Optimization of Class Weights for LDA Feature Transformations

Andrej Ljolje

AT&T Labs - Research
Florham Park, NJ 07932-0971
U.S.A.
alj@research.att.com

Abstract

One popular feature type in speech recognition is based on linear transformations of sequences of cepstral feature vectors. In general the transformation is generated in two steps: first a transformation like linear discriminant analysis (LDA) or heteroscedastic linear discriminant analysis (HLDA) is used to maximize separation between classes and reduce the dimensionality, followed by a decorrelating transformation. Here we investigate the weighting of classes when using the LDA transformation. In particular we are concerned with the special status of silence, for which the data can be arbitrarily long, and which can be represented by more than one silence/noise model. The special case of our acoustic models for commercial applications, which consist of several sub-models for each type of application, like general English, digits, names, alphabet, etc., creates a conflict when using a transformation like LDA to improve the separability of states which correspond to the same phoneme, but used within a different type of task. We also evaluate replacing sample counts with error/accuracy counts and cross-task LDA transformation estimation. The results show that it is important to take these conditions into account and demonstrate accuracy/speed improvements when appropriate care is taken in computing the LDA transformations.

Index Terms: speech recognition, acoustic modeling, LDA class weights, error count weights.

1. Introduction

Traditional speech recognition features capture short term spectral characteristics, only relying on the static features of the short speech segment. The next development consisted of augmenting the spectral features by the time derivative of the features, doubling the size of the feature vector. The natural extension was to use the second time derivative, resulting in three times as many features, but improving, in general, both the recognition accuracy and even speed due to better pruning by a more accurate acoustic model.

The more recent approach to modeling the dynamic aspects of the speech signal uses a sequence of spectral feature vectors [3]. A number of feature vectors, usually 9-15, is concatenated into an extended feature vector. Such vectors are not suitable for use in acoustic modeling, but their reduced form is. A technique like linear discriminant analysis (LDA) can be used to reduce the dimensionality of the extended feature vector and to maximize the separation of the acoustic model classes. In general the classes range form the states of the context independent phoneme hidden Markov models (HMMs) to the states of the fully context dependent phoneme HMMs. The number of classes is thus usually somewhere between 100-200 and many thousands, respectively for a conventional model for the recognition of continuous speech. Special applications, like digit recognition may differ from the general case.

In the case of all the recognition experiments described here there is an additional level of complexity. The acoustic model we use has been designed to achieve the best possible performance on ANY task it might encounter in commercial speech recognition applications[1, 2]. That means it has to perform recognition of digit strings as well as the best digit recognition systems, as well as being just as competitive when recognizing colloquial general English, or any number of other applications. This is achieved by using an extended phoneme set, creating a subset of phonemes that handle just one application. The term phoneme is used loosely, as for example, in the case of digits we use the head-body-tail structure for each digit, where the segments do not correspond to phonemes. The phonemes are all context dependent (triphonic) based on the phonetic features of the neighboring phoneme. Thus we achieve context dependency across task boundaries, allowing for more than one type of application within the same utterance. A typical example would have a digit string embedded within a general English carrier phrase. This acoustic model also contains four silence/noise (henceforth silence) HMMs, two single state, two with the three state left-to-right structure, resulting in the total of eight silence states. The model has a total of six parts covering six different types of applications with up to six versions of all English phonemes. All HMMs are three-state left-to-right except for the digits bodies which are four-state HMMs. Multiple versions, at least in principle, of the same phoneme or silence states could potentially cause problems with techniques like LDA (or equally HLDA [4]) which would try to create the transformation to best separate even the segments that are acoustically equivalent.

