

# A VECTOR TAYLOR SERIES APPROACH FOR ENVIRONMENT-INDEPENDENT SPEECH RECOGNITION

Pedro J. Moreno, Bhiksha Raj 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 analytical approach to environment compensation for speech recognition. Previous attempts at solving analytically 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 work we introduce the use of a Vector Taylor series (VTS) expansion to characterize efficiently and accurately the effects on speech statistics of unknown additive noise and unknown linear filtering in a transmission channel. The VTS approach is computationally efficient. It can be applied either to the incoming speech feature vectors, or to the statistics representing these vectors. In the first case the speech is compensated and then recognized; in the second case HMM statistics are modified using the VTS formulation. Both approaches use only the actual speech segment being recognized to compute the parameters required for environmental compensation.

We evaluate the performance of two implementations of VTS algorithms using the CMU SPHINX-II system on the 100-word alphanumeric CENSUS database and on the 1993 5000-word ARPA Wall Street Journal database. Artificial white Gaussian noise is added to both databases. The VTS approaches provide significant improvements in recognition accuracy compared to previous algorithms.

## 1. INTRODUCTION

As speech recognition systems become more accurate and more 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 non-simultaneously-recorded 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, stereo data, *a priori* knowledge about the testing envi-

ronment, 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.

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:

- It does not model the effects of the environment on the variance of speech distributions
- The noise is approximated with only limited accuracy at low SNRs

The VTS algorithms described in this paper address these problems. Specifically, they:

- Require only the segment of noisy speech signal to be recognized to perform compensation.
- Model the effect of the environment on all the statistics of the probability density function (PDF) of speech.
- Provide a unified treatment of the noise and channel reestimation problem.
- Use a better, Gaussian, model for the PDF of the log-spectra of the noise.

## 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 linearly filtered by an unknown channel:

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

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 relation can be expressed as:

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

or in more general terms<sup>1</sup>:

$$z = x + f(x, n, q)$$

where  $q$  is an unknown parameter that represents the effects of linear filtering in the log-spectral domain.

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:

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

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  $q$ ,  $\mu_n$ , and  $\Sigma_n$  need to be determined. Second, the distribution of  $z$  given the PDF of  $x$  and the parameters  $q$ ,  $\mu_n$ , and  $\Sigma_n$  has to be computed. Because of the non-linearity of the function  $f(n, x, q)$ , both problems are non-trivial. Only for very simple expressions of the function  $f(n, x, q)$  can  $p(z)$  be computed analytically. For other functions such as  $\log(1 + e^{n-x-q})$  it is not possible to compute  $p(z)$  analytically. While  $p(z)$  could be computed by Monte-Carlo methods, this approach is computationally expensive and requires previous knowledge of the parameters  $\mu_n$ ,  $\Sigma_n$  and  $q$ . VTS provides a framework that enables an analytical solution to both problems.

### 3. DESCRIPTION OF THE VTS ALGORITHMS

The key of the new VTS algorithms is to approximate the generic vector function  $f(n, x, q)$  with a vector Taylor series approximation:

$$f(x, n, q) \cong f(x_0, n_0, q_0) + \frac{d}{dx}f(x_0, n_0, q_0) \{x - x_0\} + \frac{d}{dn}f(x_0, n_0, q_0) \{n - n_0\} + \frac{d}{dq}f(x_0, n_0, q_0) \{q - q_0\} + \dots$$

where  $f(x_0, n_0, q_0)$  is the vector function evaluated at a particular vector point. Similarly,  $\frac{d}{dx}f(x_0, n_0, q_0)$  represents the matrix derivative of the vector function at a particular vector point. The higher order terms of the Taylor series involve higher order derivatives resulting in tensors.

The Taylor expansion is exact everywhere when the order of the Taylor series is infinite. However, when  $x$  has a Gaussian distribution, the function can be expanded around the mean of  $x$  and the expansion needs to be good only within a relatively

1. In fact, the function  $f(x, n, q)$  could be any function.

narrow region around the mean. We take advantage of this fact to truncate the Taylor series after just a few terms.

