INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

# Deployment of RASTA-PLP with the Siemens ZT Speech Recognition System

Michael L. Shire *

TR-97-057

December 1997

**Abstract**

RelAtive SpecTral Analysis - Perceptual Linear Prediction (RASTA-PLP) is the standard speech feature extraction method used at the International Computer Science Institute. There it has been used primarily in conjunction with a hybrid Artificial Neural Network (ANN) and Hidden Markov Model (HMM) speech recognition system. This work explores the viability of the RASTA-PLP as a candidate feature extraction method in the Siemens ZT recognition system. Experiments with a basic RASTA-PLP setup confirm that it provides good performance and is a potentially useful tool which merits further research and experimentation.

*This work was supported by and conducted at Siemens AG, München, Germany

## Contents

## List of Figures

## List of Tables

# 1 Overview

This report describes a series of experiments using the RelAtive SpecTral Analysis - Perceptual Linear Prediction (RASTA-PLP) feature extraction together with the Siemens ZT speech recognition system. First is a brief account of the experimental conditions along with a general description of the RASTA-PLP system and how it was incorporated into the Siemens recognition framework. The subsequent sections chronologically document recognition experiments and test results.

The first experiments explored various scaling and quantization methods using PLP, one of which was chosen for the remaining experiments. The next experiment provided preliminary tests using RASTA with and without delta and double delta feature coefficients. As RASTA with delta and double deltas provided the best results, the remaining banks of experiments used this feature configuration. Next was a set of tests with varying numbers of Hidden Markov Model (HMM) densities and of Linear Discriminant Analysis (LDA) coefficients. A question arose as to the inclusion of the energy parameter and a new set of tests without this parameter were done. Finally the tests were repeated using Mel Frequency Cepstral Coefficients (MFCC) to provide a reference.

## 1.1 Siemens ZT Recognition Framework

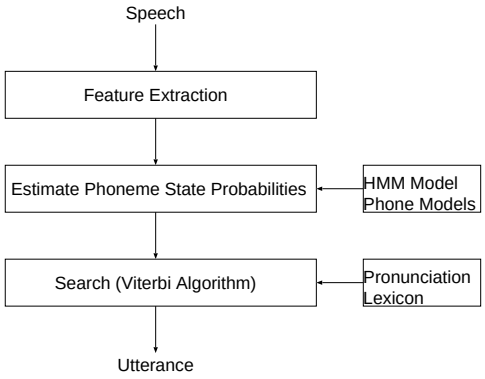


Figure 1-1: Recognition Architecture.

Figure 1-1 depicts the major processing steps for the typical Siemens ZT speech recognition framework. The first step was feature extraction. MEL features were normally used though in the experiments here, they had been replaced by RASTA-PLP features and also by MFCC. MEL and RASTA-PLP are briefly described in the following sections. The features were stored as 8 bit signed character values. MEL computed features using either float or integer calculations with the features automatically scaled, quantized, and clipped to fit into the 256 level dynamic range. RASTA-PLP on the other hand computed features entirely in float; hence additional

scaling, quantization, and clipping were added as a post operation on the features to fit them into the signed character range.

The recognition system employed a continuous density Gaussian mixture HMM. In the common system, phoneme state probabilities were modeled as a finite number of Gaussian densities:

$$p = \sum_{i=1}^{M} c_i N_i(x, \mu, \Sigma)$$

Here $N_i$ is a Gaussian distribution, $x, \mu, \Sigma$ are its input vector, mean vector, and covariance matrix respectively, and $c_i$ is the weight for each of the $M$ mixtures. To reduce the computational load some approximations were made. The summation was replaced with a maximum operation:

$$p = MAX\{c_i N_i(x, \mu, \Sigma)\}$$

Further, the covariance matrix was replaced with a diagonal matrix of variances. Hence, only the means and variances were estimated in the training process.

Each phone model consisted of a six state sequence with loops and skips as in figure 1-2. The exception to this was the silence model which consisted of a single state. Each phone was grouped into pairs yielding 3 segments per phone. For context independent models, the segments were labeled *phone.0*, *phone.1*, and *phone.2*. For context-dependent models, the segments were labeled *(previous phone)_phone_*, *phone*, *_phone_(following phone)*. See figure 1-3 for an example using "teeth." For decoding, Baum-Welch soft estimation was replaced by a Viterbi decoding. During training, labels were ascribed to time points. Each phone was initially divided equally into three segments to which the 3 segments of the phone models were assigned. The training algorithm was iterated to estimate a more accurate phone alignment. Training was done using a Maximum Likelihood optimization procedure.

### 1.2 Linear Discriminant Analysis

LDA is a data driven orthogonalization process designed to maximally discriminate classes by means of linear transformations. The goal is to develop transforms which maximally separate between classes while minimally separating within each class. The process contains three principal steps. First the vector space is rotated by a Karhunen-Loève (KL) transform which attempts to decorrelate features and align the directions of maximum variance with a new coordinate axis. This is followed by a whitening procedure which attempts to remove the effects of individual class cluster shapes. The final step rotates the KL-transformed and whitened space such that the directions of maximal separability align with the coordinate axes for the space. The end result of the final step is that the resultant feature vectors have elements which are ordered by effectiveness in ability to discriminate between classes [3].

In these experiments, LDA was applied to a supervector consisting of two frames of speech features. i.e. at each time frame, the current feature vector and the previous one were concatenated and the LDA was applied to the concatenated supervector. After LDA transformation, a given number of feature coefficients was discarded. Typically, the number of retained coefficients was less than the original feature vector dimension. An additional implementation specific item was an imposed limit on the whitening step of the LDA: Class clusters which had variances less than one thousandth of the maximum variance were not whitened. This was to counteract computation errors resulting from constant or near-constant feature components.

### 1.3 MEL Feature Extraction

Figure 1-4 shows the basic MEL feature extraction processing steps. First an FFT applied to a Hamming windowed frame of speech provided the power spectrum. A subsequent logarithm followed by Mel frequency scaling and a cepstral smoothing filterbank yielded band energies. The band energies were normalized by subtracting off the mean of coefficients for each frame. This subtraction was not done with the MFCC processing. The frames were then run through a channel compensation filter, effectively a high pass filter. Deltas and double deltas were finally added. MFCC had the added step of cepstral transformation. The MEL process gave 24 coefficients plus energy. To compute the deltas, each of the 24 coefficients were averaged in pairs giving 12 additional coefficients. A finite backwards difference gave the delta calculations and an additional finite forward difference gave the double deltas. 24 features plus energy, 12 deltas plus delta energy, and 12 double deltas plus delta delta energy, made 51 features in all.

After the deltas were calculated, the features were optionally warped by Linear Discriminant Analysis (LDA) matrices prior to quantization into the 8 bit signed character range. The necessary LDA transformation matrices were computed offline separate from the training and testing procedures.

### 1.4 RASTA-PLP Feature Extraction

The RASTA-PLP feature extraction method is similar to other feature extraction methods such as the MFCC and is also similar to the MEL features in use at Siemens. The basic processing operations are shown in figure 1-5. The principal differences resided in the addition of the human auditory model scaling and the shape of the RASTA bandpass filter. The human auditory model scaling warped the spectrum and also did some frequency smoothing so that the number of coefficients necessary for fair performance was reduced with respect to the MEL processing. While the channel compensation in the MEL processing was a high-pass filter, the RASTA filter which performed a similar function was bandpass in nature. Explicit frame normalization was also not done in RASTA-PLP. The other principal difference was that the cepstral transformation was absent in the MEL computation. Although RASTA-PLP has a

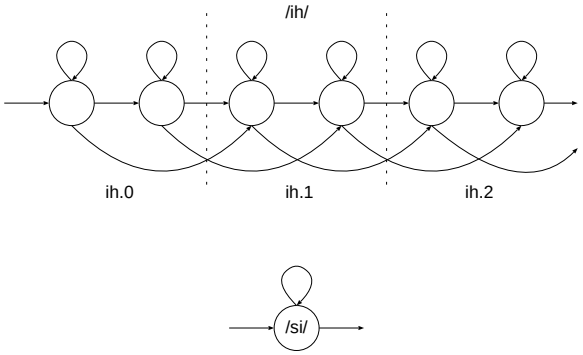


Figure 1-2: Context independent model examples for /ih/ and silence model /si/.

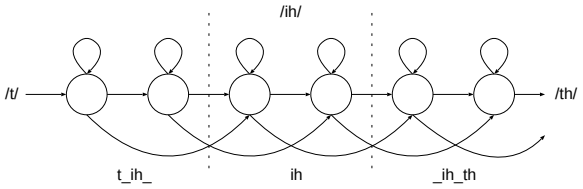


Figure 1-3: Context dependent model example for /ih/ in "teeth."

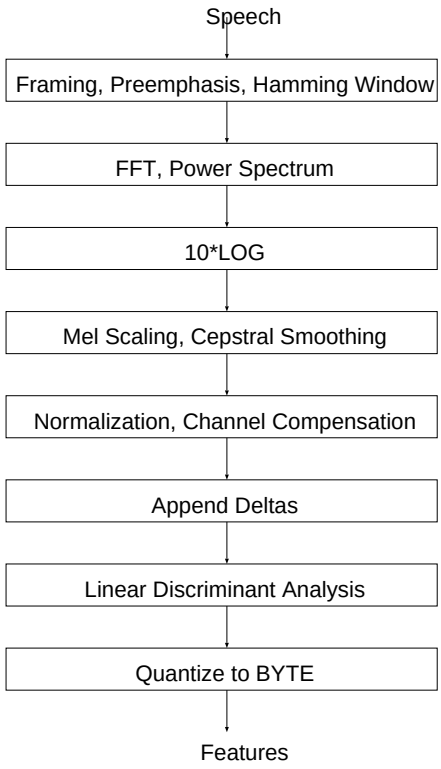


Figure 1-4: Main processing steps for the MEL feature extraction method.

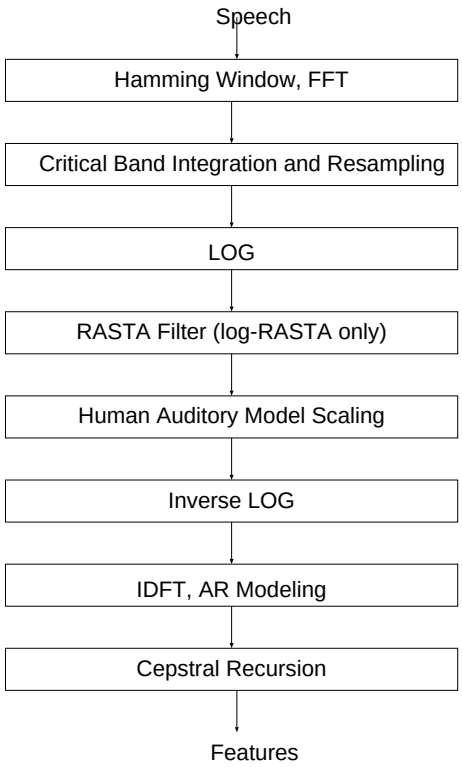


Figure 1-5: Main processing steps for the RASTA-PLP feature extraction method.

mechanism for cepstral mean subtraction, this was not done in the experiments in this report; process was kept at near-default settings. The MEL processing on the other hand did include spectral mean subtraction and this contributes to yet another difference between the results in this report. This difference is absent with respect to the MFCC. A more detailed description of the RASTA-PLP processing can be found in [1, 2].

