NONLINEAR TRANSFORMATIONS OF SPEECH FEATURES TO COMPENSATE FOR CHANNEL AND NOISE EFFECTS IN SPEECH RECOGNITION

Saurabh Prasad, Stephen Zahorian
Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, VA, USA
{sprasad, szahoria} @ odu.edu

ABSTRACT

A speech recognizer trained and tested with speech at the same SNR typically performs well. However, situations where the recognizer is trained with clean speech and used for recognizing noisy speech are commonly encountered and generally result in greatly degraded performance or lack of robustness. The features used for speech recognition setups are typically modeled by a multivariate Gaussian mixture pdf. However, additive noise and linear channel distortions alter the shape of the pdf. Hence, a recognizer trained with speech data having a multivariate Gaussian density will not perform well when the test data is non-Gaussian. Previous solutions addressing this problem restore the first two moments of the pdf (Cepstral Mean and Variance Normalization). In this paper, we propose a preprocessing step that restores the multivariate-Gaussian shape of the features representing corrupted speech. We evaluate the method with vowel classification experiments using TIMIT data for clean and the NTIMIT data for noisy speech.

1. INTRODUCTION

The effects of noise and linear 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 distorted by a simple environmental model consisting of additive noise and linear distortion. In the mismatched scenario, where the classifier is trained with clean speech and tested on noisy speech, the parameters of the trained system are not representative of the noisy speech. It has been shown that compensation for the effects of noise on the first two moments of the corrupted speech improves the recognizer's performance [3], [4].

In this paper, we propose a processing step that transforms the 'Non-Gaussian' features of the noisy speech to a multivariate Gaussian form using an appropriate non-linear transformation. The front-end first transforms each feature in the feature space to a marginally Gaussian density. This feature vector is further transformed using Principal Component Analysis (PCA) to de-correlate the features. The transformed feature vector comprised of marginally Gaussian, uncorrelated features, possesses a multivariate-Gaussian pdf. The non-linear transformation used for restoring the Gaussian pdf is obtained by the process of histogram matching (a commonly used technique in image contrast enhancement). Previous work in this direction documents the improvement in speech recognition performance by a marginal Gaussian transformation [3], in an HMM framework. Our contribution is an effort to further explore this approach and to investigate the utility of this approach to other problems using alternative frameworks (Maximum likelihood classifiers). Specifically, we explore the possibility of improving the performance of vowel classifiers using several different methods. We also suggest a simple approximation which enables the Gaussian transformation without using look-up tables. Finally, we also examine the effect of de-correlating the feature space in conjunction with the non-linear transformation, in an effort to insure that transformed feature space is multivariate Gaussian. Our experimental results lead to some new insights into the usefulness of this approach under various test conditions.

We begin by summarizing the effect of additive noise and linear channel distortions on the feature space comprised of log-energy coefficients. We then propose an efficient way to non-linearly transform the feature space so that the transformed features are multivariate Gaussian. This transformation was evaluated using vowel classification experiments and the improvement in the recognition performance of the classifier was found to be impressive when training and test data are mismatched.

2. EFFECTS OF NOISE AND LINEAR CHANNEL DISTORTION ON MFCC'S

The model for the environmental degradation [2] of speech is shown in figure 2.1. The effect of this model on the power spectral density of speech is given by

$$Y(\omega_k) = \left|H(\omega_k)\right|^2 X(\omega_k) + N(\omega_k) \qquad (1)$$

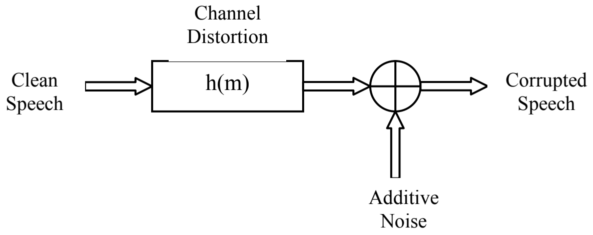

Figure 2.1: Model of the Environment

Taking the logarithm of both sides and defining the representations

$y(k) = 10 \log_{10} Y(\omega_k)$ (2)

$x(k) = 10 \log_{10} X(\omega_k)$ (3)

$n(k) = 10 \log_{10} N(\omega_k)$ (4)

$h(k) = 10 \log_{10}(|H(\omega_k)|^2)$ (5)

we obtain the following expression for the log-energy representation of the noisy speech

$$y(k) = x(k) + h(k) + 10\log_{10}(10^{\frac{x(k)+h(k)}{10}} + 10^{\frac{n(k)}{10}}) \quad (6)$$

