Combining Evidence from a Generative and a Discriminative Model in Phoneme Recognition

Joel Pinto and Hynek Hermansky

IDIAP Research Institute, Martigny, Switzerland
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
joel.pinto@idiap.ch, hynek@idiap.ch

Abstract

We investigate the use of the log-likelihood of the features obtained from a generative Gaussian mixture model, and the posterior probability of phonemes from a discriminative multilayered perceptron in multi-stream combination for recognition of phonemes. Multi-stream combination techniques, namely early integration and late integration are used to combine the evidence from these models. By using multi-stream combination, we obtain a phoneme recognition accuracy of 74% on the standard TIMIT database, an absolute improvement of 2.5% over the single best stream.

Index Terms: Phoneme recognition, Gaussian mixture model, multilayered perceptron, phoneme posterior probability, early integration, late integration.

1. Introduction

Phoneme recognition refers to identifying the underlying sequence of phonemes in a speech utterance without the use of any higher level knowledge such as a word language model or a pronunciation dictionary. Phoneme recognition has recently received renewed attention as it is useful in applications such as spoken term detection, language identification, out-of-vocabulary detection etc.

The state-of-the-art approaches to phoneme recognition includes the generative hidden Markov model (HMM) - Gaussian mixture modeling (GMM) of phonemes [1] with additional discriminative training [2]. Other discriminative models such as recurrent neural networks [3], large margin classifiers [4] or multilayered perceptrons (MLP) [5] have given higher phoneme recognition accuracies. In this work, we investigate if the estimates of the posterior probability of phonemes or the likelihood of the features given the phoneme model obtained from a generative model and a discriminative model can be combined effectively to improve the phoneme recognition accuracy.

Two contrasting approaches are investigated to model the acoustic features in an HMM state - A GMM which is a generative model and an MLP artificial neural network (ANN) which is a discriminative model. Due to the inherent difference in modeling and the training criteria, we expect the estimates from these models as ideal candidates for multi-stream combination.

2. Motivation

In this section, we explain the motivation for our work by briefly describing the HMM-GMM and HMM-ANN systems.

2.1. HMM-GMM Modeling

In the conventional generative HMM-GMM approach to speech recognition, a phoneme is modeled using a context independent or context dependent hidden Markov model with certain number of states. The acoustic observation in an HMM state is modeled using a GMM [1]. The model parameters are trained to maximize the total likelihood of the training data.

The likelihood of the data in an HMM state, which is used in Viterbi decoding, is computed using the GMM. However, the posterior probability of a phonemic state can still be computed using Bayes' rule. Suppose x_t is the acoustic feature at time t, the posterior probability of the state $s_t = j$ in phoneme $q_t = i$ is given by

$$P_g(q_t = i, s_t = j|x_t) = \frac{p_g(x_t|q_t = i, s_t = j)P(q_t = i, s_t = j)}{p_g(x_t)}, \quad (1)$$

where, the likelihood $p_g(x_t|q_t = i, s_t = j)$ is estimated using the GMM, and the prior probability $P(q_t = i, s_t = j)$ is estimated by normalizing the state occupancy counts obtained by force-aligning the training data to its true labels. The unconditional likelihood $p_g(x_t)$ is computed indirectly using

$$p_g(x_t) = \sum_{i,j} p_g(x_t|q_t = i, s_t = j)P(q_t = i, s_t = j). \quad (2)$$

2.2. HMM-ANN Modeling

In the discriminative HMM-ANN hybrid approach [6], a multilayered perceptron estimates the posterior probability of a phonemic state directly [7]. The scaled likelihood of the feature x_t in an HMM state $s_t = j$ of phoneme $q_t = i$, which is used in Viterbi decoding is derived using Bayes' rule as

$$\frac{p_m(x_t|q_t = i, s_t = j)}{p_m(x_t)} = \frac{P_m(q_t = i, s_t = j|x_t)}{P(q_t = i, s_t = j)}. \quad (3)$$

Here, the posterior probability $P_m(q_t = i, s_t = j|x_t)$ is given by the MLP, and the prior probabilities are same as in (1). Scaled likelihood is used in decoding instead of absolute likelihood because $p_m(x_t)$ in (3) cannot be computed in this approach. However, this does not affect the Viterbi decoding as $p_m(x_t)$ it is independent of the HMM state.

