

COMPENSATION FOR ENVIRONMENTAL DEGRADATION IN AUTOMATIC SPEECH RECOGNITION

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

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

ABSTRACT

The accuracy of speech recognition systems degrades when operated in adverse acoustical environments. This paper reviews various methods by which more detailed mathematical descriptions of the effects of environmental degradation can improve speech recognition accuracy using both “data-driven” and “model-based” compensation strategies. Data-driven methods learn environmental characteristics through direct comparisons of speech recorded in the noisy environment with the same speech recorded under optimal conditions. Model-based methods use a mathematical model of the environment and attempt to use samples of the degraded speech to estimate model parameters. These general approaches to environmental compensation are discussed in terms of recent research in environmental robustness at CMU, and in terms of similar efforts at other sites. These compensation algorithms are evaluated in a series of experiments measuring recognition accuracy for speech from the ARPA Wall Street Journal database that is corrupted by artificially-added noise at various signal-to-noise ratios (SNRs), and in more natural speech recognition tasks.

1. INTRODUCTION

The development of robust speech recognition systems that maintain a high level of recognition accuracy in difficult and dynamically-varying acoustical environments is becoming increasingly important as speech technology is becoming a more integral part of practical applications. Results of numerous studies have demonstrated that even automatic speech recognition systems that are designed to be speaker independent can perform very poorly when they are tested using a different type of microphone or acoustical environment from the one with which they were trained (*e.g.* [2, 1, 11, 23]), even in a relatively quiet office environment. Applications such as speech recognition over telephones, in automobiles, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

Over the years the Carnegie Mellon University (CMU) robust speech group has worked to improve speech recognition accuracy through the use of environmental compensation procedures that modify either the feature vectors of incoming speech or the internal statistics with which speech recognition systems are trained. We have also explored complementary approaches to robust recognition based on the use of arrays of multiple microphones and on the use of physiologically-motivated

approaches to initial signal processing. This paper will review the context that motivated some of our more recent approaches to environmental compensation, and will compare the performance of these approaches with previous techniques developed at CMU and elsewhere.

2. BACKGROUND: EFFECTS OF UNKNOWN NOISE AND FILTERING

There are many sources of acoustical distortion that can degrade the accuracy of speech recognition systems. For many speech recognition applications the two most important sources of acoustical degradation are *unknown additive noise* (from sources such as machinery, ambient air flow, and speech babble from background talkers) and *unknown linear filtering* (from sources such as reverberation from surface reflections in a room, and spectral shaping by microphones or by the vocal tracts of individual speakers). Other sources of degradation of recognition accuracy include transient interference to the speech signal (such as the noises produced by doors slamming or telephones ringing), nonlinear distortion (arising from sources such as carbon-button microphones or the random phase jitter in telephone systems), and “co-channel” interference by individual competing talkers. Most research in robust recognition has been directed toward compensation for the effects of additive noise and linear filtering.

Research in robust speech recognition has been strongly influenced by earlier work in speech enhancement. Two seminal speech enhancement algorithms have proved to be especially important in the development of strategies to cope with unknown noise and filtering. The first technique, *spectral subtraction*, was introduced by Boll [6] to compensate for additive noise. In general, spectral subtraction algorithms attempt to estimate the power spectrum of additive noise in the absence of speech, and then subtract that spectral estimate from the power spectrum of the overall input (which normally includes the sum of speech plus noise). The algorithm was later extended by Berouti *et al.* [5] and many others, primarily with the goal of avoiding “musical noise” by “over-subtraction” of the noise spectrum. The second major technique is *spectral normalization*, introduced by Stockham *et al.* [24] to compensate for the effects of unknown linear filtering. In general, spectral normal-

***Current address:** Pedro J. Moreno, Digital Equipment Corporation, Cambridge Research Laboratory, One Kendall Square, Bldg. 700, Cambridge, MA 02139-1562, USA.

ization algorithms first attempt to estimate the average power spectra of speech in the training and testing domains, and then apply the linear filter to the testing speech to “best” convert its spectrum to that of the training speech. Improvements and extensions of spectral subtraction and spectral normalization algorithms continue to be introduced to this date.

2.1. A Model of the Environment

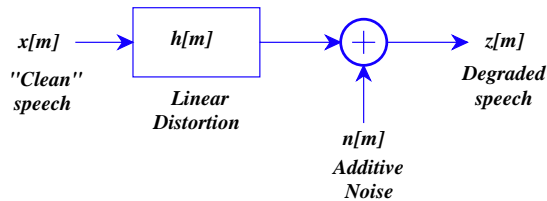


Figure 1. A model of environmental distortion including the effects of additive noise and linear filtering.

Figure 1 describes the implicit model for environmental degradation used in many signal processing algorithms developed at CMU and elsewhere. We assume that the “clean” speech signal $x[m]$ is first passed through a linear filter $h[m]$ whose output is then corrupted by uncorrelated additive noise $n[m]$ to produce the degraded speech signal $z[m]$. We characterize the power spectral densities (PSD) of the processes involved as

$$P_z(\omega) = P_x(\omega)|H(\omega)|^2 + P_n(\omega) \quad (1)$$

We can also represent the corresponding effects of noise and filtering on the input speech in the log-spectral or cepstral domains as

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

where x , n , q , and z are the logarithms and inverse transforms of the logs of $P_x(\omega)$, $P_n(\omega)$, $P_q(\omega)$, and $P_z(\omega)$, respectively. This equation can be rewritten in the form of

$$z = x + q + r(x, n, q) = x + f(x, n, q) \quad (3)$$

where $f(x, n, q)$ is referred to as the “environment function”.

