A New Algorithm for Robust Speech Recognition: The Delta Vector Taylor Series Approach

Pedro J. Moreno and Brian Eberman
email: pjm@crl.dec.com, bse@crl.dec.com

Digital Equipment Corporation
Cambridge Research Laboratory

ABSTRACT

In this paper we present a new model-based compensation technique called Delta Vector Taylor Series (DVTS). This new technique is an extension and improvement over the Vector Taylor Series (VTS) approach [7] that addresses several of its limitations. In particular, we present a new statistical representation for the distribution of clean speech feature vectors based on a weighted vector codebook. This change to the underlying probability density function (PDF) allows us to produce more accurate and stable solutions for our algorithm. The algorithm is also presented in a EM-MAP framework where some the environmental parameters are treated as random variables with known PDF's. Finally, we explore a new compensation approach based on the use of convex hulls.

We evaluate our algorithm in a phonetic classification task on the TIMIT [5] database and also in a small vocabulary size speech recognition database. In both databases artificial and natural noise is injected at several signal to noise ratios (SNR). The algorithm achieves matched performance at all SNR's above 10 dB.

1. Introduction

Over the last years several techniques have been proposed to deal with the problem of speech recognition in noisy environments. Some of them such as PMC [3], or MLLR [6] have used the recognition engine and its rich statistical representation (more than 90,000 Gaussians in systems like SPHINX-3 and HTK [9]) to model and compensate for the effects of the environment on speech recognition systems. Other techniques like CDCN [1] and POF [8] among others have used a reduced set of Gaussian mixtures (typically 256 or less) to model the clean speech feature vectors and preprocess the noisy speech features vectors to effectively "clean" the features before being processed by the recognition engine.

The use of a rich statistical representation improves performance, but has the drawback of using the whole speech recognition engine with its associated complexity. An ideal robust recognition technique should have the advantages of a rich statistical representation and at the same time being simple and fast in its operation.

The Delta Vector Taylor Series (DVTS) approach is an attempt in this direction. It tries to gain the benefits of a rich statistical representation and a low complexity technique for robust speech recognition. It tries to achieve these goals by using a different statistical representation for the speech feature vectors.

The outline of the paper is as follows. In section 2 we describe the DVTS algorithm. In section 3 we briefly describe the necessary modifications to the algorithm to make it work as a filter. In section 4 we describe our experimental results and finally in section 5 we present our conclusions.

2. New Algorithm: Delta-VTS

DVTS models the speech feature vectors as a weighted sum of multidimensional Dirac deltas

$$p(\mathbf{x}) = \sum_{k=0}^{M-1} P[k]\,\delta(\mathbf{x} - \mathbf{x}_k) \tag{1}$$

where each vector function $\delta(\mathbf{x} - \mathbf{x}_k)$ is modeled as

$$\delta(\mathbf{x} - \mathbf{x}_k) = \prod_{i=0}^{D-1} \delta(x_i - x_{i,k}) \tag{2}$$

$P[k]$ is an *a priori* probability of observing a particular delta. The sum of these probabilities must add up to one.

This novel representation of the PDF of $\mathbf{x}$ has several advantages. First of all it greatly simplifies the mathematical assumptions of the VTS [7] algorithm. It produces a simple, fast, robust and direct formulation of the EM solutions already presented in [7].

In this paper we assume a model of the environment in which speech is corrupted by unknown additive stationary noise and unknown linear filtering

$$Z(\omega) = X(\omega)|H(\omega)|^2 + N(\omega) \tag{3}$$

where $Z(\omega)$ represents the power spectrum of the degraded speech, $X(\omega)$ is the power spectrum of the clean speech, $|H(\omega)|^2$ is the transfer function of the linear filter, and $N(\omega)$ is the power spectrum of the additive noise.

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

$$\mathbf{z} = \mathbf{x} + \log(\exp(\mathbf{q}) + \exp(\mathbf{n} - \mathbf{x})) \tag{4}$$

or in more general terms

$$\mathbf{z} = \mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{n}, \mathbf{q}) \tag{5}$$

where $\mathbf{q}$ is the log-mel-spectra of $|H(\omega)|^2$. We refer to $\mathbf{f}(\mathbf{x}, \mathbf{n}, \mathbf{q})$ in equation 5 as the *environmental function*.

