Proceedings of the Institute of Acoustics

SPEECH RECOGNITION IN NOISE USING MODEL BASED ADAPTATION STRATEGIES

M. Kadirkamanathan

Speech Research Unit, Defence Research Agency, St. Andrews Rd, Malvern, WR14 3PS, England.

1 INTRODUCTION

The performance of existing speech recognition systems which are designed to operated in low noise background noise environments have been known to degrade quite significantly with increasing noise levels [5]. Among the many techniques proposed for noise robustness in recognisers, those based on adapting speech models trained in one environment to the ambient conditions have shown much success. The Hidden Markov Model (HMM) decomposition recognition[1] and the model combination[3] technique have illustrated good performances in rather extreme conditions. The Klatt noise masking[4] is another algorithm of this type.

Acoustic ambient noise is usually considered to be additive. The sampled signal is the sum of the acoustic speech signal and the acoustic ambient signal. The front-end of most speech recognisers perform a short term spectrum analysis on the sampled signal as a first step. Estimates of the signal energy in various frequency bands are calculated in dB or equivalent energy level measures. Assuming that the cross correlation term between the speech signal and the ambient noise signal is negligible compared with the autocorrelation terms, the energy estimate may be expressed as

$$O_{tk} = 20\log_{10}(10^{x_{tk}/20} + 10^{y_{tk}/20}) \quad (1)$$

$$\approx \max(x_{tk}, y_{tk}), \quad (2)$$

where O_{tk} denotes the output of channel k at time t. x_{tk} and y_{tk} denote the unobservable output of channel k of speech and noise respectively, if the other were not present. A plot of the function is shown as the continuous line in fig. 1.

One good approximation to this function is the max function described by long dashes in the figure. The max function implies that the observed output is the largest of the speech and noise energy estimates. This is why many 'model adaptation to noise' techniques have also been described as 'noise masking'.

The decomposition recognition [1] is an extension of the Viterbi HMM recognition paradigm to incorporate the noise contamination Process. The noisy speech frames are explained as a combination of the speech and the noise frames by a suitable composing function dependent upon the recogniser front-end preprocessing. The model combination technique [3] suitably corrects the speech models for the ambient noise conditions for use in a speech recogniser which otherwise operates in low noise conditions. It assumes a cepstrum transformation front-end and combines speech and noise model estimates to produce noisy speech models. Though principally similar to decomposition recognition, the speech models are adjusted for the conditions instead of modifying the recognition algorithm. Both techniques assume that the speech signal is not directly affected and do not account for the well known 'Lombard' effect.

The choice of the front-end preprocessing is also an issue in such noise robust recognition schemes. Although the cepstrum transformation is known to produce better discriminant vectors for speech in

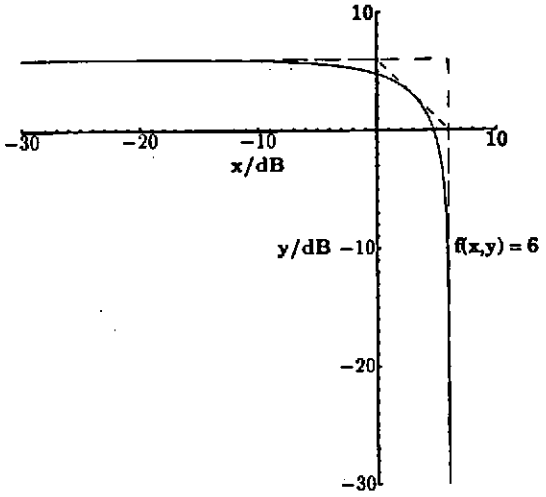

Figure 1: The solid line shows the actual contour, the long dashed line shows the max function and the short dashed one shows the 3-piece approximation.

quiet conditions, it is not known whether the same is true in high levels of noise. The model combination technique is applicable to the spectral energy as well as the cepstrum front-ends. The HMM decomposition however is limited to spectral energy front-ends owing to the difficulty in calculating complicated integrals with the cepstrum front-end.

