In Search Of Target Class Definition In Tandem Feature Extraction

Sunil Sivadas[1] *and Hynek Hermansky*[1,2]

[1]OGI School of Science & Engineering at OHSU, Portland, Oregon, USA.
[1,2] International Computer Science Institute, Berkeley, California, USA.
{sunil,hynek}@ece.ogi.edu

Abstract

In the tandem feature extraction scheme a Multi-Layer Perceptron (MLP) with softmax output layer is discriminatively trained to estimate context independent phoneme posterior probabilities on a labeled database. The outputs of the MLP after nonlinear transformation and Principal Component Analysis (PCA) are used as features in a Gaussian Mixture Model (GMM) based recognizer. The baseline tandem system is trained on 56 Context Independent (CI) phoneme targets. In this paper we examine alternatives to CI phoneme targets by grouping phonemes using apriori and and data-derived knowledge. On connected digit recognition task we achieve comparable performance to the baseline system using fewer data-derived classes.

1. Introduction

In the tandem feature extraction scheme a MLP is used as feature extractor [1, 2, 3]. The MLP is trained with softmax nonlinearity in the final layer and one-from-N target coding scheme to estimate posterior probabilities of target classes. During forward pass the softmax activation function is replaced with linear activation to obtain features that are close to Normal distribution. The features are further processed by Principal Component Analysis (PCA) to decorrelate and to optionally reduce the dimensionality and are fed to HMM. We used Context Independent (CI) phoneme classes as targets in our earlier work [1, 2, 3]. Figure 1 shows a block diagram of the tandem feature extraction scheme. The focus of this paper is to investigate the alternatives to (CI) phoneme targets. In [4] the MLP is trained to discriminate between Hidden Markov Model (HMM) states of whole word HMMs. They observed that the performance did not change a lot compared to phoneme targets. In a similar approach [5] neural networks are trained to map short-term spectral features to the posterior probability of distinctive features. They used 60 distinctive features comprising articulatory features [6] , plus some broad phonetic classes as targets. Using 44 CI phoneme targets they obtained better performance than the distinctive features on a large vocabulary task.

Motivation for this work is to find smaller set of target classes than CI phoneme classes. This results in smaller set of features and fewer parameters in the GMM classifier. The dimensionality of the feature vector can be reduced using PCA without changing the number of classes. Another method is to cluster the phoneme classes. Clustering using prior knowledge such as voiced/unvoiced and vowel/consonant results in broad category targets. We do not know whether this is an optimal way of clustering. A more structured approach is data-driven clustering.

Next section explains the various target classes we investigated in this paper and how we obtained them. Section 3 provides the details of experimental setup and results. Finally, section 4 discusses the observations and conclusions from this work.

2. Target class definition

Are CI phonemes optimum targets for tandem feature extraction? Since we use a nonlinear and complex HMM classifier with multiple states and Gaussians, we may not require the MLP to discriminate among all the classes. We start with ICSI56 phoneme set. We try four clustering methods to reduce the number of classes, 1) using apriori knowledge to cluster the phonemes to broad phonetic categories 2) decision tree based clustering [7] 3) data driven clustering of phoneme models [8] and 4) Mutual Information (MI) based clustering to reduce the phoneme confusions.

2.1. Broad phonetic categories

We cluster CI phonemes based on their phonetic properties to obtain seventeen broad phonetic categories. The categories are front vowels, central vowels, back vowels, retroflexes, diphthongs, voiced plosives, unvoiced plosives, nasals, flaps, voiced fricatives, unvoiced fricatives, affricates, glides, voiced closures, unvoiced closures, syllabics and silence. Table 1 shows the grouping of phonemes into broad phonetic categories. By training a MLP on these targets we are extracting phonetic "features".

2.2. Data derived classes

Extracting features based on hardwired phonetic attributes may not be optimal for classification of phonemes. There are many approaches in literature to derive data from classes using clustering techniques.

