A Comparison of Three Non-Linear Observation Models for Noisy Speech Features

Jasha Droppo, Li Deng, Alex Acero

Microsoft Research
Redmond, Washington 98052, USA

Abstract

This paper reports our recent efforts to develop a uni£ed, non-linear, stochastic model for estimating and removing the effects of additive noise on speech cepstra. The complete system consists of prior models for speech and noise, an observation model, and an inference algorithm. The observation model quanti£es the relationship between clean speech, noise, and the noisy observation. Since it is expressed in terms of the log Mel-frequency £lter-bank features, it is non-linear. The inference algorithm is the procedure by which the clean speech and noise are estimated from the noisy observation.

The most critical component of the system is the observation model. This paper derives a new approximation strategy and compares it with two existing approximations. It is shown that the new approximation uses half the calculation, and produces equivalent or improved word accuracy scores, when compared to previous techniques. We present noise-robust recognition results on the standard Aurora 2 task.

1. Introduction

It is well known that automatic speech recognition systems without provisions for noise robustness degrade quickly in the presence of additive or convolutional noise. The noise robustness can be constructed either in the model domain or the feature domain. Model domain techniques try to modify the acoustic model as if it were trained on speech similar to the current test utterance. By design, such a technique can not do better than a matched training and testing condition. On the other hand, feature domain techniques process the acoustic features before they arrive at the recognition system. It has been shown that feature domain techniques can achieve lower word error rates than the matched training and testing condition [1]. As a result, we continue to explore feature domain techniques.

The class of feature domain techniques is rich, including algorithms such as spectral subtraction [2], cepstral mean normalization [3], SPLICE [1] and Algonquin [4]. The systems discussed in this paper explore two extensions to the observation model in [4]. The £rst modi£cation is to provide a better £t to the data, and the other simpli£es the inference calculations.

This paper is organized as follows. Section 2 describes the overall system design, including the prior models for clean speech and noise. Section 3 describes the observation models that are used to unify the clean speech and noise prior models. Section 4 presents the inference methods used to approximate the posterior distribution of clean speech and noise, after the noisy observations are incorporated. Section 5 analyzes the effectiveness of the different system con£gurations presented in this paper.

2. System Description

The features for the system presented in this paper are log Mel-frequency £lter-bank features. This feature space has two advantages: it is linearly related to MFCC recognition parameters, and the corruption is independent across feature dimensions. This allows us to work with features that are closely related to those used for recognition, while maintaining computational ef£ciency.

The system £rst represents clean speech and noise as random vector processes. We have chosen multivariate Gaussian mixture models, with diagonal components, for clean speech x_t and noise n_t:

$$p(x_t) = \sum_{s_t} N(x_t; \mu^x_{s_t}, \sigma^x_{s_t}) p(s_t) \quad (1)$$

$$p(n_t) = N(n_t; \mu^n, \sigma^n). \quad (2)$$

The parameters of this prior model include the state-conditional means and variances, μ^x_s, σ^x_s, μ^n and σ^n, as well as the mixture component weights $p(s_t)$.

The observation model then represents the relationship among x_t, n_t, and the observed noisy signal y_t. It's noted that even for the case when the clean speech and noise features are known exactly, the noisy observation is not uniquely determined. The major cause of this uncertainty is the unknown relative phase between the speech and noise spectra. This random error will be treated differently by each of the three observation models explored in this paper.

The £rst observation model considered is the one described in [5]. It models the error as a random variable whose variance is a function of the local SNR. We will call this the SNR dependent variance model (SDVM). The advantage of using the SDVM is that it provides a good match to the data. The disadvantage is that rigorous inference takes a lot of computational power.

The second observation model we consider in this paper £rst appeared in [4]. It assumes the error is independent of x and n. We will call this the SNR independent variance model (SIVM). The advantage of using the SIVM is that it can be effectively used with fewer computations than the SDVM.

The third observation model presented for the £rst time in this paper simpli£es the above two models by ignoring the error term entirely. This yields a good £t to the data at the extreme SNR regions, and a slight mismatch in the $x \approx n$ region. We will call this the zero variance model (ZVM). It is a special case of either the SIVM or the SDVM, when the variance of the error term is set to zero. Inference with the ZVM is twice as fast as with the SIVM, with a slight improvement in word accuracy, as shown in section 5.

