IDIAP RESEARCH REPORT

On Use of Task Independent Training Data in Tandem Feature Extraction

Sunil Sivadas [a] [b]   Hynek Hermansky [a]

IDIAP–RR 03-57

October 2003

SUBMITTED FOR PUBLICATION

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] Dalle Molle Institute for Artificial Intelligence CH-1920, Martigny, Switzerland.
[b] OGI School of Science and Engineering at OHSU, Portland, Oregon, USA.

IDIAP Research Report 03-57

On Use of Task Independent Training Data in Tandem Feature Extraction

Sunil Sivadas Hynek Hermansky

October 2003

submitted for publication

Abstract. The problem we address in this paper is, whether the feature extraction module trained on large amounts of task independent data, can improve the performance of stochastic models? We show that when there is only a small amount of task specific training data available, tandem features trained on task independent data give considerable improvement over Perceptual Linear Prediction (PLP) cepstral features in Hidden Markov Model (HMM) based speech recognition systems.

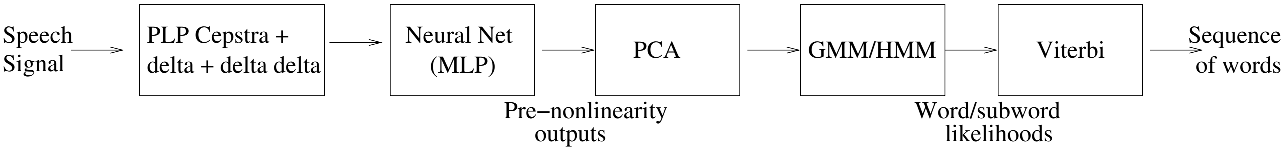

Figure 1: Block diagram of the tandem feature extraction scheme.

1 Introduction

In the tandem feature extraction scheme an MLP was successfully used as a feature extractor for small vocabulary speech recognition tasks [1, 2] and with limited success in large vocabulary tasks [3]. Here the MLP was 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 linear outputs are further processed by Principal Component Analysis (PCA) to decorrelate and to optionally reduce the dimensionality, and are used as features in a Hidden Markov Model (HMM) based recognizer. Figure 1 shows a block diagram of the tandem feature extraction scheme.

Since the MLP and HMM are trained separately, they can be trained on different databases as well. Current HMM based classifiers require large amounts of task specific training data to achieve competitive performance. The problem we address in this work is, can the features be trained on a large amount of task independent data to reduce the requirement of task specific training data for the subsequent stochastic model based classifier? By task independent database, we mean a database that is not specific to any task but contains all the variability that is encountered in the test condition. Here the MLP learns to suppress the variability in the data that is not helpful to classification of features and enhances the variability that is helpful. Since the features are already trained, we expect that the HMMs require smaller amounts of task specific training data than when training them directly on acoustic features, such as PLP cepstral coefficients. This is particularly helpful in practical situations where one has very limited task specific data. The ultimate goal of this data-guided feature extraction paradigm is to acquire permanent knowledge from a large amount of task independent training data and use the features in all kinds of speech recognition tasks. In this paper we systematically study the performance of HMM based speech recognizers as a function of the amount of task specific training data.

The next section compares the performance of features trained on task specific and task independent data. Subsequent sections study performance of the systems by varying the amount of task specific training data.

2 Using Both Task Independent and Task Specific Data

We use two databases in our experiments.

- The English part of the OGI-Multilingual Corpus [4], known as OGI-Stories, as task independent data. OGI-Stories database has 3 hours of manually transcribed telephone quality spontaneous speech. It is transcribed into 41 context-independent phonemes.

- OGI-Numbers as task specific data. OGI-Numbers contains ten continuous digits in utterances varying between one and seven digits, labeled by twenty-three phonemes. The database is split into approximately 20000 digits for training and 12000 digits for testing.

System	WER (%)
$PLP - HMM_{Stor}$	5.7
$PLP - HMM_{Dig}$	5.1
$PLP - HMM_{Stor+Dig}$	5.3
$Tand_{Stor} - HMM_{Stor}$	5.2
$Tand_{Stor} - HMM_{Dig}$	4.7
$Tand_{Dig} - HMM_{Dig}$	4.4
$Tand_{Stor+Dig} - HMM_{Stor+Dig}$	4.5

Table 1: Results using the entire task specific and task independent data.