The generative model fits the training data so that the likelihood of the data given the model is maximized. On the contrary, in discriminative modeling, decision boundaries are trained to minimize the classification error. Due to the inherent differences in the modeling as well as the training criteria, we expect the posterior probabilities derived from these models to be different and hence appropriate for multi-stream combination.

To this end, we investigate **(a)** the use of posterior probabilities of phonemes and the log-likelihoods of the data from the GMM model as features for hierarchical estimation of phoneme posterior probabilities using an MLP **(b)** multi-stream combination techniques such as early integration and late integration for combining evidence from the GMM and MLP models.

3. Basic phoneme recognizer

In this section, we describe the database, feature extraction and the basic systems using HMM-GMM and HMM-ANN modeling. Experiments were performed on TIMIT database, excluding the 'sa' dialect sentences. The training set consists of 3000 utterances from 375 speakers, cross-validation set consists of 696 utterances from 87 speakers and the test set consists of 1344 utterances from 168 speakers. The database, which is hand-labeled using 61 labels is mapped to the standard set of 39 phonemes as explained in [1], except in the way the closures are handled [8].

The speech signal is processed in blocks of 25 ms with a shift of 10 ms to extract 13 perceptual linear prediction cepstral coefficients for every frame. These coefficients, after speaker specific cepstral mean/variance normalization, are appended to their delta and delta-delta derivatives to obtain a 39 dimensional feature vector for every 10 ms of speech.

In the HMM-GMM system, each phoneme is modeled using a three state left-right context-independent HMM. The acoustic features in an HMM state are modeled using a 32 mixture Gaussian mixture model. The model parameters (transition matrices and GMM parameters) are trained using Baum-Welch algorithm followed by embedded re-estimation. In cases where single state model is to be used for analysis, we derive the emission likelihood from the three state model using Bayes' rule as

$$p_g(x_t|q_t = i) = \frac{\sum_{j=1}^{3} p_g(x_t|q_t = i, s_t = j)P(q_t = i, s_t = j)}{P(q_t = i)}, \tag{4}$$

rather than by explicit single state training. Here, the prior probabilities $P(q_t = i, s_t = j)$ and $P(q_t = i)$ are estimated from the training data. The information in the transition matrix of the HMMs are not used in the hierarchical estimation.

In HMM-ANN hybrid model, we use a multilayered perceptron to estimate the posterior probability of a phonemic state. Features are presented to the neural network with a temporal context of 90 ms. The network is trained using the standard back propagation algorithm with softmax output non-linearity and cross entropy error criteria. The learning rate and stopping criterion are controlled by the frame classification accuracy on the cross validation data. The MLP trained on PLP features consists of 351 input nodes, 1000 hidden nodes, and 39 or 117 output nodes depending on whether a single state or three state model is used for the phoneme.

While decoding, all phonemes are considered to be equally probable (*i.e.* no language model). The performance of phoneme recognition is measured in terms of phoneme accuracy (100 - *phoneme error rate*). The optimal phoneme insertion penalty is chosen to give maximum phoneme accuracy on the cross-validation data.

4. GMM log-likelihoods as features

In the HMM-ANN hybrid approach to acoustic modeling [6], the MLP is trained using spectral based features such as Mel frequency cepstral coefficients, or perceptual linear predictive co-

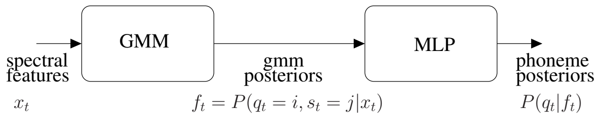

Figure 1: *GMM posterior probabilities as features for phoneme posterior estimation using an MLP.*

efficients. In this paper, we explore the use of **(a)** posterior probability of phonemes given these spectral based features, and **(b)** log-likelihood of the features given the phoneme model, which are obtained from a GMM as features to train the MLP. The block schematic of this hierarchical phoneme posterior probability estimation is shown in Fig. 1.