3. The Observation Models

All three observation models outlined in the previous section are built upon an approximate relationship between the log-spectra of the clean speech and noise. Details of this relationship can be found in [5] and [4].

In the time domain, it is assumed that the clean speech and noise mix linearly. For the log Mel-frequency £lter-bank features, this relationship is approximated by

$$y = \ln(e^x + e^n) + \epsilon. \tag{3}$$

If, within each £lter-bank, x and n are constant and have the same phase, then $\epsilon = 0$ and the relationship is exact. A stochastic ϵ was introduced to account for the fact that these assumptions are seldom true on real data. Although the expected value of ϵ may be zero, the actual samples from the random process are not.

3.1. SNR Dependent Variance Model

It has been shown previously that the variance of ϵ should be a function of x and n [5]. Under weak simplifying assumptions, the relationship between x, n, and y should be

$$e^y = e^x + e^n + 2\alpha e^{\frac{x+n}{2}}. \tag{4}$$

The random variable α accounts for the effects of the unseen random phase between x and n. We empirically found that α is well modeled by a zero-mean Gaussian with a variance of $\sigma_\alpha^2 = 0.15$. If $x \gg n$, then the £rst term dominates Eq. 4. When $n \gg x$, the second term dominates Eq. 4. It is only when x and n have similar magnitudes, that the third term becomes important.

This model produces the conditional observation probability function [5]

$$\ln p(y|x,n) = y - \frac{x+n}{2} - 8\ln 2\pi\sigma_\alpha^2 - \frac{(e^y - e^x - e^n)^2}{8\sigma_\alpha^2 e^{x+n}},$$

and provides a good approximation for all regions of the true distribution simultaneously. Unfortunately, although [5] presented a slow technique that works well, there is not yet any practical technique that achieves the theoretical advantage of using the SDVM.

3.2. SNR Independent Variance Model

Another option is to model ϵ as $N(\epsilon; 0, \psi^2)$, a zero-mean random variable independent of x and n.

The variance ψ^2 can be tuned to change the shape of the approximate observation probability. One must choose a value for ψ that trades off modeling the true variance where $x \approx n$ with a small value that agrees when $x \gg n$ or $x \ll n$. For very small values of ψ, the extreme SNRs are well modeled. Moderate values of ψ model the region near $x = n$ well. The SIVM produces the conditional observation probability function

$$p(y|x,n) = N(y; \ln(e^x + e^n), \psi^2) \tag{5}$$

and works well with simple inference techniques such as the iterative vector Taylor series (VTS) approximation [4].

3.3. Zero Variance Model

Finally, we introduce a new derivation that is a special case of either the SIVM or SDVM. It assumes the error term of Eq. 3 is equal to zero. This is a good approximation when $x \gg n$ or $n \gg x$, but under-estimates the true variance of ϵ around $x = n$.

This zero variance approximation is the same as used for spectral subtraction. Indeed, the classical spectral subtraction formula appears if we set $\epsilon = 0$ and solve Eq. 3 for x:

$$x = \ln(e^y - e^n).$$

The inherent problem is that, as the estimate of n approaches y, the estimate of x becomes unstable. Additionally, care must be taken to avoid estimating $n \geq y$. These problems are overcome by the technique described below.

Instead of estimating x or n and inferring the other variable, we can estimate the local SNR r, de£ned as

$$r = x - n.$$

Unlike the spectral subtraction formula above, we are free to estimate any real value for r. These can be mapped into estimates of x and n through,

$$x = y - \ln(e^r + 1) + r, \text{ and} \tag{6}$$
$$n = y - \ln(e^r + 1). \tag{7}$$

These formulas satisfy the intuition that as the SNR r gets more positive, x approaches y from below. As the SNR r gets more negative, n approaches y from below. The result is no longer unstable because, regardless of the SNR, x and n are always separated by exactly r.

Two immediate advantages of the ZVM are dimensionality reduction and implicit correlation modeling in the speech and noise posterior. In the SIVM or SDVM models, inference must be performed on both x and n. For the ZVM, inference is performed only on r. This yields an immediate halving in computational cost. Additionally, it is known that the posterior distribution of speech and noise should be correlated. With the SIVM or SDVM models, this correlation must be explicitly modeled. The ZVM includes this correlation in the posterior automatically, when appropriate.