In the above equations, the vectors z that represent the observed speech are considered to have been obtained by additive perturbations of the original speech features x . The environment function $f(x, n, q)$ represents the effects of additive noise and linear filtering on the feature vectors characterizing the incoming speech. In general, our goal is to obtain an estimate of x , the representation of $x[m]$ in the feature space, from z , the corresponding representation of $z[m]$.

Performing compensation in the cepstral domain (as opposed to the spectral domain) has the advantage that a smaller number of parameters needs to be estimated. In addition, cepstral-based features are widely used by current speech recognition systems. On the other hand, for some of the compensation pro-

cedures considered, the statistical models are more accurate or more easily developed in the log-spectral domain.

2.2. Approaches to Environmental Compensation

In this section we review several types of approaches to the problem of joint compensation for the effects of noise and filtering. We find it convenient to group these algorithms into three classes: (1) empirical compensation by direct cepstral comparison, (2) model-based compensation, and (3) compensation via cepstral high-pass filtering. In the case of the first two of these approaches there is also the second independent choice of whether to use a compensation procedure that modifies the feature vectors of the incoming speech or one that modifies the internal statistics of the recognition system itself.

Empirical compensation by direct cepstral comparison is totally data driven, and requires a “stereo” database that contains time-aligned samples of speech that had been simultaneously recorded in the training environment and in representative testing environments. The success of empirical compensation approaches depends on the extent to which the putative testing environments used to develop the parameters of the compensation algorithm are in fact representative of the actual testing environment.

In contrast, model-based compensation assumes a structural model of environmental degradation, such as the one depicted in Fig. 1. Compensation is then provided by applying the appropriate inverse operations. The success of model-based approaches depends on the extent to which the model of degradation used in the compensation process accurately describes the true nature of the degradation to which the speech had been subjected.

As the name implies, compensation by high-pass filtering implies removal of the steady-state components of the cepstral vector. The amount of compensation provided by high-pass filtering is more limited than the compensation provided by the two other types of approaches, but the procedures employed are so simple and effective that they should be included in virtually every current speech recognition system.

From a historical standpoint, research at CMU on algorithms to provide joint compensation for the effects of noise and filtering has proceeded in two phases. In the initial phase (which spanned the period of approximately 1988-1994) we were primarily concerned with understanding the basic properties of the environment function and with the development of compensation procedures that were relatively simple but that provided significant improvements in recognition accuracy compared to the accuracy that could be obtained from independent compensation for the effects of noise and filtering. During the second phase of algorithm development (roughly since 1994) our efforts focussed on the development of algorithms that could achieve greater recognition accuracy under the most difficult conditions through the use of more accurate mathematical characterizations of the effects of noise and filtering. We describe in the following section many of the results

obtained during the initial phase of this investigation. In Sec. 4 we review and discuss the second series of algorithms which provide greater recognition accuracy by virtue of more detailed modeling of the statistics of degraded speech.

3. INITIAL APPROACHES TO ENVIRONMENTAL COMPENSATION

3.1. Empirical Compensation: SDCN, FCDCN, MFCDCN, and MPDCN

As noted above, empirical cepstral comparison procedures assume the existence of “stereo” databases containing speech that had been simultaneously recorded in the training environment and one or more prototype testing environments. Our initial work on empirical compensation made use of cepstral features, and the effects of the environment function were expressed through additive cepstral “compensation vectors”. These compensation vectors were calculated by computing the frame-by-frame differences between the cepstral vectors representing speech in the training and testing environments:

$$\hat{x} = z + \hat{f}(x, n, q) = \bar{x} - \bar{z} \quad (4)$$

where $\hat{f}(x, n, q)$ is a set of vectors that serve to estimate the environment function. In general, these vectors can depend on instantaneous SNR, the specific vector-quantized (VQ) cluster location that is nearest to the incoming feature vector, the presumed phonemic identity, and the specific testing environment.

Applying the compensation is equally simple, as the compensation vector is just added to the incoming cepstral vector to produce an estimate of the original cepstral vector.

The goal of compensation is normally to provide relief from the effects of both additive noise and linear filtering, which affect different speech frames differently. For example, at high SNRs, the environment function $f(x, n, q)$ primarily represents the effects of linear filtering, because under these circumstances the impact of additive noise is negligible. At the lowest SNRs, the vectors primarily compensate for the effects of additive noise, because under these circumstances Eq. (3) is dominated by the effects of the additive noise. At intermediate SNRs, the compensation vectors perform a combination of compensation for the effects of noise and filtering. Compensation using direct cepstral comparison is generally rather simple to apply, although its utility is limited by the coverage of the stereo training data.

The empirical approach to cepstral comparison can be most easily understood by considering the simplest cepstral comparison algorithm developed at CMU, *SNR-Dependent Cepstral Normalization* (SDCN) [1]. Compensation vectors for the SDCN algorithm are developed using a “stereo” database consisting of speech that has been simultaneously recorded in the training and testing environments. Individual frames are partitioned into subsets according to the SNR in each frame (as inferred from the total frame energy) in the testing environ-

ment. Compensation vectors corresponding to a given range of SNRs are estimated by calculating the average difference between cepstral vectors in the training and testing environments for all frames with that particular range of SNRs. The ensemble of compensation vectors constitutes an empirical characterization of the differences between the training and testing environments. When a new test utterance is presented to the classifier, the SNR is estimated for each frame of the input speech, and the appropriate compensation vector is added to the cepstral coefficients derived from the input speech for that frame.

The *Fixed Codeword-Dependent Cepstral Normalization* (FCDCN) algorithm [1] produces greater recognition accuracy by developing a more fine-grained set of compensation vectors for a particular testing environment. Compensation vectors for FCDCN are obtained by first partitioning the frames of speech from a stereo development corpus according to SNR, as with SDCN. A second partitioning of the development corpus is then obtained by vector quantizing (VQ) the cepstral coefficients at each SNR in the testing environment. Individual compensation vectors are developed for each VQ cluster location at each SNR.

