Enhanced Phone Posteriors for Improving Speech Recognition Systems

Hamed Ketabdar and Hervé Bourlard, Fellow, IEEE

Abstract—Using phone posterior probabilities has been increasingly explored for improving automatic speech recognition (ASR) systems. In this paper, we propose two approaches for hierarchically enhancing these phone posteriors, by integrating long acoustic context, as well as phonetic and lexical knowledge.

In the first approach, phone posteriors estimated with a Multi-Layer Perceptron (MLP), are used as emission probabilities in HMM forward-backward recursions. This yields new enhanced posterior estimates integrating HMM topological constraints (encoding specific phonetic and lexical knowledge), and context. In the second approach, temporal contexts of the regular MLP posteriors are post-processed by a secondary MLP, in order to learn inter and intra dependencies between the phone posteriors. These dependencies are phonetic knowledge. The learned knowledge is integrated in the posterior estimation during the inference (forward pass) of the second MLP, resulting in enhanced phone posteriors.

We investigate the use of the enhanced posteriors in hybrid HMM/ANN and Tandem configurations. We propose using the enhanced posteriors as replacement, or as complementary evidences to the regular MLP posteriors. The proposed methods have been tested on different small and large vocabulary databases, always resulting in consistent improvements in frame, phone and word recognition rates.

Index Terms—Regular MLP posteriors, Integrating context and phonetic/lexical knowledge, Enhanced posteriors, Tandem, Hybrid HMM/ANN.

I. 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 methods allow for discriminant training, as well as for the possibility of using small acoustic context by presenting few frames at MLP input. Regarding the use of posteriors as features, one successful approach is Tandem [2]. 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 features 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, phone information is not limited to acoustic evidence in a few frames of spectral feature. Information about phones are extended over a long temporal context of at least 200-300ms [3], [4], and there are generally no distinct boundaries between phones. Phonemes have specific duration constraints (phonetic knowledge), follow specific sub-lexical and lexical rules (lexical knowledge), etc. These long contextual and phonetic/lexical 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 [5], [6], [7]. 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 [8], 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 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 at every frame. Many 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 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) [8]. 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 phonetic knowledge. The learned 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 phonetic/lexical knowledge for improving phone posterior estimation in ASR. The framework can be also extended to the estimation of higher level (e.g., word) local 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 Conversational Telephone Speech (CTS) [9] and Numbers’95 [10] and 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 [11]), 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.

II. 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. It estimates $p(q^i_t | x_t)$, where $x_t$ is a spectral feature frame at time $t$, and $q^i_t$ 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 $p(q^i_{t+c} | x_{t-c}^{t+c})$, where $c$ is typically equal to 4. $x_{t-c}^{t+c}$ represents a short temporal context obtained by concatenating acoustic feature vectors in $\{x_{t-c}, \ldots, x_t, \ldots, x_{t+c}\}$. This is in fact very limited context\(^1\).

Posterior probabilities have also been used as local measures for different ASR purposes, such as (1) estimating confidence measures [7], [15], [16], [17], (2) beam search pruning [18], or (3) word lattice rescoring [19].

B. 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 such as few frames of PLP based spectral features [20]. In some cases, Tandem inputs are concatenated outputs from other sub-band classifiers (e.g. TRAPs) [21], [22]. 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. As illustrated in Fig. 1, the MLP phone posterior estimates $p(q^i_t | x_t)$ are gaussianized by a static nonlinearity (usually logarithm) and whitened by the Karhunen-Loeve transform (KLT) derived from training data. $p(q^i_t | x_t)$ is a vector of phone posterior probabilities at time $t$ with the components $p(q^i_t | x_t)$ for $i \in \{1, \ldots, i, \ldots, N_q\}$. Such gaussianized and whitened posterior probabilities form the feature vector for the subsequent HMM/GMM training/inference

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

III. ENHANCING POSTERIOR PROBABILITY ESTIMATION

In the previous section, we have studied the estimation and use 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 in a temporal context of at least 200-300ms, and there are generally no sharp boundaries between phones [3, 4], therefore taking into account long contextual information can be useful. Moreover, some prior linguistic 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, phonetic and lexical knowledge) can be integrated in the posterior estimation to improve the posterior estimates. We propose two different approaches for integrating these higher level knowledge in the posterior 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.

