EXPLOITING CONTEXTUAL INFORMATION FOR IMPROVED PHONEME RECOGNITION

Joel Pinto [1,2], *B. Yegnanarayana* [3], *H. Hermansky* [1,2], *Mathew Magimai.-Doss* [1]

[1] IDIAP Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3] International Institute of Information Technology, Hyderabad, India
joel.pinto@idiap.ch, yegna@iiit.ac.in, hynek@idiap.ch, mathew@idiap.ch

ABSTRACT

In this paper, we investigate the significance of contextual information in a phoneme recognition system using the hidden Markov model - artificial neural network paradigm. Contextual information is probed at the feature level as well as at the output of the multilayerd perceptron. At the feature level, we analyze and compare different methods to model sub-phonemic classes. To exploit the contextual information at the output of the multilayered perceptron, we propose the hierarchical estimation of phoneme posterior probabilities. The best phoneme (excluding silence) recognition accuracy of 73.4% on the TIMIT database is comparable to that of the state-of-the-art systems, but more emphasis is on analysis of the contextual information.

Index Terms— Phoneme recognition, contextual information, hierarchical systems, matched filters.

1. INTRODUCTION

Phoneme recognition refers to identifying the sequence of phonemes present in a given speech signal. Phoneme recognition can be useful in applications such as spoken document retrieval, named entity extraction, out-of-vocabulary detection, language identification, and spoken term detection. Hence, there is an increased interest in the speech research community to develop phoneme recognition systems with accuracies as high as possible.

The state-of-the-art approaches to phoneme recognition include the traditional hidden Markov model (HMM) - Gaussian mixture modeling of phonemes [1] with additional discriminative training [2] techniques. Recently, conditional random fields [3] and large margin classifiers [4] based acoustic modeling have shown to give good recognition accuracies. The best result on TIMIT so far has been achieved by using the hidden Markov model - artificial neural network (ANN) paradigm [5]. We further investigate this approach and explore ways to incorporate the contextual information.

The best recognition accuracy obtained in this work is comparable to those obtained in the state-of-the-art systems [2][3][4][5]. The objective of this work is to investigate the significance of contextual information in the HMM-ANN approach to phoneme recognition. Here, the contextual information refers to the knowledge at two levels **(a)** sequence of feature vectors at the input of the multilayered perceptron (MLP), and **(b)** sequence of phoneme posterior probabilities at the output of the MLP.

We incorporate contextual information at the feature level by estimating the posterior probability of sub-phonemic classes instead of whole phoneme and analyze two approaches for its estimation. We also analyze the contextual information at the output of the MLP and exploit it in an hierarchical system using an MLP or a single layered perceptron (SLP). The SLP is viewed multidimensional matched filter and this interpretation is an extension of [6].

2. BASIC PHONEME RECOGNIZER

The basic phoneme recognition system is based on the hidden Markov model - artificial neural network (HMM-ANN) paradigm [7]. A multilayered perceptron estimates the posterior probability of phonemes given the acoustic evidence $P(q_t = i|x_t)$, where q_t denotes the phoneme index at frame t, x_t denotes the feature vector taken with a window of certain frames. A neural network with sufficient capacity and trained on enough data estimates the true Bayesian *a posteriori* probability [8][7]. The scaled likelihood in an HMM state is given by the Bayes rule as (1), where we assume equal prior probability $P(q_t = i)$ for each phoneme $i = 1, 2 \ldots 39$. The state transition matrix is fixed with equal self and next state transition probabilities. Viterbi algorithm is applied to decode the phoneme sequence.

$$\frac{p(x_t|q_t = i)}{p(x_t)} = \frac{P(q_t = i|x_t)}{P(q_t = i)} \tag{1}$$

Experiments were performed on TIMIT database, excluding the 'sa' dialect sentences. The training data 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. The TIMIT 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. In our case, when a closure occurs before its own burst, the closure and the burst are merged (*e.g.* /tcl t/ → /t/). On the other hand, if a closure precedes any phoneme other than its own burst, the closure is mapped to its burst (*e.g.* /pcl t/ → /p t/).

The speech signal is processed in blocks of 25 ms with a shift of 10 ms to extract 13 perceptual linear prediction cepstral coefficients every frame. These coefficients after 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.