Assuming that the features of the clean speech are normal $N_x(\boldsymbol{\mu_x},\Sigma_x)$, the above expression can be used to obtain the density function of y(k). It has been shown [2] that $p(\mathbf{y}|\ \boldsymbol{\mu_x},\ \Sigma_x,\ \mathbf{n},\ \mathbf{h})$ is not Gaussian. In fact, if we express the relation between **y** and **x** in the vector notation as

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{x},\boldsymbol{h},\boldsymbol{n}) \quad (7)$$

then, the mean vector and covariance matrix of the distorted feature space can be expressed as

$$\mu_y = \mu_x + E\{\boldsymbol{g}(\boldsymbol{x},\boldsymbol{h},\boldsymbol{n})\}$$

$$= \mu_x + \int_{\boldsymbol{X}} \boldsymbol{g}(\boldsymbol{x},\boldsymbol{h},\boldsymbol{n}) N_x(\mu_x, \Sigma_x) d\boldsymbol{x} \quad (8)$$

$$\Sigma_y = E\{(\boldsymbol{x}+\boldsymbol{g}(\boldsymbol{x},\boldsymbol{h},\boldsymbol{n}))(\boldsymbol{x}+\boldsymbol{g}(\boldsymbol{x},\boldsymbol{h},\boldsymbol{n}))^T\} - \mu_y \mu_y^T \quad (9)$$

Numerical methods are needed to solve the above equations to estimate the new mean vector and covariance matrix. Hence, any environmental compensation technique using the above equations will need to estimate a good model for the noise and channel distortion and then use numerical methods. Our compensation scheme circumvents this issue by conditioning the distorted feature space back to the multivariate Gaussian reference pdf by using an appropriate non-linearity and Principal Component Analysis.

In the process of deriving MFCCs from the above described log-energy coefficients, the features will tend to have their Gaussian pdf restored to some extent by virtue of the central limit theorem. However, since the log energy terms are highly correlated, in practice, the MFCCs of “non-clean” speech typically are not normal.

3. THE NON-LINEAR TRANSFORMATION

The non-linear transformation used to make the pdfs of the features Gaussian is based upon the principle of histogram matching. This is a popular technique used in image contrast enhancement applications, mapping probability densities of pixel intensities to a reference shape. The underlying theory is based on the transformation of random variables – given a random variable **x** with a pdf $p_x(x)$, we need to find an appropriate transformation $\mathbf{z} = T(\mathbf{x})$ such that **z** has a desired pdf $p_z(z)$. The pdfs of **x** and **z** in this case can be related by

$$p_z(z) = p_x(x)\left|\frac{dx}{dz}\right| \quad (10)$$

For the case where $p_z(z)$ is uniform (Histogram Equalization), and assuming the original density $p_x(x) = 0$ for $x < 0$), the transformation $T(\mathbf{x})$ is given by

$$z = T(x) = \int_0^x p_x(w)dw \quad (11)$$

For the general case, the algorithm for the transformation is as follows

1. ‘Equalize’ the levels of the input feature-data using the transformation described above.

   $$u = T(x) = \int_0^x p_x(w)dw \quad (12)$$

   The new random variable ‘u’ is then uniformly distributed.
2. Specify the desired density function $p_z(z)$ and obtain the transformation function

   $$v = G(z) = \int_0^x p_z(w)dw \quad (13)$$

   The new random variable, ‘v’ is also uniformly distributed.
3. Apply the inverse transformation function $z = G^{-1}(u)$ to the values obtained in step 1. The resulting random variable z will have the desired density function $p_z(z)$.

It is straightforward to derive an analytic expression for the transformation for cases when the desired density function is Exponential, Laplacian, Uniform etc. However, when the desired density function is Gaussian, the transformation in step 2 of the above algorithm cannot be analytically expressed. A good approximation to the function G^{-1} was given by Marsaglia [5]. Using Polya’s approximation to the normal he proposed the following analytic expression for generating a standard normal random variable **z** (density $ce^{-z^2/2}, 0 < z < 1$) from a uniform random variable:

$$z = -[1.553\ln(1-u^2)]^{1/2} \qquad (14)$$

where **u** is uniform between 0 and 1. We use this approximation for performing step 3 of our non-linear transformation.