VTS-0 uses only the zeroth-order terms of the Taylor series and VTS-1 uses the zeroth-order and first-order terms. Higher orders of VTS are also possible when greater approximation accuracy is required.

### 3.1. Modeling speech statistics using VTS

To confirm that the Taylor series approximations are a good alternative to the Monte-Carlo approach, simulations were performed using artificial data. A one-dimensional set of vectors was produced using Monte-Carlo simulation, and these clean signal vectors were contaminated with noise at different signal-to-noise ratios (SNRs) and passed through a linear channel producing a set of noisy vectors.

Figure 1 shows how the resulting means of the noisy vector set  $x$  can be approximated quite well by the Taylor series. In this figure we show the mean of the simulated noisy input signal, as well as the mean computed using the Taylor series expansion of orders 0 and 2. As we see, the zeroth-order provides a reasonably good approximation. However at lower SNRs the second-order Taylor series expansion provides an even better approximation of the actual distribution.

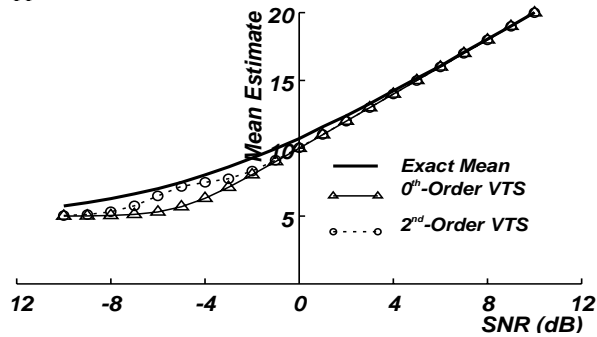


Figure 1. Effects of noise on the mean of the incoming signal. The exact values of the mean and estimates of the mean obtained from the zeroth-order and second-order VTS expansion are compared over a range of SNRs.

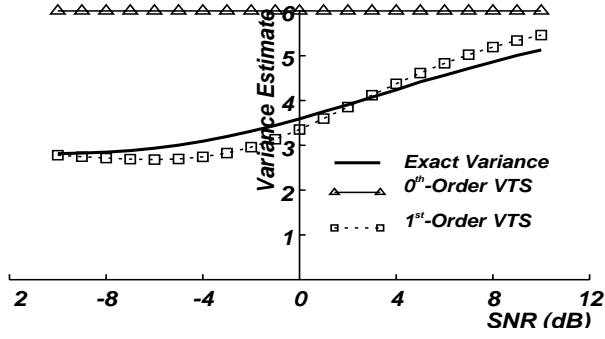
Similarly, in Figure 2 we present the zeroth-order and first-order Taylor series approximations to the variance. The first-order approximation is closer to the real variance than the zeroth-order approximation. Odd orders of the Taylor series do not contribute to the approximation of the mean.

### 3.2. Statistics of clean and noisy speech

The statistics of clean speech can be modeled as a mixture of Gaussian distributions. The parameters describing these statistics are estimated using basic EM methods.

The goal of the VTS algorithm is to estimate the pdf of noisy speech given the pdf of clean speech, a segment of noisy speech and the Taylor series expansion that relates noisy speech to clean speech. Once the pdf of the noisy speech is computed, minimum mean square estimation (MMSE) can be used to predict the unobserved clean speech sequence.

Alternately, if HMMs are used to describe the pdf of clean speech we can use the Taylor series approach to compute the noisy HMMs and perform recognition on the noisy signal



**Figure 2.** Effects of noise on the variance of the signal. The exact values of the variance and estimates of the variance obtained from the zeroth-order and first-order VTS expansion are compared over a range of SNRs.

itself. In this paper we only report results obtained using the first approach.

