MULTI-BAND AND ADAPTATION APPROACHES TO ROBUST SPEECH RECOGNITION

Sangita Tibrewala[1] *and Hynek Hermansky*[1,2]

[1]Oregon Graduate Institute of Science and Technology, Portland, Oregon, USA.
[2]International Computer Science Institute, Berkeley, California, USA.
Email: *sangita,hynek@ee.ogi.edu*

ABSTRACT

In this paper we present two approaches to deal with degradation of automatic speech recognizers due to acoustic mismatch in training and testing environments. The first approach is based on the multi-band approach to automatic speech recognition (ASR). This approach is shown to be inherently robust to frequency selective degradation. In the second approach, we present a conceptually simple unsupervised feature adaptation technique, based on recursive estimation of means and variances of the cepstral parameters to compensate for the noise effects. Both techniques yield significant reduction in error rates.

1. INTRODUCTION

Automatic speech recognizers exhibit rapid degradation in performance when there is a mismatch between training and testing acoustic environments. There are various sources that can cause acoustic distortion, e.g. presence of additive environmental noise such as machinery, background speakers etc. as well as convolutive distortions due to use of different communication channels such as different microphones, telephone channels and reverberation.

One approach towards addressing this problem, is to understand the robust mechanism of the human speech recognition system and try to incorporate it into the ASR model. Fletcher's work [2] on articulatory index suggests that in humans, the linguistic message gets decoded independently in different frequency sub-bands and the final decoding is based on merging the decisions from the sub-bands. This supports the notion that if some frequency sub-band carries unreliable information (possibly due to noise) it could be de-emphasized in the final merging. Our first approach towards robust ASR, the multi-band approach [9, 3] is motivated by this idea. We show that this approach is robust towards noise which selectively corrupts only a few regions of the frequency spectrum.

Another approach towards addressing the problem, is to characterize the adverse effect of the interfering noise. Additive and convolutive noise is known [11] to cause a shift in means and change in variances of the log-spectral components (cepstral coefficients). The cepstral mean subtraction (CMS) technique as well as the RASTA [8] technique were introduced to compensate for the changes in means of the parameters at the feature level. Though several dynamic adaptation schemes ([10, 11, 1]) are being used to adapt both the means and variances of features and acoustic models to changing environmental conditions, most of these approaches are based on maximum likelihood estimation (MLE) methods in the HMM framework. Since in a connectionist model, input parameters are usually standardized by subtracting the mean and dividing by the standard deviation for each parameter, to make the training of the multi-layer perceptron (MLP) faster, a conceptually simple compensation technique can be applied.

In our second approach we present a simple recursive estimation of means and variances for each incoming test frame. Similar techniques have been used by Cook et al. [5] and Gauvain et al. [6].

2. EXPERIMENTAL SETUP

Our experiments are based on the telephone-quality Bellcore isolated digits database. The database has a 13-word vocabulary consisting of eleven digits (including 'oh') and two control words ('yes', 'no'). The training set consists of 150 speakers and 50 speakers comprise the test set. Each speaker has uttered the vocabulary once. The classifiers used are the phoneme-based HMM/MLP hybrid classifiers [4].

3. MULTI-BAND ASR APPROACH

The multi-band approach [9, 3] to automatic speech recognition is illustrated in Fig. 1. In our multi-band model the frequency spectrum is divided into 7 sub-

bands, each sub-band comprising of about two critical bands. The features in each sub-band are the cepstra obtained from the all-pole modeling of the PLP [7] cube-root compressed and equalized (equal-loudness equalization) critical band energies. Independent probability estimation for each class (phoneme) is conducted in each sub-band using a MLP. The class-conditional log-likelihoods for each frame from each sub-band classifier are then non-linearly merged using another MLP to obtain the merged probability estimates for each class. This merging MLP is trained on the sub-band log-likelihoods of the training data.