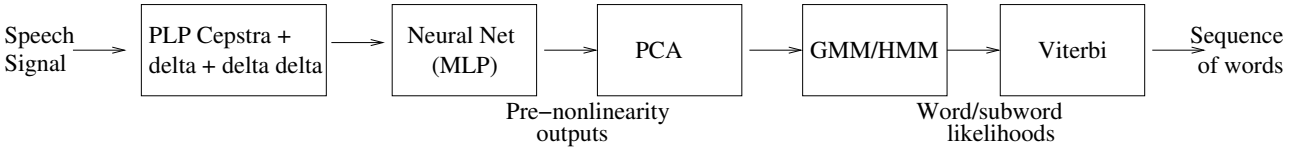

Figure 1: Block diagram of the tandem feature extraction scheme.

Broad phonetic category	Phonemes
front vowels	iy, ih, eh, ae
central vowels	ix, ux, ax
back vowels	uw, uh, ah, ao, aa
retroflexes	er, axr
diphthongs	ey, ay, oy, aw, ow
unvoiced plosives	p, t, k
voiced plosives	b, d, g
nasals	m, n, nx, ng
flaps	dx
unvoiced fricatives	f, s, th, sh, hh
voiced fricatives	v, dh, z, zh
affricates	ch, jh
glides	l, y, r, w
voiced closures	bcl, gcl, dcl
unvoiced closures	pcl, kcl, tcl
syllabics	em, en, el
silence	h#, q

Table 1: Grouping of phonemes into broad categories.

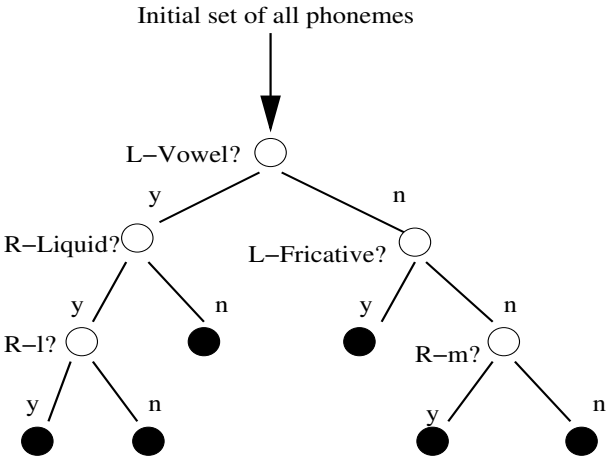

Figure 2: Example of a phonetic decision tree.

2.2.1. *Decision tree based clustering*

Here phoneme models are clustered using a phonetic decision tree [7]. A phonetic decision tree is a binary tree with yes/no questions attached to each node. We use HTK [9] to build the decision tree. Initially the 56 CI phoneme are modeled using single state, single gaussian models. Each phoneme is renamed to have the same central phoneme. For example phoneme "aa" is renamed as "aa-phn+aa" and phoneme "p" as "p-phn+p", so that all phonemes are placed in a single cluster at the root of the tree corresponding to the central phoneme "phn". The decision tree asks whether the phoneme to the left/right of the central phoneme is in a certain set, e.g. "Is the phoneme to the left or right a plosive?". Typically a linguist expert derives the questions sets. Examples are: "Vowel", "Fricative" , "Stop", etc. The question that gives the maximum increase in log likelihood is chosen at each node. This process is repeated until the increase in log likelihood falls below a specified threshold. We vary the number of classes from two to fifty five by changing the threshold. Figure 2 shows an example of splitting classes using decision tree. Broad phonetic categories are a special case of classes obtained using decision tree clustering.

2.2.2. *Data driven clustering of phoneme models*

Initially all the phoneme models are placed in individual clusters. Here again the phonemes are modelled using single state, single gaussian HMM. The pair of clusters which, when combined, would form the smallest resultant cluster are merged. This process repeats until the number of clusters have reached the specified number. The size of the cluster is defined as the greatest distance between any two phoneme models. Euclidean distance between the class conditional means weighted by the inverse of the variance is used as the distance metric. We use HTK to implement the clustering.