The *Phone-Dependent Cepstral Normalization* (PDCN) algorithm [13] is similar in philosophy, but it makes use of a different type of partitioning of the input frames. Compensation vectors are developed that depend on the presumed phoneme to which a given frame belongs. Phoneme hypotheses are obtained by running an initial pass of the HMM decoder without compensation. The PDCN algorithm is somewhat similar in concept to the method proposed by Beattie and Young [3].

Although all of the compensation algorithms described above were designed to work in the specific testing environment used to develop the compensation vectors, a degree of environmental independence can be obtained if several stereo training databases are available using different testing environments. Environment-independent compensation is performed by first determining which of the environments used to develop compensation vectors most closely resembles the actual testing environment. The ensemble of compensation vectors that is appropriate for that most likely environment is then applied to the incoming data. If the incoming speech is not from one of the environments used to develop compensation vectors, recognition accuracy can be further improved by interpolating among the several “closest” environments. Environmental classification need not be perfect for these algorithms to be effective. The “multiple-environment” versions of FCDCN and PDCN are referred to as MFCDCN and MPDCN.

These approaches are also similar to complementary work performed at other sites include piecewise-linear mapping and noise-adaptive prototypes developed at IBM [4, 18] and the probabilistic optimal filtering (POF) algorithm developed at SRI [19]. The POF algorithm, for example, is typically realized with many more free environmental parameters than are commonly used in algorithms like MFCDCN or MPDCN to characterize the environment function. POF also makes additional

use of temporal correlation across frames, which are not exploited by MFCCDN or MPDCN.

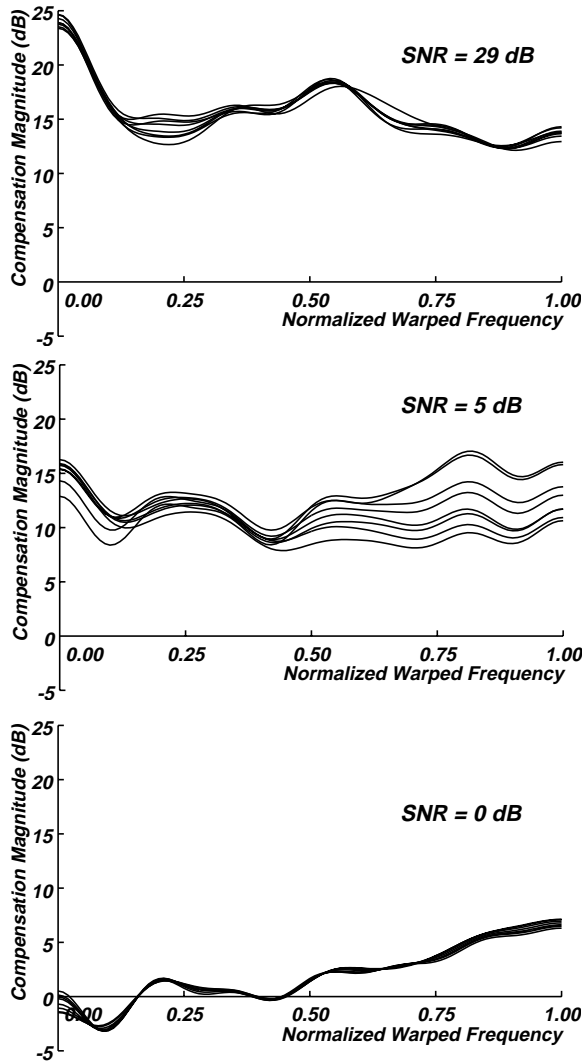


Figure 2. Power spectra of compensation vectors used by the FCDCN algorithm. The compensation vectors are based on three different SNRs and eight VQ codeword locations at each SNR. The training environment is a close-talking Sennheiser HMD-224 microphone, while the testing environment is a unidirectional desktop PCC-160 microphone.

Figure 2 illustrates some typical compensation vectors produced by the MFCCDN algorithm. A close-talking Sennheiser HMD-224 microphone was used for the training data, and the unidirectional desktop PCC-160 desktop microphone was used in the testing environments. Fig. 2 depicts MFCCDN compensation vectors, plotted at the extreme SNRs of 0 and 29 dB, as well as at 5 dB. Spectral representation of compensation vectors are plotted for 8 VQ cluster locations at each value of SNR. The curves are obtained by calculating the cosine transforms of the cepstral compensation vectors, $\hat{f}(x, n, q)$, which provide an estimate of the effective spectral profile of the compensation vectors. The horizontal frequency axis is warped

nonlinearly according to the mel scale [7]. The maximum frequency corresponds to the Nyquist frequency, 8,000 Hz. We note that the spectral profiles of the compensation vectors vary with SNR. This confirms our assertion that the vectors needed to compensate for the effects of linear filtering (which are dominant at high SNRs) are different from the vectors needed to compensate for the effects of additive noise (which dominate at low SNRs). Furthermore, at intermediate SNRs (such as 5 dB), additional improvement in recognition accuracy can be obtained by developing separate compensation vectors for the different VQ clusters within a given SNR. Compensation vectors for speech frames with SNRs that are greater than 10 dB are very similar in appearance to the compensation vectors shown for 29 dB.

3.2. Model-Based Compensation: CDCN

The compensation algorithms described in the previous section depend on frame-by-frame empirical comparisons of cepstral coefficients in the training and testing domains. An alternate approach to compensation is the use of a parametric model of degradation, combined with optimal estimation of the parameters of the model. For example, Ephraim [8] has presented a unified view of statistical model-based speech enhancement that can be applied to speech enhancement (for human listeners), speech coding, and enhanced robustness for automatic speech recognition systems. Varga and Moore [25] and Gales and Young [9] have also developed algorithms that modify the parameters of HMMs to characterize the effects of noise on speech. Sankar and Lee [22] have used a linear transform to reduce distortions between training and testing environments of the incoming features or model parameters of the HMM. Most of the above approaches have been developed primarily to ameliorate the effects of pure additive noise on speech. Acero's *Codeword-Dependent Cepstral Normalization* (CDCN) algorithm [2, 1] is similar in principle, except that it was developed explicitly to provide for joint compensation for the effects of additive noise combined with linear filtering.

