Proceedings of the Institute of Acoustics

THE EFFECTS OF REAL WORLD NOISE ON SPEAKER RECOGNITION PERFORMANCE

John Openshaw (1), John S. Mason (1) & John Oglesby (2)

(1) University of Wales, Swansea, Singleton Park, SWANSEA, SA2 8PP, UK
(2) BT Laboratories, Martlesham Heath, IPSWICH, IP5 7RE, UK

1. INTRODUCTION

In recent years there has been much interest in the recognition of speech in adverse conditions [1]. In general performance is shown to decrease rapidly with increasing levels of background noise, unless specific techniques are introduced to reduce its effects. Here, we address the complimentary task of speaker recognition [2], in adverse conditions, focusing on the effects of additive background noise. To date, little work in this area has been reported, but as for speech recognition, robustness to environmental noise must be achieved before the technology can be used widely.

Historically, investigations into the effects of background noise have typically used Gaussian white noise (GWN) to corrupt the speech signal. In this study the effects of a range of noise types, including GWN, are investigated using the defacto standard feature representations of mel-frequency cepstral coefficients (MFCC) and perceptually based linear predictive cepstral coefficients (PLP). The relative level of difficulty, as measured by speaker identification performance, for the different noises at various signal to noise ratios (SNRs), is evaluated for the two features.

One approach to maintaining performance in noisy conditions is to aim to decrease the sensitivity of the chosen feature representation to the effects of background noise, while retaining the required speaker-specific information. Here the approach proposed in [3] is evaluated. This technique attempts to decrease the sensitivity of cepstral based features to contamination by arbitrary noise, i.e. no prior knowledge of the noise type is assumed. This technique is investigated for a range of real world noise types to determine its effectiveness.

2. EXPERIMENTS

To assess the impact that differing background noise types have on speaker recognition performance, text-dependent speaker identification experiments are used [4]. Text-dependent vector quantised (VQ) codebook models are created and tested with speech data which is, in general, corrupted by additive noise. One codebook is produced for each vocabulary item for each individual speaker, using a modified version of the LBG clustering algorithm. The incoming test token is matched against all codebooks of the appropriate word from all speakers and the model with the smallest accumulated mean square distance across the test utterance is used to indicate the speaker, see Figure 1.

The codebook models are created always using clean speech, while the test utterances typically have noise added prior to feature extraction. Here, two feature types are used, namely standard MFCCs [5] and PLPs [6].

3. SPEECH DATABASE

The identification task is derived from a subset of the BT Millar speech database. This database was collected in a quiet environment, using a high quality microphone, across five sessions. The speech was

recorded at 20 kHz using 16 bits (linear) per sample. In these experiments the data is bandpass filtered to telephone bandwidth and downsampled to 8 kHz prior to feature extraction.

During collection, each of 63 speakers responded to a visual prompt to say the digits *1* through *9*, *oh*, *zero* and *nought* (in a random order) a total of five times in each of five sessions. The database correspondingly contains 25 repetitions of each of the vocabulary items from each speaker. The sessions took place over a period of approximately three months and speakers were encouraged to divide their sessions evenly across this period.

