ISCA Archive
http://www.isca-speech.org/archive

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18-22, 2001

SPEAKER VERIFICATION UNDER ADDITIVE NOISE CONDITIONS WITH NON-STATIONARY SNR USING PMC

Lit Ping Wong and Martin J. Russell

School of Electronic and Electrical Engineering
The University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
L.P.Wong, M.J.Russell@bham.ac.uk

ABSTRACT

In real speaker verification applications, additive or convolutive noise creates a mismatch between training and recognition environments, degrading performance. Parallel Model Combination (PMC) is used successfully to improve the noise robustness of Hidden Markov Model (HMM) based speech recognisers [5]. This paper presents the results of applying PMC to compensate for additive noise with non-stationary signal-to-noise ratios (SNRs) in HMM-based text-dependent speaker verification. Speech and noise data were obtained from the YOHO [6] and NOISEX-92 databases [13] respectively. Speaker recognition Equal Error Rates (EER) are presented for noise-contaminated speech at different SNRs and different noise sources. For example, average EER for speech in operations room noise at 6dB SNR dropped from approximately 20% un-compensated to less than 5% using PMC. Finally, it is shown that PMC improves verification by an average EER reduction of 18.29% under varying SNRs (50% of the actual speech duration under 0dB SNR).

1. INTRODUCTION

The goal of speaker verification is to confirm the claimed identity of a subject by exploiting individual differences in their speech. It is useful to distinguish between text-dependent speaker verification, where the decision is made using speech corresponding to known text, and text-independent speaker verification, where the speech is unconstrained [4]. The present study is concerned with the former. Text-dependent verification is clearly the simpler problem, and is amenable to word-level acoustic modelling techniques from automatic speech recognition, and in particular the use of HMMs.

It is well known that noise contamination of speech signals results in increased speech and speaker recognition errors, due to the consequent mismatch between training and test conditions, and loss of information. Hence considerable effort has been applied to the development of robust noise compensation techniques. These techniques generally fall into two categories, speech pre-processing and adaptation of the recognition stage. The first class of methods attempt to pre-process the corrupted speech such that the resulting parameters are representative of clean speech. Techniques in this category include spectral subtraction [2], and spectral mapping [11]. Examples of techniques that modify the recognition stage include noise masking [8], HMM decomposition [12] and Parallel Model Combination (PMC) [5]. Methods based on pre-processing are often computationally simpler, but have the disadvantage that any relevant information that is discarded in the pre-processing stage cannot be recovered for use in recognition. The work in this paper focuses on the second approach, and in particular on the use of Parallel Model Combination (PMC) in which speech and noise HMMs are compiled into a single composite HMM prior to recognition.

PMC has been applied successfully to HMM-based automatic speech recognition, where it has been shown to improve recognition performance on speech contaminated with additive noise [5]. This paper presents the results of experiments to investigate the utility of PMC for noise-robust HMM based text-dependent speaker verification. Clean speaker verification data from the YOHO database [6], designated as test data, was mixed with two different types of noises from the NOISEX-92 database [13] at a range of different signal-to-noise ratios. This data was then processed using clean speech models (trained from clean enrolment speech) combined with the appropriate noise models at the appropriate mixing levels using PMC. The results show that PMC gives significant reductions in equal error rates. For example, un-compensated EERs of 20%, 33% and 45% (at +6dB, 0dB and –6dB respectively) are reduced to 5%, 13% and 30% respectively using PMC. Further experiments were conducted to measure the sensitivity of the PMC error rate to changes in the value of the mixing level. The results show that performance is relatively insensitive to mixing level, and that restriction to seven different mixing levels only results in a small increase in error relative to the experiments with correctly matched mixing levels. Based on this conclusion, a composite model consisting of seven PMC HMMs was used to evaluate speech with varying SNRs. For each speech segment, the SNR was gradually altered from +18dB to –18dB and finally back to +18dB. This produced verification data with SNRs of below 0dB for approximately 50% for the actual speech duration. Results show that PMC gives an average reduction in EER of 18.29%. For example, using speech contaminated with operations room noise, un-compensated EER dropped from 34.78% to 16.84% compensated.

