Sections 4 and 5.

In this work, the second MLP has been trained on the same database as the first MLP. An alternative (although not experimented here) will be to use the second MLP 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 the specific task.

4 USING ENHANCED PHONEME POSTERIORS AS FEATURES

As discussed in Section 2.2, posterior probabilities have been used as more discriminant features in speech recognition systems. The most well known sample of these systems is Tandem [5]. In Tandem approach, posterior probabilities are used as features for training and inference in a HMM/GMM back-end module. In this section, the use of the enhanced posteriors as features in Tandem configuration is investigated. We propose new Tandem configurations for HMM-based and MLP-based enhanced phone posteriors. We show that using the enhanced posteriors as features, or as complementary features can improve the performance of Tandem system. Since HMM-based and MLP-based enhanced posteriors have different properties, we study their cases separately.

4.1 HMM-Based Enhanced Posteriors

In Section 3.1, we have studied the integration of prior and contextual knowledge using a HMM. This integration leads to estimating more informative posteriors. We also mentioned to the issue of integrating partially incorrect prior knowledge leading to deviation form the data. Considering this, a safe compromise is using the enhanced posteriors as complementary features along with the original MLP posteriors. In the other words, the enhanced posteriors should be combined with the MLP posteriors. Considering a configuration similar to Tandem, the combined evidences are then used as features for training and inference. In this way, the raw evidences (MLP posteriors) representing the data are preserved, while there is also access to the posteriors enriched by the prior knowledge and context.

Fig. 7 is showing a diagram of the normal Tandem system using MLP posteriors as features, and Tandem system using enhanced posteriors as complementary to the MLP posteriors. The emission probabilities in the HMM module which integrates prior knowledge are provided by the MLP. The enhanced posteriors are obtained by post-processing MLP posteriors in the HMM to integrate prior and contextual knowledge. In our experiments, we have used phone duration information (modeling phones with few number of states) as the prior knowledge. Normal Tandem configuration uses only the MLP estimated posteriors as features, while in our method, the enhanced posteriors are combined with the MLP posteriors. The combined evidence is then used as features for training/inference in the HMM/GMM back-end.

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

$$Comb_t^i = \frac{p(q_t = i|x_t) + p(q_t = i|x_{1:T}, M)}{2} \tag{12}$$

where $Comb_t^i$ shows the combined evidence for phone i at frame t. In case of concatenation rule, the MLP and enhanced posterior vectors at frame t are concatenated. The dimension of the resulting vector is reduced by applying KLT transform. The performance of normal Tandem system, and Tandem system with complementary features will be compared later in this section.

We have used OGI Numbers'95 database [12], and a reduced vocabulary version of the DARPA Conversational Telephone Speech-to-text (CTS) task (1'000 words) [13] for the experiments. For the OGI Numbers'95 database, 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. There are 27 context-independent phones including silence. The acoustic vector x_t is the PLP cepstral coefficients [23] 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. For the estimation of regular MLP phone posteriors, we trained an MLP with 351 input nodes (9 frames of acoustic features), 1200 hidden units and 27 output units

