Nonlinear and Linear Transformations of Speech Features to Compensate for Channel and Noise Effects

Saurabh Prasad and Stephen A. Zahorian

Department of Electrical and Computer Engineering
Old Dominion University, Virginia, U.S.A.
saurabh.prasad@ieee.org, szahoria@odu.edu

Abstract
Automatic speech recognizers perform poorly when training and test data are systematically different in terms of noise and channel characteristics. One manifestation of such differences is variations in the probability density functions (pdfs) between training and test features. Consequently, both automatic speech recognition and automatic speaker identification may be severely degraded. Previous attempts to minimize this problem include Cepstral Mean and Variance Normalization and transforming all speech features to a univariate Gaussian pdf. In this paper, we present a quantile based Cumulative Density Function (CDF) matching technique for data drawn from different distributions. This method can be used to compensate for the systematic marginal (i.e. each feature individually) differences between training and test features. We further propose a linear covariance normalization technique to compensate for differences in covariance properties between training and test data. Experimental results are given that illustrate these techniques for speech recognition and automatic speaker identification.

1. Introduction
The effects of noise and channel distortions on robust speech recognition have been documented in various studies in the past [1], [2]. It has been shown that the feature space is non-linearly transformed by a simple environmental model consisting of additive noise and linear distortion (i.e., filtering). In a mismatched scenario, where a classifier is trained with clean speech but tested with noisy and/or channel degraded speech, the parameters of the trained system are not representative of the test speech. It has been shown that a normalization which restores the first two moments of corrupted speech partially compensates for additive noise and improves the recognizer’s performance [3], [4].

In this paper, we propose a method that transforms the features of degraded speech in order that they more closely resemble those of the training speech using a combination of non-linear and linear transformations. A linear (i.e. matrix multiply) transformation is first applied to the features such that test data covariance matches that of the training data. After restoring the inter-feature information, we perform a non-linear transformation that restores marginal pdfs. The non-linearity is obtained using quantile-based cumulative density function (CDF) matching approximated as a polynomial. Previous work along these lines documents the improvement in speech recognition performance by a marginal Gaussian transformation using histogram matching [3]. Although these transformations do help to restore the statistics of degraded speech features, they typically assume a single mode Gaussian for the target pdf, and also make a fundamental assumption about the independence of features. Our contribution is the introduction of an efficient algorithm for non-linearly transforming features that can restore arbitrary training pdfs, including Gaussian mixtures. We also introduce the concept of restoring the statistical information across feature axes in a multi-dimensional sense by a linear transformation, thereby reducing the inter-feature variations introduced by the channel.

We begin by summarizing the effects of additive noise and linear channel distortions on a feature space comprised of Mel Frequency Cepstral Coefficients (MFCCs). We then propose an efficient way to non-linearly transform the feature space so that the transformed features more closely resemble the original features. Finally, we propose an algorithm for compensating for the cross-feature variations caused by noise and channel differences, using linear transformations. Two types of tests were used for studying the effect of these transformations on the recognition accuracy. One was a vowel recognizer using a Bayesian minimum error rate classifier and the other was a speaker ID system, consisting of a binary-pair partitioned neural network as the classifier. The results show a significant improvement in the recognition performance of these systems with mismatched training and test conditions.

2. Background
Consider a model for the environmental degradation [2] of speech consisting of linear channel distortion and additive noise. It has been shown [3] that pdfs of clean speech features are non-linearly scaled by this degradation model. Restoration techniques used to recover clean speech features from noisy observations require good estimates of the channel noise and distortion parameters.

Alternatively, in [5,7] a CDF matching algorithm to a Gaussian density is proposed, but the fitting between quantiles is restricted to at most two modes of freedom. This restricts the quality of the transformed pdfs, especially for low SNRs and severe channel distortion.