2. PARALLEL MODEL COMBINATION (PMC)

PMC is based on the premise that noise compensation should occur during the pattern processing stage of speech or speaker recognition and not during parameterisation. In particular, the decision that a component of the data is 'noise' should emerge from the recognition process, rather than precede it. In this way, all of the information contained in the speech signal is retained and exploited for correct verification. Although similar to HMM decomposition [12], PMC has the advantage that it is able to operate in the cepstral domain and therefore inherits the advantages of parameter decorrelation. Methods similar to this work have been investigated for text-independent speaker recognition [14]

Let σ_S and σ_N be single Gaussian states of a 'clean' speech HMM and a noise HMM respectively, in the cepstral domain. Suppose that the means and variances of σ_S and σ_N are denoted by $\{\mu_S^c, \Sigma_S^c\}$ and $\{\mu_N^c, \Sigma_N^c\}$. PMC creates a combined cepstral state $\sigma_S \otimes \sigma_N$ by inverse-transforming $\{\mu_S^c, \Sigma_S^c\}$ and $\{\mu_N^c, \Sigma_N^c\}$ into the linear, spectral domain (via the log spectral domain), combining them appropriately into a single distribution, and then mapping this distribution back into the cepstral domain. During this process, PMC makes a number of assumptions. It is assumed that speech and noise are independent, that they are additive, and, during the combination process, that the sum of two log-normal distributions is log normal. Further extensions to the basic PMC routine (with the aid of an additional general speech HMM) enable it to compensate for both additive and convolutive noises [5]. However, for this study, only additive noise compensation will be considered.

Both the static and dynamic (velocity and acceleration) parameters, which are normally included in a cepstrum-based representation for speech recognition can be compensated using PMC. However, in this study, only static parameters were compensated, as it was believed that compensating for the delta and acceleration parameters would only bring marginal improvements in performance as most of the speaker dependent characteristics will be removed by the differencing process. The speech mean and covariance matrix in the log spectral domain are given by:

$$\mu_S^l = C^{-1} \mu_S^c \tag{1}$$

$$\Sigma_S^l = C^{-1} \Sigma_S^c (C^{-1})^T \tag{2}$$

where C is the cosine transform defined by the matrix

$$C_{ij} = \cos(i(j - 0.5)\pi / B) \tag{3}$$

Similar expressions exist for the parameters of the noise state. The mean and covariance matrices for the state $\sigma_S \otimes \sigma_N$ are then given by:

$$\mu_{S \otimes N}^l = \log(\exp(\mu_S^l) + g \exp(\mu_N^l)) \tag{4}$$

$$\Sigma_{S \otimes N}^l = \log(\exp(\Sigma_S^l) + g \exp(\Sigma_N^l)) \tag{5}$$

where g is a gain matching term which determines the signal-to-noise ratio.

3. EXPERIMENTAL DATA

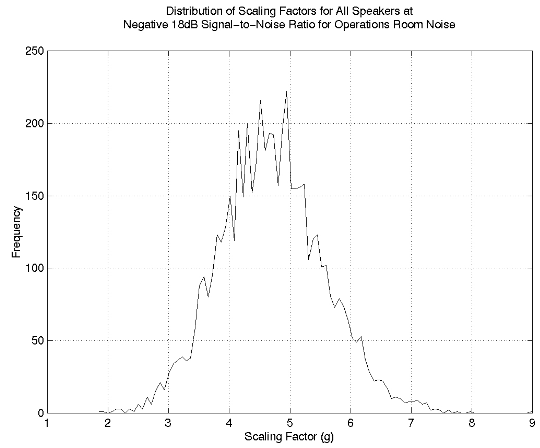

Figure 1: Distribution of the scaling factor

The experiments used speech data from three corpora: TIMIT, YOHO [6] and NOISEX-92 [13]. TIMIT was used to initialise a set of phoneme-level HMMs, because no phoneme transcriptions were provided with YOHO.