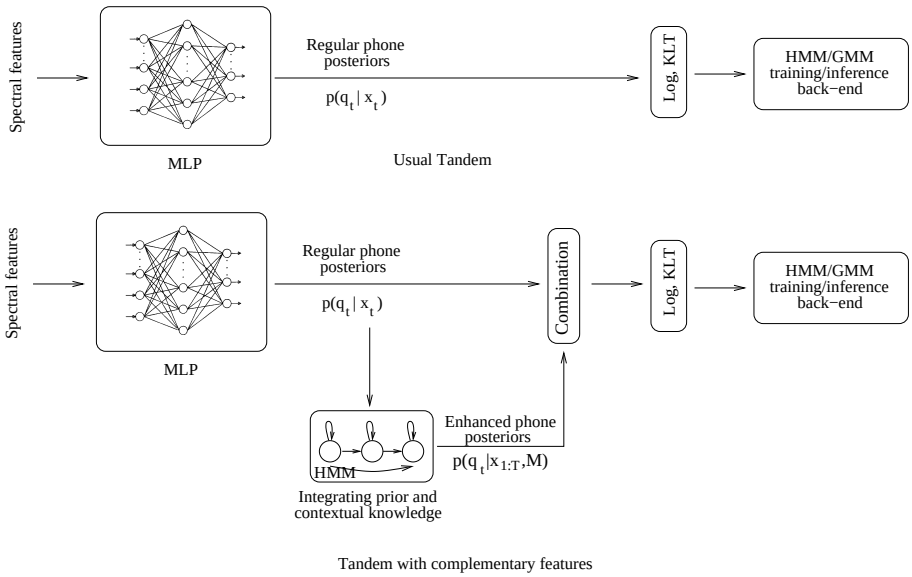

Figure 7: (top) Usual Tandem, and (bottom) Tandem system using enhanced posteriors as complementary features. Usual Tandem uses MLP posteriors (after some transformations) as features. The new Tandem system uses a combination of the MLP and enhanced posteriors as features. In the new Tandem configuration, enhanced posteriors are estimated using a HMM module integrating phone duration information. The enhanced posteriors are then combined with the MLP posteriors, some transformations applied, and the resulting features are used for training and inference in a HMM/GMM back-end.

corresponding to the 27 context-independent phones. After training, the phone posteriors for the training set and test set were estimated and scaled by their respective priors (estimated from the training segmentation) to obtain scaled-likelihoods.

The idea was also evaluated on conversational telephone speech (CTS) task. There are 1000 words and 46 phones in this task. The training set contains 16 hours of male CTS speech randomly selected from the Fisher Corpus and the Switchboard Corpus. The tuning set consists of 1.2 hours of data and the test set consists of 1.3 hours of data. The acoustic features are 13 PLP coefficients concatenated with their first two derivatives. It was computed with vocal tract normalization (VTLN) [31], and mean and variance normalization. For the estimation of regular posteriors, an 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. After training, the phone posteriors for the training set and the test set were estimated.

For both databases, the MLP posteriors obtained were then used to estimate the enhanced phone posteriors as explained in Section 3.1. The prior knowledge used to obtain enhanced posteriors is the phonetic duration knowledge. This was achieved by considering 3 states per phone model in the HMM module integrating prior knowledge.

We first start with the comparison of the enhanced and MLP posteriors at the frame level. Table 1 is showing the fame recognition results for the enhanced and regular MLP posteriors (for the two databases). For both databases, the enhanced posteriors show lower frame error rates than the MLP posteriors. In addition, we also study the entropy for each type of posteriors. The entropy can provide a measure of consistency/confusion in the posteriors. The entropy of phone posteriors is measured at

each frame, and averaged over the whole database:

$$E_t = -\sum_i p(q_t^i|x_{1:T}, M) \log_2 p(q_t^i|x_{1:T}, M) \quad (13)$$

$$AvE = \frac{\sum_{t=1}^{\mathcal{T}} E_t}{\mathcal{T}} \quad (14)$$

where E_t is the entropy of posteriors at frame t, and $\mathcal{T}$ is the total number of frames in the database. The obtained average entropy values AvE for the enhanced and MLP posteriors are shown in Table 2. Lower entropy of the enhanced posteriors shows that they have more consistency than regular MLP posteriors.

Database	MLP posteriors	Enhanced posteriors
CTS	35.2%	**33.3%**
Numbers	17.6%	**16.2%**

Table 1: Frame error rates (FER) on Numbers'95 and CTS tasks, for regular MLP posteriors and HMM-based enhanced phone posteriors. Enhanced posteriors have lower FER than the regular MLP posteriors. Frame error rates are obtained on cross-validation partition of the databases.

Database	MLP posteriors	Enhanced posteriors
CTS	1.64	**0.33**
Numbers	0.67	**0.18**

Table 2: Average entropy of enhanced and regular MLP posteriors for different databases. The measures are obtained by computing the entropy of posteriors at each frame, and averaging over the whole database. Enhanced posteriors have lower average entropy indicating higher consistency than the regular posteriors.