A three layered MLP is used to estimate the phoneme posterior probabilities. The network is trained using the standard back propagation algorithm with cross entropy error criteria. The learning rate and stopping criterion are controlled by the frame classification rate on the cross validation data. In the basic system, the MLP consists of 1000 hidden neurons, and 39 output neurons (with softmax non-linearity) representing the phoneme classes.

The performance of phoneme recognition is measured in terms of phoneme accuracy (100 - *phoneme error rate*). While decoding, all phonemes are considered equally probable (no language model).

1-4244-1484-9/08/$25.00 ©2008 IEEE

The optimal phoneme insertion penalty is chosen to give maximum phoneme accuracy on the cross-validation data. A window duration of 9 frames on the feature vector gives the best phoneme recognition of 68.1%. The recognition accuracy for different window durations are reported in [9]. The context at this stage is only to address the fact that MLP does a record (not sequence) based classification, and feature vectors bear sequential information. In section 3, we try to exploit the contextual information in a more explicit way.

3. CONTEXTUAL INFORMATION FOR PHONEME RECOGNITION

Human speech production is a continuous process, where, depending on the linguistic message to be communicated, the articulators (lips, tongue, vocal cords etc.) are appropriately moved to produce a sequence of information bearing sounds. Due to the inherent inertia in the production mechanism, any sound in this sequence is influenced by its neighboring context. This effect is known as coarticulation.

3.1. Context modeling at the feature level

Due to coarticulation effect, the phoneme has an left segment which is influenced by the preceding phoneme, a center part corresponding to the phoneme, and a right segment which is influenced by the following phoneme. One way to exploit this contextual information is to model the left, middle and right parts of the phonemes using three separate MLP classifiers. For this, each phoneme is segmented equally into three states. For training the left classifier, only the frames belonging to the left part of the phoneme are used. Similarly, the right and middle classifiers are trained independently. Each MLP estimates the posterior probability $P(q_t = i|x_t, s_t = j)$, where q_t denotes the phoneme and s_t denotes the state at time t. The state index can take values $j = 1, 2, 3$ corresponding left, middle, and right phonemic state. The scaled likelihood in an HMM state $(q_t = i, s_t = j)$ is derived using Bayes rule as (2). The state prior probability $P(q_t = i, s_t = j)$ is independent of t and all states are assumed to be equally likely. $P(s_t = j|x_t)$ can be estimated using an MLP, but in this work we make a strong assumption of conditional independence *i.e.* $P(s_t = j|x_t)$ is equivalent to $P(s_t = j)$.

$$\frac{p(x_t|q_t = i, s_t = j)}{p(x_t)} = \frac{P(q_t = i|x_t, s_t = j)}{P(q_t = i, s_t = j)} P(s_t = j|x_t) \quad (2)$$

To validate this formulation, we plot the cumulative distribution function (CDF)[1] of the posterior probability for the phoneme /uw/ obtained from the middle MLP classifier in two conditions: (i) when actually the phoneme /uw/ is uttered and (ii) any other phoneme is uttered as shown in Fig. 1. In the best case, the posterior value should be unity when phoneme /uw/ is uttered and zero otherwise. It is clear from the figure that by independent modeling, we get a CDF slightly closer to the best case than by a single model for the whole phoneme.

In the above case, the sub phonemic classes are not discriminated against each other. Another way to exploit the contextual information is to train a single MLP classifier whose output represents the sub phonemic classes [5]. In this case, the MLP classifier learns to discriminate between the sub phonemic classes and estimates the posterior probability of each state $P(q_t = i, s_t = j|x_t)$. The scaled likelihood in an HMM state is given by (3), where equal prior probability for each state is assumed.