8 PLP cepstral features, its first and second derivates are calculated from the speech signal. The features are then mean and variance normalized over an utterance. The MLP uses 9 frames of normalized cepstral features (9x24=216) as input. It has 500 hidden units and one node per phoneme. The MLP trained on OGI-Stories ($Tand_{Stor}$) has 41 output nodes and the MLP trained on OGI-Numbers ($Tand_{Dig}$) has 23 output nodes. To make the number of features comparable to cepstral features, only the 24 dimensions corresponding to the largest 24 eigenvalues are retained at the output of $Tand_{Stor}$ after PCA. We train Hidden Markov Model (HMM) using HTK [5]. We use 3 state context-dependent HMMs, each state modeled by mixture of 8 Gaussians. HMMs are trained on both OGI-Stories (HMM_{Stor}) and OGI-Numbers (HMM_{Dig}). The Word Error Rates (WER) using various combinations of training and testing using available databases are tabulated in Table 1. From Table 1 the following things can be observed.

- Tandem features perform better than PLP cepstral features irrespective of the type of training data.

- Training HMMs on the task specific data is better than training on task independent data.

- The Tandem system trained on task independent data ($Tand_{Stor} - HMM_{Stor}$) performs better than the PLP system trained on task independent data ($PLP - HMM_{Stor}$) and comparable to the PLP system trained on task specific data ($PLP - HMM_{Dig}$).

- The best performance is obtained by training both the MLP and HMM on task specific data ($Tand_{Dig} - HMM_{Dig}$).

3 Limited Amount of Task Specific Training Data

Since the MLP is trained on large amounts of task independent data, we expect the knowledge acquired by the MLP to be helpful in reducing the amount of training data required by HMM without sacrificing the performance. The tandem features are trained once on the entire task independent data and only the HMMs are trained on varying amounts of data. The dash-dot (red) line and dash-dash (blue) line in Figure 2 show the WER as a function of the amount of HMM training data. It can be seen that the performance of the HMM trained on cepstral features degrades faster with reduction in training data. To confirm that this is actually due to the training of features and not due to discriminative features, we train the MLP and HMM on the same amount of task specific data. The solid (green) line in Figure 2 shows the WER when both the MLP and HMM are trained on same amount of task specific data. From the figure it can be observed that the performance of tandem and cepstral features are comparable when the HMMs are trained on the entire task specific data. Also, the difference is greatest when there is less training data. The best performance is obtained when the tandem features are trained on task specific data. This explains why the WER for $Tand_{Dig} - HMM_{Dig}$ is lower than $Tand_{Stor} - HMM_{Dig}$. When the training

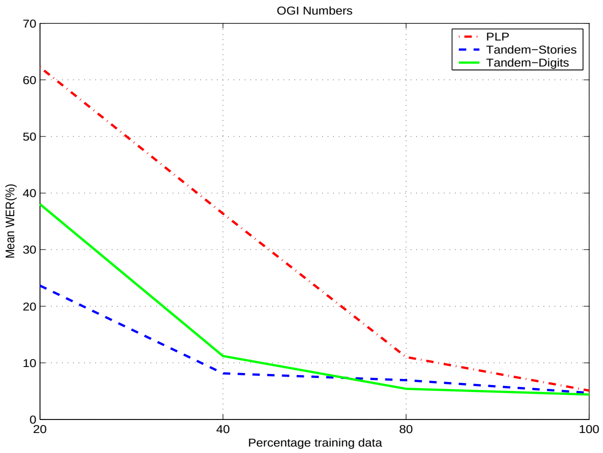

Figure 2: Word Error Rates (WER) for continuous digits recognition task as a function of the amount of training data.

data for MLP is reduced severely, it starts over-fitting the data and performance on test data suffers. This is evident by the cross-over of $Tand_{Dig} - HMM_{Dig}$ performance around 60% training data.

