# ROBUST DIGIT RECOGNITION IN NOISY ENVIRONMENTS: THE IBM AURORA 2 SYSTEM

*George Saon, Juan M. Huerta and Ea-Ee Jan*

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
E-mail: {saon,huerta,ejan}@watson.ibm.com, Phone: (914)-945-2985

## ABSTRACT

In this paper we describe some experiments on the Aurora 2 noisy digits database. The algorithms that we used can be broadly classified into noise robustness techniques based on a linear-channel model of the acoustic environment such as CDCN [1] and its novel variant termed Alignment-based CDCN ($\mathrm{ACDC_N}$, proposed here), and techniques which do not assume any particular knowledge about the structure of the environment or noise conditions affecting the speech signal such as discriminant feature space transformations and speaker/channel adaptation. We present recognition experiments for both the clean training data and the multi-condition training data scenarios.

## 1. INTRODUCTION

In this paper we describe the system and techniques for the Aurora 2 noisy digits database and the results obtained. We developed two sets of acoustic models: the first set of models was trained on clean data only and the second set was trained on the multi-condition training data. For the system trained on clean data only, we applied CDCN [1] and a novel variant of this technique called Alignment-based CDCN ($\mathrm{ACDC_N}$). For both systems, we applied discriminant feature space transformations and speaker/channel adaptation. In section 2 we present an overview of the system characteristics and in section 3 we describe the techniques based on a linear model of the environment: CDCN and Alignment-based CDCN. Section 4 deals with discriminant feature space transformations and section 5 is about speaker/channel adaptation techniques: feature space MLLR and projection-based feature space MLLR. Finally, in section 6 we present a summary of the results obtained.

## 2. BASELINE SYSTEM DESCRIPTION

### 2.1. Front-end

Speech is coded into 25 ms frames, with a frame-shift of 10 ms. Each frame is represented by a feature vector of 13 Mel frequency-warped cepstral coefficients (MFCC) computed from a 24-filter Mel filterbank spanning the 0 Hz - 4.0 kHz frequency range. Prior to the Mel binning, the power spectra are smoothed via periodogram averaging (i.e., we compute five 2 ms frames and average the result in spectral domain to obtain one 10 ms frame) [12]. Every 9 consecutive cepstral frames are spliced together and projected down to 39 dimensions using linear discriminant analysis (LDA). The range of this transformation is further diagonalized by means of a maximum likelihood linear transform (MLLT). More details about these transformations will be given in section 4.

### 2.2. Acoustic models

The system uses 22 context-independent phones, each phone being modeled by a 3-state HMM with self-loops and forward transitions (no skips). The output distributions for the 66 sub-phonetic classes (called fenemes) are given by a mixture of at most 50 diagonal covariance Gaussian mixture components totaling around 3.2K Gaussians. Inter-word silence and silence at the beginning and end of the utterances are modeled by two separate phones. The systems trained on clean and on multi-style data are identical in terms of parameters, the only difference being the training data.

### 2.3. Search strategy

Given the simplicity of the task, we wrote a dedicated Viterbi decoder operating on an HMM network obtained by expanding the words in the acoustic vocabulary in terms of their phones (and fenemes). In order to decode multiple words, we allow transitions from the end state of one word to the start states of all the other words or to a sentence boundary state. The scores of these transitions are controlled by a word insertion penalty (set to 0 for the multi-style trained system and to -120 for the clean system, both in log domain). The Viterbi path computation is exact: it is done without any form of pruning.

## 3. COMPENSATION BASED ON A LINEAR MODEL OF THE ENVIRONMENT

The codeword dependent cepstral normalization technique (CDCN) was originally proposed by Acero and is described in [1]. This technique makes use of the assumption that the speech cepstra are affected by a linear channel and an uncorrelated stationary additive noise. We implemented a simplified version of this technique described in the next subsection. We also propose a novel enhancement of CDCN, the Alignment-based CDCN, in which a word hypothesis is employed to iteratively estimate the environment characteristics making use of the HMM models.

### 3.1. A simplified CDCN implementation

Let $X(\omega_k)$ denote the power spectral density (PSD) of the clean speech signal and similarly, let $N(\omega_k)$ and $H(\omega_k)$ be the PSD of the additive noise and the channel characteristic. Then, if $y, x, q$ and $n$ represent respectively, the cepstra of the of the noisy speech, the clean speech, the filter characteristic and the noise signal, it is possible to show that $y = x+q+r(x,y,q)$ and $y = n+s(x,n,q)$ where:

$$r(x,n,q) = IDFT \log(1 + e^{DFT(n-q-x)}) \qquad (1)$$

The CDCN algorithm assumes that the effect of the environment follows a model similar to the one described above and estimates $n$ and $q$ based on a Gaussian Mixture Model (GMM) of the clean environment where the correction terms (1) are applied to the means of the GMM.

