NON-LINEAR NOISE COMPENSATION FOR ROBUST SPEECH RECOGNITION USING GAUSS-NEWTON METHOD

Yong Zhao and Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, Georgia Institute of Technology

ABSTRACT
In this paper, we present the Gauss-Newton method as a unified approach to optimizing non-linear noise compensation models, such as vector Taylor series (VTS), data-driven parallel model combination (DPMC), and unscented transform (UT). We demonstrate that the commonly used approaches that iteratively approximate the noise parameters in an EM framework are variants of the Gauss-Newton method. Through the formulation of the Gauss-Newton method for estimating noise means and variances, the noise estimation problems are reduced to determining the Jacobians of the noisy speech distributions. For the sampling-based compensations, we present two methods, sample Jacobian average (SJA) and cross-covariance (XCOV), to evaluate the Jacobians. Experiments on the Aurora 2 database verify the efficacy of the Gauss-Newton method to three noise compensation models. 

Index Terms: Gauss-Newton method, non-linear compensation, robust speech recognition, vector Taylor series

1. INTRODUCTION
In the past two decades, non-linear compensation methods that specifically cope with the additive noise and convolutional distortion have achieved promising results in various robust speech recognition applications. These compensation approaches typically utilize a non-linear noise mismatch function, defined in the feature space, that characterizes the joint effects of additive and convolutional noise[1]. Due to the non-linearity of the mismatch function, approximations must be taken to estimate the noisy speech distributions. For the vector Taylor series (VTS) compensation, a first-order Taylor series approximation of the mismatch function is often used [2]. In sampling-based methods, like unscented transform (UT) [3] and data-driven parallel model combination (DPMC) [4], the noisy speech distribution is estimated by drawing samples from the clean speech and noise distributions.

Though these methods have been shown to be highly effective, the non-linearity of the compensation models induces remarkable complexities to the model optimization procedure. One popular approach to estimating the noise parameters is to directly differentiate the conventional expectation maximization (EM) auxiliary function and iteratively approximate the root of the resulting non-linear derivative function. The approach was originally proposed in [2] for estimating the means of both the noise and the channel in the VTS compensation. In [5], [6], [7], this approach was generalized to incorporate the adaption of dynamic features, the estimation of noise variances, and the noise estimation for UT.

In this paper, we demonstrate that the above noise estimation approaches are variants of the Gauss-Newton method for estimating noise means and variances, and thus their convergence rates are approximately quadratic. The unified formulation of the Gauss-Newton method reduces the noise estimation problems to determining the Jacobians of the noisy speech distributions with respect to the clean speech and noise distributions. For the sampling-based compensation, we present two methods, sample Jacobian average (SJA) and cross-covariance (XCOV), to evaluate the Jacobians. Experiments are performed on the Aurora 2 database to verify the efficacy of the Gauss-Newton method to three noise compensation models, VTS, UT, and DPMC.

2. NON-LINEAR NOISE COMPENSATION
Assuming that a clean speech signal is corrupted by both additive noise and convolutional distortion. In the mel-cepstral domain, the noisy speech observations can be characterized by the following mismatch function [1]

\[ y = x + h + C \log(1 + \exp(C^\top (n - x - h))) \equiv g(x, n, h) \] (1)

where \( C \) and \( C^\top \) are the discrete cosine transformation (DCT) matrix and its pseudo-inverse, and \( x, n, \) and \( h \) denote the static MFCC feature vectors of the clean speech, additive noise, channel distortion, and noisy speech, respectively.

Due to the non-linearity of the mismatch function, it is difficult to derive the distribution of the corrupted speech feature vector, even if the clean speech and noise are Gaussian distributed. One commonly used approximation is to assume that the resulting noisy speech is also distributed in Gaussian. Under this assumption, a number of compensation models can be derived, among which we review two such forms, VTS and sampling-based compensations.

