IDIAP RESEARCH REPORT

Significance of Contextual Information in Phoneme Recognition

Joel Pinto [a] [b] S.R.M. Prasanna [c] B. Yegnanarayana [d]
Hynek Hermansky [a] [b]

IDIAP–RR 07-28

March 2007

soumis à publication

[a] IDIAP Research Institute, Martigny, Switzerland
[b] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[c] Dept. of ECE, Indian Institute of Technology (IIT) Guwahati, India
[d] International Institute of Information Technology, Hyderabad, India

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Rapport de recherche de l'IDIAP 07-28

Significance of Contextual Information in Phoneme Recognition

Joel Pinto S.R.M. Prasanna B. Yegnanarayana Hynek Hermansky

March 2007

soumis à publication

n this paper, we investigate the significance of contextual information at various stages in the development of a phoneme recognition system using an artificial neural network. A phoneme is treated as made up of three sub-phonemic states representing left contextual information, right contextual information and the steady state of the phoneme. Contextual information is probed at the level of sequence of feature vectors and at the output of the multi layered perceptron. By a series of incremental improvements, we obtain a phoneme recognition accuracy of 73.4% on TIMIT database using a reduced phoneme set of 39 phonemes.

1 Introduction

Phoneme recognition refers to identifying the sequence of phonemes present in a given speech signal. Phoneme recognition can be useful in applications like spoken document retrieval, named entity extraction, out-of-vocabulary (OOV) detection and language identification. In spoken document retrieval and named entity extraction, phoneme recognition is used to index speech as a compact and easy-to-search form. Phoneme sequence obtained by phoneme recognition can be compared against those obtained as by-product in a conventional automatic speech recognition (ASR) for OOV detection. Also, phoneme recognition can be used for extracting phonotactic information in speech for language identification. Therefore, a good phoneme recognition has direct impact on the performance of the aforementioned applications. Hence, there is increased interest in speech research community to develop a phoneme recognition systems with accuracy as high as possible [1][2].

Phoneme recognition is evaluated by comparing the recognised phonemes to labeled reference phoneme sequence. The estimated accuracy of a phoneme recognizer depends on several factors including : (a) accuracy of labeling, (b) accuracy of pronunciation, (c) representation of speech in terms of features, (d) models used for classification, and (e) exploiting the knowledge at various levels like production, acoustic-phonetic and linguistic levels. By careful consideration to several of these factors, a phoneme accuracy of about 75% is realized on TIMIT database with reduced phoneme set of 39 phonemes [1].

It is less likely to achieve any significantly higher accuracy over the currently available systems by addressing any single issue. Our work aims at exploring aspects for phoneme recognition which may compliment the existing systems. It may be possible for careful integration of some of our research findings to the existing systems to further improve the accuracy. In this direction, we try to investigate the contextual information at various levels in a phoneme recognizer to achieve small incremental improvement in the recognition accuracy.

The objective of this study is to show that by exploiting the contextual information in a systematic way, it is indeed possible to get additional improvement in the phoneme recognition, which in turn may help to improve the accuracy of the speech recognizer. We study the effect of contextual information using a basic hidden Markov model - artificial neural network (HMM-ANN) phoneme recognition system [3]. The phoneme recognizer consists of perceptual linear prediction (PLP) coefficients for representation, multilayered perceptron (MLP) for estimating the phoneme posterior probabilities and Viterbi decoder for finding the phoneme sequence. The contextual information refers to the knowledge at three levels **(a)** sequence of feature vectors, **(b)** sequence of phoneme posterior probabilities, and **(c)** sequence of phonemes level.

Hand labeled TIMIT database with a reduced phoneme set of 39 classes is used for this study. The rest of the paper is organized as follows : The database, feature extraction, hybrid phoneme recognizer and the baseline results are discussed in Section 2. The experimental studies related to the proposed contextual information are discussed in Section 3. Section 4 provides a summary and directions for future work.

2 Basic Phoneme Recognizer

In this section we discuss the HMM-ANN approach to phoneme recognition. The experimental details and the results for a basic phoneme recognition system are also described briefly.

2.1 Hybrid HMM-ANN

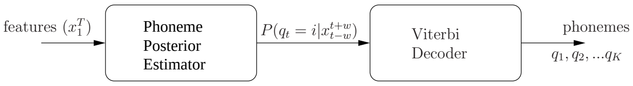

FIG. 1 – *Block diagram of hybrid HMM-ANN phoneme recognition system.* x_1^T *is the feature vector,* $P(q_t = i|x_{t-w}^{t+w})$ *is the posterior probability, and* $q_1, q_2 \ldots q_K$ *are the phonemes decoded.*

The hybrid HMM-ANN [3] phoneme recognition system consists of two blocks as shown in Fig. 1. In the first block, a multilayerd perceptron is used to estimate the posterior probabilites of phonemes using sufficiently long temporal context of feature vectors. Neural network classifiers estimate the Bayesian *a posteriori* probability provided that, the network is complex enough, trained on sufficient training data, and the classes are taken with the correct *a priori* probabilities. The proof for this can be found in [4]. In the second block, these posterior probabilities are taken as emission probabilites in the states of the phoneme HMM, and Viterbi algorithm is applied to find the best phoneme sequence. In all the experiments, the transition matrix is fixed with equal self and next state transition probabilities.