This article mainly focuses on the scope of the various model adaptation techniques. The performances of the HMM decomposition recognition and the model combination technique are evaluated in a digit recognition experiment. A number pre-recorded noise samples are mixed with clean speech samples for testing the recognizers. These performances are compared with baseline results. The scope of the 'adaptation to noise' techniques in general are also assessed by considering the noise masking of frames corresponding to each phoneme.

2 HMM DECOMPOSITION: MAXIMUM LIKELIHOOD RECOGNITION

The HMM decomposition recognition is the derivation of the maximum likelihood recognition within the decomposition [2]. Representing the ambient noise process also as an HMM, it calculates the most likely state sequences among both the speech states and the noise states given the contaminated speech data. The algorithm is described in [1] and the general framework in [2].

Let $f()$ be the equivalent composing function of a speech frame x_t and a noise frame y_t to produce the observed frame, $O_t = f(x_t, y_t)$. The Viterbi scoring function is derived for the composed speech-noise states as, with the usual notations,

$$\Phi_{t+1}(i,k) = \max_{\substack{1\le j\le N_1 \\ 1\le l\le N_2}} \Phi_t(j,l) a1_{ji} a2_{lk} b_{ik}(O_{t+1}) \qquad (3)$$

where the observation probability function of a joint state

$$\tilde{b}_{ik}(O_t) = \oint_{C_t} b1_i(x) b2_k(y), \tag{4}$$

where $C_t \equiv f(x,y) = O_t$. In general the noise model may consist of many states but for stationary background noises one is sufficient.

The front-end of the speech recogniser employs a 27 channel filterbank which produces log energy outputs in dB every 10ms. The composition function is therefore as stated in eqn. 1. In order to reduce computations approximations of this function are used in the recognition system. Varga and Moore [1] use the max function as an approximation to $f()$. The max function deviates from the actual function by about 4.2dB at the break point as illustrated in Fig. 1. This is reduced to about 1.2dB with a three piece approximation as illustrated in fig. 1. in short dashed lines.

2.1 The max approximation

The masking function approximation is a popular approximation for the combination of speech frames and noise frames.

$$f(x,y) \equiv \max(x,y). \tag{5}$$

The joint observation probability function for a pair of gaussian states with diagonal co-variances is therefore

$$\tilde{b}_{ij}(O_t) = \prod_{k=1}^{K} \Big[g(\mu_{ik},\sigma_{ik},O_{tk})\, e(\mu_{jk},\sigma_{jk},O_{tk}) + e(\mu_{ik},\sigma_{ik},O_{tk})\, g(\mu_{jk},\sigma_{jk},O_{tk}) \Big], \tag{6}$$

where $g()$ and $e()$ denote the gaussian and the error functions respectively.

2.2 The three piece approximation

The three piece approximation function is a closer fit to the actual combination function. It is defined as

$$f(x,y) \equiv \begin{cases} x & x \geq y+6 \\ \frac{x+y-6}{2} & y-6 < x \leq y+6 \\ y & y > x+6 \end{cases} \tag{7}$$

The joint observation probability function is therefore

$$\begin{aligned} \tilde{b}_{ij}(O_t) = \prod_{k=1}^{K} \Big[& g(\mu_{ik},\sigma_{ik},O_{tk}) e(\mu_{jk},\sigma_{jk},O_{tk}-6) + e(\mu_{ik},\sigma_{ik},O_{tk}-6)\, g(\mu_{jk},\sigma_{jk},O_{tk}) \\ & + g(\mu_{ik}+\mu_{jk},\sigma_{ik}+\sigma_{jk},O_{tk}).\big[e(\mu^*_{ijk},\sigma^*_{ijk},O_{tk}) - e(\mu^*_{ijk},\sigma^*_{ijk},O_{tk}-6) \big] \Big] \end{aligned} \tag{8}$$

where

$$\mu^*_{ijk} = \frac{\sigma^2_{jk}\mu_{ik} + \sigma^2_{ik}(O_{tk}-\mu_{jk})}{\sigma^2_{ik}+\sigma^2_{jk}} \tag{9}$$

$$\sigma^{*2}_{ijk} = \frac{\sigma^2_{ik}\,\sigma^2_{jk}}{\sigma^2_{ik}+\sigma^2_{jk}} \tag{10}$$

3 CEPSTRUM MODEL COMBINATION

Given the speech and noise models with cepstrum coefficients, the distributions are suitably combined in the filter-bank output energy domain, and then transformed back to the cepstrum domain to obtain the corrected models. [3]. The main attraction of this method is the use of cepstrum data which are known to offer better discrimination for clean speech. The method is outlined as follows.

