NOISE ROBUST SPEECH RECOGNITION USING GAUSSIAN BASIS FUNCTIONS FOR NON-LINEAR LIKELIHOOD FUNCTION APPROXIMATION

Chris Pal[1,2*] *, Brendan Frey*[1,2] *and Trausti Kristjansson*[1,2]

[1]University of Waterloo, Dept. Computer Science, Waterloo, Ontario, Canada
[2]Univerisity of Toronto, Dept. Electrical and Computer Engineering, Toronto, Ontario, Canada

ABSTRACT

One approach to achieving noise and distortion robust speech recognition is to remove noise and distortion with algorithms of low complexity prior to the use of much higher complexity speech recognizers. This approach has been referred to as cleaning. In this paper we present an approach for speech cleaning using a time-varying, non-linear probabilistic model of a signals log Mel-filter-bank representation. We then present a new non-linear probabilistic inference technique and show results using this technique within the probabilistic cleaning model. In this approach we represent distributions for underlying noise, speech and channel characteristics as Gaussian mixtures and use Gaussian basis functions to model the non-linear likelihood function. This allows us to efficiently compute complex multi-modal probability distributions over speech and noise components of the underlying signal. We show how this method can be used to clean speech features and present results using the Aurora 2 speech recognizer trained on clean speech data. We present competitive initial results from a minimum mean square error version of this approach for a subset of the Aurora 2 noisy digits recognition tasks.

1. INTRODUCTION

Recently there has been particular interest in techniques for removing noise and distortion from speech using relatively low complexity algorithms prior to the use of more complex recognizers [1, 2, 3]. For example, in the work of [1] four algorithms are combined to clean noisy speech. In this work, the components of the approach consist of: variable frame rate analysis, peak isolation, peak-to-valley ratio locking and harmonic demodulation.

In the SPLICE technique [2] noise characteristics are embedded in a piecewise linear mapping between "stereo" clean and distorted cepstral vectors. In other work, formal probabilistic models have been used. In the ALGONQUIN method [3] an iterative form of Laplace's method [4] is used to compute approximate probability distributions over noise, speech and channel components in the log Mel-filter-bank feature domain. As in the ALGONQUIN approach, we use a well-defined probabilistic model. In section 2 we illustrate our model using a graphical representation and present computations in terms of inference operations in this graphical model [5].

Part of the motivation for cleaning in the log-spectral domain is that the channel properties interact linearly. This simplifies adaptation to channel properties. However, when working in the log-spectral domain one of the main challenges is to accurately approximate the non-linear relationships between clean speech and noise in the log- spectrum. Other approaches have used the vector Taylor series [6] or variational approximations [3] to deal with this non-linearity. In section 2.1 we present the form of this non-linear relationship as derived and presented in [6, 4, 3]. In section 3 we present our approach to modeling the non-linear relationship using Gaussian basis functions. We show how this allows us to perform efficient computations producing Gaussian mixtures representing marginal probability distributions for noise, speech and channel components of the underlying signal.

Finally, in section 4 we present results from the initial instantiation of our approach where we compute minimum mean squared error (MMSE) estimates from the joint noise and speech marginal probability distributions. We use the recognizer distributed with the Aurora 2 [7] noisy TI digits database trained on clean speech. We present results for a subset of the Aurora 2, Test B noise cases. We use an instantiation of our model operating independently over log Mel-filter-bank frames and a version incorporating HMM like time dynamics.

2. A PROBABILISTIC MODEL FOR NOISE, CLEAN SPEECH AND CHANNEL

The *structure* of our probabilistic model is the same as that of ALGONQUIN [3]. In section 3 we describe the alternative method we use for inference in this model. Here we briefly describe the probabilistic model and illustrate it as

*Thanks to the Natural Science and Engineering Research Council (NSERC) of Canada and Bell University Labs

a Bayesian Network [8]. It is possible to phrase the cleaning task as a matter of computing MMSE estimates from marginal probability distribution for clean speech. Another advantage of working in the log spectrum as opposed to the commonly used Mel-frequency cepstrum is that MMSE estimates for clean speech can be performed independently in each dimension without affecting the other dimensions. Thus, the dimensions can be decoupled and treated as being conditionally independent given a discrete latent variable. This structure can be illustrated graphically as a Bayesian network with a latent discrete variable variable C_t indexing the mixture component c at time t. We then group the clean speech x_i, noise n_i and channel h_i variables into joint three-dimensional continuous variables $\mathbf{z}_i = [x\ n\ h]_i^T$. We can then treat the interaction of $y_i = g(\mathbf{z}_i^T)$ separately within each dimension i separately. The model is illustrated in figure (1) where we have also illustrated how the models for frames are coupled are coupled in time, t through the discrete variable C.