In the regular VTS approach $p(\mathbf{x})$ is represented by a mixture of Gaussians and as a result $p(\mathbf{z})$ cannot be computed analytically. In fact, even though the resulting PDF is not Gaussian we still model it as Gaussian for lack of a better solution. In DVTS the solution for $p(\mathbf{z})$ can be computed directly with **no** approximations as

$$p(\mathbf{z}) = \sum_{k=0}^{M-1} P[k]\, \delta(\mathbf{z} - \mathbf{z}_k) \tag{6}$$

where $\mathbf{z}_k$ is equal to $\mathbf{x}_k + \mathbf{f}(\mathbf{x}_k, \mathbf{n}, \mathbf{q})$. Notice that this is due the the highly localize nature of $p(\mathbf{x})$ around the $\mathbf{x}_k$ sample vector points.

If the noise is assumed to have a Gaussian distribution $p(\mathbf{n}) = \mathcal{N}_{\mathbf{n}}(\mu_{\mathbf{n}}, \Sigma_{\mathbf{z}})$ we still obtain a direct solution with **no** approximations for $p(\mathbf{z})$ as

$$p(\mathbf{z}) = \sum_{k=0}^{M-1} P[k]\, \mathcal{N}_{\mathbf{z}}(\mathbf{z}_k, \Sigma_{\mathbf{z}}) \tag{7}$$

The result for $p(\mathbf{z})$ can be interpreted as a PDF modelled via Parzen windows with Gaussian kernels all sharing the same covariance. Under these assumptions equation 7 is trivially obtained from the definition of $p(\mathbf{x})$ and the vector relation $\mathbf{z} = \mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{n}, \mathbf{q})$.

The goal of the algorithms is to find the environmental parameters $\theta = \{\mathbf{q}, \mathbf{n}, \Sigma_{\mathbf{z}}\}$ [1] that given an ensemble of noisy speech features vectors $\mathcal{Z} = \{\mathbf{z}0, \ldots, \mathbf{z}_{T-1}\}$ maximize the log likelihood that this ensemble has been produced by the PDF of the noisy speech feature vectors $p(\mathbf{z})$. If additional information is available about the *a priori* PDF's of the environmental parameters θ, the maximization problem can be reformulated using a MAP framework. In this paper we will assume previous knowledge of the environmental parameters θ as modelled by a simple Gaussian distribution with diagonal covariance Σ_θ.

Notice that the use of diagonal covariances is done for simplicity and the formulation can be trivially extended to a full covariance matrix. Also notice that the treatment of the environment is not tied to our particular choice of environmental function. Any environmental function can be used provided it is smooth and its derivatives exist up to a certain order. As in any EM formulation we start by defining the $\mathcal{Q}$ function as

$$\mathcal{Q}(\theta, \theta') = \sum_{t=0}^{T-1} \sum_{k=0}^{M-1} P[k|\mathbf{z}_t, \theta]\, \log(p(\mathbf{z}_t, k|\theta') + p(\theta')) \tag{8}$$

by taking derivative with respect to θ' we obtain the system of equations $\eta = (\Lambda + \Sigma_\theta)\theta'$ where η is equal to

$$\begin{pmatrix} \sum_{t=0,k=0}^{T-1,M-1} P[k|\mathbf{z}_t]\, (\mathbf{I} - \mathbf{F}_k)^t \Sigma_{\mathbf{z}}^{-1} (\mathbf{z}_t - \mathbf{z}_k) \\ \sum_{t=0,k=0}^{T-1,M-1} P[k|\mathbf{z}_t]\, (\mathbf{F}_k)^t \Sigma_{\mathbf{z}}^{-1} (\mathbf{z}_t - \mathbf{z}_k) \end{pmatrix} \tag{9}$$

and Λ is equal to $\sum_{t=0,k=0}^{T-1,M-1} P[k|\mathbf{z}_t]\, \mathbf{M}_{\mathbf{k}}$ which in turn is equal to

$$\begin{pmatrix} (\mathbf{I} - \mathbf{F}_k)^t \Sigma_{\mathbf{z}}^{-1} (\mathbf{I} - \mathbf{F}_k) & (\mathbf{I} - \mathbf{F}_k)^t \Sigma_{\mathbf{z}}^{-1} \mathbf{F}_k \\ \mathbf{F}_k^t \Sigma_{\mathbf{z}}^{-1} (\mathbf{I} - \mathbf{F}_k) & \mathbf{F}_k^t \Sigma_{\mathbf{z}}^{-1} \mathbf{F}_k \end{pmatrix} \tag{10}$$

where $\mathbf{F}_k$ is equal to the derivative of $\mathbf{f}(\mathbf{x}_k, \mathbf{n}', \mathbf{q}')$ with respect to $\mathbf{q}'$ or $\mathbf{n}'$ (they are the same with different signs), *i.e.*, $\nabla_{\mathbf{q}'} \mathbf{f}(\mathbf{x}_k, \mathbf{n}', \mathbf{q}')$.

