MULTIVARIATE-GAUSSIAN-BASED CEPSTRAL NORMALIZATION FOR ROBUST SPEECH RECOGNITION

Pedro J. Moreno, Bhiksha Raj[1], Evandro Gouvêa and Richard M. Stern
Department of Electrical and Computer Engineering & School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

In this paper we introduce a new family of environmental compensation algorithms called Multivariate Gaussian Based Cepstral Normalization (RATZ). RATZ assumes that the effects of unknown noise and filtering on speech features can be compensated by corrections to the mean and variance of components of Gaussian mixtures, and an efficient procedure for estimating the correction factors is provided. The RATZ algorithm can be implemented to work with or without the use of "stereo" development data that had been simultaneously recorded in the training and testing environments. "Blind" RATZ partially overcomes the loss of information that would have been provided by stereo training through the use of a more accurate description of how noisy environments affect clean speech. We evaluate the performance of the two RATZ algorithms using the CMU SPHINX-II system on the alphanumeric census database and compare their performance with that of previous environmental-robustness developed at CMU.

1. INTRODUCTION

As speech recognition systems become more accurate and sophisticated, robustness with respect to noise, channel effects, and other perturbations caused by the acoustical environment becomes increasingly important. Over the past few years, researchers at CMU and other sites have developed a series of techniques to address this problem (*e.g.*[1,2,3,4]). These techniques can be classified into two broad groups, data-driven techniques such as *multiple fixed codeword-dependent cepstral normalization* (MFCDCN) [2] and model based techniques such as the original *codeword-dependent cepstral normalization* algorithm (CDCN) [1]. The data-driven algorithms make few assumptions about the effects of the environment on the speech cepstra. They rely on empirical comparisons of the acoustical characteristics of speech that is simultaneously recorded using a close-talking microphone (CLSTLK) and in the target environment. Such databases are commonly referred to as "stereo" data. The second class of techniques assumes a particular structural model of the acoustical degradation. For example, much of the work of Acero and colleagues (*e.g.* [1]) assumes that degraded speech can be modeled as "clean" speech that is corrupted by additive noise and linear filtering. Approaches that make use of a structural model of degradation often use maximum likelihood methods to learn the effects of the environment. They do not require the use of stereo data.

The algorithms we propose in this paper can be looked on as a combination of some of the best features of empirical compensation procedures like MFCDCN and approaches which use structural models of degradation like CDCN. In the initial formulation, the proposed methods perform compensation based on empirical comparisons, like MFCDCN, but using the more formal representation of probability densities and the optimal estimation procedures that were used in previous model based procedures like CDCN. Nevertheless, there is no explicit model for environmental degradation (unlike the model based approaches). We only assume that the environment modifies some of the parameters used to describe the feature distributions of clean speech.

Our new techniques can exploit the information provided by stereo data if available. However, stereo databases are not always easy to collect. We will demonstrate that the representational structure of the algorithms permit nearly-optimal compensation, even in the absence of stereo data.

In Sec. 2 we describe the effects of environmental degradation on the probability density functions (pdfs) of the feature vectors used for recognition. The new algorithms are described in Sec. 3, and they are evaluated using simulated and real speech data in Sec. 4.

1: Tata Institute of Fundamental Research
CSC Group, Bombay, India

2. EFFECT OF THE ENVIRONMENT ON SPEECH STATISTICS

In this section we describe how even well-behaved environments, such as those modeled by linear channels and additive stationary noise, modify the statistics of clean speech in very unpredictable ways. Even though we can formulate equations that analytically describe how the pdfs of clean speech change, the solutions for these equations are mathematically intractable. For this reason we model the effects of the environment as changes in the parameters of the statistics that characterize clean speech while keeping the same distribution structure.

For analytical purposes, we adopt the simple model of degradation proposed by Acero [1]. In this model, degraded speech is characterized by passing clean speech through a linear channel and contaminating the filtered output by additive stationary noise. For simplicity, we will also assume that the feature vector is unidimensional, although all conclusions developed can be easily expanded to an arbitrary *N*-dimensional space such as the log spectral domain.