Speech in the YOHO database is sampled at 8k samples per second. The corpus consists of recordings of 138 subjects speaking connected digit phrases in an office environment. It was chosen because of its established use in evaluation of speaker recognition systems, and enables the results obtained from the present experiment to be compared with those from other laboratories [3].

The NOISEX-92 database contains a range of different types of noise, including speech noise, car noise, and operations room noise. It has been used previously to investigate noise compensation techniques for automatic speech recognition [5].

4. EXPERIMENTAL METHOD

4.1. Mixing Speech and Noise

4.1.1. Stationary SNR

The speech level was measured using a software implementation of the procedure described in [7], which accounted for the silence intervals present in speech and only calculated the level for voiced segments. Before applying this process, both verification speech and noise data were amplitude scaled to ¼ of the maximum 16-bit integer value to prevent saturation when noise was added. Enrolment data was also scaled to prevent trained clean speech HMMs from exhibiting any mismatch due to level differences. Then, noise signals were added to verification speech data to give noise contaminated speech signals at–18dB to +18dB at 6dB SNR intervals. A second set of enrolment data, contaminated with noise was also created according to the procedure above. This was used to produce matched HMMs.

Obviously, each verification speech segment will have different number of samples. Therefore, the number of samples read from the noise data will vary. However, since data in NOISEX-92 is repetitive, there are no noticeable differences between the noise contaminated verification segments.

The actual scaling factor used to obtain a particular SNR varied between speakers and speech segments. Hence for each SNR the average scaling factor g was computed across all speakers. This scaling factor was used in the construction of the PMC models for that SNR. An example of the distribution of the scaling factor for a single SNR and different speakers and speech segments is shown in Fig. 1.

4.1.2. *Non-Stationary SNR*

Extensions were made to the procedure described in section 4.1.1, which allowed noise signals to be added at varying SNRs across the duration of each speech segment. For this study, we varied the noise SNR according to the graph shown in Fig. 2. This yielded noise-contaminated speech segments that have SNRs below 0dB for ≈ 30% of the entire speech segment. In addition, each segment has a silence period at the beginning, which contributes ≈ 40% to the entire duration. This meant that ≈ 50% of the actual spoken phrase has SNRs of below 0dB. Further illustrations are shown in Fig. 3.

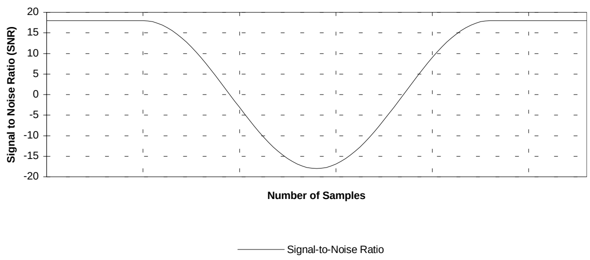

Figure 2: Variation in SNR

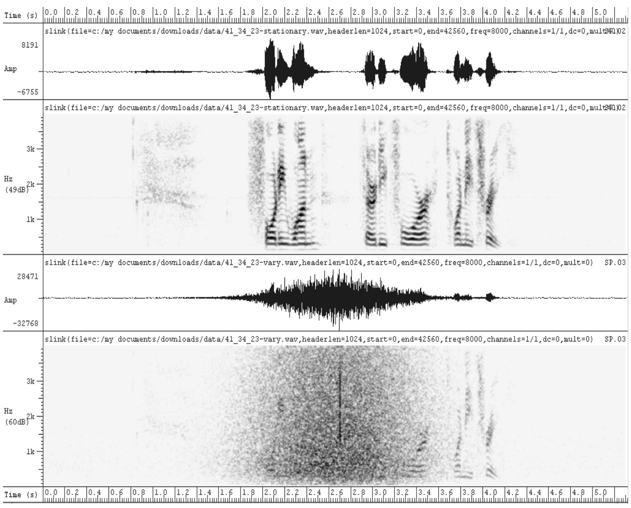

Figure 3: Clean speech and noise contaminated speech waveforms, and spectrograms

4.2. HMM System

4.2.1. *Standard HMM*

HMM training and recognition was performed using the Hidden Markov Model Toolkit (HTK) system, which was easily modified to accommodate for PMC. Speech was parameterised into 39 dimensional representation based on Mel-Frequency Cepstral Coefficients (MFCC), using a 25ms Hamming window. This included the 0^{th} order energy term, plus velocity (Δ) and acceleration (Δ^2) of each coefficient.