After the frame level studies, we investigate the performance of enhanced posteriors for word recognition. As discussed before, for word recognition studies in Tandem configuration, the enhanced phone posteriors at each frame t are combined with the original MLP posteriors[3]. Two combination rules which are summation (average) and concatenation have been tried. The resulting combined evidences are processed by Log and KLT transforms, as done for normal Tandem feature extraction. For comparison purpose, we have also extracted the regular Tandem features by performing Log and KLT transforms on the regular MLP posteriors at each frame.

For each type of features (regular Tandem and combined evidence), we trained a HMM/GMM system using HTK toolkit [32]. In case of Numbers database, 80 context-dependent phone models with 12 mixtures per state, and 3 states per phone is used. In case of CTS database, models were trained through 40 iterations: 5 iterations for the context-independent models, 5 iterations for the context-dependent models, 5 iterations for the clustered context-dependent models, and then 5 iteration each for incrementing mixtures from 1 to 32 (2, 4, 8, 16, 32). During the recognition, a bi-gram language model is used.

We compare the results of recognition studies for the normal Tandem, which uses only MLP posteriors as features, and the Tandem system which uses combined evidence (MLP and enhanced posteriors) as features. Table 3 is showing the results in terms of word error rate for the two databases, and different combination rules. As illustrated in the table, the combined evidences obtained from the two streams of posteriors consistently perform better than the MLP posterior features alone. Using enhanced posteriors (encoding prior and contextual knowledge) in combination with MLP posteriors has helped to provide better evidences for Tandem.

[3]In practice, using HMM-based enhanced posteriors alone in the Tandem configuration did not improve word recognition performance.

Database	MLP posteriors	**MLP + Enh**	**MLP & Enh**
CTS	44.2%	**43.8%**	**41.3%**
Numbers	4.7%	**4.3%**	**4.3%**

Table 3: Word error rates (WER) on Numbers and CTS tasks, for MLP posteriors, and MLP posteriors combined with the enhanced posteriors, using addition (MLP+Enh) and concatenation (MLP & Enh) as combination rules. The combined evidences perform better than regular MLP posteriors in Tandem configuration.

4.2 MLP-Based Enhanced Posteriors

The enhanced posteriors obtained by a secondary MLP can be also used as features in Tandem configuration. In this case, unlike HMM-based enhanced posteriors, the integrated prior knowledge is learned from the data. Therefore, there is less risk of being biased by partially wrong prior assumptions. This allows using the enhanced posteriors as features directly (without the need for combination with the regular posteriors). In this way, the configuration for using the MLP-based enhanced posteriors would be similar to the normal Tandem configuration. The only difference is that the regular phone posteriors are replaced with the enhanced phone posteriors. We compare the performance of regular and enhanced posteriors as features in the Tandem configuration. The databases, specifications of spectral features extraction, and regular MLP posterior estimation is the same as the case of HMM-based posterior experiments (see Section 4.1).

In order to enhance phone posterior estimates for the Numbers database, a second MLP for post-processing 19 frames of regular posteriors is used (as explained in Section 3.2). It has 513 (19x27) input nodes, 1000 hidden nodes and 27 output nodes. For enhancing phone posteriors in the CTS database, a second MLP with 690 (15x46) input nodes, 2000 hidden nodes and 46 output nodes is used to post-process 15 frames of regular posteriors. The size of the temporal posterior context, and the structure of the second MLP is obtained empirically for all the experiments.

As before, we start with frame level performance study of enhanced posteriors. Table 4 is showing frame error rates of the regular and enhanced posteriors, for Numbers and CTS databases (cross validation portion). Again, lower error rates can be observed for the enhanced posteriors in both databases.

The same as Section 4.1, we do entropy studies on the MLP-based enhanced posteriors. Table 5 shows the average entropies for the enhanced and regular posteriors. Enhanced posteriors have less entropy than the regular posteriors. This indicates that there is more consistency in the enhanced posteriors, as compared to the regular posteriors.

Database	Regular posteriors	Enhanced posteriors
CTS	35.2	**31.5%**
Numbers	17.6	**15.4%**

Table 4: Frame error rates (FER) on Numbers'95 and CTS tasks, for regular (first MLP) and enhanced (second MLP) phone posteriors. Enhanced posteriors have lower FER than the regular posteriors. Frame error rates are obtained on cross-validation partition of the databases.