In the case of the MLP trained on spectral based features, the input to the network is taken with a temporal context of 90 ms. In our case, we present a longer context of 210 ms. The MLP is expected to learn the information in the trajectories of the estimates (posterior probabilities or likelihoods) from the GMM model across different phonemes.

Table 1 shows the phoneme recognition accuracies for the baseline HMM-GMM system as well as the proposed hierarchical setup. Results are shown for three state as well as single state modeling of a phoneme in the GMM stage of the hierarchy. In both these cases, the MLP models a phoneme as a whole. It can be seen that by using by GMM log-likelihoods as features, we obtain about 7-8% absolute improvement over the baseline HMM-GMM system. It can also be seen that the likelihoods are more effective as features to the neural network as compared to the posterior probabilities. This is due the large dynamic range of the GMM likelihoods, resulting in high posterior probability for phonemes even in the case of a misclassification.

Table 1: *Phoneme recognition accuracy using GMM posteriors and likelihoods as features compared to direct HMM-GMM decoding.*

classifier	3-state	1-state
HMM-GMM	64.1	62.1
hierarchy, GMM posteriors	68.4	67.1
hierarchy, GMM log-likelihoods	71.0	70.3

Table 2 shows the phoneme recognition accuracy with the standard hybrid setup where the MLP is trained using PLP features with a temporal context of 90 ms. In the case of single state modeling, by using GMM log-likelihoods as features, we obtain 2% improvement over the baseline HMM-ANN system. We also compare these results to hierarchical estimation method proposed in [8][9], with a setup similar to Fig. 1, except that the GMM is replaced by another MLP, which estimates the posterior probabilities of phonemes given the PLP features. A detailed analysis on the improvement in performance using an hierarchy of two MLPs has been earlier performed in [8]. As seen from the results, the hierarchical posterior estimation using MLP posteriors outperform GMM likelihoods as features. However, the proposed method can be useful in combination with the posterior probabilities derived from the MLP.

The proposed setup can be considered as the converse of the Tandem system [10], where the posterior probability of phonemes derived from an MLP are modified (logarithm fol-

Table 2: *Phoneme recognition accuracy using MLP posteriors as features compared to direct HMM-ANN decoding.*

classifier	3-state	1-state
HMM-ANN	71.6	68.1
hierarchy, MLP posteriors	73.4	71.5

lowed by principal component analysis) and used as features in the standard HMM-GMM system. In an attempt to understand the Tandem system, experiments were designed in [11] where, posterior probability of phonemes from the MLP were replaced by those derived using a GMM. One of the conclusions in this work was that the effectiveness of the Tandem system features comes from the better estimates of the posterior probabilities using the MLP. Experimental results in this paper also suggest that an MLP gives a better estimate of the phoneme posterior probabilities as compared to a GMM. However, the log-likelihoods from a GMM, when taken with a temporal context of around 200 ms are effective as features to an MLP.

5. Multi-stream Combination

The posterior probability of phonemes obtained from an MLP trained on PLP features forms one stream of information and the likelihoods of the features given phonemes, and derived using a GMM forms another stream of information. The two streams can be combined using early and late integration multi-stream combination techniques.

Multi-stream combination is useful when the individual streams bear complimentary information. To get some idea on how much the two streams of information differ, we compare the frame level agreement between these streams and the ground truth. Phonemes are identified at every frame by maximum *a posteriori* classification using the posterior probabilities from the GMM and MLP models. The recognized phonemes are compared to the ground truth labels and their agreement/disagreements are computed.

