ROBUST SPEAKER VERIFICATION IN NOISY ENVIRONMENTS

Hong Tang, Xiaoyuan Zhu, Bruce Millar,
Iain Macleod and Michael Wagner

TRUST[1] Project
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT - This paper considers the problem of robust parametric model estimation and classification in noisy acoustic environments. Characterisation and modelling of external noise sources is in itself an important issue in noise compensation. The techniques described here provide a mechanism for deriving parametric models of acoustic backgrounds along with the signal model so that noise compensation is tightly coupled with signal model training and classification. Prior information about the acoustic background process is provided using a maximum likelihood parameter estimation procedure. A *max* noise model is defined for use in the cepstral domain. We describe experimental evaluation of the benefits gained by applying this approach to text-independent speaker verification. With short speech utterances and various noise environments, useful improvements in speaker verification performance were obtained.

INTRODUCTION

The growing need for automation in complex work environments and the increased use of voice-operated services in many commercial areas have motivated recent efforts to translate laboratory speech-processing algorithms to practical use. In the area of speaker recognition, the TRUST Project is achieving good results in relatively constrained environments with short spoken interactive commands (Zhu *et al.*, 1994a and Chen *et al.*, 1994). Normal workplace environments have various forms of extraneous noise, however. Speaker verification performance normally deteriorates, often dramatically, in such environments. One of the biggest obstacles to the practical use of speech processing equipment is the presence of acoustic noise. In many applications, acoustic noise can be nonstationary and may depart considerably from traditional broadband or impulsive noise models.

A typical speaker verification system is shown in Figure 1. There are three main stages at which noise compensation may be employed: (i) in the recording environment, which can to some extent be controlled by physical adjustment of microphone, for example; (ii) before the feature extraction stage, where signal processing techniques may be used; and (iii) during the comparison stage, in a way that will be highly dependent on the algorithms used.

The first step in reducing the effects of such noise is to use a noise-cancelling microphone which takes the difference signal between two matched transducers (one facing toward and the other facing away from the speaker). The next step is to develop a comprehensive approach to dealing with the effects of acoustic noise in speech processing applications. We have discussed

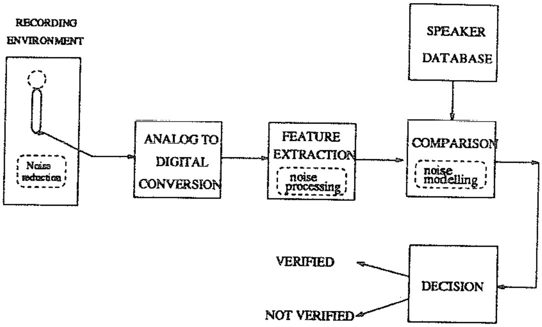

Figure 1: Incorporation of noise processing in the speaker verification system

this issue in Tang *et al.* (1994). This paper concentrates on the third option: the comparison stage.

Our overall aim is to develop a formalism which integrates models of signal and background and which has potential applications to a variety of signal classification problems. Rose, Hofstetter and Reynolds (1993) proposed an integrated model that exploits prior information about background noise to obtain more robust estimates of the parameters of Gaussian mixture classifiers. Prior information about background sources was integrated by providing a mechanism for integrating a broad class of statistical models of background directly with the model for the underlying signal. Their integrated model was shown to be effective for speaker identification in noisy acoustic environments.

In this paper, we describe a mechanism for deriving parametric models of acoustic backgrounds along with the signal model so that noise compensation is tightly coupled with noise and signal models during training and classification. Gaussian mixtures are used because of their ability to provide mathematically tractable approximations to arbitrarily shaped underlying densities. They have demonstrated good performance as parametric speaker models in speaker verification studies. The techniques used in this paper have been applied to text independent speaker verification from short spoken interactive utterances corrupted by various forms of environmental noise.

SPEECH MODELLING WITH AN ACOUSTIC BACKGROUND

It is often the case that the only available observations of a signal source are taken in a noisy environment. It may further be the case that signal classification is to be performed in an environment with dissimilar background noise characteristics. In such applications, it would be desirable to be able to separate the effects of the noisy acoustic background from the signal when estimating signal model parameters and to estimate the effect of combining the signal model with alternate noise models.

