JOINT ESTIMATION OF NOISE AND CHANNEL DISTORTION IN A GENERALIZED EM FRAMEWORK

T. Kristjansson†, B. Frey‡

University of Waterloo†
University of Toronto‡
{ttkristj,frey}@uwaterloo.ca

L. Deng, A. Acero

Microsoft Research
Redmond, WA
{deng,alaxac}@microsoft.com

ABSTRACT

The performance of speech cleaning and noise adaptation algorithms is heavily dependent on the quality of the noise and channel models. Various strategies have been proposed in the literature for adapting to the current noise and channel conditions. In this paper, we describe the joint learning of noise and channel distortion in a novel framework called ALGONQUIN. The learning algorithm employs a generalized EM strategy wherein the E step is approximate. We discuss the characteristics of the new algorithm, with a focus on convergence rates and parameter initialization. We show that the learning algorithm can successfully disentangle the non-linear effects of noise and linear effects of the channel and achieve a relative reduction in WER of 21.8% over the non-adaptive algorithm.

1. INTRODUCTION

It is well known that recognition rates of speech recognition systems suffer considerably when there is a mismatch between training and deployment conditions. One approach to dealing with this mismatch is to restore or clean the noisy features such that they resemble those of the training environment. Methods that fall into this category are Spectral Subtraction (SS)[2], Cepstral mean normalization (CMN) and Algonquin[3], to name a few.

The performance of a feature cleaning method is greatly dependent on how well the noise and channel distortion are estimated and modeled. For example, it is possible to use point estimates or single or multiple mixture gaussian distributions[4]. In general, for methods that employ noise and channel models of some sort, the correct estimation of the model parameters is crucial.

Estimation of the noise and channel model parameters is complicated by the fact that the observations contain a combination of speech, noise and channel distortion. Various ad-hoc methods are used to deal with this problem such as using low powered frames to estimate the parameters of the noise model, and high powered frames to estimate the parameters of the channel model.

In this paper, we discuss a principled method for jointly learning the parameters of the noise and channel models. The method is able to learn the noise and the channel model simultaneously, by employing an accurate speech model and a model for the combination of speech, noise and channel. In [5] we described the use of noise adaptation to enhance performance. In this paper we extend the framework to include learning of channel distortion.

In the first section we introduce the Algonquin framework and discuss the estimation of the posterior $p(\mathbf{x}, \mathbf{n}, \mathbf{h}|\mathbf{y})$. Then we discuss the learning of the noise and channel parameters within the Generalized EM framework. In the Analysis section we discuss convergence characteristics and pathological cases when run on synthetic data. In the Results section, we show that the algorithm performs well for real speech data.

2. THE ALGONQUIN FRAMEWORK

In the Algonquin framework[3], the MMSE estimate of the clean speech $\hat{\mathbf{x}}$ is estimated.

$$\hat{\mathbf{x}} = \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\mathbf{y}) \tag{1}$$

where

$$p(\mathbf{x}|\mathbf{y}) \propto \int_{\mathbf{n},\mathbf{h}} p(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h}) p(\mathbf{x}, \mathbf{n}, \mathbf{h}) \tag{2}$$

The crux of the method is the way in which the posterior $p(\mathbf{x}, \mathbf{n}, \mathbf{h}|\mathbf{y})$ is approximated by a simplified posterior function $q(\mathbf{x}, \mathbf{n}, \mathbf{h})$. As described below, a variational approach is used to find the parameters of the approximate joint distribution $q(\mathbf{x}, \mathbf{n}, \mathbf{h})$. This procedure constitutes the E step of the Generalized EM algorithm.

In the Fourier domain, the relationship between speech X, channel H, noise N and noisy observation Y is:

$$Y(f) = H(f)X(f) + N(f). \tag{3}$$

After taking the magnitude squared and logarithm we arrive at the equivalent equation in the log-spectrum domain:

$$\mathbf{y} \approx \mathbf{g}(\begin{bmatrix}\mathbf{x}\\\mathbf{n}\\\mathbf{h}\end{bmatrix}) = \mathbf{x} + \mathbf{h} + \ln(\mathbf{1} + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h})). \quad (4)$$

where $\ln()$ and $\exp()$ operate on the individual elements of their vector arguments.

