ON NOISE ESTIMATION FOR ROBUST SPEECH RECOGNITION USING VECTOR TAYLOR SERIES

Yong Zhao and Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, Georgia Institute of Technology

ABSTRACT

In this paper, we propose a novel noise variance estimation method using the fixed point method for the VTS-based robust speech recognition. Noise parameters are re-estimated over a given utterance using an EM algorithm. The derivative of the auxiliary function with respect to the noise variance is resolved, and the fixed point algorithm estimates the noise variance by recursively approximating the root of the resulting derivative. The method leads to a re-estimation formula with a flavor like the standard ML variance estimation, and the iteration procedure is step-size free. We also investigate improving the noise estimation for efficient VTS adaptation. Several fast noise estimation methods are examined including estimation from non-speech areas and incremental adaptation. In the evaluation over Aurora 2 database, the proposed noise variance estimation method obtains a significant improvement in recognition accuracy over the method using sample variance. Further experiments show that the VTS ML estimation over non-speech areas is an effective fast adaptation method. The final refined approach achieves 8.75% WER, 13% relative improvement over the conventional VTS adaptation.

Index Terms— Robust speech recognition, vector Taylor series, noise estimation

1. INTRODUCTION

The goal of noise robust speech recognition is to maintain satisfactory recognition accuracy under mismatched operating conditions. In the past several years, vector Taylor series (VTS), which provides a linear approximation to relate the noisy speech and its clean counterpart, has been shown successful in various robust speech recognition methods. In [1], VTS has been proposed for feature compensation. The noisy speech and the clean speech are modeled by a joint Gaussian distribution through the VTS expansion, and the noisy features are compensated by computing a minimum mean square error (MMSE) estimate of the clean signal. Recognition then takes places in the clean environment. On the other hand, VTS can be used in model adaptation approaches [2]. The clean acoustic models are modified through VTS to match the noisy environment. The recognizer then decodes the given noisy speech based on the derived noisy acoustic models.

The success of the VTS approximation relies on the accurate estimation of noise parameters. Moreno [1] provided an expectation maximization (EM) framework for the estimation of additive and channel noise means. Estimating noise variance in an EM fashion is difficult. In [3], Liao proposed a gradient-descent method to obtain the noise variance estimate. The main drawback of this method is that it does not guarantee the gradient-based update would increase the auxiliary function, and thus a backing off method is used. In [4], the Newton's method has been proposed to optimize the auxiliary function. However, the second order derivative of the auxiliary function leads to a complicated computation. In theory, both methods need to examine the step size to ensure convergence.

In this paper, we present a fixed point approach to recursively approximate the noise variance. The method leads to a re-estimation formula with a flavor like the standard maximal likelihood (ML) estimation for the speech variance.

The paper also explores improving the noise estimation for efficient VTS adaptation. One of the solutions is to have a properly-initialized noise estimate to achieve computational efficiency as well as recognition accuracy. We analyze two schemes for noise parameter initialization, estimating from non-speech portions and incremental adaptation.

The remaining of the paper is organized as follows: Section 2 gives an overview of the VTS adaptation algorithm. Section 3 introduces our proposed approach to estimate the noise variance. Section 4 analyzes the VTS adaptation procedure and presents different techniques for improvement. Experiments and results are reported in Section 5.

2. VECTOR TAYLOR SERIES ADAPTATION

Assuming that a clean speech signal $x(t)$ is corrupted by both additive noise $n(t)$ and convolutional distortion $h(t)$, the resulting noisy speech $y(t)$ can be expressed as:

$$y(t) = x(t) * h(t) + n(t) \tag{1}$$

In the mel-cepstral domain, the noisy speech can be established as a nonlinear function of its clean counterpart [1] [2],

$$y = x + h + C\ln(1+\exp(C^{-1}(n-x-h))) \equiv x + g(x,n,h) \tag{2}$$

where C is the discrete cosine transformation matrix, and x, n, h, and y denote the static feature vectors of the clean speech, additive noise, channel distortion, and noisy speech, respectively.