In the power spectral domain the degraded speech can be expressed as:

$$Z(\omega) = X(\omega) * |H(\omega)|^2 + N(\omega) \quad (1)$$

where $Z(\omega)$ represents the power spectrum of the noisy speech, $X(\omega)$ is the power spectrum of the clean speech, $H(\omega)$ is the transfer function of the linear channel, and (ω) is the power spectrum of the additive noise. In the log-spectral domain this relation can be expressed as:

$$z = x + q + log(1 + e^{n-x-q}) \quad (2)$$

$$r(x, n, q) = log(1 + e^{n-x-q}) \quad (3)$$

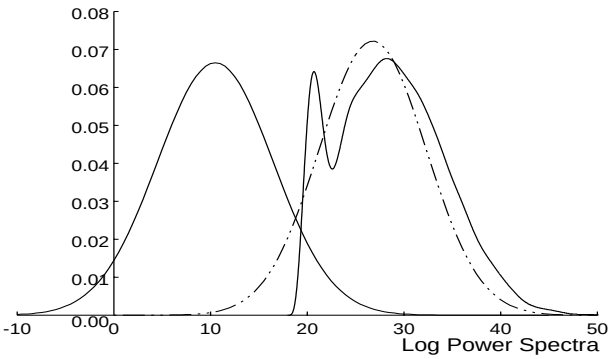

Figure 1. Effect of noise on pdfs of *x* and *z*. The curve on the left represents the pdf of *x*. The dashed curve represents the pdf for *z* that would be obtained if *z* were assumed to be Gaussian. The solid curve is the actual pdf for *z*, obtained using Monte Carlo simulations.

where z, x, q, and n represent the logs of $Z(\omega)$, $X(\omega)$, $|H(\omega)|^2$, and (ω), respectively, for some particular value of ω.

Assuming knowledge of the pdf of the clean speech, $p(x)$, and its mean μ_x and variance σ_x^2, the effect of the degradation will affect the mean and variance of *z* in the following manner:

$$\begin{aligned}\mu_z &= E[z] = \mu_x + q + \mu_{r(x,n,q)} \\ \sigma_z^2 &= E[(z-\mu_z)^2] = \sigma_x^2 + \sigma_{r(n,x,q)}^2 \\ &\quad + 2[E\{x\, r(x,n,q)\} - \mu_x \mu_{r(x,n,q)}]\end{aligned} \tag{4}$$

For simplicity we assume that *x* is Gaussian, and we assume that the power spectrum of the noise and the transfer function of the channel are known and deterministic. In this simplified special case the new equations for the mean and variance are:

$$\mu_z = \mu_x + q + \int_X N_x(\mu_x, \sigma_x)\, r(x,n,q)\, dx \tag{5}$$

$$\sigma_z^2 = \int_X N_x(\mu_x, \sigma_x)\, (r(x,n,q))^2 dx - \mu_z^2 \tag{6}$$

We are not aware of any analytical solutions for these equations. In fact the distribution of *z* can be shown to be non Gaussian:

$$p(z) = \frac{(1-e^{n-z})^{-1}}{\sqrt{2\pi}\sigma_x} e^{\frac{(z-q+\log(1-e^{n-z})-\mu_x)^2}{-2\sigma_x^2}} \tag{7}$$

Equations (5) and (6) were obtained under the unrealistic assumption that (ω) is known *a priori*. In practice, (ω) must be estimated, producing a random estimate for *n* to which we assign the pdf $p(n)$. Assuming that *n* and *x* are statistically independent, the new expression for μ_z becomes:

$$\mu_z = \mu_x + q + \int_X N_x(\mu_x, \sigma_x) \int_N N_n(\mu_n, \sigma_n)\, r(x,n,q)\, dx\, dn \tag{8}$$

Figure 1 shows the effect of the non-linear relationship between noise, channel and the clean signal on the signal statistics. The curve on the left represents the pdf of *x*, which is assumed to have a mean of 10.5 and a standard deviation of 6. The data were passed through a channel of value $q = 8$ and contaminated with Gaussian noise with mean 6 and standard deviation 0.02. The solid curve in Figure 1 is the actual pdf for *z*, obtained using Monte Carlo simulations.

The dashed curve in Figure 1 describes the pdf for *z* that would be obtained if *z* were assumed to be Gaussian. If the means of *n* and *x* are close to one another, the Gaussian assumption tends to be inaccurate. Nevertheless, we will adopt the Gaussian assumption throughout this paper for the simplicity that it provides.

Figure 2 describes the effects of corruption by noise on the variance of the degraded signal. The curves are plotted for three values of the relation $\mu_x - \mu_n$, 0, 10, and 20 dB. We note that the range of variances of the corrupted signal is compressed as σ_n^2/σ_x^2 increases. The variance of *z* increases as the variance of *n* increases, but it is always lower than the original variance.