Notice that in taking derivatives we have approximated $\mathbf{z}_k = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k, \mathbf{n}', \mathbf{q}')$ by its first order vector Taylor series approximation $\mathbf{z}_k \sim \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k, \mathbf{n}', \mathbf{q}') + \nabla_{\mathbf{n}'} \mathbf{f}(\mathbf{x}_k, \mathbf{n}', \mathbf{q}')(\mathbf{n}' - \mathbf{n}) + \nabla_{\mathbf{q}'} \mathbf{f}(\mathbf{x}_k, \mathbf{n}', \mathbf{q}')(\mathbf{q}' - \mathbf{q})$. Otherwise when taking the derivative of the $\mathcal{Q}$ function we would not obtain a solution. In effect, we have linearized a nonlinear relationship ($\mathbf{z}_k = \mathbf{x}_k + \mathbf{f}(\mathbf{x}_k, \mathbf{n}', \mathbf{q}')$) around our current estimates of $\mathbf{q}$ and $\mathbf{n}$.

To solve for θ' we need to solve the above system of equations. Notice that the $\Lambda + \Sigma_\theta$ matrix can be robustly inverted due to its symetricities. The addition of Σ_θ to Λ has the practical property of conditioning the Λ matrix making its inversion a much more stable problem.

The EM algorithm [2] provides an iterative solution to the maximization problem. It also guarantees that the likelihood function does not decrease at each iteration. For further details of the EM algorithm applied to a similar derivation (the VTS approach) see [7].

To summarize, to maximize the $\mathcal{Q}(\theta, \theta')$ function with respect to the environmental parameters θ' our algorithm works in three stages:

1. Training or learning of the distribution of $p(\mathbf{x})$. This step is performed using the well known EM [2] algorithm and is done with sufficient amounts of clean training data.
2. Estimation of environmental parameters θ using EM. Given a stream of noisy speech feature vectors $\mathcal{Z} = \{\mathbf{z}0, \ldots, \mathbf{z}_{T-1}\}$ and the PDF of the clean feature vectors $p(\mathbf{x})$, find the unknown environmental parameters θ that maximize the likelihood of observing the noisy data.
3. Compensation of the noisy feature vectors. Given the noisy speech feature vectors $\mathcal{Z}$ and their estimated PDF $p(\mathbf{z})$ estimate the unobserved clean speech feature vectors $\mathcal{X} = \{\mathbf{x}0, \ldots, \mathbf{x}_{T-1}\}$.

The compensation is done in two different ways. In the first one we compute the conditional expectation as a convex hull over the original K clean speech feature vectors

[1] In general we will refer to the environmental parameters with the θ symbol

$\mathbf{x}_k$ that describe $p(\mathbf{x})$.

$$\begin{aligned}\mathcal{E}[\mathbf{x}_t|\mathbf{z}_t] &= \int_{\mathcal{X}} \frac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t)\,\mathbf{x}_t}{p(\mathbf{z}_t)} d\mathbf{x}_t \\ &= \int_{\mathcal{X}} \frac{p(\mathbf{x}_t)\,\mathcal{N}_{\mathbf{z}_t}(\mathbf{x}_t + \mathbf{f}(\mathbf{x}_t, \mathbf{n}, \mathbf{q}), \Sigma_{\mathbf{z}})\,\mathbf{x}_t}{p(\mathbf{z}_t)} d\mathbf{x}_t \\ &= \sum_{k=0}^{M-1} P[k|\mathbf{z}_t]\,\mathbf{x}_k \end{aligned} \quad (11)$$

The original VTS compensation is done using a zeroth order vector Taylor approximation *i.e.*, replacing $\mathbf{x}_t$ by $\mathbf{z}_t - \mathbf{f}(\mathbf{x}_k, \mathbf{n}, \mathbf{q})$ and moving $\mathbf{z}_t$ out of the integral

$$\mathcal{E}[\mathbf{x}_t|\mathbf{z}_t] \sim \mathbf{z}_t - \sum_{k=0}^{M-1} P[k|\mathbf{z}_t]\,\mathbf{f}(\mathbf{x}_k, \mathbf{n}, \mathbf{q}) \quad (12)$$

We report results using both compensation algorithms.

3. Online Compensation

The algorithm so far has been described as a batch compensation process. In practice this imposes several restrictions for real-time operation. It is therefore important to study the possibility of an *online* version of the algorithm.