Our CDCN implementation is based on 24-dimensional cepstral feature vectors from which 13-dimensional cepstra are derived after compensation. A 24-dimensional diagonal covariance GMM was trained on a subset of the clean training data. In our simplified implementation, the initial estimate of the noise vector is maintained across iterations (i.e., the noise is not re-estimated), and the correction terms (1) and $q$ are recomputed and re-estimated by keeping $n$ constant. In this way, we circumvent the main CDCN assumption which is that one mode of the GMM has to represent the noise model. This assumption is difficult to meet in practice where the noise may exhibit a rich variety of SNRs and spectral characteristics. We present the results of this simplified version of CDCN in Table 2 for the clean system.

### 3.2. Alignment-based CDCN

Standard CDCN makes use of a single GMM description of the clean speech. This description of the speech signal might be too general and thus might adversely affect the estimation of the environmental parameters. By considering Gaussian models which are closer to the phonetic characteristics of the speech frames, one could produce more accurate estimates of $n$ and $q$. In order to achieve this, we propose making use of a feneme dependent GMM i.e., we have separate feneme distributions similar to the HMM models used in the recognizer. We first perform an initial recognition pass and from the resulting output hypothesis we produce an alignment of the corresponding fenemes $g_1, \ldots, g_T$ to the speech frames $x_1, \ldots, x_T$.

Once this frame to feneme association is established, the correction vectors are then estimated in terms of the means of the GMMs. The correction for Gaussian mixture component $k$ of the feneme mixture model $g_j$ is:

$$r^{(g_j)}(\mu_k^{(g_j)}, n, q) = IDFT \log(1 + e^{DFT(n-q-\mu_k^{(g_j)})}) \quad (2)$$

The posterior probabilities are computed for each feneme making use of the correction terms for each Gaussian. In the maximization step, for each frame $x_t$, the corresponding feneme means and corrections are employed, weighted by the feneme posterior probability.

The clean cepstra are obtained by means of a MMSE estimate, as in conventional CDCN. The clean cepstra can be employed to obtain a new hypothesis and the whole compensation cycle can be repeated using the new alignment. The noise and channel vectors can also be initialized using the estimates from the previous iteration. The Gaussian mixture models need to match the type of features used in this compensation step; therefore, for the purpose of compensation, a set of GMMs was trained on 24-dimensional cepstra by means of a single-pass retraining scheme using the 39-dimensional HMM mixture distributions. The recognition results for the clean models are indicated in Table 2.

## 4. FEATURE SPACE TRANSFORMATIONS

### 4.1. Linear discriminant analysis

Linear discriminant analysis [2, 3] is a standard technique in statistical pattern classification for dimensionality reduction with a minimal loss in discrimination. Its application to speech recognition has shown consistent gains for both small and large vocabulary tasks [10]. In the following, we will recall some of its basic principles.

Consider a set of $N$ independent vectors $\{x_i\}_{1 \le i \le N}$, $x_i \in \mathbb{R}^n$, each of the vectors belonging to one and only one class $j \in \{1, \ldots, J\}$ through the surjective mapping of indices $l : \{1, \ldots, N\} \to \{1, \ldots, J\}$. Let each class $j$ be characterized by its own mean $\mu_j$, covariance $\Sigma_j$, and sample count $N_j$, where the standard definitions hold:

$$\mu_j = \frac{1}{N_j} \sum_{i \in l^{-1}(j)} x_i, \qquad \Sigma_j = \frac{1}{N_j} \sum_{i \in l^{-1}(j)} x_i x_i^T - \mu_j \mu_j^T$$

and $\sum_{j=1}^{J} N_j = N$. The class information is condensed into 2 scatter matrices called:

- *within-class* scatter: $W = \frac{1}{N} \sum_{j=1}^{J} N_j \Sigma_j$ and
- *between-class* scatter: $B = \frac{1}{N} \sum_{j=1}^{J} N_j \mu_j \mu_j^T - \overline{\mu\mu}^T$

The goal of LDA is to find a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^p$, $y = f(x) = \theta x$, with $\theta$ a $p \times n$ matrix of rank $p \le n$, such that the following ratio of determinants is maximized:

$$J(\theta) = \frac{|\theta B \theta^T|}{|\theta W \theta^T|} \quad (3)$$

The maximizer of (3) is given by the transposed eigenvectors corresponding to the $p$ largest eigenvalues of the generalized eigenvalue problem: $Bx = \lambda W x$.