A. HMM-based Integration of Prior and Contextual Knowledge

Topological constraints in a HMM encode specific prior phonetic and lexical knowledge. For instance, modeling phones with a minimum number of states imposes the knowledge about duration of phones, or left-to-right connection of phone models imposes specific 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 \( p(q_t|x_t) \) obtained by MLP, to a more informative posterior \( p(q_t^M|x_{1:T}, M) \), where \( q_t^M \) is the event of having phone \( i \) at time \( t \), \( x_{1:T} = \{x_1, \ldots, x_t, \ldots, x_T\} \) is the acoustic context as available possibly in the whole utterance, and \( M \) is the 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 \( p(q_t^M|x_t) \) as state emission probabilities in the HMM. Each state \( s_k \) of the set of HMM states \( S = \{s_1, \ldots, s_k, \ldots, s_{N_s}\} \) \((N_s \text{ total number of HMM states})\) is associated with one of MLP outputs representing a phone posterior probability. The state emission probabilities are used in HMM forward-backward recursions [8] to integrate HMM topological constraints (encoding specific prior knowledge). This gives the estimates of HMM state posteriors \( p(s_k|x_{1:T}, M) \), where \( s_k \) is the event of having state \( k \) at time \( t \). The state posteriors will then be integrated to enhanced phone posteriors \( p(q_t^M|x_{1:T}, M) \) 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 posterior (and consequently phone posterior) 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 \( t \), \( s_k^t \), given the whole observation sequence \( x_{1:T} \) and the HMM model \( M \) encoding specific prior knowledge (topological/temporal constraints):

\[
\gamma(k, t) = p(s_k^t|x_{1:T}, M) \tag{1}
\]

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 \( s_k^t \) 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 refer to \( \gamma(k, t) \) as “state posterior” in this paper.

The state posteriors \( \gamma(i, t) \) can be estimated using forward \( \alpha \) and backward \( \beta \) recursions [8] using local emission likelihoods.
As shown in Eq. (5), we can express the scaled $\alpha$ recursion as follows:

$$\alpha^{scale}(k, t) = \frac{p(x_t|s_t^k)}{\prod_{\tau=t+1}^T p(x_\tau)} \sum_j \frac{p(s_{\tau+1}^j|s_t^k)p(x_{\tau+1}, s_{\tau+1}^j)}{p(s_{\tau+1}^j)} \beta^{scale}(j, \tau)$$

Similarly, we can define the “scaled” $\beta$ and $\beta$ recursions as follows:

$$\beta^{scale}(k, t) = \frac{p(x_t|s_t^k)}{\prod_{\tau=t+1}^T p(x_\tau)} \sum_j \frac{p(s_{\tau+1}^j|s_t^k)p(x_{\tau+1}, s_{\tau+1}^j)}{p(s_{\tau+1}^j)} \beta^{scale}(j, \tau)$$

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:

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

These scaled likelihoods may 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 $\alpha$ as:

$$\alpha^{scale}(k, t) = \frac{p(x_t|s_t^k)}{\prod_{\tau=t+1}^T p(x_\tau)}$$

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.
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 $p(q^e_i|x_{1:T}, M)$ can be expressed in terms of state posteriors $\gamma(k, t)$ as follows:

$$
p(q^e_i|x_{1:T}, M) = \sum_{k=1}^{N_s} p(q^e_i | s^k_i, x_{1:T}, M) \gamma(k, t)
$$

where $p(q^e_i|x_{1:T}, M)$ is the enhanced phone posterior for phone $i$ at time $t$. Probability $p(q^e_i | s^k_i, x_{1:T}, M)$ 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 $p(q^e_i|x_{1:T}, M)$ integrating context and prior knowledge is obtained. In the reminder of the paper, we call them as “HMM-based enhanced posteriors”.

Figure 2 is showing the configuration for the HMM-based integration of prior and contextual knowledge. As it is shown, the regular phone posterior vectors $p(q_i|x_t)$ are initially estimated using an MLP. $p(q_i|x_t)$ is a vector of phone posteriors at time $t$ with the components $p(q^e_i|x_t)$ for $i \in \{1, \ldots, N_p\}$. 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 $p(q^e_i|x_{1:T}, M)$, $p(q^e_i|x_{1:T}, M)$ 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.