In the word recognition studies, we compare the performance of regular and enhanced posteriors as features in the Tandem configuration. Unlike the case of HMM-based posteriors, MLP-based enhanced posteriors can be used directly as features, without being necessarily combined with regular posteriors. As the usual case of Tandem approach, the enhanced and regular posteriors are gaussianized and decorrelated using Log and KLT transforms. The result of the transformation is used as features for training and inference in a HMM/GMM back-end. Details of implementation for the HMM/GMM

Database	Regular posteriors	Enhanced posteriors
CTS	1.64	**1.29**
Numbers	0.67	**0.40**

Table 5: Average entropy of enhanced and regular phone posteriors for different databases. The measures are obtained by computing the entropy of posteriors at each frame, and taking average over the whole database. Enhanced posteriors have lower entropy indicating higher consistency than the regular posteriors.

back-end is the same as Section 4.1.

Table 6 is showing the word recognition performances for regular and enhanced posteriors. It can be observed that the enhanced posteriors are consistently performing better than the regular posteriors for the two databases.

Database	Regular posteriors	Enhanced posteriors
CTS	44.2%	**42.5%**
Numbers	4.7%	**4.3%**

Table 6: Word error rates (WER) on Numbers'95 and CTS tasks, for regular and enhanced phone posteriors. Enhanced posteriors are obtained by post-processing regular posteriors using a secondary MLP. The phone posteriors are used in Tandem configuration for the recognition. Enhanced phone posteriors perform consistently better than the regular posteriors for the two databases.

In addition to the use of MLP-based enhanced posteriors as a replacement for the regular MLP posteriors, we have investigated their usage as complementary features to the regular MLP posteriors (as done for HMM-based enhanced posteriors). The configuration for using the combined evidences is the same as shown in Figure 7, except that the HMM-based enhanced posteriors are replaced with the MLP-based enhanced posteriors. The same addition and concatenation rules have been tried. Table 7 is showing the word recognition results when the MLP-based enhanced posteriors are used as complementary features. As illustrated in Table 7, usage of the MLP-based enhanced posteriors as complementary features improves the performance even more than using them instead of regular posteriors. Therefore, they perform best when they are used in combination with the regular MLP posteriors.

Database	MLP posteriors	**MLP + Enh**	**MLP & Enh**
CTS	44.2%	**41.2%**	**42.3%**
Numbers	4.7%	**4.2%**	**4.2%**

Table 7: Word error rates (WER) on Numbers and CTS tasks, for MLP posteriors, and MLP posteriors combined with enhanced posteriors using addition (MLP+Enh) and concatenation (MLP & Enh) as combination rules. Enhanced posteriors are obtained at the output of the second MLP. Combined evidences perform better than regular MLP posteriors.

We also further studied 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. Phoneme posteriors have simpler and possibly more 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. In our study, initially we tried to reduce the complexity of the second MLP in terms of the number of hidden nodes. The optimum complexity is obtained empirically. Reducing the complexity below this optimum slightly degrades the performance of enhanced posteriors, however they still perform better than the regular posteriors. The degradation in the performance of enhanced posteriors is very small even for large decrease in the complexity of the second MLP. In addition,

we have studied using a Single Layer Perceptron (SLP) as the second ANN. Although the obtained enhanced posteriors are performing better than the regular posteriors, their performance is slightly lower than the case of using MLP as the second ANN. It implies that there is still nonlinearly separable patterns at the output of the first MLP (regular posteriors) which can not be learned by the SLP.

Building upon the same idea of ANN hierarchy, a third MLP has also been tried in order to post-process the output of the second MLP. Using a third MLP, the frame error rate and entropy results are improved, but no considerable improvement in phone and word recognition is observed.

5 USING ENHANCED PHONEME POSTERIORS AS LOCAL SCORES

Another conventional usage of posteriors in ASR is as local scores for decoding (e.g. hybrid HMM/ANN method). In this section, we investigate the use of the enhanced posteriors as scores for decoding, and we compare them with the regular MLP posteriors. Since HMM-based and MLP-based enhanced posteriors have different properties, we study them separately.