To verify whether this observation holds for another task, we use the Speech In Noisy Environments (SPINE) database [6]. It involves a medium-sized vocabulary of about 5000 words. The data consists of conversations between two communicators working on a collaborative, Battleship-like task in which they seek and shoot at targets. Each person is seated in a sound chamber in which a previously recorded military background noise environment is accurately reproduced. The speech is sampled at 16KHz. PLP cepstral features are extracted from a frame of 25 ms of speech, every 10ms. The feature vector consists of 13 PLP coefficients augmented by deltas and double-deltas. They are then normalized over the utterance to zero mean and unit variance. The input to each MLP is a window of 9 successive feature vectors. The training set is divided into two parts, one is used to train MLP and the other to train HMM to simulate the task specific and task independent data. Figure 3 shows the results on SPINE data. The trend is similar to the small-vocabulary test data, except that the WER is higher due to the higher complexity of the task.

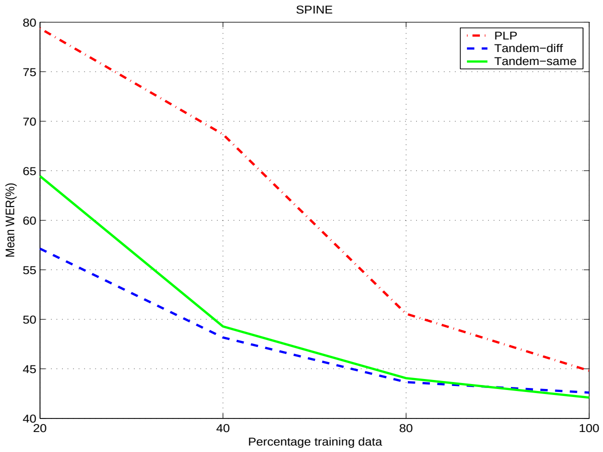

Figure 3: Word Error Rates (WER) for large vocabulary recognition task as a function of the amount of training data.

System	WER (%)
$PLP - HMM_{Dig_{20\%}}$	62.2
$PLP - HMM_{Stor}$	5.7
$PLP - HMM_{Stor+Dig_{20\%}}$	5.6
$Tand_{Dig_{20\%}} - HMM_{Dig_{20\%}}$	38.1
$Tand_{Stor} - HMM_{Dig_{20\%}}$	23.6
$Tand_{Stor} - HMM_{Stor}$	5.2
$Tand_{Stor+Dig_{20\%}} - HMM_{Stor+Dig_{20\%}}$	5.0

Table 2: Results using task independent data and small amount of task specific data.

To study the situation when the availability of task specific data is very limited, as in many practical situations, we use only 20% of the task specific data. From Figure 2 it can be seen that the difference in performance between cepstral features and tandem is the largest when the HMMs are trained with the least amount of data.

3.1 Using task independent data together with a small amount of task specific data

We train both HMM and MLP using the entire task independent data and 20% of the task specific data. Table 2 lists the WER for various combinations of training data. The following observations can be made from Table 2.

- Using small amounts of the task specific training data to train tandem features and HMM, the WER is reduced by 39% relative to HMM trained on cepstral features with the same amount of training data.
- Using the MLP trained on task independent data to extract features, and training the HMM on small amounts of task specific data, we obtain relative WER reduction of 62% compared to the cepstral system.
- By training the MLP and HMM on the combination of task independent data and a small amount of task specific data, the WER is reduced by 11%.

4 Conclusion

In this paper we addressed the problem of how features trained on large amounts of task independent training data reduces the requirement of task specific training data for the HMM. With small amounts of task specific training data, the tandem system outperforms the cepstral system. This may be due to the knowledge acquired by the tandem features from the task independent data. We showed that the performance of tandem features is superior to cepstral features even when all the available training data is used to train HMM.

5 Acknowledgements

The authors want to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)", as well as DARPA through the EARS (Effective, Affordable, Reusable Speech-to-Text) project.

References

[1] H. Hermansky, D. Ellis and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems”, *in Proc. ICASSP’00,* Istanbul, Turkey, June 2000.

[2] 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.

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

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

[6] http://elazar.itd.nrl.navy.mil/spine/spine1