**Zeroth-order Vector Taylor Series expansion (VTS-0):** The zero order Taylor series expansion of  $f(x, n, q)$  results in a Gaussian distribution for the noisy speech  $z$  when  $x$  is Gaussian

$$p(z) = N_z(\mu_z, \Sigma_z)$$

The mean vector and covariance matrices that represent the noisy speech statistics are computed as

$$\mu_z = E(z) = E(x + f(n_0, x_0, q_0)) = \mu_x + f(n_0, x_0, q_0)$$

$$\Sigma_z = \Sigma_x$$

**First-order Vector Taylor Series expansion (VTS-1):** In the case of the first-order Taylor series expansion of  $f(x, n, q)$  the resulting distribution of  $z$  is also Gaussian when  $x$  is Gaussian. The new mean vector is computed as

$$\begin{aligned} \mu_z = & E(x + f(n_0, x_0, q_0)) + \\ & E\left(\frac{d}{dx}f(x_0, n_0, q_0) \{x - x_0\}\right) + \\ & E\left(\frac{d}{dn}f(x_0, n_0, q_0) \{n - n_0\}\right) + \\ & E\left(\frac{d}{dq}f(x_0, n_0, q_0) \{q - q_0\}\right) \end{aligned}$$

In a similar fashion, the new covariance matrix can be expressed as

$$\begin{aligned} \Sigma_z = & \left(I + \frac{d}{dx}f(n_0, x_0, q_0)\right)^T \Sigma_x \left(I + \frac{d}{dx}f(n_0, x_0, q_0)\right) + \\ & \left(\frac{d}{dx}f(n_0, x_0, q_0)\right)^T \Sigma_n \frac{d}{dx}f(n_0, x_0, q_0) \end{aligned}$$

where  $\Sigma_n$  is the variance of the noise.

For both VTS-0 and VTS-1 the parameters  $q$  and  $\mu_n$ , and hence the parameters  $\mu_z$  and  $\Sigma_z$ , are estimated iteratively using a modified version of the EM algorithm. VTS-1 also esti-

mates the variance of noise,  $\Sigma_n$ . The algorithms proceed as follows:

1. Obtain initial estimates of  $q$ ,  $\mu_n$  and  $\Sigma_n$ .
2. Expand the function  $f(x, n, q)$  around the mean vector of each Gaussian in the distribution of  $x$ ,  $\mu_{x,k}$  and the estimates of  $\mu_n$  and  $q$ .
3. Estimate the parameters of the distribution of  $z$ ,  $\mu_{z,k}$  and  $\Sigma_{z,k}$ .
4. Perform a single iteration of the EM algorithm to re-estimate the values of  $q$  and  $\mu_n$ . In the case of VTS-1  $\Sigma_n$  is also re-estimated.
5. If the likelihood of the observed noisy data has not converged, return to Step 2.

Because the distribution of  $x$  is assumed to be a Gaussian mixture, the resulting distribution computed for  $z$  is also a Gaussian mixture distribution with a one-to-one correspondence between each Gaussian in the distribution of  $x$  to a Gaussian in the distribution of  $z$ .

In all cases, the covariance matrices of the clean speech, the noisy speech, and the additive noise are assumed to be diagonal in order to reduce the computational complexity of the algorithm. Non-diagonal matrices would result in a computationally expensive tensor formulation.

### 3.3. 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

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

$$\hat{x}_{MMSE} = \int (z - f(x, n, q)) p(x|z) dx$$

The results obtained depend on which order Taylor series approximation is used. The zeroth-order approximation produces

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

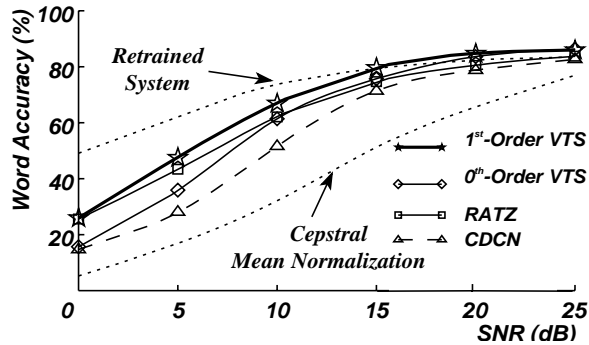
A similar value is obtained for the first order approximation.

