TOWARDS NON-STATIONARY MODEL-BASED NOISE ADAPTATION FOR LARGE VOCABULARY SPEECH RECOGNITION

T. Kristjansson, B. Frey

University of Waterloo,
Dep. of Computer Science, 200 University Ave.
Waterloo, Ontario, N2L 3G1, Canada
{ttkristj,frey}@uwaterloo.ca

L. Deng, A. Acero

Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA
{deng,alexac}@microsoft.com

ABSTRACT

Recognition rates of speech recognition systems are known to degrade substantially when there is a mismatch between training and deployment environments. One approach to tackling this problem is to transform the acoustic models based on the channel distortion and noise characteristics of the new environment. Currently, most model adaptation strategies assume that the noise characteristics are stationary. We present results for using multiple noise distributions for the Whisper large vocabulary speech recognition system. The Vector Taylor Series method for adaptation of the distributions is used, and either a weighted average of the noise states or the locally best noise states is used. Our results indicate that for certain types of noise, significant gains in recognition accuracy can be achieved.

1. INTRODUCTION

In order to achieve high recognition rates, the characteristics of the training data should match the deployment data closely. However, a variety of external noise conditions and transfer function characteristics are encountered in real world applications, and it is impossible to train acoustic models for all possible conditions. New applications, e.g. for hand-held devices, impose even more stringent demands on noise compensation than previous applications.

Various methods are used to compensate for mismatch between training and deployment conditions of a speech recognition system. The two main ways of adapting to a new environment are to “clean” the features passed to the speech recognizer, and to alter the acoustic models employed in the recognizer. These techniques are called feature-based methods and model-based methods, respectively. Cepstral Mean Normalization [8] and Spectral Subtraction [9], and their extensions [6] are examples of feature-based methods.

Model adaptation attempts to alter the acoustic models, such that they resemble closely the models attained under matched conditions. One method for accomplishing this is to sample the background noise of a new environment, artificially mix this noise signal with the training set, and retrain the system. However, this is impractical, due to the large storage and time requirements.

A more efficient way of model adaptation is to estimate the noise and transfer function distributions of a new environment, and update the acoustic models, by combining the existing clean speech acoustic models with the estimated environment model.

2. PARALLEL MODEL COMBINATION

The goal of PMC is to update the observation distributions such that they resemble those of a system matched to the deployment environment.

The model for combining noise and speech in the signal domain is shown in Eq.(1).

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

where x is the clean speech signal, h is the impulse response of the transfer function, n is the noise signal, and y is the resulting noisy signal.

If we apply the power Mel spectrum transformation to Eq.(1), we arrive at:

$$|Y(f_i)|^2 = |X(f_i)|^2 |H(f_i)|^2 + |N(f_i)|^2 + \epsilon \tag{2}$$

where ϵ is an error term due to introduction of cross terms in the Mel scale binning of frequencies, and omission of cross terms when performing the power operation.

In the power spectrum domain, finding the corrupted speech distributions is a matter of convolving the noise and signal distributions.

Modern recognizers use Mel Frequency Cepstrum Components, and delta and delta delta coefficients. The acoustic models are mixtures of Gaussian distributions in the MFCC domain. In the power spectrum domain, the distributions are mixtures of log-normals.

The combination of the signal, noise and channel distributions can still be accomplished using numerical integration. This however, is very costly and various methods have been proposed to approximate this combination [1, 4].

2.1. Vector Taylor Series

One such approximation is the Vector Taylor series (VTS), developed by Moreno [5] for log-spectrum features and extended by Acero [10, 7] for MFCC features. VTS has been shown to perform very well for stationary noise. It also has the advantage of being fast.

