CEPSTRAL COMPENSATION BY POLYNOMIAL APPROXIMATION FOR ENVIRONMENT-INDEPENDENT SPEECH RECOGNITION

Bhiksha Raj, Evandro B. Gouvêa, Pedro J. Moreno, and Richard M. Stern

Department of Electrical and Computer Engineering & School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

Speech recognition systems perform poorly on speech degraded by even simple effects such as linear filtering and additive noise. One possible solution to this problem is to modify the probability density function (PDF) of clean speech to account for the effects of the degradation. However, even for the case of linear filtering and additive noise, it is extremely difficult to do this analytically. Previously attempted analytical solutions to the problem of noisy speech recognition have either used an overly-simplified mathematical description of the effects of noise on the statistics of speech, or they have relied on the availability of large environment-specific adaptation sets. Some of the previous methods required the use of adaptation data that consists of simultaneously-recorded or "stereo" recordings of clean and degraded speech. In this paper we introduce an approximation-based method to compute the effects of the environment on the parameters of the PDF of clean speech.

In this work, we perform compensation by Vector Polynomial approximationS (VPS) for the effects of linear filtering and additive noise on the clean speech. We also estimate the parameters of the environment, namely the noise and the channel, by using piecewise-linear approximations of these effects.

We evaluate the performance of this method (VPS) using the CMU SPHINX-II system and the 100-word alphanumeric CENSUS database. Performance is evaluated at several SNRs, with artificial white Gaussian noise added to the database. VPS provides improvements of up to 15 percent in relative recognition accuracy.

1. INTRODUCTION

As speech recognition systems become more accurate and sophisticated, robustness to noise, channel, and other environmental effects becomes increasingly important. In the past few years, researchers at CMU and other sites have developed a series of techniques to address this problem. Many of these environment compensation algorithms take advantage of the availability of "stereo data", *i.e.* speech databases that are simultaneously recorded in high-quality and degraded environments (*e.g.* [1][2]). Other algorithms make use of large amounts of adaptation data from the degraded environment (*e.g.* [5]). Still other algorithms (*e.g.* [6]) use knowledge of noise statistics and extensive computation to adapt the HMMs of clean speech to a new environment. Unfortunately, artifices such as stereo data, *a priori* knowledge about the testing environment, and/or the computational resource requirements of such algorithms are frequently unavailable.

From a practical point of view, algorithms that can compensate for the effects of the environment with almost no previous knowledge, and that only require a small segment of the speech signal to perform the compensation, are far more attractive than those that require environment-specific training information of any sort. Such compensation algorithms tend to be based on an analytic characterization of the nature of the degradation, rather than a mere empirical characterization of a large number of examples.

This analytic characterization can be thought of as an *environment function* that transforms the log spectrum of the clean speech into the log spectrum of the noisy speech. The parameters of this function would depend on the nature of the degradation. For example, in the case of linear filtering and additive noise, these parameters would be the log spectra of the noise and the impulse response of the channel. These parameters are usually not known, and need to be estimated during the compensation process.

The CDCN algorithm [3] is an example of this class of model-based algorithms that has been applied with success to several databases. Nevertheless, the CDCN algorithm has some limitations in that it does not model the effects of the environment on the variance of speech distributions, and the noise is estimated with only limited accuracy at low SNRs.

In [7] we introduced a Vector Taylor Series (VTS) method, which approximated the environment function by a Taylor series truncated after two terms, resulting in a straight line approximation. Although this modeled the effects of the environment on the variance of speech distributions, the coefficients of the straight line (the slope and the intercept) were not optimized according to any criterion.

We believe that it is more important to approximate the means and variances of the Gaussians describing the noisy speech than to approximate the environment function itself. In this paper, we do so using a two-fold approximation. To estimate the means and variances of the Gaussians describing the noisy speech, we use a generic piecewise polynomial approximation for the environment function, and we approximate the Gaussian probability density function by a polynomial in the range of interest. To estimate the noise and channel parameters, we use a straight-line approximation for the environment function in an analogous manner to VTS. However, instead of defining the line using a Taylor series, we constrain its coefficients to provide the same means and variances

as are obtained using the piecewise polynomial approximation. We refer to this approach as compensation by Vector Polynomial approximationS (VPS).

2. A MODEL OF THE ENVIRONMENT

As in previous papers we assume a model of the environment in which speech is corrupted by unknown additive stationary noise and unknown linear filtering:

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

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 $N(\omega)$ is the power spectrum of the additive noise.