2.2.3. *Mutual information based clustering*

First a hybrid HMM/MLP [10] is trained to estimate the phoneme posterior probability using the manually labelled training data. Using the frame level phoneme classification results on the training data a confusion matrix is obtained. A confusion matrix (CM) is a matrix of hits and misses for all phonemes. A joint Probability Distribution Function (PDF) is estimated from the confusion matrix by dividing each element in it with the total number of phoneme segments. We compute the Mutual Information (MI), $I(X;C)$ between the feature vector X and phoneme C from the joint PDF. $I(X;C)$ is the reduction in uncertainty of the phoneme C due to the knowledge of X [11]. The pair of phonemes which, when combined, would result in the maximum reduction in $I(X;C)$ are merged to form new classes. For example, "em" and "en" are merged in the first step to form a new class "em_en".

Feature	WER (%)
PLP Cep+Δ+$\Delta\Delta$	6.2
Baseline Tandem	5.7

Table 2: Word Error Rates (WER) on connected digit recognition task.

Phoneme model	Number of classes			
	12	24	36	48
HMM-1 state	10.1	8.7	7.6	6.3
HMM-3 states	7.6	6.8	6.2	5.9
MLP	7.1	6.4	6.0	5.7

Table 3: Results of MI based clustering.

The process is repeated until all the phonemes are paired.

3. Experiments and Results

We used two databases in our experiments. One is the English part of OGI Stories database [12]. It is approximately 3 hours of hand-labelled speech data. It is labelled by ICSI56 phoneme set. This is used to train the MLP and in deriving new categories from data. The other is the database on which recognition is performed, namely, OGI Numbers database. It contains ten continuous digits in utterances varying between one and seven digits, labelled by twenty-three phonemes. The database is split into approximately 20000 digits for training and 12000 digits for testing. Note that all the clustering schemes are performed on a database independent of the final recognition task.

Baseline tandem system is trained on ICSI56 CI phonemes. Input to the MLP is nine frames, four frames from past and four frames in future, of 8 PLP cepstral coefficients, 8 delta and 8 double delta features (24x9=216) after utterance based mean subtraction. Phoneme label corresponding to the center frame is used as the target class. MLP has 216 input units, 500 hidden units and 56 output units.

The single state, single gaussian HMMs used in clustering are trained on 8 PLP cepstral coefficients, 8 delta and 8 double delta features. After clustering the classes using the aforementioned techniques, the CI phonemes are mapped to the new classes and an MLP is trained on each of the new classes. Each MLP has same number of input and hidden units as the baseline system, only the number of output units vary from two to fifty-five.

3.1. Results

Connected digit recognition experiments are performed on OGI Numbers database. The 23 context independent phonemes are modelled using 5 state left-to-right HMMs with 3 Gaussians/state and diagonal covariance matrix. Table 2 gives the Word Error Rate (WER) for 8 PLP cepstral coefficients, 8 delta and 8 double delta features after utterance based mean subtraction and the baseline tandem system.

Figure 3 shows the WER for various clustering schemes and different number of classes. It can be seen that WER rolls off much faster with increasing number of clusters using MI based clustering than decision tree based clustering and data driven clustering of phoneme models. The WER for MI based clustering saturates at 29 classes, the WER=6.0% (not statistically significant compared to 5.7% at 95% confidence). The best WER of 5.5% is obtained using 34 categories. The WER for tree-based and data-driven clustering of phoneme models continues to improve with increase in number of classes. This could be attributed to the complexity of the phoneme models used in clustering. Both tree based and data driven clustering use single state, single gaussian model (because of the limitations of the software), where as MI based clustering use a MLP to generate the confusion matrix. To verify this we trained two sets of HMMs to generate phoneme confusion matrices. A single state, single gaussian model and three state, eight gaussian components per state model on OGI stories. Phoneme recognition is performed on training data and phoneme confusion matrix is obtained as explained in section 2.2.3. Table 3 compares the WER obtained for 12, 24, 36 and 48 categories obtained using HMM and MLP. It can be seen that the more the complexity of the model used in estimating MI the better the WER.

Table 4 compares the performance of seventeen broad categories with the same number of classes obtained using MI based clustering. It shows that "better" target categories can be obtained using data driven methods than grouping based on phonetic properties.