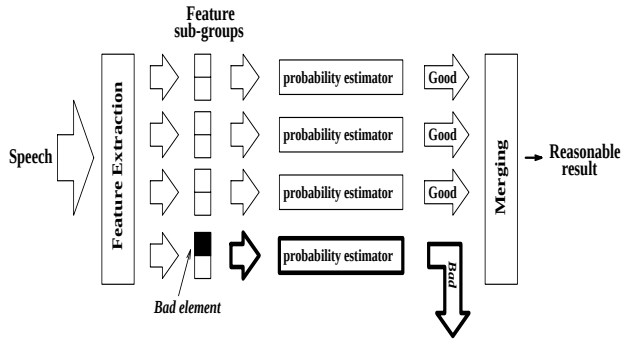


Figure 1: Multi-band Model

3.1. Experimental Results

It has been shown [9, 12] that for well-matched training and test conditions, there is no loss of performance from the multi-band approach. To test the performance of the model with real-noise we used some of the noise samples from the NOISEX-92 database, viz. destroyer engine, factory2, pink, white,volvo, babble and high-frequency radio channel noise. Each noise was added to the test speech data after being scaled so that the performance of the conventional full-band ASR system degraded from baseline error of 2.5% to about 25%.

Additive noise	Conventional	Multi-band
clean (no noise)	2.5	1.2
destroyer-engine	26.6	18.5
factory noise	26.2	8.2
pink	24.3	11.7
babble	24.6	10.3
volvo	24.6	6.2
volvo (lab recorded)	25.2	15.4
white	24.8	34.6
high-frequency radio	25.8	36.9

Table 1: Word error (%) on Bellcore isolated digits

Noise	B1	B2	B3	B4	B5	B6	B7
clean	45.5	14.9	18.6	14.0	11.8	18.0	25.7
engine	93.1	34.5	43.4	53.1	65.4	82.0	81.4
factory	94.3	51.7	25.1	18.9	13.2	19.4	27.4
pink	94.9	53.1	40.0	34.6	25.7	32.3	66.5
babble	93.1	48.9	50.0	28.6	15.8	20.3	37.1
volvo	93.8	55.5	18.8	14.0	12.0	18.3	25.5
volvo2	78.6	64.6	59.5	56.0	37.7	44.8	47.5
white	94.3	42.2	52.5	58.0	52.8	71.7	87.2
radio	86.8	48.0	58.6	73.4	65.2	86.9	91.4

Table 2: Word error (%) in the 7 subbands (B1 to B7 refer to the 7 subbands respectively, engine refers to destroyer-engine noise, volvo2 refers to the lab recoded volvo noise and radio refers to the high-frequency radio noise).

From the results (Table 1) we see that the multi-band approach improves the performance by 50% on average for the first six noise cases as compared to the conventional system. These noise cases are such that they corrupt some sub-bands much more than the other sub-bands as shown in Table 2 (e.g., for the case of factory noise, sub-bands 1 and 2 (B1,B2) are more corrupted than the other 5 sub-bands). In these cases a further improvement in performance can be achieved by leaving out the highly corrupted sub-bands from the merging process.

The white noise and high-frequency radio channel noise results in degradation of all sub-bands (last 2 rows of Table 2) and hence the multi-band approach is found to be ineffective for these noise cases.

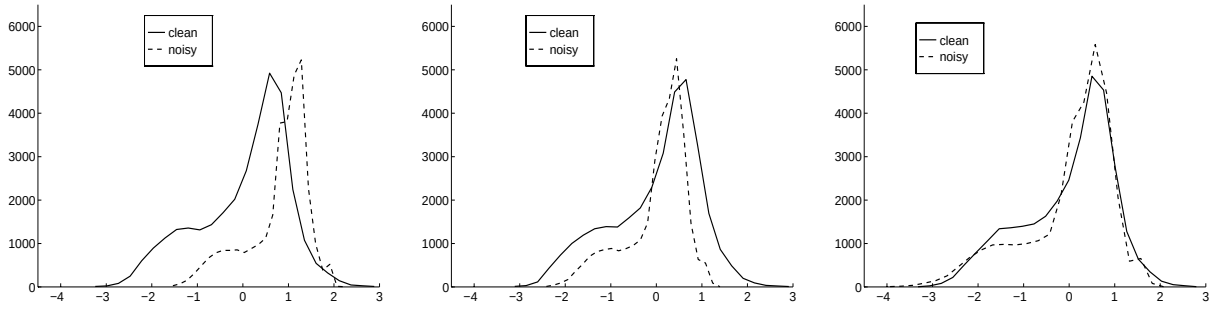
The results in Table 1 support the notion of inherent noise robustness of the multi-band approach to frequency-selective degradation. Also, it is noteworthy that the improvement in performance does not require any additional processing.