If the actual linear signal x_t is corrupted by additive background noise y_t in the time domain and the parameter vectors are log filterbank channel energies, $X = log(x)$ and $Y = log(y)$, where x and y denote the energies of x_t and y_t, respectively, then one can approximate the noise process using a *max* operator (Nadal, Nahamoo & Picheny, 1989)

$$log(x+y) \approx log(max(x,y)) \approx max(X,Y). \tag{1}$$

The speech features used in our experiments were mel-frequency cepstral coefficients (MFCCs), computed as

$$MFCC_i = \sum_{k=1}^{K} X_k cos[i(k-\frac{1}{2})\frac{\pi}{K}], \qquad i = 1,2,\cdots,I, \tag{2}$$

where I is the number of cepstral coefficients, K is the number of triangular bandpass filters (Davis & Mermelstein, 1980), and X_k represents the log-energy output of the kth filter. Therefore, our speech features are a linear combination of log filterbank energies. If we assume that the background acoustic noise is independent of the speech signal, we can use the *max* noise model in the cepstral domain.

The philosophy of the proposed approach is as follows: if we have *a priori* knowledge of the acoustic background, we can model the noise in such environments, and use the noise model along with the speech model in the speaker verification stage for noise compensation. First, the speech and noise models were trained with clean speech and pure noise respectively. The model parameters of the mixture-Gaussian VQ were estimated with the method described by Zhu *et al.* (1994b). In the speaker verification procedure, the speech features of input noisy speech were compared with the speech and noise models. Frame-based speech and noise likelihood scores were input to the *max* operator. The claimed speaker's overall likelihood score was calculated by accumulating scores for those frames which had speech scores larger than their noise scores. As signal-to-noise ratio varies throughout speech owing to its intrinsic dynamic energy range, those frames which contain little speech energy are likely to be ignored. As this technique switches only when the noise characteristics dominate the speech characteristics it may be expected to operate most effectively in relatively high noise environment.

Figure 2 shows the design of speech modelling with an acoustic background, where the speech model has M mixtures and the noise model has N mixtures. The output speaker likelihood score is then compared with a pre-set threshold to decide whether to accept or reject the input speaker's claimed identity.

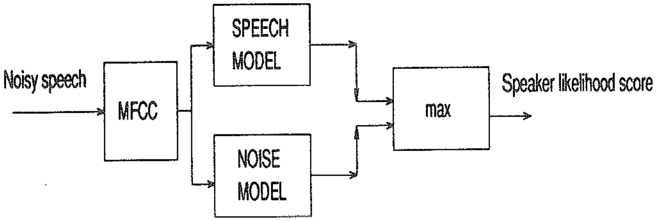

Figure 2: Design of speech modelling with acoustic background.

EXPERIMENTAL RESULTS

The proposed noise compensation approach has been applied to the problem of text independent speaker verification. The techniques were evaluated as part of an automated system for verifying the speaker with a short utterance. The goal of these experiments was to determine to what extent speaker verification performance in noisy acoustic environments can be improved when prior knowledge of the acoustic background process is incorporated through the use of distinct signal and background models.

The speaker verification experiments were performed on the TRUST speech database (Millar *et al.*, 1994), which contains natural speech commands from 45 male and female subjects. Speech data were acquired in three sessions, with five repetitions of 30 utterances in each session. An additional 10 utterances were recorded in sessions 2 and 3, with 2 repetitions in each session. The intervals between the sessions ranged from one to several weeks.

The speech used in these experiments was digitised to 16-bit linear samples at a sampling rate of 20000 samples/s. The speech pre-processing and MFCC generation was as described by Zhu *et al.* (1994c). Two sets of the same utterances were used for each speaker: the first set was clean speech and was used for training the speaker models; the second set (recorded in a different session) was degraded by added noise and used to evaluate verification performance. In order to use speech-only-utterance to train speakers' models, silences within each utterance in training data were cut.

Various acoustic background environments were simulated by adding Speech Babble and Office Noise from the NOISEX-92 corpus (Varga, 1993) to the digitised speech. The discrete time samples of the noise signals were resampled, scaled and added to the speech giving speech/noise SNR ratios of $20dB$, $10dB$ and $5dB$.

In our experiments, speaker models were trained with clean speech from session 1. For testing, 170 utterances from session 2 were degraded with various types and levels of noise. In order to evaluate the proposed noise compensation approach, we used a conventional mixture-Gaussian VQ classifier (MG-VQ) and our proposed mixture-Gaussian VQ with noise model classifier (MG-NM) to evaluate speaker verification performance with noisy speech. The speaker verification results from 10 speakers are shown in Table 1. The tabulated results represent the equal-error-rate (EER) percentage for speaker verification with utterances of about one second duration and thresholds derived a posteriori. We used $M = 32$ mixtures in the speech model and $N = 4$ mixtures in the noise model.