2.2 Experimental Details

Experiments were performed on TIMIT database [5]. The dialect sentences ('sa') were excluded from the training and test data. 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 is hand-labeled using 61 labels. These are mapped to a reduced set of 39 labels as explained in [6], 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 (PLP) cepstral [7] coefficients every frame. The resulting features after cepstral mean/variance normalization are appended to its delta and delta-delta derivative to obtain a 39 dimensional feature vector for every 10 ms of speech. In the experimental studies, 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 depends on the frame classification rate on the cross validation data.

The performance of phoneme recognition is measured in terms of phoneme accuracy (100 - PER, where PER is the phoneme error rate). Phoneme insertion penalty is the only free parameter in the recognition system. The optimal phoneme insertion penalty is that which gives best accuracy on the cross-validation data. The silence class is not considered for evaluation and accuracy is reported for the remaining of the 38 phonemes. The hybrid decoding toolkit in [8] was used.

2.3 Baseline Results

For the basic system, the MLP trained to estimate the phoneme posterior probability consists of 1000 hidden neurons, and 39 output neurons representing the output phoneme classes. The feature vector presented at the input of MLP consists of a window of certain number of frames to capture the trajectory of features in the feature space.

The accuracy of phoneme recognition for a basic system is shown in Table. 1 for different values of window duration. A window size of nine frames corresponding to 90 ms seems to give the best phoneme recognition accuracy. It is clear from the table that most of the improvement is obtained by going from no-context to a context of 30 ms. 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. Some of the ways to exploit the actual contextual information is given in Section. 3.

TAB. 1 – *Accuracy of phoneme recognition for different feature level context frames presented at the input of the MLP. One frame corresponds to* 10 *ms of time duration.*

Context Frames	Phoneme Accuracy	Context Frames	Phoneme Accuracy
1	61.92	9	68.12
3	66.79	11	68.12
5	67.27	13	68.02
7	67.92	15	67.61

3 Contextual Information for Phoneme Recognition

Human speech production is a continuous process, where, depending on the linguistic message to be communicated, the articulators (*e.g.* lips, tongue, vocal chord etc.) are appropriately moved to produce a sequence of information bearing sounds. However, due to the inherent inertia in the production mechanism, any sound in this sequence is influenced by its neighbouring context. This effect is known as coarticulation. In addition to coarticulation, there is a contextual information at the linguistic level arising due to the distribution of phoneme sequences in a language.

3.1 Context Modeling (hand-labeled data)

Due to contextual effect, the phoneme has an initial segment which depends on its left context, a center part corresponding to the phoneme, and a right context which depends on 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 divided equally into three parts using the hand labeled phoneme segmentation. For training the left MLP classifier, only the frames belonging to the left part of the phoneme are used. Similarly, the right and middle MLP classifiers are trained independently.

To validate this hypothesis, we plot the cumulative distribution function (CDF)[1] of the posterior probability of a phoneme (*e.g.* 'uw') obtained from the middle MLP classifier in two conditions : (i) when actually the phoneme 'uw' is uttered as shown in Fig. 2a and (ii) any other phoneme is uttered as shown in Fig. 2b. In the ideal case, the posterior value should be unity when phoneme 'uw' is uttered and zero otherwise. The corresponding CDF is also shown in both the figures. It is clear from the figure that by independent modeling, we get a CDF slightly closer to the ideal case than by a single model for the whole phoneme. Only frames corresponding to middle part of the phonemes are considered while estimating the CDF in order to avoid any errors due to wrong labeling at the phoneme boundary.

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

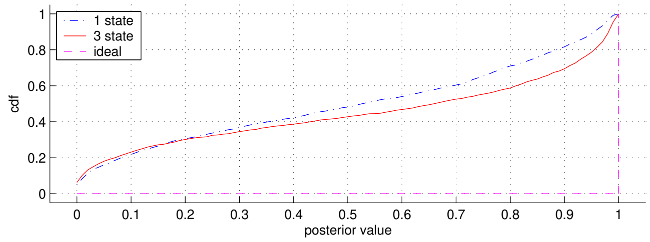

Fig. 2a – *CDF for the posterior probability of phoneme 'uw' when the uttered phoneme is 'uw'.*

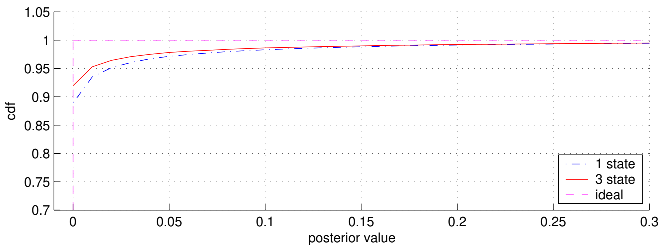

Fig. 2b – *CDF for the posterior probability of phoneme 'uw' when other phonemes are uttered.*

In the above approach, the sub-phonemic classes are not discriminated against. Another approach would be to train a single classifier with all the sub-phonemic classes as output neurons [1]. In this case, the MLP classifier learns to discriminate between the sub-phonemic classes, also refered to as phoneme states.

