ACCOUNTING FOR UNCERTAINITY IN OBSERVATIONS: A NEW PARADIGM FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Trausti T. Kristjansson, Brendan J. Frey

Probabilistic and Statistical Inference Group
University of Toronto

ABSTRACT

We introduce a new paradigm for Robust Automatic Speech Recognition that directly incorporates information about the uncertainty introduced by environmental noise. In contrast to the *feature cleaning* and *model adaptation* paradigms, where the noise compensation mechanism is separate from the recognizer, the new paradigm unifies the noise compensation mechanism and the recognizer. The Algonquin framework serves to demonstrate the importance of retaining *soft information*, i.e. information about the degree of uncertainty in the observations. The Algonquin framework employs Gaussian mixture models to model both noise and speech. Uncertainty introduced by the noise process is captured by the variance of the noise model. The Algonquin framework also allows us to isolate the effect of retaining or discarding soft information. Our initial results indicate that substantial improvements in recognition rates can be achieved through the use of soft information.

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.

The two most common approaches to noise robust speech recognition are *feature cleaning* and *model adaptation*. The goal of feature cleaning 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), Cepstral Mean Normalization (CMN) and Algonquin[1], to name a few. The complexity of feature cleaning methods can be low. However, these methods produce point estimates of the clean speech and information about the uncertainty in the observations is thus lost.

A second method is to alter the acoustic models of the recognizer. In this case, given a model of the noise and channel environment, the goal is to update the acoustic models of the recognizer, such that they approximate the models that would have resulted from training directly on the speech in the current noisy environment. Methods that fall into this category are PMC[2] and VTS[3]. The advantage of this approach is that the models reflect the inherent uncertainty that is introduced by the noise process. A disadvantage is the relative computational complexity of this approach.

In this paper we introduce a new approach that gives the advantages of both above mentioned paradigms.

2. FUNDAMENTALS

For speech recognition systems based on Hidden Markov Models, the most common decoding method is based on the Viterbi algorithm. The Viterbi algorithm returns the most likely state sequence $\mathbf{s}$, given a sequence $\mathbf{X}$ of observation vectors $\mathbf{x}$:

$$\hat{\mathbf{s}} = \underset{\mathbf{s}}{argmax}\ p(\mathbf{s}, \mathbf{X}) = \underset{\mathbf{s}}{argmax}\ p(\mathbf{s}|\mathbf{X}) \qquad (1)$$

An HMM based speech recognizer is comprised of acoustic models $p(\mathbf{X}|s)$ as well as language models and state transition matrices of the HMMs which combine to give the transition probabilities between states $p(s_i|s_{i-1} \ldots s_0)$, where i indexes the time frame.

When there is noise and channel distortion in the environment, we observe corrupted features $\mathbf{Y}$ instead of $\mathbf{X}$. The environmental noise process introduces both bias and fundamental uncertainty. Bias shifts the classification boundaries, but can be accounted for. However uncertainty increases the overlap of class conditional likelihood distributions, and thus the classification error increases. Despite this, the optimal classification strategy is based on using the posterior of the noisy speech $p(\mathbf{s}|\mathbf{Y})$. By the data processing inequality[4] it is impossible to gain more information about s by manipulating $\mathbf{y}$ e.g. by cleaning $\mathbf{y}$ to produce $\hat{\mathbf{x}}$.

Intuitively, the effect of noise is to reduce our certainty that an observation belongs to one class rather than the other. Figure 1 shows observation scores for a particular speech frame. At the top, the log observation likelihoods $\log(p(\mathbf{x}|s))$ of the clean speech frame are shown. The bottom two plots show observation scores for the same frame with noise at 5dB SNR. The middle plot shows the log