Initially, a total of 57 3-state, 4-component Gaussian mixture monophone HMMs were created using the TIMIT database. These were later expanded to 78 tied-state triphones to cover all possible pronunciations of connected digit pairs, e.g. 29_30_31. The triphones were subsequently used to create speaker dependent models using speech from the YOHO database. For each speaker, all 24 utterances in each of 4 enrolment sessions were used for training. Two sets of speaker dependent models were used, one trained using clean enrolment speech and the other using noise contaminated enrolment speech. The latter was used for matched recognition.

4.2.2. *Composite HMM*

To perform recognition on speech with varying SNR, a '3 dimensional' composite PMC-like model is constructed in which the (i,j,k)th state corresponds to the ith state of the speech model composed with the jth state of the noise model at the kth SNR. Seven SNRs are considered (see section 5.4), ranging from –18dB to +18dB in 6dB steps. The changes between these seven SNRs are modeled as a Markov process, whose transition probability matrix is shown in Fig. 4. Although this matrix was fixed in these experiments, there is clearly an opportunity to adapt the parameters to reflect the changes in SNR which occur in a particular application.

$$
\begin{array}{r c}
 & \begin{array}{ccccccc} +18dB & +12dB & +6dB & 0dB & -6dB & -12dB & -18dB \end{array} \\
\begin{array}{r} \textit{Entry} \\ +18dB \\ +12dB \\ +6dB \\ 0dB \\ -6dB \\ -12dB \\ -18dB \end{array} &
\begin{bmatrix}
0.3 & 0.2 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.4 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.4 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.4 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.4 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.4 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.4 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.4
\end{bmatrix}
\end{array}
$$

Figure 4: SNR transition probability of the composite HMM

4.3. Recognition and Scoring

Of the 138 speakers in YOHO, 118 were used as authorised speakers and 20 were used to train a general speaker model (GSM) [10]. Furthermore, all 4 utterances in the 10 verification sessions were used to calculate the False Reject Rate (FRR). To calculate the False Accept Rate (FAR) for each speaker, 40 utterances were randomly chosen from the authorised speaker set except the speaker on test to form an impostor set. The authorised speaker model (AS) was then used to recognise utterances from the impostor set.

For both FRR and FAR, the GSM was used to normalise scores from the speaker dependent models [1]. The decision rule used, for a particular threshold t, is as follows:

If $\dfrac{P(X \mid S)}{P(X \mid GSM)} \geq t$ then 'Accept', else 'Reject'.

The Equal Error Rate (EER) corresponds to the value of t for with FRR = FAR.

5. EVALUATION

Six sets of experiments were carried out. First, the performance of the baseline system was tested using clean speech, and matched recognition. The second experiment investigated the degradation in performance when noise contaminated speech was used in verification without PMC. Next, the same experiment was performed using PMC. The second and third experiments considered both operations room noise (as an example of 'typical' noise) and speech noise (as 'worst case'). Then, the dependency of performance on the gain matching term, g, was investigated. Verification performance where the SNR was unknown and only 7 PMC models were used is reported. Finally, experiments using the composite model described in section 4.2.2, tested on noise contaminated speech at varying SNRs were performed.

5.1. Baseline System Performance

5.1.1. Clean Speech Recognition

The EER achieved using clean verification speech was 0.57%. The target, therefore, was to get the EER as close to this value for all SNRs.

5.1.2. Matched Recognition

To provide a more substantial baseline across all SNRs, matched recognition was performed using HMMs trained on noise contaminated enrolment data. It was observed that the results were worse than PMC compensated recognition as reported in section 5.3. Furthermore, for SNRs of below ≈ 6dB, performance was worse than un-compensated performances as reported in section 5.2. It was believed that the speaker models were not optimally trained. This was because the Baum-Welch process could not find the best state alignment for the noise-contaminated data during embedded training. The suggested solution would be to perform isolated phoneme training on time stamped data.