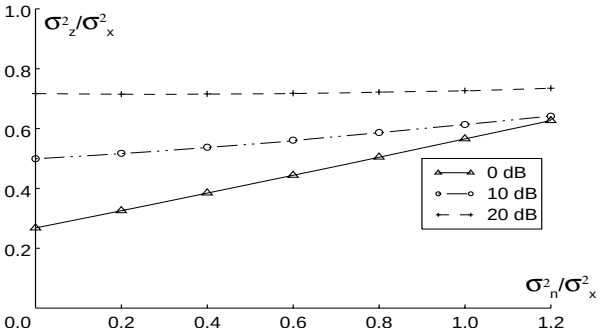

Figure 2. Effect of noise on the variance of *y*. The curves are plotted for three values of the relation $\mu_x - \mu_n$, 0, 10, and 20 dB.

We conclude that when we assume that the corrupted distributions have a normal shape, the effects of the environment on signal statistics can be modeled by additive correction terms to:

- the mean of *z*, thus shifting its pdf
- the variance of *z*, thus compressing its pdf

In the algorithms developed in this paper the effects of noise and filtering are modeled by corrections to μ_z and σ_z^2.

3. COMPENSATION USING RATZ

In this section we describe the new compensation algorithms, which are referred to as Multiva**r**iate-Gaussian-Based Cepst**r**al Normali**z**ation (RATZ). We first describe the version of RATZ that exploits the empirical differences between clean and degraded speech in stereo databases. We then describe how similar compensation can be effected without the need for stereo training data.

3.1. RATZ using stereo databases

The implementation of RATZ that exploits the information in stereo data ("stereo RATZ") assumes that the statistics of speech can be represented by a multivariate Gaussian mixture distribution. It also assumes that the effects of the environment on the statistics of clean speech can be modeled as additive compensations for mean vectors and covariance matrices.

The algorithm works in three following stages which are describes as follows:

- Estimation of the statistics of clean speech

- Estimation of the statistics of noisy speech
- Compensation of noisy speech

Estimation of the statistics of clean speech. The pdf for the features of clean speech is modeled as a mixture of multivariate Gaussian vectors. Under these assumptions the distribution of clean speech can be written as:

$$\begin{aligned} \boldsymbol{x} &= \left[x_0 \ \ldots . x_{P-1} \ x_P \right]^T \\ p(\boldsymbol{x}) &= \sum_{k=0}^{M-1} P[k] \, N_x(\mu_{x,k}, \Sigma_{x,k}) \end{aligned} \tag{9}$$

where $P[k]$, $\mu_{x,k}$ and $\Sigma_{x,k}$ represent respectively the *a priori* probabilities, mean vector and covariance matrix of each multivariate Gaussian mixture element *k*. These parameters are learned through traditional maximum likelihood methods.

Estimation of the statistics of noisy speech. The effects of the environment on the statistics of clean speech are modeled as:

$$\mu_{z,k} = \mu_{x,k} + \boldsymbol{r}_k \qquad \Sigma_{z,k} = \Sigma_{x,k} + R_k \tag{10}$$

resulting in a new set of statistics describing the noisy speech vector $\boldsymbol{z}$. We assume that the *a posteriori* probabilities of the mixtures $P(k|\boldsymbol{z}_i)$ are a feature of the speech distribution and are not affected by the channel. While this assumption is not strictly correct, it is convenient and reasonable as a first order approximation.

The shift parameters r_k and R_k are learned using a traditional maximum likelihood approach that attempts to maximize the probability that the observed noisy data set is generated by the transformed statistics. We define a likelihood function, *L(Z)* over all the noisy observed cepstral vectors $\boldsymbol{z}_i$ given the unknown parameters to optimize , θ, as:

$$L(Z = \{\boldsymbol{z}_0, \ldots . \boldsymbol{z}_i \ldots . \boldsymbol{z}_{N-1}\} \,|\, \theta) = \sum_{i=0}^{N-1} \log\left(p(\boldsymbol{z}_i|\,\theta)\right) \tag{11}$$

To find the set of parameters θ that maximize $L(Z)$ we can use stereo data and the assumption that the *a posteriori* probabilities $P(k|\boldsymbol{z}_i)$ do not change due to the environment. The latter assumption enables us to avoid the iterative reestimation needed in traditional EM techniques when stereo data are not available.