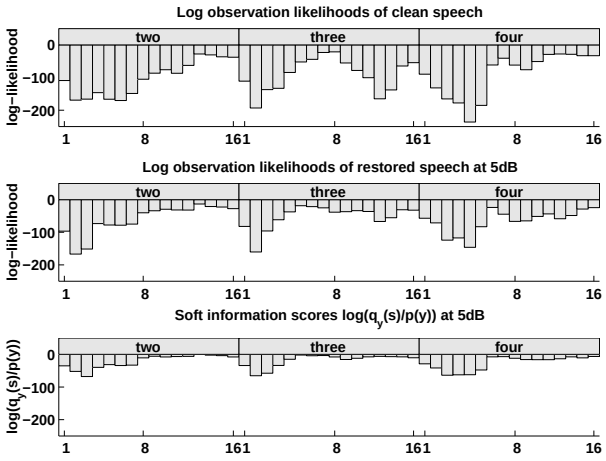

Fig. 1. The plots show the log observation likelihood for the states s in the word models for 'two','three' and 'four' for frame 30 of file MAH_3A. This file contains the utterance 'three'. The plots show clean speech $\log(p(\mathbf{x}|s))$, cleaned speech $\log(p(\hat{\mathbf{x}}|s))$ and the soft information score which is approximately equal to $\log(p(\mathbf{y}|s)) + const.$

observation likelihood $\log(p(\hat{\mathbf{x}}|s))$ of the cleaned speech frame and the bottom plot shows the soft information score which is approximately equivalent to $\log(p(\mathbf{y}|s))$[1]. Notice that $p(\hat{\mathbf{x}}|s)$ looks more like $p(\mathbf{x}|s)$, and the range of soft information scores is smaller. Since a cleaning method is forced to choose a single $\hat{\mathbf{x}}$, it may amplify the error in the case of a wrong choice, instead of remaining neutral as to which class the observation belongs.

Speech recognition systems employ complex speech models including language models and word or phone HMMs that encode the transition probabilities $p(s_i|s_{i-1} \dots s_0)$ between states. Some cleaning methods such as Algonquin also use speech models. In the case of Algonquin, speech is modeled by a Gaussian Mixture Model (GMM). By deferring the hard decision to the decoding step of the recognizer, we avoid making a decision based on the much weaker state transition model of the cleaning algorithm[2]. As we will see below, we remove the effect of the "language model" of the cleaning algorithm by dividing by $p(s)$ in Eqn. (5).

Model adaptation replaces $p(\mathbf{X}|s)$ by an approximation to $p(\mathbf{Y}|s)$. Thus model adaptation methods preserve information about the uncertainty introduced by the environmental noise.

We propose two alternatives to the model adaptation method, that also preserve the information about the uncertainty of the observations. The first relies on estimating $p(s_i|\mathbf{y}_i)/p(s_i)$ and returning this value to the recognizer:

$$p(\mathbf{s}|\mathbf{Y}) = p(s_0) \prod_i \frac{p(\mathbf{y}_i|s_i)}{p(\mathbf{y}_i)} \cdot p(s_i|s_{i-1}) = p(s_0) \prod_i \frac{p(s_i|\mathbf{y}_i)}{p(s_i)} \cdot p(s_i|s_{i-1}) \quad (2)$$

Thus, if we can approximate $p(s_i|\mathbf{y}_i)/p(s_i)$ we can preserve soft information. The Algonquin framework allows us to do this. We present results for this approach below.

Alternatively, we can estimate $p(\mathbf{x}_i|\mathbf{y}_i)/p(\mathbf{x}_i)$, since:

$$\begin{aligned} p(\mathbf{s}|\mathbf{Y}) &= p(s_0) \prod_i \frac{p(\mathbf{y}_i|s_i) \cdot p(s_i|s_{i-1})}{p(\mathbf{y}_i)} \\ &= p(s_0) \prod_i \frac{\int p(\mathbf{y}_i|\mathbf{x}_i) p(\mathbf{x}_i|s_i) d\mathbf{x}}{p(\mathbf{y}_i)} \cdot p(s_i|s_{i-1}) \\ &= p(s_0) \prod_i \int \frac{p(\mathbf{x}_i|\mathbf{y}_i)}{p(\mathbf{x}_i)} p(\mathbf{x}_i|s_i) d\mathbf{x} \cdot p(s_i|s_{i-1}) \end{aligned} \quad (3)$$

In this case, the goal is to estimate $\frac{p(\mathbf{x}_i|\mathbf{y}_i)}{p(\mathbf{x}_i)}$ in a form that allows for the integral to be calculated easily, e.g. in a Gaussian form. Some noise cleaning methods e.g. Algonquin, employ speech priors $p(\mathbf{x}_i)$ and estimate the posterior $p(\mathbf{x}_i|\mathbf{y}_i)$, and can thus be used in this context.