2.1. Vector Taylor series compensation
Assuming that \( x \) and \( n \) are independent and Gaussian distributed as \( N(\mu_x, \Sigma_x) \) and \( N(\mu_n, \Sigma_n) \), respectively, and \( h = \mu_h \) is a constant, the first-order VTS approximation of the static noisy speech \( y \) may be expressed as

\[ y \approx y_{\mu(0)} + G_x^{(0)}(x - \mu_x) + G_n^{(0)}(n - \mu_n) \] (2)

where \( G_x^{(0)} \) and \( G_n^{(0)} \) are Jacobian matrices given by

\[ G_x^{(0)} = \frac{\partial y}{\partial x} \bigg|_{\mu(0)} = CG_x^{(0)}C^\top \] (3)

\[ G_n^{(0)} = \frac{\partial y}{\partial n} \bigg|_{\mu(0)} = I - G_x^{(0)} \] (4)

and \( \mu(0) \) denotes Taylor expanding around \( \mu_x, \mu_n, \) and \( \mu_h \). The Jacobian \( G_x^{(0)} \) in the log-spectral domain is a diagonal matrix whose diagonal entries are given by

\[ \frac{1}{1 + \exp(C^\top (\mu_x - \mu_n - \mu_h))} \].

The approximation of (2) yields the following distribution of \( y [2], [8] \)

\[ \mu_x^{\text{vts}} = y_{\mu(0)} = g(\mu_x, \mu_n, \mu_h) \] (5)

\[ \Sigma_x^{\text{vts}} = G_x^{(0)} \Sigma_x G_x^{(0)\top} + G_n^{(0)} \Sigma_n G_n^{(0)\top} \] (6)
where

\[ m \]

responding to the Gaussian component of the parameters \( \theta \),

given a clean acoustic HMM set and an estimate of the noise parameter, mismatch function (1), and then estimates the distribution of the noisy speech. Let \( y \) be the noisy speech observation corresponding to the \( m \)-th sample pair of \( x \) and \( n \). We have

\[
\mu_{x,m} = \frac{1}{M} \sum_{m=1}^{M} x^{(m)}
\]

\[
\Sigma_{x,m} = \frac{1}{M} \sum_{m=1}^{M} (y^{(m)} - \mu_{x,m})(y^{(m)} - \mu_{x,m})^T
\]

The transformation of the delta parameter takes a similar form.

2.2. Sampling-based compensation

Though the VTS compensation has been shown successful in many robust speech recognition applications, the choice of a linear approximation may undermine the full exploitation of the non-linear mismatch function. One approach to addressing this concern is to use the sampling-based compensation methods. The sampling-based method randomly draws samples from the clean speech and noise distributions, obtains the noisy speech observations through the governing mismatch function (1), and then estimates the distribution of the noisy speech. Let \( y^{(m)} \) be the noisy speech observation corresponding to the \( m \)-th sample pair of \( x^{(m)} \) and \( n^{(m)} \). We have

\[
\mu_{y,m} = \frac{1}{M} \sum_{m=1}^{M} y^{(m)}
\]

\[
\Sigma_{y,m} = \frac{1}{M} \sum_{m=1}^{M} (y^{(m)} - \mu_{y,m})(y^{(m)} - \mu_{y,m})^T
\]

An instance of the sampling-based compensation is DPMC [4], where the samples are drawn using the Monte Carlo sampling technique. The advantage of DPMC is that as the sampled observations increase, the noise compensated models approach the exact distribution asymptotically. However, DPMC is computationally prohibitive; normally, 25–1000 sample points need to be generated per Gaussian.

Another sampling approach is to use UT [3]. UT draws a minimal number of artificially-chosen samples, called sigma points, from the mismatch function to approximate the statistics of the transformed distribution. Let \( z \) denote a combined vector by joining the clean speech feature and noise feature of \( d \) dimensions, \( z = [x^T, n^T]^T \). The following 4d sigma points are chosen

\[
\sum_{m=1}^{M} z^{(m)} = \begin{cases} \mu_z + \sqrt{2\Sigma_m}, & \text{if } 1 \leq m \leq 2d \\ \mu_z - \sqrt{2\Sigma_m}, & \text{if } 2d < m \leq 4d \end{cases}
\]

where \( \sqrt{\Sigma_m} \) indicates the \( m \)-th column of the square root of \( \Sigma \).

3. GAUSS-NEWTON METHOD FOR NOISE ESTIMATION

Given a clean acoustic HMM set and an estimate of the noise parameters \( \hat{\theta} = \{ \mu_n, \mu_g, \Sigma_n, \Sigma_{\Delta n}, \Sigma_{\Delta g} \} \), applying the above compensation formulae to each Gaussian component of the acoustic models produces the corresponding noisy speech models. The transformed HMM set would better match the target noise environment and obtain an improved recognition performance for the noisy speech. Additionally, the noise parameters \( \hat{\theta} \) can be estimated over the given utterance under an EM framework [2].