Assuming the errors in the above approximation are Gaussian, the observation likelihood is

$$p(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = \mathcal{N}(\mathbf{y}; \mathbf{g}(\begin{bmatrix}\mathbf{x}\\\mathbf{n}\\\mathbf{h}\end{bmatrix}), \mathbf{\Psi}), \quad (5)$$

The ALGONQUIN framework employs Gaussian mixture models to model the speech, noise and channel impulse response in the log-spectrum domain, thus, the joint distribution over noisy speech $\mathbf{y}$, speech $\mathbf{x}$, speech class c^x, noise $\mathbf{n}$, noise class c^n, channel $\mathbf{h}$ and channel class c^h is:

$$\begin{aligned} p(\mathbf{y},\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h) &= \mathcal{N}(\mathbf{y}; \mathbf{g}(\begin{bmatrix}\mathbf{x}\\\mathbf{n}\\\mathbf{h}\end{bmatrix}), \mathbf{\Psi}) \\ &\cdot \pi^x_{c^x}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}^x_{c^x}, \mathbf{\Sigma}^x_{c^x}) \cdot \pi^n_{c^n}\mathcal{N}(\mathbf{n}; \boldsymbol{\mu}^n_{c^n}, \mathbf{\Sigma}^n_{c^n}) \\ &\cdot \pi^h_{c^h}\mathcal{N}(\mathbf{h}; \boldsymbol{\mu}^h_{c^h}, \mathbf{\Sigma}^h_{c^h}). \quad (6)\end{aligned}$$

For the current frame of noisy speech $\mathbf{y}$, ALGONQUIN approximates the posterior using a simpler, parameterized distribution, q:

$$p(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h|\mathbf{y}) \approx q(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h). \quad (7)$$

The "variational parameters" of q are adjusted to make this approximation accurate, and then q is used as a surrogate for the true posterior when computing $\hat{\mathbf{x}}$ and learning the noise and channel models. See [6] for a review of variational inference techniques.

The q function is a mixture of gaussians:

$$q(\mathbf{x},\mathbf{n},\mathbf{h}) = \sum_{\{c^x,c^n,c^h\}} \rho_{c^x c^n c^h} q(\mathbf{x},\mathbf{n},\mathbf{h}|c^x,c^n,c^h) \quad (8)$$

where $\rho_{c^x c^n c^h}$ are the mixture weights. The form of each component $q(\mathbf{x},\mathbf{n},\mathbf{h}|c^x,c^n,c^h)$ is:

$$q(\mathbf{x},\mathbf{n},\mathbf{h}|c^x,c^n,c^h) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\\mathbf{n}\\\mathbf{h}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\eta}^x_{c^x c^n c^h}\\\boldsymbol{\eta}^n_{c^x c^n c^h}\\\boldsymbol{\eta}^h_{c^x c^n c^h}\end{bmatrix}, \begin{bmatrix}\mathbf{\Phi}^{xx}_{c^x c^n c^h} \mathbf{\Phi}^{xn}_{c^x c^n c^h} \mathbf{\Phi}^{xh}_{c^x c^n c^h}\\ \mathbf{\Phi}^{xn}_{c^x c^n c^h} \mathbf{\Phi}^{nn}_{c^x c^n c^h} \mathbf{\Phi}^{nh}_{c^x c^n c^h}\\ \mathbf{\Phi}^{xh}_{c^x c^n c^h} \mathbf{\Phi}^{nh}_{c^x c^n c^h} \mathbf{\Phi}^{hh}_{c^x c^n c^h}\end{bmatrix}\right),$$

where $\boldsymbol{\eta}^x_{c^x c^n c^h}$, $\boldsymbol{\eta}^n_{c^x c^n c^h}$ and $\boldsymbol{\eta}^h_{c^x c^n c^h}$ are the approximate posterior means of the speech, noise and channel for classes c^x, c^n and c^h, and $\mathbf{\Phi}^{xx}_{c^x c^n c^h}$ etc. specify the covariance matrices for the speech, noise and channel for classes c^x, c^n and c^h.