From a practical point of view, every 9 consecutive 13-dimensional cepstral vectors are spliced together forming 117-dimensional feature vectors which are then clustered to make possibly multiple full covariance Gaussians for each HMM state (totaling around 300 Gaussians). Subsequently, a 39×117 transformation, $\theta$, is computed using the LDA objective function (3) and the parameters $(N_j, \mu_j, \Sigma_j)$ of the Gaussians.

### 4.2. Maximum likelihood linear transforms

If the dimensions in the LDA subspace are highly correlated then a diagonal covariance modeling constraint will result in distributions with large overlap between classes. In this case, a maximum likelihood feature space transformation [6, 5] which aims at minimizing the loss in likelihood between full and diagonal covariance models is known to be very effective. The objective for MLLT is to find a transformation $\psi$ that minimizes the difference:

$$\hat{\psi} = \underset{\psi \in \mathbb{R}^{p \times p}}{\operatorname{argmin}} \sum_{j=1}^{J} -N_j \left( \log |\psi \Sigma_j \psi^T| - \log |\operatorname{diag}(\psi \Sigma_j \psi^T)| \right) \quad (4)$$

The $\Sigma_j$'s in (4) are obtained by reclustering the vectors in the LDA space for each HMM state (roughly 1K full-covariance Gaussians). Finally, we estimate the 3.2K diagonal covariance Gaussians in the 39-dimensional LDA+MLLT space. The performance of our baseline LDA+MLLT models is indicated in the rows 4–6 of Table 1 for the multi-style trained system and in Table 2 for the clean system.

| System | Test | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB | Clean | 0-20dB |
|---|---|---|---|---|---|---|---|---|---|
| HTK$^+$ Baseline | TSA | 24.6 | 61.7 | 87.8 | 94.9 | 96.9 | 97.7 | 98.5 | 87.8 |
| | TSB | 25.9 | 60.5 | 84.7 | 93.1 | 95.8 | 97.2 | 98.5 | 86.3 |
| | TSC | 21.6 | 50.6 | 82.5 | 92.9 | 95.9 | 96.9 | 98.5 | 83.8 |
| Baseline | TSA | 28.9 | 67.2 | 88.8 | 95.9 | 97.7 | 98.3 | 99.1 | 89.6 |
| | TSB | 20.2 | 60.7 | 85.0 | 93.9 | 97.2 | 98.3 | 99.1 | 87.0 |
| | TSC | 35.3 | 70.8 | 89.2 | 95.4 | 97.3 | 98.1 | 99.1 | 90.2 |
| FMLLR | TSA | 31.9 | 71.5 | 89.6 | 96.3 | 97.9 | 98.6 | 99.3 | 90.8 |
| | TSB | 24.2 | 64.0 | 86.2 | 94.5 | 97.2 | 98.4 | 99.3 | 88.1 |
| | TSC | 32.6 | 73.5 | 89.3 | 95.4 | 97.9 | 97.9 | 99.1 | 90.8 |
| FMLLR-P | TSA | 32.2 | 73.4 | 92.1 | 97.2 | 98.5 | 99.0 | 99.5 | 92.0 |
| | TSB | 24.1 | 67.4 | 89.3 | 96.2 | 98.0 | 98.9 | 99.5 | 90.0 |
| | TSC | 32.1 | 74.9 | 92.0 | 96.8 | 98.4 | 98.7 | 99.5 | 92.1 |

Table 1: Word recognition accuracies for the multi-style system averaged across noise types. $^+$Please refer to [8] for HTK baseline

## 5. SPEAKER/CHANNEL ADAPTATION

Speaker adaptation, as exemplified by MLLR, is a key technique that is used in most state-of-the art systems. In this step, a linear transform is found such that, when it is applied to either the Gaussian means [9] or, as in constrained MLLR, to the feature vectors themselves [4], the likelihood of the acoustic data associated with an utterance is maximized with respect to an initial word hypothesis. The utterance is then re-decoded after applying the transform.

