APPROACHES TO ENVIRONMENT COMPENSATION IN AUTOMATIC SPEECH RECOGNITION

Pedro J. Moreno, Bhiksha Raj, Richard M. Stern

Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.

ABSTRACT

This paper describes a series of cepstral-based compensation procedures that render the SPHINX-II continuous speech recognition system more robust with respect to acoustical changes in the environment. The first two algorithms, SNR based MultivaRiate gAussian based cepsTral normaliZation (SNR-based RATZ) and STAtistical Reestimation of HMMs (STAR), compensate for environmental degradation based on comparisons of simultaneously-recorded data in the training and testing environments ("stereo data"). They differ in that RATZ modifies the incoming feature vectors to a recognition system while STAR modifies the internal representation of speech by the system. We also describe N-CDCN, an improved version of codeword-dependent cepstral normalization (CDCN) which does not require stereo training data but nevertheless achieves performance levels comparable to RATZ and other algorithms that require stereo training. Use of these compensation algorithms significantly reduces the error rates for SPHINX-II. The algorithms are tested in a variety of databases and environmental conditions.

INTRODUCTION

Robustness with respect to environmental variability remains a continuing problem for speech recognition technology (*e.g.* [3]). For example, the use of microphones other than the ARPA standard Sennheiser HM-414 headset (CLSTLK) severely degrades the performance of speech recognition systems like the SPHINX-II, even in relatively quiet environments [1, 4].

Traditional algorithms to compensate for environmental variation have either relied on the availability of simultaneously-recorded data in the training and testing environments ("stereo data"), or have utilized structural models to define the degradation. For example, multiple fixed codeword-dependent cepstral normalization (MFCDCN) [4] uses stereo data to compute correction vectors to compensate for the effects of the environment. Dual-channel codebook adaptation (DCCA) [4], on the other hand, modifies the statistical representation used in the HMMs for speech on the basis of comparisons obtained from stereo data. The other approach to environmental compensation is through the use of structural models of degradation. On the other hand, codeword-dependant cepstral normalization (CDCN) [1] assumes that speech is degraded by unknown additive noise and unknown linear filtering. It makes use of expectation-maximization (EM) techniques to determine the parameters characterizing these distortions.

In this paper we describe three new cepstral-domain compensation strategies, SNR based MultivaRiate gAussian based cepsTral normaliZation (SNR-based RATZ), STAtistical Reestimation of HMMs (STAR), and new CDCN (N-CDCN). SNR-based RATZ and STAR both make use of stereo data. They differ in that RATZ modifies the incoming feature vectors to a recognition system while STAR modifies the internal representation of speech by the system. RATZ and STAR are similar in philosophy to MFCDCN and DCCA, respectively [4], but they achieve improved performance through the use of better mathematical models which introduce strong structural constraints into the assumed distribution for speech. N-CDCN is a modification and improvement of the original CDCN algorithm, which is based on a structural model of degradation.

EFFECT OF THE ENVIRONMENT ON SPEECH STATISTICS

In this section we describe how even well-behaved environments, such as those that can be modeled by unknown linear filtering 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 analytical purposes, we adopt the simple model of degradation proposed by Acero [1]. In this model, degraded speech is characterized by passing high-quality clean speech through a linear filter 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 extended to an arbitrary *N*-dimensional space such as the log spectral domain. The degraded speech can be characterized as:

$$Z(\omega) = X(\omega)\,|H(\omega)|^2 + N(\omega) \tag{1}$$

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

$$z = x + q + log\,(1 + e^{n-x-q}) \qquad r(x, n, q) = log\,(1 + e^{n-x-q}) \tag{2}$$

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

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

$$\mu_z = E[z] = \mu_x + q + \mu_{r(x,n,q)} \qquad \sigma_z^2 = E[(z-\mu_z)^2] = \sigma_x^2 + \sigma_{r(n,x,q)}^2 + 2\,[E\{x\,r(x,n,q)\} - \mu_x\,\mu_{r(x,n,q)}] \tag{3}$$

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

$$\mu_z = \mu_x + q + \int_X N_x(\mu_x, \sigma_x)\, r(x,n,q)\, dx \qquad \sigma_z^2 = \int_X N_x(\mu_x, \sigma_x)\,(x + q + r(x,n,q))^2 dx - \mu_z^2 \tag{4}$$

These equations become difficult to solve. In fact we are not aware of any analytical solutions for these equations.