The goal of variational inference is to minimize the relative entropy (Kullback-Leibler divergence) between q and p:

$$\mathcal{K} = \sum_{\{c^x,c^n,c^h\}} \int_{\{\mathbf{x},\mathbf{n},\mathbf{h}\}} q(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h) \cdot \ln \frac{q(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h)}{p(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h|\mathbf{y})}. \quad (9)$$

This is a particularly good choice for a cost function, because minimizing $\mathcal{K}$ is equivalent to maximizing

$$\mathcal{F} = \ln p(\mathbf{y}) - \mathcal{K} = \sum_{\{c^x,c^n,c^h\}} \int_{\{\mathbf{x},\mathbf{n},\mathbf{h}\}} q(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h) \cdot \ln \frac{p(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h,\mathbf{y})}{q(\mathbf{x},\mathbf{n},\mathbf{h},c^x,c^n,c^h)}. \quad (10)$$

The form of the re-estimation formulas for the parameters of the q distribution can be found in [3].

3. JOINT LEARNING OF NOISE AND CHANNEL DISTORTION

The generalized EM algorithm alternates between: 1) updating one set of variational parameters $\rho^{(t)}_{c^x c^n c^h}$, $\boldsymbol{\eta}^{n\,(t)}_{c^x c^n c^h}$, *etc.* for each frame $t = 1, \ldots, T$, and 2) maximizing $\mathcal{F}$ with respect to the noise and channel model parameters $\pi^n_{c^n}$, $\boldsymbol{\mu}^n_{c^n}$ and $\mathbf{\Sigma}^n_{c^n}$ and $\pi^h_{c^h}$, $\boldsymbol{\mu}^h_{c^h}$ and $\mathbf{\Sigma}^h_{c^h}$. Since $\mathcal{F} \leq \sum_t \ln p(\mathbf{y}^{(t)})$, this procedure maximizes a lower bound on the log-probability of the data, up to approximations in the optimization procedure (see Figure 2).

Setting the derivatives of $\mathcal{F}$ with respect to the noise model parameters to zero, we obtain the following M step updates [1]:

$$\begin{aligned} \pi^n_{c^n} &\leftarrow \frac{1}{T}\sum_t \sum_{c^x,c^h} \rho^{(t)}_{c^x c^n c^h}, \\ \boldsymbol{\mu}^n_{c^n} &\leftarrow \Big(\sum_t \sum_{c^x,c^h} \rho^{(t)}_{c^x c^n c^h} \boldsymbol{\eta}^{n\,(t)}_{c^x c^n c^h}\Big) / \Big(\sum_t \sum_{c^x,c^h} \rho^{(t)}_{c^x c^n c^h}\Big), \\ \mathbf{\Sigma}^n_{c^n} &\leftarrow \frac{\sum_t \sum_{c^x,c^h} \rho^{(t)}_{c^x c^n c^h} \big(\mathbf{\Phi}^{nn\,(t)}_{c^x c^n c^h} + \gamma\big)}{\sum_t \sum_{c^x,c^h} \rho^{(t)}_{c^x c^n}}. \quad (11)\end{aligned}$$

where

$$\gamma = \mathrm{diag}\big((\boldsymbol{\eta}^{n\,(t)}_{c^x c^n c^h} - \boldsymbol{\mu}^n_{c^n})(\boldsymbol{\eta}^{n\,(t)}_{c^x c^n c^h} - \boldsymbol{\mu}^n_{c^n})^\mathsf{T}\big). \quad (12)$$

The update equations for the parameters of the channel model are analogous.

[1] A detailed exposition of this derivation can be found at http://newgist.uwaterloo.ca/~trausti/RobustASR/AdaptiveAlgonquin.html

4. ANALYSIS

In order to disentangle and learn the parameters of the noise and channel models, the algorithm relies on an accurate speech model $p(\mathbf{x})$ as well as a model for how speech, noise and the channel are combined $p(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h})$.

In order to assess the susceptibility of the algorithm to pathological behaviour such as slow or stalled convergence, local minima, and saddle points, we ran the algorithm on synthetic data. The data was generated by sampling from the speech model, noise model and channel model and combining the samples according to Eqn. (4).

Figure 1 shows a pathological case for synthetic data. The figure shows the value of μ_h and μ_n as a function of iteration. In this case, the algorithm learns the channel model quickly, i.e. within about 3-5 iterations.

The true noise model has a relatively flat characteristic with a value of around 6. The noise model is initialized with zero mean and variance 10. After 50 iterations the μ_n values for the higher order log-spectrum coefficients are still at 0.

This is due to the algorithm finding and incorrect but plausible "explanation" of the observed signal. The combination of the model for the sound /s/ and the noise model at iteration 50 (see Fig. 1) produces a relatively good fit to the observed output. The algorithm eventually recovers and learns the correct model. This shows that the algorithm is susceptible to poor initialization. We did not observe such pathological behaviour when the algorithm was run on real speech data and the noise model was initialized with the mean and variance of the first 20 frames of the speech file.