It is worth noting that the delta calculations were also different between the RASTA-PLP and MEL calculated features. RASTA-PLP deltas were calculated using a linear regression over a 9 frame window. MEL deltas were calculated using a finite single step backward difference (for the deltas) and an additional finite single step forward difference (for the double deltas). Furthermore, the deltas and double deltas were averaged pairwise producing additional sets of coefficients each half the dimension of the original feature vector.

There are three principal types of processing available in the RASTA-PLP software: PLP, Log-RASTA, and Jah-RASTA. Experiments in this report only used the PLP and Log-RASTA types of processing. In the following, experiments with PLP used the first type of processing and experiments noting use of RASTA used only the second type. RASTA-PLP then refers to the complete general feature extraction method. PLP differs from Log-RASTA by the lack of filtering of the spectra in the logarithm domain. This log-spectral filtering was introduced to add robustness to convolutional noise. Jah-RASTA (not deployed) attempts to handle both additive and convolutional noise.

As RASTA-PLP was a separate entity to the Siemens recognition framework at the time of the experiments, some additional processing steps were introduced to allow the recognizer to use its features. The first additional process was the scaling and quantization of the produced features. The Siemens recognizer was hard-coded to use 8-bit signed character range features, and some of the processing was done in integer as well. Because RASTA-PLP computed and produced features in floating point it was necessary to scale and quantize the features into the 8-bit character range. The first series of experiments lightly explored types of scaling parameters. The second additional processing was LDA computation. The LDA for the MEL features was an integral part of its computation and was not an original part of the RASTA-PLP processing. LDA was therefore computed using separate programs provided by Josef Bauer that operated on the stored feature files.

A possible shortcoming of the non-inclusion of LDA in RASTA-PLP was that the quantization effects took effect before the computation of the LDA matrices. In other words, the LDA matrices were not computed using the original floating point values, but rather on the features after clipping and quantization. Similarly, the LDA transformation operated on the quantized RASTA-PLP coefficients in producing the LDA features. This is in contrast to the LDA matrix computation and transformation for the MEL features which where internally computed in floating point. This added still another possible source of variability in the observed performance.

Strictly speaking, the collection of mis-matches disallow any fair direct comparison between RASTA-PLP and MEL/MFCC tests. Further the restricted use of RASTA-PLP (i.e. lack of J-RASTA tests, no cepstral mean subtraction, etc.) further limited the competitiveness of this feature extraction method. The results for both are shown for expository purposes only and serve merely as an indication of performance in relation to the original standard and tuned feature extraction method. The goal of these experiments was to demonstrate the viability of RASTA-PLP as a plausible feature extraction method in conjunction with the Siemens ZT Recognizer.

### 1.5 Experimental Conditions

The VM American English Digits database served as the training corpus for all of the following experiments. It was composed of the numbers zero through nine and the word *oh*. The training set contained 7423 utterances and the test set contained 2228 utterances from mutually independent sets of speakers. Additionally cross database experiments were performed using isolated digits from the Bellcore and Macrophone databases. The Bellcore test set contained 2200 utterances and the Macrophone contained 782 utterances. As this was an isolated word recognition task, insertions and deletions were non-existent and all occurring errors were substitutions. Therefore, recognition performance is uniformly reported as percent correct with the understanding that the implied error rate is entirely due to substitutions.

## 2 Feature Scaling and Quantization Experiments with PLP

The Siemens speech-recognition system, for historical reasons, processed speech feature files that had been quantized to the range of 8 bit signed characters. The MEL features included fixed scalings and quantizations to produce 8 bit ZFE formatted feature files. The first obstacle for testing the RASTA-PLP features was then to scale and quantize the RASTA-PLP features into the range of 8 bit signed characters. This section explores a few different varieties of scaling methods and selects a particular scaling for the remaining experiments.

The experiments in this section used strict PLP features alone with no delta features added. The features included a single gain feature and 8 PLP cepstral coefficients for a total of 9 coefficients per feature vector. The HMM recognizer used the context-independent phone models and an arbitrary number of Gaussian mixtures. After initialization, the HMM Viterbi training was iterated five times before recognition.

LDA was applied to the quantized 8-bit ranged PLP features. The necessary LDA transformation matrices were computed from these quantized features as well. After LDA computation with a supervector of 2, the 9 most significant features were retained.

| feature | max | min | avg | std (n=934135) |
|---|---|---|---|---|
| 0 | 6.48875 | 1.64749 | 2.67646 | 0.93661 |
| 1 | 0.37246 | -1.23896 | -0.42983 | 0.15469 |
| 2 | 0.63651 | -1.39489 | -0.34395 | 0.21509 |
| 3 | 0.46590 | -1.36073 | -0.33973 | 0.15771 |
| 4 | 0.61487 | -1.40495 | -0.30803 | 0.19148 |
| 5 | 0.80038 | -0.87748 | -0.16324 | 0.11600 |
| 6 | 0.61108 | -0.76360 | -0.10962 | 0.10948 |
| 7 | 0.62558 | -0.65672 | -0.02765 | 0.10962 |
| 8 | 0.62680 | -0.48057 | 0.00777 | 0.08319 |

Table 2-1: Statistics of the PLP features for the VMAE training set.

## 2.1 Scaling and Quantization Methods

The following experiments explored five different types of scaling and quantization. The first was a direct linear scaling. The later ones involved scaling based on the statistics of the feature set. For example, table 2-1 lists the maximum, minimum, mean, and standard deviation of the PLP features for the VMAE training set.

### 2.1.1 Direct Linear Scaling

This first type of scaling was a naive direct linear scaling of all of the features with a single constant. 8 bit signed characters offered a range of 256 quantization levels. Scaling the features better extended the feature distribution over these quantization levels. After the scaling, the float values were truncated to integer and stored as character ZFE feature format. Values that exceeded the limits were clipped to the maximum and minimum range values.

From table 2-1, all of the features for the VMAE training set had a magnitude less than 7. Further, all with the exception of the gain feature fell within a magnitude less than 1.5. Multiplying all features with a single constant could therefore extend the dynamic range of the features while preserving the relative between-feature magnitude differences. However since each feature had a different distribution, a single scaling constant yielded more dynamic range for some features than for others. The subsequent scaling methods attempted to remedy this.

### 2.1.2 Maximum-Minimum Scaling

In this scaling, the dynamic range was set to the maximum and minimum values determined from the training set. Therefore, the maximum feature value was assigned to 127 and the minimum feature value was assigned to -128 with a linear partitioning

for values in between (figure 2-1). Again, the float values were truncated to integer for quantization. Each feature was adjusted individually according to its own respective maximum and minimum. Values in the test set outside of this range were clipped to highest or lowest value permitted (i.e. 127 or -128).

### 2.1.3 Mean-Std Scaling

The maximum-minimum scaling sought to confine the whole of each feature distribution inflexibly into the character range. Hence, less dynamic range was partitioned to the center bulk of each feature distribution and more was allotted to the tails of the distribution. Since the importance of allowing more quantization levels to the tails versus the center was not yet clear, the mean-std scaling method was contrived to allow more flexibility. In this method, the dynamic range was set to a chosen number of standard deviations from the mean. The values 127 and -128 were assigned to a chosen number of standard deviations above and below the mean respectively. As before, each feature was scaled independently using its own statistics and its range was partitioned in a linear fashion with outlying values clipped to the limits of the signed character range.

### 2.1.4 Arctangent Scaling

For the mean-std scaling method, the chosen number of standard deviations granted more dynamic range either to tails or the center bulk of the feature distribution. As a compromise, the arctangent scaling method was contrived as an experiment in balancing the two. The previous mean-std scaling without the clipping and quantization was followed by an arctangent operation (figure 2-3). The limits of an arctangent are $-\pi/2$ to $\pi/2$ and these were fixed to -128 and 127 respectively. Near 0, the arctangent function has a linear characteristic and has a compressing behavior at higher argument values. Depending on the scaling factor, the arctangent partitions the center of the distribution in a roughly linear fashion and the tails are compressed. Thus the entire distribution is smoothly squeezed into the allotted character range before quantization to integer values.

### 2.1.5 Histogram Equalization

In this scaling experiment, the feature distribution was mapped roughly to a uniform distribution (figure 2-4). The 255 levels were partitioned out over the distribution such that each level had the same number of feature examples in its bin from the training set. A side effect of this was that the center bulk of the feature distribution was allotted more dynamic range than the tail ends. Each feature was equalized individually according to its own distribution.

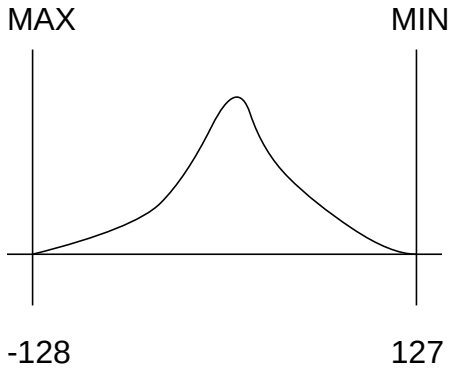


Figure 2-1: Maximum-minimum scaling.

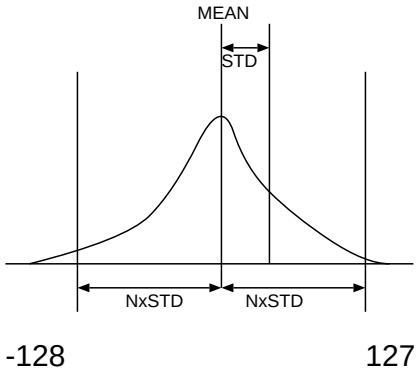


Figure 2-2: Mean-Std scaling.

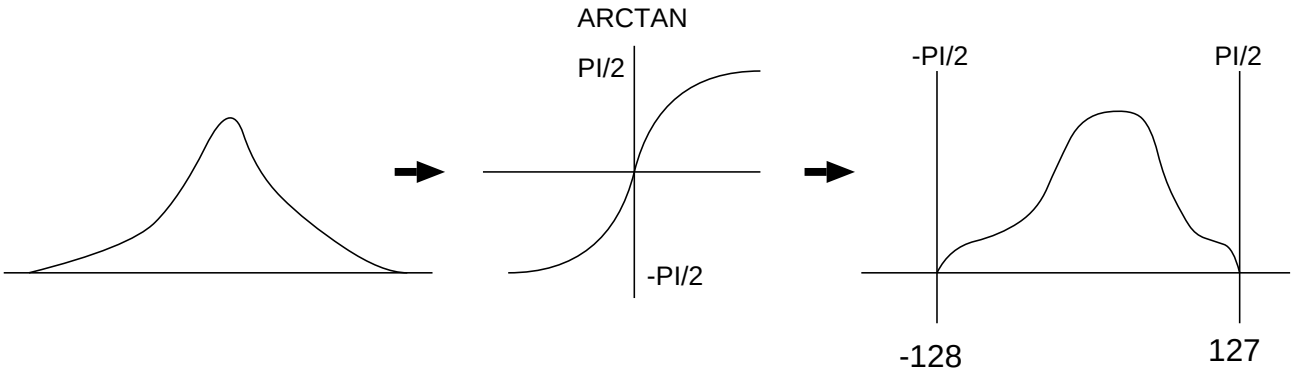


Figure 2-3: Arctan Scaling.