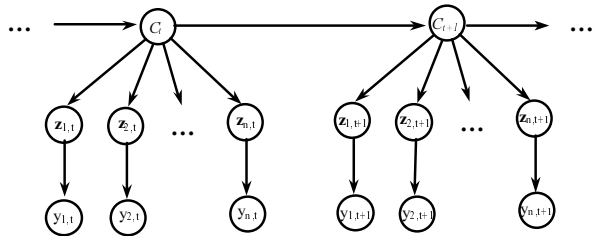

Fig. 1. A Bayesian network illustrating our probability model for speech frames coupled in time.

2.1. The Non-Linear Conditional Distribution

In this model, the computation of the marginal distribution of the joint $\mathbf{z}_i = [x\ n\ h]_i^T$ variable involves a nonlinear conditional distribution $P(y_i|[x\ n\ h]_i^T)$. In [3, 6, 4] an approximation for this non-linear function is derived as the following:

$$\mathbf{y} \approx \mathbf{g}([\mathbf{x}\ \mathbf{n}\ \mathbf{h}]^T) = \mathbf{h} + \mathbf{x} + \ln(\mathbf{1} + \exp(\mathbf{n} - \mathbf{h} - \mathbf{x})), \quad (1)$$

where the ln() and exp() operations apply to individual elements of the vector argument. Under the assumption of Gaussian approximation errors, the corresponding conditional distribution in the probability model can be written as:

$$p(\mathbf{y}|[\mathbf{x}\ \mathbf{n}\ \mathbf{h}]^T) = \mathcal{N}(\mathbf{y}; \mathbf{g}([\mathbf{x}\ \mathbf{n}\ \mathbf{h}]^T), \mathbf{\Psi}), \quad (2)$$

where $\mathbf{\Psi}$ is the covariance matrix of errors. For more details about this derivation and approximation see [3, 6, 4].

As in [3] we encode prior knowledge about $p(\mathbf{x}, \mathbf{n}, \mathbf{h})$ by grouping these components into three-dimensional priors within each of the log-Mel-filterbank dimensions i. In this way we can parameterize the prior as a mixture model in the following way:

$$\begin{aligned} p(c) &= \pi_c \\ p(x_i, n_i, h_i|C) &= \mathcal{N}([x\ n\ h]_i^T; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c). \end{aligned} \quad (3)$$

The joint distribution, illustrated in figure (1) is then given by:

$$\begin{aligned} &p(\mathbf{y}, \mathbf{x}, \mathbf{n}, \mathbf{h}, C) = \\ &\mathcal{N}(\mathbf{y}; \mathbf{g}([\mathbf{x}\ \mathbf{n}\ \mathbf{h}]^T); \mathbf{\Psi}) \mathcal{N}([\mathbf{x}\ \mathbf{n}\ \mathbf{h}]^T; \boldsymbol{\mu}_c, \mathbf{\Sigma}_c) \pi_c. \end{aligned} \quad (4)$$

The difficulty with probabilistic inference in this model lies in computing the marginal distribution $p(\mathbf{x}, \mathbf{n}, \mathbf{h}|\mathbf{y}, C)$ In 3 we show how the use of Gaussian basis functions to approximate the likelihood function produces easily manipulated Gaussian distributions for marginals on $[x\ n\ h]_i^T$.

3. APPROXIMATING THE NON-LINEAR LIKELIHOOD USING GAUSSIAN BASIS FUNCTIONS

We are interested in computing $p(\mathbf{x}, \mathbf{n}, \mathbf{h}|\mathbf{y}, C)$ using an efficient and accurate approximation. Consider the approximate relationship in (1). We can rewrite this as:

$$\begin{aligned} \mathbf{n} &= \mathbf{h} + \mathbf{x} + \ln(\exp(\mathbf{y} - \mathbf{h} - \mathbf{x}) - \mathbf{1}) \\ &= (\mathbf{h} + \mathbf{x} - \mathbf{y}) + \ln(\exp(-(\mathbf{h} + \mathbf{x} - \mathbf{y}) - \mathbf{1}) + \mathbf{y} \\ &= \mathbf{f}(\mathbf{h} + \mathbf{x} - \mathbf{y}) + \mathbf{y}. \end{aligned} \quad (5)$$