The posteriors obtained from the context modeling are taken as the emission likelihoods in the HMM states of the respective phoneme model, and Viterbi algorithm is applied to get the phoneme sequence. Table. 2a shows the recognition accuracy obtained for independent sub-phoneme modeling (three MLP case) and a single MLP modeling. The accuracy in both these cases is better than that obtained in the basic system.

Tab. 2a – *Phoneme recognition accuracy for context modeling with uniform segmentation. Decoding with three state HMM.*

Classifier	Accuracy(%)
one MLP with 117 classes	69.87
three MLPs each 39 classes	70.13

Tab. 2b – *Phoneme recognition accuracy for context modeling after forced alignment. Decoding with three state HMM.*

Classifier	Accuracy(%)
one MLP with 117 classes	71.67
three MLPs each 39 classes	69.70

3.2 Context Modeling (force aligned labels)

While modeling the context in Section. 3.1, it was assumed that the models representing the context within a phoneme captures the true information of the context even when each phoneme is divided equally into three parts. One can obtain a more accurate state segmentation by force aligning the true phoneme sequence using the posteriors obtained by hand-labeled data. The aligned labels are then used to re-train the MLP classifier. The phoneme recognition accuracy for the forced aligned case is given in Table. 2b. It can be seen that single MLP estimating the state posteriors gives an improved accuracy when trained on force-aligned labels. On the other hand, independently trained network does not show any improvement. This is because in the case of independent training, the sub-phonemic classes are not discriminated against each other and hence they are insensitive to the exact boundaries. However, in the case of single MLP case, this helps the separability among the classes.

3.3 Context Modeling at the posterior level

In Sections 3.1 and 3.2, the posteriors obtained by context modeling are taken as state emission probabilities in the hybrid decoding framework. Even though this yields better accuracy compared to the basic system, information can still be contained in the trajectories of the state posteriors. This can be captured by training an MLP to estimate the phoneme posterior probability given the state posterior trajectories.

The 117 state posteriors obtained from the network trained to discriminate the sub-phonemic classes are presented to another MLP classifier with 3000 hidden neurons and 39 output phoneme classes. If n frames of context is taken, the input MLP layer will have $n \times 117$ neurons. Table 3 shows phoneme recognition accuracy for various values of n. The posterior probability obtained from this merging classifier gives a recognition accuracy of 73.4% compared to 68.12% from the basic system.

Tab. 3 – *Recognition accuracy for different values of the context at the phoneme posterior level. A frame corresponds to* 10 *ms of time interval.*

context frames	phoneme accuracy	context frames	phoneme accuracy
1	69.64	15	72.70
3	69.69	17	72.77
5	70.28	19	73.21
7	70.82	21	72.29
9	71.34	23	73.42
11	71.69	25	73.42
13	72.42		

4 Discussion and Conclusions

The context present in the speech signal may be viewed at different levels such as : **(a) Production Level** - Context arising due to the inertia of the articulators. Speech information characterizing a particular sound extends beyond its boundary to the neighbouring sounds. **(b) Linguistic Level** - Context arising due to the linguistic sequence produced to convey the information. This captures the phonotactic information **(c) Semantic Level** - Context arising due to meaning in the spoken message. Human speech recognition can seemlessely integrate the different contextual information to understand the linguistic message. In the case of automatic speech recognition, while semantic context is never exploited, there have been attempts to exploit production and linguistic contests.

In this work, we analyse context information in a hybrid phoneme recognition system. At this stage, it is difficult to attribute with certainty the improved accuracy using an hierarchical MLP to any particular reason. However, it is interesting to note that at the feature vector level, the maximum accuracy was reached for a context of about 90 ms, while at the posterior level, the best accuracy is reached for a context of about 230 ms. As this duration roughly correspond to the duration of about three phonemes, one may attribute the improvement to capturing the linguistic information. However, a detailed analysis needs to be carried out to ascertain this fact by carefully designed experiments.

In this work, we attempt to understand the significance of contextual information at various levels in a hybrid phoneme recognition system. Experimental results indicate the importance of contextual information. We believe that more focused effort is required to exploit this information to further improve the recognition accuracies.

5 ACKNOWLEDGEMENTS

This work was supported by the European Union under the the DIRAC integrated project, contract number FP6-IST-027787 as well as the Indo-Swiss project from the Swiss National Science Foundation, and the DARPA GALE program. The authors acknowledge Guillermo Aradilla, IDIAP and Petr Schwartz, Speech@FIT, Brno for their help.

Références

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

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

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

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

[5] W.M Fisher, G.R Doddingtion, and K.M Goudie-Marshall, "The DARPA Speech Recognition Research Database : Specifications and Status," *Proc. of DARPA Workshop on Speech Recognition*, pp. 93–99, Feb 1986.

[6] 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, Nov 1989.

[7] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of Acoustical Society of America*, vol. 87, no. 4, April 1990.

[8] "The STK Toolkit," *http ://www.fit.vutbr.cz/speech/sw/stk.html*.