3.4. Joint PDF for the ZVM

The joint PDF for the ZVM is a distribution over the clean speech x, the noise n, the observation y, the SNR r, and the speech state s.

$$p(y, r, x, n, s) = p(y|x,n)p(r|x,n)p(x,s)p(n).$$

The observation and SNR are both deterministic functions of x and n. As a result, the conditional probabilities $p(y|x,n)$ and $p(r|x,n)$ can be represented by Dirac delta functions:

$$p(y|x,n) = \delta\left(\ln\left(e^x + e^n\right) - y\right) \tag{8}$$
$$p(r|x,n) = \delta\left(x - n - r\right). \tag{9}$$

This allows us to marginalize the continuous variables x and n, as follows:

$$\begin{aligned}
& p(y,r,s) = \int dx \int dn\, p(y,r,x,n,s) \\
= & \int dx \int dn\, p(y|x,n)p(r|x,n)p(x,s)p(n) \\
= & \int dx \int dn\, \delta\left(\ln\left(e^x + e^n\right) - y\right) \delta\left(x - n - r\right) p(x,s)p(n) \\
= & \left. p(x,s)\right|_{x=y-\ln(e^r+1)+r} \left. p(n)\right|_{n=y-\ln(e^r+1)} \\
= & N(y - \ln\left(e^r + 1\right) + r; \mu_s^x, \sigma_s^x) p(s) \\
& N(y - \ln\left(e^r + 1\right); \mu^n, \sigma^n)
\end{aligned} \tag{10}$$

The only remaining continuous hidden variable is r. The behavior of this joint PDF is intuitive. At high SNR, $r \gg 0$, and

$$p(y, r, s) \approx N(y; \mu_s^x, \sigma_s^x) p(s) N(y - r; \mu^n, \sigma^n)$$

That is, the observation is assumed to be clean speech, and the noise is at a level r units below the observation. The converse is true for low SNR, where $r \ll 0$.

4. Estimation of Clean Speech

The inference algorithm produces posterior distributions for the hidden variables x and n, given the observation y.

All three systems use an iterative VTS approximation algorithm to the non-linear observation model. This algorithm proceeds as follows. For each mixture component,

1. Initialize expansion point at the mean of the prior.
2. Approximate the non-linear observation model with a vector Taylor series.
3. Update expansion point as mean of the approximate posterior.

Steps 2 and 3 are repeated until convergence; typically £ve iterations are suf£cient. The £nal step is to compute the MMSE estimate of the hidden variables given the observation, e.g.,

$$\hat{x} = E[x|y] = \sum_s E[x|y, s] p(s|y) \quad (11)$$

$$p(s|y) = \frac{p(y|s)p(s)}{\sum_s p(y|s)p(s)} \quad (12)$$

It is the nature of this iterative VTS that it converges to a local extremum of the model's posterior. The estimate $\hat{x}$ is only the MMSE estimate given the observation and the approximate model. If there is a discrepancy between the mean and mode of the exact model (as in the SDVM, below), then the inference is inaccurate.

4.1. SIVM

The iterative VTS inference algorithm for the SIVM was presented in [4]. It consists of an iterative VTS algorithm applied to the non-linear term in Eq. 5.

As Figure 1 shows, the SIVM produces a posterior whose mean and mode tend to coincide. So, if it converges to the correct mode, the expected value of the posterior is also correct.

4.2. SDVM

In general, iterative VTS with the SDVM produces poor feature enhancement accuracy. The root cause appears to be a discrepancy between the posterior mean and mode of the SDVM, as demonstrated in Figure 1. Although the posterior mean is the optimal estimate for clean speech, iterative VTS will converge to the mode instead.

An alternative, computationally intensive, inference algorithm for the SDVM model was presented in [5].

4.3. ZVM

This section derives an iterative VTS approximation for inference under the ZVM. Under this approximation, the non-linear function in Eqs. 6 and 7 becomes an af£ne function of r:

$$\ln(e^r + 1) \approx f(r_s^0) + F(r_s^0)(r - r_s^0). \quad (13)$$

The vector function $f(r_s^0)$ and the matrix function $F(r_s^0)$ represent the £rst two terms in the Taylor series expansion of $f(r) = \ln(e^r + 1)$ around the state-conditional point $r = r_s^0$.

$$\begin{aligned} f(r_s^0) &= \ln(e^{r_s^0} + 1) = f_s^0 \\ F(r_s^0) &= \text{diag}\left(\frac{1}{1 + e^{-r_s^0}}\right) = F_s^0 \end{aligned}$$

