Proceedings of the Institute of Acoustics

NOISE-ADAPTIVE HIDDEN MARKOV MODELS

S. V Vaseghi, B. P Milner
School of Information Systems, University of East Anglia, Norwich.

ABSTRACT

This paper considers the problems of adaption of HMM models to ambient noise. The objective of noise adaption is to obtain the model parameters that would be obtained if the training and operating environments were matched. Noise adapted HMMs can potentially approach the performance of HMMs trained and operated in matched environments. The HMM parameters affected by noise are state transition and observation densities. We focus on methods of adaptation of state observation densities to noise.

1. INTRODUCTION

Performance of HMM speech recognition systems, trained on noise-free examples, degrades severely in the presence of noise. HMMs perform well when the ambient noise in operating and training environments are the same, but performance deteriorates when the environments are different. For most applications it is impractical to match the environments because the operating noise changes with time and space, and it is necessary to employ a noise compensation scheme. Noise compensation methods for speech recognition can be classified into three broad categories. In one category noise filtering methods such as spectral subtraction or Wiener filtering are used to take out an estimate of the noise from noisy speech observation parameters [Lim & Oppenheim 1978] [Porter & Boll 1984] [Ephraim, Malah, Juang 1989]. In the second category the focus is on the development of distance measures which are robust to noise [Mansour, Juang 1989], [Carlson, Clements 1992]. In the third category noise-free speech models are adapted to noise, and the noisy signal is left unmodified [Roe 1987] [Nadas, Nahamoo, Picheny 1989], [Varga, Moore 1990] [Gales, Young 1992].

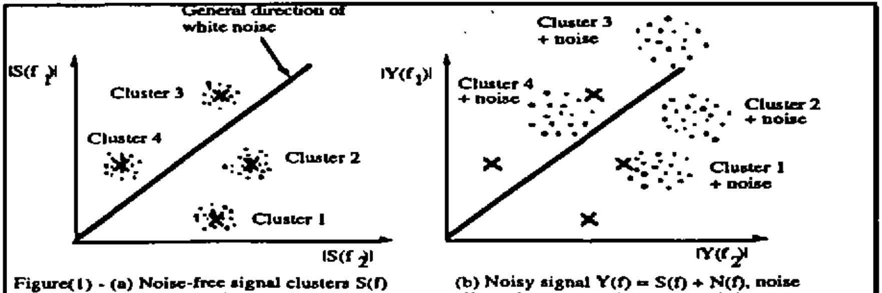
This paper considers methods of adaption of noise-free HMM parameters to ambient noise. In recent years several methods for noise adaptive recognition systems were proposed. Roe developed a noise adaptive code book in which autocorrelation coefficients are adapted to reflect changes due to addition of noise and lombard effects. Nadas *et al.* describe a system for adaption of a probabilistic mixture model of speech. The method is based on the observation that at any time the energy in a frequency band is dominated by either the signal or the noise. Varga and Moore describe an HMM based signal and noise decomposition in which the signal and noise are modelled by separate HMM models. This is particularly useful if the noise evolves in time in such a way that it has to be modelled by a set of 'statistical states'. Gales and Young developed an improved method to HMM based speech and noise decomposition. Noise obscures speech events and affects the way people articulate a word. In noise, people speak louder and there are increases in duration, pitch and higher frequency energy of speech [Pisoni *et al* 1985]. The noise-induced stress (also known as lombard effects) can be as harmful to recognition as the noise itself, however in this paper the focus is on the effects of additive noise. In the following we consider the effects of noise on signal observation space and describe methods for adaption of state observation parameters of HMM models.

2. A SIGNAL AND NOISE LINEAR SPACE MODEL

Speech spectral features may be viewed as points in a multi-dimensional space fig(1). Repetitions of an utterance forms clusters of points whose centroids represents the average characteristics of the utterance. Each cluster may be modelled by a multivariate Gaussian density $\mathcal{N}(\mathbf{x}_t, \mu, \mathbf{C})$. An observation vector, $\mathbf{x}_t$, can be classified using the maximum likelihood criterion as

$$\text{Label}(\mathbf{x}_t) = \underset{k}{\text{argmax}}\ \{p_k \mathcal{N}(\mathbf{x}_t, \mu_k, \mathbf{C}_k)\} \qquad (1)$$

where p_k, the cluster prior probability, is a measure of the fraction of training vectors associated with cluster k.



Figure(1) - (a) Noise-free signal clusters S(f) and centroids denoted by X (b) Noisy signal Y(f) = S(f) + N(f), noise effects the mean and variance of clusters. Using noise-free models to classify noisy data results in errors due to mismatching. In this case cluster 4 would be classified as 3 and cluster 1 as 2.

