# VECTOR POLYNOMIAL APPROXIMATIONS FOR ROBUST SPEECH RECOGNITION

*Bhiksha Raj, Evandro B. Gouvêa, and Richard M. Stern*
Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

## 1. THE VECTOR POLYNOMIAL APPROXIMATIONS METHOD (VPS)

Over the past few years, researchers at CMU and other sites have developed a series of techniques to address the problem of robustness to noise, channel, and other environmental effects. While some of these methods need "stereo-data" (that are simultaneously recorded in high quality and degraded environments) (*e.g.* [1]), others use knowledge of the statistics of noise and extensive computation to adapt the HMMs of clean speech to a new environment [3], and yet others need large amouts of adaptation data [2].

In contrast, algorithms that can compensate for the effects of a new environment with almost no previous knowledge about it are very attractive. Such algorithms usually make use of an analytical characterization of the nature of the degradation. In the proposed paper we present and discuss one such algorithm, Vector Polynomial approximationS (VPS) [5]. In this algorithm, we analytically characterize the effect of the environment as an *environment function* $f(n, x, h)$ that transforms the log spectrum of the clean speech, $x$, into the log spectrum of the noisy speech, $z$, using the parameters $n$, representing additive noise, and $h$, for the channel:

$$z = x + f(n, x, h) \qquad (1)$$

We assume that the probability density function (PDF) of the log spectra of the clean speech signal can be well represented by a mixture of multivariate Gaussian distributions and that the distribution of the log spectra of noise is Gaussian.

VPS works on the principle that the parameters of the PDF of the enviromentally-corrupted speech signal can be well approximated by an appropriately derived polynomial function of the parameters of the channel, of the PDF of the log spectra of clean speech, and of the noise. These estimates can also be incorporated into an EM (Expectation-Maximization) formulation to estimate the values of the parameters of the environment function, *i.e.*, the channel $h$ and the noise $n$.

We compute in alternation parameters of the environment function (representing the effects of the noise and the channel) and of the distributions (the means and variances of the Gaussian mixture models representing the noisy signal).

After convergence has been achieved, we perform compensation of the noisy speech. This compensation is performed based on an MMSE estimate. The clean speech is estimated from the observed noisy speech, using the polynomial approximation estimates of the parameters of the PDF of the noisy speech, as:

$$\hat{x}_{MMSE} = z - \sum_{k=0}^{M-1} P[k|z](\mu_{z,k} - \mu_{x,k}) \qquad (2)$$

Note that an alternative approach would be to correct the means and variances of the HMMs instead of performing the MMSE estimate on the features representing the incoming clean speech.

## 2. SIMULATIONS AND EXPERIMENTAL RESULTS

Figure 1 displays the percentage error of our polynomial approximations for the means, which differ only marginally from the actual values, according to results of Monte Carlo simulations.

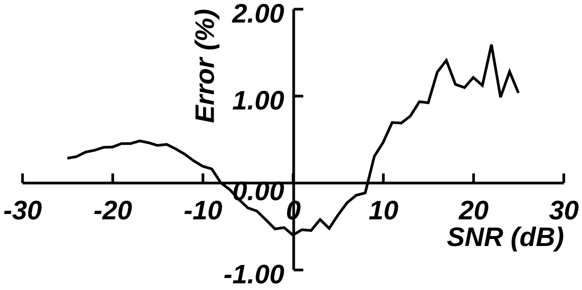

**Figure 1:** Percentage error between estimated means and actual means, as a function of the Signal-to-Noise ratio (SNR).

Figure 2 compares recognition accuracy obtained using VPS with that using other CMU compensation algorithms. All experiments were performed using the CMU Census database [1]. VPS performs better than all our previous algorithms, and it is also 20 percent faster than our second best algorithm (VTS)..

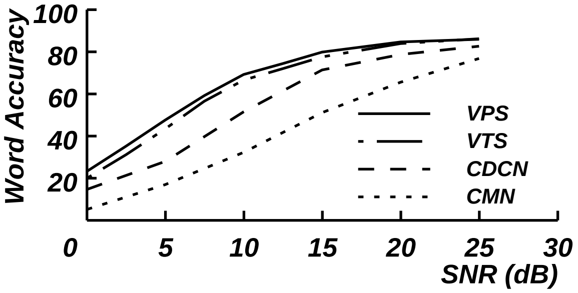

**Figure 2:** Comparison of recognition accuracy obtained for the CENSUS database using the VPS, VTS, and CDCN algorithms as a function of SNR. The dotted curve indicates baseline performance using cepstral mean normalization.

## 3. REFERENCES

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

2. 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.

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

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

5. B. Raj, E. B. Gouvêa, and R. M. Stern (1996), “Cepstral Compensation by Polynomial Approximation for Environment-Independent Speech Recognition”, Proc. ICSLP-96.