Expressed in this way, we can more easily see that we have a function that maintains its form but is linearly translated with the observed signal $\mathbf{y}$. It is then possible to approximate the likelihood corresponding to the conditional in (2) using Gaussian basis functions with linearly translated means. We can do so in each dimension i as follows

$$\begin{aligned} &p(x_i, n_i, h_i|y_i) = \\ &\sum_l w_l \mathcal{N}\left(\begin{bmatrix} x \\ n \\ h \end{bmatrix}_i ; \begin{bmatrix} \mu_x \\ \mu_n \\ \mu_h \end{bmatrix}_i + \begin{bmatrix} y \\ y \\ 0 \end{bmatrix}_i , \mathbf{\Sigma}_{i,l} \right), \end{aligned} \quad (6)$$

where $\mathbf{\Sigma}_{i,l}$ is the covariance matrix (although in this case it is not a true covariance matrix, but a parameter of the basis function) for each Gaussain l in dimension i. This form of the approximation for the likelihood is also particularly compatible with the Gaussian mixtures used for our noise and clean speech prior model. Let $\mathbf{z}_i = [x\ n\ h]_i^T$. Then, consider the marginal distribution on $p(\mathbf{z}_i|y_i, C)$ in one of the dimensions independently from the other dimensions. This marginal distribution can be computed as follows:

$$p(\mathbf{z}_i|y_i, C) = \sum_l \sum_c \frac{1}{\mathcal{Z}_i} \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_{i,l,c}, \mathbf{\Sigma}_{i,l,c}), \quad (7)$$

where $\mathbf{\Sigma}_{i,l,c}$ and $\boldsymbol{\mu}_{i,l,c}$ are the posterior means and covariance matrices in each dimension. They are computed for each dimension i (omitting i for clarity) as follows:

$$\begin{aligned}\mathbf{\Sigma}_{l,c} &= (\mathbf{\Sigma}_l^{-1} + \mathbf{\Sigma}_c^{-1})^{-1}, \\ \boldsymbol{\mu}_{l,c} &= \mathbf{\Sigma}_{l,c}(\mathbf{\Sigma}_l^{-1}\boldsymbol{\mu}_l + \mathbf{\Sigma}_c^{-1}\boldsymbol{\mu}_c),\end{aligned} \tag{8}$$

where $\boldsymbol{\mu}_l$ is computed for each dimension i as:

$$\boldsymbol{\mu}_{i,l} = \left(\begin{bmatrix} \mu_x \\ \mu_n \\ \mu_h \end{bmatrix}_{i,l} + \begin{bmatrix} y \\ y \\ 0 \end{bmatrix}_i \right) \tag{9}$$

The normalization constant $\mathcal{Z}_i$, can be computed for each dimension i as follows:

$$\mathcal{Z}_i = \sum_l \sum_c w_l \pi_c \frac{|\mathbf{\Sigma}_{i,l,c}|^{\frac{1}{2}}}{|\mathbf{\Sigma}_{i,l}|^{\frac{1}{2}}|\mathbf{\Sigma}_{i,c}|^{\frac{1}{2}}} \frac{\mathcal{T}_{i,l,c}}{2\pi}, \tag{10}$$

where

$$\mathcal{T}_{i,l,c} = \exp\left[-\frac{1}{2}(\boldsymbol{\mu}_l \mathbf{\Sigma}_l \boldsymbol{\mu}_l' + \boldsymbol{\mu}_c \mathbf{\Sigma}_c \boldsymbol{\mu}_c' - \boldsymbol{\mu}_{l,c}\mathbf{\Sigma}_{l,c}\boldsymbol{\mu}_{l,c}')\right]_i \tag{11}$$