The CDCN algorithm assumes the model of environmental degradation shown in Fig. 1. The algorithm attempts to reverse the effects of the linear filter with transfer function $H(\omega)$ and the additive noise with power spectrum $P_n(\omega)$ by solving two independent problems. The first problem is that of estimating the parameters q and n , which define the environment function in Eq. (3). This is accomplished using ML parameter estimation. The second problem is estimation of the uncorrupted cepstral vector x for a particular input frame, given the corrupted observation vector z and the environment parameters q and n . MMSE parameter estimation is used for this task. In effect, these two operations determine the values of q and n that when applied in inverse fashion map the set of input cepstra z into a set of compensated cepstral coefficients x that are as “close” as possible to the VQ codeword locations encountered in the training data. CDCN is typically implemented on a sentence-by-sentence basis.

Although model-based compensation is somewhat more computationally intensive than compensation based on empirical comparisons, the bulk of the computational cost is incurred in estimating the environment parameters q and n . Since distortion due to noise and filtering changes relatively slowly, it is generally not necessary to compute new values for these parameters for every incoming speech frame. The compensation itself must be applied to each incoming frame, but this does not entail great computational cost.

Model-based compensation can provide effective compensation if the assumptions built into the structural model are valid, even if only a small amount of speech is available in the testing environment. For example, in our implementations of CDCN, we typically apply compensation on a sentence-by-sentence basis.

3.3. Cepstral High-Pass Filtering: RASTA and CMN

We comment in passing on the third major adaptation technique, cepstral high-pass filtering, which provides a remarkable amount of robustness at almost zero computational cost. The development of these algorithms was originally motivated by a desire to emphasize the transient aspects of speech representations.

In the well-known *Relative Spectral Processing* or *RASTA* processing [10], a high-pass (or band-pass) filter is applied to a log-spectral representation or cepstral representation of speech. *Cepstral mean normalization* (CMN) is an alternate way to high-pass filter cepstral coefficients. High-pass filtering in CMN is accomplished by subtracting the short-term average of cepstral vectors from the incoming cepstral coefficients.

Algorithms like RASTA and CMN are effective in compensating for the effects of unknown linear filtering in the absence of additive noise because under these circumstances the ideal cepstral compensation vector $\hat{f}(x, n, q)$ is a constant that is independent of SNR and VQ cluster identity. Such a compensation vector is, in fact, equal to the long-term average difference between all cepstra of speech in the training and testing environments. The high-pass nature of both the RASTA and CMN filters forces the average values of cepstral coefficients to be zero in the training and testing environments individually, which, of course, implies that the average cepstra in the two environments are equal to each other.

Cepstral high-pass filtering can also be thought of as a degenerate case of compensation based on direct cepstral comparison. Consider, for example, the compensation vectors with frequency response depicted in Fig. 2. Cepstral high-pass filtering produces the same effect that would have been achieved if all of the compensation vectors for a particular testing environment are combined into a *single* compensation vector, weighted in proportion to the percentage of frames having the set of physical parameters (or presumed phoneme identity) corresponding to each of the original compensation vectors. As Fig. 2 indicates, actual cepstral compensation vectors depend

on the SNR, VQ codeword location, and/or phonemic identity of the individual frames of the testing utterances. Hence neither CMN nor RASTA can compensate directly for all of the combined effects of additive noise and linear filtering.

In general, cepstral high-pass filtering is so inexpensive and effective that it is currently embedded in some form in virtually all systems that are required to perform robust speech recognition.

3.4. Performance of Compensation Algorithms

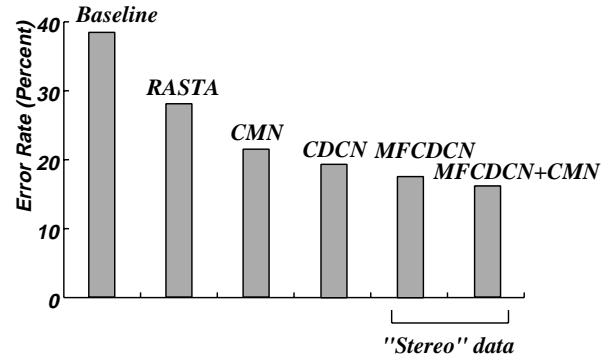


Figure 3. Comparison of the effects of CMN, the original RASTA algorithm, CDCN, MFCDCN, and a combination of MFCDCN and CMN on the recognition accuracy obtained for the “secondary microphones” of the 1992 ARPA 5000-word Wall Street Journal task. The system was trained using a close-talking microphone.

The performance of the compensation algorithms described above has been described and compared in several previous papers including [23]. Figure 3 compares the recognition accuracy obtained using CMN, the original RASTA algorithm, CDCN, MFCDCN, and a combination of MFCDCN and CMN [12]. These comparisons were obtained using SPHINX-II and data from the “secondary” microphones of Version 0 of the 5000-word 1992 Wall Street Journal evaluation set (WSJ0).