4. FEATURE ADAPTATION

In connectionist model (HMM/MLP hybrid framework) for ASR the input feature vector is usually scaled to zero mean and unity variance as

$$\bar{x}_k(t) = \frac{x_k(t) - \mu_k^{train}}{\sigma_k^{train}} \quad (1)$$

where $x_k(t)$ is the kth input parameter at time (frame) t, μ_k^{train} and σ_k^{train} are the mean and the standard deviation of the kth input parameter computed over the entire training data. This normalization is generally carried out to make training of the MLP faster and avoid problems of getting stuck in a local minima.



(a) no adaptation (b) mean adaptation (c) mean and variance adaptation

Figure 2: Histogram of a cepstral coefficient of clean data and data contaminated with white noise

In the presence of acoustic mismatch during training and testing conditions, the observed signal y can be modeled as $y = (x + n) * h$ where x is the input signal, n is the additive noise and h is the convolutive linear channel distortion. Based on this model, it has been shown [11] that the effects of the environment show up as a shift in means and change in variance of the input cepstral parameters. These effects are illustrated in Fig. 2(a) which compares the histogram of a cepstral coefficient from matched (clean) condition to that when additive white noise is present. Both these parameters have been normalized as in Eq. 1 using the mean and variance computed over the training data.

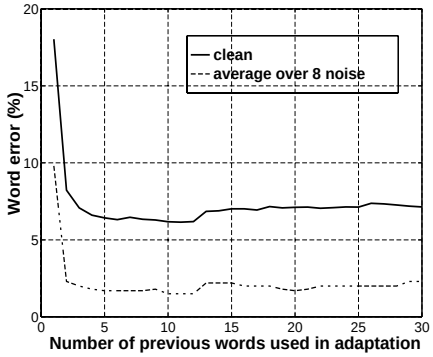


Figure 3: Average reduction in word error (%) using adaptation as compared to the case with no adaptation

In order to compensate for the change in mean and variance the new values can be re-estimated over the entire current utterance. However for real-time application the delay introduced in processing by waiting till the end of the utterance is unacceptable. This delay can be avoided by estimating the mean and variances over the past utterances (words in our case). Fig. 3 represents the effect of using increasing number of words to estimate the mean and variance. It is seen that using as few as 5 words results in stable recognition errors for both clean and noisy conditions.

The above technique requires buffering of feature vectors corresponding to 5 words. Also it is computationally inefficient. In order to reduce the memory requirements we recursively estimate the mean and variances for each incoming frame using an integrator as

$$\mu_k(t) = \alpha\mu_k(t-1) + (1-\alpha)x_k(t) \quad (2)$$

$$s_k(t) = \alpha s_k(t-1) + (1-\alpha)x_k^2(t) \quad (3)$$

$$\sigma_k^2(t) = s_k(t) - \mu_k^2(t) \quad (4)$$

$$\bar{x}_k(t) = \frac{x_k(t) - \mu_k(t)}{\sigma_k(t)} \quad (5)$$

where $\mu_k(t)$ is the mean of the parameter x_k at time frame t, α is the forgetting factor used to forget the effect of past frames. The average sum of squares s_k is estimated in a similar manner as shown in Eq 3. Finally the variance σ_k^2 is estimated as in Eq. 4 followed by the normalization in Eq. 5. Series of experiments showed that $\alpha = 0.995$ results in stable estimates.

Fig. 2(b) represents the effect on the distribution of the cepstral coefficients in Fig. 2(a) when only mean compensation is carried out according to Eq. 2 and the variance is that of the training (σ_k^{train}). Fig. 2(c) represents the effect when both mean and variance are compensated. It supports the fact that both the mean and variance need to be adapted to make the distribution under noisy condition similar to that of the clean case.

4.1. Experimental results

Table 3 shows the word recognition error without adaptation (Eq. 1) and using adaptation of mean and vari-

ance (Eqs. 2- 5). The results indicate that for matched training and test conditions the simple adaptation technique does not affect performance but under noise conditions it results in approximately 75% reduction in error rates on an average.