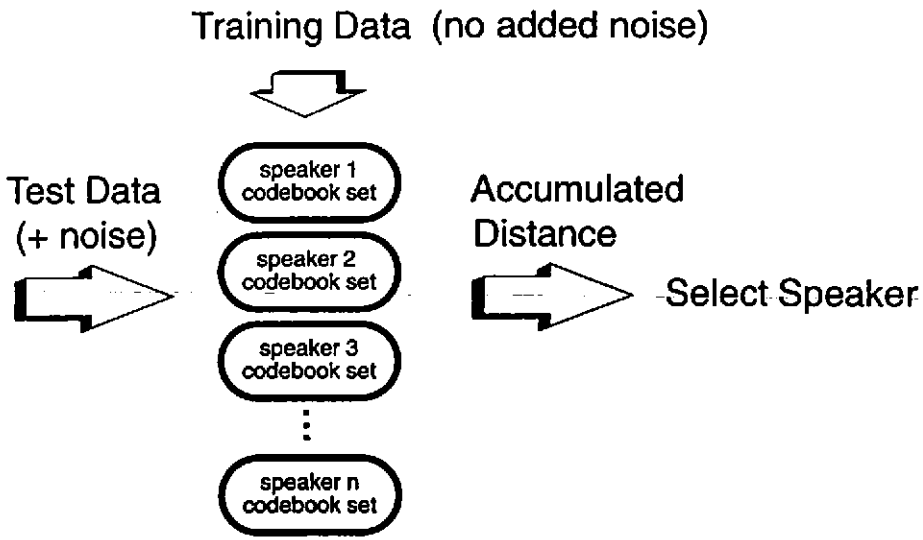

Figure 1: The VQ codebook based speaker identification system.

The database is divided into training and testing sets. The first ten versions, i.e. the first two collection sessions for each individual are reserved for training, with the remainder used for testing. All fifteen repetitions reserved for testing are used as the standard test set. However, two standard training sets are defined:

Set 1: the first three repetitions from the first recording session;

Set 2: the entire training corpus of the first ten repetitions from the first two sessions.

A subset of speakers is adopted. The data from twenty males, all of approximately the same age is used. Unless otherwise stated, all experiments use training set 1. The vocabulary is reduced to be the ten digits *1* through *9* and *zero*, with the exception of the results for the individual digits in Figure 3. This use of the data is chosen to reduce the amount of computation required while maintaining a relatively difficult speaker identification task. There are $20 \times 25 \times 10 = 5000$ test tokens for this identification task, which gives rise to a 95% confidence interval of 1.1% for an example 20% error rate.

4. NOISE DATABASE

The background noise data was taken from the BT Piper database. This database was recorded in situ, directly over the UK telephone network and used a variety of handsets. The data was collected digitally in A-law format at 8kHz and converted to 16bits (linear) for use with the Millar data. For each different noise, approximately 60 seconds of background signal was recorded. This was edited down to 30 seconds, such that the entire recording was representative of a specific type of environment. Five noise types were chosen for this work:

- a conversation between two people on the television (TV);
- crowd babble in a restaurant (Babble);
- background 'contemporary' music (Music);
- an office environment with keyboard clicks from typing at a workstation (Keys);
- car noise when travelling at around 70 m.p.h. (Car).

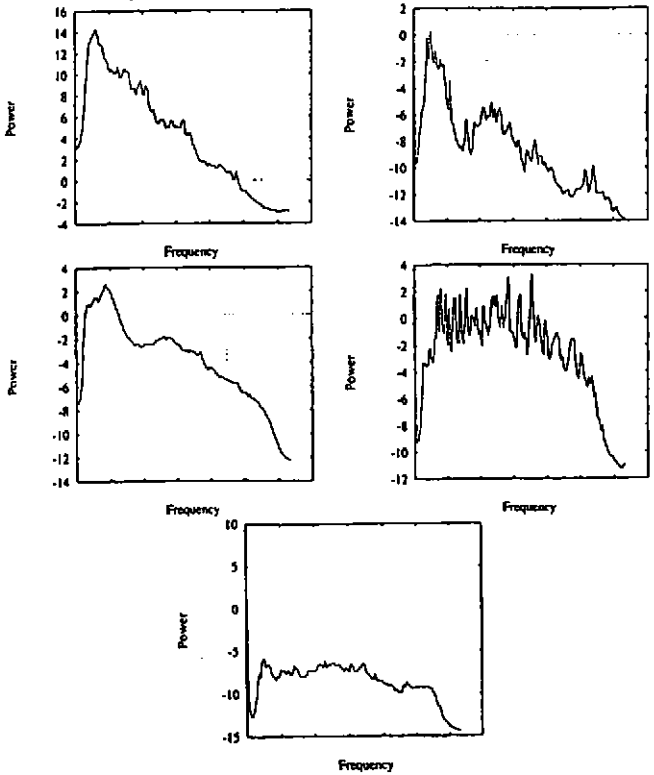

Figure 2: The spectrum of the five noises from the Piper Database. The noises are (from top to bottom and left to right) : Car, TV, Babble, Music and Keys.