We note that both the RASTA and CMN algorithms provided substantially better recognition error rates than the rates obtained using the then-current “baseline” processing using LPC-derived cepstra. Use of the model-based CDCN algorithm provided an additional 10 percent relative reduction in error rate compared to CMN. A 24.3 percent reduction in relative error rate was obtained by adding MFCDCN processing to CMN, although MFCDCN requires the use of a stereo training database. We believed that the CDCN and MFCDCN algorithms provided greater recognition accuracy than cepstral high-pass filtering because they provide for different types of compensation under differing conditions, either through an ensemble of empirically-derived functions or through a parametric model of the degradation process. This is equivalent to recognizing that there are a number of different environment functions represented by the curves of Fig. 2, rather than just a single condition-independent function.

4. CURRENT COMPENSATION APPROACHES

Although the compensation algorithms described in Sec. 3 above provided substantial improvements in recognition accuracy in a number of environments, they still exhibited many obvious shortcomings. The goals of our more recent work on environmental compensation focused included greater recognition accuracy at lower SNRs, better performance with small amounts of environment-specific adaptation data, and the elimination of the need for “stereo” data.

As before, we developed algorithms based both on empirical comparisons of speech in the training and testing environments, and on mathematical models of the effects of degradation. The major empirically-derived algorithms that emerged from these efforts thus far have been *Multivariate Gaussian-Based Cepstral Normalization* (RATZ) and *Statistical Re-estimation of HMMs* (STAR). The most important model-based algorithms developed in recent years are the *Vector Taylor Series* (VTS) and *Vector Polynomial Approximations* (VPS) algorithms. RATZ, STAR, and the initial development of VTS are all described in detail by Moreno in [16]; an additional discussion of VTS and VPS may be found elsewhere in these Proceedings [21].

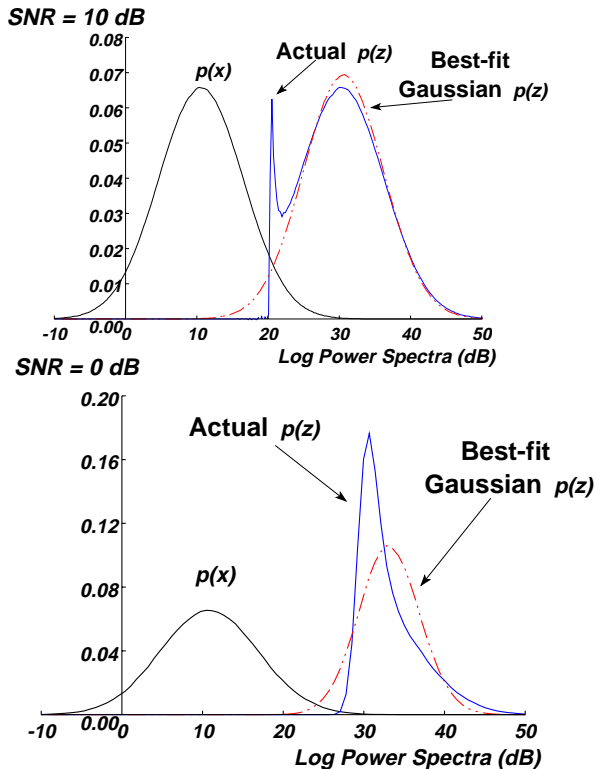


Figure 4. The effect of noise on the probability density functions of the logs of power spectra.

Much of our recent work was motivated by straightforward observations of the effects of noise on the probability distributions of features commonly used in speech recognition systems. For example, Fig. 4 demonstrates how additive noise and linear filtering can affect the probability densities of log spectral features of speech.

These simulations show the densities of normally-distributed “clean” log spectra $p(x)$ and the corresponding densities of “noisy” log spectra $p(z)$ after exposure to additive noise and linear filtering. We note that as SNR decreases, the mean of $p(z)$ shifts and its variance decreases. It can easily be seen that the resulting density is no longer Gaussian, and similar shape distortions are easily observed for the case of features that were originally characterized by Gaussian mixtures as well. Unfortunately, there is no tractable analytical expression for the means and variances of the random vector z in Eq. (3) which characterizes the degraded speech. Hence, the goal of compensation algorithms developed by our group and other sites is always to obtain a reasonably accurate estimate of $p(z)$ or its moments by empirical observation, by parametric models, or by series approximation of the environment function itself.

4.1. Empirical Compensation: RATZ and STAR

In this section we describe the RATZ and STAR algorithms in somewhat greater detail. The RATZ algorithm modifies the cepstral vectors of incoming speech, while the STAR algorithm modifies the internal statistical models used by the recognition system. Nevertheless, RATZ and STAR have a very similar conceptual framework, as is elaborated in [15]. While RATZ can be considered to be a generalization of algorithms like MFCDCN and STAR can be considered to be an extension of the codebook adaptation algorithms described in [13], the mathematical framework for them has been developed more carefully and accurately.

RATZ and STAR both assume that the probability density function for clean speech can be characterized as a mixture density

$$p(x) = \sum_k a_k(t) N_k(\mu_k, \sigma_k) \quad (5)$$

where the mixture coefficients a_k are fixed for the case of RATZ, and assumed to vary as a function of time to represent the Markov transitional probabilities for the case of STAR.

Environmental compensation is introduced by modifying the means and variances of the probability density functions:

$$\hat{\mu}_k = \mu_k + r_k \quad \text{and} \quad (6)$$

$$\hat{\Sigma}_k = \Sigma_k + R_k \quad (7)$$

where r_k and R_k are the factors that compensate the means and variances respectively.

As described in [14], direct solutions for the parameters r_k and R_k can be obtained if “stereo” data are available. For example in the case of RATZ, these vectors can be obtained from the equations

$$\mathbf{r}_k = \frac{\sum_{i=0}^{N-1} (z_i - \mathbf{x}_i) P(k|x_i)}{\sum_{i=0}^{N-1} P(k|x_i)} \quad (8)$$

and

$$\mathbf{R}_k = \frac{\sum_{i=0}^{N-1} (z_i - \mu_{z,k})(z_i - \mu_{z,k})^T P(k|x_i)}{\sum_{i=0}^{N-1} P(k|x_i)} \quad (9)$$

where the parameter i refers to the analysis frame and evolves over time.