3. Algorithm development
This section describes the algorithm presented for the statistical conditioning of features and motivates its use for removing the mismatch induced by noise and channel distortion. The Quantile-based CDF Matching (QCM) algorithm computes and uses a nonlinear scaling such that the statistical properties of the scaled features have a desired (for example those of clean speech) statistical structure. The nonlinearity is approximated by a polynomial using least
squares fitting. Although ideally a multivariate nonlinear transformation should be used to nonlinearly scale all features so that they would have specified first and second order moments, the determination of such a transformation is likely to be very complex. Therefore, as an approximation to this ideal, we present a linear transformation (Cross feature Linear Normalization, CLN) that rotates and linearly scales the features to a previously determined orientation for each category. In essence, these normalizations make the boundaries learned by the classifier from clean data applicable to the transformed noisy data, thus improving classifier accuracy.

3.1. Quantile-based CDF matching (QCM)

The non-linear scaling transformation we present to restore the feature statistics to a defined or learned form is based on Quantile based CDF matching. Marginal Gaussian normalization has been a topic of active research for the past few years. It has been shown [4] that a marginal transformation of features to a Gaussian target pdf (histogram equalization) improves the recognition accuracy of ASR systems. However, this previous work typically assumes the target pdf is single-mode Gaussian, which is not necessarily a good representation for feature pdfs. We propose a generalized statistical normalization that is capable of restoring the global feature statistics (even multi-mode Gaussians) of the noisy speech features. The fact that we do not restrict the reference pdf to a single-mode Gaussian makes it an attractive feature normalization scheme for Automatic Speech Recognition and Speaker Identification. Further, this implementation has the practical advantage of not needing a large data-set to determine the transformation.

The algorithm for QCM is as follows:

- Define (or learn) the target Cumulative Density Function
- Partition into equi-probable bins
- Sort the data to be transformed in ascending order.
- Plot this data into equi-probable bins
- Find the mean of each bin ($\mu_{\text{data}}$)
- Find the mean of each bin ($\mu_{\text{target}}$)
- Determine the transformation (based on QCM) required to transform the test speech to a target CDF.
- Fit this non-linear transformation to an n’th order polynomial, using least squares fitting.

The problem of fitting the non-linearity to an n’th order polynomial $y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$ is formulated in the usual least squares sense - given m data points $(x_i,y_j)$, we wish to find the coefficients $a_0, a_1, a_2, \ldots, a_n$ such that the following cost function is minimized

$$E^\lambda = \sum_{j=1}^{m} \left( y_j - \left(a_0 + a_1x_j + a_2x_j^2 + \ldots + a_nx_j^n\right) \right)^2 \quad (1)$$

The coefficients $a_0, a_1, a_2, \ldots, a_n$ are obtained by solving the equations

$$\partial E^\lambda / \partial a_k = 0, k = 1:n \quad (2)$$

Figures 1 and 2 illustrate this overall QCM process. Figure 1 depicts the partitioning of a bimodal CDF into equi-probable bins. The derived nonlinearity (approximated as a 7th order polynomial) was used to nonlinearly scale data, simulated using $10^7$ randomly generated data points, from an original bimodal density to a specified different bimodal density, as depicted in Figure 2.

The procedure described above can be used to compensate for mismatched training and test data. Assuming that we have a large number of clean feature vectors available for ‘Learning the Target CDF’ and ‘Training the Classifier,’ all we need is a small number of ‘training’ feature vectors at the same SNR as that of the test speech. We can use these to determine the transformation (based on QCM) required to restore each feature CDF to the target CDF (learned from the clean frames), and then fit this transformation to a polynomial of appropriate order (the lower the SNR, the higher the polynomial order required for a good fit). This polynomial model is then used for nonlinear scaling of each feature.

We found that a few hundred frames at the same SNR as the test speech are sufficient to accurately learn the non-linearity required to transform the test speech to a target CDF. Once learned, the nonlinearity is used as a fixed scaling. This implementation has an advantage over previously proposed histogram transformation algorithms, which typically need a large accumulated data-set to perform the transformation. Further, since we are not making a ‘single-mode’ Gaussian pdf assumption, we can more effectively restore the noisy speech features to their true probabilistic
structure (for example, MFCC \( c_0 \) as bimodal), instead of a single-mode Gaussian with zero mean and unit standard deviation \([4]\).