5.1 HMM-Based Enhanced Posteriors

HMM-based enhanced posteriors can be used as local scores for decoding, in the same way as regular posteriors are used in HMM/ANN configuration. Unlike the case of using HMM-based enhanced posteriors as features, there are few issues regarding the use of these posteriors as local scores for decoding. The main issue is the fact that the knowledge which is integrated in the enhancement process is the same as the knowledge which is taken into account in the topological constraints of the decoder. For instance, the same duration knowledge as integrated in the enhancement process, is taken into account in the hybrid decoder configuration. This means that we should not expect performance improvement when the HMM-based enhanced posteriors are used for decoding, since no additional knowledge is integrated in the enhancement process. The experiments also confirm that the performance of the enhanced and regular posteriors for decoding are the same. However, there is a side advantage in using enhanced posteriors for decoding.

The advantage is revealed when we compare the sensitivity to ad-hoc tuning factors (e.g. phone deletion penalty) for the decoder using the enhanced posteriors, and the decoder using regular posteriors [33]. Phoneme deletion penalty is a tuning factor and an engineering trick which is used for numerical compensation of scores for different paths during decoding [32]. It can significantly affect the recognition performance of standard HMM/ANN and HMM/GMM systems[4]. We have setup some experiments to investigate this issue.

We have used OGI Numbers'95 database for the experiments. Specifications of the database, spectral features and regular MLP posteriors estimation is the same as mentioned in Section 4.1. We have used a fully constrained model (as explained in Section 3.1) to get estimates of enhanced posteriors. This means that we integrate full lexical and phonetic knowledge in the posterior estimation. The obtained enhanced posteriors are then used as local scores for decoding. We have used NOWAY [34] as the hybrid decoder. For comparison, regular phone posteriors are also used in the same decoder. In order to compare the sensitivity of the systems (one using regular posteriors, and the other one using enhanced posteriors), we vary the phone deletion penalty value in the decoder and observe the change of performance for the two systems. Figure 8 shows the results. Comparing the two curves, we can conclude that the decoder using enhanced posteriors is much less sensitive to tuning than the one using regular posteriors (standard hybrid HMM/MLP system). HMM-based enhanced posteriors

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

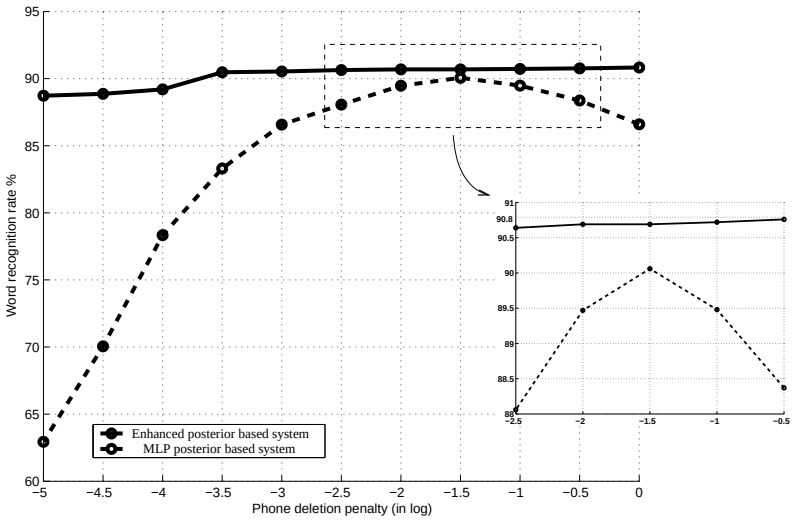

Figure 8: Comparing the sensitivity to tuning phone deletion penalty, for the decoder using enhanced posteriors and the one using MLP posteriors. Phoneme deletion penalty is varied for the two decoders and the performances are observed (on OGI Numbers'95 database). The inside diagram is a zoom of performance curves for small values of phone deletion penalty (fine tuning). The decoder using enhanced posteriors is much less sensitive to tuning ad-hoc parameters than the one using regular MLP posteriors.

tend to have very close to binary values (similar to a decision), because they are estimated by integrating some extra knowledge, while the MLP posteriors can change more smoothly between 0 and 1. Therefore, the accumulated scores obtained by enhanced posteriors during decoding tend to be discrete, while it is continuous for the case of regular MLP posteriors. The tuning operation which slightly changes the scores, affects the decision made based on continuous scores more than the one made based on discrete scores. This means that the decoder using enhanced posteriors is much less sensitive to tuning ad-hoc parameters.