For the static noise mean and variance, the following auxiliary function needs to be maximized

\[
Q(\hat{\theta} | \theta) = -\frac{1}{2} \sum_{t} \sum_{j,k} \gamma_{jk}(t) [\log |\Sigma_{y,jk}| + (y_t - \mu_{y,jk})^T \Sigma_{y,jk}^{-1} (y_t - \mu_{y,jk})]
\]

where \( \theta \) and \( \hat{\theta} \) are the existing and the new parameter set, respectively. \( \gamma_{jk}(t) \) denotes the posterior probability being in the \( k \)-th Gaussian component of the \( j \)-th state at time \( t \) given \( \theta \).

We may differentiate the \( Q \) function and solve for its zeros as a typical M-step in the EM algorithm. Unfortunately, the derivative of \( Q \) is a non-linear function of the noise parameter, and the maximization problem does not have a closed-form solution. One popular approach in the literature is to approximate the root of the derivative through iterative refinements [2], [9], [5], [6]. In this section, we demonstrate such approaches are variants of the Gauss-Newton method.

3.1. Estimating noise and channel means

We note that the \( Q \) function, if \( \Sigma_{y,jk} \) is fixed, say \( \hat{\Sigma}_{y,jk} = \Sigma_{y,jk} \), is in a form of weighted non-linear least squares for \( (\mu_n, \mu_g) \), where observations \( y_t \) are fitted with non-linear models \( \hat{\mu}_{y,jk} \). Thus the noise parameters can be estimated using the Gauss-Newton method, which can be derived from Newton’s method via an approximation [10]. The gradient and Hessian of the \( Q \) function with respect to \( \mu_n \) are as follows

\[
\frac{\partial Q}{\partial \mu_n} = \sum_{j,k \in \Omega} G^T \frac{\partial}{\partial \mu_n} \Sigma^{-1} \gamma_{jk} G_n \Sigma_{y,jk}^{-1} G_n + \frac{\partial^2 \mu_{y,jk}}{\partial \mu_n^2} \Sigma_{y,jk}^{-1} c_{y,jk}
\]

where \( \frac{\partial^2 \mu_{y,jk}}{\partial \mu_n^2} \), a 3-dimensional matrix, is not exact and used here for illustration. The sufficient statistics \( \gamma_{jk} \) and \( c_{y,jk} \), as well as \( S_{y,jk} \) (will be used later), are defined as

\[
\gamma_{jk} = \sum_t \gamma_{jk}(t)
\]

\[
c_{y,jk} = \sum_t \gamma_{jk}(t)(y_t - \mu_{y,jk})
\]

\[
S_{y,jk} = \sum_t \gamma_{jk}(t)(y_t - \mu_{y,jk})(y_t - \mu_{y,jk})^T
\]

Besides, \( G_{n,jk} \) and \( G_{h,jk} \) denote the Jacobian matrices of \( \mu_n \) with respect to \( \mu_n \) and \( \mu_h \) for Gaussian component \( (j, k) \), respectively

\[
G_n = \frac{\partial \mu_n}{\partial \mu_n}, \quad G_h = G_n = \frac{\partial \mu_n}{\partial \mu_n}
\]

Note that they are in definition different from \( G^{(0)}_{n,jk} \) in (3) and \( G^{(0)}_{n} \) in (4), though being identical in the special case of the VTS compensation.

The Gauss-Newton method is formed by ignoring the second term in the Hessian matrix (14), thus leading to

\[
\hat{\mu}_{n,jk} = \mu_n + \left[ \sum_{j,k \in \Omega} \gamma_{jk} G_n \Sigma_{y,jk}^{-1} G_n + \frac{\partial^2 \mu_{y,jk}}{\partial \mu_n^2} \Sigma_{y,jk}^{-1} c_{y,jk} \right]^{-1} \sum_{j,k \in \Omega} G^T \Sigma_{y,jk}^{-1} c_{y,jk}
\]

The estimation of the channel mean \( \mu_h \) is similar and omitted here.

The Hessian approximation can be reliable when the first term in (14) dominates over the second term. This condition benefits our scenario for noise compensation. When we begin from well-initialized noise parameters, the residual \( e_{y,jk} \), and thus the second term, becomes sufficiently small.