2. Experimental Setup

The acoustic models have all been trained on the same training data, which consists of over a thousand hours of speech from numerous databases covering the six types of applications. Only three domains are tested here, general English, alphabet and digits. The training speech initially contained over 1000 hours of silence in addition to the speech. This was considered wasteful and most of the silence was removed, preserving only a short silence segment before and after utterances including all of the silence within utterances. Due to the computational load imposed by running so many experiments on such a large database, the training process was minimized. The best available MMI trained model was used to segment the training database. Every time a different LDA transformation was generated, the MFCC features were transformed, a

new estimate of the decorrelating transformation [5, 6] was found and a new model was built using the existing segmentations, without further training. Computational efficiency was also the reason for using LDA rather than HLDA, as it has a closed form solution and is consequently much easier to compute and more than an order of magnitude faster than HLDA.

All the testing was done on three different test sets. One was alpha-digits (AD), with over a thousand strings, each with seven alpha-digits. The language model did constrain for the length of the alpha-digit sequence. The additional two tasks were general English (GE), which we call TASK 1 and TASK 2 when describing the results. Both are collections of the responses to a system greeting in two different customer care applications, both with over five thousand utterances by real customers.

The LDA transformation consists of collecting the sample statistics for each class. Those statistics are then weighted based on their frequency and then the transformation is computed. The frequency can be based on the number of frames, counting all the frames aligned with the input class in forced alignment of the training data and an existing acoustic model. We chose number of segments as weight, which, unlike the number of frames, ignores the length of segments/states/classes, as it better matches the application of speech recognition. It also produced better results in the past experiments when direct manipulation of the weights given different classes was not used. Based on the results here the conjecture is that the weight given to silence when counting silence states, and not silence frames, gave a better result, and the change in the rest of the segments was insignificant.

3. Experimental Results

In all the experiments here we use some or all of the 9219 states of the context dependent acoustic model as classes for the purpose of training the LDA transformation. In the first set of experiment we chose to use only the states that correspond to the General English (GE) subset of the model and the silence states. First we compare the performance of the LDA transformation where the eight silence states are kept as separate classes with the case where they are merged into a single class. The weight given to the different classes is the state counts, or the number of class occurrences in the training database. If the weight of all the GE states , not including silence, is normalized to 1.0, then the natural weight of the silence states, based on the sample statistics was W = 0.27. The performance is displayed as an accuracy/speed curve for the three test sets in Figure 1.

It is clear from the results that it can be beneficial not to separate states that belong to the same class into multiple competing classes, when computing the LDA transformation. In this experiment we used the arbitrary weight for silence based on the training database. However, in this case, the amount of the silence data was arbitrarily reduced to what remained in the final training data.

Next we experiment with the weight given to the silence class (the eight silence classes will be merged into a single class in all the following experiments) relative to the normalized weight of 1.0 given to all the other states used in estimating the LDA transformation. In this case it is still only the GE states. The effect of varying the silence class weight is shown in Figure 2.

Although there is some variability across different tasks, it is clear that low weight of 0.27 or the high weight of 2.0 are clearly inferior to the silence weights around 1.0.

Having established the optimal operating point for the silence

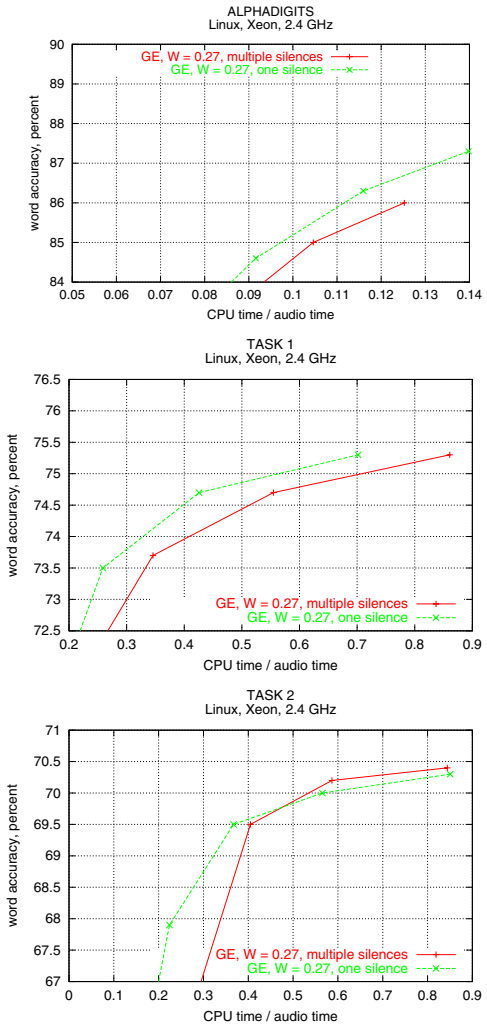

Figure 1: Word recognition performance on different recognition tasks using one versus all eight classes for silence