3. VARIATIONAL ESTIMATE OF $p(s|\mathbf{y})/p(s)$

The Algonquin framework is ideally suited to demonstrate the importance of retaining soft information in the decoding of speech. While some noise robustness methodologies, such as spectral subtraction, use point estimates for the noise process, Algonquin used Gaussian mixture models to model both speech and noise. The "uncertainty" introduced by the noise process is captured in the variance parameters of the noise model.

Algonquin uses a variational method to produce an approximation $q_{\mathbf{y}}(\mathbf{x})$ to the posterior $p(\mathbf{x}|\mathbf{y})$. The approximate posterior is used to calculate a point estimate of the clean speech features $\hat{\mathbf{x}}$ through an MMSE estimate:

$$\hat{\mathbf{x}} = \int \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x} \approx \int \mathbf{x} \sum_{\mathbf{s}} q_{\mathbf{y}}(s) q_{\mathbf{y}}(\mathbf{x}|s) d\mathbf{x} \quad (4)$$

Using the cleaned features $\hat{\mathbf{x}}$, the observation likelihood calculated by the recognizer is thus $p(\hat{\mathbf{x}}|s)$. In order to use soft information, we require the evaluation of

$$\frac{p(s|\mathbf{y})}{p(s)} \approx \frac{q_{\mathbf{y}}(s)}{p(s)} \quad (5)$$

which is substituted for $p(\hat{\mathbf{x}}|s)$ in the recognizer. As we will see below $p(s|\mathbf{y}) \approx q_{\mathbf{y}}(s)$ so all the components required to

[1]The plot shows $log(q_{\mathbf{y}}(s)/p(s))$ which is approximately equal to $log(p(\mathbf{y}|s)/p(\mathbf{y})) = log(p(\mathbf{y}|s)) + const.$

[2]The MMSE version of Algonquin uses a GMM to model speech and therefore does not use state transition probabilities.

calculate the soft information score in Eqn. (5) are available from the calculation of the point estimate in Eqn. (4).

We will now fill in some of the relevant details of the Algonquin framework. See [1] for a more thorough introduction, and [5, 6] for a description of noise and channel adaptive extensions.

As noted before, the Algonquin framework employs Gaussian mixture models to model the speech and noise in the log-spectrum domain, thus, the joint distribution over noisy speech $\mathbf{y}$, speech $\mathbf{x}$, speech class s^x, noise $\mathbf{n}$, noise class s^n is:

$$\begin{aligned} p(\mathbf{y},\mathbf{x},\mathbf{n},s^x,s^n) &= p(\mathbf{y}|\mathbf{x},\mathbf{n})p(s^x)p(\mathbf{x}|s^x)p(s^n)p(\mathbf{n}|s^n) \\ &= \mathcal{N}(\mathbf{y};\mathbf{g}\left(\left[\mathbf{x}\mathbf{n}\right]\right),\mathbf{\Psi}) \\ &\cdot \pi^x_{s^x}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}^x_{s^x},\mathbf{\Sigma}^x_{s^x})\cdot\pi^n_{s^n}\mathcal{N}(\mathbf{n};\boldsymbol{\mu}^n_{s^n},\mathbf{\Sigma}^n_{s^n}). \quad (6) \end{aligned}$$

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

$$p(\mathbf{x},\mathbf{n},s^x,s^n|\mathbf{y}) \approx q_{\mathbf{y}}(\mathbf{x},\mathbf{n},s^x,s^n). \quad (7)$$

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 calculating the soft information score $q_{\mathbf{y}}(s)/p(s)$. See [7] for a review of variational inference techniques.

The q function is a mixture of Gaussians:

$$q_{\mathbf{y}}(\mathbf{x},\mathbf{n}) = \sum_{\{s^x,s^n\}} q_{\mathbf{y}}(s^x,s^n)q_{\mathbf{y}}(\mathbf{x},\mathbf{n}|s^x,s^n) \quad (8)$$

where the $q_{\mathbf{y}}(s^x,s^n)$s serve as mixture weights. Note that

$$p(s^x|\mathbf{y}) \approx q_{\mathbf{y}}(s) = \sum_{s^n} q_{\mathbf{y}}(s^x,s^n) \quad (9)$$

which is used in the calculation of the soft information score in Eqn. (5).