Assume that x, n, and h are Gaussian with means μ_x, μ_n, and μ_h and covariance matrices Σ_x, Σ_n, and Σ_h, respectively. the mean and variance of y can be estimated by the first-order VTS expansion around the means of x, n, and h.

$$\mu_y = \mu_x + g(\mu_x, \mu_n, \mu_h) \tag{3}$$

$$\Sigma_y = G\Sigma_x G^T + (I-G)\Sigma_n (I-G)^T \tag{4}$$

where the Jacobian matrix G is given by

$$G = C diag\left(\frac{1}{1+exp(C^{-1}(\mu_n - \mu_x - \mu_h))}\right) C^{-1} \tag{5}$$

Conventionally, if Σ_x and Σ_n are diagonal, Σ_y is also forced to be diagonal to keep the same functional form [2].

For the delta portion of the MFCC features, the following adaptation formulas have been used:

$$\mu_{\Delta y} = G\mu_{\Delta x} \tag{6}$$

$$\Sigma_{\Delta y} = G\Sigma_{\Delta x} G^T + (I-G)\Sigma_{\Delta n}(I-G)^T \tag{7}$$

The delta/delta portion of the MFCC features takes a similar form.

Given a clean acoustic HMM set and an initialization of the noise parameters, applying the above adaptation formulas to each Gaussian component of the models produces the corresponding noisy speech models. The generated HMM set matches the testing noise environment and obtains an improved performance against the noisy speech. On the other hand, the noise parameters may be re-estimated over the given utterance using an EM algorithm, as described in [1] [5] [3]. Here we give the noise re-estimation formulas for μ_n and μ_h as a reference,

$$\mu_n = \mu_n^0 + \Big[\sum_j \sum_k \gamma_{jk}(I - G_{jk})^T \Sigma_{y,jk}^{-1}(I - G_{jk})\Big]^{-1} \sum_j \sum_k (I - G_{jk})^T \Sigma_{y,jk}^{-1} c_{y,jk} \quad (8)$$

$$\mu_h = \mu_h^0 + \Big[\sum_j \sum_k \gamma_{jk} G_{jk}^T \Sigma_{y,jk}^{-1} G_{jk}\Big]^{-1} \sum_j \sum_k G_{jk}^T \Sigma_{y,jk}^{-1} c_{y,jk} \quad (9)$$

where we define the following sufficient statistics for the k-th Gaussian in the j-th state, with mean vector $\mu_{y,jk}$ and covariance matrix $\Sigma_{y,jk}$, of the noisy speech HMM set

$$\gamma_{jk} = \sum_t \gamma_{jk}(t) \quad (10)$$

$$c_{y,jk} = \sum_t \gamma_{jk}(t)(y_t - \mu_{y,jk}) \quad (11)$$

Note that in the derivation of formulas (8) and (9), the fixed point principle is implied [1], i.e., in each iteration, the Jacobian matrix G_{jk} is considered constant although it is a function of the noise parameters. In the next section, we will use the same principle to estimate the noise variance.

3. NOISE VARIANCE ESTIMATION USING THE FIXED POINT METHOD

It is difficult to estimate the additive noise variance by maximizing the auxiliary Q function since the model variances, which are affine functions of the noise variance, appear in the form of determinant and matrix inverse in the auxiliary function. In this paper, we present a fixed point approach to recursively approximate the maximum point of the Q function. The derivative of the Q function with respect to the noise variance is linearized by computing its denominator-like parts using the previous estimate of the noise variance, and then the noise variance is refined by solving the resulting linear equation.