The Gauss-Newton method can also be derived by linearly approximating \( \mu_{y,jk} \), which has been described in [2], [5]. However, the derivation herein shows one remarkable advantage of the Gauss-Newton method that it saves the calculation of the second-order derivatives, whilst achieving an approximately quadratic convergence rate, similar to that of Newton’s method.
In [11], we conducted a comparative study between the Gauss-Newton method and another popular noise estimation approach that views the compensation models from a generative perspective, giving rise to an EM algorithm analogous to the ML estimation for factor analysis (FA). As the convergence rate of typical EM methods is linear, the FA-based method is inferior to the Gauss-Newton method in efficiency. [11] also experimentally verified this convergence discrepancy.

3.2. Estimating noise variances

In the literature, several methods have been proposed to estimate the noise variances, such as the gradient descent method in [9], and Newton’s method in [5]. In [6], we presented an approach to recursively approximate the noise variances, promising better performance and less computational complexity. Here, we show the optimization procedure conforms to the Gauss-Newton principle.

For clarity, we use a 1-dimensional case to sketch the approximation of the Hessians of the compensation models introduced in Section 2, the main question left is to find the Jacobian matrices of the compensation models, $G_x$ and $G_n$. For the VTS compensation, it is straightforward

$$ G_{vts}^x = G_x^{(0)}; \quad G_{vts}^n = G_n^{(0)} $$

For the sampling-based compensation, directly determining the Jacobians is problematic, since the noisy speech mean is not a simple closed-form function of the noise means. To tackle this problem, we first relate the Jacobians of the compensation models to the expected value of the Jacobians of the mismatch function, and then describe two alternatives to evaluate $G_x$ and $G_n$.

Consider the random vector $z = [x^T, n^T]^T$ depending on its distribution parameters as $z = \sqrt{\Sigma_z} z + \mu_z$, where $\hat{z}$ is distributed as $\mathcal{N}(0, I)$. This mapping procedure separates out the effect of the distribution parameters from the Gaussian randomness, as $\mu_z$ and $\Sigma_z$ are independent of $z$. Thus it follows from the chain rule that

$$ \frac{\partial \mu_z}{\partial \mu_z} = E \left[ \frac{\partial y}{\partial \mu_z} \right] = E \left[ \frac{\partial y}{\partial z} \frac{\partial z}{\partial \mu_z} \right] = E \left[ \frac{\partial y}{\partial \mu_z} \right] $$

(26)

The identity (26) produces the expression suitable for the numerical evaluations of the Jacobians, which are described in the following.

4.1. Sample Jacobian average (SJA)

Clearly, we can evaluate $E \left[ \frac{\partial y}{\partial \mu_z} \right]$ by replacing the operation of expectation with the sample mean. In [7], this approach has been used for the noise estimation in the UT compensation model. We have

$$ G_{sja}^x = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial y}{\partial x} \mid_{x^{(m)}, \mu_z^{(m)}} = \frac{1}{M} C \left[ \sum_{m=1}^{M} G_x^{(m)} \right] C^T $$

(27)

$$ G_{sja}^n = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial y}{\partial n} \mid_{x^{(m)}, \mu_z^{(m)}} = I - G_{sja}^x $$

(28)

The second equality in (27) follows from (3) and uses the fact that $G_x^{(m)}$ depends linearly on $G_x^{(m)}$. Since $G_x^{(m)}$ is diagonal, the degree of freedom in evaluating $G_x^{(m)}$ is reduced to the size of the filter bank.

4.2. Cross-covariance method (XCOV)

There is an underlying assumption in the SJA method that the mismatch function characterizing the corrupted speech observations is in a closed-form. However, in some instances, the dependence of the mismatch function on its control parameters is so complicated that it is difficult to reliably evaluate the Jacobians of the mismatch function. An alternative to evaluating the Jacobians of the compensation models is in the form

$$ G_{x}^{xcoev} = \Sigma_{yx} \Sigma_x^{-1}; \quad G_{n}^{xcoev} = \Sigma_{yn} \Sigma_n^{-1} $$

(29)

The above equation comes from the following theorem:

**Theorem 1** Assume $x$ to be a multivariate Gaussian random variable distributed as $p(x) = \mathcal{N}(\mu_x, \Sigma_x)$, and $y = g(x)$ is function of $x$. If $p(x) \mid g(x) \rightarrow 0$, as $|x| \rightarrow \infty$, we have

$$ E \left[ \frac{\partial y}{\partial x} \right] = \Sigma_{yx} \Sigma_x^{-1} $$

(30)