The convergence rate is greatly dependent on the variance of the initial noise and channel models. Convergence is much slower if the initial variance is set to a small value.

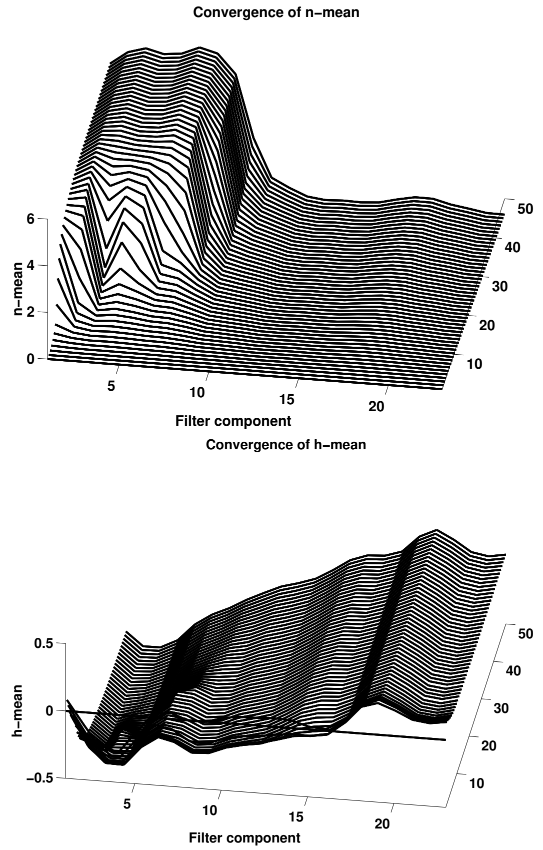

Fig. 1. Convergence of μ_h and μ_n as function of iteration, for joint estimation of noise and channel distortion (simulated data). This shows a pathological case where the noise model has been poorly initialized. Such behaviour was not observed for real speech data.

5. RESULTS

We used set C of the ETSI Aurora task to evaluate the performance and convergence characteristics of the algorithm. The Aurora task consists of spoken digits (TI digits), mixed with various noise types at multiple signal to noise ratios. In addition, set C has been filtered (ITU MIRS frequency characteristic) to simulate channel distortion. The results reported here are for Subway Noise at 10dB SNR. The test set consisted of 1001 files each containing from 1 to 5 spoken digits.

Figure 3 shows the accuracy results for adaptation of the channel model alone (diamonds), the noise model alone (triangles) and joint estimation of noise and channel distortion (squares). In each case the initial noise model was estimated from the first 20 frames of the a speech file. The initial channel distortion was initialized to $\mu_h = 0$ with $\sigma_h^2 = 1$. In these experiments, the noise and channel models were single multivariate gaussians.

First note the results when the algorithm was constrained to adapt only the channel model (i.e. noise model was estimated from first 20 frames and not adapted). In this case, the channel model was initialized to $\mu_h = 0$ and $\sigma_h^2 = 1$. The accuracy goes from 74.42% to a maximum of 86.09%. The non-adaptive algorithm that does not take into account the channel distortion ($\mu_h = 0$, $\sigma_h^2 = 1 \cdot 10^{-4}$) achieves accuracy of 84.36% for this condition.

A second case was run where only the noise model was adapted. The initial noise model was estimated from the first 20 frames, and the variance was multiplied by 3, in order to speed up convergence. The channel model was set to $\mu_h = 0$ and $\sigma_h^2 = 1 \cdot 10^{-4}$. The recognition accuracy goes from 81.95% to a maximum of 87.23% at iteration 19. The recognition rate declines after iteration 19. This interesting effect may be due to the algorithm attempting to compensate for the channel with the noise model.