The distribution of r based on this approximation is derived by substituting the Taylor series approximation for $\ln(e^r + 1)$ into Eq. 10.

$$\begin{aligned} p(y, r, s) &\approx N(y - f_s^0 + F_s^0 r_s^0 - (F_s^0 - I)r; \mu_s^x, \sigma_s^x) \\ &\quad N(y - f_s^0 + F_s^0 r_s^0 - F_s^0 r; \mu^n, \sigma^n) p(s) \\ &= N(r; \hat{\mu_s^r}, \hat{\sigma_s^r}) N(a_s; b_s, C_s) p(s) \\ &= p(r|y, s) p(y|s) p(s) \end{aligned}$$

Standard Gaussian manipulation formulas are used to bring $p(y, r, s)$ into this factored form.

$$\begin{aligned} p(r|y, s) &= N\left(r; \hat{\mu_s^r}, \hat{\sigma_s^r}\right) \\ \left(\hat{\sigma_s^r}\right)^{-1} &= (F_s^0 - I)^T (\sigma_s^x)^{-1} (F_s^0 - I) + F_s^0 (\sigma^n)^{-1} F_s^0 \\ \hat{\mu_s^r} &= \hat{\sigma_s^r} (F_s^0 - I)^T (\sigma_s^x)^{-1} (y - f_s^0 + F_s^0 r_s^0 - \mu_s^x) \\ &\quad + \hat{\sigma_s^r} F_s^0 (\sigma^n)^{-1} (y - f_s^0 + F_s^0 r_s^0 - \mu^n)) \\ p(y|s) &= N(a_s; b_s, C_s) \\ a_s &= y - f_s^0 + F_s^0 r_s^0 \\ b_s &= \mu^n + F_s^0 (\mu_s^x - \mu^n) \\ C_s &= F_s^0 \sigma_s^x F_s^0 + (F_s^0 - I)^T \sigma^n (F_s^0 - I) \end{aligned}$$

As in the other iterative VTS algorithms, we use the expected value $E[r|y, s] = \hat{\mu_s^r}$ as a new expansion point for Eq. 13 and iterate.

After convergence, we compute an estimate of x from the parameters of the approximate model:

$$\begin{aligned} \hat{x} &= \sum_s E[x|y, s] p(s|y) \\ E[x|y, s] &\approx y - \ln(e^{\hat{\mu_s^r}} + 1) + \hat{\mu_s^r} \end{aligned}$$

Here, Eq. 6 has been used to map $E[r|y, s] = \hat{\mu_s^r}$ to $E[x|y, s]$. Since the transformation is non-linear, our estimate for $\hat{x}$ is not the optimal MMSE estimator.

Figure 1 shows an example of the true posterior under the ZVM, together with the approximate posterior. Note that the iterative VTS has converged to the mode of the true posterior of the model.

5. Results

The experiments presented here were conducted using the data, code, and training scripts provided within the original Aurora 2 task [6]. The task consists of recognizing strings of English digits embedded in a range of arti£cial noise conditions. The acoustic model used for recognition is the "clean" acoustic model trained with the standard Aurora 2 scripts on uncorrupted data. It contains eleven whole word models, plus `sil` and `sp`, with a total of 546 diagonal 39 dimensional Gaussian mixture components. To conform with our observation models, the feature generation was modi£ed slightly from the reference "FE V2.0" implementation. In particular, we replaced the log