In this signal space the effects of noise on a signal cluster is two fold; (a) the centroid of the noisy signal cluster is moved in the general direction of noise, and (b) the variance of signal cluster increases fig(1.b). Both effects result in a decrease in intercluster distance and an increase in recognition error. The objective of noise compensation is to correct the discrepancy between the mean of the noisy signal and noise-free models. This can be done in two ways : either the noisy signal is filtered in order to move the centroid of noisy signal cluster towards the noise-free models as in spectral subtraction, or the noise-free models are adapted in order to move the centroids of the models towards the noisy signal.

3. EFFECTS OF NOISE ON OBSERVATION LIKELIHOOD USING HMMS TRAINED ON NOISE-FREE EXAMPLES

In a HMM the probability that a speech utterance, **y**, is associated with a model λ_S, summed over all state sequences, **q**, and mixture sequences, **h**, is given by

$$p_{\lambda_s}(\mathbf{y}) = \sum_q \sum_h p_{\lambda_s}(\mathbf{q})\; p_{\lambda_s}(\mathbf{h}|\mathbf{q})\; p_{\lambda_s}(\mathbf{y}|\mathbf{h},\mathbf{q}) \qquad (2)$$

The state observation probability, $p_{\lambda_s}(\mathbf{y} \mid \mathbf{h}, \mathbf{q})$, commonly modelled by a mixture Gaussian density, is the function through which the signal and noise influence the likelihood calculations. In the following the effects of noise on observation likelihood *for the case in which the likelihood function itself is trained on noise-free signal* is considered. The noisy signal spectral energy, Y(ω), is the sum of signal, S(ω), and noise, N(ω)

$$Y(\omega) = S(\omega) + N(\omega) \qquad (3)$$

The pdf of noise-free signal magnitude spectrum is assumed as

$$p_{s(\omega)}(s(\omega)) = \frac{1}{\sqrt{2\pi\sigma_s(\omega)}} \exp\left(\frac{[s(\omega)-\mu_s(\omega)]^2}{2\sigma_s^2(\omega)}\right) \qquad (4)$$

where $\mu_s(\omega)$ is the mean magnitude spectrum at ω, and $\sigma^2{}_{s(\omega)}$ is the variance of sample spectrum about the mean. The pdf of noisy signal based on noise-free signal statistics is

$$p_{s(\omega)}(Y(\omega)) = \frac{1}{\sqrt{2\pi}\sigma_s(\omega)} \exp\left(\frac{[s(\omega)+N(\omega)-\mu_s(\omega)]^2}{2\sigma_s^2(\omega)}\right) \qquad (5)$$

Taking expectation of log likelihood and substituting $N(\omega) = N'(\omega) + \mu_N(\omega)$, where $N'(\omega)$ is a zero mean noise magnitude spectrum gives

$$E\{\log[p_{s(\omega)}(Y(\omega))]\} = \log[p_{s(\omega)}(s(\omega))] - \frac{\sigma_N^2(\omega)+\mu_N^2(\omega)}{2\sigma_s^2(\omega)} \qquad (6)$$

Eq(6) shows that the decrease in probability depends on the noise power spectrum $\mu_N(\omega)$, the variance of noise sample spectrum $\sigma_N(\omega)$, and variance of signal sample spectrum $\sigma_s(\omega)$. In spectral subtraction, the term $\mu_N(\omega)$ may be subtracted out, whereas the term $\sigma_N(\omega)$ constitutes an irrevocable degradation. In a probabilistic formulation of the noise problem the effects of both the mean and the variance of noise on speech distribution may be taken into account.

4. ADAPTIVE HIDDEN MARKOV MODELS

The main parameters of an HMM are the state transition and observation densities. State transitions provide a mechanism for modelling variations in articulation rates. State observation model the feature space associated with each state. we consider adaption of state observation parameters, as these are most affected by noise and are easier to adapt. In the following we assume the observation parameters for the ith state consists of M mixture Gaussian densities denoted by the parameter set ($p_{ik}, \mu_{ik}, \Sigma_{ik}, k = 1, \ldots, M$). We make the assumption that the main effect of noise on each component of the mixture is a shift of the mean vector, μ_{ik}, and an increase of the variance, Σ_{ik}, and ignore the effect of noise-induced cluster redistribution on the parameter p_{ik}.