To see how the VTS method works, we first complete the transformation of Eq.(2) into the cepstrum domain. Taking the log and multiplying by the cosine transform $\mathbf{C}$ we arrive at:

$$\begin{aligned} \mathbf{C}ln|Y|^2 &= \mathbf{C}ln|\mathbf{X}|^2 + \mathbf{C}ln|\mathbf{H}|^2 \qquad (3) \\ &+ \mathbf{C}ln(1 + exp(|\mathbf{N}|^2 - |\mathbf{H}|^2 - |\mathbf{X}|^2) \end{aligned}$$

which can be rewritten as:

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{g}(\mathbf{n} - \mathbf{x} - \mathbf{h}) \qquad (4)$$

where $\mathbf{g}(\mathbf{z}) = \mathbf{C}ln(1 + e^{\mathbf{C^{-1}z}})$. In order to find the updated distributions, each GMM component is transformed independently.

Equation (4) is linearized using the Vector Taylor Series:

$$\begin{aligned} \mathbf{y^{VTS}} &= \mathbf{x_0} + \mathbf{h_0} + \mathbf{g}(\mathbf{n_0} - \mathbf{x_0} - \mathbf{h_0}) \qquad (5) \\ &+ \frac{\delta \mathbf{g}}{\delta \mathbf{x}}(\mathbf{x} - \mathbf{x_0}) + \frac{\delta \mathbf{g}}{\delta \mathbf{h}}(\mathbf{h} - \mathbf{h_0}) + \frac{\delta \mathbf{g}}{\delta \mathbf{n}}(\mathbf{n} - \mathbf{n_0}) \end{aligned}$$

The mean and variance of each component of the updated distributions are found by evaluating the expected value $E(\mathbf{y^{VTS}})$ and variance $\Theta(\mathbf{y^{VTS}})$ of equation $\mathbf{y^{VTS}}$ expanded at the modes of each component of the speech, noise and channel distortion distributions. Thus, the standard VTS method replaces the observation distributions for clean speech $p(\mathbf{x}|s_i)$ with observation distributions for noisy speech $p(\mathbf{y}|s_i)$. For a more detailed exposition of the VTS method, see [5, 7, 10].

The standard VTS method uses a single multivariate distribution to model the noise. In the following we will be using multiple component/state noise distributions. Therefore, the observation distributions will be dependent on the speech state as well as on the noise state $p(\mathbf{y}|s_i, r_j)$. To produce these observation distributions, the standard VTS method is applied once for each noise state.

3. HIDDEN MARKOV MODEL DECOMPOSITION

Varga and Moore [1] introduced the decomposition of speech and noise by the method of Hidden Markov Model Decomposition. Figure 1 shows the Bayesian net representation of two hidden Markov models [2], and the observation sequence that results from combining the outputs of the two models. The speech HMM is represented by the state sequence $s_1, s_2, \ldots, s_M$ with observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$. Similarly, the noise HMM is represented by the state sequence $r_1, r_2, \ldots, r_M$ with observations $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_M$. The resulting distributions of the observed noisy speech $\mathbf{y}_1$, $\mathbf{y}_2, \ldots, \mathbf{y}_M$, can be obtained by the methods discussed above. We used VTS for this purpose.

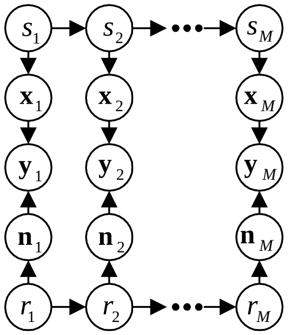

Fig. 1. The speech HMM is shown at the top, and the noise HMM is at the bottom.

Parallel Viterbi decoding[1] results in the optimum state sequences of both speech and noise:

$$\{\hat{S}, \hat{R}\} = \underset{k,l}{\mathrm{argmax}} P(Y, S_k, R_l) \qquad (6)$$

where $S_k = \{s_{1,k}, s_{2,k}, \ldots, s_{M,k}\}$ is a particular speech state sequence and R_l is a particular noise state sequence.

Although Parallel Viterbi leads to an optimum decoding of the speech *and* noise state sequences, we are not usually interested in the state sequence of the noise process. It may therefore be more reasonable to marginalize over all noise state sequences for a given speech state sequence:

$$\{\hat{S}\} = \underset{k}{\mathrm{argmax}} \sum_l P(Y, S_k, R_l) \qquad (7)$$

Using a multi-state noise model incurs a cost, both in the evaluation of observation likelihoods, and in the decoding. The increase in complexity of the likelihood evaluation is linear in the number of states of the noise model. For example, for a 4 state noise model, the number of observation likelihoods that need to be evaluated is 4 fold.