As in all EM algorithms accumulators can be defined where the contributions of each speech feature vector can be stored. In most EM problems the conversion from a batch based algorithm to an online algorithm is as simple as multiplicating by a forgetting factor each of the mentioned contributions. Another popular alternative is to process windows of data and pass from window to window the previous accumulators adding them to the current ones weighted by a appropriate factor.

For every window k we compute a Λ_k as

$$\Lambda_k = r_1\Lambda_k + r_2\Lambda_{k-1} \quad (13)$$

where $r_1 + r_2$ must add up to one. Once the Λ_k is computed we can solve the $\eta_k = (\Lambda_k + \Sigma_\theta)\theta'_k$ equation and search for the environmental parameters θ'_k on this particular window of speech applying a MMSE estimator.

4. Experimental Results

To study the validity of the proposed algorithm on noisy speech we performed a series of experiments on the TIMIT [5] phonetic classification task injecting artificially produced noise at several signal-to-noise ratios. Phonetic modeling, classification, and word recognition was performed using the Stochastic Trajectory Modelling approach (STM) [4].

STM models were trained using 3695 utterances from 462 different speakers. The testing set contained 250 utterances from 50 different speakers not seen in the training set. The speech was first parametriced into log-mel power spectra and compensation was performed in this domain. Once the noisy data was compensated it was transformed via a Discrete Cosine Transform into the cepstrum space where STM classification and recognition was performed.

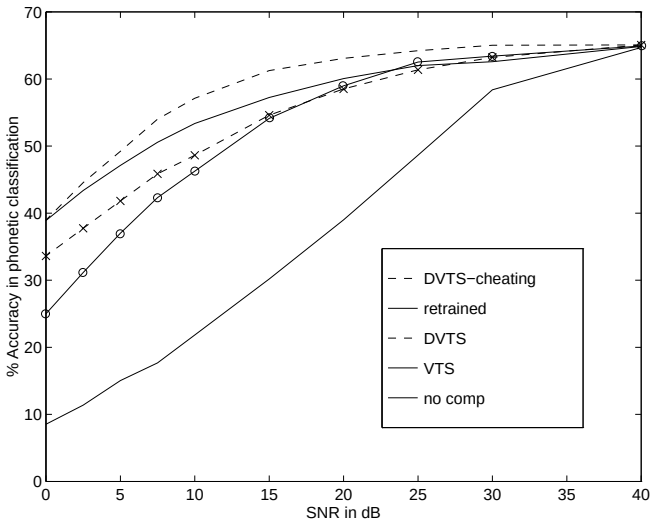

Figure 1: Performance of the DVTS algorithm at several SNR's on the TIMIT phonetic classification task.

Similar experiments were also done injecting naturally occurring noise collected in an office environment.

Figure 1 shows phonetic classification performance *vs.* signal to noise ratio (SNR) at different dB levels. The upper continuous line curve represents the performance of the system when the STM models are trained and tested in matched conditions, *i.e.*, at the same SNR. The lower continuous line represents the performance of the STM phonetic classifier when the system is trained with clean speech and tested on noisy speech. These two lines represent upper and lower bounds respectively for the performance of any robustness algorithm.

The upper dotted line represents an ideal experiment in which the clean utterance feature vectors were used as $p(\mathbf{x})$. This case would occur with simultaneous recordings of clean and noisy speech.I represents an ideal situation in which the statistics of the clean data are exactly the clean speech feature vectors. The dotted line with stars represents the performance of the DVTS algorithm using 1000 Dirac deltas trained from 2000 training utterances. Finally, for comparison, the continuous line with bullets represents the performance of a VTS algorithm with 256 Gaussians.

As we can see the DVTS algorithm outperforms the VTS at SNR's below 15 dB and both of them achieve matched performance at SNR's above 20 dB.

Figure 2 shows the same experiments but comparing the effect of a convex hull compensation module *vs* the normal pseudo MMSE estimator described in section 2. The dotted line represents the performance of DVTS using the pseudo MMSE estimator. The dotted line with bullets shows the performance of the convex hull compensation. Although this should have certain error repair properties, performance is not as good as the original pseudo conditional expectation formulation employed in VTS. The difference is larger at high SNR's. We believe this drop in