In the log-spectral domain this can be expressed as:

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

or in more general terms:

$$z = x + h + f(n-x-h) \quad (3)$$

where h equals $\log(|H(\omega)|^2)$. We refer to $f(n-x-h)$ in equation (3) as the *environment function*.

We also assume that the PDF of the log spectra of the speech signal can be well represented by a summation of multivariate Gaussian distributions:

$$p(x) = \sum_{k=0}^{M-1} p[k] N_x(\mu_{x,k}, \Sigma_{x,k}) \quad (4)$$

Furthermore, we assume that the statistics of the noise can be well represented by a single Gaussian $N_n(\mu_n, \Sigma_n)$.

The problem of compensation is twofold. First, the parameters h, μ_n, and Σ_n need to be determined. Second, the distribution of z given the PDF of x and the parameters h, μ_n, and Σ_n must be computed. Because of the nonlinearity of the function $f(n-x-h)$, both problems are nontrivial. Only for very simple expressions of the function $f(n-x-h)$ can $p(z)$, the PDF of z, be computed analytically. It is not possible to compute $p(z)$ analytically for environmental functions such as the function $\log(1 + e^{n-x-h})$ that is valid for the model of Eq. (1). While $p(z)$ could be computed by Monte-Carlo methods, this approach is computationally expensive and requires previous knowledge of the parameters μ_n, Σ_n and h. Polynomial approximations provide a convenient framework that enables an analytical solution to both problems.

3. DESCRIPTION OF THE VPS ALGORITHM

3.1. Approximations For The Environment Function And The Gaussian Pdf

If we let $v = n - x - h$, we can see that the environment function is monotonically increasing, with asymptotes at $f(v) = 0$ for $v \to -\infty$ and $f(v) = v$ for $v \to \infty$. Therefore, its first derivative is a cumulative density function. The second derivative is observed to be a bell-shaped density function. We approximate the environment function by approximating the bell-shaped density function with a triangular density function and integrating it twice, to produce a piecewise cubic approximation. Figure 1 shows a comparative plot of the actual function and our approximation. As can be seen from this figure, the approximation cannot be distinguished from the actual function.

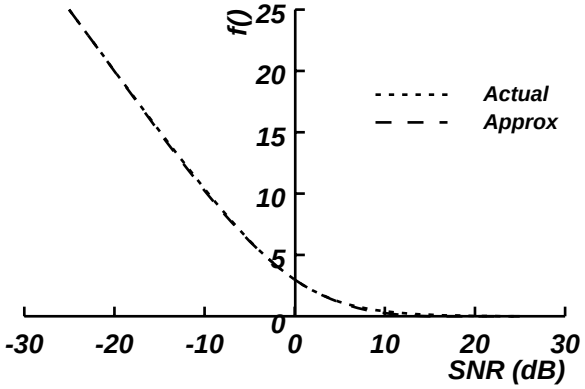

Figure 1: Comparative plot of the environment function and the approximation function.

We approximate the Gaussian distribution by a uniform distribution with the same mean and variance as the Gaussian. Possible extensions could use any bell shaped function, such as the convolution of a triangle function with a rectangular function.

3.2. Simulations

To confirm that the polynomial approximations are a good alternative to the Monte Carlo approach, Monte Carlo simulations were performed where Gaussian "signals" and Gaussian noise were produced at different signal-to-noise ratios (SNRs) and passed through a linear filter producing a set of noisy vectors. We compare statistics of these noisy data with results obtained using the VPS method, as well as results obtained from using a first order Taylor series approximation, as in [7].

Figure 2 shows how the resulting means of the noisy data set z can be approximated extremely well by VPS. In this figure we show the mean of the simulated noisy input signal, as well as the mean computed using the polynomial approximation and the first-order vector Taylor series expansion (*cf.* [7]). As we see, the Taylor series provides a reasonably good approximation, but the polynomial approximation outperforms it and cannot be distinguished from the actual mean itself.

Similarly, in Figure 3 we present the VPS and first-order Taylor series approximations to the variance. The polynomial approximation is somewhat closer to the real variance than the first-order Taylor series approximation.