Table 3: *Frame level agreement/disagreement (in percentage) between the posterior probabilities estimated from GMM and MLP models.*

	gmm correct	gmm wrong
mlp correct	47.5	22.1
mlp wrong	6.7	23.7

Table 3 shows the percentage of times the posteriors from GMM and MLP agree/disagree in the TIMIT test set. As seen in the table, using oracle analysis [12] (cheating experiment), we can obtain a maximum frame accuracy improvement of 6.7% compared to the best stream. In an attempt to exploit this towards improving the phoneme recognition accuracy, we explore the following multi-stream combinations.

5.1. Early integration

In early integration, the streams of information (GMM likelihoods and MLP posteriors in our case) are simply concatenated and presented to the classifier (another MLP in our case) as

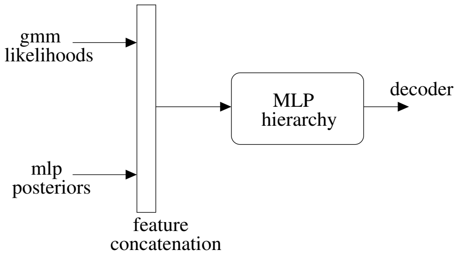

Figure 2: *Block diagram of early integration scheme for combining evidence from GMM and MLP*

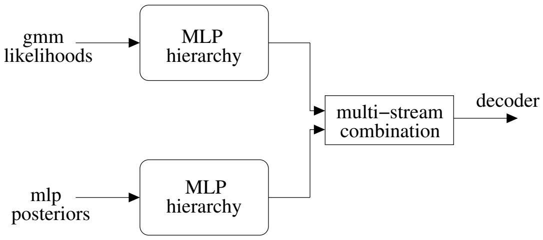

Figure 3: *Late integration scheme for combining hierarchical phoneme posterior probabilities derived from GMM likelihoods and MLP posteriors*

shown in Fig. 2. The MLP is trained to learn the relation in the two streams taking a context of 210 ms.

In the case of single state modeling, by using early integration, we get a higher recognition accuracy of 72.5% as compared to 70.3% obtained using only GMM log-likelihoods and 71.5% obtained by using only MLP posterior as features. By using GMM posteriors instead of likelihoods, we obtain a recognition accuracy of only 71.9%, which further supports our earlier observation that GMM likelihoods are indeed better features than posteriors for hierarchical posterior estimation. Henceforth, we only use the GMM likelihoods as features.

In Fig. 2, due to feature concatenation, there is an increase in the number of input nodes of the MLP compared to the MLP in Fig. 1. Hence, the number of hidden nodes in the MLP classifier in early integration is appropriately modified so that the model parameters are same as in both the cases. This ensures that any improvement in the performance is indeed due to the combination of the features and not due to the increase in the MLP model capacity. In our experiments, we consider a context of 21 frames corresponding to 210 ms. Hence, for single state modeling, the number of input nodes in the MLP in early integration scheme is 1638 (39.21+39.21). The neural network has 2050 hidden nodes and 39 output nodes corresponding to the number of phonemes. The neural network that takes either of the two streams has 819 (39.21) input nodes, 4000 hidden nodes and 39 output nodes.

5.2. Late integration

In the late integration multi-stream scheme, individual streams are presented to separate classifiers and the classifier outputs are appropriately combined. As shown in Fig. 3, GMM log-likelihoods are presented to a MLP classifier to obtain a stream of phoneme posterior probabilities. Similarly, MLP posterior

Table 4: *Phoneme recognition accuracy using late integration scheme for multi-stream combination. Results shown for sum, product, inverse entropy (IE) and Dempster Shafer (DS) combination as well as individual GMM and MLP streams*

	gmm	mlp	sum	prod	I.E.	D.S
1-state	70.3	71.5	73.6	74.0	73.5	73.7
3-state	71.0	73.4	74.2	74.6	74.4	74.6

probabilities (conditioned on spectral features) are presented to another MLP classifier to derive another stream of phoneme posterior probabilities. As in the early integration scheme, the input to these MLPs are taken with a context of 210 ms. The output of the two classifiers, which are probabilities are combined using various multi-stream combination techniques such as sum, product [13][14], inverse entropy [15], and Dempster Shafer [16] combination.