5.2. Un-Compensated System Performance

The results without noise compensation for SNRs varying between –18dB and +18dB are shown in Fig. 5. The EER deteriorated at an average rate of 10% for every 6dB reduction in SNR. This trend continued until performance levelled off at 50% EER ('random' error for a two class problem). As expected, the performance degradations were much worse for 'speech' noise than for 'operations room' noise.

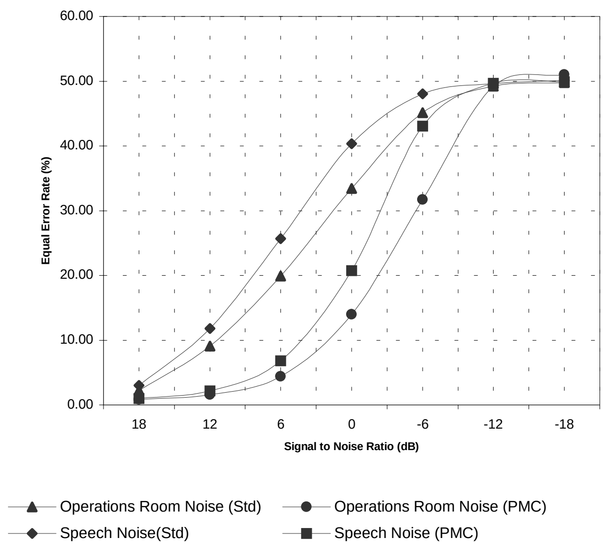

Figure 5: Graph of compensated results against un-compensated results

5.3. Compensated System Performance

PMC compensation was performed using the average scaling factor obtained during speech and noise mixing to set the gain matching term g (section 4.1). With PMC, performance at 18dB was close to the clean speech baseline result. Although performance degradation was still observed as SNR was reduced, this was not as severe as the un-compensated models. An average of 50% performance increase was observed between standard and PMC recognition. However, this was only true until -12dB where the performance curve of PMC rejoins that of standard recognition.

5.4. Dependency on g

To investigate the dependency of the EER on the gain matching term g, experiments were conducted using operations room and speech noise in which the value of g was held constant while the SNR was varied between +18dB and –18dB. Values of g which were considered correspond to +18, +12, +6, 0, -6, -12 and –18dB (Fig. 6,7). As one would expect, the results show that sensitivity to the absolute value of g increases as SNR increases. However, the results show that performance of the system is relatively insensitive to the exact value of g, and that for an SNR of RdB, good performance is obtained with values of g corresponding to SNRs of R±6dB. This suggests that good speaker verification performance can be obtained over SNRs