### 2.2 Experimental Results

Figure 2-5 lists the results of experiments with all of the scaling conditions. The first item lists the baseline recognition results with MEL using 51 coefficients with no LDA and 24 coefficients with LDA. The next 3 items (labeled as LIN50, LIN100, and LIN150) chart results using the naive linear scaling with scale factors of 50, 100, and 150. Following this is the result for the minimum-maximum scaling (labeled as MIN/MAX in the chart). The items marked STD5 and STD9 denote the results for the mean-std scaling where the number of standard deviations that impose the maximum limit were set to 5 and 9 respectively. The HISTEQ item is the result for the histogram equalization experiment. The final 3 items (marked ATAN1, ATAN3, and ATAN5) display the results for the arctan scaling with a prescale of 1, 3, and 5 standard deviations respectively.

On first inspection, the quantized PLP features performed satisfactorily with most results reaching into the 90 percent range. The highest result, 95.2% for STD9 scaling with LDA, was only 3.5% below baseline for this task. This was with only 9 PLP coefficients without deltas and keeping only 9 coefficients after LDA. By observing the effects of adjusting the scaling parameter from 5 to 9 for the mean-std tests and from 1 to 3 to 5 in the arctangent tests, the tails of the distribution seemed important for good LDA results. However observing the scaling parameter move from 50 to 100 to 150 in the linear tests, scaling much beyond the tails began to retard the performance a little for the LDA tests. Additionally, the histogram equalization performed poorly, likely because most of the dynamic range was partitioned principally among the centers of the distributions.

As the best result was with the mean-std scaling, the remaining experiments in this work used this means of scaling and quantization but with a scaling parameter of 10 instead of 9.

## 3 Preliminary Experiments with RASTA

The experiments in this section explored the performance of Log-RASTA. RASTA was tested both with and without LDA and optionally with the addition of delta and double delta feature calculations. Additionally, some tests with a varying number of LDA coefficients are presented. The Mean-Std scaling and quantization technique with a scaling parameter of 10 standard deviations was used to constrict the RASTA features to 8 bit signed character features. The means and standard deviations were recomputed over the VMAE training set using Log-RASTA processing.

Eight RASTA coefficients and gain (9 feature coefficients in all) served as the core features for the experiments here. The delta coefficients were computed individually for each feature and the gain. RASTA with deltas contained 18 features and RASTA with deltas and double deltas contained 27 features. The LDA computation used a supervector of 2 frames but only the same number of feature coefficients were kept

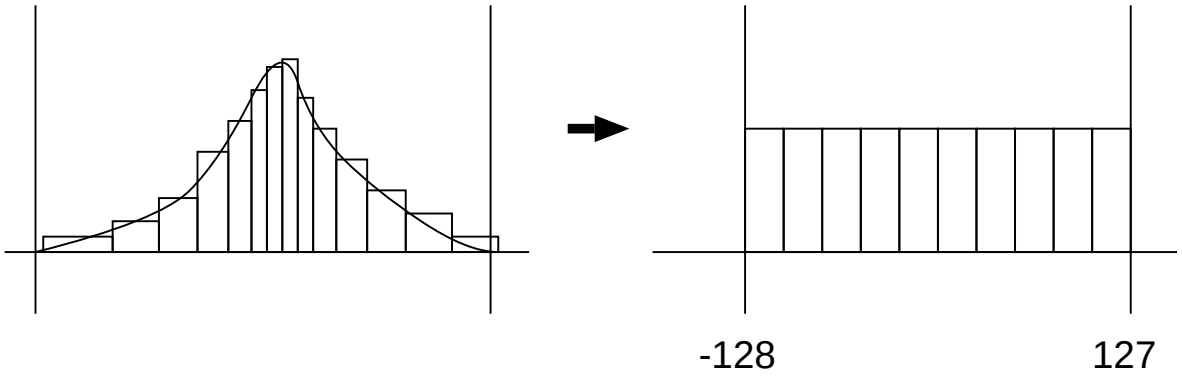


Figure 2-4: Histogram Equalization Scaling.

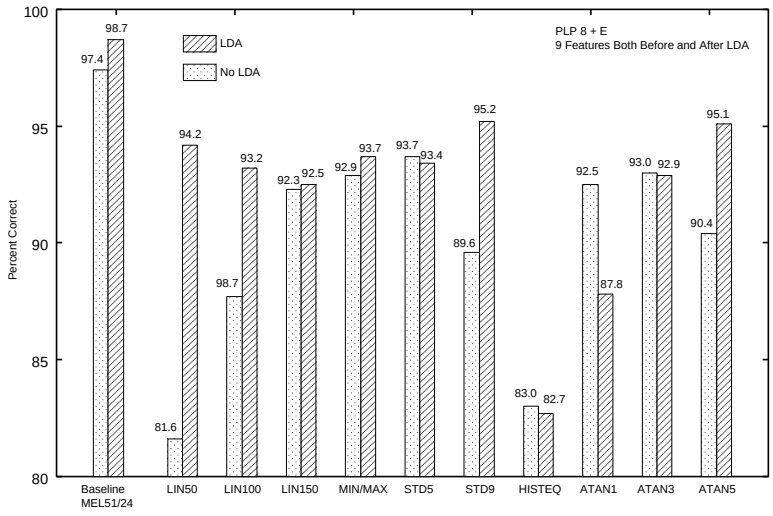


Figure 2-5: Results of Scaling Experiments.

as with before the LDA computation. Therefore, RASTA without deltas contained 9 features both before and after LDA, RASTA with deltas had 18, and RASTA with deltas and double deltas had 27. This is in contrast to the MEL features which had 51 features without LDA and 24 with LDA.

### 3.1 Tests with Addition of Deltas

Two sets of tests were conducted and are presented here. In each, the RASTA features were tested alone and then succeedingly with deltas and double deltas. The first test used context-independent monophone models and the second test set used context dependent diphone models. Each experiment involved a completely new training and testing given the particular number of features. The HMM was allowed an unspecified number of Gaussian mixture models for training. The training algorithm determined the number of models used. The number of mixtures arrived at by the training algorithm is mentioned in the results in the corresponding figures of each test.

#### 3.1.1 Monophone Models

The results for the monophone model tests are presented in figure 3-1. The first item indicates the results for the baseline experiments. The remaining items indicate the RASTA tests with deltas indicated by the "d" and "dE" and double deltas indicated by the "dd" and "ddE." The number of mixtures that the HMM training algorithm selected is printed on the bar for the respective experiments. While not equivalent, the numbers of mixtures for each test set fell into the same order of magnitude with a maximum difference of only 4.5% of the lowest value. We assume that here the difference in the number of mixture models is negligible for comparing the best attainable scores. This difference between numbers of mixtures is corrected for in later experiments.

The succeeding addition of deltas and then double deltas improved recognition performance. Additionally, tests with LDA performed better than those without LDA. The 9 RASTA features without deltas performed adequately for this task with the LDA. The addition of the deltas improved this performance significantly. Without the LDA and with 18 features, the performance was superior to the 51 baseline MEL features without LDA and not significantly lower than the 24 baseline MEL LDA features. Addition of the deltas and the double deltas improved results further. The 27 RASTA features both with and without LDA exceeded the respective baseline results. The differences, though, in maximum scores with LDA were not statistically significant. Again, the 27 RASTA features without LDA performed almost as good as the baseline LDA score.

One additional experiment was conducted which is not included in figure 3-1. Since LDA operated on a supervector of two consecutive frames, it had the ability to incorporate delta information into the feature set without explicit delta calculation. RASTA was then tested with the base feature and the double deltas but without

the deltas. With 2252 mixtures, it scored 96.1% without LDA. With 2218 mixtures, it scored 98.7% with LDA. The use of base features and double deltas then scored better than RASTA with deltas but not as well as RASTA with explicit deltas and double deltas. With LDA, however, the differences were not statistically significant.

### 3.1.2 Diphone Models

The results for the diphone model tests are presented in figure 3-2 with items corresponding as in the monophone model tests. Again, the number of mixtures differ but remain on the same order of magnitude. The use of diphone models raised the scores of all feature combinations, indicating that use of context-dependent models is superior to the use of context-independent models. As with the monophone case, the succeeding addition of deltas and double deltas improved the recognition performance very significantly over no deltas. In this best case however, RASTA with deltas and double deltas LDA scored equal to the baseline LDA score. The additional test with 24 LDA features instead of 27 for RASTA shows an equal score and indicates that the recognizer had reached or nearly reached its upper limit in performance. We also note that the difference between this maximum score and RASTA with full deltas and without LDA was not statistically significant. For this task it appears that RASTA may have served adequately well without the extra LDA computation.

## 3.2 Varying the Number of LDA Coefficients

As the LDA orders the relative "importance" of each feature coefficient for discrimination purposes, it became instructive to vary the number of retained coefficients to note the effects on recognition performance. Starting with the "best" feature combination of RASTA with deltas and double deltas and LDA, we lowered the number of retained coefficients from 27 to 24, 18, and 9. This was a sparse sampling but offered a glimpse at the discriminative weights of the ranked retained coefficients. A heavier sampling was pursued in later experiments.

This experiment used the monophone models. The scores are displayed in figure 3-3 along with the baseline score. As indicated, the RASTA performance was maintained down to 18 LDA features and the difference in scores for 27, 24, and 18 together with the baseline was not statistically significant. Further, the performance of RASTA with 18 retained coefficients was almost the same as the 24 coefficient baseline score.

# 4 RASTA Experiments

This series of tests served to provide a broader mapping of the basic performance of RASTA. The test cases included same corpus tests, additive noise tests, and cross database tests. Each series included tests which examined the performance using a

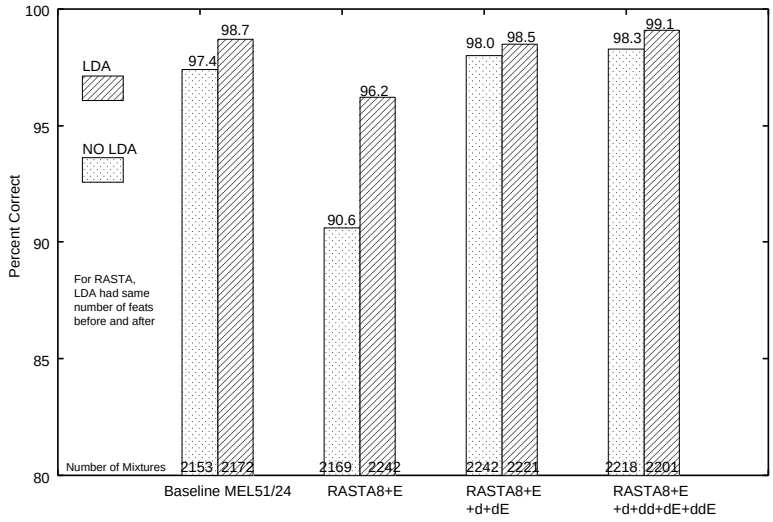


Figure 3-1: Monophone model RASTA experiment results.