4.1 Adaption of State Observation Using Linear Speech Features

The computational complexity and ease of adaption of state observation probabilities depends on the choice of speech feature vectors. The adaption of state observation is relatively easier for linear feature vectors such as filter bank outputs or correlation coefficients compared to nonlinear features such as log spectral energy or cepstrum. For linear features the distribution of noisy signal $\mathcal{N}(\mu_y, \Sigma_y)$ is the sum of the distributions of the signal $\mathcal{N}(\mu_s, \Sigma_s)$ and the noise $\mathcal{N}(\mu_n, \Sigma_n)$ expressed as

$$\mu_y = SNR^2 \mu_s + \mu_n \qquad (7)$$

$$\Sigma_y = SNR^2 \Sigma_s + \Sigma_n \qquad (8)$$

where SNR is the signal to noise ratio. The drawbacks of using linear spectral features are :
(a) a relatively greater number of spectral features are required compared to cepstral features, and (b) the recognition results using spectral features are not as good as the results based on cepstral features.

4.2 Adaption of State Observation using Log Spectral Energy and Cepstral Features

A problem in adaption of models that use log spectral energy or cepstral features is that due to logarithmic operation the distribution of noisy features can not be obtained by simple addition of the distributions of signal and noise. Nadas *et al* based their adaption of log spectral energy prototypes on the observation that at any time speech spectral energy is dominated by either the signal or the noise. They describe the noisy speech density in terms of the densities of signal and noise as

$$h(z)=f(z)G(z)+F(z)g(z) \qquad (9)$$

where h(z) the observation density is expressed in terms of the signal density f(z) and the noise density g(z). F(z) and G(z) are cumulative density functions. This scheme can also be incorporated into a noise-adaptive HMM model. A drawback with this method is its reliance on the use of log spectral energy features which are not as efficient as cepstral features.

Most speech recognition systems use cepstral feature vectors for compactness, good invariance and supposedly small correlation among feature vector elements. Cepstral parameters are cosine transform of log spectral parameters. Of interest are the mappings between the statistics of spectral and cepstral parameters. Such a mapping would allow us to adapt in linear spectral domain and transform the results into cepstral domain. When cepstral parameters, c_i, are modelled by a Gaussian density, then the distribution of the log spectral energy, x_i, is also Gaussian, and the distribution of spectral energy, $s_i = e^{x_i}$, is lognormal

$$f(s_i=e^{x_i})=\frac{1}{\sqrt{2\pi}\,\sigma_{x_i}s_i}e^{\left(-(\ln s_i-\mu_{x_i})^2/2\sigma_{xi}^2\right)} \qquad (10)$$

The equations relating the mean and variance of a Gaussian distributed log spectrum (μ^l, Σ^l) and those of the log normal distributed spectrum (μ, Σ) [Gales and Young] are :

$$\mu_i^l=\log(\mu_i)-\frac{1}{2}\Sigma_{ii} \qquad (11)$$

$$\Sigma_{ij}^l=\log\left(1+\Sigma_{ij}/\mu_i\mu_j\right) \qquad (12)$$

The mean vector, μ^c, and covariance matrix, Σ^c, of cepstrum and log spectrum are related by the cosine transform C as

$$\mu^c = C\,\mu^l \qquad (13)$$

$$\Sigma^c_{ij} = C\,\Sigma^l_{ij}\,C^T \qquad (14)$$

The effects of noise on the mean and variance of clean model may be calculated in the linear spectral energy domain using equations (7) (8) , and the results transformed into cepstral domain.

Another form of commonly used speech features are the LPC-cepstral parameters which are obtained from linear predictor coefficients using the recursion

$$c_k=-a_k-\sum_{n=1}^{k-1}\left(1-\tfrac{n}{k}\right)a_n c_{k-n}$$

where a_k and c_k are the linear predictor and cepstral coefficients respectively. The LPC vector, $\mathbf{a}$, is derived from the autocorrelation vector, $\mathbf{r}_s$, and the effect of noise on autocorrelation vector is additive i.e. $\mathbf{r}_y= \mathbf{r}_s +\mathbf{r}_n$. A method of adaption of the mean cepstral vector, $\mathbf{c}_s$, is to add the effect of noise to the corresponding autocorrelation vector $\mathbf{r}_s$, and then transform the result into the noise adapted cepstral vector $\mathbf{c}_y$. This method compensates for noise-induced shifts in cepstral means but does not provide for the effects of noise on signal variance.

4.3 Cepstral Compensation

HMM models can be adapted to noise by addition of a compensation cepstral vector to the mean of each mixture component. The cepstral compensation vector may be word based or state based. A word based compensation vector may be obtained as the average difference vector between the clean and noisy cepstral vectors

$$c_{comp}=\frac{1}{N}\left(\sum_N c_{noisy}-\sum_N c_{clean}\right)$$

A state-based compensation vector can be derived in a similar manner as the average difference between the clean and noisy vectors associated with each state. The direction and magnitude of compensation vector depends on SNR. The SNR varies globally across a word as a function of the average signal and noise energies, and also locally with the changes in energy across an utterance. The problem of local variations in SNR across the segments of a word can be dealt with by using state based compensation vectors. However the global changes in SNR makes the use of this compensation scheme for varying SNR impractical.