The modified estimation formulas for the correction terms are:

$$\boldsymbol{r}_k = \frac{\sum_{i=0}^{N-1} (\boldsymbol{z}_i - \boldsymbol{x}_i) \, P(k|\boldsymbol{x}_i)}{\sum_{i=0}^{N-1} P(k|x_i)} \tag{12}$$

$$R_k = \frac{\sum_{i=0}^{N-1} (z_i - \mu_{z,k})(z_i - \mu_{z,k})^T P(k|x_i)}{\sum_{i=0}^{N-1} P(k|x_i)} \tag{13}$$

Compensation of noisy speech. This step is accomplished using a modified minimum mean square error (MMSE) estimation method. The estimator attempts to maximize the expected value of the unobserved clean speech data:

$$\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{r}(\boldsymbol{x})$$

$$\begin{aligned} \hat{\boldsymbol{x}} = E[\boldsymbol{x}|\,\boldsymbol{z}] &= \int_{\boldsymbol{X}} \boldsymbol{x} \; p(\boldsymbol{x}|\,\boldsymbol{z}) \, d\boldsymbol{x} = \boldsymbol{z} - \int_{\boldsymbol{X}} \boldsymbol{r}(\boldsymbol{x}) \, p(\boldsymbol{x}|\,\boldsymbol{z}) \, d\boldsymbol{x} \\ \hat{\boldsymbol{x}} &\cong \boldsymbol{z} - \sum_{k=0}^{M-1} P(k|\,\boldsymbol{z}) \, \boldsymbol{r}_k \end{aligned} \tag{14}$$

Since the dependence of the correction factor ***r(x)*** on $\boldsymbol{x}$ makes the equation intractable, we approximate $\boldsymbol{r}(\boldsymbol{x})$ by $\boldsymbol{r}_k$, the correction factor associated with the mean of each Gaussian mixture.

A novel feature of this method is that it attempts to update the covariance matrix to reflect more accurately the effect of the environment on the speech statistics.

3.2. RATZ without stereo databases

In this section we extend RATZ to conditions where no stereo data are available ("blind RATZ"). This entails only minimal changes in the reestimation formulas for the shifts in means and covariance matrix elements. In this case we do not have information about the *a posteriori* probabilities since no stereo data are available. Therefore $P(k|\,\boldsymbol{z}_i)$ cannot be replaced by $(k|\,\boldsymbol{x}_i)$ in the estimation algorithm. Using normal EM techniques we iteratively estimate $P(k|\,\boldsymbol{z}_i)$ until convergence is achieved.

The new reestimation formulas are:

$$\hat{\boldsymbol{r}}_k^{l+1} = \frac{\sum_{i=0}^{N-1} z_i \hat{P}^l(k|\,z_i)}{\sum_{i=0}^{N-1} \hat{P}^l(k|\,z_i)} - \mu_{x,k}$$

$$\hat{R}_k^{l+1} = \frac{\sum_{i=0}^{N-1} (\boldsymbol{z}_i - \hat{\boldsymbol{r}}_k^l - \mu_{x,k})(\boldsymbol{z}_i - \hat{\boldsymbol{r}}_k^l - \mu_{x,k})^T \hat{P}^l(k|\,\boldsymbol{z}_i)}{\sum_{i=0}^{N-1} \hat{P}^l(k|\,\boldsymbol{z}_i)} - \Sigma_{x,k}$$

The compensation step remains the same as in stereo RATZ.

4. EXPERIMENTAL RESULTS

In this section we compare the ability of stereo RATZ and blind RATZ to learn the statistics of degraded speech. We show that clean signal vectors estimated using stereo RATZ and blind RATZ are very similar, despite the more limited information available to the blind RATZ algorithm. Finally, we compare the recognition accuracy obtained using the two RATZ algorithms with the performance of previous noise robustness algorithms developed at CMU.

4.1. Performance of RATZ in simulated noise

Artificial feature data were produced by a random number generator creating two-dimensional sample vectors with four equiproba-

ble mixture Gaussian distributions with the following means and variances:

$$\mu_1 = [1 \quad 1] \qquad \mu_2 = [1 \quad -1]$$
$$\mu_3 = [-1 \quad -1] \qquad \mu_4 = [-1 \quad 1]$$

$$\Sigma_i = \begin{bmatrix} 0.5625 & 0.0 \\ 0.0 & 0.5625 \end{bmatrix} \qquad i = 1, ...,4$$

The data were contaminated by artificial noise which had a two-dimensional Gaussian distribution with a mean value of 0.50 and a variance of 0.001 for both dimensions. The covariance matrix was diagonal.