classes when computing the LDA transformation, we next investigate other alternatives for selecting weights of all classes, but preserving the relative settings for the silence. Since the target is the improvement of recognition accuracy, and LDA provides improved separation of classes, we decided to make the weights correspond to the class error rates. Hopefully, this would give higher weight to the "problem" classes improving their separability, without impairing the performance of the already "good" classes. Since the classes in this case are states of the context dependent HMM states, we needed to find their recognition error rate.

The original best acoustic model was used to segment the reference transcriptions, providing the reference state sequences. It was also used to perform recognition of all the training speech, and after segmenting the hypothesized word sequences the same way the reference transcriptions were segmented we were able to compute the state error rates. The error rate is computed as the sum of all cases when the state was deleted, inserted, substituted by another state, or another state was substituted by it, divided by the total number of state occurrences. Similar counts are found for accuracies too, which are also used as class counts, and compared

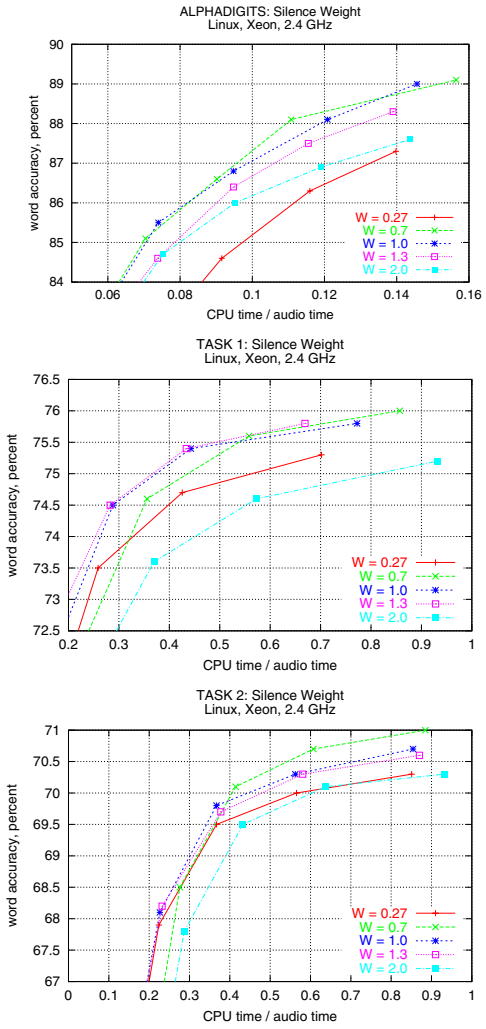

Figure 2: Word recognition performance using different weights for one silence class