## 4. EXPERIMENTAL RESULTS

The effectiveness of the VTS algorithms was evaluated by artificially contaminating utterances from the CMU census database [3] and from the ARPA Wall Street Journal task with white noise at different SNRs. The SPHINX-II continuous speech recognition system was used.

In Figure 3 we compare the effectiveness of the zeroth-order VTS algorithm and the first-order VTS algorithm to the effectiveness of another model-based compensation algorithms

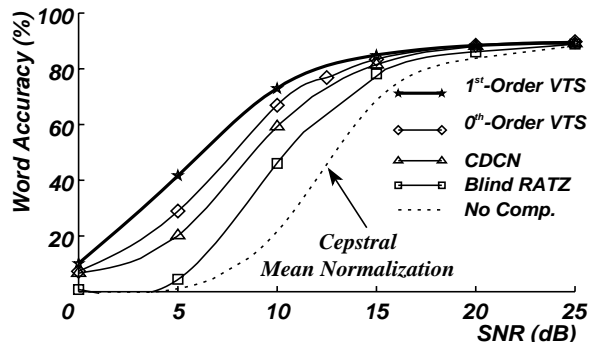
CDCN [3] (which does not require stereo data), and the empirical algorithm, RATZ [4] (which does require stereo data).



**Figure 3.** Comparison of recognition accuracy obtained for the CENSUS database using the zeroth-order and first-order 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.

The VTS-0 algorithm performs better than CDCN at all SNRs, and the VTS-1 algorithm is observed to perform even better than the VTS-0 algorithm. In fact, at all SNRs, VTS-0 outperforms RATZ, which is an algorithm that assumes the availability of stereo data.

In Figure 4 we present results from a similar experiment using the 5,000-word evaluation set of the 1993 ARPA Wall Street Journal test set. As before, the data were contaminated by artificial white noise at different SNRs. Again, the zeroth-order VTS algorithm outperforms the CDCN algorithm at all SNRs.



**Figure 4.** Comparison of recognition accuracy obtained for the 1993 ARPA 5000-word WSJ0 database using the zeroth-order and first-order VTS, CDCN, and RATZ algorithms as a function of SNR. The dotted curves are as in Fig. 3.

## 5. DISCUSSION

A truncated Taylor series is a special case of a polynomial approximation to a function. It is well known that for polynomial approximation of any order of a function, better polynomials exist than the Taylor series. Hence, we speculate that using more generic polynomial approximations that are opti-

mized to minimize the error for the parameters of the distribution of  $z$  may give us even better performance. Simulations indicate that the more general polynomials provide much better estimates of the mean and variance of  $z$  than the Taylor series.

In fact, the VTS algorithms may be viewed as a special case of algorithms based on more generic polynomial expansions. Similarly, CDCN may be viewed as a special case of the VTS approach worked for a zeroth order polynomial in the cepstral domain.

## 6. SUMMARY

In this paper we introduce an efficient approximation that analytically handles the problem of compensating for the effects of noisy and filtered speech with a bare minimum of testing data and no “stereo” training data. The algorithms presented provide significant improvement over previous work. We provide an easily expandable framework for further improving the performance of these algorithms at greater computational expense by increasing in the order of the Taylor series approximation.

## ACKNOWLEDGEMENTS

The authors thank Evandro Gouvea, Matthew Siegler and Uday Jain for useful discussions, and especially Matthew Siegler for helping us with the simulations. 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. 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, 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.