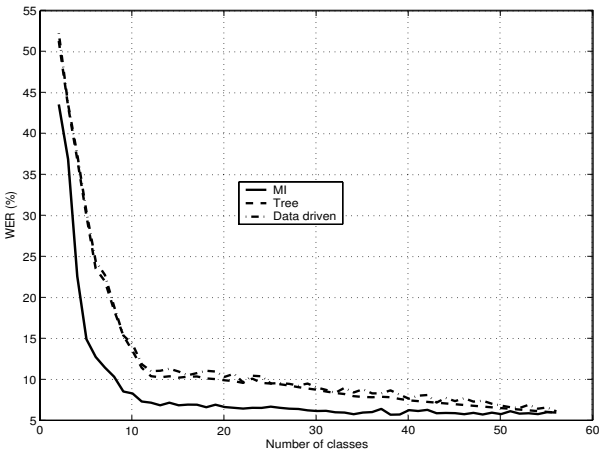

Figure 3: Word Error Rates (WER) for different clustering schemes and number of classes.

Clustering method	WER (%)
Broad categories	8.1
MI based	6.9

Table 4: Word Error Rates (WER) on connected digit recognition task for seventeen categories obtained by MI based clustering.

Clustering	WER (%)
Random-1	10.9
Random-2	12.4
Random-3	12.5

Table 5: Results using random clustering of phonemes.

3.1.1. Random clustering

To test whether it is the "meaningful" clustering approach that is providing the improvement, we cluster the phonemes randomly. A random number generator produces the indices of phonemes to be grouped. We generated three different random mappings to reduce the number of classes to seventeen. As shown in Table 5 the WER increases.

3.1.2. PCA versus clustering

The number of features can be reduced using PCA. We reduce the dimensionality of the baseline tandem features from 56 and compare the results with same number of classes obtained using MI based clustering. From Table 6 it can be seen that the performance of clustering is superior to PCA. This shows that clustering retains more information for discriminating among phonemes than PCA.

4. Conclusions

The choice of CI phonemes as target classes for MLP is arbitrary. We have investigated alternative target definitions by grouping phonemes based on different clustering schemes. The MLP is trained on these new target classes. Grouping phonemes to reduce the mutual information between classes and features based on the phoneme confusion matrix has shown promising results. We have obtained performance comparable to 56 CI phoneme targets using 34 data derived classes. This results in a classifier with fewer parameters without sacrificing the performance.

	Number of features			
	12	24	36	48
PCA	7.2	6.8	6.4	5.9
MI	7.1	6.4	6.0	5.7

Table 6: PCA vs. MI based clustering.

5. Acknowledgements

The research was supported by DARPA EARS program under MDA-972-02-01-0024.

6. References

[1] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems", *Proc. ICASSP'00,* Istanbul, June 2000.

[2] S. Sivadas, P. Jain and H. Hermansky, "Discriminative MLPs in HMM-based recognition of speech in cellular telephony", *in Proc. ICSLP'00,* Beijing, China, October 2000.

[3] D.W.P. Ellis, R. Singh and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition", *in Proc. ICASSP'01,* Salt Lake City, Utah, USA, May 2001.

[4] D.W.P. Ellis and M. J. R. Gomez, "Investigations into Tandem Acoustic Modeling for the Aurora Task", *in Proc. Eurospeech'01,* Copenhagen, Denmark, September 2001.

[5] B. Launay, O. Siohan, A. Surendran and C.H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition", *in Proc. ICASSP'02,* Orlando, Florida, USA, May 2002.

[6] K. Kirchhoff, "Robust Speech Recognition Using Articulatory Information", *TR-98-037,* ICSI, 1998.

[7] S.J. Young, J. Odell and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling", *in Proc. ARPA Workshop on Human Language Technology,* Berlin,1994.

[8] S.J. Young and P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition", *in Proc. Eurospeech'03,* Berlin, September 1993.

[9] S. Young, "The HTK Hidden Markov Model Toolkit: Design and Philosophy", *Technical Report TR.153,* Department of Engineering, Cambridge University, UK, 1993.

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

[11] Thomas M. Cover and Joy A. Thomas, *Elements of Information theory,* John Wiley & Sons, Inc., 1991.

[12] R. Cole, M. Noel, and T. Lander, "Telephone speech corpus development at CSLU", *Proc. ICSLP 94, September 1994.*