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 the 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 general 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 [24].

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. This means there is a trade off between the smoothness obtained by integrating prior knowledge, and deviation from data. Considering this issue, as it is studied in Section IV-A, 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 IV-A.

B. MLP-Based Integration of Phonetic and Contextual Knowledge

In III-A, we have studied the integration of phonetic and lexical knowledge (encoded in HMM topology) in the posterior estimation. The HMM topology specifies the knowledge based on the solid prior assumptions about phones duration and the lexical use of phones in the words. The alternative to this solid prior assumptions is learning the knowledge from data. In this section, we study a second approach for integrating phonetic knowledge which realizes the idea of learning phonetic knowledge from 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 3. 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. \{$p(q_{t-c}|x_{t-c}), \ldots, p(q_t|x_t), \ldots, p(q_{t+c}|x_{t+c})$\}, where ‘$c$’ shows a temporal context (typically 6-9). To form this input, the posterior vectors in the mentioned 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 phonetic information, such as phone trajectory shape, co-articulation between phones, and phone duration information. Therefore, the second MLP learns phonetic knowledge from data, and integrates these knowledge in the phone posterior estimation during the inference (forward pass). This leads to enhancement of phone posterior. 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 the role of the second MLP, and can be considered as a multi-dimensional linear matched filter [25]. 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. 3. The single layer perceptron can be viewed as a multi-dimensional matched filter derived jointly for all the components of phone posteriors. The rational behind this is that at the output 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.

In the MLP-based integration of phonetic and lexical knowledge, the risk of using the knowledge which is not matching the reality of data is less than HMM-based integration. It is due to the fact that the 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 IV and V.

IV. USING ENHANCED PHONE POSTERIORS AS FEATURES

As discussed in Section II-B, posterior probabilities have been used as more discriminant features in speech recognition systems. The most well known sample of these systems is Tandem [2]. 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.

A. HMM-Based Enhanced Posteriors

In Section III-A, 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 a HMM/GMM back-end. 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. We have studied addition (average) and concatenation as the combination rules. In case of addition (average), the combined evidence is written as:

$$\text{Combi} = \frac{p(q_i = i|x_t) + p(q_i = i|x_{1:T}, M)}{2} \quad (12)$$

where $\text{Combi}$ 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.

Fig. 4 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 the following experiments, the knowledge integrated by the HMM is minimum phone duration information. The topology of this HMM consists of phones modeled with 3 states, imposing a minimum phone duration of 3 frames. The phone models are connected with uniform transitions. A minimum phone duration equal to 3 frames is a usual assumption in many traditional speech recognition systems, and also matches the phone duration statistics we obtained from the phonetic segmentation of data. Longer or shorter minimum durations degrade or do not improve the results. We have also tried other topologies integrating different types of prior knowledge. The minimum phone duration was the best concerning the overall word recognition performance. However, this does not limit the general ability of the HMM-based enhancement method for integrating other adequate types of knowledge in different applications.

We have used a reduced vocabulary version of the DARPA Conversational Telephone Speech-to-text (CTS) task (1000 words) [9], and OGI Numbers’95 database [10] for the experiments. CTS is the main task in the experiments of this paper. 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/test set was a subset selected from the the NIST 2003 evaluation set. Only those utterances that covered the top most frequent 1000 words with lower than 10% out-of-vocabulary 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 sets contained similar ratio of the number of utterances from Fisher corpus to the number of utterances from the Switchboard corpus. The acoustic features are 13 PLP coefficients concatenated with their first two derivatives. It was computed with vocal tract normalization (VTLN) [26], 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. The structure of the MLP is obtained by cross validation. After training, the phone posteriors for the training set and the test set were estimated.

For the OGI Numbers’95 database, the training set contains 3233 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 1206 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 [20] 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 corresponding to the 27 context-independent phones. In all the experimental setup in the paper, the structure of MLP is obtained using cross-validation. 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.

For both databases, the MLP posteriors were then used to estimate the enhanced phone posteriors as explained in Section III-A. The prior knowledge used to obtain enhanced posteriors is the phonetic duration knowledge as explained before.