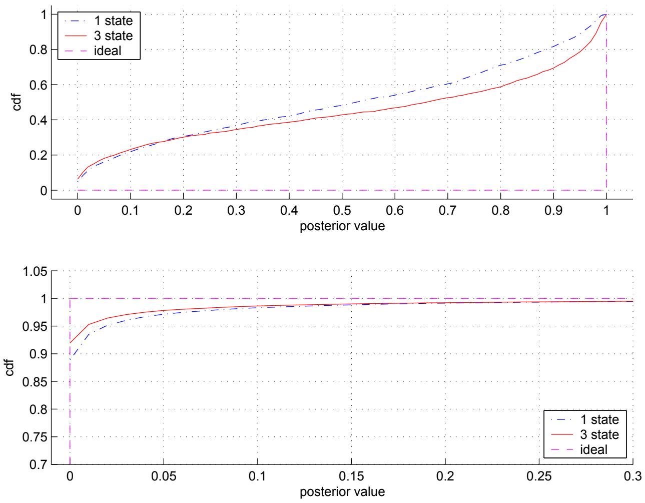

Fig. 1. *CDF of posterior probability of phoneme /uw/ when phoneme /uw/ is said (top) and when other phonemes are said (bottom).*

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

Results in Table. 1 show that for models trained on uniformly segmented labels, independent modeling of sub-phonemic classes (three MLP case) is slightly better than joint modeling (single MLP case) of sub-phonemic classes. Both these approaches perform better than 68.12% obtained by modeling the whole phoneme.

Table 1. *Phoneme recognition accuracy for context modeling with uniformly segmented state labels and force aligned labels.*

classifier	labels for training MLP	
	uniform	force aligned
one MLP with 117 classes	69.87	71.67
three MLPs each 39 classes	70.13	69.70

In the above analysis, a phoneme was equally segmented into three sub phonemic states. One can obtain a more accurate state segmentation by force aligning the posteriors obtained from an MLP trained on hand-labeled data to the true phoneme sequence. The new labels are then used to re-train the MLP classifiers. As shown in Table. 1, independent modeling of sub-phonemic classes do not show any additional improvement in accuracy by using force aligned labels for training as they are insensitive to exact state segmentation. On the other hand, joint modeling of sub-phonemic classes shows significant improvement as class separability is increased.

3.2. Context Modeling at the posterior level

In section 3.1, the state posteriors obtained by sub-phonemic modeling are taken as state emission probabilities in the hybrid decoding framework. As shown in Table. 1, by joint modeling of sub-phonemic classes and a classifier trained on force-aligned labels, a

[1] We choose to plot CDF over the probability density function (PDF) as both its x and y axis are between zero and one.

recognition accuracy of 71.67% is obtained. This compares favorably to the basic system accuracy of 68.12% but useful information can still be contained in the trajectories of the state posteriors.

To validate this hypothesis, we train another MLP to estimate the posterior probability of a phoneme given the state posterior trajectories $P(q_t = i|Q_t)$ obtained from sub-phonemic modeling. Here, q_t denotes the output phoneme index and Q_t denotes the state posterior probabilities at time t taken with a window of certain frames. Hidden layer size was arbitrarily fixed at 3000 neurons, but not much change in recognition accuracies was observed by reducing the size. A window duration of 23 frames gave the maximum accuracy of 73.4%[2][9]. Here, the hierarchical approach is used to estimate the posterior probability of a phoneme as a whole, not much improvement is observed by sub-phonemic modeling.

Hierarchical estimation of posterior probability of phonemes is a non linear system (black box). However, as both input and output to the hierarchy are phonemes, we can study the system by applying carefully designed inputs and analyzing its outputs. This analysis is explained in the following section.

4. ANALYSIS

The factors that could contribute to the improvement in the phoneme recognition accuracy using an hierarchical approach are (a) information across state posteriors of a phoneme, (b) information across state posteriors of other phonemes, and (c) a context of approximately 230 ms taken while combination.

4.1. Information across state posteriors of a phoneme

To study the contribution of using three state posteriors at the input of the combining classifier, we use single state posteriors and compare the recognition accuracies.The single state posteriors could be obtained using a single state MLP (whose output neuron represents a phoneme) or by summing up the state posteriors of phonemes obtained from a three state MLP (whose output neuron represents a phonemic state). A context of 230 ms is presented at the input of the combining classifier. The phoneme recognition results for the hierarchy as well as direct hybrid decoding are given in Table 2.