Take the derivative of the Q function with respect to the static noise variance,

$$\frac{\partial Q}{\partial \Sigma_n} = -\frac{1}{2}\sum_t \sum_j \sum_k \gamma_{jk}(t) \frac{\partial}{\partial \Sigma_n}\Big[log|\Sigma_{y,jk}| + (y_t - \mu_{y,jk})^T \Sigma_{y,jk}^{-1}(y_t - \mu_{y,jk})\Big] + const \quad (12)$$

Observing $\Sigma_{y,jk}$ depends on Σ_n as Eq. (4), there are two kinds of terms being differentiated in the above function, the normalizing determinant, which depends on Σ_n in a form as $log|C + AXB|$, and the main probability term, depending on Σ_n in a form as $a^T(C + AXB)^{-1}b$, where we denote Σ_n by X, $G\Sigma_{x,jk}G^T$ by C, $(I - G)$ by A, $(I - G)^T$ by B, and $(y_t - \mu_{y,jk})$ by a and b, respectively. By applying the following identities for the derivatives,

$$\frac{\partial}{\partial X}\log|C + AXB| = A^T(C + AXB)^{-T}B^T \quad (13)$$

$$\frac{\partial}{\partial X}a^T(C + AXB)^{-1}b = -A^T(C + AXB)^{-T}ab^T(C + AXB)^{-T}B^T \quad (14)$$

the derivative function (12) transforms to

$$\frac{\partial Q}{\partial \Sigma_n} = -\frac{1}{2}\sum_j \sum_k (I - G_{jk})^T \Sigma_{y,jk}^{-1}(\gamma_{jk}\Sigma_{y,jk} - S_{y,jk}) \Sigma_{y,jk}^{-1}(I - G_{jk}) \quad (15)$$

where we define the sufficient statistic

$$S_{y,jk} = \sum_t \gamma_{jk}(t)(y_t - \mu_{y,jk})(y_t - \mu_{y,jk})^T \quad (16)$$

The resulting derivative function (15) is a nonlinear function of the noise variance. In an approach described in [3], the derivative is exploited by using a gradient descent method to maximize the Q function. An alternative approach is to equate the derivative to 0 and find the root as the maximum point of the Q function; however, the obstacle is that the derivative function, due to its non-linearity, has no closed form solution. Here we present the fixed point method to recursively approximate the root of the derivative function.

The proposed method takes advantage of the structure of the derivative function. First of all, the function (15) is a sum over the homogeneous terms that correspond to each pair of state and mixture component. Each summand can be regarded as a rational function of the noise variance, where the numerator (the central item of each summand) is first-order and the denominator (the second and fourth items) is a square of the noise variance plus the clean speech model variance. If we equate to 0 only one rational summand of the derivative, the denominator vanishes and leaves behind $(\gamma_{jk}\Sigma_{y,jk} - S_{y,jk}) = 0$, which can be readily solved.

The solution to one Gaussian component case motivates a fixed point approach. In the scheme, we compute the denominator parts using the previous estimate of the noise variance Σ_n^0, and obtain a linear equation to update the noise variance estimate. Substituting Eq. (4) for $\Sigma_{y,jk}$ in the numerator terms of Eq. (15) and moving the constant terms to the right-hand side of the equation, we have

$$\sum_j \sum_k \gamma_{jk} A_{jk} \Sigma_n A_{jk}^T = \sum_j \sum_k B_{jk} \quad (17)$$

where the following notations are defined

$$A_{jk} = (I - G_{jk})^T \Sigma_{y,jk}^{0^{-1}}(I - G_{jk}) \quad (18)$$

$$B_{jk} = (I - G_{jk})^T \Sigma_{y,jk}^{0^{-1}}\big(S_{y,jk} - \gamma_{jk} G_{jk}\Sigma_{x,jk}G_{jk}^T\big) \Sigma_{y,jk}^{0^{-1}}(I - G_{jk}) \quad (19)$$

where superscript ^{0} marked on $\Sigma_{y,jk}$ denotes the dependence on the previous estimate of the noise variance Σ_n^0, and reflects the recursive nature of the optimization procedure. The estimation procedure may repeat multiple rounds until some convergence criterion has been reached, i.e., we update A_{jk} and B_{jk} by substituting the noise variance estimate of the m-th iteration into Eqs. (18) and (19), and obtain a new variance estimate for the $(m + 1)$-th iteration by resolving Eq. (17). The iterations of variance estimation may proceed simultaneously with the iterations of mean estimation as Eqs. (8) and (9).