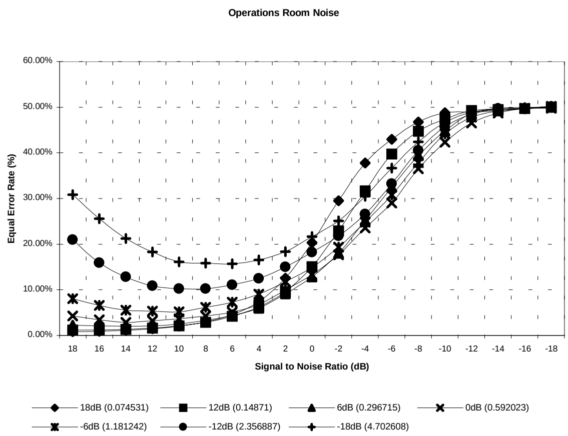

Figure 6: Dependency on g using operations room

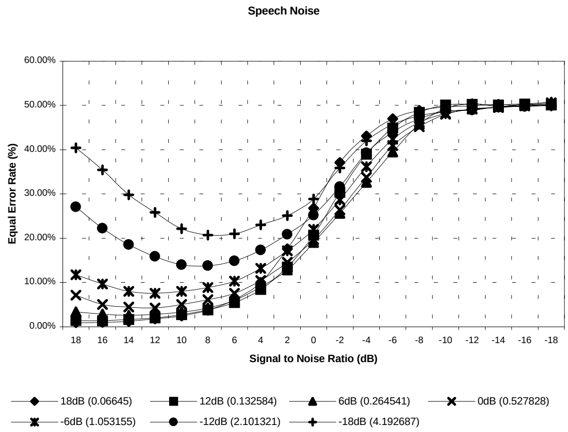

Figure 7: Dependency on g using speech noise

ranging from +18dB to –18dB using PMC models with g corresponding to no more that 7 different SNRs. For example, consider operations room noise. At –6dB the best performance is actually achieved using the value of g which matches 0dB SNR, and the error rates for values of g corresponding to –12dB, 0dB and +6dB are all within 10% of the error rate for the 'correct' value of g for –6dB.

6. UNKNOWN SNR

The next experiment considered the case where the SNR was unknown. Given a section of noise contaminated speech y, the probability $P(y|M_g)$ was calculated for the seven authorised speaker models M_g for values of g corresponding to SNRs of –18dB, -12dB, -6dB, 0dB, +6dB, +12dB and +18dB. The model M_g for which $P(y|M_g)$ is greatest was used for verification. Similarly, the probabilities $P(y|GSM_g)$ were calculated for the general speaker models GSM_g with values of g corresponding to the same seven SNRs and the model GSM_g for which $P(y|GSM_g)$ is greatest was used for verification. Fig. 8 shows the results of this experiment for speech mixed with operations room and speech noise at SNRs between –18dB and +18dB at 2dB

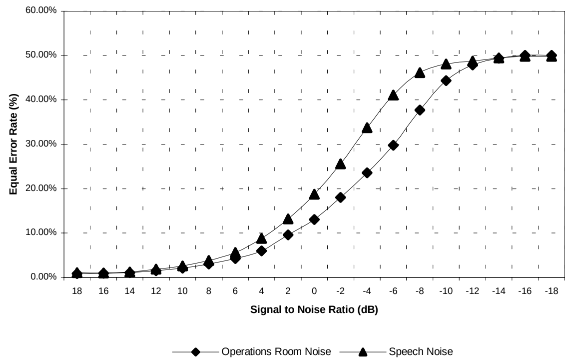

Figure 8: Optimally compensated results

intervals. When compared with Fig. 5, results show that given unknown SNR, performance for all two noise sources lie between ≈0.1% of predicted values. As expected, result shows that at 0dB, both speech and operations room noise have a low compensated EER of 18.74% and 13.07% respectively as compared to 39.39% and 33.42% un-compensated.

7. NON-STATIONARY SNR

The final experiment was conducted to investigate verification performance when noise sources have non-stationary SNRs. This was to test the hypothesis in the previous section that 7 PMC HMM models can model a whole range of SNRs. A composite HMM model described in section 4.2.2 was used. This was used on verification data prepared as outlined in section 4.1.2. DET curves are plotted (Fig. 9,10) and results show that for both noise sources, the composite model performs better than uncompensated models. The compensated EER for both operations room noise and speech noise is 16.84% and 19.92% respectively, compared to non-compensated EER of 34.78% and 38.55%. Performance is expected to improve further by a more careful selection of the SNR transition matrix, as opposed to a random selection of probabilities (section 4.2.2). By analysing the SNR variation in a given noise source, we would be able to characterise the SNR more accurately.