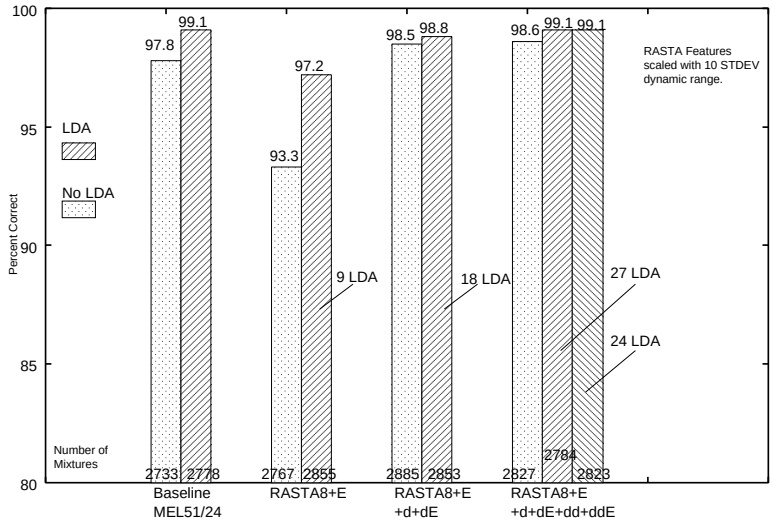


Figure 3-2: Diphone model RASTA experiment results.

varying number of target HMM Gaussian mixtures. This was done both with and without LDA. Additionally, they included tests using a varying number of LDA coefficients for a specific number of target densities. From the previous section, RASTA with deltas and double deltas in conjunction with diphone models for the HMM provided the highest performance. This combination was therefore used for all of the following experiments in this section.

The training regimen differed from that in the previous sections. We adjusted a few HMM parameters and modified the training procedure as follows: The HMM, once initialized, was iterated 5 times with an unspecified number of target densities; the HMM training algorithm automatically selected the number of densities it would use. It was then iterated 3 times to the first and highest target of 2000 densities. A test was then completed for this specified number of target densities. For each remaining target number, the HMM was iterated 3 more times to the next lower desired target number with a test after each series of iterations.

### 4.1 Same Corpus Experiments

In this experiment, the training and testing utterance were both from the VMAE isolated digits corpus. There were 7423 training utterance and 2228 test utterance from mutually exclusive speakers. Figure 4-1 shows the percent correct scores for a range of target numbers of mixtures. Performance was not severely impaired until the number of target densities fell below 500. Inclusion of LDA alleviated some of this performance loss. Figure 4-2 shows the percent correct varying the number of retained LDA features using an HMM with 1000 Gaussian mixtures. Here reducing the number of features did not impair the performance to a great degree until the number of coefficients was below 15. Recognition rate was relatively high even with 9 coefficients.

### 4.2 Added Car Noise Experiments

Here, car noise was added to the test utterances with an SNR of 10db and 0db. The car noise came from the NOISEX database. The noise was only added to the test utterances and not to the training utterances. This tested the robustness of the features and the recognizer to noise. The LDA matrices were also kept identical to those arrived at in the clean speech condition from the previous subsection. Figure 4-3 indicates, as expected, that the recognition took a performance hit as the noise increased. Additionally, the addition of LDA in most cases lowered the recognition rate even further, though not by a statistically significant amount. Note that Log-RASTA was designed to handle convolutional noise. Jah-RASTA, which was designed to also handle additive noise, may have shown an improvement.

Figure 4-4 documents the performance of the recognizer as fewer and fewer coefficients were retained after LDA. Again, the HMM was fixed to use 1000 mixture

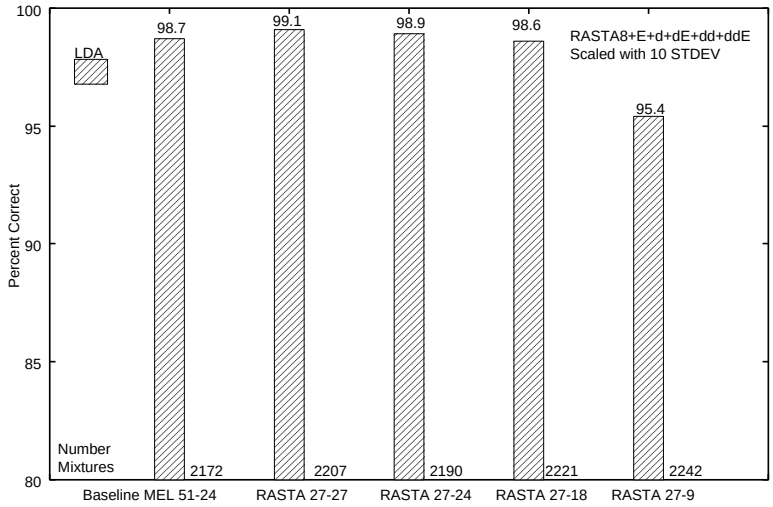


Figure 3-3: Results for varying number of retained LDA coefficients.

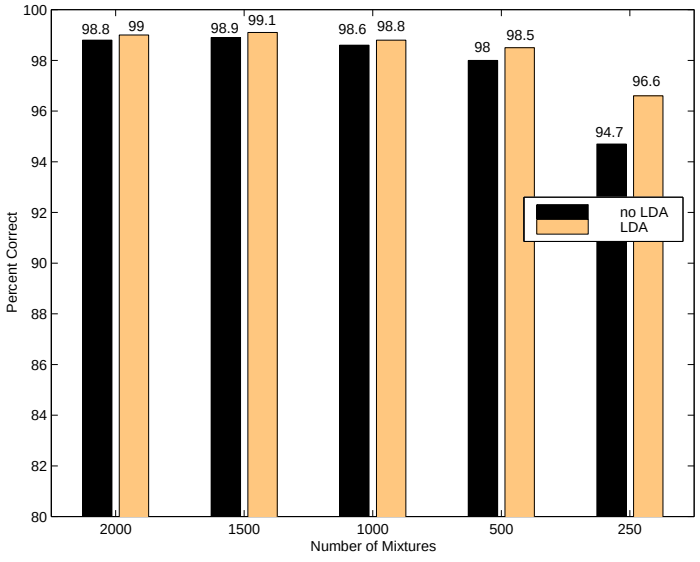


Figure 4-1: Results of varying the number of target mixtures, same corpus. (8 RASTA Coefficients + E + d + dE + dd + ddE)

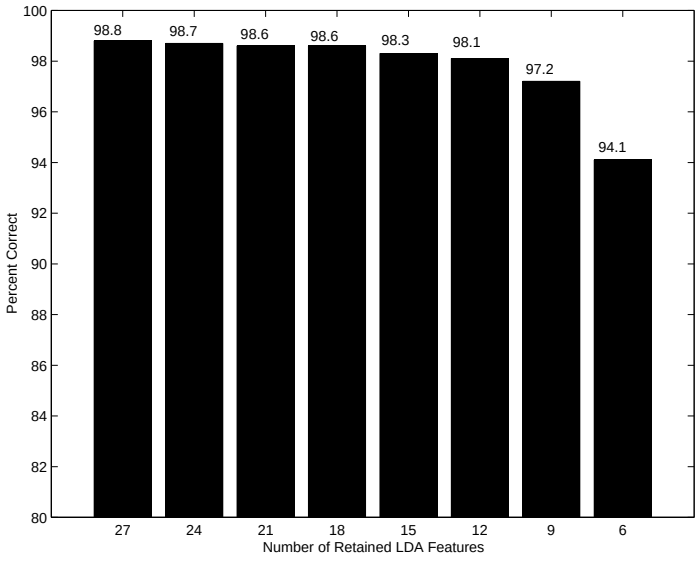


Figure 4-2: Results of varying the number of LDA features at 1000 mixtures, same corpus. (8 RASTA Coefficients + E + d + dE + dd + ddE)

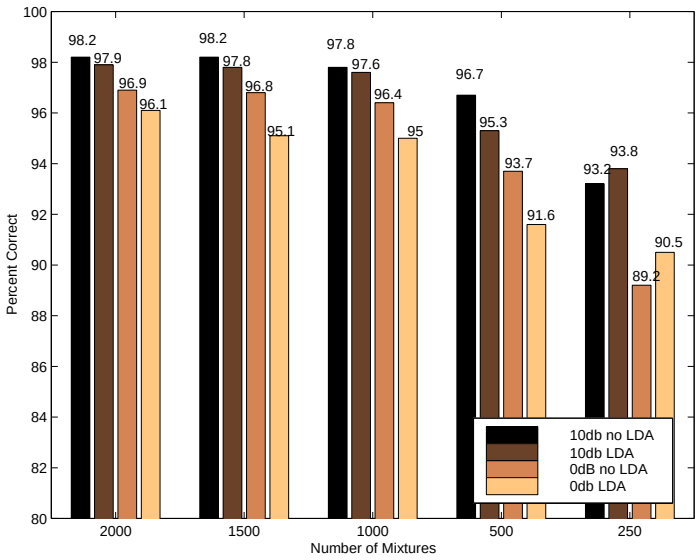


Figure 4-3: Results of varying the number of target mixtures, noise conditions. (8 RASTA Coefficients + E + d + dE + dd + ddE)

prototypes. Eliminating up to two-thirds of the features did not drastically deter adequate performance.

### 4.3 Cross-Database Experiments

Cross database tests shed light on the robustness of the feature extraction to different speakers and different recording conditions. Figure 4-5 displays test results using digits from both the Bellcore and Macrophone databases. The HMM was kept trained on the VM American English digits training set with the LDA transformation matrices kept the same as well. Again, the performance was smaller but not as appreciably as in the noise condition tests. RASTA maintained an adequate level of robustness for these tasks. Addition of LDA gave mixed results. For the Bellcore tests, LDA improved performance. For the Macrophone tests, LDA decreased performance. Varying the number of retained LDA features (figure 4-6) again showed that as much as two-thirds of the features could be eliminated without seriously degrading the recognition rate.

## 5 RASTA Experiments Without Gain Feature

The results in the previous section revealed a reduction in the effectiveness of LDA in conjunction with RASTA for the cross database tests and the noise condition tests. Changes in database and additive noise conditions often affect the energy parameter significantly, due in part to different recording conditions and additional noise energy respectively. Possibly the energy parameter hampered the performance. In these experiments, the energy parameter was eliminated from the feature vectors. The delta energy and the double delta energy were retained. In all other respects, the experiments here were precisely the same as those in the previous.