Table 2. *Recognition accuracy for different inputs to hierarchical posterior estimation. The output of the MLP models the whole phoneme and single state decoding is applied.*

experiment	input to the MLP hierarchy		
	1-state	1-state (sum)	3-state
no hierarchy	68.12	70.17	71.67
hierarchy	71.55	73.01	73.42

It can be inferred from Table 2 that by three state modeling of a phoneme and subsequent three state decoding, the improvement in accuracies come from better modeling of the sub-phonemic states as well as the decoding process itself [5]. The better modeling is evident from the improvement in recognition accuracies by 3-state modeling and single state decoding (70.17%) over single state posteriors and single state decoding (68.12%). The contribution of the decoding process is evident from the increase in accuracy by three state decoding (71.55%) over single state decoding (70.17%) on the posterior obtained from the same three state MLP.

Another inference from Table 2 is that there is an improvement in recognition accuracies by using hierarchical combination of posteriors than directly using in hybrid decoding. In the case of phoneme posteriors obtained by summing state posteriors, the hierarchy performs at 73.01% which is close to combination directly on state posteriors 73.42%. This suggests that information in the state posteriors may be less significant as compared to other factors. In the following section, we investigate the information across different phoneme posteriors.

4.2. Information across phoneme posteriors

To study the contribution of information across different phoneme posteriors at the input of the hierarchy, we distort the input to suppress any information across the phoneme posteriors. In the first experiment (expt-A), at every frame, the maximum phoneme posterior is assigned a value of 0.9 and the rest are assigned random values such that they sum up to 1.0. In the second experiment, (expt-B), at every frame, the maximum phoneme posterior retains its value, but the rest are assigned random values. Table 3 shows the phoneme recognition accuracies for these experiments.

Table 3. *Phoneme recognition accuracies for hierarchical posterior estimation using multilayered and single layered perceptron.*

experiment	no hierarchy	MLP hierarchy	SLP hierarchy
baseline	68.12	71.55	70.40
expt-A	62.77	70.27	69.23
expt-B	64.24	70.75	69.60

In expt-A and expt-B, a decision on the phoneme identity at every frame is already made based on the maximum posterior probability. In expt-A, the input to the combining classifier can be considered as a sequence of discrete phoneme symbols (e.g. phoneme decisions every 10 ms, /b/, /b/, /k/, /b/, /b/, /ah/, /ah/...) presented with a long context. The performance of the hybrid recognition performance drops due to coarse quantization. However, the MLP used for for combination, still outperforms the baseline performance by 2.15%. The only knowledge here is a context of 23 frames provided to the hierarchy, which shows the influence of only the long context presented to the hierarchy. Another important observation is that in expt-A, as the input to the hierarchy is 23 frames of 39 phoneme posteriors, all the data points are near the the basis of (23 * 39) dimensional space. This indicates that a linear classifier may be sufficient for hierarchical combination of the phoneme posteriors. To validate this hypothesis, we investigate the single-layered perceptron (SLP) which is discussed in the next section.

4.3. Single layer perceptron (SLP)

A single layered perceptron linear classifier is used for hierarchical estimation of the phoneme posterior probabilities. Unlike the MLP, there exists a closed form solution for the SLP weights [10], but we use gradient descent approach with softmax output nonlinearity and cross entropy error criteria. As shown in Table 3, with single state

[2] By using a bigram phoneme language model on hierarchically estimated posteriors, we obtain an accuracy of 73.85%. Furthermore, by considering silence class while evaluation, as done in some of the prior works, we obtain an recognition accuracy of 75.0%.

phoneme posteriors at its input, the SLP hierarchy compares favorably to the baseline (no hierarchy) accuracy of 68.12% but performs poorer compared to the MLP hierarchy by about 1%.

By extending the SLP hierarchy for three state phoneme posteriors, we obtain an accuracy of 72.01% compared to 71.67% obtained by hybrid decoding and 73.40% obtained by using an MLP hierarchy. Despite its poor performance compared to MLP hierarchy, SLP is still useful in understanding hierarchical estimation as it can be viewed as a matched filter and this interpretation is an extension of the work described in [6].

4.4. SLP as a matched filter

In the work [6], a novel approach for phoneme spotting is proposed. A matched filter for each phoneme is derived independently by averaging its phoneme posterior trajectory. The width of the matched filter captures the duration of the phoneme and height captures the prior probability of the phoneme. The phoneme posteriors are convolved with their respective matched filters and peaks are picked to spot phonemes.