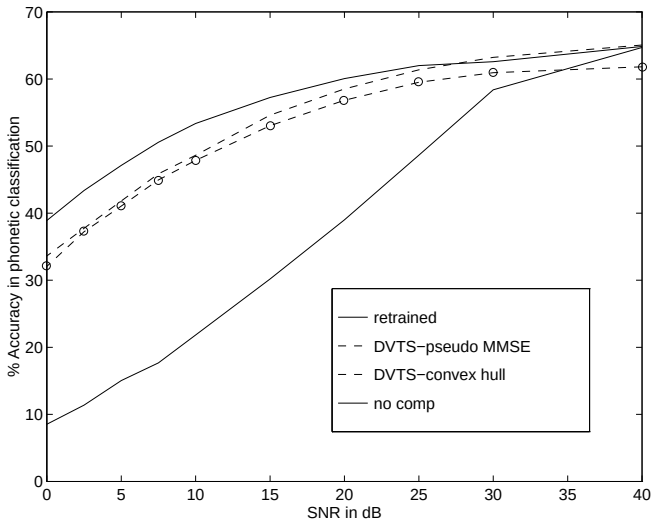

Figure 2: Comparison of two different compensation schemes. Pseudo MMSE *vs* convex hull MMSE.

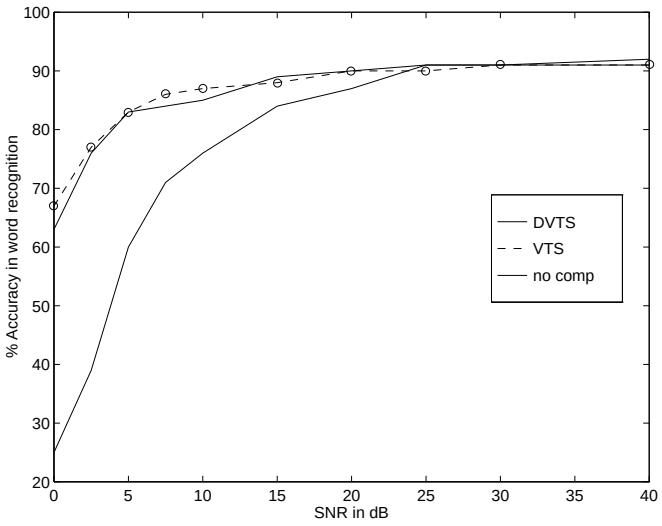

Figure 3: Performance of the DVTS algorithm at several SNR's on a local 107 words database.

performance is due to quantization effects. At high SNR's most of the distortion introduced by the environment is a simple shift and a MMSE estimator in this form does not suffer the quantization error effects introduced by a convex hull formulation. At lower SNR's the environment has a more complex effect on the signal and quantization effects are less severe. We believe that a larger number of deltas in $p(\mathbf{x})$ should alleviate this problem.

In all cases the algorithm produced significant improvements in phonetic classification performance. Notice that almost matched performance is obtained when the $p(\mathbf{x})$ statistics are built from the clean uncontaminated speech feature vectors. This experiment shows the dependency of the technique on a proper choice of deltas to build $p(\mathbf{x})$.

Figure 3 shows our results on a 107 word vocabulary recognition task recorded in our lab. DVTS performs very similarly to VTS an both achieve significant gains in recognition accuracy at all SNR's.

5. Conclusions

In this paper we have introduced our preliminary results with a new environmental compensation technique able to cope with the effects of unknown environments on speech feature vectors. The technique uses a novel statistical representation that enables us to simplify many of the approximations needed with Gaussian mixtures PDF's. The algorithm presented here provides significant improvement over previous work, specially at lower SNR's. We have also described how to turn this compensation algorithm into an online adaptable filter, able to adapt to the changing conditions of the environment.

6. REFERENCES

1. A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. PhD thesis, CMU, Department of Electrical and Computer Engineering, 1990.
2. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B. Vol. 39*, pages 1–38, 1977.
3. M. J. F. Gales. *Model-Based Techniques for Noise Robust Speech Recognition*. PhD thesis, University of Cambridge., 1995.
4. W. Goldenthal. *Statistical Trajectory Models for Phonetic Recognition*. PhD thesis, MIT, Department of Aeronautics and Astronautics, 1994.
5. L. Lamel, R. Kassel, and S. Senef. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *DARPA Speech Recognition Workshop SAIC-86/1546*, pages 100–109. Palo Alto, CA, February 1986.
6. C. J. Leggetter and P. C. Woodland. Speaker adaptation of hmms using linear regression. Technical Report CUED/F-INFENG/TR. 181, Cambridge University Engineering Dept., 1994.
7. P. J. Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, CMU, Department of Electrical and Computer Engineering, 1996.
8. L. Neumeyer and M. Weintraub. Probabilistic optimum filtering for robust speech recognition. In *Proceedings: ICASSP 94. 1994 International Conference on Acoustics, Speech, and Signal Processing*, volume I, pages 417–420, May 1994.
9. D. S. Pallet and J. G. Fiscus. 1996 preliminary broadcast news benchmark tests. In *ARPA Speech Recognition Workshop*, 1997.