Equations (4) were obtained under the unrealistic assumption that (ω) is known *a priori* and deterministic. 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:

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

This equation is more difficult to solve than equations (4). The computation for σ_z becomes even more complicated.

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), and the variance of *z* (thus compressing its pdf). This is the approach that is followed in the SNR-based RATZ and STAR algorithms described in this paper.

COMPENSATION ALGORITHMS

In this section we briefly describe the three new algorithms RATZ, N-CDCN and STAR. SNR-based RATZ and STAR assume the availability of a database of "stereo" training sentences, with one channel containing speech recorded using a high-quality microphone and a second channel containing degraded speech samples.

SNR-based RATZ

SNR-based RATZ uses frame energy information, represented by the zero^th component of the cepstral vector (x_0), to model the statistics of the cepstra of the clean speech hierarchically: the x_0 cepstral coefficients are defined to have a Gaussian mixture distribution, and the cepstral vectors corresponding to each x_0 Gaussian are further assumed to have a mixture Gaussian distribution:

$$x = \begin{bmatrix} x_0 & x_P^T \end{bmatrix}^T \qquad x_P = \begin{bmatrix} x_1 & \ldots\mathrm{x}_{P-1} & x_P \end{bmatrix}^T \qquad p(x) = \sum_{k=0}^{N-1} P[k]\, N_{x_0}(\mu_{x_0,k}, \sigma_{x_0,k}) \sum_{j=0}^{M-1} P[j|\,k]\, N_{x_P}(\mu_{x_P,jk}, \Sigma_{x_P,jk}) \tag{6}$$

where $P[k]$, $\mu_{x_0,k}$and $\sigma_{x_0,k}$ represents the *a priori* probability, mean and variance of each energy component, and $P[j|\,k]$, $\mu_{\boldsymbol{x}_P,jk}$and $\Sigma_{\boldsymbol{x}_P,jk}$ represent the *a priori* probability, mean vector and covariance matrix of each multivariate Gaussian mixture conditioned on the frame energy. These parameters are learned through traditional EM methods.

The effects of the environment on the statistics of clean speech are modeled as shifts in the means and variances:

$$\mu_{z_0,k} = \mu_{x_0,k} + r_{0,k} \qquad (\sigma_{z_0,k})^2 = R_{x_0,k} + (\sigma_{0,k})^2 \qquad \mu_{\boldsymbol{z}_P,jk} = \mu_{\boldsymbol{x}_P,jk} + \boldsymbol{r}_{P,jk} \qquad \Sigma_{\boldsymbol{z}_P,jk} = \Sigma_{\boldsymbol{x}_P,jk} + R_{P,jk} \tag{7}$$

resulting in a new set of statistics describing the degraded speech vector $\boldsymbol{z}$.

The shift parameters 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. When stereo data are available the reestimation formulas use it by modelling the *a posteriori* probabilities using the stereo clean vectors.

The degraded speech is compensated by using an MMSE technique to shift them back to the clean speech statistics:

$$\boldsymbol{r}_{jk} = [r_{0,k}, \boldsymbol{r}_{P,jk}^{T}]^{T} \qquad \hat{\boldsymbol{x}} = E(\boldsymbol{x}|\boldsymbol{z}) = \int_{\boldsymbol{X}} \boldsymbol{x}.p(\boldsymbol{x}|\boldsymbol{z})\,d\boldsymbol{x} = \boldsymbol{z} - \int_{\boldsymbol{X}} \boldsymbol{r}(\boldsymbol{x})\,p(\boldsymbol{x}|\boldsymbol{z})\,d\boldsymbol{x} \cong \boldsymbol{z} - \sum_{k}\sum_{j} P(j,k|\boldsymbol{z})\,\boldsymbol{r}_{jk} \tag{8}$$

STAR

STAR assumes that any effect of noise, channel, or environment can be accurately formulated as shifts in the means and corrections to the variances of HMMs:

$$\mu_{\mathbf{z},k} = \mu_{\mathbf{x},k} + \boldsymbol{r}_k \qquad \Sigma_{\mathbf{z},k} = \Sigma_{\mathbf{x},k} + R_k \tag{9}$$

This is approach is similar to that in [2]. This model is applied to all of the four streams of data that SPHINX-II uses, that is, cepstral, delta-cepstral, double delta-cepstral coefficients, and a fourth three-dimensional stream that contains the cepstral component c_0, its difference Δc_0, and its double difference $\Delta^2 c_0$.