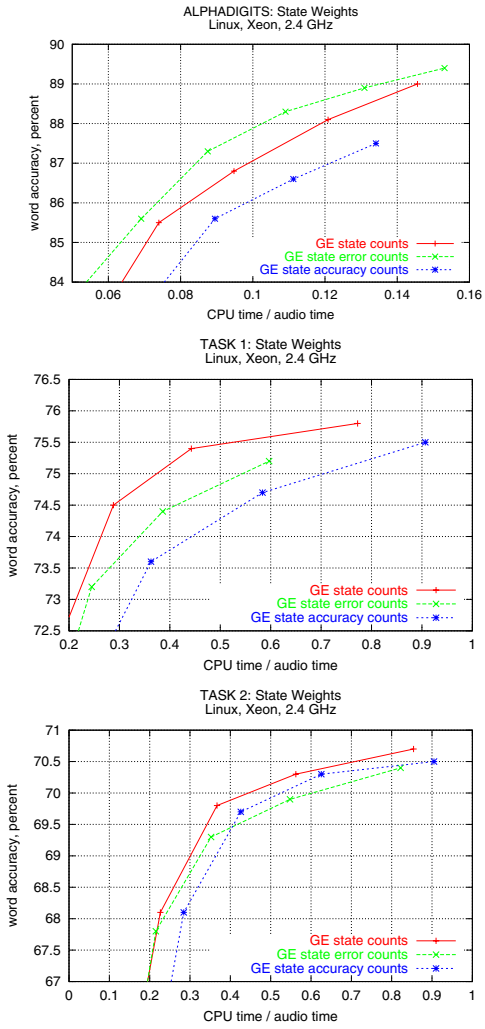

Figure 3: Word recognition performance when class weights for computing the LDA transformation are GE class counts, class errors or class accuracy

to the error counts and the sample counts. Similar counts apply to the silence classes, but they are still merged, albeit at different relative weights, into one silence class, whose weight is set to be the same as the total weight of all the other classes used in computing the LDA transformation. The recognition results are shown in Figure 3.

The results show two different characteristics. When the LDA transformation is trained using the same subset of the states as is use for testing the model (GE in both cases), sample counts provide the best performance. However when the testing is performed on AD test set after training on GE data, the error counts provide the best basis for training the LDA transformation. In order to investigate this phenomenon further we train LDA transformation using only alpha-digit states, including silence, still preserving the same silence class weights and merging into a single silence class. The results based on alpha-digits trained LDA transformation is shown in Figure 4. These results point to several different conclusions. The first one is that using the AD error counts is bad. This might be caused by the fact that many more errors are caused by digits than alphabet part of the model creating an imbalance. Also it seem clear that cross training is beneficial. LDA trained on AD works well when recognizing GE, but not for recognizing AD, and similarly, LDA trained on GE works well for recognizing AD, but not as well on GE, comparatively speaking. In addition the best result for the AD task is achieved when LDA is trained using the GE data weighted by the error counts.

We finally test what happens when all the data is included to compute the LDA transformations, ignoring the potential problem of same phoneme separation across different task specific phoneme sets. Those results are compared in the same plots with both the task dependent and cross task trained LDA transformation, with class weights based on sample and error counts. They are shown in Figure 5.

There is no single setting that excels in all the testing conditions. However some general trends can be observed. For a skewed task like alpha-digits it appears that the best performance is achieved by either training the LDA transformation using GE or all the available data, especially if using the error counts. For the general English tasks, LDA computed on the AD data works well,

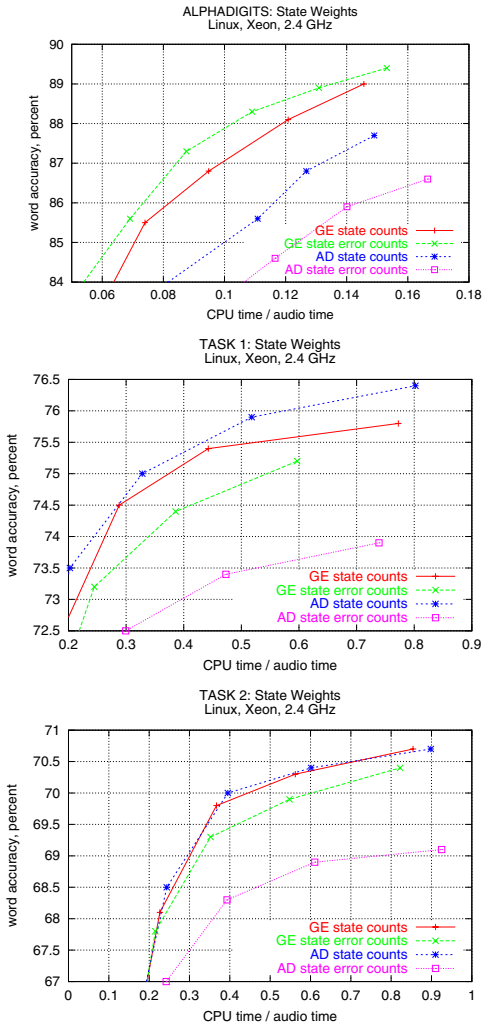

Figure 4: Word recognition performance when class weights are either GE or AD class counts or class errors