The marginally Gaussian features obtained by the above mentioned preprocessing can further be made multivariate-Gaussian by de-correlating them. This follows from a basic result in probability theory that n uncorrelated marginally Gaussian random variables also exhibit a multivariate Gaussian pdf. To this effect, we used PCA to de-correlate features.

4. EXPERIMENTS AND RESULTS

The non-linear transformation described above was evaluated with vowel classification experiments. The 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. The extracted features were passed through the above pre-processing steps, transforming their pdfs to Gaussian.

The classifiers we used for the vowel classification experiments were three variants of Bayesian minimum-error rate classifiers (which we refer to MXL-1, MXL-2 and MXL-3). The classifiers [7] assume multivariate Gaussian densities of the features. For MXL-1 the features are assumed to be uncorrelated and to have equal variances. Thus the decision rule is based on the minimum Euclidean distance to class centroids, plus a term to incorporate the apriori probabilities of the classes. For the case of MXL-3, a common covariance matrix is assumed among all classes. Thus the decision is based on the minimum Mahalanobis distance, plus a term to incorporate apriori probabilities. For MXL-2, no assumptions are made other than multivariate Gaussian, and thus covariance matrices are computed separately for each class and used in the decision process.

Figure 4.1b illustrates the typical transformation applied to each feature in the feature vectors from both TIMIT and NTIMIT. A sample histogram of a feature extracted from the NTIMIT corpus (Fig. 4.1a) is seen to deviate considerably from Gaussian form. After the transformation of section 3 we can condition the distorted feature pdfs to a Gaussian shape. Features extracted from the TIMIT database are already very close to Gaussian (as expected). Hence their transformation is primarily a shift and linear scaling.

The above transformation was incorporated into the front-end of the various vowel classifiers. We first tested the matched scenario – where the speech data used for both training and testing was passed through similar noise and channel distortion. This matched case serves as a control for the case of interest, the mismatch scenario – where, the classifier was trained with clean speech and tested on noisy speech. In both cases, the transformation was done on the training and the test data separately with the reference histogram a Gaussian with a zero mean and standard deviation of 0.2. We have also examined the effect of feature space dimensionality on classification performance. This is particularly relevant for the investigation of the effects of the PCA transformation.

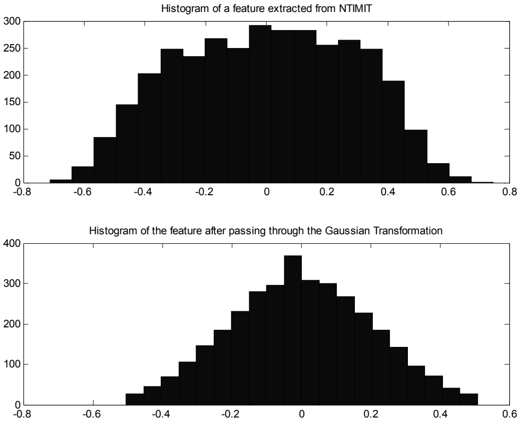

Fig. 4.1a: Sample histograms of a corrupted feature before and after transformation

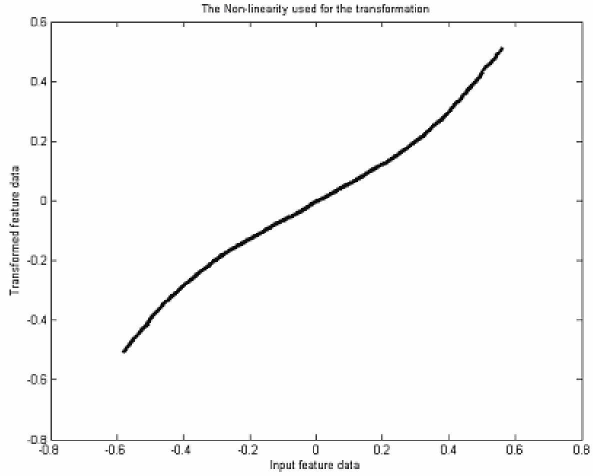

Fig. 4.1b: The nonlinearity used for the transformation

The effects of the non-linear transformation on the recognition results for both scenarios are illustrated by the data in tables 4.1 through 4.4. The baseline performance corresponds to classification without any preprocessing. Cepstral Mean and Variance Normalization (CMVN) represent compensation for the effects of noise and channel variations on the first and second central moments of the corrupted speech features. Table 4.1 illustrates the results in the matched case (using NTIMIT corpus). Though CMVN improves the recognition rate, the proposed nonlinear transformation consistently degrades the performance by a small amount. Tables 4.2 through 4.4 show the results of the nonlinear transformation on the MXL classifier for three different distance measures. The mismatch is seen to lead to very poor baseline results in all cases. CMVN improves the recognition rate somewhat. However the proposed histogram matching results in a dramatic improvement in performance. The use of PCA together with the nonlinear transformation results in further improvements in performance only when working in lower dimensional feature spaces.