### 5.1. Feature space MLLR

The goal of feature space MLLR is to affinely transform the adaptation data $x_1, \ldots, x_T$, $x_t \in \mathbb{R}^n$, such as to maximize their likelihood, i.e. find $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$ with $\hat{x}_t = Ax_t + b$ such that the auxiliary function of the EM algorithm is maximized:

$$\sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_t(j) \left[ \log |A| - \frac{1}{2} (\hat{x}_t - \mu_j)^T \Sigma_j^{-1} (\hat{x}_t - \mu_j) \right] + \mathcal{C} \quad (5)$$

where $\mathcal{C}$ is a constant with respect to $A$ and $b$. For simplicity of notation, the summation over the HMM states and over the mixture components within a state has been collapsed into a single sum over all the Gaussians in the model. $\gamma_t(j)$ represents the posterior probability of component $j$ at time $t$ given the complete observation sequence. We define the *sufficient statistics* for feature space MLLR by:

- $K = \sum_{t=1}^{T} \sum_{j=1}^{N} \gamma_t(j) \Sigma_j^{-1} \mu_j x_t^T$ and
- $G_i = \sum_{t=1}^{T} \sum_{j=1}^{N} \frac{\gamma_t(j)}{\sigma_{ji}^2} x_t x_t^T, \; i = 1 \ldots n$

where $\Sigma_j = \mathrm{diag}(\sigma_{j1}^2, \ldots, \sigma_{jn}^2)$. By writing the gradient of (5) in terms of these statistics, the rows of $A$ can be found independently through iteratively solving a set of quadratic equations (for more details the reader is referred to [4]).

### 5.2. Projection-based feature space MLLR

Until now, we have studied the case when $A$ is a full rank transformation operating in the complete $n$-dimensional space. To consider the projection to a $p$-dimensional subspace with $p \leq n$, the following structure will be imposed on the model parameters [11]:

$$\mu_j = \begin{bmatrix} \mu_j^{(p)} \\ \mu_0^{(n-p)} \end{bmatrix}, \; \Sigma_j = \begin{bmatrix} \Sigma_j^{(p)} & 0 \\ 0 & \Sigma_0^{(n-p)} \end{bmatrix}, \; 1 \leq j \leq N \quad (6)$$

meaning that, after the transformation is applied, the rejected dimensions are supposed to be identically (Gaussian) distributed across all the mixture components. This is a similar assumption to the one made in heteroscedastic discriminant analysis [7]. Correspondingly, $A$ can be decomposed into two parts, $A = [A^{(p)T} | A^{(n-p)T}]^T$, where $A^{(p)}$, of dimension $p \times n$, will be the useful projection and $A^{(n-p)}$, of dimension $n - p \times n$, will provide the complementary dimensions. Its role is to provide a full rank completion to $A^{(p)}$ in order to be able to make meaningful likelihood comparisons across feature spaces of equal dimension ($n$).

From an experimental point of view, for both feature space MLLR (FMLLR) and its projection variant (FMLLR-P), we first accumulate sufficient statistics for the test data of each speaker, then we find the feature space transform and then we redecode the transformed acoustic data. The difference between the two is as follows: for FMLLR, the statistics are accumulated in the 39-dimensional LDA+MLLT space and the transform is $39 \times 39$ whereas for FMLLR-P the statistics are accumulated in an 80-dimensional LDA space for a model obtained by augmenting the 39-dimensional Gaussians with the mean and variance of all the training data in the $80 - 39$ LDA complement. The resulting FMLLR-P transform is $39 \times 80$. The performance of FMLLR and FMLLR-P is indicated in Table 1 for the multi-style trained model and in Table 2 for the clean system.

## 6. DISCUSSION

Tables 1 and 2 show the results obtained using multi-condition and clean models respectively, after applying the techniques described in this paper. For the multi-style trained system, Table 1 has four sections corresponding to the HTK baseline, our baseline, FMLLR and FMLLR-P performance. Both the baseline and the adapted systems include the LDA and MLLT transformations described previously. It can be seen that, in the 0 to 20 dB SNR range, both FMLLR and FMLLR-P result in significant reductions