The third case shown in Figure 3 is that of joint adaptation of noise and channel. In this case, the accuracy goes

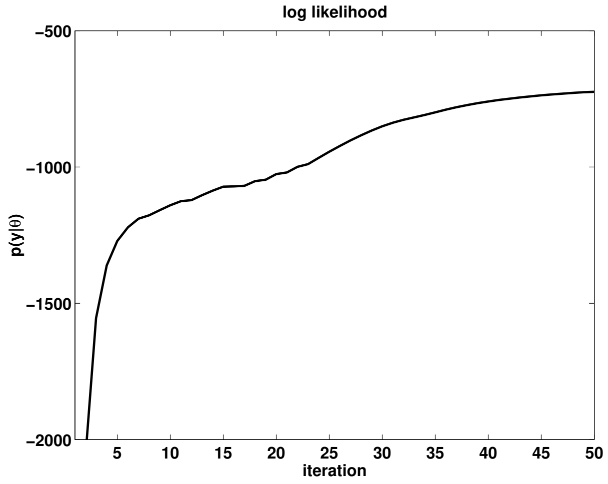

Fig. 2. Convergence of log-likelihood as a function of iteration for the data shown in Figure 1.

from 73.01% to a maximum of 87.78% at iteration 37. This is 0.55% higher than the accuracy for noise adaptation alone at iteration 19 (87.23%). In comparison to the non-adaptive algorithm, the absolute drop in word error rate is 3.4% and the relative drop of is 21.8%. This illustrates well the effectiveness of joint noise and channel adaptation.

These results indicate that the algorithm can successfully and simultaneously learn the additive distortion due to the channel and the non-linear distortion due the noise, and thus successfully untangle these two types of distortion.

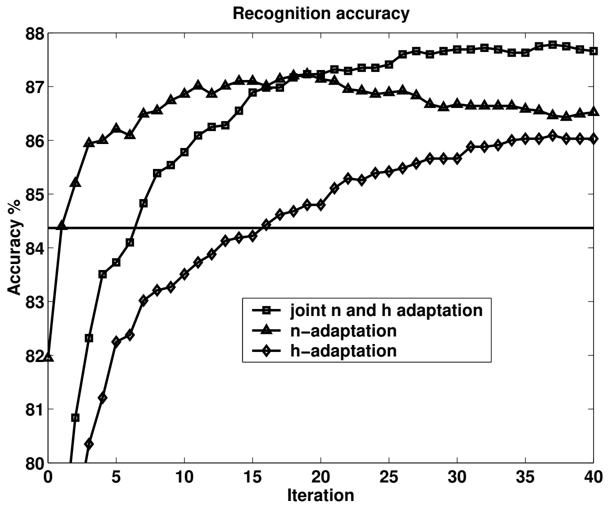

Fig. 3. Recognition accuracy as a function of iteration. Diamond-line shows accuracy when adapting $\mathbf{h}$ alone, triangle-line shows $\mathbf{n}$-adaptation and square-line shows accuracy for joint $\mathbf{n}$ and $\mathbf{h}$ adaptation. Horizontal line shows result for non-adaptive algorithm.

6. CONCLUSION AND FUTURE WORK

In this paper we have introduce a principled way of jointly learning the noise and channel distortion characteristics. The method is based on the Algonquin framework, and employs a Generalized EM strategy. We examined pathological cases but found that for real speech data, the algorithm can successfully and simultaneously learn the parameters of the noise and channel models. We showed that recognition accuracy is substantially improved over the recognition accuracy of the non-adaptive algorithm.

For this method to be practical, the number of iterations has to be small. Our current implementation uses a poor initialization for the channel model. We are currently exploring better strategies for initializing the parameters of the channel model, and other methods for speeding up convergence.

7. REFERENCES

[1] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," *Procedings of ICSLP*, 2000.

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

[3] B.J. Frey, L. Deng, A. Acero, and T. Kristjansson, "Algonquin: Iterating laplace's method to remove multiple types of noise and channel distortion from log-spectra used in robust speech recognition," *Eurospeech*, 2001.

[4] T. Kristjansson, L. Deng, A. Acero, and B.J. Frey, "Towards non-stationary noise adaptation for large vocabulary speech recognition," *In Proc. of ICASSP*, 2001.

[5] B.J. Frey, T. Kristjansson, L. Deng, and A. Acero, "Leaning dynamic noise models from noisy speech for robust speech recognition," *Advances in Neural Information Processing Systems*, 2001.

[6] M.I. Jordan, Z. Grahamani, T.S. Jaakkola, and L.K. Saul, *Learning in Graphical Models*, chapter An introduction to variational methods for graphical models, Klewer Academic Publishers, Norwell MA., 1998.

[7] B.H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, pp. 275 – 294, 1991.

[8] H. Attias, J.C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Advances in Neural Information Processing Systems 13*, 2000.