Here speech utterances with average SNRs of between clean (>30dB) and 2dB are used. At an SNR of around 2dB the performance is so poor that more extreme conditions were not used. The signal level at which to add the noise is determined from the word energy level and the noise energy level. For the word this is determined individually for each utterance, whereas the noise signal level is averaged across the entire 30 seconds. Therefore, the SNR of each individual word is not necessarily the quoted average figure, given that the exact figure is dependent upon the portion of the noise segment that is added to the speech. This approach is chosen to reflect a realistic environment where the long term average SNR is known. The spectrum of each of the noises is shown in Figure 2.

5. PERFORMANCE IN CLEAN CONDITIONS

The baseline speaker identification performance using MFCC features is shown in Figure 3, i.e. with no additional noise added to the test data. Performance is shown to level off at a codebook size of 8 for both the training sets, with the larger training set consistently out performing the small set for a given size codebook, as would be expected. A large variation between different digits is observed, with similar trends for the two training sets.

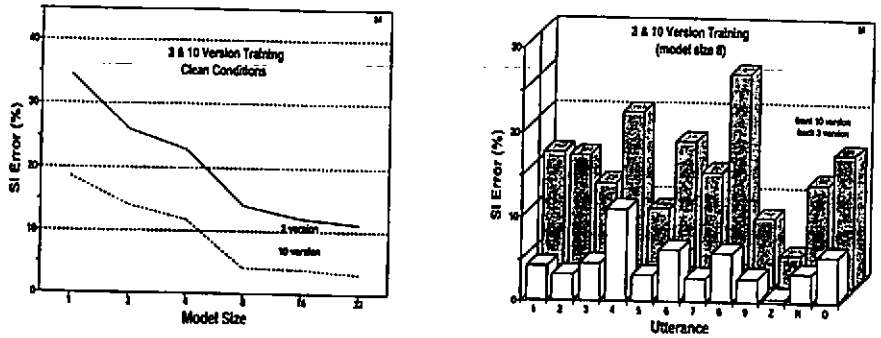

Figure 3: Speaker identification performance under clean conditions versus codebook size (left) and for each individual vocabulary item with the codebook size set at 8 (right).

6. PERFORMANCE IN NOISE

The performance in noisy conditions is now considered. Figure 4 shows the speaker identification performance using MFCC features and codebooks of size 8 at 20 dB SNR for the range of noises considered. The same general trend of improving performance with codebook size is observed, but there are marked differences in absolute performance across the noise types. The Keys noise is in general the worst with the GWN degrading performance by a similar amount. The Car, Babble and TV noise degrade performance significantly from that obtained using clean speech, but do not have such a dramatic effect as the Keys and GWN. Also shown in Figure 4 is the variation in performance with SNR values from clean (> 30 dB) down to 2 dB. Again a distinct ranking is observed, which is reasonably consistent across all SNRs. An interesting observation is that the error rate falls off approximately linearly with the SNR in dB for noise levels in the range 30 dB down to 2 dB, showing a smooth degradation as conditions worsen.

Figure 5 shows the results of the same set of experiments, but now using PLP features. Again a distinct ranking of the different noise conditions is observed across all the codebook sizes, which is essentially the same as for the MFCC case. One difference is that the PLP features appear to be more severely affected by the Keys noise, with errors in excess of 70% for all book sizes at 20 dB SNR.