The above equations give a general form for estimating the noise variances, even in the form of full covariance matrices. For the esti-

mation of the dynamic noise variance, one needs to replace the static parameters with the corresponding dynamic parts.

The noise variance update formula to some extent reflects the physical meaning of the noise variance. $\left(S_{y,jk} - \gamma_{jk} G_{jk} \Sigma_{x,jk} G_{jk}^T\right)$ in B_{jk} can be regarded as the residual between the sample variance of the noisy speech and the transformed variance of the clean speech model. Then the weighted sum of the residual variances over all Gaussian mixtures turns into an estimate of the noise variance.

The computational cost of the fixed point noise estimation can be greatly reduced if covariance matrices Σ_n, Σ_x, and Σ_y are assumed diagonal. Since Σ_n, the variable to be differentiated, is diagonal, the derivative function (15) is diagonal as well, which means that Eq. (17) effectively holds for its diagonal positions, other positions being 0.

In an approximation, we further assume that a matrix product in a form like GDG^T, where D denotes a diagonal matrix, can be diagonalized without much loss of accuracy. This is an assumption consistent with the diagonalization of Σ_y. Then A_{jk} is diagonal, and the central term of B_{jk} is diagonal, which again leads to B_{jk} being diagonal. Thus, Eq. (17) can be decomposed into a set of equations on separate dimensions, and the noise variance at the i-th dimension is estimated by

$$(\Sigma_n)_{ii} = \Big[\sum_j \sum_k \gamma_{jk} (A_{jk})_{ii}^2 \Big]^{-1} \sum_j \sum_k (B_{jk})_{ii} \tag{20}$$

where $(.)_{ii}$ denotes the i-th diagonal component of a matrix.

Another advantage of the diagonalization procedure is that the computation of the matrix product GDG^T can be reduced to a multiplication between a matrix and a vector. The i-th diagonal component of GDG^T is

$$(GDG^T)_{ii} = \sum_j (G)_{ij}^2 (D)_{jj} \tag{21}$$

Since the denominator of the noise variance estimate in (20) is positive, a negative-valued weighted sum $\sum_j \sum_k (B_{jk})_{ii}$ would lead to a negative noise variance. This occasionally happens when the noisy speech has a high SNR level. Under this condition, variance flooring to the noise model is necessary.

4. IMPROVED VTS ADAPTATION SCHEMES

The standard VTS adaptation algorithm is an unsupervised utterance-by-utterance procedure. The algorithm is summarized as follows according to [5]:

1. For each utterance, initialize the additive noise parameters using the start and end frames, and set the channel mean vector to 0.
2. Transform the clean acoustic models using the adaptation formulas and decode the utterance.
3. Refine the noise estimate by maximizing the likelihood with respect to the assumed hypothesis.
4. Transform the models again, decode the utterance to obtain a final recognition transcription.

The steps described above include two pass decoding and one iteration of EM re-estimation. If the best performance was desired, a multiple-iteration EM of Step 3 and a multi-pass decoding loop between Steps 3 and 4 may be necessary.

For a large vocabulary system, the main computational cost lies on the recognition pass and the transformation of the model set. The accumulation of the sufficient statistics (Eqs. (10), (11), and (16)), and the multiple-iteration EM re-estimation of parameters (Eqs. (8), (9), and (20)), are computationally less expensive since they can be optimized by restraining the operations on those seen models in the assumed hypothesis. Hence, for each utterance, the complexity of the VTS adaptation is roughly proportional to the number of decoding passes (one recognition along with one transformation of the full model set), less relevant to the number of re-estimation iterations.