The computation complexity of exact evaluation of Eq.(6) using Viterbi is the same as that of Eq.(7) using a hybrid Viterbi/forward algorithm. However, due to the long term relationships imposed by the language model, large vocabulary speech recognition systems use an approximate Viterbi algorithm based on token passing. In this scenario, the additional cost of Viterbi may be much greater than that of the Viterbi/forward hybrid algorithm.

In the experiments reported below, we assess the difference between these two schemes, using the simplified model shown in Figure 2.

In this simplified model, we have removed the dynamic distributions of the noise process. In this case, the evaluation of Eq.(6) reduces to picking the most likely noise state for each combined speech/noise state,

$$p(\mathbf{y}|s_i, \hat{r}_i) = \max_k p(\mathbf{y}|s_i, r_{i,k}) \tag{8}$$

and the evaluation of Eq.(7) reduces to marginalizing over the noise states

$$p(\mathbf{y}|s_i) = \sum_k p(\mathbf{y}|s_i, r_{i,k}). \tag{9}$$

The experiments give a lower bound on the expected accuracy of the two schemes when the dynamic distributions for the noise process are used.

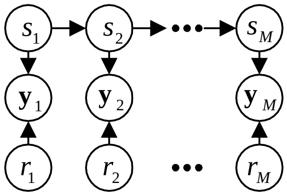

Fig. 2. Network used in experiments. VTS has been used to find the speech and noise state conditional observation distributions and the dynamic links of the noise distribution have been removed.

4. EXPERIMENT AND RESULTS

Experiments were conducted using the Whisper Large Vocabulary Speech Recognition system. For the experiments, a vocabulary size of 5000 words was used with a tri-gram language model. The acoustic models consisted of 6000 senones (shared observation models), each with 20 gaussian mixtures. The system was trained on 16000 sentences of clean speech data from the Wall Street Journal data set. The test set consisted of 167 sentences from the Wall Street Journal spoken by female speakers. Noise was artificially added to each test sentence at 10 dB SNR.

4.1. Noise

Two types of noise were used in the experiments, Babble noise and Roller coaster noise.

Babble noise consists of a large number of speakers talking simultaneously. The state sequence of the 4 state model for Babble noise is shown in Figure 3. As can be seen, the noise tends to stay in a state for a few frames before jumping to another state, however, the transitions between states do not exhibit any apparent structure.

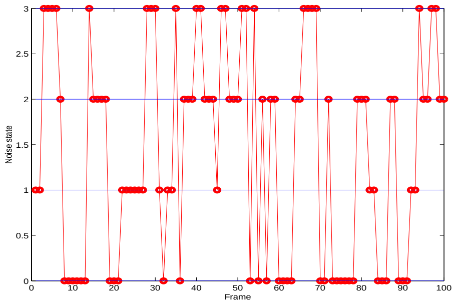

Fig. 3. State sequence for babble noise.

Roller coaster noise is a repetitive sound of a roller coaster moving on a track, similar to the sound of a locomotive. As can be seen in Figure 4 the state sequence is highly regular, moving from one state to the next in a systematic manner. The state sequences for the 8 component models exhibited similar structure.

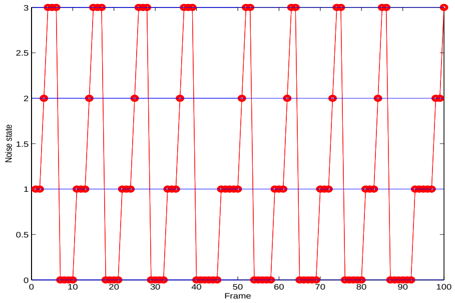

Fig. 4. State sequence for roller coaster noise.