Correction factors for the statistics of each of these four streams are computed using the following modified Baum-Welch estimation formulas:

$$\hat{\boldsymbol{r}}_{\mathbf{x},k} = \left\{ \sum_{t=0}^{T-1} \gamma_t(k)\,(\boldsymbol{z}_t - \boldsymbol{x}_t) \right\} \Big/ \sum_{t=0}^{T-1} \gamma_t(k) \tag{10}$$

where $\gamma_t(k)$, the probability of being in state k at time t is conditioned on statistics λ of the *clean* speech channel of the stereo training data. We assume that the *a priori* probabilities do not change due to the effects of noise or environment. We note that $\gamma_t(k)$ is computed using clean speech and clean models. As a result, it does not change from estimation to estimation, and no iteration is needed.

The estimation formulas for corrections to the covariance matrices are similar:

$$\hat{R}_k = \left(\sum_{t=0}^{T-1} \gamma_t(k)\,(\boldsymbol{z}_t - \mu_{\mathbf{x},k} - \hat{\boldsymbol{r}}_k)\,(\boldsymbol{z}_t - \mu_{\mathbf{x},k} - \hat{\boldsymbol{r}}_k)^T \right) \Big/ \left(\sum_{t=0}^{T-1} \gamma_t(k) \right) - \Sigma_{\mathbf{x},k} \tag{11}$$

N-CDCN

New Codebook-Dependent Cepstral Normalization (N-CDCN) is an improved version of its predecessor, CDCN. The original CDCN has been able to achieve a respectable amount of error reduction [1] for many types of acoustical degradations without requiring the use of stereo training data. Nevertheless, it has not found much use in current CMU speech systems because it requires empirical information and retraining of the HMMs. N-CDCN alleviates these problems while retaining CDCN's ability to compensate for the combined effects of additive noise and distortion produced by unknown linear filtering. As before, compensation is performed on a sentence-by- sentence basis.

N-CDCN assumes a structural model of the degradation where speech is contaminated by additive stationary noise after being filtered by an unknown linear filter. This can be expressed as:

$$z[k] = x[k] * h[k] + n[k]$$

A statistical description of the cepstral space characterizing "clean" speech, as parametrized by a mixture of multivariate Gaussians, is estimated by EM methods. Given a noisy utterance, parametrized by a sequence of cepstral vectors, and given the previously learned statistics describing the clean speech, N-CDCN iteratively estimates noise and linear-filtering vectors that maximize the likelihood of observing the degraded speech. The initial values of the filter and noise estimates are the values that maximize the likelihood in the case of no estimation error.

Finally, an MMSE technique is used to estimate the unobserved clean speech vectors given the observed degraded speech, the previously-estimated noise and filter vectors, and the statistics describing clean speech.

PERFORMANCE ANALYSIS

In this section we describe the results of a series of experiments that compare the recognition accuracy of the algorithms described in Sec. with previous algorithms developed at CMU. The experiments measure recognition accuracy obtained using speech from the CENSUS database [1] that was recorded using the omnidirectional desktop Crown PZM 6FS microphone. The CENSUS database consists of strings of letters and digits stereo that were simultaneously recorded using the close-talking Sennheiser HMD-414 microphone and the Crown PZM6FS desktop microphone.

Table 1 shows the results of these experiments. The system was trained on clean speech from the close-talking Sennheiser HMD-414 microphone and tested using noisy speech from the desktop Crown PZM-6FS microphone. The word error rate of the SPHINX-II system when training and testing on clean speech was 14.7%. It can be seen that SNR-based RATZ and STAR provide slightly better performance than FCDCN. The performance of N-CDCN is better than that of its predecessor, CDCN, and it has the additional advantage of not requiring that the system be retrained.