In order to find the approximate posterior q, the Algonquin framework uses variational inference. The goal of variational inference is to minimize the relative entropy (Kullback-Leibler divergence) between q and p:

$$\mathcal{K} = \sum_{\{s^x,s^n\}}\int_{\{\mathbf{x},\mathbf{n}\}} q_{\mathbf{y}}(\mathbf{x},\mathbf{n},s^x,s^n) \cdot \ln\frac{q_{\mathbf{y}}(\mathbf{x},\mathbf{n},s^x,s^n)}{p(\mathbf{x},\mathbf{n},s^x,s^n|\mathbf{y})}. \quad (10)$$

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

$$\mathcal{F} = \ln p(\mathbf{y}) - \mathcal{K} = \sum_{\{s^x,s^n\}}\int_{\{\mathbf{x},\mathbf{n}\}} q_{\mathbf{y}}(\mathbf{x},\mathbf{n},s^x,s^n)\cdot\ln\frac{p(\mathbf{x},\mathbf{n},s^x,s^n,\mathbf{y})}{q_{\mathbf{y}}(\mathbf{x},\mathbf{n},s^x,s^n)}. \quad (11)$$

4. EXPERIMENTS AND RESULTS

Our tests were conducted on the Aurora 2 data set produced by ETSI. The experiments were run on Set A of the dataset. This set consists of 1001 files containing spoken digits, for each of the 28 noise conditions. Each file contains from 1 to 5 digits. Four noise types (subway, car, babble and exhibition) were artificially added to the clean speech files at seven SNR levels (Clean, 20dB, 15dB, 10dB, 5dB, 0dB and -5dB).

The Aurora data set is supplied with a standard Mel-frequency Cepstrum Coefficient (MFCC) front end and the CU-HTK speech recognizer. For the experiments reported here, we used filter-bank parameters without delta or acceleration features. These features were produced by altering the standard front end such that it writes out the log-Mel-spectrum values just prior to taking the DCT. It is known that MFCC parameters perform considerably better than filter-bank parameters. Due to the larger complexity of experiments when performed in the MFCC domain we only report results on filter-bank parameters.

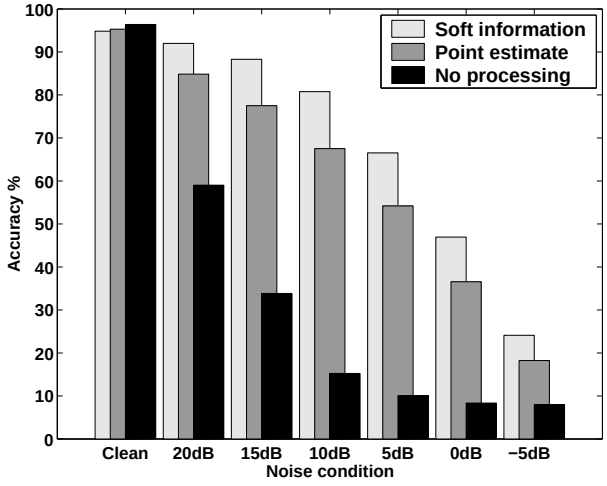

Fig. 2. Average recognition accuracy as a function of signal to noise ratio. The three conditions shown are *soft information* i.e. using $q_{\mathbf{y}_i}(s_i)/p(s_i)$ in the recognizer, using MMSE Algonquin for feature cleaning and no processing of the corrupted features.

The standard HTK recognizer accepts observation features $\mathbf{x}$ and calculates acoustic scores internally based on the acoustic models $p(\mathbf{x}|s)$. Experiments based on feature cleaning and model adaptation can be performed without altering the recognizer, i.e. by supplying $\hat{\mathbf{x}}$ or altering $p(\mathbf{x}|s)$ respectively. However, our algorithm relies on the fusion of the noise adaptation stage and calculation of acoustic scores. Thus, the HTK recognizer had to be altered to accept the scores $q_{\mathbf{y}_i}(s_i)/p(s_i)$ calculated by our algorithm. These

scores were substituted for $p(\mathbf{x}_i|s_i)$ in the recognizer.