5.2 MLP-Based Enhanced Posteriors

The MLP-based enhanced posteriors can also be used in the same way as regular posteriors for decoding. In this case, they are used as local scores instead of the regular posteriors in the hybrid HMM/MLP configuration. We compare the performance of regular and enhanced posteriors for decoding. The comparison is done for the OGI Numbers and CTS databases. The specifications of databases, regular MLP posteriors, and enhanced posterior estimation are the same as mentioned in Section 4.2. We have used JUICER [35] as the hybrid decoder. In case of Numbers database, phones are modeled with 5 states in the decoder. In case of CTS database, phones are modeled with 5 states, and a bi-gram language model is used. Table 8 is showing the word recognition performances for regular and enhanced posteriors. It can be observed that the enhanced posteriors are performing significantly better than the regular posteriors for the two databases.

We have also done phone recognition experiments to compare the enhanced and regular posteriors for phone recognition in a hybrid decoder. For the experiments, TIMIT database [14] is used. The training data set 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 phones. The acoustic features are PLP, delta and double delta features. For estimating regular posteriors, we have used an MLP with 351 input nodes (9 frames of PLPs), 1000 hidden nodes and 39 (corresponding to the number of phones) output nodes.

Database	Regular posteriors	Enhanced posteriors
CTS	53.6%	**49.2%**
Numbers	9.9%	**8.8%**

Table 8: Word error rates (WER) on Numbers'95 and CTS tasks, for regular and enhanced phone posteriors. The phone posteriors are used in hybrid HMM/MLP configuration for decoding. Enhanced posteriors perform significantly better than the regular posteriors.

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

Table 9: Frame error rates (FER) and phone error rates (PER) for regular and enhanced phone posteriors, on TIMIT database. Lower FER and PER can be observed for enhanced posteriors as compared to the regular posteriors.

In order to estimate enhanced posteriors, 19 frames temporal contexts of the regular posteriors are post-processed by a secondary MLP (as explained in Section 3.2). This MLP has 741 (39x19) input nodes, 1000 hidden nodes and 39 output nodes (corresponding to the number of phones). For the phone recognition, we have used NOWAY [34] which is a hybrid HMM/ANN decoder. In this decoder, each phone is modeled with 5 states, and a bi-gram phone level language model is used. Frame and phone recognition results are shown in Table 9. The enhanced posteriors perform significantly better than the regular posteriors for frame and phone recognition.

6 OTHER USAGES OF THE ENHANCED POSTERIORS

In this paper, the most conventional usages of the posteriors, i.e. as features for Tandem, and as local scores for decoding was investigated for the case of enhanced posteriors. However, the usage of the enhanced posteriors is not limited to these cases. In this section, we briefly study some other related works:

6.1 Higher Level Posteriors

In [29], we have shown how the HMM-based posterior estimation approach can be used to estimate higher level (e.g. word) posteriors at every frame. Basically, the same as phone posteriors, word posteriors can be obtained by integrating posteriors of states belonging to a word in the HMM model:

$$\begin{aligned} p(w_t^i|x_{1:T}, M) &= \sum_{k=1}^{N_s} p(w_t^i, s_t^k|x_{1:T}, M) \\ &= \sum_{k=1}^{N_s} p(w_t^i|s_t^k, x_{1:T}, M)p(s_t^k|x_{1:T}, M) \\ &= \sum_{k=1}^{N_s} p(w_t^i|s_t^k, x_{1:T}, M)\gamma(k,t) \end{aligned} \tag{15}$$

where w_t is a word at time t and w_t^i represents the event "$w_t = i$". $p(w_t^i|s_t^k, x_{1:T}, M)$ represents the probability of being in a given word i at time t knowing to be in the state k at time t. Assuming that there is no parameter sharing between words, it is deterministic and equal to 1 or 0.