No. Feat	Baseline	CMVN	histogram matching
5	51.8%	52.8%	49.7%
10	60.7%	63.2%	60.4%
14	62.3%	63.6%	61.4%

Table 4.1: Vowel Classification accuracy, illustrating the nonlinear transformation on the vowel classifier - **Matched case,** MXL-3

No. Feat.	Baseline	CMVN	Histogram matching	Histogram matching + PCA
5	25.3%	33.5%	41.9%	48.0%
10	28.9%	44.2%	52.9%	54.5%
14	29.4%	45.4%	54.1%	54.9%

Table 4.2: Vowel Classification accuracy, illustrating the nonlinear transformation on the vowel classifier - **Mismatched case,** MXL-1

No. Feat.	Baseline	CMVN	Histogram matching	Histogram matching + PCA
5	23.2%	23.4%	47.0%	49.9%
10	39.7%	39.7%	53.9%	54.5%
14	41.7%	41.7%	53.4%	53.8%

Table 4.3: Vowel Classification accuracy, illustrating the nonlinear transformation on the vowel classifier - **Mismatched case,** MXL-2

No. Feat	Baseline	CMVN	Histogram matching	Histogram matching + PCA
5	22.8%	22.7%	47.6%	51.0%
10	39.2%	39.4%	56.1%	57.3%
14	45.1%	45.0%	58.1%	58.2%

Table 4.4: Vowel Classification accuracy, illustrating the nonlinear transformation on the vowel classifier - **Mismatched case,** MXL-3

5. CONCLUSION

For mismatched speech data (clean versus telephone quality) the Gaussian histogram equalization significantly improves the recognizer's performance. Here, the non-linear transformation can be viewed as a conditioning process, restoring the discrimination information that was lost because of the mismatch. We can further conclude from our experiments that use of marginal Gaussian transformations results in significantly improving the recognition rate. The multivariate Gaussian transformation (using PCA after the marginal transformations) is only beneficial if lower dimensional feature spaces are used.

The proposed transformation scheme does not improve the recognition performance in the situation when the training and testing is done by speech at the same SNR. In fact, the transformation in this case actually degrades the recognition of the test-data by a small amount. Such a behavior can be attributed to a fundamental issue in pattern classification [8], that any transformation of a feature space can not improve discrimination. Improvements that were obtained in this work were possible only because separate nonlinear transformations were applied to the training and test data.

In additional work, we plan to approximate the nonlinear transformation using a polynomial curve fitting to the histogram derived transformation nonlinearity. In the current implementation, there is a slightly quantization effect in the transformed features, which might have a small adverse effect on performance. More fundamentally, a separate nonlinear transformation should be computed for each channel (.i.e., for each sentence in the case of NTIMIT), rather than single nonlinear transformation for all of the testing data.

6. REFERENCES

[1] Alejandro Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*, PhD Thesis, ECE Dept., Carnegie Melon University, 1990.

[2] Pedro J Moreno. *Speech Recognition in Noisy Environments*, PhD Thesis, ECE Dept., Carnegie Melon University, 1996.

[3] Angel de la Torre, Jose C Segura, Carmen Benitez, Antonio M Peinado, Antonio J Rubio. "Non-Linear Transformations of the feature space for robust speech recognition", *Proceedings of ICASSP-2002*, pp. I-401 – I-404.

[4] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, *Spoken Language Processing*, Prentice Hall, 2001.

[5] George Marsaglia, "The Exact Approximation method for generating Random Variables in a Computer", *Journal of the American Statistical Association*, Vol. 79, No. 385 (Mar. 1984), pp. 218-221.

[6] S. A. Zahorian and Z. B. Nossair, "A Partitioned Neural Network Approach for Vowel Classification Using Smoothed Time/Frequency Features," *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 4, pp. 414-425, July 1999.

[7] Richard O. Duda, Peter E. Hart, David G. Stork. *Pattern Classification,* Second Edition. Wiley Interscience, 2000.

[8] Joseph P. Campbell, JR., "Speaker Recognition: A Tutorial", *Proceedings of the IEEE,* vol. 85, NO. 9, pp. 1437-1462, September 1997.