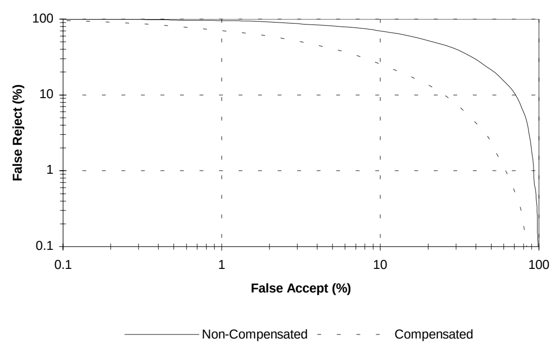

Figure 9: DET curve for operations room noise

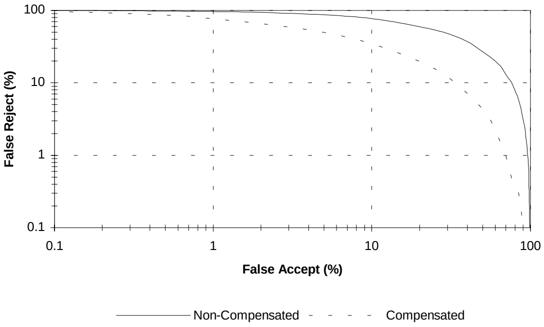

Figure 10: DET curve for speech noise

8. CONCLUSION

This paper reports the results of experiments, which investigate the utility of Parallel Model Combination (PMC) for noise-robust HMM-based text-dependent speaker verification. Speech data was taken from the YOHO corpus, and noise data from NOISEX-92. The system used context-sensitive phoneme-level HMMs with 4-component Gaussian mixture states. The results show that baseline EER of 0.57% for 'clean' speech degrades rapidly as a consequence of contamination with 'speech' or 'operations room' noise. For example the clean speech EER of 0.57% drops to 33% and 41% respectively at 0dB SNR. Using PMC with the correctly matched gain factor *g*, the corresponding figures are 14% and 21%, and both EERs are below 10% at +6dB SNR. It has also been shown that performance is relatively insensitive to the exact value of the gain factor *g*, provided that it corresponds to a SNR that is within ±6dB of the true SNR. Therefore, a small number of models can be used on the whole range of SNRs from –18dB to +18dB. Finally, it has been shown that using a composite model corresponding to 7 PMC HMMs, verification using noise contaminated speech data with varying SNRs was improved using PMC. Results show that for un-compensated verification using 'operations room' and 'speech' noise at varying SNRs, EERs of 34.78% and 38.55% were reported. This dropped to 16.84% and 19.92% respectively when compensated.

9. REFERENCES

[1] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," Digital Signal Processing, Vol. 10, pp. 42-54, 2000.

[2] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Transactions on Acoustics Speech and Signal Processing, Vol. ASSP-27, 1979.

[3] J. P. Campbell Jr., "Speaker Recognition: A Tutorial," Proceedings of the IEEE, Vol. 85, No. 9, September 1997.

[4] S. Furui, "An Overview of Speaker Recognition Technology," ESCA workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[5] M.J. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination," IEE Transactions on Speech and Audio Processing, Vol. 4, No. 5, pp. 352-359, September 1996.

[6] A. Higgins, "YOHO speaker verification," Presented at the Speech Research Symposium, Baltimore, MD, 1990.

[7] ITU-T Recommendation, "Objective Measurement of Active Speech Level," pp. 56, September 1993.

[8] D.H. Klatt, "A digital Filter Bank for Spectral Matching," Proceedings of the ICASSP, pp. 573-576 , 1979.

[9] T. Matsui, T. Kanno, S. Furui, "Speaker Recognition Using HMM Composition in Noisy Environments," Computer Speech and Language, Vol. 10, pp. 107-116, 1996.

[10] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, Vol. 10: (1-3), pp. 19-41, 2000.

[11] H.B.D Sorensen, "A Cepstral Noise Reduction Molti-Layer Neural Network," Proceedings of the ICASSP, S14.14, pp. 933-936, 1991.

[12] A.P. Varga and R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," Proceedings of the ICASSP 90, 1990.

[13] A.P. Varga, H.J.M. Steeneken, M. Tomlinson and D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition," Technical Report, DRA Speech Research Unit, Malvern, England, 1992.

[14] O. Bellot, D. Matrouf, T. Merlin and Jean-Francois Bonastre, "Additive and Convolutive Noises Compensation for Speaker Recognition", Proceedings of the ICSLP 2000 Beijing, China, 2000.