If “stereo” data are not available, the recursive EM technique must be used to obtain estimates of the parameters \mathbf{r}_k and \mathbf{R}_k using a very similar equation.

A further extension of RATZ described in [15] is referred to as SNR-RATZ. SNR-RATZ uses a more structured model for $p(z)$ whereby the number of Gaussians used to define the statistics for X_0 can be different from the number used for the other cepstral components. The statistics of the remaining components of \mathbf{x} , are tied to the individual Gaussians that comprise the component X_0 to which they belong, so they can exhibit different statistics for different SNRs. The means, variances and *a priori* probabilities of the individual Gaussians are learned by standard EM methods, as before.

In the case of the STAR algorithm, the correction parameters \mathbf{r}_k and \mathbf{R}_k are computed as in Eqs. (8) and (9). It is assumed that the *a posteriori* probabilities in these equations do not change due to the effects of noise or filtering and can be computed from the clean speech. The correction factors are then applied to the cepstra, delta-cepstra, and double delta-cepstra produced by SPHINX-II, along with the cepstral component c_0 , its difference, and its double difference. (In practice, we have observed that adapting the cepstral double-delta statistics does not affect the recognition performance.) Once the correction terms are computed, the Gaussians are adapted to the new environment as in Eqs. (6) and (7).

Figure 5 compares the recognition accuracy obtained using RATZ and FCDCN. These and subsequent results were obtained by testing on the ARPA 5000-word Wall Street Journal evaluation test data, after the data had been corrupted by additive noise and linear filtering, with word recognition accuracy plotted as a function of SNR. In addition to the results obtained using RATZ and FCDCN, we also plot the recognition accuracy obtained using CMN alone, and using a system that had been completely retrained on the degraded speech for each separate condition of degraded speech. These two curves provide a reasonable estimate of the best and worst performance to be expected for any specific combination of recognition system, recognition task, and feature set. Hence these

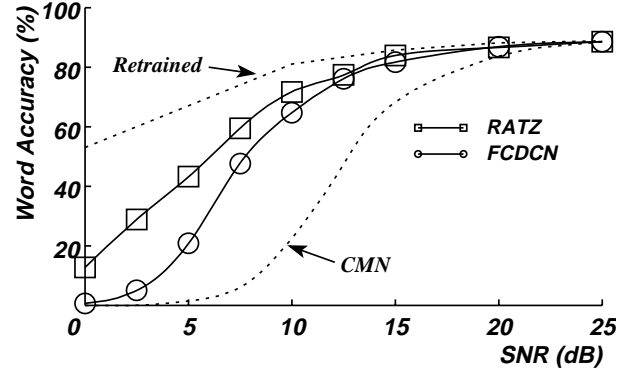


Figure 5. Comparison of recognition accuracy obtained using the RATZ and FCDCN, the empirically-derived compensation algorithm that previously had exhibited the best performance. The system was evaluated on speech samples from the 1993 5000-word ARPA Wall Street Journal evaluation test set, after the speech had been corrupted by linear filtering and additive noise. Results are plotted as a function of SNR.

curves provide reasonable bounds on the degree of performance to be expected from an environmental compensation algorithm such as the ones we describe. We note that the recognition accuracy obtained using RATZ is greater than the recognition accuracy obtained using FCDCN, particularly at the lower SNRs, and that the recognition accuracy obtained using both procedures approaches best possible performance for SNRs down to about 15 dB. We believe that better performance is obtained using RATZ because it includes a more detailed model that characterizes the effects of the environment on the variances of the speech features. This allows the compensation procedure to reflect (in a limited way) the changes in variance of the features due to the effects of the noise.

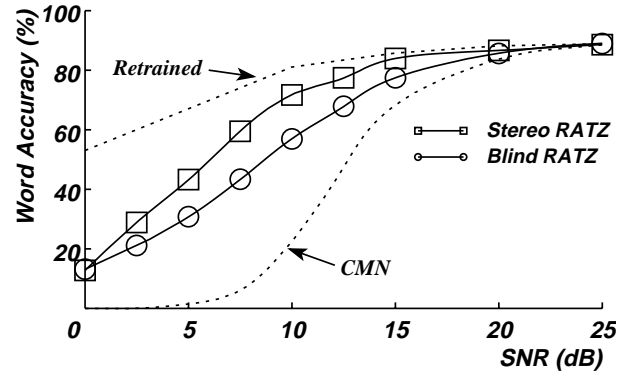


Figure 6. Comparison of recognition accuracy obtained using RATZ with compensation factors derived from “stereo” data and using a “blind” version of RATZ without access to stereo data.

Figure 6 compares recognition accuracies obtained using the original version of RATZ and using a “blind” implementation that does not make use of “stereo” data. Although it is not a surprise that better recognition accuracy is obtained using stereo data are available, the word accuracy obtained using Blind RATZ is quite a bit better than that obtained using CMN, and is in fact comparable to the word accuracy observed for FCDCN using stereo data, as plotted in Fig. 5.

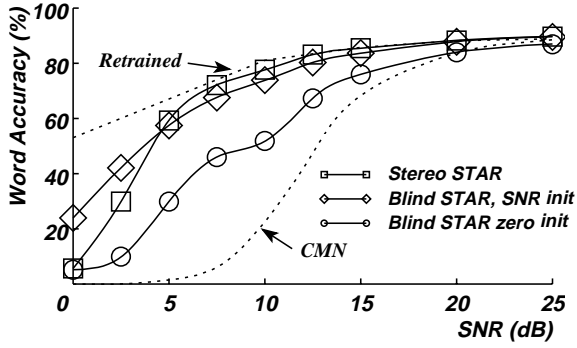


Figure 7. Comparison of recognition accuracy obtained using RATZ with compensation factors derived from “stereo” data and using a “blind” version of RATZ without access to stereo data.