In this way, a word posterior at each frame $p(w_t^i|x_{1:T}, M)$ encoding phonetic and lexical knowledge can be obtained. In [29], we have shown the application of these frames based word posteriors in keyword spotting. At each frame t, we estimate a posterior for the keyword and a posterior for the garbage unit. Comparing the two posteriors at each frame, we can have frame based decisions on detecting the keyword. Counting the number of these decisions provides a score which is related to the detected length of the keyword. This score is compared with a length based threshold to enable the final decision on detecting the keyword. The main advantage of this techniques is simple relation between thresholds and keyword characteristics such as length. This allows to predetermine thresholds for new keywords (with no corresponding development set), which can be important in practical keyword spotting systems. In the conventional keyword spotting approaches, the threshold is an ad-hoc parameter which is not related to characteristics of the keyword in a simple way, and should be tuned using a relatively huge development set. Therefore, it is not simple to predetermine the threshold for a new keyword.

6.2 Out-Of-Vocabulary Word Detection

Another application of the enhanced posteriors is revealed when we measure the divergence between the regular and enhanced posteriors [36]. The MLP posteriors $p(q_t|x_t)$ can be interpreted as sensory information representing data as a sequence of phone evidences. On the other hand, the HMM-based enhanced posteriors $p(q_t|x_{1:T}, M)$, can be interpreted as the MLP phone posteriors enriched by the M (phonetic and lexical knowledge), and context $x_{1:T}$. Therefore, the difference between the two posteriors can indicate cases that data (represented by MLP posteriors) does not match the assumed prior phonetic and lexical knowledge. Since the two posteriors are estimated for every frame, we can have a frame level measure of deviation, thus a frame level measure of match/mismatch (consistency) between data and phonetic/lexical knowledge. The deviation can be measured using Kullback-Leibler divergence:

$$\begin{aligned} KL(\overline{S_t}, \overline{C_t}) &= \sum_i S_t^i log_2 \frac{S_t^i}{C_t^i} \\ &= \sum_i p(q_t^i|x_t) log_2 \frac{p(q_t^i|x_t)}{p(q_t^i|x_{1:T}, M)} \end{aligned} \tag{16}$$

Where $\overline{S_t}$ is the regular MLP posterior vector at frame t, and $\overline{C_t}$ is the enhanced posterior vector at frame t. S_t^i and C_t^i show the i^{th} element of the posterior vectors at frame t.

One of the applications of measuring this inconsistency is detecting Out-Of-Vocabulary (OOV) words [36] in posterior based ASR. In case of an OOV word, the lexical knowledge does not match an existing sample of data, resulting in large values of divergence between the two posteriors. In general, the difference between the two posteriors can be used to detect any inconsistency in data or model.

7 SUMMARY AND CONCLUSION

In this paper, we first briefly discussed current approaches for estimating phone posteriors, and using them as local scores or as features in ASR systems. Indeed, several approaches in that direction have been shown to have a potential for improving state-of-the-art ASR systems. However, we also believe that further progress in that direction will critically depends on improving the quality of these posterior estimates.

Considering this fact, we proposed and discussed two approaches for enhancing phone posterior estimates, by integrating context and prior (phonetic and lexical) knowledge. The first approach uses an HMM module to integrate this additional knowledge. The prior knowledge is encoded in the topology

of the HMM. The regular MLP posteriors are used in HMM forward-backward recursions to integrate context and prior knowledge, yielding enhanced phone posterior estimates. In the second approach, a secondary MLP is used to post-process a temporal context of regular MLP posteriors, and learn long term dependencies between these posteriors. These long term dependencies are prior phonetic knowledge. During the inference (forward pass of the second MLP), the learned prior knowledge is integrated in the phone posterior estimation, resulting in enhanced phone posteriors at the output of the second MLP.