Figures 5-1, 5-3, and 5-5 show the results of varying the number of target mixture densities with the energy coefficient eliminated. Similarly, figures 5-2, 5-4, and 5-6 show the performance varying the number of retained LDA feature coefficients using a 1000 mixture HMM. The performance curves were similar to those in the previous section in that fewer mixture densities slowly decrease performance as did retaining fewer feature coefficients.

Comparing with results of from the last section, the performance was generally maintained or increased by the elimination of the energy feature in tests without using LDA. This effect often did not become noticeable until the number of mixture densities was low to around 500. With LDA, the elimination of energy generally faired poorly particularly in the noise condition and cross database tests. This remained true until the number of retained LDA coefficients became small (12 coefficients and fewer).

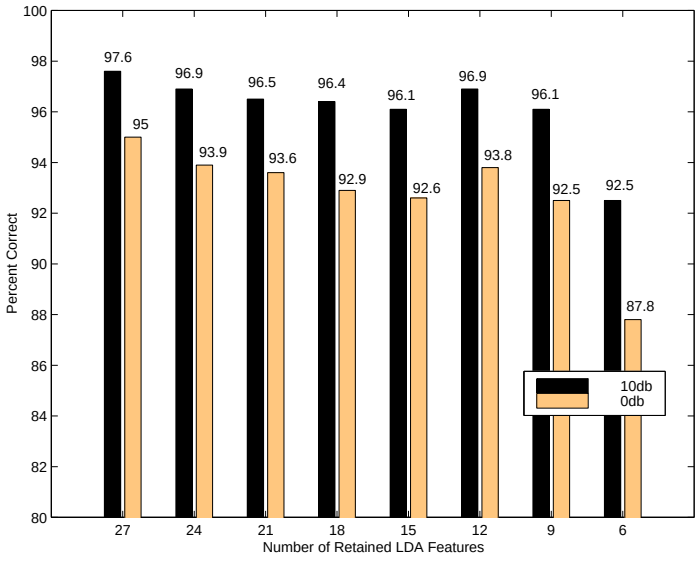


Figure 4-4: Results of varying the number of LDA features at 1000 mixtures, noise conditions. (8 RASTA Coefficients + E + d + dE + dd + ddE)

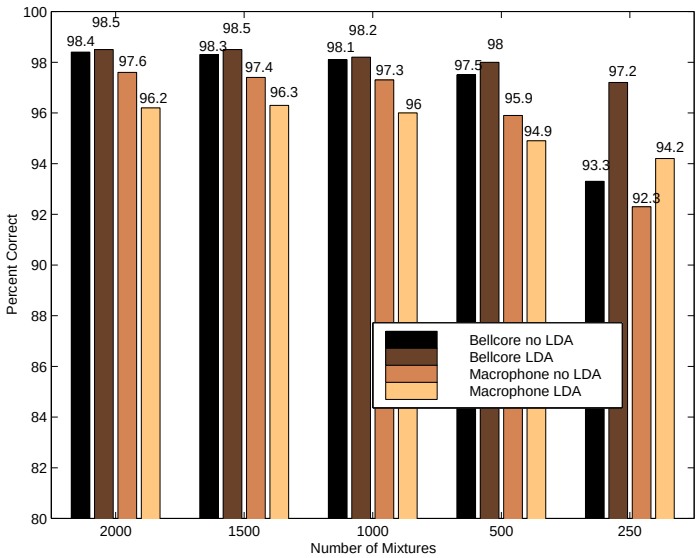


Figure 4-5: Results of varying the number of target mixtures, cross database. (8 RASTA Coefficients + E + d + dE + dd + ddE)

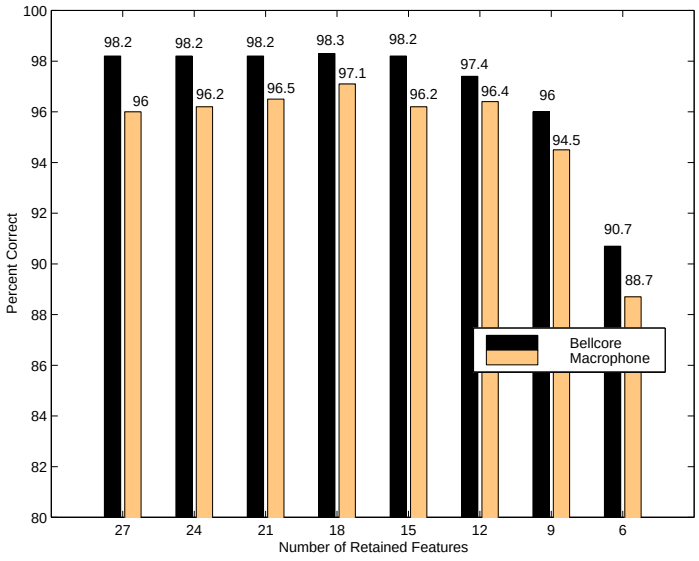


Figure 4-6: Results of varying the number of LDA features at 1000 mixtures, cross database. (8 RASTA Coefficients + E + d + dE + dd + ddE)

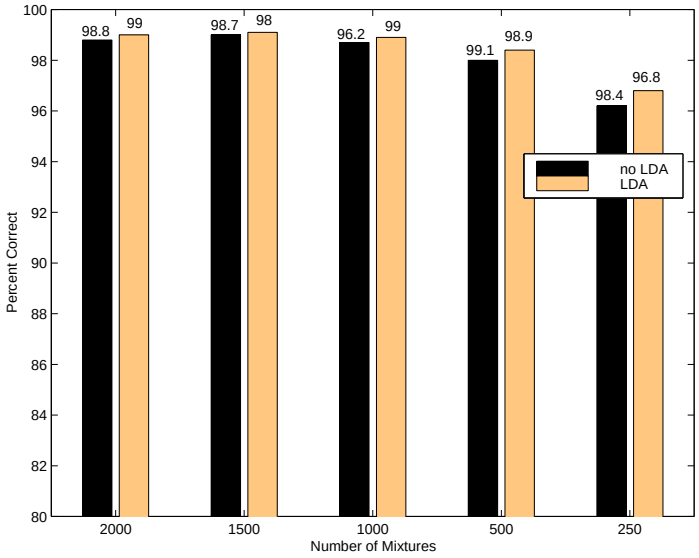


Figure 5-1: Results of varying the number of target mixtures, same corpus. (8 RASTA Coefficients + d + dE + dd + ddE, no energy)

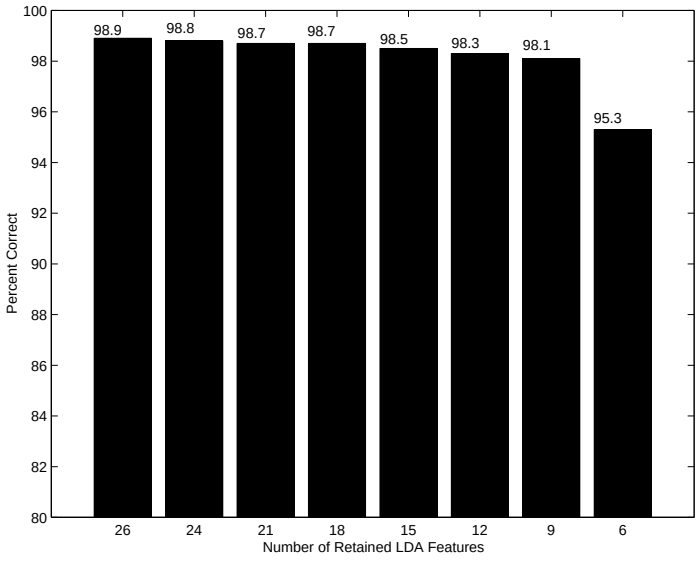


Figure 5-2: Results of varying the number of LDA features at 1000 mixtures, same corpus. (8 RASTA Coefficients + d + dE + dd + ddE, no energy)

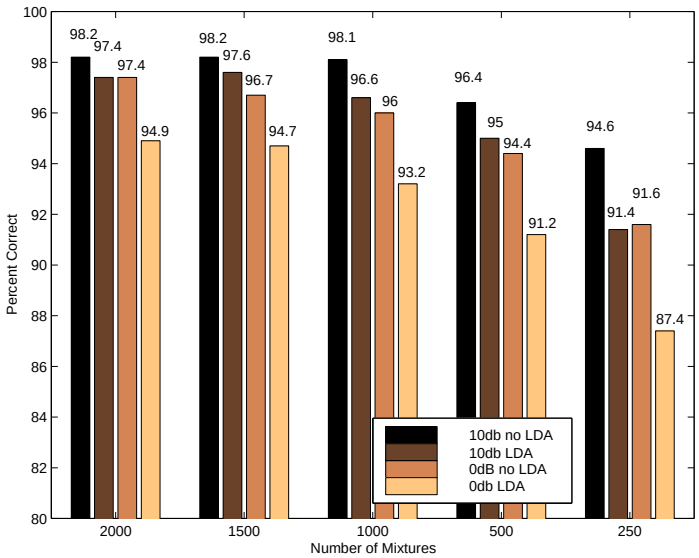


Figure 5-3: Results of varying the number of target mixtures, noise conditions. (8 RASTA Coefficients + d + dE + dd + ddE, no energy)

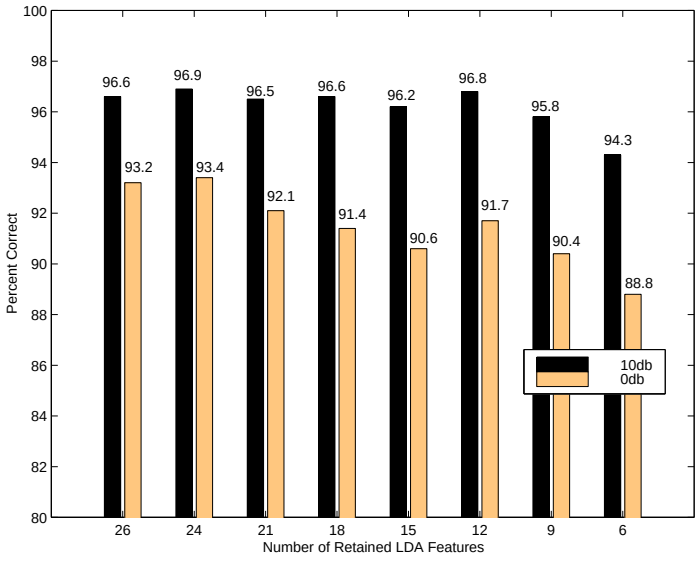


Figure 5-4: Results of varying the number of LDA features at 1000 mixtures, noise conditions. (8 RASTA Coefficients + d + dE + dd + ddE, no energy)