4.2. Multiple component/state observation models

In order to evaluate the effect of introducing multiple state/component models, 1, 4 and 8 component gaussian mixture noise models were trained. 20 - 50 models were trained for each noise type and model size, and the best model was used. This was done to avoid local minima.

In the experiment labeled MAX in Table 1, the maximum likelihood noise state was chosen in the observation likelihood evaluation, according to Eq.(8). For example, for the 4 state noise model, 4 senones were stored, one for each noise state.

In the experiments labeled SUM in Table 1, we marginalized over the noise states (see Eq.(9)). This can be incorporated into the observation likelihood evaluation of a GMM based speech recognizer by increasing the number of mixtures. Since each senone has 20 gaussian components, the

Babble Noise, 10dB SNR			
	1 st./cmp.	4 st./cmp.	8 st./cmp.
MAX	15.10%	14.51%	9.68 %
Δ WER		-3.91 %	- 36.0%
SUM	15.10%	14.81%	9.53%
Δ WER		-1.92%	-37.0%
Roller Coaster Noise, 10dB SNR			
	1 st./cmp.	4 st./cmp.	8 st./cmp.
MAX	6.68%	6.39%	6.50 %
Δ WER		-4.34 %	- 2.69%
SUM	6.68%	6.50%	6.46%
Δ WER		-2.69%	-3.29%

Table 1. Word error rate for Babble noise (upper half) and Roller Coaster noise (lower half) at 10db SNR, for maximization over noise states (MAX) and marginalizing over noise states (SUM), and different number of component/state noise models

resulting distributions had 20, 80 and 160 components, for the 1, 4 and 8 component cases, respectively.

For Babble noise, the word error rate was 31.09% when clean speech acoustic models were used (i.e. mismatched training and deployment conditions). When the system was trained and tested on speech corrupted by babble noise (i.e. matched training and deployment conditions), the error rate dropped to 8.56%.

The upper half of Table 1 shows the effect of using 1, 4 and 8 component/state models for Babble noise. The word error rate drops by 1.92% when 4 states are used. When 8 component noise models are used, the word error rate drops by 37% percent (from 15.1% to 9.53%). This is a significant drop in word error rate, due to more accurate modeling of the noise process. Comparing the maximization method Eq.(6) to the method of marginalizing over noise states Eq.(7) shows that the two methods perform similarly well.

For Roller-Coaster noise, the word error rate was 10.04% for the mismatched condition and 6.31% for the matched condition. The lower half of Table 1 shows the effect of using 1, 4 and 8 component/state models for Roller Coaster noise. In this case there is a drop in error rate when more accurate noise models are used. However, the word error rate for a single component is close to the matched condition, and the advantage of more accurate noise models is not as significant. Again, SUM and MAX perform similarly well.

5. DISCUSSION

We have shown that the use of more accurate, multiple component/state noise model improves recognition accuracy of a large vocabulary speech recognizer when the input speech signal is corrupted by additive non-stationary noise.

We expect that introducing the dynamic links into the noise model will improve word error rate for noise types that have considerable temporal structure, such as roller coaster noise.

Our experiments also indicate that the the two decoding methods, i.e. MAX and SUM, perform similarly well. However, finding the optimal noise state sequence as opposed to marginalizing over noise states results in different decoding algorithms with different complexity. Future work will asses this difference in the context of a large vocabulary speech recognition system.

6. REFERENCES

[1] A.P. Varga and R.K. Moore, “Hidden markov model decompostion of speech and noise,” *proceedings ICASSP*, pp. 845–848, 1990.

[2] G. Zweig and S. Russell, “Speech recognition with dynamic baysian networks,” *Fifteenth National Conference on Artificial Intelligence,AAAI’98*, pp. 173–180, 1998.

[3] Y. Ephraim, “Statistical-model-based speech enhancement systems,” *Proceedings of the IEEE*, vol. 80, no. 2, pp. 1526 – 1554, October 1992.

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

[5] P.J. Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, April 1996.

[6] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Reconition*, Klewer Academic Publishers, 1992.

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

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

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

[10] X.D. Huang, A. Acero, and H. Hon, *Spoken Language Processing*, Prentice Hall, 2000.