To understand how these partial computations in each dimension are coupled it is helpful to think of the graph in figure (1). We can then think of computing the likelihoods from each dimension separately for the discrete variable C. For a model with no time dynamics, the "new" prior $\tilde{\pi}_{i,c}$ for a given dimension i coupled with the other dimensions can be computed as follows:

$$\tilde{\pi}_{i,c} = \pi_c \prod_{j \neq i} \sum_l w_{j,l} \frac{|\mathbf{\Sigma}_{j,l,c}|^{\frac{1}{2}}}{|\mathbf{\Sigma}_{j,l}|^{\frac{1}{2}}|\mathbf{\Sigma}_{j,c}|^{\frac{1}{2}}} \frac{\mathcal{T}_{j,l,c}}{2\pi}, \tag{12}$$

where we have used determinants for notational simplicity; However, it should be noted that one can use the simpler equivalent computations when matrices are diagonal. Time dynamics are easily incorporated into the model using a conditional distribution or transition matrix for $P(C_{t+1}|C_t)$. This conditional distribution can then be estimated using the standard hidden markov model (HMM) computations [9] and the "likelihoods" computed within (12) from each time step. Similarly, probabilistic inference using the time coupled C_t can be performed using HMM or the equivalent message passing techniques [5]. We can represent either a time coupled prior or a static but dimensionally coupled prior using $\tilde{\pi}_{i,c}$.

The marginal distribution $p(\mathbf{z}_i|\mathbf{y}, \mathbf{z}_{j\neq i}, C)$, for the noise, speech and channel in a given dimension, coupled with the other dimensions (and possibly time) can be expressed as:

$$p(\mathbf{z}_i|\mathbf{y}, \mathbf{z}_{j\neq i}, C) = \frac{1}{\tilde{\mathcal{Z}}_i} \sum_l \sum_c w_l \tilde{\pi}_{i,c} \frac{|\mathbf{\Sigma}_{i,l,c}|^{\frac{1}{2}}}{|\mathbf{\Sigma}_{i,l}|^{\frac{1}{2}}|\mathbf{\Sigma}_{i,c}|^{\frac{1}{2}}} \frac{\mathcal{T}_{i,l,c}}{2\pi} \mathcal{N}(\mathbf{z}_i; \boldsymbol{\mu}_{i,l,c}, \mathbf{\Sigma}_{i,l,c}), \tag{13}$$

where $\tilde{\mathcal{Z}}_i$ is the new normalization constant and is easily computed. Further, the MMSE estimate is also easy to compute as it is simply a weighted combination of the means in (13).

In addition to using Gaussian functions to approximate the likelihood it is also possible to use Heaviside functions to approximate the likelihood in regions that extend off to infinity and Heaviside functions "cutting" Gaussians at the mean near the start of the bend. This is illustrated in Figure (2). As such, when computing the MMSE estimate there are "chopped" Gaussians in the posterior distribution and for computing these values we can make use of:

$$\begin{aligned}&\int_\alpha^\beta \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) x\, dx = \\ &\frac{\sigma}{\sqrt{2\pi}}\left[-exp\left(-\frac{(\beta-\mu)^2}{2\sigma^2}\right) + exp\left(-\frac{(\alpha-\mu)^2}{2\sigma^2}\right)\right] \\ &+ \frac{\mu}{2}\left[erf\left(\frac{(\beta-\mu)}{\sqrt{2}\sigma}\right) + erf\left(\frac{(\mu-\alpha)}{\sqrt{2}\sigma}\right)\right],\end{aligned} \tag{14}$$

where we let α and β be ∞,$-\infty$ or μ_{l^*}, the location of the end of the Heaviside function that does not extend to infinity. For additional speed, these integrals can be pre-evaluated using one dimensional lookup tables.

3.1. Examples of the Approximation

There are a number of ways to fit Gaussian basis functions to the likelihood that we wish to approximate. For clarity we illustrate the two dimensional likelihood function over x and n. We, have used rejection sampling to draw points from this the function and we then fit mixture models to the sampled function using the Expectation Maximization (EM) algorithm to fit the approximation. Figure (2) (right) shows points sampled from the function for $y = 7$ and the approximation used in our experiments is shown in Figure (2) (left). As the likelihood function we are approximating extends to negative infinity along both the noise and clean speech axes we deal with this in the following way. In figure (2) the approximation consists of five diagonal covariance Gaussians modeling the bend of the likelihood, two Gaussians truncated using Heaviside functions and two one-dimensional Gaussians that extend to negative infinity.

4. RESULTS ON THE AURORA2 NOISY DIGITS

Here we present the results of our experiments on the Aurora 2 database for the restaurant noise condition, Test B. In all cases we present results where a 256 component, diagonal covariance Gaussian mixture model trained on the clean speech was used for the speech model. In table 1 we compare the impact of using 5, 15 and 20 Gaussians to model

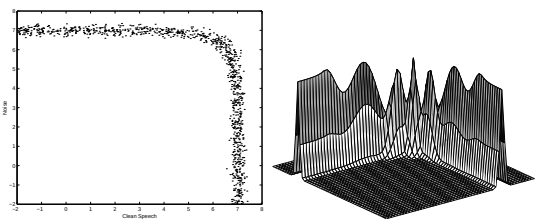

Fig. 2. (left) The sampled likelihood. (right) The likelihood function approximated by five diagonal covariance Gaussians.