Figure 6 compares recognition accuracy obtained using the STAR algorithm as developed from “stereo” training data, along with two implementations of a “blind” STAR algorithm. The latter two curves describe the effects of two different sets of initial values for the correction factors: The curve with the circular symbols represents truly “blind” performance in that the re-estimation process used to obtain the correction factors was initialized using clean speech. The intermediate curve with the diamond symbols represents results obtained by initializing on speech that has been corrupted by noise at a comparable SNR to the test data.

Although the performance obtained using the “blind” implementation is somewhat dependent on the initial conditions, it is clear that the STAR algorithm trained with “stereo” data provides much better recognition accuracy than RATZ. STAR, in fact, provides approximately the best possible recognition accuracy for SNRs down to about 7 dB.

We believe that STAR is superior, especially at low SNRs, because signal processing algorithms such as RATZ that attempt to correct for the effects of noise do not account completely for the changes of ideal classification boundaries that occur due to the effects of noise on the variances of the distributions. Furthermore, additional approximation errors are introduced in the MMSE process used to actually perform the compensation (once the parameters are estimated), leaving a residual mismatch between the estimates of “clean” speech and the original HMMs. In contrast, classifier adaptation algorithms such as STAR modify the variances as well as the means in the internal representation of the incoming features. This is a better approximation to the ideal condition where training and testing are performed in the same environment.

4.2. Model-Based Compensation: VTS and VPS

The Vector Taylor Series (VTS) [16, 17] and Vector Polynomial Expansion (VPS) [20, 21] procedures are model-based algorithms that develop series approximations to the nonlinear environment function $f(x, n, q)$ defined in Eq. (3)

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

For example, the VTS algorithm approximates the environment function $f(x, n, q)$ using the first several terms of its Taylor series:

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

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

The series expansion of the environment function is particularly convenient because the means and variances of the series approximations are quite easily obtained. The EM algorithm is then used to find the values of n and q that maximize the likelihood of the observations, and the statistics of the incoming cepstral vectors are re-estimated using MMSE techniques.

The VPS algorithm is described in detail elsewhere in these Proceedings [21]. Briefly, the VPS approach replaces the Taylor series expansion used in VTS with a more general approach to approximating the environment function. VPS is shown to provide a more accurate approximation to the environment function than VTS. In pilot evaluations described in [21], VPS provided somewhat better recognition accuracy compared to VTS, and at a reduced computational cost. The original versions of the VTS and VPS algorithms were implemented only to modify the incoming speech feature vectors. It is expected that the difference in error rates between VPS and VTS will increase when implementations of these algorithms that modify the internal statistical models are completed.

Figure 8 shows how the resulting means and variances of the noisy vector set z can be approximated quite well by the Taylor series expansion. It can be seen that the zeroth-order VTS expansion provides a reasonably good approximation to the mean of z , but at lower SNRs the second-order expansion provides an even better approximation. Similarly, the first-order approximation is closer to the real variance than the zeroth-order approximation.

Figure 9 compares the recognition accuracies obtained using three model-based compensation procedures: the zeroth-order and first-order approximations of VTS and CDCN. It can be seen that at all SNRs the first-order VTS algorithm outperforms the zeroth-order VTS algorithm, which in turn outperforms CDCN. In fact, the zeroth-order VTS algorithm also outperforms RATZ, which is an algorithm that assumes the availability of “stereo” data.

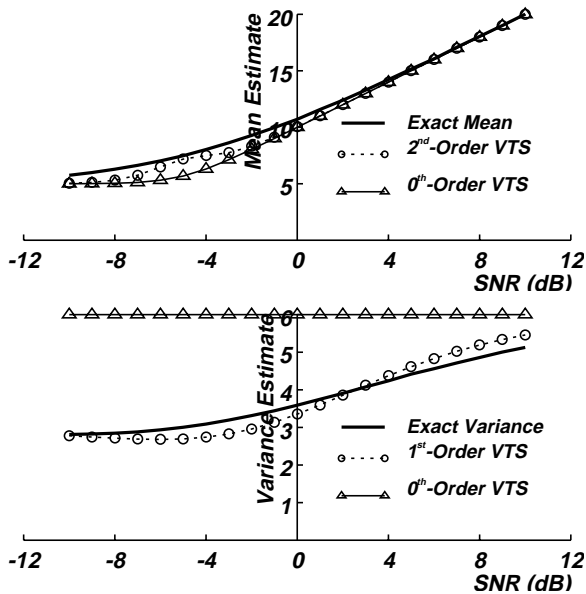


Figure 8. Simulated effects of noise on the means and variances of incoming features, as well as their VTS approximates of selected orders.

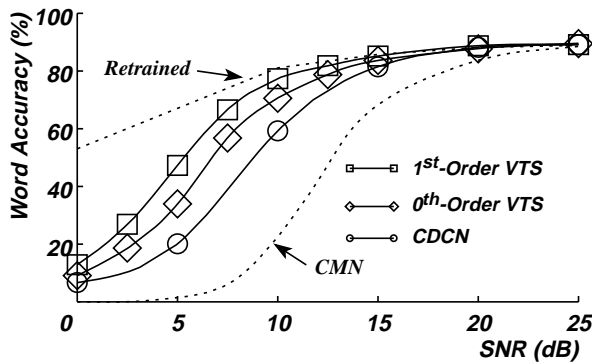


Figure 9. Comparison of recognition accuracies obtained on the ARPA 5000-word Wall Street Journal task using the zeroth-order VTS, first-order VTS, and CDCN.