Compensation Method	Stereo Data Needed?	Error Rate
CMN	No	29.5%
FCDCN	Yes	21.9%
STAR	Yes	21.5%
SNR-based RATZ	Yes	21.0%
CDCN	No	24.3%
N-CDCN	No	20.7%

Table 1: Word error rates obtained using speech from CENSUS database using different compensation algorithms in conjunction with SPHINX-II.

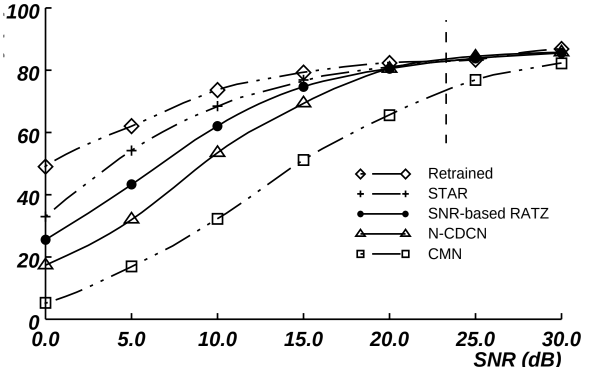

Figure 1. Comparison of recognition accuracy obtained using STAR, SNR-based RATZ, and N-CDCN with no compensation at all (CMN) and with full retraining using noisy speech data.

In a separate set of experiments we evaluated the performance of the various compensation algorithms in low signal-to-noise ratio (SNR) conditions. Clean speech from the CENSUS database was contaminated with artificially-produced additive Gaussian noise at different global SNRs. Figure 2 compares recognition accuracy obtained using the census database for the STAR, SNR-based RATZ, and N-CDCN algorithms, as a function of global SNR. For comparison purposes, we also provide the recognition accuracy obtained when the system was completely retrained at each SNR using the speech with added noise, as well as the baseline accuracy obtained using cepstral mean normalization (CMN) alone. In contrast, the experimental results described in Table 1 are based on data recorded at an SNR of 23.0 dB, which is indicated by the vertical dashed line in Figure 2.

As can be seen in Figure 2, the STAR algorithm outperforms all other algorithms, and the difference becomes especially evident at low SNRs. Algorithms that attempt to correct the noisy cepstra using MMSE techniques introduce additional classification errors over the optimal classifier based on noisy cepstra. Hence, a technique like STAR, that attempts to “classify” the noisy speech based on noisy cepstral statistics can outperform techniques that attempt to clean noisy data using MMSE methods, and classify the data using clean statistics.

SUMMARY AND CONCLUSIONS

In this paper we described three new procedures, SNR-based RATZ, STAR, and N-CDCN, that improve speech recognition accuracy in unknown acoustical environments. SNR-based RATZ and STAR are based on the availability of stereo training data, while N-CDCN performs blind compensation.

We also presented a brief analytical study with simulations to support our modelling of how noise and filtering can affect clean speech statistics. We showed how the addition of a hierarchical structure to data driven methods like SNR-based RATS improves performance compared to previous methods developed at CMU, especially at low SNRs. Finally, we also show how methods that attempt to modify the statistics of the HMMs can perform even better, especially at low SNRs.

ACKNOWLEDGMENTS

The authors thank Uday Jain, Matthew Siegler and Evandro Gouvêa for useful suggestions. Pedro J. Moreno has been supported by a Fulbright fellowship awarded by the *Ministerio de Educación y Ciencia*, Spain. 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., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic, 1993.
2. Gales, M.J.F. & Young, S.J., “Cepstral Parameter compensation for HMM recognition in noise”. *Speech Communication*. 1993.
3. Juang, B.-H., “Speech Recognition in Adverse Environments”, *Computer Speech and Language*, **5**:275-294, 1991.
4. Liu, F-H., “Environmental Adaptation for Robust Speech Recognition”. Ph.D. Thesis, ECE Department, CMU, July 1994.
5. Moreno, P.J., Raj, B., Gouvêa, E. and Stern, R.M., “Multivariate Gaussian Based Cepstral Normalization”, *ICASSP-95*, 1995.