We first start with the comparison of the enhanced and MLP posteriors at the frame level. Table I is showing the fame recognition results (non-italic numbers) for the enhanced and regular MLP posteriors (for the two databases). All the error rates in this paper are expressed in percentage. For both databases, the enhanced posteriors show lower frame error rates than the MLP posteriors. We have also performed the statistical significance test [27] in order to verify the reliability of improvements obtained by the enhanced posteriors as compared to the regular posteriors. This test shows if the improvements are due to a genuine advantage of one system over the other, or just an effect of chance. The test is based on a bootstrap method for assigning measures of accuracy to statistical estimates, and it gives a bootstrap estimate of the probability of error reduction (improvement). The results of the test are shown inside parentheses in Table I. The probability of improvement has been expressed in percentage as the confidence on the reliability of improvements. The test indicates a very high confidence (100%) that the improvements in the error rates reflects a real superiority of the enhanced posteriors.

In addition, we 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:

$$E_t = - \sum_{i} p(q_i^t | x_1:T, M) \log_2 p(q_i^t | x_1:T, M)$$

$$AvE = \frac{\sum_{t=1}^{T} E_t}{T}$$

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

<table>
<thead>
<tr>
<th>Database</th>
<th>MLP posteriors</th>
<th>Enhanced posteriors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>35.2</td>
<td>33.3 (100.0)</td>
</tr>
<tr>
<td>Numbers</td>
<td>17.6</td>
<td>16.2 (100.0)</td>
</tr>
</tbody>
</table>

**TABLE I**

Frame error rates on Numbers’95 and CTS, for regular MLP posteriors and HMM-based enhanced phone posteriors. The numbers in parentheses are statistical significance of improvements.

<table>
<thead>
<tr>
<th>Database</th>
<th>MLP posteriors</th>
<th>Enhanced posteriors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>1.64</td>
<td>0.33</td>
</tr>
<tr>
<td>Numbers</td>
<td>0.67</td>
<td>0.18</td>
</tr>
</tbody>
</table>

**TABLE II**

Average entropy of enhanced and regular MLP 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\(^2\). 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. We have also extracted the regular baseline Tandem features by performing Log and KLT transforms on the regular MLP posteriors. For comparison purpose, we also report the standard acoustic feature baseline results obtained by using the traditional PLP features (already used for MLP training) in the HMM/GMM back-end.

For each type of features (PLPs, regular Tandem and combined evidence), we trained a HMM/GMM system using HTK toolkit [28]. 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.

Table III is showing the results in terms of word error rate percentage (for the two databases), using PLPs, regular MLP posteriors, and combined evidence (MLP and enhanced posteriors). The first column shows the standard baseline PLP acoustic feature results. The second column shows the baseline Tandem where the regular MLP posteriors are used as features. The third and forth columns show the performance of the combination of the enhanced and regular posteriors (using different combination rules). The statistical significance of

\(^2\)In practice, using HMM-based enhanced posteriors alone in the Tandem configuration did not improve word recognition performance.
improvements between the baseline Tandem and the system using combined evidence is shown in the parentheses. The combined evidence is performing consistently better than the baseline Tandem and acoustic PLP features. In the CTS case, the best combination rule is concatenation, resulting in about 3% relative improvement with 100% confidence. The very high confidence indicates the true superiority of the new system using combination (concatenation) of enhanced and MLP posteriors. Using enhanced posteriors (encoding prior and contextual knowledge) in combination with MLP posteriors has helped to provide better evidences for Tandem. Smaller improvement in case of using addition rule is due to imperfect combination strategy. The same statistical significance test on the results for Numbers database gives probabilities of improvement ranging between 81% to 92.4% indicating moderately high confidence on the improvements.

<table>
<thead>
<tr>
<th>Database</th>
<th>PLP</th>
<th>MLP posteriors</th>
<th>MLP + Enh</th>
<th>MLP &amp; Enh</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>44.4</td>
<td>44.2</td>
<td>43.8 (76.1)</td>
<td>41.3 (100.0)</td>
</tr>
<tr>
<td>Numbers</td>
<td>7.3</td>
<td>4.7</td>
<td>4.3 (81.1)</td>
<td>4.3 (92.4)</td>
</tr>
</tbody>
</table>

**TABLE III**

Word error rates for PLPs, MLP posteriors, and MLP posteriors combined with the enhanced posteriors, using addition (MLP+Enh) and concatenation (MLP & Enh) as combination rules.