We have compared these enhanced posteriors with the regular MLP posteriors. The entropy studies indicate that there is more consistency in the enhanced posteriors. Frame recognition studies show consistently lower error rates for the enhanced posteriors. In the word recognition studies, again we have observed that the enhanced posteriors perform consistently better than the regular posteriors as complementary features in Tandem configuration, as well as local scores in hybrid HMM/MLP configuration. The HMM-based enhanced posteriors should be used in combination with the regular posteriors for improving the performance, while the MLP-based enhanced posteriors can be used as a replacement to regular posteriors.

We believe that the present paper introduced a principled general framework for enhancing posterior estimates in ASR systems. Based on this work, we can estimate a more informative phone (or even higher level) posterior at every frame. Some of the advantages and applications of the new posteriors were investigated. Many ASR algorithms get phone evidences at the frame level as input. The new enhanced posteriors can thus be widely general purpose since they are estimated for *phones* at every *frame*. One can think of other applications of regular MLP posteriors in ASR, and use the new enhanced posteriors instead or in combination with the regular posteriors.

8 ACKNOWLEDGEMENTS

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, Petr Fousek, Joel Pinto, John dines, and Mathew Magimai Doss (from IDIAP), Steve Renals (University of Edinburgh), and Nelson Morgan (ICSI) for helpful discussions.

References

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

[2] Mangu, L., Brill, E., and Stolcke, A., "Finding consensus in speech recognition: word error minimization and other applications of confusion networks", *Computer, Speech and Language*, Vol. 14, pp. 373-400, 2000.

[3] Abdou, S. and Scordilis, M.S., "Beam search pruning in speech recognition using a posterior-based confidence measure", *Speech Communication*, Vol. 42, pp. 409-428, 2004.

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

[5] 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.

[6] Bilmes, J., "Maximal mutual information based reduction strategies for cross-correlation based joint distribution modelling", *Proc. ICASSP98*, SP14.6, Seattle, 1998.

[7] Yang, H. H., Sharma, S., van Vuuren, and Hermansky, H., "Relevance of Time Frequency Features for Phonetic and Spectral/Channel classification", *Speech Communiations*, Aug. 2000.

[8] Wessel, F., Schlter, R., Macherey, K., Ney, H., "Confidence Measures for Large Vocabulary Continuous Speech Recognition", *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 3, pp. 288-298, March 2001.

[9] Wessel, F. et al., "Using Word Probabilities as Confidence Measures", *ICASSP'98*, Vol. 1., pp 225-228, May 1998.

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

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

[12] 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.

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

[14] 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.

[15] Povey, D., Saon, G., Mangu, L., Kingsbury, B., and Zweig G., "EARS progress update: improved MPE, inline lattice rescoring, fast decoding, Gaussianization and Fisher experiments", *EARS STT Workshop*, St. Thomas, US Virgin Islands, Dec. 2003.

[16] Abdel-Haleem, Y., "Conditional Random Fields for Continuous Speech Recognition", *PhD Thesis*, University of Sheffield, November 2006.

[17] Povey, D. and Woodland P.C., "Minimum Phone Error and I-Smoothing for Improved Discriminative Training", *Proc. ICASSP'02*, Orlando.

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

[19] 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.

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

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

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

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

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

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

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

[27] 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.

[28] 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.

[29] Ketabdar, H., Vepa, J., Bengio, S., and Bourlard, H., "Posterior Based Keyword Spotting with A Priori Thresholds", *Proc. Interspeech'06*, Pittsburgh, USA, 2006.

[30] Lehtonen, M., Fousek, P., Hermansky, H., "Hierarchical Approach For Spotting Keywords", *IDIAP Research Report*, np. 05-41, 2005.

[31] Zhan, P., Waibel, A., "Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition", *Language Technologies Institute Technical Report: CMU-LTI-97-150, Carnegie Mellon University*, Pittsburgh, USA.

[32] 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.

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

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

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

[36] Ketabdar, H., Hannemann, M., Hermansky, H., "Detection of Out-of-Vocabulary Words in Posterior Based ASR", *Proc. Interspeech'07*, Belgium, 2007.