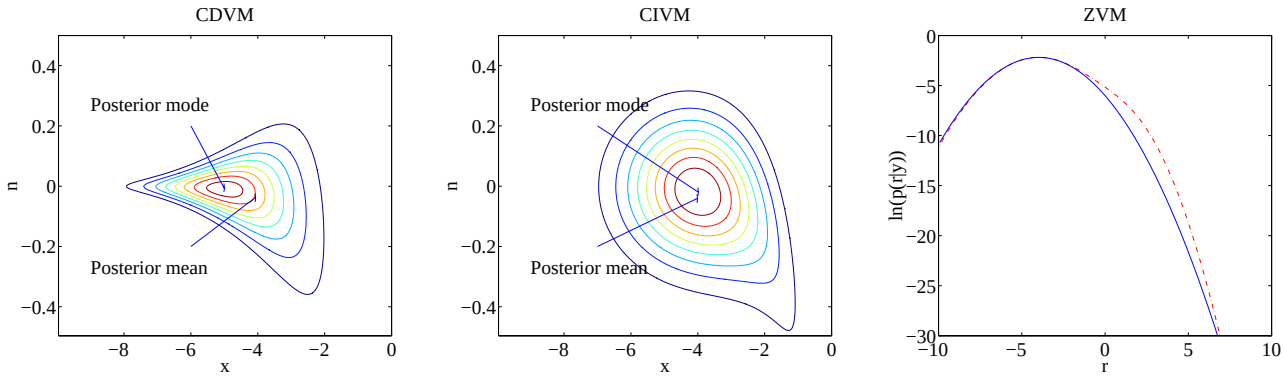

Figure 1: Exact posteriors under the three observation models. Parameters are $\mu^x = -4$, $\mu^n = 0$, $\sigma^x = 2$, $\sigma^n = 1$, $y = 0$. The SDVM and SIVM plots show $p(x, n|y)$. The ZVM plot shows the true $p(r|y)$ (solid) together with the iterative VTS approximation (dashed).

energy feature with c_0, and changed from using spectral magnitude to using power spectral density as the input to the Mel-frequency £lter-bank.

The prior GMM for clean speech was trained from the data provided in the "clean1" subset of the multi-style training set. The number of components in the system is varied to explore system performance at different operating points.

For each utterance, we trained an utterance-speci£c noise model on the features that the noise would have produced in the absence of speech. This paper is concerned with comparing observation models, and does not address the issue of how to track or learn the parameters of the noise model. It has been previously demonstrated how the noise parameters can be learned from the current utterance [7], although a rough speech/non-speech detector may be enough to make this system practical.

Table 1: Average word accuracy using the iterative VTS algorithm. Task is Aurora 2, Set A, clean acoustic model. Baseline accuracy is 63.66%.

Components	4	16	64	256
SDVM, $\sigma_\alpha^2 = 0.15$	46.30	64.17	66.89	69.00
SIVM, $\psi^2 = 0.050$	82.70	85.73	86.68	87.12
SIVM, $\psi^2 = 0.025$	82.87	85.90	86.82	87.23
ZVM	83.21	86.29	87.04	87.38

Table 1 presents word accuracy results for the three systems. The poor results for the SDVM are mostly due to the discrepancy between the posterior mean and mode described above. Note that when ψ^2 approaches zero, the performance of the SIVM approaches that of the ZVM.

6. Summary

This paper quanti£es the differences between three different observation models within a uni£ed Bayesian framework.

The SDVM provides the best theoretical £t to the true observation model. A computationally intensive numerical integration solution exists, but when using the faster iterative VTS technique, the discrepancy between mean and mode of the posterior causes serious errors.

In contrast, the SIVM does not £t the true observation model globally. However, after choosing an appropriate ψ^2, it works quite well in practice with the iterative VTS technique.

The ZVM is faster and produces better noise-robust features than the SIVM. It is faster because the dimensionality of the inference problem is cut in half. It produces better features because the gains in modeling the extreme SNR regions well outweigh the loss in the $x \approx n$ region. Although one could theoretically get the same result from the SIVM by setting $\psi^2 = 0$, in practice the existing derivations can not handle this case due to the singularity. The new ZVM derivation presented in this paper successfully overcomes this problem.

7. References

[1] L. Deng, A. Acero, M. Plumpe, and X. D. Huang, "Large vocabulary speech recognition under adverse acoustic environments," in *Proc. 2000 ICSLP*, Beijing, China, October 2000, pp. 806–809.

[2] S. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, pp. 114–120, 1979.

[3] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave on automatic speaker identi£cation and veri£cation," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304–1312, 1974.

[4] B. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. 2001 Eurospeech*, Aalbork, Denmark, September 2001.

[5] J. Droppo, A. Acero, and L. Deng, "A nonlinear observation model for removing noise from corrupted speech log Mel-spectral energies," in *Proc. ICSLP*, Denver, CO, Sept. 2002, pp. 182–5.

[6] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condidions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, September 2000.

[7] T. Kristjansson, B. Frey, and L. Deng, "Joint estimation of noise and channel distortion in a generalized EM framework," in *Proc. ASRU 2001*, Madonna di Campiglio, Italy, Dec. 2001.