SNR	5	15	20
20 dB	97.67	97.53	97.57
15 dB	96.50	96.41	96.43
10 dB	90.62	90.60	90.75
5 dB	75.25	74.77	75.87
0 dB	47.14	47.37	48.08
Avg	81.45	81.34	81.74

Table 1. Comparison of *Word Accuracy* using two noise components and using five, fifteen and twenty components to model the likelihood bend.

the likelihood curve. In table 2, 5 diagonal covariance Gaussians were used to approximate the likelihood function. We show results labeled 1S obtained from fitting a single Gaussian noise model estimated for each file using the first 20 frames of the file. These are compared with results using a one, two and four component model fitted using the first and last 20 frames of the file (labeled 1,2,4). The column labeled 4T represents results using the four component noise model, but incorporating time dynamics for the speech model as described in section 3. In table 3 we compare these results with recent results from Microsoft's non-adaptive SPLICE presented in [2], the results presented in [1] (UCLA), and the results from the recognizer using the noisy speech with no cleaning.

SNR	1S	1	2	4	4T
20 dB	97.14	97.73	97.67	97.68	97.88
15 dB	94.96	95.84	96.50	96.32	96.47
10 dB	88.16	89.05	90.62	90.64	91.40
5 dB	69.07	72.38	75.25	76.51	77.43
0 dB	40.33	42.12	47.14	48.97	50.72
Avg	77.93	79.42	81.45	82.02	82.78

Table 2. Comparison of *Word Accuracy* as the number of noise components increases and time dynamics are incorporated into the model.

SNR	BASE	UCLA	SPLICE	4T
20 dB	89.35	97.18	98.59	97.88
15 dB	74.52	95.64	97.51	96.47
10 dB	52.75	91.25	94.29	91.40
5 dB	28.86	75.47	81.73	77.43
0 dB	12.59	44.67	51.61	50.72
Avg	52.59	80.84	85.48	82.78

Table 3. Comparison of *Word Accuracy* using the recognizer with no cleaning, results from UCLA, Microsoft's non-adaptive results and our model using time dynamics.

5. CONCLUSIONS AND DISCUSSION

We have presented a new inference methodology in a non-linear probabilistic model with time dynamics. Our MMSE cleaning implementation has produced competitive results on a subset of the Aurora 2 digit recognition tasks. We have found it important to ensure "good" likelihood models are fit when using EM. Finally, we have observed that a relatively small number of Gaussians can be used as a reasonable approximation of the likelihood curve.

6. REFERENCES

[1] X. Cui Q. Zhu, M. Iseli and A. Alwan, "Noise robust feature extraction for asr using the aurora 2 database," *Eurospeech 2001, Special Event: Noise Robust Recognition, Web only paper*, vol. 1, pp. 185–188, 2001.

[2] L. Deng J. Droppo and A. Acero, "Evaluation of the splice algorithm on the aurora 2 database," *Eurospeech 2001, Special Event: Noise Robust Recognition, Web only paper*, vol. 1, pp. 217–220, 2001.

[3] A. Acero B. Frey, L. Deng and T. Kristjansson, "Algonquin: Iterating laplace's method to remove multiple type of acoustic distortion for robust speech recognition," *Eurospeech*, 2001.

[4] P. Moreno, "Speech recognition in noisy environments," *Carnegie Mellon University, Pittsburg PA*, vol. Doctoral dissertation., 1996.

[5] J. Pearl, *Probabilistic Inference in Intelligent Systems*, Morgan Kaufmann, San Mateo, California, 1988.

[6] T. Kristjansson A. Acero, L. Deng and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech," *Proceedings of the International Conference on Spoken Language Processing*, pp. 869–872, October 2000.

[7] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000, Automatic Speech Recognition: Challenges for the Next Millennium, Paris France*, September 18-20 2000.

[8] S. Russell, "Expressive probability models for speech recognition and understanding," *Proc. International Workshop on Automatic Speech Recognition and Understanding (ASRU), Keystone, Colorado (invited paper)*, 1999.

[9] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall Inc., Englewood Cliffs NJ, 1993.