Though the algorithm has achieved promising results for robust speech recognition, its prohibitive computational load prevents its application in practice. One of the solutions is to have a properly-initialized noise estimate to achieve computational efficiency as well as recognition accuracy. Here we discuss two schemes for noise parameter initialization, estimating from non-speech areas and inheriting the noise estimate from the previous utterance.

4.1. Estimating over non-speech segments

Sample average over non-speech frames (SA-NS) can be used as an estimate of the noise mean and variance. A more elaborate variant is to apply the same VTS noise estimation process on non-speech areas (VTS-NS) of an utterance. If the silence of clean speech is modeled by one Gaussian component (this model may be separate from the HMM set), the estimation formulas (8) and (17) simplify to

$$\mu_n = \mu_n^0 + \frac{1}{N} (I - G_{sil})^{-1} c_{y,sil} \tag{22}$$

$$\Sigma_n = (I - G_{sil})^{-1} \left(\frac{S_{y,sil}}{N} - G_{sil} \Sigma_{x,sil} G_{sil}^T \right) (I - G_{sil})^{-T} \tag{23}$$

where parameters being labeled with subscript $_{sil}$ denote they come from the silence model. The sufficient statistics $c_{y,sil}$ and $S_{y,sil}$ are accumulated over N frames as (11) and (16) with the posterior probability $\gamma_{jk}(t)$ set to 1.

If used stand-alone (Steps 1 and 2 only), the VTS-NS estimation can be regarded as a fast adaptation scheme. The scheme in principle is analogous to the Jacobian approach described in [6]. In both schemes, the difference between the reference and observed noise cepstra is exploited to predict the compensation of the whole model set.

4.2. Incremental adaptation

The initial noise parameters can choose the estimate from the previous utterance. In [7], this incremental adaptation scheme (VTS-INC) was examined in combination with various noise compensation approaches. Provided the noise statistic changes slowly, incremental adaptation can approximate the standard VTS adaptation while saving a significant amount of computation.

5. EXPERIMENTS AND RESULTS

The proposed noise estimation methods were evaluated on the Aurora 2 database [8] of connected digits. The clean training set is used to estimate the baseline ML HMMs. 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. For each noise type, a subset of the clean speech utterances is contaminated at SNRs ranging from 20 to -5 dB at a 5 dB step size, which, including the clean condition, constitute 7 different SNR levels.

The acoustic models were trained using the standard Aurora 2 recipe for the simple back end. 39-dimensional MFCC features with the 0th cepstral coefficient for the energy term are used in the experiments. The cepstra are computed based on spectral magnitude. The first and last 20 frames of each utterance, assumed from the non-speech area, are used for initializing the noise means and variances. 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. The baseline VTS system follows the noise re-estimation procedure described in [5], which is also summarized in Section 4, i.e., the noise and channel mean are re-estimated according to Eqs. (8) and (9), while the noise variance is estimated using sample variance over non-speech frames.

We compare the performance of the proposed noise variance estimation method and that of the estimation using sample variance. As shown in Table 1, the second column is the baseline VTS system, and the third column uses the fixed point approach to estimate noise variance. As different portions of HMM variance parameters are gradually added for the adaptation, the fixed point method consistently reduces WER from 17.88% to 8.95%. Compared with the best result (10.03%) of the sample variance method, the fixed point method yields 11% relative improvement, demonstrating the efficacy of the proposed method.

Table 1. WER(%) for the baseline and proposed variance estimation methods.