Effect of incorporating speech/background models during speaker verification (Speech models trained in clean environment)								
Noise type	Speech Babble				Office Noise			
S/N ratio	∞	$20dB$	$10dB$	$5dB$	∞	$20dB$	$10dB$	$5dB$
MG-VQ	6.18	12.67	23.10	29.52	6.18	16.18	28.05	35.30
MG-NM (mean % speech)	5.99 (99.75)	12.09 (87.12)	18.70 (62.31)	22.54 (41.38)	6.33 (99.66)	15.34 (84.23)	24.72 (69.47)	29.10 (47.47)

Table 1. Speaker verification equal-error-rates for the MG-VQ and MG-NM classifiers and mean percentages of contributing speech frames in the MG-NM classifier under speech babble and office noise environments.

From Table 1, we can see that classifier MG-NM improved speaker verification results in all our tested noise environments, especially in lower SNR environments. This was confirmed by an analysis of variance, which showed that there was a significant difference (in all cases $p < 0.001$) between the two rows of columns $5dB$ and $10dB$.

Table 1 also gives the mean percentage of the numbers of speech frames which contributed to the speaker scores in the MG-NM classifier. Speech frames were detected by the *max* operator. Speaker verification results of the MG-NM classifier also deteriorated under larger noisy environments since fewer speech frames were used for verifying speakers. However, it outperformed the MG-VQ classifier, because excessively noise frames were not allowed to influence the speaker's overall likelihood score.

CONCLUSION

Conventional speaker verification techniques are rather sensitive to environmental acoustic noise. To improve the verification process in this respect, we have proposed and tested an acoustic noise compensation approach, based on prior modelling of background noise with Gaussian mixtures and the use of a *max* operator during calculation of the likelihood score for the claimed speaker. As a result of extensive experimentation, we conclude that with short speech utterances and various noise environments, noticeable improvements in speaker verification performance have been obtained and significant improvements in high noise environments can be achieved as expected.

ACKNOWLEDGEMENT

This research has been carried out on behalf of the Harry Triguboff AM Research Syndicate.

NOTES

[1] Technology for Robust User-conscious Secure Transactions.

REFERENCES

Chen, F., Macleod, I., Millar. B. & Wagner M. (1994) *Hybrid threshold approach in text-independent speaker verification* , Proc. Int. Conf. Spoken Language Processing, Yokohama, Japan, pp.1855-1858.

Davis, S. & Mermelstein, P. (1980) *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans. Acoust., Speech and Signal Processing, Vol.28, No.4, 357–366.

Millar, B., Chen, F., Macleod, I., Ran, S., Tang, H., Wagner, M. & Zhu, X. (1994) *Overview of speaker verification studies towards Technology for Robust User-conscious Secure Transactions*, Proc. Int. Conf. on Speech Science and Technology, Perth, Australia.

Nadal, A., Nahamoo, D. & Picheny, M.A. (1989) *Speech recognition using noise-adaptive prototypes*, IEEE Trans. Acoust., Speech and Signal Processing, Vol.37, No.10, 1495–1502.

Rose, R.C., Hofstetter, E.M. & Reynolds, D.A. (1994) *Integrated models of signal and background with application to speaker identification in noise*, IEEE Trans. Speech and Audio Processing, Vol.2, No.2, 245–257.

Tang, H., Zhu, X., Macleod, I., Millar. B. & Wagner M. (1994) *A dynamic-window weighted-RMS averaging filter applied to speaker verification*, Proc. Int. Conf. Spoken Language Processing, Yokohama, Japan, 1603–1606.

Varga, A. (1993) *Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems*, Speech Communication, Vol.12, 247–251.

Zhu, X., Gao, Y., Ran, S., Chen, F., Macleod, I., Millar, B. & Wagner, M. (1994a) *Text-independent speaker recognition using VQ, mixture Gaussian VQ and ergodic HMMs*, Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, 55–58.

Zhu, X., Millar, B., Macleod, I., Wagner, M., Chen, F. & Ran, S. (1994b) *A comparative study of mixture-Gaussian VQ, ergodic HMMs, and left-to-right HMMs for speaker recognition*, Proc. IEEE Int. Symposium on Speech, Image Processing and Neural Networks, Hong Kong, 618–621.

Zhu, X., Millar, B., Macleod I. & Wagner, M. (1994c) *Speaker verification: beyond the absolute threshold*, Proc. Int. Conf. on Speech Science and Technology, Perth, Australia.