1. If the speech models are based on cepstrum data apply the inverse cepstrum transform on both the mean vector and the covariance matrix.
2. Compute the mean vector and the covariance matrix of the lognormal distribution with appropriate scaling. These estimates will be in the absolute energy domain.
3. Do steps 1 and 2 on the noise model.
4. Add the two estimates together with appropriate gains if necessary.
5. Again assuming a good lognormal fit to the mixed distribution, calculate the mean vector and the covariance matrix for the log energy domain.
6. Apply cepstrum transformation is required to obtain cepstrum distributions of noisy speech. Use the diagonal covariances only if desired.

4 EXPERIMENT

The recognition performances were obtained for the HMM decomposition with the masking function and the three piece approximations. The model combination algorithm was also tested in the following cases:

1. Cepstrum speech models and cepstrum noise models.
2. Cepstrum speech models and filterbank noise models.
3. Filterbank speech models with diagonal covariances and filterbank noise models.

It seems appropriate to use filterbank noise models as there is no good reason to make cepstrum estimates of the noise. Hence the two cases are evaluated. The third combination is also tested to compare directly the cost in performance which may be associated with the approximations.

A digit recognition task was performed using speaker independent models. The background noises used in the experiment were pink, f16 aircraft and ambient noise in a car production factory hall. All these were obtained from the NATO-RSG10 Noise database. The RSRE-PIN89 database was used for speaker independent data in which each subject spoke exactly 19 quadruplet pin numbers in quiet conditions. The noisy data was created artificially by mixing speech and noise at prespecified SNR levels.

The speech and noise models were trained separately. The train set for speaker independent models consisted of the first 80 speakers in PIN89. Single state models were trained for all noises.

The speaker independent recognition was performed on the next 120 Speakers in PIN89. The sound levels were measured in accordance with the SV6 speech voltmeter which calculates rms energy of active speech. The measurement bandwidth was 10kHz.

5 RESULTS

The figures illustrate the percentage accuracy of recognition. The pink noise baseline results were calculated using a front-end which obtained 8 cepstrum and 8 delta cepstrum coefficients. A state variance floor of 2.5dB was imposed for all models.

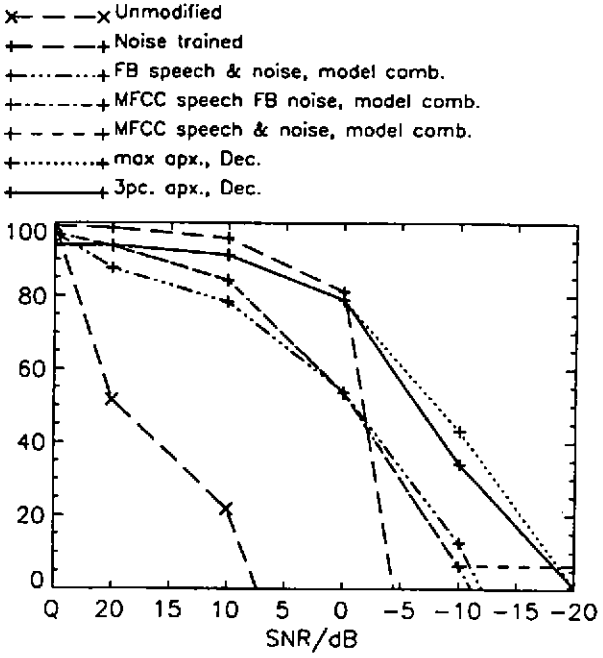

Figure 2: Recognition in Pink Noise.