From Figures 4 and 5 a number of more general observations can be made. In both cases the error rate is significantly affected by the change from clean (> 30 dB) to 30 dB SNR, a level of noise that would typically go unnoticed by a human listener. This highlights the sensitivity of the automatic speaker identification system to noise. Perhaps the most surprising result is that the TV noise, which is essentially two people talking, has the amongst the lowest impact upon performance. Similarly the most consistent and well defined noise, GWN, is one of the harder noise types to cope with.

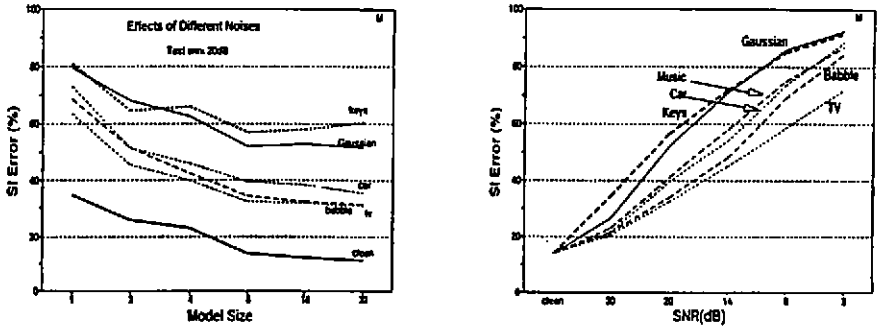

Figure 4: Speaker identification performance in noise contaminated conditions using MFCC features. Performance is shown versus models size for the different noise types at 20dB SNR (left) and versus SNR for a fixed codebook size of 8 (right).

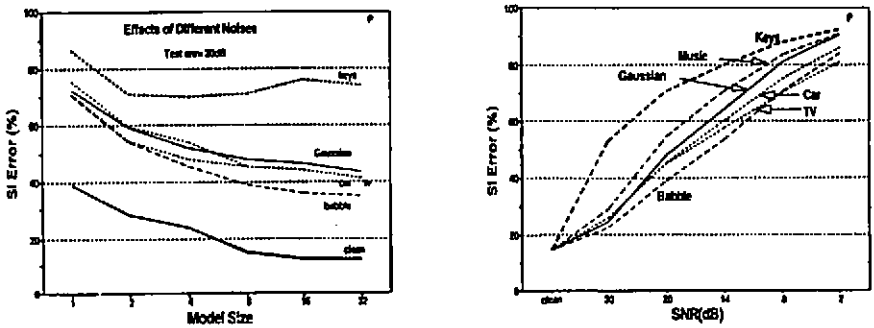

Figure 5: Speaker identification performance in noise contaminated conditions, as for Figure 4 but now using PLP features.

Considering the contrast between the results for the different features it is shown that PLP features give the best performance for only GWN. This is significant due to the large amount of work on noise robust features that make use of GWN as the noise corruption. Our results make clear that such features may not be optimal in environments more typical of real world situations.

7. INCREASING FEATURE ROBUSTNESS

In this section a modification to the MFCC processing is introduced in an attempt to increase robustness to noise. In general cepstra may be defined as:

$$Cepstra = DCT(\log[P(f)]),$$

where $P(f)$ is some representation of the short term power spectrum and DCT is the discrete cosine transform. As proposed in [3] we adjust the formulation to introduce an addition in the power spectrum, which takes the form:

$$Cepstra_{robust} = DCT(\log[P(f) + K(f)]),$$

where $K(f)$ is chosen here to be independent of frequency:

$$K(f) = k \ \forall f.$$

For small values of the offset value k little difference is introduced, whereas as k increases the effect of additional noise on the spectra is reduced, giving an inherently more noise robust feature. However as the value of k is increase the class information component in the feature may reduce. The key issues that arise from this type of modification are:

- what value of k is optimal for different noise environments;
- what level of class discrimination information remains in a feature modified in this way; and
- what combination of background noise and offset value k give rise to improved recognition.

The experiments reported here investigate the use of this form of feature. The offset is introduced for all feature calculations, i.e. both training and testing data. The speaker identification performance is then assessed for the different noise types in cross testing conditions, as above.