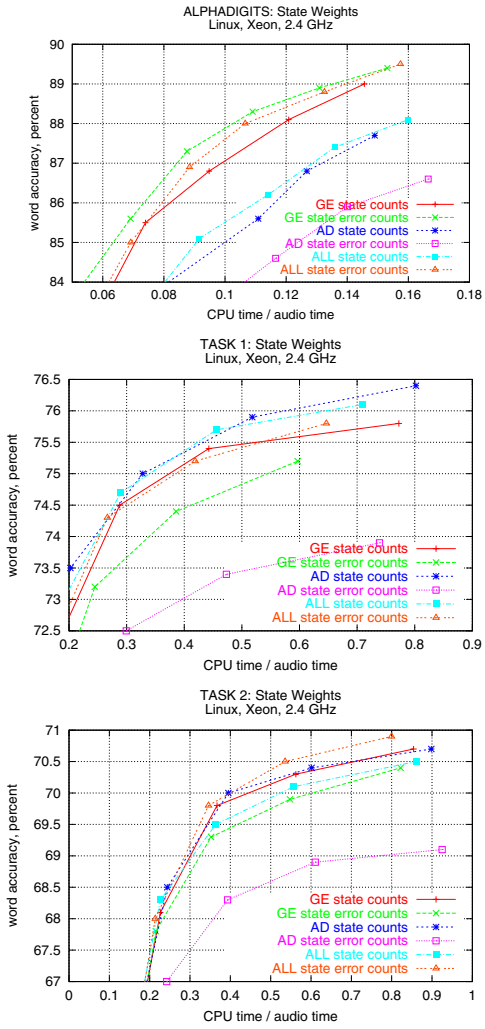

Figure 5: Word recognition performance when all the data, GE subset or the AD subset are used for LDA transformations and are weighted using either class counts or class error counts

providing the sample counts are used. However, recognition of the GE tasks is also accurate when using all the available data, with one task slightly more accurate when using error counts, the other with using sample counts.

4. Conclusions

We investigated the effects of different setups when computing the LDA transformation, as used in conjunction with a decorrelating transformation. We compared single vs. multiple silence classes (single is better), different weights given to the silence class (about 1.0 is best), using sample counts and error counts for weights of individual classes, with error counts slightly advantageous, especially if based on a balanced class set. A surprise was that excellent performance is achieved when computing the LDA transformation on the data from a task significantly different than the test set used for evaluating the model. Fortunately, using the data from all the tasks provides close to optimal performance on all the tasks, especially if error counts are used instead of sample counts.

5. References

[1] Ljolje, A., "Multiple task-domain acoustic models," In *Proceedings ICASSP,* 2003.

[2] Ljolje, A., Goffin, V. and Saraclar, M., "Low Latency Real-Time Vocal Tract Length Normalization", *Text, Speech and Dialog '04*, Brno, Czech Republic, 2004.

[3] Saon, G., Padmanabhan, M., Gopinath, R. and Chen, S., "Maximum Likelihood Discriminant Feature Spaces ", In *Proceedings ICASSP,* , 2000.

[4] Kumar, N. and Andreu, A., "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition", *Speech Communications*, 26, pp. 283-297, 1998.

[5] Gales, M. J. F., "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech and Language,* 12, pp. 75-98, 1998.

[6] Gales, M. J. F., "Semi-Tied Covariance Matrices for Hidden Markov Models," *IEEE Transactions on ASSP,* Vol. 7, 1999.