| System | Test | -5dB | 0dB | 5dB | 10dB | 15dB | 20dB | Clean | 0-20dB |
|---|---|---|---|---|---|---|---|---|---|
| HTK$^+$ | TSA | 7.9 | 17.0 | 39.5 | 67.7 | 87.3 | 95.3 | 99.0 | 61.3 |
| Baseline | TSB | 7.7 | 13.7 | 31.9 | 59.0 | 81.3 | 92.8 | 99.0 | 55.7 |
| | TSC | 11.5 | 24.2 | 50.2 | 74.2 | 87.8 | 94.3 | 99.1 | 66.1 |
| | TSA | 10.0 | 19.8 | 45.3 | 70.6 | 84.2 | 92.7 | 99.4 | 62.4 |
| Baseline | TSB | 10.8 | 21.4 | 48.2 | 75.7 | 90.4 | 95.9 | 99.4 | 66.3 |
| | TSC | 11.5 | 24.7 | 49.9 | 69.5 | 82.1 | 89.8 | 99.4 | 63.2 |
| | TSA | 13.0 | 25.7 | 51.9 | 78.5 | 93.1 | 97.4 | 99.6 | 69.4 |
| CDCN | TSB | 12.9 | 30.4 | 57.6 | 83.2 | 95.6 | 97.9 | 99.4 | 72.9 |
| | TSC | 15.1 | 28.0 | 52.4 | 77.4 | 91.8 | 96.9 | 99.4 | 69.3 |
| | TSA | 14.9 | 31.4 | 62.6 | 84.5 | 94.6 | 97.6 | 99.4 | 74.1 |
| $\mathrm{ACDC_N}$ | TSB | 14.4 | 33.2 | 64.7 | 87.0 | 96.1 | 98.0 | 99.4 | 75.8 |
| | TSC | 14.8 | 33.0 | 61.1 | 83.3 | 93.0 | 97.0 | 99.4 | 73.5 |
| | TSA | 12.5 | 26.9 | 56.8 | 83.2 | 94.5 | 97.8 | 99.4 | 71.8 |
| FMLLR | TSB | 12.8 | 31.4 | 62.4 | 86.7 | 96.3 | 98.1 | 99.4 | 75.0 |
| | TSC | 14.8 | 28.7 | 56.4 | 81.5 | 93.0 | 97.3 | 99.4 | 71.4 |
| | TSA | 12.9 | 26.9 | 56.0 | 82.6 | 94.3 | 97.8 | 99.4 | 71.5 |
| FMLLR-P | TSB | 12.9 | 31.3 | 61.6 | 86.2 | 96.3 | 98.1 | 99.4 | 74.7 |
| | TSC | 14.9 | 28.8 | 56.0 | 80.8 | 92.8 | 97.2 | 99.4 | 71.1 |
| $\mathrm{ACDC_N}$+ | TSA | 14.7 | 31.5 | 63.3 | 85.0 | 94.8 | 97.7 | 99.4 | 74.5 |
| FMLLR | TSB | 14.3 | 33.3 | 65.6 | 87.6 | 96.2 | 98.1 | 99.4 | 76.2 |
| | TSC | 14.8 | 33.3 | 61.7 | 83.7 | 93.2 | 97.1 | 99.4 | 73.8 |

Table 2: Word recognition accuracies for the clean model system averaged across noise types. $^+$Please refer to [8] for HTK baseline

of the word error rate, with FMLLR-P yielding 16% relative larger gains than FMLLR.

For the system considered in Table 2, the CDCN and $\mathrm{ACDC_N}$ techniques were also tried. We observe that the gains for the range of 0 to 20 dB SNR are considerably larger for $\mathrm{ACDC_N}$ compared to CDCN. FMLLR and FMLLR-P seem to give comparable gains in these conditions, in other words, the superiority of FMLLR-P over FMLLR that we observed for the multi-condition system is not observed for the clean system. Finally, $\mathrm{ACDC_N}$ was used to compensate the 24 dimensional cepstra followed by LDA+MLLT and FMLLR; the results are presented in the last 3 rows of Table 2. We observe that the combined gains of FMLLR and $\mathrm{ACDC_N}$ are just marginally better than those obtained using each technique separately.

## 7. REFERENCES


[1] A. Acero. Acoustical and Environmental Robustness in Automatic Speech Recognition. Ph.D. Thesis, Department of ECE, Carnegie Mellon University, Pittsburgh PA, 1990.

[2] R. O. Duda and P. B. Hart. Pattern classification and scene analysis. *Wiley*, New York, 1973.

[3] K. Fukunaga. Introduction to statistical pattern recognition. *Academic Press*, New York, 1990.

[4] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report CUED/F-INFENG, Cambridge University Engineering Department, 1997.

[5] M. J. F. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 7:272–281, 1999.

[6] R. A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. *Proceedings of ICASSP'98*, Seattle, 1998.

[7] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communcation*, 26:283–297, 1998.

[8] H. G. Hirsh and D. Pearce. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proceedings of ISCA ASR 2000*, Paris, 2000.

[9] C. J. Leggetter and P. C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG, Cambridge University Engineering Department, 1994.

[10] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen. Maximum likelihood discriminant feature spaces. *Proceedings of ICASSP 2000*, Istanbul, 2000.

[11] G. Saon, G. Zweig and M. Padmanabhan. Linear feature space projections for speaker adaptation. *Proceedings of ICASSP 2001*, Salt Lake City, 2001.

[12] S. Dharanipragada, R. Gopinath and B. Rao. Techniques for capturing temporal variations in speech signals with fixed-rate processing. *Proceedings of ICSLP'98*, Sydney, 1998.