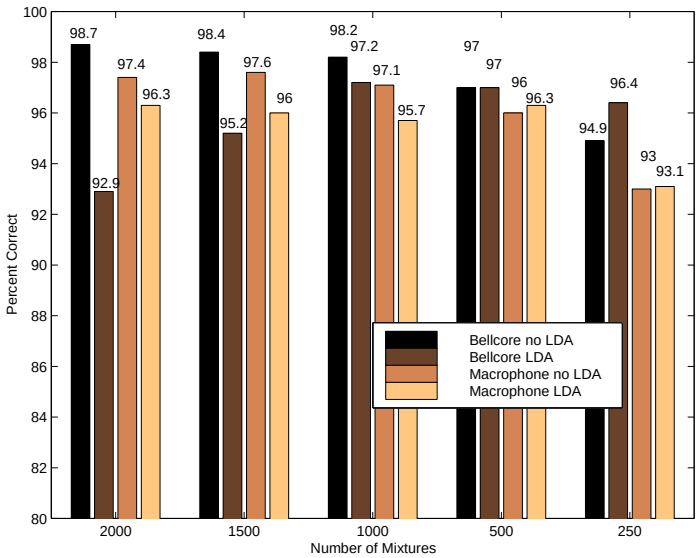


Figure 5-5: Results of varying the number of target mixtures, cross database. (8 RASTA Coefficients + d + dE + dd + ddE, no energy)

## 6 MFCC Experiments

The experiments in this section, again mimic those of sections 4 and 5 with the RASTA features replaced by Mel Frequency Cepstral Coefficients (MFCC). Rather than the MEL features typically used at Siemens, MFCC's were employed to provide a mechanism for comparison with features closer to "standard" MFCC features used frequently in the field. The main addition to MFCC over MEL was simply the addition of cepstral coefficient computation. Also MFCC did not include the MEL step of frame normalization through spectral mean subtraction. There was a slight modification to the training procedures for these experiments. In the previous sections, the HMM was initialized and iterated to an arbitrary number of densities before further iterations to the highest number of target mixtures (2000 mixtures). Here, the HMM was initialized and iterated directly to the highest number of target mixtures. Remaining iterations proceeded identically to those of the previous sections. The MFCC features consisted of 9 base features (energy and 8 cepstral coefficients) with deltas and double deltas computed for each of the base features independently.

Figures 6-1 through 6-6 display results for tests similar to those of the previous sections and with similar trends in performance with respect to decreasing numbers of mixtures and LDA coefficients. RASTA generally performed better than the MFCC baseline system in tests without LDA computation. The chief exception to this was the test with 0db additive noise. MFCC with LDA, however, tended to outperform RASTA with LDA, particularly in those experiments where LDA decreased RASTA's performance. The performance of MFCC fell below the performance of RASTA in cases of a low retained number of LDA coefficients.

## 7 Discussion

Figures 7-1 through 7-9 recapture the results of the previous section. They combine the results for the three feature configurations and are divided according to testing condition. It is worth noting that as either feature coefficients or numbers of mixture are decreased, the performance should degrade monotonically and gracefully. This property was not accurately pictured in the plots as there were sometimes dips and rises in the performance curve counter to what was expected. For most cases, the dips and rises were not significantly different and can be attributed to random fluctuation. This likely arose from the training procedure which iterated by fixed amounts and contained no cross-validation procedure. Therefore there were probably instances where models were undertrained or overtrained to the training set.

The first three figures (7-1, 7-2, and 7-3) show the results for the tests that vary the number of target mixtures without using LDA. For higher numbers of target mixtures, RASTA, both with and without the energy feature, performed consistently as well or better than the baseline MFCC coefficients. This was particularly true for the cross database tests and the same corpus tests. The RASTA advantage, however decreased

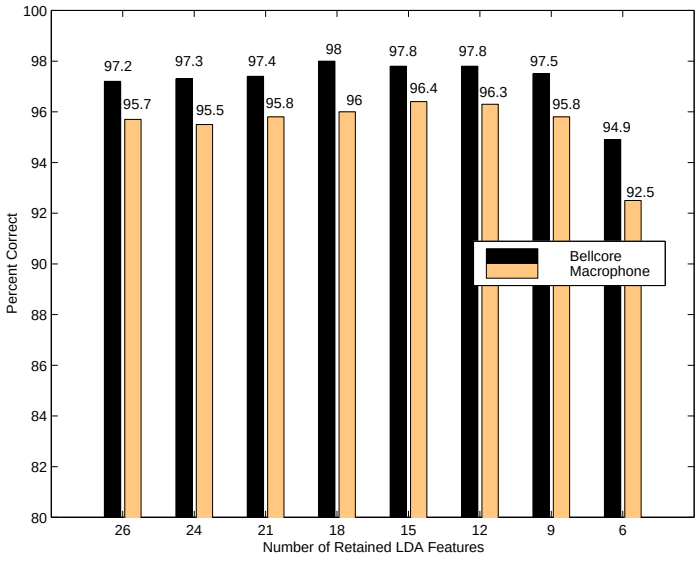


Figure 5-6: Results of varying the number of LDA features at 1000 mixtures, cross database. (8 RASTA Coefficients + d + dE + dd + ddE, no energy)

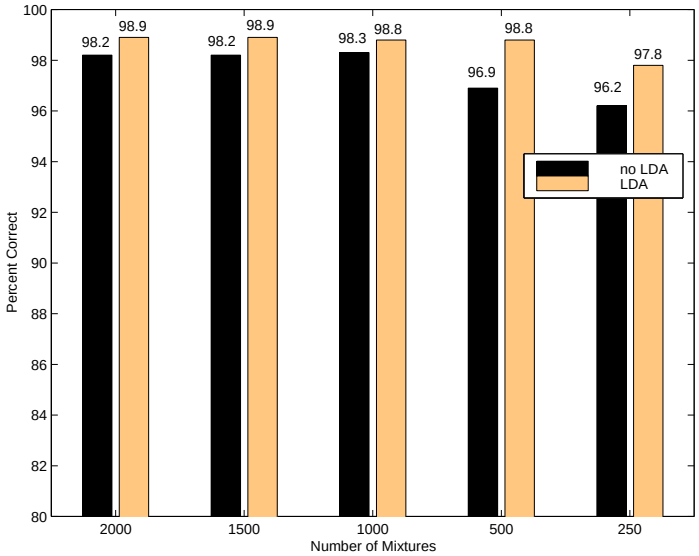


Figure 6-1: Results of varying the number of target mixtures, same corpus. (8 MFCC + E + d + dE + dd + ddE)

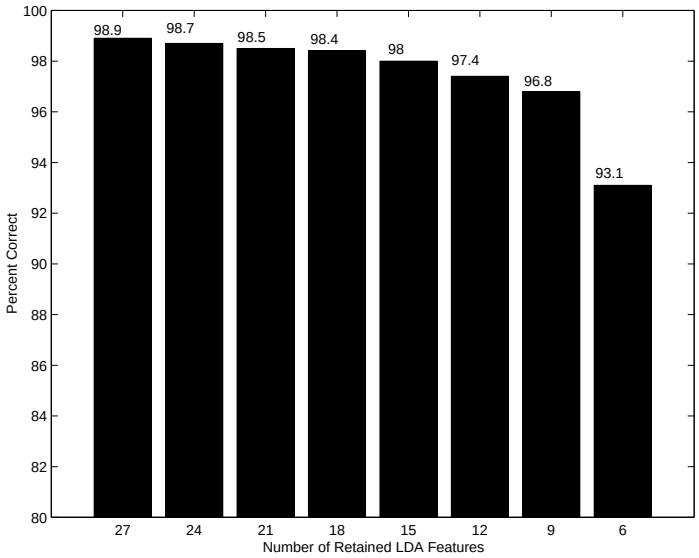


Figure 6-2: Results of varying the number of LDA features at 1000 mixtures, same corpus. (8 MFCC + E + d + dE + dd + ddE)

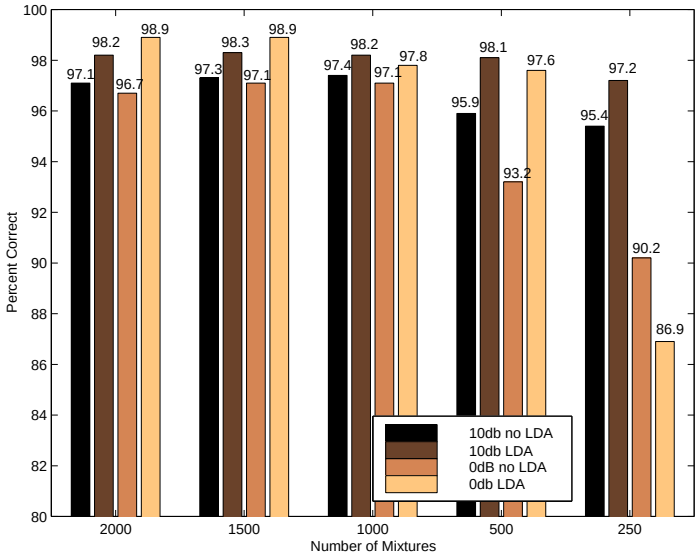


Figure 6-3: Results of varying the number of target mixtures, noise conditions. (8 MFCC + E + d + dE + dd + ddE)

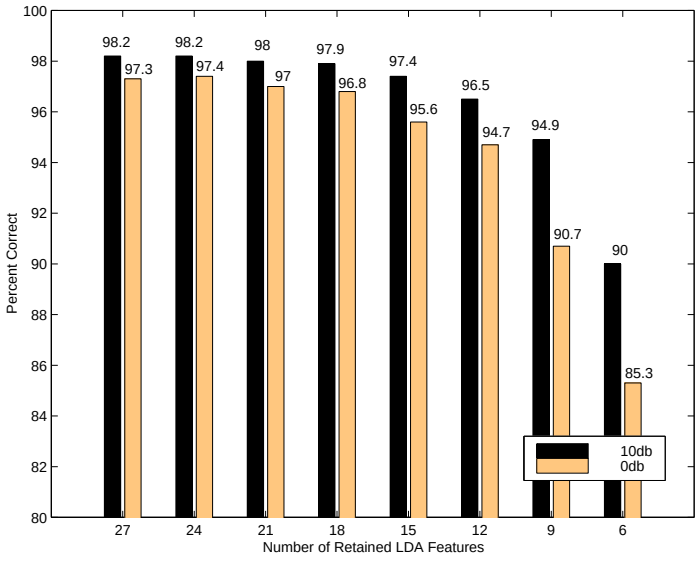


Figure 6-4: Results of varying the number of LDA features at 1000 mixtures, noise conditions. (8 MFCC + E + d + dE + dd + ddE)

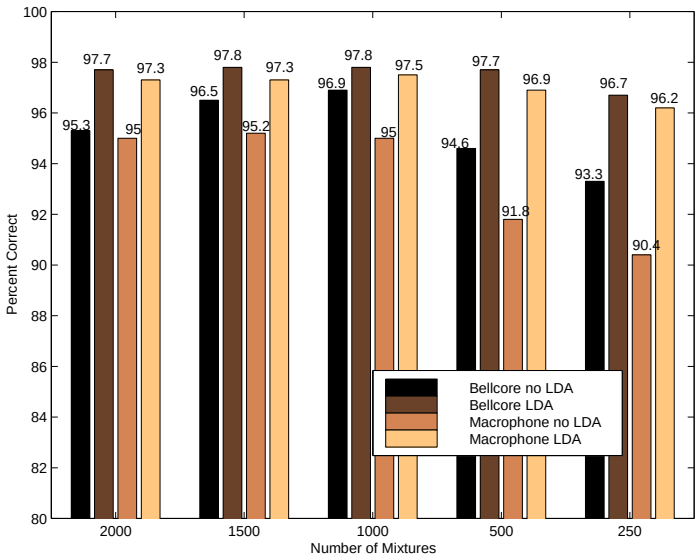


Figure 6-5: Results of varying the number of target mixtures, cross database. (8 MFCC + E + d + dE + dd + ddE)

with smaller numbers of mixtures. RASTA without energy tended to produce similar results to RASTA with energy for higher numbers of mixtures, but consistently faired better with smaller numbers of mixtures for these three tests. This tended to confirm that eliminating the raw gain from the feature vector may reduce some of the bias in the feature vectors.