Twelve speech models are used in the Aurora task, 'zero' though 'nine', 'oh' and silence. Each model has 16 states and the silence model has 3 states, for a total of 179 states. Thus, $q_{\mathbf{y}_i}(s_i)/p(s_i)$ had to be calculated for each state s_i for each frame i, as described above.

The Algonquin algorithm requires a GMM speech model $p(\mathbf{x})$. $p(\mathbf{x})$ was constructed from the HMM models trained by HTK. The mixture means $\boldsymbol{\mu}^x_{s^x}$ and variance $\boldsymbol{\Sigma}^x_{s^x}$ (see Eqn. (6)) were copied directly from the acoustic models of the recognizer. To find the mixture weights $\pi^x_{s^x}$, a 179×179 state transition matrix was first constructed from the language model and the transition matrices of the HTK HMM word models. Then the stationary distribution of the transition matrix was found and multiplied by the mixture weights of the Gaussian components of the acoustic models. This resulted in a 552 component Gaussian mixture model.

The noise model consisted of a single component multivariate Gaussian. A different model was estimated for each utterance, from the first 20 frames of that file.

The calculation of the soft information score in Eqn. (5) and the point estimate Eqn. (4) share almost all of the same steps. We can therefore provide a comparison that differs only in this aspect (i.e. point estimate vs. soft information), while holding all other aspects constant, such as methodology, approximation errors, speech and noise models etc.

Figure 2 compares the techniques of passing a point estimate of the clean speech to that of taking uncertainty into account by using $q_{\mathbf{y}_i}(s_i)/p(s_i)$. As can be seen recognition accuracy improves considerably for all SNRs except for clean speech where it is slightly reduced. For example, at 15dB, the average accuracy goes from 77.50% accuracy to 88.29% which is an increase in accuracy of 10.78% and a relative drop in Word Error Rate (WER) of 47.60%. As expected, the use of soft information is most effective at intermediate SNRs. For clean speech, there is a drop in accuracy of 0.49% or 10.56% relative WER. At "infinite" SNR, we should ideally leave the parameters unchanged. Approximation error and error in estimation of the noise parameters seems to have a greater adverse effect on the soft information method.

The relative reduction in WER is shown in Table 1. The average relative reduction in WER for noise conditions 20dB through 0dB is 36.06%.

	Subway	Car	Babble	Exhib.	Ave.
clean	-16.97	-9.86	-6.59	-8.81	-10.56
20dB	49.90	36.99	45.97	50.78	45.91
15dB	47.88	39.92	52.62	50.00	47.60
10dB	36.10	40.58	46.60	42.72	41.50
5dB	19.71	27.75	35.33	28.82	27.91
0dB	6.80	13.75	28.85	20.03	17.36
-5dB	1.61	1.73	16.74	10.10	7.55
Ave.	32.08	26.79	41.87	38.47	36.06

Table 1. Relative reduction in word error rate in percent for soft information method compared to using a point estimate of clean speech.

5. CONCLUSION AND FUTURE WORK

In this paper we have clearly shown the importance of incorporating information about the uncertainty introduced by the noise process into the speech decoder. We derived two forms of the *soft information* paradigm and showed how the Algonquin method can be used within this paradigm.

The results given here are based on log-Mel-spectrum features. The recognition rates for these features are lower than when using Algonquin with MFCC features[1, 5]. We are confident that the gains in recognition accuracy due to the use of soft information will carry over to other methods that are compatible with this new paradigm, in particular, we expect similar gains in accuracy for Algonquin when used in the MFCC domain.

5.1. Acknowledgments

We would like to thank Dale Schuurmans, Alex Acero, Li Deng and Chris Pal for helpful discussions and suggestions.

6. REFERENCES

[1] 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.

[2] M.J.F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, University of Cambridge, September 1995.

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

[4] Cover and Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.

[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] T. Kristjansson, B.J. Frey, L. Deng, and A. Acero, "Joint estimation of noise and channel distortion in a generalized em framework," *Proc. ASRU*, 2001.

[7] 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.