6 ESTIMATED MASKING OF PHONES

Assuming the noise masking model, it is possible to calculate how much of the speech patterns are masked by a particular noise at various levels. This data gives a useful insight into the difficulty associated with recognising words in noise.

For each phone in spoken English the expected number of times each channel output will be not masked by noise may be calculated. The analysis is done with pink noise as its long term spectrum resembles closely those of most real noises. The phones are gathered from the RSRE ARM database. The expected total number of SRUbank channels which will contain speech and not noise at various levels is as shown in table 1. The first SRUbank channel has been left out as it includes the DC output. The phoneme notations comply with the SAMPA recommendations.

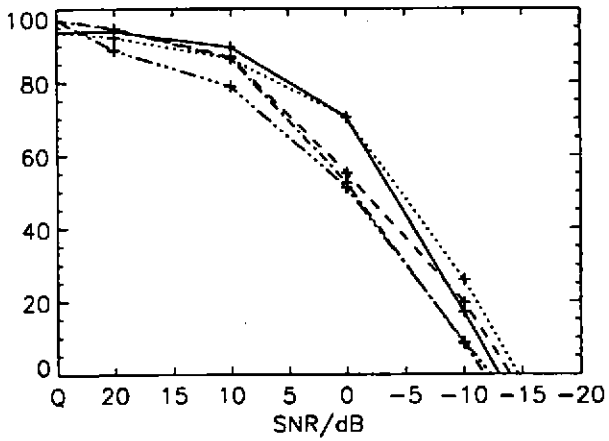

Figure 3: Recognition in F16 aircraft Noise.

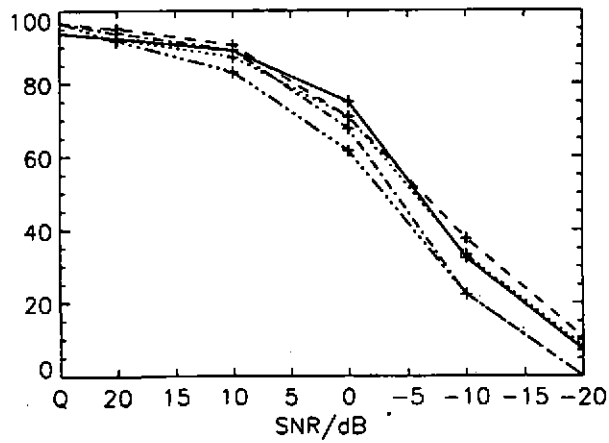

Figure 4: Recognition in Car Production Hall Noise.