The next three figures (7-4, 7-5, and 7-6) also show the results for the tests that vary the target number of mixtures, but this time with the use of LDA. Here, the trends were reversed. RASTA faired poorer than the baseline MFCC, often significantly. LDA seemed to affect RASTA in an inconsistent manner. For the same corpus tests and the Bellcore tests, LDA caused an improvement. For the remaining tests, LDA decreased the performance. This was not apparent in the MFCC tests which showed uniform improvements.

The final three figures (7-7, 7-8, and 7-9) collect the results for tests varying the number of retained LDA feature coefficients using a 1000 mixture HMM. The consistent trend was that fewer number of retained coefficients decreased recognition performance. For RASTA however, performance was maintained with fewer coefficients. This was especially true for RASTA without the energy feature where performance was maintained to as few as 9 retained coefficients.

## 8 Conclusions and Future Work

An underlying experimental issue affecting all of the RASTA-PLP experiments here was the manner and type of scaling and quantization. This was particularly pertinent for the LDA computation and scaling. In the MEL case, LDA was integrated into the code and computations were done in floating point. Prior MEL experiments by Josef Bauer demonstrated that that performing the calculations in integer rather than float affected results little. This is possibly due to the automatic range scaling in the MEL system. This was not apparent for experiments with RASTA. The scaling and quantization definitely parameterized the recognition performance of RASTA and is easily seen in figure 2-5. For the remaining tests, a specific scaling was chosen based upon some statistics of the features from the training set. With different noise conditions and different databases, the feature distributions were almost surely different which implies that the chosen scaling options may not have been optimal. The remedy for this would have been to bypass the scaling and quantization step and modify the recognition and LDA computation programs to accept float values. As this was not done at the time of these experiments, this issue remains to be explored.

An early possible problem with using RASTA with HMMs was the use of the RASTA filter. This filter performed a temporal differentiation and re-integration. The affect of this was the introduction of past context into the present frame of features which counters HMM assumptions of independent states. This however seemed not to be a problem with the recognition performance for this task. Indeed, the RASTA features with monophone models faired as well as or better than the baseline MEL and

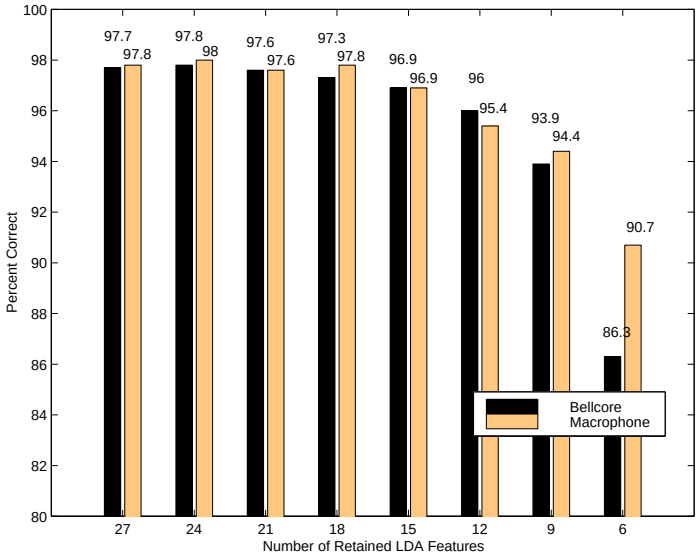


Figure 6-6: Results of varying the number of LDA features at 1000 mixtures, cross database. (8 MFCC + E + d + dE + dd + ddE)

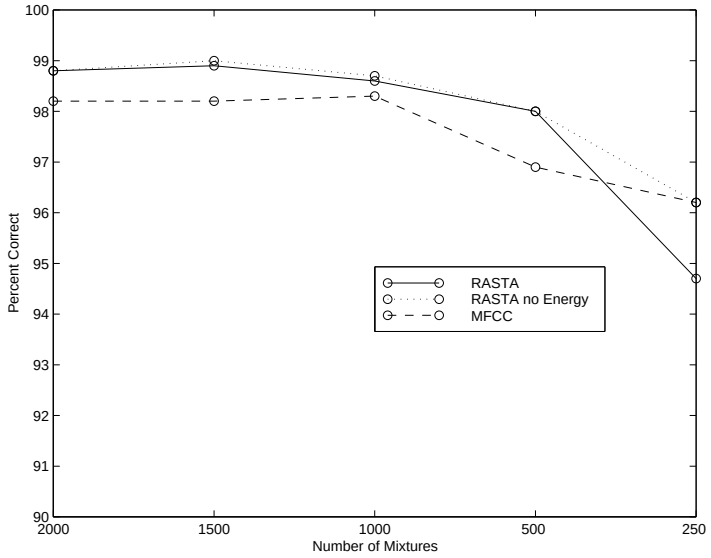


Figure 7-1: Same corpus results, varying number of mixtures, no LDA.

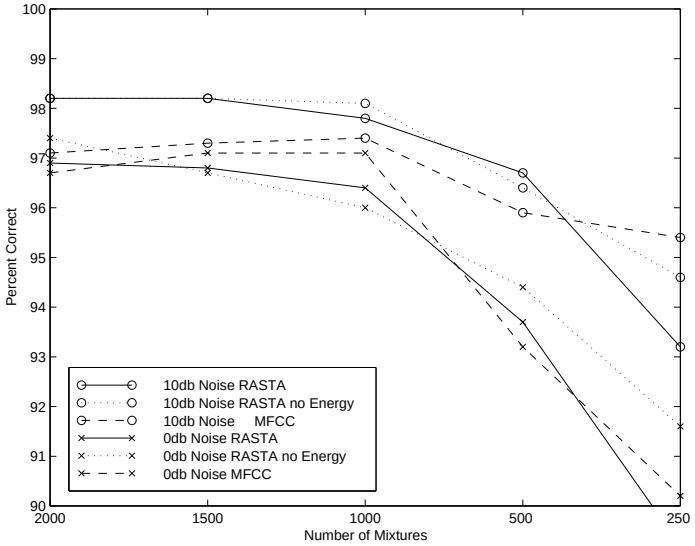


Figure 7-2: Noise condition results, varying number of mixtures, no LDA.

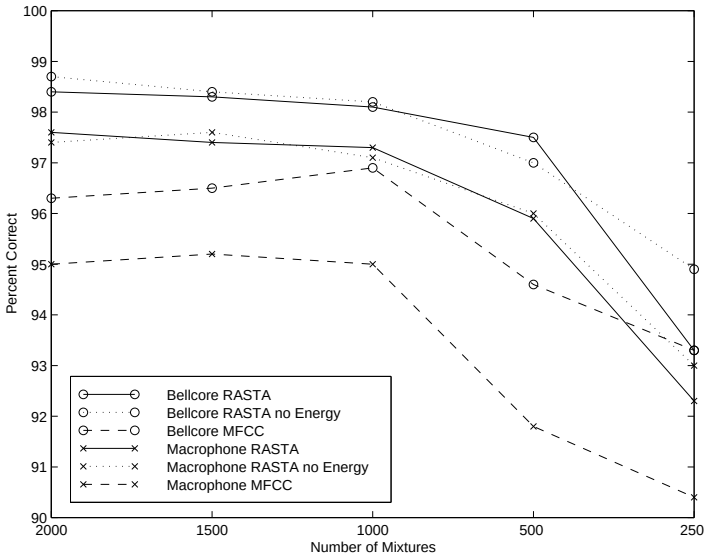


Figure 7-3: Cross database results, varying number of mixtures, no LDA.

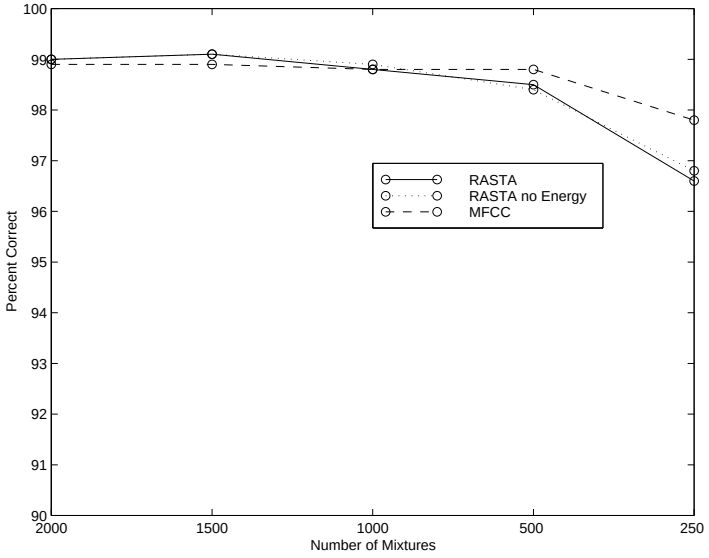


Figure 7-4: Same corpus results, varying number of mixtures, with LDA.

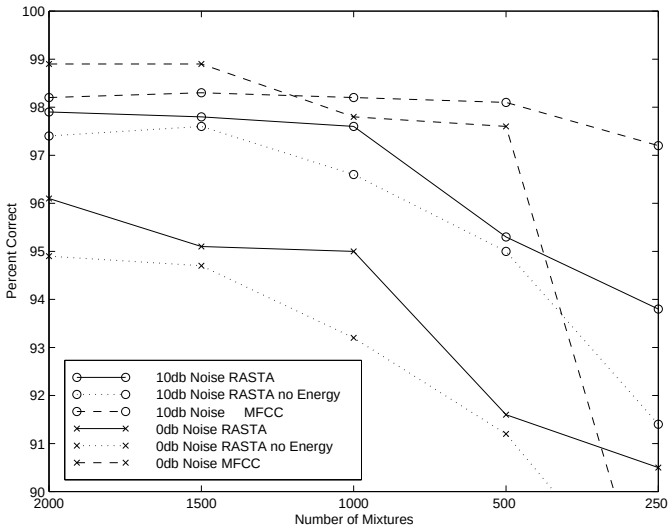


Figure 7-5: Noise condition results, varying number of mixtures, with LDA.