In this work, the HMM used for integrating prior knowledge has fairly simple topology, imposing 3 frames minimum phone duration. The HMM topology is composed of uniformly connected phone models in which each phone is modeled with 3 states, summing up to 138 (46x3) states and only 2300 parameters overall. In such a simple topology, forward-backward recursions are not computationally expensive. Taking into account the huge number of parameters (approximately 190 millions) and computationally expensive large vocabulary decoding in the HMM/GMM back-end, the additional complexity imposed by the posterior enhancement process is practically negligible. In practice, the enhancement process increases the overall Tandem execution duration only by 0.35% and overall complexity (in terms of number of parameters) by 0.001%.

**B. 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 knowledge is learned from the data. Therefore, there is less risk of being possibly 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 IV-A).

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 III-B). 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 by cross validation for all the experiments. The size of the temporal context is close to the reported 200ms duration for phone temporal information.

As before, we start with frame level performance study of enhanced posteriors. Table IV is showing frame error rates percentage 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. Results of the significance test (100% confidence) shows very high reliability of the improvements.

The same as Section IV-A, we do entropy studies on the MLP-based enhanced posteriors. Table V 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.

**TABLE IV**

Frame error rates for regular (first MLP) and enhanced (second MLP) phone posteriors.

**TABLE V**

Average entropy of enhanced and regular posteriors.

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. Details of implementation for the HMM/GMM back-end is the same as Section IV-A. For comparison purpose, we also report baseline PLP acoustic feature performance. Table VI is showing the word recognition performances for PLPs, regular posteriors (baseline Tandem) and enhanced posteriors. It can be observed that the enhanced posteriors are consistently performing better than the regular posteriors and also PLP features. As before, the numbers inside the parentheses are showing the statistical significance of the improvements obtained by enhanced posteriors as compared to the regular posteriors. The probability of improvement is very

\[ \text{In practice, a temporal context of 9-15 frames resulted in very similar results.} \]
high specially for the CTS database, indicating high reliability of the improvements obtained by the system using enhanced posteriors.

<table>
<thead>
<tr>
<th>Database</th>
<th>PLP</th>
<th>Regular posteriors</th>
<th>Enhanced posteriors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>44.4</td>
<td>44.2</td>
<td>42.5 (99.6)</td>
</tr>
<tr>
<td>Numbers</td>
<td>7.3</td>
<td>4.7</td>
<td>4.3 (80.5)</td>
</tr>
</tbody>
</table>

TABLE VI
WORD ERROR RATES FOR PLPS, REGULAR PHONE POSTERIORS AND ENHANCED PHONE POSTERIORS. ENHANCED POSTERIORS ARE OBTAINED BY POST-PROCESSING REGULAR POSTERIORS USING A SECONDARY MLP.

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 4, 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 VII is showing the word recognition results when the MLP-based enhanced posteriors are used as complementary features. As illustrated in Table VII, 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. The statistical significance of the improvements is also very high specially for the CTS database.

<table>
<thead>
<tr>
<th>Database</th>
<th>PLP</th>
<th>MLP posteriors</th>
<th>MLP + Enh</th>
<th>MLP &amp; Enh</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>44.4</td>
<td>44.2</td>
<td>41.2 (99.9)</td>
<td>42.3 (99.0)</td>
</tr>
<tr>
<td>Numbers</td>
<td>7.3</td>
<td>4.7</td>
<td>4.2 (80.2)</td>
<td>4.2 (82.4)</td>
</tr>
</tbody>
</table>

TABLE VII
WORD ERROR RATES FOR PLPS, MLP POSTERIORS, AND MLP POSTERIORS COMBINED WITH ENHANCED POSTERIORS BY ADDITION (MLP+ENH) AND CONCATENATION (MLP & ENH).