5. SUMMARY

In the sections above, we have reviewed many of the major approaches taken by the CMU speech group for robust speech recognition over the past 10 years. The current ensemble of compensation algorithms, including RATZ, STAR, and VTS, has demonstrated significant improvements in recognition accuracy compared to what had been obtained previously using algorithms such as CDCN, FCDCN, and MFCDCN. In general these improvements were obtained using a general approach that was similar to past work, but using more elaborated models of the effect of degradation. We believe that the newer algorithms exhibit improved performance in part because they model more accurately environmental effects on feature variance, and in part because they compensation algorithms are now more tightly linked to the mathematical representation used by the HMMs.

We also note the following specific comments:

- Relative improvements in recognition accuracy provided by the newer and more mathematically-detailed algorithms are greater at the lower SNRs.
- Algorithms (such as STAR) that modify the statistical models used by classifiers provides greater recognition accuracy than algorithms (such as RATZ) that modify the incoming feature vectors.
- If “stereo” data are not available, model-based algorithms (such as VTS) provide greater recognition accuracy than empirical approaches (such as RATZ), at the expense of somewhat greater computational complexity.

ACKNOWLEDGMENTS

We thank all of the members of the CMU speech group and especially Alejandro Acero, Evandro Gouvêa, Uday Jain, Fu-Hua Liu, Vipul Parikh, and Matthew Siegler for their contributions to this work. 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. A. Acero and R. M. Stern (1990). “Environmental Robustness in Automatic Speech Recognition,” *Proc. ICASSP-90*, pp. 849-852, 1990.
2. A. Acero (1993). *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.
3. V. L. Beattie (1992). *Hidden Markov Model State-Based Noise Compensation*, Ph.D. Thesis, Churchill College, Cambridge University.
4. J. R. Bellegarda, P. V. de Souza, A. J. Nadas, D. Nahamoo, M. A. Picheny, L. R. Bahl, (1992). “An Efficient Procedure for Robust Speaker Adaptation Using a Piecewise Linear Acoustic Mapping”, *Signal Processing VI - Theories and Applications. Proc. EUSIPCO-92, Sixth European Signal Processing Conference*.
5. M. Berouti, R. Schwartz, and J. Makhoul (1979). “Enhancement of Speech Corrupted by Acoustic Noise,” *Proc. ICASSP-79*.
6. S. F. Boll (1979). “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” *IEEE Trans. Acoustics, Speech and Signal Processing*, **27**: 113-120.
7. S. B. Davis, and P. Mermelstein (1980). “Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. Acoustics, Speech and Signal Processing* **28**: 357-366.
8. Y. Ephraim (1992). “Statistical-model-based Speech Enhancement Systems,” *Proc. IEEE* **80**: 1526-1555.
9. M. J. F. Gales and S. J. Young (1993). “Cepstral Parameter

- Compensation for HMM Recognition in Noise,” *Speech Communication* **12**: 231-239.
10. H. Hermansky, and N. Morgan, N. (1994). “RASTA Processing of Speech”, *IEEE Trans. Speech Audio Process.* **2**: 578-89.
 11. B.-H. Juang, (1991). “Speech Recognition in Adverse Environments,” *Computer Speech and Language* **5**: 275-294.
 12. F.-H. Liu, R. M. Stern, X. Huang and A. Acero (1993). “Efficient Cepstral Normalization For Robust Speech Recognition,” *Proc. DARPA Speech and Natural Language Workshop* Princeton, NJ, Morgan Kaufmann, M. Bates, Ed.
 13. F.-H. Liu, R. M. Stern, A. Acero, A., and P. J. Moreno (1994). “Environment Normalization for Robust Speech Recognition using Direct Cepstral Comparison.” *Proc. ICASSP-94*.
 14. P. J. Moreno, B. Raj, E. Gouvea, and R. M. Stern (1995). “Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition”, *Proc. ICASSP-95*.
 15. P. J. Moreno, B. Raj, B., and R. M. Stern (1995). “A Unified Approach to Robust Speech Recognition,” *Proc. Eurospeech-95*.
 16. P. J. Moreno (1996). *Speech Recognition in Noisy Environments*, Ph. D. Dissertation, ECE Department, CMU, May 1996.
 17. P. J. Moreno, B. Raj, and R. M. Stern, R. M. (1996). “A Vector Taylor Series Approach for Environment-Independent Speech Recognition,” *Proc. ICASSP-96*.
 18. A. Nadas, D. Nahamoo, and M.A. Picheny (1989). “Speech Recognition Using Noise-Adaptive Prototypes,” *IEEE Trans. Acoustics, Speech and Signal Processing*, **37**: 1495-1503.
 19. L. Neumeyer and M. Weintraub (1994). “Probabilistic Optimum filtering for Robust Speech Recognition,” *Proc. ICASSP-94*.
 20. B. Raj, E. Gouvea, P. J. Moreno, and R. M. Stern (1996). “Cepstral Compensation by Polynomial approximation for Environment-Independent Speech Recognition,” *Proc. ICSLP-96*.
 21. B. Raj, E. Gouvea, and R. M. Stern, “Cepstral Compensation using Statistical Linearization” (1996). These Proceedings.
 22. A. Sankar and C.-H. Lee (1994). “Stochastic Matching for Robust Speech Recognition,” *IEEE Signal Processing Letters* **1**: 124-125.
 23. R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima (1996). “Signal Processing for Robust Speech Recognition”, in *Speech Recognition*, C.-H. Lee and F. Soong, Eds., Boston: Kluwer Academic Publishers.
 24. T. G. Stockham, T. M. Cannon and R. B. Ingebreetsen (1975). “Blind Deconvolution Through Digital Signal Processing,” *Proc. IEEE* **63**: 678-692.
 25. A. P. Varga and R. K. Moore (1990). “Hidden Markov Model Decomposition of Speech and Noise,” *Proc. ICASSP-90*.