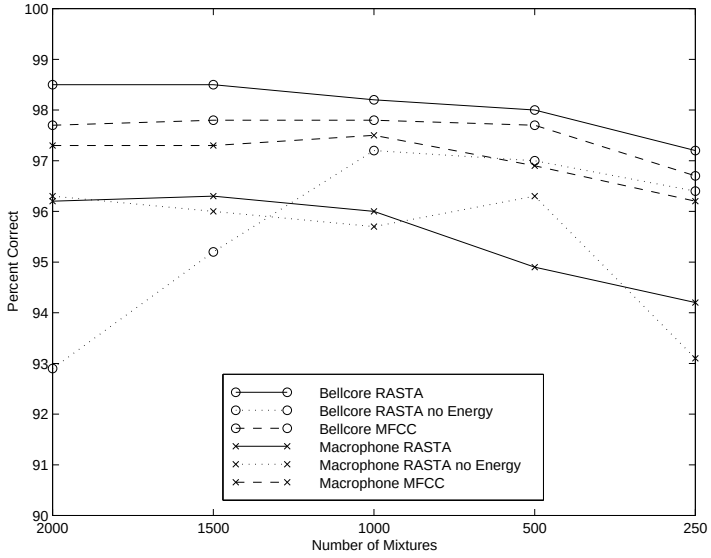


Figure 7-6: Cross database results, varying number of mixtures, with LDA.

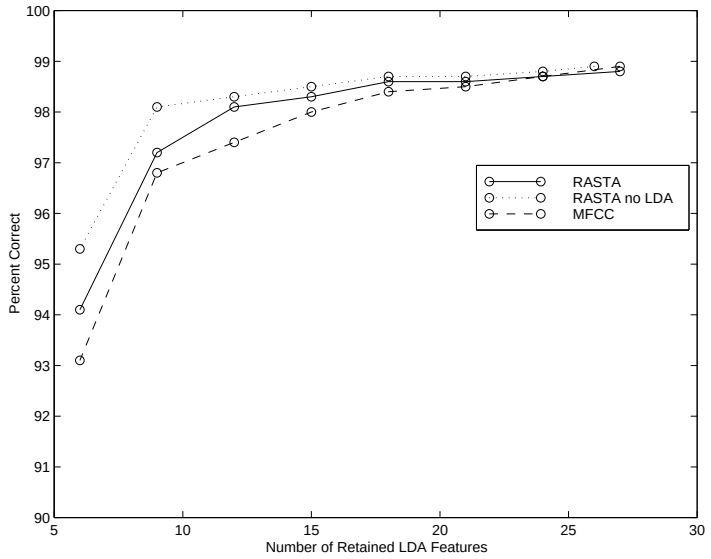


Figure 7-7: Same corpus results, vary number of retained LDA features.

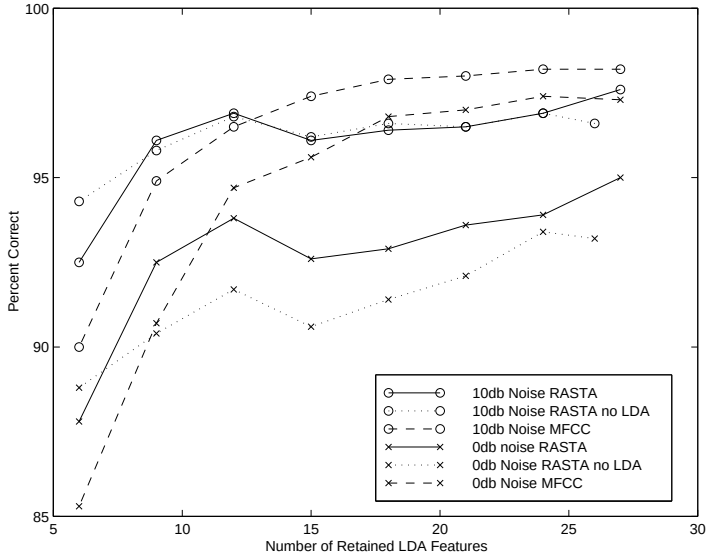


Figure 7-8: Noise condition results, vary number of retained LDA features.

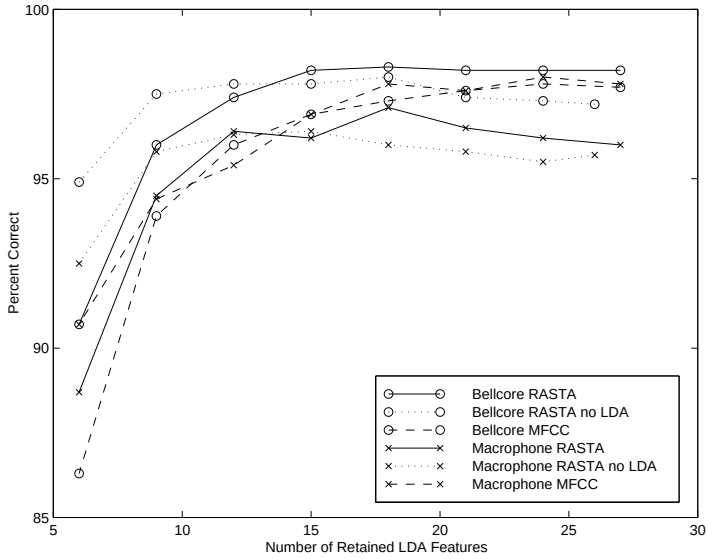


Figure 7-9: Cross database results, vary number of retained LDA features.

MFCC reference scores for cases without LDA. The use of context-dependent diphone models provided a better match for RASTA features and the results improved further.

LDA proved to be of undetermined benefit in its use with RASTA. For the same corpus tests and some cross-database tests, addition of LDA improved recognition performance. In other tests such as with additive noise LDA noticeably decreased the performance. This was not the case for the MFCC which showed uniform improvement when LDA was added. Pre-quantization may have been responsible for this. Further experiments where the LDA computations operated directly on float values would be necessary to adequately assess the benefits of LDA with RASTA. It would then be possible to observe if other effects are responsible for the questionable LDA results. For instance, the added noise or the change of recording environment may have upset the class feature distributions and therefore the LDA transformations. Further, the long integration window of the RASTA filter may have lead to wider class clusters with more possibility of class overlap. A detailed analysis of the class distributions both with and without quantization effects should shed light on some of these issues.

Despite the experimental limitations, RASTA-PLP still demonstrated good functional use as a feature extraction method in conjunction with the Siemens ZT recognizer. It performed as well or better than the standard or variant MEL reference scores for all cases without use of LDA. Further, the tests which varied the number of LDA coefficients showed that the performance of RASTA can be maintained with fewer coefficients and hence less computational burden.

It should be re-emphasized that the mismatch in experimental conditions disallow any direct comparison between RASTA-PLP and the MEL or MFCC feature extraction methods. The results of the series of experiments merely demonstrate the RASTA-PLP can achieve good performance in relation to what has already been achieved using MEL features. The recorded performance even with the imposed handicap of pre-quantized LDA demonstrate its value as an experimental research tool. With that, the RASTA-PLP computation was integrated into Siemens ZT feature extraction library for further experimentation. It is recommended that experiments be performed using the LDA directly on the float-valued features to resolve the inconsistent LDA performance issue. It is also recommended that the recognizer be modified to handle float-valued features to eliminate the problem of selecting which scaling and quantization to use for the RASTA-PLP features. Lastly, it is encouraged that Jah-RASTA be implemented and experimented with in the cases of different noise conditions.

## 9 Acknowledgments

Josef Bauer and Joachim Köhler were an instrumental help in acquainting me with the Siemens recognition environment and in conducting experiments. Josef Bauer assisted greatly in the experimental setup and was responsible for most of MFCC results. I

would like to express my deepest appreciation to Nelson Morgan, Alfred Hauenstein, Harald Höge, Wolfgang Küpper, and Peter Kleinschmidt for facilitating my stay at Siemens, AG. My stay would have been neither as productive nor as enjoyable were it not for Josef Bauer, Joachim Köhler, Erwin Marschall, Martin Holzapfel, Udo Bub, and the rest of my esteemed colleagues in the speech group. Additional thanks go to Ute Ziegenhain and Petra Witschel for encouraging my brief education in the German language. Vielendank.

## 10 References

## References

[1] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, Apr. 1990.

[2] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, Oct. 1994.

[3] T. W. Parsons. *Voice and Speech Processing*, chapter 7. McGraw-Hill Book Company, New York, 1986.

# 11 Addendum: Incorporation of RASTA into the Siemens ZT EAR Tools

RASTA-PLP has been incorporated into the EAR library and can be accessed through member functions in the MEL class library as well as the HSam2Feat feature extraction routine. No new parameters were added to the the front-end parameter file although the meaning of the some of the parameters has been altered.

The full functionality of RASTA-PLP available through the command line arguments of rasta 2.4 are not immediately available and would require further modifications. Only PLP and Log-RASTA processing are accessible as front end feature extraction techniques; Jah-RASTA remains incorporated but unavailable through the parameter file. As the incorporated code did not have direct access to speech files and to the command line, some selectable modes of operation in rasta 2.4 are fixed in the incorporated code. Among these are the restriction to online-processing mode, forced inclusion of deltas and double deltas, and elimination of control over various parameters such as integration constants, delta calculation window lengths, and RASTA-PLP mixture fractions. Where the internal original parameters were inaccessible, they were kept at default values. The only exception to this was the delta calculation regression filter length which was fixed to seven instead of nine. All rasta modes, switches, and variables are available for future access; it would be necessary to modify the MEL parameter file and the front-end library to include the new options. A description and explanation of all RASTA-PLP parameters can be found using rasta 2.4 without any arguments or by consulting the rasta 2.4 documentation.

Table 11-1 shows the front end parameters used for RASTA-PLP operation along with a brief description of the parameter and a typical value that would be used. Where possible these MEL parameters were mapped to analogous RASTA-PLP parameters. These parameters were selected as those which would likely be needed in typical use. Additional parameters may be added at a future date by modifying the front-end library to allow for more parameters.

| Parameter | Usage |
|---|---|
| intFlag | Set to 3 to select RASTA-PLP operation. |
| samp_freq | Sample frequency in Hz. [8000] |
| time_win_ms | The frame size in ms. [25] |
| time_shift_ms | The step size in ms. [10] |
| featDim | Number of features per frame [27] |
| subDC_flag | Turns on high pass filtering of speech to remove DC component [0] |
| noise_reduc_type | Allowed values are 0 to 3. Setting the lowest bit to 1 adds a small amount of single bit noise to the speech to safeguard against long strings of 0's which may cause numerical errors. Setting the second lowest bit to 1 turns off computation of the gain, delta gain, and double delta gain. [0] |
| channel_comp_type | Set to 0 for PLP and set to 1 for Log-RASTA [0] |
| channel_reset_mode | Set to 1 to turn on history [0] |
| chanFileName | History file name. This is not of the same format as the MEL chanFileName. [history.out] |

Table 11-1: Parameters and definitions relating to RASTA-PLP operation.