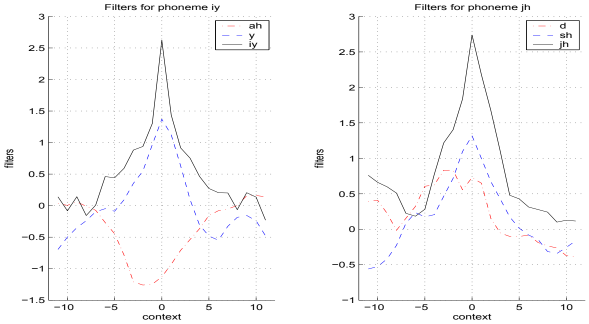

Fig. 2. *The matched filters for the phoneme /iy/ and /jh/. The plot also shows top three contributing phonemes in the filter.*

The single layered perceptron can be viewed as multidimensional linear matched filters derived jointly for all phonemes by minimizing the cross entropy error criteria. Fig 2 shows the matched filter for the phonemes /iy/ and /jh/. The SLP matched filter of a phoneme (*e.g.* /iy/) captures the contribution of different phoneme posteriors at the input of the SLP (in the window duration of 23 frames) to the posterior probability of phoneme /iy/. Phoneme /iy/ has a negative contribution from the phoneme /ah/. In the matched filter for the phoneme /jh/, there is a contribution from phoneme /d/ and its peak precedes the center of /jh/, which is consistent with the production of /jh/. A similar phenomenon is observed for the phoneme /ch/, which has a similar contribution from phoneme /t/.

5. SUMMARY AND CONCLUSIONS

In this paper, we further investigate the hidden Markov model - artificial neural network paradigm for phoneme recognition and analyze the contextual information at the features level as well as the output of the MLP (phoneme or state posterior probabilities). At the feature level, we probed two ways to estimate the state posteriors of phonemes which are (a) independent modeling of the three subphonemic classes (three MLP case) and (b) joint modeling of subphonemic classes (single MLP case). Experiments suggest that after force alignment, the joint modeling gives the best performance, but this could be due to the strong assumption of conditional independence.

We also analyzed the contextual information in the phoneme and state posterior probabilities. We show that hierarchical estimation of phoneme posterior probabilities using MLP or SLP gives better recognition accuracies compared to direct hybrid decoding (no hierarchy). The major factor for this is a context of approximately 230 ms, even though the information across the phoneme/state posterior trajectories are also important.

Hierarchical estimation of the phoneme posterior can also be viewed as a classifier combination, where the MLP or SLP uses the output of the first classifier over a window of 23 frames and makes a new decision. The inferior performance of the SLP hierarchy could be attributed to its inability to learn simple voting rules (such as max) for classifier combination. Nevertheless, the SLP can be interpreted as a linear multidimensional matched filters which enables us to study the relations between input and output phoneme posteriors in the hierarchical classifier. While such an relation certainly exists in the case of an MLP, it is difficult to plot or analyze it due to the presence of the hidden layer.

6. ACKNOWLEDGEMENTS

This work was supported by the European Union under the the DIRAC integrated project, contract No. FP6-IST-027787, the DARPA GALE program, contract No. HR0011-06-C-0023 and the Indo-Swiss project KEYSPOT from the Swiss National Science Foundation. The authors thank S.R.M. Prasanna, IIT Guwahati, India and John Dines, IDIAP for the fruitful discussions.

7. 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] Y. Hifny and S. Renals, "Speech Recognition using Augmented Conditional Random Fields," *IEEE Trans. Speech. Audio. Process.*, vol. 1, no. 11, 2007.

[4] F. Sha and L.K. Saul, "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," *Proc. of ICASSP*, 2006.

[5] P. Schwarz, Matejka. P, and J. Cernocky, "Hierarchical Structures of Neural Networks for Phoneme Recognition," *Proc. of ICASSP 2006*, pp. 325–328, 2006.

[6] M. Lehtonen, P. Fousek, and H. Hermansky, "Hierarchical approach for spotting keywords," *IDIAP Research Report*, , no. 05-41, 2005.

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

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

[9] J. Pinto, S.R.M. Prasanna, B. Yegnanarayana, and H. Hermansky, "Significance of Contextual Information in Phoneme Recognition," *IDIAP Research Report*, , no. 07-28, 2007.

[10] C.M. Bishop, "Neural Networks for Pattern Recognition," 1995.