5. RESULTS

Experiments are based on a data set of spoken English alphabet. For each of the 26 letters, the HMM model was trained using 52 speakers with 3 utterances per speaker. The test data set consisted of a similar number of utterances from a different set of speakers. The feature vector (c, δc, δlogE), consists of 25 features comprising of 12 cepstral coefficients, 12 differential cepstral coefficients and differential log energy. The baseline HMM recogniser chosen is an 8-state left-right HMM without skip-state transition, with 7 mixture multivariate Gaussian density per state and diagonal covariance matrices. The recognition rate of this HMM system in a noise free environment is about 87 %. The relatively low accuracy is due to several confusable subsets in the alphabet vocabulary. The following experiments were performed and the results are tabulated in table-1.

Recognition in Matched Conditions The results for Matched conditions provide an upper bound on recognition accuracy of noise compensated HMMs, as it is unlikely that a noise adapted HMM can outperform one trained and operated in matched conditions.

Recognition in Noise Using Noise-Free Models : A set of HMM models were trained on noise-free examples and tested on examples which had white Gaussian noise added. The recognition accuracy deteriorates rapidly as the SNR falls below 30 dB.

Spectral Subtraction - A non-linear fraction , $\alpha(SNR)$, of an estimate of noise spectral template was subtracted from noisy signal spectrum. The choice of $\alpha(SNR)$ has a significant effect on recognition rate and it was empirically chosen to produce optimal results.

Noise-Adaptive HMMs- In these experiments noisy signal was left unmodified and noise-free state observation statistics were adapted to noise. The adaption of mean and variance were performed in linear spectral energy domain and the results were then transformed into cepstral statistics. As table-1. indicates noise adaptive HMMs outperform front end compensation methods such as spectral subtraction, but the performance is not as good as than those obtained under matched conditions.

SNR	Recognition Accuracy (%)			
(dB)	Matched	Clean	SS	AHMM
30	84.3	79.9	82.9	83.1
20	77.0	52.9	64.5	68.3
15	72.6	29.7	45.7	49.2
10	65.0	9.6	23.6	35.3
5	51.3	4.7	9.2	22.4
0	48.4	4.4	8.0	16.0

Table-1 : Recognition accuracy Vs SNR for HMMs. KEY : AHMM = Adaptive HMMs,
SS = Spectral Subtraction

6. CONCLUSION

The objective of adaption of HMMs is to obtain the statistical parameters and hence performance that would be obtained if the training and operating environments were matched. Several schemes for adaption of observation parameters of an HMM were considered. The main drawback of noise-adaptive models is the relatively large computational increase involved in adaptation of every state of each model.

REFERENCES

Carlson, B. A., Clements, M. A. (1992),"Speech Recognition in Noise Using a Projection-Based Likelihood Measure for Mixture Density HMM's", IEEE Proc. ICASSP-92, Pages I-237-I240, San Francisco.

Ephraim, Y., Malah D., Juang B. H. , (1989), "On the Application of Hidden Markov Models for Enhancing Noisy Speech", IEEE Trans. ASSP, vol. 37,pages 1846-1856, December .

Gales,M.J.F., Young, S.,(1992),"An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise", IEEE Proc., ICASSP_92, pages I-223-I-226, San Francisco.

Lim, J. S., and Oppenheim, A. V., (1978),"All-pole modelling of degraded speech", IEEE Trans. Acoust., Speech and Signal Proc., vol. ASSP-26, pages 197-210, June .

Mansour, D., Juang, B. (1989),"A Family of Distortion Measure Based Upon Projection Operation For Robust Speech Recognition", IEEE Trans. ASSP, Vol 37, pages 1659-1671, November.

Nadas, A., Nahamoo, D., Pichney, A., (1989),"Speech Recognition Using Noise-Adaptive Prototypes", IEEE Trans. ASSP, vol. 37, No. 10, pages 1495-1503, October.

Pisoni et al,(1985)" Some acoustic-Phonetic Correlates of Speech Produced in Noise", IEEE Proc. ICASSP-85, pages 1581-1584, Florida.

Porter, J. E., Boll, S. F. (1984),"Optimal estimators for spectral restoration of noisy speech", IEEE Proc. ICASSP-84, San Diego, California, pages 18A.2.1.-18A.2.4., March.

Roe, D. B., (1987)," Speech Recognition with a Noise-Adapting Codebook", IEEE Proc. ICASSP-87, pages 1139-1142, Dallas, Texas.

Varga, A.P., Moore, R.K.(1990),"Hidden Markov Model Decomposition of Speech and Noise",IEEE Proc. ICASSP-90, pages 845-848, NewMexico.