The theorem is related to Bussgang’s theorem [12], and can be derived using integration by parts. The evaluation formulae (29) transform the computation of the Jacobians to the determination of the cross-covariances, which can be calculated, say $\Sigma_{yx}$, as

$$ \Sigma_{yx} = \frac{1}{M} \sum_{m=1}^{M} (y^{(m)} - \mu_y^{(m)}) (x^{(m)} - \mu_x^{(m)})^T $$

(31)

The advantage of the XCOV method is that it does not require any explicit gradient information. Interestingly, we observe that in the whole noise estimation procedure, we need four quantities for each Gaussian component to completely characterize the mismatch function, the mean $\mu_x$, the variance $\Sigma_x$, and the cross-covariances $\Sigma_{yx}$ and $\Sigma_{yn}$. This treatment may facilitate the study of more complicated distortion models.
The proposed noise estimation algorithms were evaluated on the Aurora 2 database [13]. The test set consists of three different parts. Test Set A and Test Set B each contain 4 types of additive noises, and the data in Test Set C are contaminated with 2 types of additive noises as well as channel distortion. The training set was used to estimate the baseline ML HMMs. The acoustic models were trained using the standard Aurora 2 recipe for the simple back end. Each feature frame is characterized by 39-dimensional MFCCs with the 0-th cepstral coefficient for the energy term. The cepstra are computed based on spectral magnitude. The baseline ML system yields the 0-th cepstral coefficient for the energy term. The cepstra are computed based on spectral magnitude. The baseline ML system yields word error rate (WER) of 41.57% by averaging over SNRs between 20 and 0 dB of three test sets.

For each utterance, we conducted noise compensation in two decoding passes. The first pass uses the noise parameters initialized with the first and last 20 frames, and the channel mean set to 0. The second pass refines the noise estimate using the Gauss-Newton methods and re-decodes the utterance.

Three compensation models, VTS, UT, and DPMC, are compared in Table 1. In the DPMC compensation, 100 Monte Carlo samples per Gaussian are generated for SJA, and 800 for XCOV.

| TABLE 1 | WER (%) of different noise compensation models using the Gauss-Newton method. |
|---------|-------------------------------|------------------------------|-------------------|
|         | VTS  | UT  | DPMC  |
|         | SJA  | XCOV | SJA  | XCOV  |
| 1 pass  | 12.86| 12.60| 12.49| 13.94| 14.15|
| 2 passes| 8.35 | 8.22 | 8.35 | 8.43  | 8.53 |

All of the compensation systems achieve significant improvements over the baseline ML system. The one-pass results, where noise parameters have not been treated with the Gauss-Newton optimization, may reflect the intrinsic efficacy of the compensation models. From this perspective, we can say UT has advantage over VTS and DPMC in terms of the modeling accuracy. All of the compensation models produce a similar performance after two passes, though UT-SJA is slightly better than other systems. This implies that the noise estimation algorithm substantially diminishes the difference of the modeling powers in these compensation models.

Fig. 1 plots the performance variations with respect to the number of Monte Carlo samples per Gaussian in the DPMC compensation. It is observed that the two Jacobian evaluation approaches achieve a similar performance given sufficient samples, whereas DPMC-SJA converges significantly faster than DPMC-XCOV. This can be attributed to the lower degree of freedom used in the SJA method than the XCOV method, as discussed in Section 4.

6. CONCLUSION

In this paper, we have shown the Gauss-Newton method as a unified approach to optimizing various non-linear noise compensation models. The formulation of the Gauss-Newton method reduces the noise estimation problems to determining the Jacobians of the noisy speech distributions with respect to the clean speech and noise distributions. We presented two methods, SJA and XCOV, to evaluate such Jacobians for the sampling-based compensations. From the perspective of XCOV, we showed that in the noise estimation procedure, the nonlinear compensation model can be completely characterized by four statistics for each Gaussian component, the mean $\mu_y$, the variance $\Sigma_y$, and the cross-covariances $\Sigma_{yx}$ and $\Sigma_{xy}$. As such, the proposed noise estimation method can be generalized to allow for more complicated compensation models. Experimental results on the Aurora 2 database verified the efficacy of the Gauss-Newton method to three noise compensation models, VTS, UT, and DPMC.

7. REFERENCES