	Sample variance	Fixed point
Mean	17.88	
+ Static variance	13.34	13.54
+ Δ variance	10.03	9.86
+ Δ^2 variance	10.10	8.95

We next examine the optimal configuration for decoding passes and re-estimation iterations. As shown in Fig. 1, using initial noise estimate by SA-NS for the first pass recognition produces 12.86% WER. The standard VTS adaptation procedure with 2 decoding passes and 1 re-estimation iteration per (*#dec=2, #res=1*) reduces WER to 9.57%. The performance is maximized by adding another iteration (*#dec=2, #res=2*) or another decoding pass (*#dec=3, #res=1*). Considering the computational cost, discussed in 4, the former configuration is believed a better choice and chosen as the default setting through our experiments. It is also observed that with the increase of decoding passes and re-estimation iterations, the performance does not improve as well. That is possibly due to the unsupervised nature of the VTS adaptation scheme.

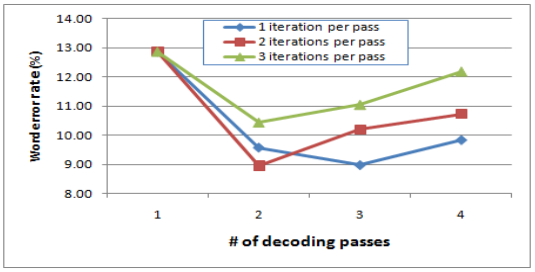

Fig. 1. WER (%) for the VTS adaptation algorithm as a function of re-estimation iterations and decoding passes.

Table 2 compares the performance of several VTS adaptation schemes discussed in Section 4. The second column runs one pass decoding for each utterance, which represents a fast adaptation scheme for SA-NS and VTS-NS, and a standard incremental adaptation of VTS-INC. The third column shows their best performances after fully estimated (*#dec=2, #res=2*). For VTS-INC, we reset the noise estimate whenever speaker or noise environment changes. We also disable the update of channel mean for VTS-INC as it has been found that the channel estimation is not stable for our incremental adaptation setup, greatly degrading the performance.

In one pass case, VTS-NS significantly outperforms the other two methods. Interestingly, it achieves 19% relative improvement over SA-NS though their implementations differ modestly. After fully estimated, VTS-NS produces the lowest WER of 8.75%, 13% relative improvement over the baseline system (10.03%).

The incremental adaptation presents a mediocre performance in one pass case and improves slightly after fully estimated. We attribute this to two reasons. Fist, under low-SNR cases, estimation over non-speech areas may provide a more robust estimate than does VTS-INC, which risks a poor state/frame alignment. Second, full estimation make the adapted model overfitting to the current utterance and introduces a mismatch with the next utterance.

Table 2. WER (%) for various noise initialization and estimation schemes.

Initial estimate	One pass	Fully estimated
SA-NS	12.86	8.95
VTS-NS	10.40	8.75
VTS-INC	12.62	11.71

6. CONCLUSION

In this paper, we propose a noise variance estimation algorithm using the fixed point method for the VTS adaptation. The algorithm gives a general form for estimating the noise variance, even in full covariance matrix. The re-estimation formula reflects the physical meaning of the noise variance, and the iteration procedure is step-size free. We also investigate the tradeoff between the recognition accuracy and the computational cost for the VTS adaptation. Several fast noise estimation schemes are described to remedy the computation requirement of the VTS adaptation. It has been found that VTS-NS can act as an effective method for fast adaptation. The VTS adaptation with VTS-NS as initial noise estimation achieves the lowest WER of 8.75%.

7. ACKNOWLEDGEMENT

We would like to thank Dr. Jinyu Li at Microsoft for valuable discussions.

8. REFERENCES

[1] P. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[2] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *ICSLP*, Beijing, China, 2000.

[3] H. Liao, *Uncertainty Decoding for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, 2007.

[4] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech and Language*, vol. 23, pp. 389–405, 2009.

[5] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series," in *IEEE ASRU*, 2007.

[6] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashit, "Jacobian approach to fast acoustic model adaptation," in *ICASSP*, Munich, Germany, 1997.

[7] F. Flego and M.J.F. Gales, "Incremental predictive and adaptive noise compensation," in *ICASSP*, Taipei, Taiwan, 2009.

[8] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR*, 2000.