Additive noise	No adaptation	Adpatation
clean (no noise)	2.5	1.8
destroyer-engine	26.6	5.2
factory noise	26.2	4.2
pink	24.3	3.7
babble	24.6	5.7
volvo	24.6	3.5
volvo (lab recorded)	25.2	10.0
white	24.8	9.7
high-frequency radio	25.8	9.2

Table 3: Word error (%) on Bellcore isolated digits

4.1.1. Comparison of mean-only and mean and variance adaptation

In order to further investigate the importance of mean and variance adaptation over mean-only adaptation as suggested by Figs. 2(b), 2(c) we carried out a series of experiments with the two techniques. The techniques were tested with the 8 noise cases as mentioned in Section 3.1. The noise was added at different signal-to-noise (SNR) ratio ranging from 20dB to -5dB.

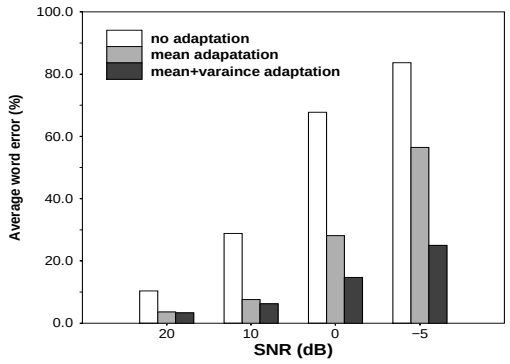


Figure 4: Average word error rates of the 8 noise cases for different SNR and adaptation schemes

From Fig. 4 it is seen that both the adaptation techniques result in significant reduction in error rates over the no adaptation case. However, the additional variance adaptation results in further significant reduction in error rates as the SNR decreases.

5. CONCLUSION

In this paper we have described two approaches to robust ASR under conditions of environmental degradation. The adaptation approach to noise compensation is effective under the assumption that the current noise characteristics are similar to that over the period of estimation. The multi-band approach on the other hand works reasonably well for frequency selective degradation without any adaptation.

6. ACKNOWLEDGMENTS

This work has been partially supported by the Center for Language and Speech Processing at the Johns Hopkins University, NSF/ARPA (IRI-9314959) and DoD (MDA-904-94-C-6169).

7. REFERENCES

[1] A. Acero and R.M. Stern. Environmental robustness in automatic speech recognition. *Proc. ICASSP-90*, 1:849–852, 1990.

[2] J.B. Allen. How do humans process and recognize speech? *IEEE Trans. on Speech and Audio Processing*, 2(4):567–577, 1994.

[3] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and re-combination of partial frequency bands. *Proc. ICSLP 96*, 1:426–429, 1996.

[4] H. Bourlard and N. Morgan. *Connectionist Speech Recognition — A Hybrid Approach,*. Kluwer Academic Publishers, Boston, 1994.

[5] G.D. Cook, J.D. Christie, P.R. Clarkson, M.M. Hochberg, B.T. Logan, and A.J. Robinson. Real-time recognition of broadcast radio speech. *Proc. ICASSP-96*, 1:141–144, September 1996.

[6] J.L. Gauvain, J.J. Gangolf, and L. Lamel. Speech recognition for an information kiosk. *Proc. ICLSP-96*, 2:849–852, 1996.

[7] H. Hermansky. Pereceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[8] H. Hermansky. Recognition of speech in additive and convolutive noise based on RASTA spectral processing. *Proc. ICASSP-93*, 1:141–144, May 1995.

[9] H. Hermansky, S. Tibrewala, and M. Pavel. Towards ASR on partially corrupted speech. *Proc. ICSLP 96*, 1:462–465, 1996.

[10] C.H. Lee and J.L. Gauvain. *Bayesian Adaptive Learning and MAP estimation of HMM*. Kluwer Academic Publishers, Boston, 1993.

[11] P.J. Moreno, B. Raj, E. Gouvea, and R.M. Stern. Multivariate-gaussian-based cepstral normalization for robust speech recognition. *Proc. ICASSP-95*, 1:141–144, May 1995.

[12] S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. *Proc. ICASSP 97*, II:1255–1258, 1997.