Table. 4 shows the phoneme recognition accuracies for different multi-stream combinations. As seen from the results, for both single state as well as three state (*i.e.* input to the MLP classifiers are three state posteriors/likelihoods) modeling of a phoneme, multi-stream combination gives significant improvement over the single best stream. The improvement in recognition is seen in all the multi-stream combinations with product combination giving the best performance. Results also show that posterior probability of phonemes from MLP forms the single most reliable stream.

5.3. Oracle analysis

We also perform frame level agreement/disagreement analysis between the two individual streams used in late integration combination by comparing them to the ground truth phoneme labels. The results are tabulated in Table 5.

Table 5: *Frame level agreement/disagreement (in percentage) between the posterior probabilities estimated using GMM log-likelihoods and MLP posteriors as features.*

	gmm correct	gmm wrong
mlp correct	64.1	9.2
mlp wrong	8.3	18.4

To estimate the best recognition accuracy using late integration, we perform oracle analysis [12], where we pick the stream with the maximum posterior probability for the ground truth phoneme. Using oracle analysis, we observe a recognition accuracy of 80.50%, which is the best phoneme recognition accuracy that can be achieved using late integration multi-stream combination. The remaining error is attributed to the case where both the GMM and MLP agree and are wrong.

In this work, we have not used any phoneme language model. For comparison with other works, we use a bigram phoneme language model on hierarchically estimated posteriors and obtain an accuracy of 75.0 %. Furthermore, by considering silence class while evaluation, as done in some of the prior works, we obtain an recognition accuracy close to 76.0%.

6. Conclusions

We show that log-likelihood of the typical spectral based features modeled by a GMM can be used as a feature to estimate the posterior probability of phonemes using a neural network. Multi-stream combination using GMM log-likelihoods as one feature stream and posterior probability of phonemes from an MLP as another gives significant improvement in the phoneme recognition accuracies compared to the single best stream.

7. Acknowledgements

This work was supported by the Swiss National Science Foundation under the Indo-Swiss joint research program KEYSPOT, the European Union under the DIRAC integrated project, contract No. FP6-IST-027787 as well as DARPA under the GALE program, contract No. HR0011-06-C-0023.

8. References

[1] K.-F. Lee and H.-W. Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models," *IEEE Trans. Acoust. Speech. Signal Process.*, vol. 37, no. 11, pp. 1641–1648, 1989.

[2] Q. Fu, X. He, and L. Deng, "Phone-Discriminating Minimum Classification Error (P-MCE) Training for Phonetic Recognition," *Proc. of Interspeech*, 2007.

[3] A. Robinson, "An Application of Recurrent Nets to Phone Probability Estimation," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 298–305, March 1994.

[4] F. Sha and L. Saul, "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2006.

[5] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical Structures of Neural Networks for Phoneme Recognition," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2006.

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

[7] M. Richard and R. Lippmann, "Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities," *Neural Computation*, vol. 3, pp. 461–483, 1991.

[8] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai.-Doss, "Exploiting Contextual Information for Improved Phoneme Recognition," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2008.

[9] H. Ketabdar and H. Bourlard, "Hierarchical Integration of Phonetic and Lexical Knowledge in Phone Posterior Estimation," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2008.

[10] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2000.

[11] D. Ellis and M. Gomez, "Investigations into Tandem Acoustic Modeling for the Aurora Task," *Proc. of Eurospeech*, 2001.

[12] H. Misra, J. Vepa, and H. Bourlard, "Multi-stream ASR: An Oracle Perspective," *Proc. of Interspeech*, 2006.

[13] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, Mar. 1998.

[14] L. Kuncheva, "A Theoretical Study of Six Classifier Fusion Strategies," *IEEE Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, Feb. 2002.

[15] H. Misra, H. Bourlard, and V. Tyagi, "Entropy Based Combination of Tandem Representations for Noise Robust ASR," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2003.

[16] F. Valente and H. Hermansky, "Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence," *Proc. of IEEE Conf. Acoust. Speech. Signal Process. (ICASSP)*, 2007.