Figure 3 compares the effectiveness of both algorithms in estimating clean speech vectors in the presence of this noise. The filled circles in the two panels of Figure 3 indicate the locations of the clean speech vectors, and the filled squares indicate the locations of the noisy speech before compensation was applied. The dashed arrows and continuous arrows indicate the corrections provided by the stereo RATZ and blind RATZ algorithms, respectively. It can be seen that blind RATZ provides a compensation that is almost as complete as that provided by stereo RATZ. Similar results have been observed on real speech statistics.

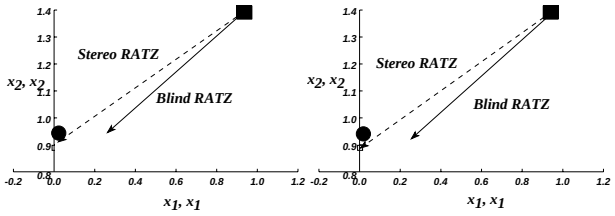

Figure 3. Comparison of compensation vectors obtained using stereo RATZ and blind RATZ. Filled circles and filled squares denote the locations of clean speech and uncompensated degraded speech, respectively.

4.2. Performance of RATZ on speech recognition in noise

The effectiveness of the RATZ algorithms was evaluated using the CMU census database [1], a continuous speaker-independent database consisting of strings of letters, numbers, and a few control words with a total vocabulary size of 107 words. The training set consisted of 1018 sentences stereo recorded over a noise-canceling close-talking microphone (CLSTK) and the desktop Crown-PZM6FS microphone (CRPZM). The testing set consisted of 140 stereo-recorded sentences. The adaptation set used by all compensation algorithms had a size of 400 stereo sentences randomly chosen from the training set. The SPHINX-II continuous speech recognition system was used.

In Table 1 we compare the recognition error rate of several versions of RATZ to the error rates of previous algorithms developed at CMU [1,2]. The system was trained on clean speech and the adaptation set was used to learn the noisy speech distributions. Compensation algorithms were applied to the noisy data before recognition. A recognition error rate of 12.4% was achieved using the CLSTK microphone with no compensation.

It can be seen that the RATZ algorithms perform better than virtually all of the previous CMU algorithms. We also note that accounting for the change in the variance improves performance. Finally, blind RATZ, which does not make explicit use of the stereo training data performs almost as well as stereo RATZ, and performs better that some of the previous algorithms that use stereo training, such as SDCN.

Compensation Algorithm	%ERROR
NONE	32.6
SDCN	27.0
FCDCN	22.0
CDCN	25.2
Blind RATZ without variance comp.	25.1
Blind RATZ with variance comp.	24.5
Stereo RATZ with variance comp.	21.2

Table 1. Comparison of stereo RATZ and blind RATZ with other algorithms developed at CMU.

5. SUMMARY

In this paper we introduce a new family of algorithms to deal with the problem of speech recognition in noisy environments. We analyze the effects of noise on the statistics of common speech features, and we note that the actual pdfs of the features are frequently non-Gaussian. Nevertheless, we model the speech features as Gaussian vectors with means that were shifted and variances that are compressed in the presence of noise. We compare the performance of two implementations of the algorithm, stereo RATZ and blind RATZ, and we demonstrate that blind RATZ can perform at a comparable level to stereo RATZ without the need for training with stereo databases. The algorithms were tested using the non-closetalking alphanumeric census database and found to provide superior error-rate reduction.

ACKNOWLEDGEMENTS

The authors thank Alex Acero, Uday Jain, and Fu-Hua Liu for useful suggestions. Pedro J. Moreno has been supported by a Fulbright fellowship awarded by the Ministerio de Educación y Ciencia, Spain. Bhiksha Raj has been supported by a United Nations Development Program fellowship. Evandro Gouvêa has been supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil. This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

1. Acero A. (1990). Acoustical and Environmental Robustness in Automatic Speech Recognition. Ph. D. Dissertation, ECE Department, CMU, Sept. 1990

2. Fu-Hua Liu. (1994). Environmental Adaptation for Robust Speech Recognition. Ph. D. Dissertation, ECE Department, CMU, July 1994.

3. Neumeyer L. and Weintraub M. (1994). Probabilistic Optimun Filtering for Robust Speech Recognition. ICASSP-94.

4. Fu-Hua Liu, Acero A. and Stern R. M. (1992). Efficient Joint Compensation of Speech for the Effect of Additive Noise and Linear Filtering. ICASSP-92.

5. McLachlan G. J. and Dasford K. E. “Mixture Models: Inference and Applications to Clustering”. M. Dekker (1988).