MLP-based enhancement approach involves using a second MLP in the Tandem configuration. Thanks to the computational efficiency due to the regular and parallel structures of the MLPs, and more importantly taking into account the huge computational load needed for large vocabulary decoding in the Tandem configuration, the additional computational load imposed is practically negligible. The experiments have shown that the overall execution duration increases only by 0.19% when a second MLP is used. The total number of parameters in the second MLP (approximately 1.5 millions) is in the same order as the number of parameters in the first MLP (approximately 0.5 millions). However, considering the huge number of parameters in the HMM/GMM backend (approximately 190 millions), the overall increase in the complexity is not practically noticeable (only by 0.76%).

V. USING ENHANCED PHONE 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.

A. 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 for the decoder using the enhanced posteriors, and the decoder using regular posteriors [29]. Phoneme deletion penalty is a tuning factor and an engineering trick which is used for numerical compensation of scores for different paths during decoding [28]. It can significantly affect the recognition performance of standard HMM/ANN and HMM/GMM systems. 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 IV-A. We have used a fully constrained model (as explained in Section III-A) 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 [30] 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 5 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 tend to have very close to binary values (similar

\footnote{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.}
Fig. 5. Comparing the sensitivity to tuning phone deletion penalty, for the decoder using enhanced posteriors and the one using MLP posteriors. Phone deletion penalty is varied for the two decoders and the performances are observed (on OGI Numbers’95 database). The diagram inside 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.

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.

### B. 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 IV-B. We have used NOWAY [30] for Numbers database and JUCIER [31] for CTS database as the hybrid decoder. In case of Numbers database, phones are modeled with 3 states in the decoder. In case of CTS database, phones are modeled with 3 states, and a bi-gram language model is used. The number of states in the decoder follows the convention in traditional ASR decoders. In practice also the best performance was obtained by using 3 states. Table VIII 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. As before, the numbers in the parentheses show the statistical significance of improvements, indicating very high superiority of the system using enhanced posteriors. The confidence of improvements are over 99%.

As before, the additional computational load imposed by the MLP-based enhancement process is very small compared to the computation load needed in the decoder, and it can be practically ignored. The experiments have shown that the computation duration for the hybrid system using enhanced posteriors increase only by 1.92%.

<table>
<thead>
<tr>
<th>Database</th>
<th>Regular posteriors</th>
<th>Enhanced posteriors</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>53.6</td>
<td>49.2 (100.0)</td>
</tr>
<tr>
<td>Numbers</td>
<td>9.9</td>
<td>8.8 (99.1)</td>
</tr>
</tbody>
</table>

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 [11] 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. The structure of the MLP has been obtained by cross validation.

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 III-B). This MLP has 741 (39x19) input nodes, 1000 hidden nodes and 39 output nodes (corresponding to the number of phones). Again, the size of the temporal context and the number of hidden nodes have been obtained by cross validation. For the phone recognition, we have used NOWAY [30] which is a hybrid HMM/ANN decoder. In this decoder, each phone is modeled with 3 states, and a bi-gram phone level language model is used. Frame and phone recognition results are shown in Table IX. The enhanced posteriors perform significantly better than the regular posteriors for frame and phone recognition.

<table>
<thead>
<tr>
<th>Error rates</th>
<th>Regular posteriors</th>
<th>Enhanced posteriors</th>
</tr>
</thead>
<tbody>
<tr>
<td>FER</td>
<td>29.9</td>
<td>27.4 (100.0)</td>
</tr>
<tr>
<td>PER</td>
<td>31.2</td>
<td>28.5 (99.3)</td>
</tr>
</tbody>
</table>

### VI. 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 depend 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 as well as phonetic/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 phonetic knowledge. During the inference (forward pass of the second MLP), the learned knowledge is integrated in the phone posterior estimation, resulting in enhanced phone posterior 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. The statistical significance analysis of the improvements has also shown that in most of the cases the improvements obtained by the system using enhanced posteriors is very highly reliable with a confidence of over 99% (in 11 out of 18 experiments). In the rest of the cases, the confidence is still moderately high ranging between 80-92%. As shown in the paper, the increase in the computational load due to the enhancement process is practically negligible.

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, therefore the new local enhanced phone posterior can be used in a wide range of ASR systems. 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.

VII. ACKNOWLEDGEMENTS

The authors want to thank 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 thank Samy Bengio, Hynek Hermansky, Petr Fousek, Joel Pinto, Mathew Magimai Doss, Steve Renals, and Nelson Morgan (ICSI) for helpful discussions.

REFERENCES