The results at an SNR of 14 dB across the range of noise types are given in Figure 6. The performance in clean conditions show that there is a steady decrease in identification performance, demonstrating the reducing discrimination ability of the feature as k increases. What is also shown is the relatively large increase in performance, particularly for the larger training set, across the whole range of noise types. Specifically for the results with the larger training set, the speaker identification performance improves by over 20% by the addition of the offset set at 1000. Even larger gains are seen for the GWN and Keys noise types. For the smaller training set, the improvements in performance for the noise corrupted test data are still evident, again with the GWN and Keys showing most benefit.

Comparing back to the spectra of Figure 2, it is evident that the flattest spectrum is that of the Keys noise. The spectrum of the GWN is also flat. It is therefore not surprising that the contamination that can be compensated for the most should have spectra that reflect the frequency independent adjustment to the power spectrum. A result that is less easy to understand is why the modified features show such an improvement for the Car noise, which has a significant low frequency component.

In all cases the improvements in performance for noisy conditions are obtained at the price of reduced performance for the clean case. For the larger training set, the best performance on the noisy data is achieved when $k = 1000$, at which point the clean performance has dropped from 4% down to 14% errors.

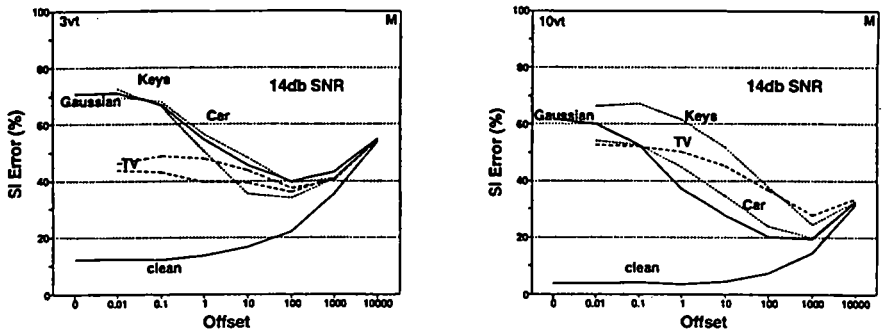

Figure 6: Speaker identification performance using the modified feature representation versus offset value k, for the three repetitions training set (left) and the ten repetitions training set (right).

8. CONCLUSIONS

Speaker identification experiments have been performed using test speech data corrupted by different levels and types of background noises. The results demonstrate that GWN has a significant impact on speaker identification performance. Of the five types of noise used, only keyboard clicks proved more difficult to cope with, at a given SNR. Interestingly, a conversation on the television produced amongst the lowest degradation in performance. It is therefore concluded that it is non-trivial to predict the degradation that specific types of noise will have on speaker identification performance.

The results also show that MFCC features appear to outperform PLP features in all noise environments, except for GWN. This exposes a significant limitation of GWN as a noise model and indicates that optimal features for robust speaker recognition will be dependent upon the type of background noise present. Therefore features that perform well in one noise environment may not be optimal for other environments.

A modification to MFCC features has been investigate and shown to give large performance improvements across all the types of noise corruption considered here at a 14 dB SNR. This improvement in noisy conditions is however at the price of reduced clean speech performance.

9. REFERENCES

[1] ESCA Workshop on Speech Processing in Adverse Conditions, November 1992.

[2] ESCA Workshop on Automatic Speaker Recognition Identification and Verification, April 1994.

[3] "On the limitations of cepstral features in noise", J. P. Openshaw & J. S. Mason, Proc. ICASSP-94, April 1994.

[4] "A vector quantisation approach to speaker recognition", F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang, Proc. ICASSP-85, March 1985.

[5] "Comparisons of parametric representations for monosyllabic word recognition in continuously spoken sentences", S. B. Davis et al, IEEE Trans. ASSP-28, pp 357.

[6] "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)", Proc. Eurospeech-91, September 1991.