Phone	20dB	15dB	10dB	5dB	0dB	-5dB	-10dB	-15dB
i	25.5	22.5	15.9	9.3	4.0	0.4	0.0	0.0
I	25.9	25.0	21.1	13.6	5.0	0.6	0.0	0.0
E	26.0	25.3	22.1	17.6	9.8	2.1	0.1	0.0
{	26.0	25.2	22.0	17.9	11.2	2.8	0.1	0.0
A	26.0	25.2	21.8	17.5	11.4	4.7	0.4	0.0
Q	26.0	24.7	20.0	13.7	7.4	2.5	0.1	0.0
O	25.4	21.2	14.4	8.4	5.2	1.5	0.0	0.0
U	26.0	25.3	22.7	15.6	6.8	1.9	0.2	0.0
u	25.1	20.9	15.2	9.1	3.1	0.4	0.0	0.0
3	26.0	25.6	22.7	18.2	10.5	2.4	0.1	0.0
V	26.0	24.4	20.3	14.1	6.7	1.3	0.0	0.0
@	25.6	22.4	16.6	9.4	2.7	0.2	0.0	0.0
eI	26.0	25.6	21.9	16.9	9.9	2.2	0.1	0.0
aI	26.0	25.5	22.0	18.1	11.1	3.0	0.1	0.0
OI	26.0	23.7	18.1	10.4	3.6	0.4	0.0	0.0
aU	26.0	25.0	21.4	17.0	10.7	2.5	0.1	0.0
@U	25.4	22.0	16.8	11.5	5.5	1.0	0.0	0.0
I@	26.0	24.9	21.9	18.3	10.8	2.4	0.1	0.0
e@	26.0	25.5	22.2	17.9	11.3	2.8	0.1	0.0
l	25.8	23.5	18.1	11.7	5.4	1.0	0.0	0.0
r	25.9	23.5	19.0	15.2	8.1	1.5	0.0	0.0
w	25.1	20.5	15.0	10.6	6.0	1.5	0.0	0.0
j	26.0	25.4	20.3	14.0	6.6	1.0	0.0	0.0
m	25.6	20.9	13.9	6.2	1.9	0.2	0.0	0.0
n	24.9	18.3	9.6	3.0	0.7	0.0	0.0	0.0
N	25.4	20.6	12.8	5.2	1.3	0.0	0.0	0.0
p	24.5	18.0	7.7	0.8	0.0	0.0	0.0	0.0
b	25.1	20.1	13.5	7.9	3.3	0.5	0.0	0.0
t	25.3	21.8	16.0	6.7	0.8	0.0	0.0	0.0
d	23.9	17.2	8.5	1.8	0.1	0.0	0.0	0.0
k	24.8	19.5	13.0	5.4	0.7	0.0	0.0	0.0
g	25.9	24.5	20.3	14.0	5.3	0.7	0.0	0.0
s	25.9	24.7	18.0	8.3	2.4	0.1	0.0	0.0
z	25.7	23.9	15.9	6.1	1.1	0.0	0.0	0.0
S	25.7	24.0	19.1	10.1	3.9	0.3	0.0	0.0
Z	25.9	24.0	18.1	10.8	2.9	0.2	0.0	0.0
f	25.8	24.4	17.8	3.6	0.0	0.0	0.0	0.0
v	26.0	24.6	19.3	10.3	3.3	0.2	0.0	0.0
T	26.0	25.6	19.4	5.0	0.2	0.0	0.0	0.0
D	26.0	25.6	19.3	12.3	6.8	1.5	0.0	0.0
h	26.0	24.5	19.0	9.7	2.7	0.2	0.0	0.0
tS	26.0	25.8	21.1	11.1	3.8	0.3	0.0	0.0
dZ	25.4	22.1	12.8	5.1	1.2	0.1	0.0	0.0

Table 1: Masking of phones in the ARM database.

7 DISCUSSION

Table 1 suggests that below 0 SNR recognisers have to depend almost entirely on the vowels for discrimination as many of the consonants are completely masked out. The digit recognition results obtained indicate a more optimistic picture, probably based on this explanation.

Though the HMM decompositions appears to offer better performance for pink noise, the difference disappears when using real background noises. The 3 piece approximation appears to be slightly better than the masking function for HMM decomposition.

The difference between using filterbank models and cepstrum models for noise appears to be negligible. Both estimates seem to be accurate enough. The combination of filterbank models seem to offer the poorest of the results thereby illustrating the cost of the approximations used in the combination which are noticeable when comparing them with the HMM decomposition results. This is also observable when comparing the cepstrum combined models with the noise trained baseline results. It is interesting to note that the baseline models could not be trained at -10dB SNR. Though only stationary noises were used in the experiment these techniques are applicable to non-stationary noises as well.

References

[1] Varga A. P., Moore R. K., *Hidden Markov model decomposition of speech and noise*, Proc. ICASSP90, pp845-848, Albuquerque 1990.

[2] Kadirkamanathan M., Varga A. P., *Simultaneous model re-estimation from contaminated data by 'composed' hidden Markov modelling*, Proc. ICASSP91, pp729-734, Toronto 1991.

[3] Gales M. J. F., Young S. J., *An improved approach to the hidden Markov model decomposition*", Proc. ICASSP92, pp729-734, San Fransisco 1992.

[4] Klatt D. H., *A digital filter bank for spectral matching*, Proc. ICASSP76, pp573-576, Philadelphia 1976.

[5] Dautrich B. A., Rabiner L. R., Martin T. B., *On the effects of varying filter bank parameters on isolated word recognition*, IEEE Trans. on ASSP, ASSP-32, pp 793-806.