### 3.2 Cross-feature Linear Normalization (CLN)

Since in most speech/speaker recognition applications, class conditional density functions are Gaussians or their mixtures, the covariance matrices have a direct influence on the shape of the equi-probable contours and hence the decision boundaries learned by the classifier. It can hence be inferred that a classifier that learns its boundaries from data having a covariance matrix \( \Sigma_i \) for the \( i \)th category will not perform well when the test data has a covariance matrix \( \Sigma_j \). It is desired to restore the shape of the equi-probable contours before making a labeling decision. In this section, we describe one method for combining the CDF matching restoration with a linear transformation, which can be used to make the second order statistics (covariance matrix) of test data match that of training data. In particular, consider the following algorithm:

- **Learn the covariance matrix** \( \mathbf{R}_i^{\text{clean}} \) of data for each category from the clean speech frames.
- Collect a suitable number of test data points (\( y \)), and compute the sample covariance matrix, \( \mathbf{R}_i^{\text{test}} \).
- ‘Diagonalize’ and then ‘whiten’ the collected test data to so that the covariance matrix is Identity:
  \[
  \mathbf{y}' = \Lambda_y^{-1/2} \mathbf{U}_y^T \mathbf{y}
  \]

In (3), \( \mathbf{U}_y \) is a matrix whose columns are the eigenvectors of \( \mathbf{R}_i^{\text{test}} \), and \( \Lambda_y \) contains the corresponding eigen-values on the diagonal.

- Perform another transformation as :
  \[
  \hat{\mathbf{y}} = \left( \mathbf{R}_i^{\text{clean}} \right)^{1/2} \mathbf{\tilde{y}}
  \]

It can be shown easily that

\[
E\{\mathbf{y}' \mathbf{y}'^T\} = \mathbf{R}_i^{\text{clean}}
\]

assuming that we started with zero-mean vectors. To insure zero mean, the mean vector can be subtracted before this transformation and added back later. These steps ensure that the collected test-data has a covariance matrix approximating the covariance matrix of the clean data corresponding to that category.

In a classification situation, we do not have the category label, and hence do not know which \( \mathbf{R}_i^{\text{clean}} \) to use as the target matrix. However, the approach just described can be applied to the data set as a whole, with no need for category labels, and the global covariance of the test data can be forced to match the global covariance of the training data. To demonstrate feasibility, we used the following approach:

- Evaluate the sample covariance matrix \( \mathbf{R}_i^{\text{test}} \) over a sufficient number of frames from the category being recognized.
- Among the possible target covariance matrices \( \mathbf{R}_i^{\text{clean}} \) (\( i \in [1:C] \) where \( C \) is the number of categories), choose the one such that the matrix norm \( \| \mathbf{R}_i^{\text{test}} - \mathbf{R}_i^{\text{clean}} \| \) is minimum.

The above proposition can be easily extended to the case of isolated word recognition or speaker ID problems, where each category is modeled by a mixture of Gaussians. In this case, we need to build the Gaussian mixture model of the test frames, and repeat the above technique for each component in the mixture.

Experimental results in the next section illustrate the potential benefit of this cross-feature normalization.

### 4. Experiments and results

In this section we present a few small experiments and results intended only to show the feasibility and possible application of QCM and CLN. Figure 3 illustrates processing incorporating both the QCM and CLN algorithms for removing the training-test mismatches in speech recognition / speaker-identification systems. During training at high SNR, the front-end learns the covariance matrices of the categories and the marginal CDFs of the features and the back-end learns the optimal decision boundaries for the labeled training vectors. During testing (at any SNR), the CLN unit removes the distortion and rotation effects on the test data. The front-end then scales the feature vectors such that the marginal CDFs are close to those learned from the clean speech frames. After this conditioning, the back-end makes the classification decision.