3.3. Estimation Of Environment Parameters

The statistics of clean speech can be modeled as a mixture of Gaussian distributions as specified in Eq. (4). The parameters

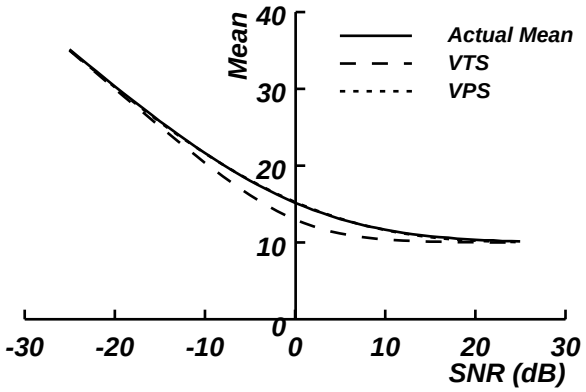

Figure 2: Effects of noise on the mean of the incoming signal. The exact values of the mean and estimates of the mean obtained from VPS and first-order VTS expansion are compared over a range of SNRs.

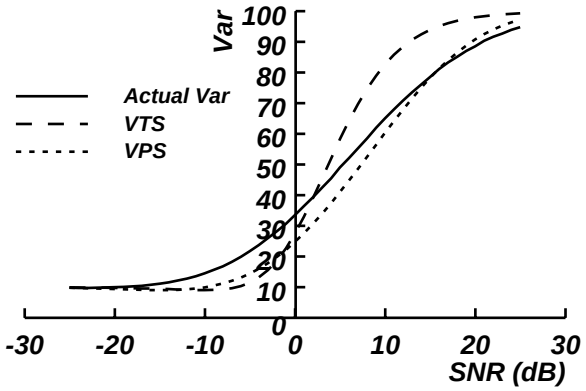

Figure 3: Effects of noise on the variance of the signal. The exact values of the variance and estimates of the variance obtained from VPS and first-order VTS expansion are compared over a range of SNRs.

describing these statistics are estimated using basic EM methods. In previous work [4], it was shown that it is reasonable for Gaussian densities for clean speech to be assumed to transform into Gaussian densities for noisy speech. In the present work we use this assumption to estimate the environment parameters, namely μ_n, Σ_n and h. This assumption implies a linear transformation between the log spectra of clean and noisy speech for each Gaussian density component of the PDF of clean speech. Therefore, to estimate the noise and channel parameters, we approximate the environment function (Eq. (3)) by the linear transformation

$$f(n-x-h) = A_k(n-x-h) + B_k \tag{5}$$

This leads to the estimation of log spectra of noisy speech as:

$$\hat{z} = (1-A_k)(x+h) + A_k n + B_k \tag{6}$$

from which we can easily derive the estimate of mean and variance of the k[th] Gaussian of the noisy speech as functions of A_k and B_k:

$$\mu_{z,k} = (1-A_k)(\mu_{x,k} + h) + A_k\mu_n + B_k \tag{7}$$

$$\Sigma^2_{z,k} = (1-A_k)^2\Sigma^2_{x,k} + A_k^2\Sigma_n^2 \tag{8}$$

Since we can obtain estimates for $\mu_{z,k}$ and $\Sigma_{z,k}$ from the polynomial approximation, we can use Eqs. (7) and (8) to compute A_k and B_k.

The algorithm for estimating μ_n, Σ_n and h proceeds as follows:

1. Obtain initial estimates of h, μ_n and Σ_n.
2. Compute values for $\mu_{z,k}$ and $\Sigma_{z,k}$ using the estimates of μ_n, Σ_n, and h for all the Gaussians.
3. Obtain the values of A_k and B_k using the estimates of $\mu_{z,k}$, $\Sigma_{z,k}$, h, μ_n and Σ_n.
4. Perform a single iteration of the EM algorithm to re-estimate the values of h, μ_n and Σ_n.
5. If the likelihood of the observed noisy data has not converged, return to Step 2.

The covariance matrices for all the Gaussian components of the clean speech and the noisy speech, and for the additive noise are assumed to be diagonal in order to reduce the computational complexity of the algorithm.

3.4. Compensation Of Noisy Speech

Once the parameters of the distribution of z are computed, an MMSE estimate is used to calculate the clean speech given the observed noisy speech

$$\hat{x}_{MMSE} = E(x|z) = \int x p(x|z)dx \tag{9}$$

$$\hat{x}_{MMSE} = \int (z-h-f(n-x-h))p(x|z)dx \tag{10}$$

Using the polynomial approximation estimates of $\mu_{z,k}$ and $\Sigma_{z,k}$, this can be approximated as:

$$\hat{x}_{MMSE} = z - \sum_{k=0}^{M-1} P[k|z](\mu_{z,k} - \mu_{x,k}) \tag{11}$$

Note that an alternative approach would correct means and variances of HMMs instead of performing the MMSE estimate of clean speech.