![Diagram](image)

**Fig. 3:** Illustrating the proposed implementation of the QCM and CLN algorithms

In all experiments, the speech sampling rate was 16 kHz and a 1024 point FFT was used. DCTCs, very similar to MFCCs but computed somewhat differently \([6]\) were used as the features. Cepstral mean normalization was used for pre-processing DCTCs for the baseline results. Table 1 illustrates the effect of incorporating the QCM processing for speaker ID. A binary-pair partitioned neural network was used for classifying 50 speakers from the NTIMIT database at various SNRs. A few hundred frames of the features from one speaker were used to learn the polynomial coefficients to transform the data at that SNR to a reference Gaussian using QCM. This set of polynomial coefficients for each feature were then used for marginally transforming all feature vectors (training and test) to a global reference Gaussian pdf. It is evident that QCM consistently improves the recognition accuracy. We also used the SPIDRE corpus to simulate a severe channel mismatch in training and test conditions on a speaker ID experiment with 45 speakers. We used two sentences from one handset for training the neural network and used a sentence from a different handset for testing the speaker’s identity. Here too, QCM gave some improvement.
Conditions, and was found to be slightly beneficial. Correlated the feature space, both for training and test phase in conjunction with HEQ. PCA defines independence. In some pilot work, we performed PCA in the histogram matching also assumes that the features are restricted to the single mode target pdf) in the implementation of QCM. Previous ASR work using restricted to the single mode target pdf (HEQ) because of the generalized transformation (not approximating to the normal [9]. Table 2 illustrates the improvement of QCM over Histogram equalization technique using Polya's approximation to the normal [9].

We illustrated the advantage of the algorithm for speaker ID and vowel classification. We also introduced the concept of compensating for cross-feature distortion and rotation by restoring the covariance matrix of the test data. In additional work, we plan to employ the CLN algorithm for automatic speaker identification. It is hoped that a cross-feature statistical normalization will yield improvements in recognition accuracy for the speaker ID problem also.

6. Acknowledgements

This work was partially funded by JTASC grant JW900.

7. References


Table 2: Accuracy of speaker identification, with and without QCM, for NTIMIT and SPIDRE databases.

<table>
<thead>
<tr>
<th># Features</th>
<th>Baseline</th>
<th>HEQ</th>
<th>QCM</th>
<th>QCM+CLN</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>33.4%</td>
<td>47.5%</td>
<td>46.4%</td>
<td>60.8%</td>
</tr>
<tr>
<td>10</td>
<td>42.7%</td>
<td>53.9%</td>
<td>55.9%</td>
<td>91.6%</td>
</tr>
<tr>
<td>14</td>
<td>42.4%</td>
<td>53.4%</td>
<td>56.9%</td>
<td>92.9%</td>
</tr>
</tbody>
</table>

Table 2 illustrates the results of QCM and CLN algorithms for a monophone classification experiment. A mismatch was simulated by training on TIMIT and testing on NTIMIT files. For these experiments, 10 vowels were used, extracted from approximately 100 speakers (a total of 3864 training examples, and 1077 testing tokens). The classifiers we used for the vowel classification experiments were three variants of Bayesian minimum-error rate classifiers. The classifiers [8] assume multivariate Gaussian densities of the features. The results shown in Table 2 use the minimum Mahalanobis distance, plus a term to incorporate a-priori probabilities to make a decision. The reference CDFs for QCM were learned from TIMIT, and a 7th order polynomial was used to transform the incoming test features from NTIMIT to have the same pdfs as those of TIMIT. Both QCM and QCM in conjunction with CLN boost the performance of the recognizer by compensating for the channel noise and distortion in NTIMIT. Results for vowel classification with other distance measures showed a similar improvement in performance.

We also implemented the previously proposed [4,5] histogram equalization technique using Polya’s approximation to the normal [9]. Table 2 illustrates the general improvement of QCM over Histogram equalization (HEQ) because of the generalized transformation (not restricted to the single mode target pdf) in the implementation of QCM. Previous ASR work using histogram matching also assumes that the features are independent. In some pilot work, we performed PCA in the training and test phase in conjunction with HEQ. PCA decorrelated the feature space, both for training and test conditions, and was found to be slightly beneficial.

5. Conclusions

In this paper, we demonstrated the potential of “cleaning up” speech features distorted by noise and channel effects, using a non-linear transformation. The transformation is derived by Quantile based CDF matching and least-squares polynomial fitting. The polynomial can be determined from a small number of speech frames at the same SNR as the test speech and can then be applied to all test frames.