4. EXPERIMENTAL RESULTS

The effectiveness of the VPS algorithm was evaluated by artificially contaminating utterances from the CMU census database [3] with white noise at different SNRs. We used the SPHINX-II continuous speech recognition system.

In Figure 4 we compare the effectiveness of the VPS algorithm to the effectiveness of other model-based compensation algorithms, CDCN [3] and VTS [7] (which do not require stereo data), and RATZ [8], our best empirical algorithm that compensates input speech feature vectors using stereo data.

The VPS algorithm outperforms CDCN and RATZ at all SNRs, and it provides an improvement in relative recognition accuracy of up to 15 percent compared to VTS. We note that this improvement over VTS is obtained at zero additional computational cost. We believe

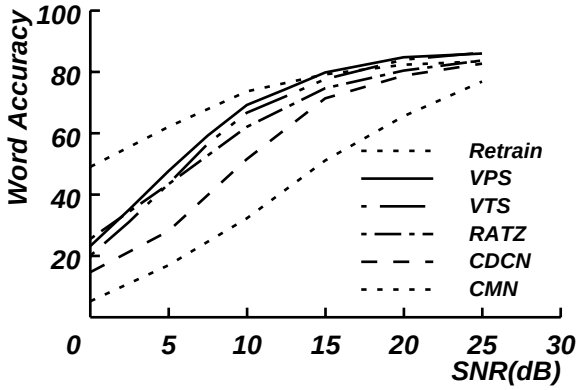

Figure 4: Comparison of recognition accuracy obtained for the CENSUS database using the VPS, VTS, CDCN, and RATZ algorithms as a function of SNR. The dotted curves indicate baseline performance using cepstral mean normalization only, as well as results obtained by completely retraining the system in the new environment.

that the gain would be significantly greater if the compensation were performed on the HMMs rather than on the incoming data, because of the more precise estimates of the parameters of the Gaussians in the HMMs.

5. DISCUSSION

The algorithm for estimating μ_n, Σ_n, and h is independent of the actual polynomial approximation method used to estimate $\mu_{z,k}$ and $\Sigma_{z,k}$. The algorithm could be used, without any modification, even if $\mu_{z,k}$ and $\Sigma_{z,k}$ were estimated using Monte Carlo methods or numerical integration. The algorithm can therefore be used to eliminate the requirement of samples of noise and separately-computed channel estimates for algorithms such as PMC [6]. The estimates for the mean and variance can be improved by using better approximations for the environment function and the Gaussian densities.

6. SUMMARY

In this paper we introduce an efficient approximation-based method to compensate for the effects of noise and linear filtering on the parameters of the PDF of clean speech. We also introduce a linear approximation based algorithm to estimate the parameters of the environment given estimates for the parameters of the PDF of noisy speech.

7. ACKNOWLEDGEMENTS

The authors thank Matthew Siegler and Uday Jain for useful discussions. We specially thank Matthew Siegler for helping us with the simulations. 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.

8. REFERENCES

1. F.-H. Liu (1994). *Environmental Adaptation for Robust Speech Recognition*. Ph. D. Dissertation, ECE Department, CMU, July 1994.

2. L. Neumeyer and M. Weintraub (1994). "Probabilistic Optimum Filtering for Robust Speech Recognition". Proc. ICASSP-94.

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

4. P. J. Moreno, B. Raj, E. B. Gouvêa, and R. M. Stern (1995). "Multivariate Gaussian Based Cepstral Normalization for Robust Speech Recognition". Proc. ICASSP-95.

5. C. J. Leggetter and P. C. Woodland (1995). "Flexible Speaker Adaptation using Maximum Likelihood Linear Regression", Proc. ARPA Spoken Language Systems Technology Workshop, January, 1995.

6. M. Gales and S. Young (1995). "A fast and flexible implementation of Parallel Model Combination". Proc. ICASSP-95.

7. P. J. Moreno, B. Raj, and R. M. Stern (1996). "A Vector Taylor Series Approach for Environment Independent Speech Recognition", Proc. ICASSP-96.

8. P. J. Moreno (1996). *Speech Recognition in Noisy Environments.* Ph. D. Dissertation, ECE Department, CMU, April 1996.