SPEECH ENHANCEMENT FOR ROBUST SPEAKER VERIFICATION

Peter Kootsookos [1] Ah Chung Tsoi, and Brian Lovell

Department of Electrical Engineering
University of Queensland, Australia

ABSTRACT – We examine the performance of Kalman filtering and smoothing techniques in the context of a working verification system to see the effect on interspeaker and intraspeaker variability.

The efficacy of Kalman noise reduction on speech contaminated by several types of noise in the context of three different speaker verification techniques – dynamic time warping, vector quantization and recurrent neural network – is investigated. Although the neural network system appears to benefit the most from Kalman noise reduction, there is also significant improvement for the other two systems. Vector quantisation had the least noise sensitivity and the best overall performance.

INTRODUCTION

Any normal work place environment suffers from extraneous sounds. As a result, the problem of robust speech processing, robust meaning insensitive to noise disturbances, has received much attention in the literature [2,4,5,9,10].

We intend to produce a speaker verification system for use in a noisy environment and so we must find ways to deal with extraneous sounds.

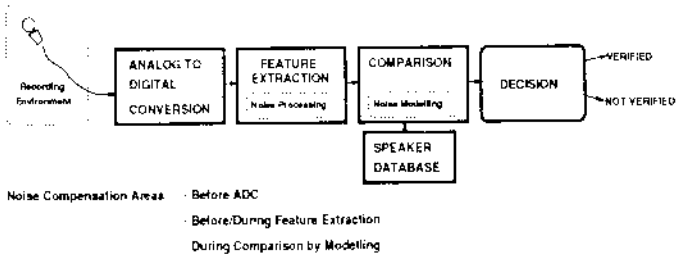

Figure 1. Where noise processing can be incorporated into the proposed system.

Our speaker verification system is illustrated as a sequence of block-diagrams in Figure 1. There are three main places where noise may be compensated for [4]: in the recording environment, which is largely out of our control, before the feature extraction stage, where standard signal processing techniques may be used, and during the comparison stage, which will be highly dependent on the algorithm used. This work concentrates on the second option [2,3,9].

We examine the use of standard signal processing techniques, namely Kalman filtering and smoothing, to improve the quality of the raw speech data before it's passed to the feature extraction section of the system.

KALMAN FILTERING AND SMOOTHING

The Kalman filter [1] is the best (in the least mean-square error sense) linear estimator of the state x_k of our speech generation system given the noisy measurements z_k. See [1] for a complete treatment of the derivation and properties of the Kalman filter and [6] for implementation details.

[1] Peter Kootsookos is now a Research Fellow with the Centre for Robust and Adaptive Systems (CRASys), 89 Labs, DSTO, Salisbury, SA 5108.

Our current verification system uses the standard linear predictive (or auto-regressive) model for speech [8]

$$s_k = -\sum_{i=1}^{M} a_i s_{k-i} + u_k$$
$$= \hat{s}_k + u_k$$

where s_k is the noiseless speech, u_k is the noise source driving our speech model, z_k is the measured speech, v_k is the measurement noise, and M is the order of the linear predictive model. Because of this special structure, the state vector for the system may be selected as

$$x_k = [s_k \; s_{k-1} \; \cdots \; s_{k-M}]^T \tag{1}$$

Remark 1 The form of the state vector x_k means that using the Kalman filter as a state estimator will yield estimates of the noiseless speech signal s_k. □

Remark 2 Because of the special form of our signal model (the fact that it is an all pole model and single input/single output), computational simplifications arise. Also, fixed lag smoothed estimates (with lags up to $N - M$, if the order of the signal model is M) are directly available from the state estimate (the output of the Kalman filter) without further calculation. □

RESULTS

If noise sequence $\{n_k\}$ is added to a signal $\{y_k\}$ then the signal to noise ratio (SNR) of the resulting signal $\{z_k\}$ is given by

$$\text{SNR} = \frac{\sum_{k=1}^{L} |y_k|^2}{\sum_{k=1}^{L} |n_k|^2} \tag{2}$$

where both signal and noise are taken to be of length L, starting at time $k = 0$.

But the intelligibility of speech signals is not necessarily directly related to standard mathematical measures such as the SNR improvement, nor does an improvement in SNR necessarily imply that automatic comparison techniques will perform better. As a result, we must examine the performance of Kalman techniques for noise reduction in the context of working verification systems using a variety of noise models.

Types of Noise

In any normal environment, many types of noise other than white, Gaussian noise may be present. Common examples are harmonic noise (as from the electricity mains or engines), interfering speech, and impulsive noise (as from door slams). In this section we examine the performance of Kalman filtering noise reduction using a number of noise sources in the context of a dynamic time warping verification system.

Figures 2, 3 and 4 plot the dynamic time-warping distance measure versus SNR for various noise sources. The interfering signals used were, respectively: Harmonic noise, a sinusoid of random frequency was used to simulate harmonic noise. Interfering speech, four speech segments from four different speakers from the database were mixed (added) together to form the interfering speech signal, and Impulsive noise, an exponentially decaying impulse of duration 300 samples was used to simulate impulsive noise. The length was chosen so that at least three frames were affected by the impulse.

Figure 2 shows that when the interfering noise signal was harmonic, the noisy dynamic time warping distance was negligibly affected, even for relatively high noise levels (a signal-to noise ratio of 2.0 or so).

Remark 3 Figures 2 and 3 show that harmonic and impulsive (but not necessarily repetitive and impulsive) noise do not appear to greatly affect the dynamic time-warping distance measure. □

Figure 4 displays the interfering speech signal to noise ratio versus dynamic time warping results. Clearly, of the four main types of interfering signal identified earlier, this form has the most effect.

Remark 4 For the dynamic time warping distance measure used, the most performance degrading form of noise is interfering speech. □

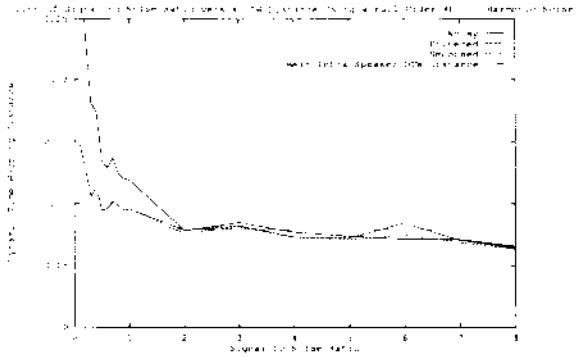

Figure 2. SNR versus DTW distance for the full order Kalman filter/smoother (Order = 10) with harmonic noise.

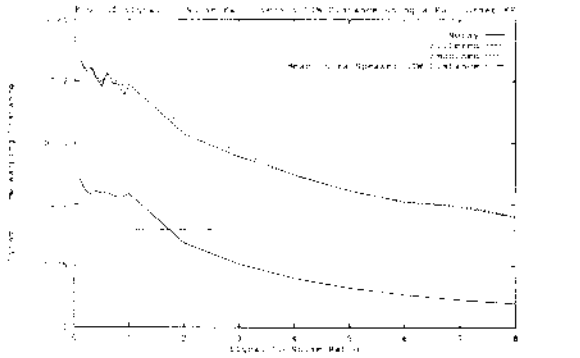

Figure 3. SNR versus DTW distance for the full order Kalman filter/smoother (Order = 10) with impulsive noise.

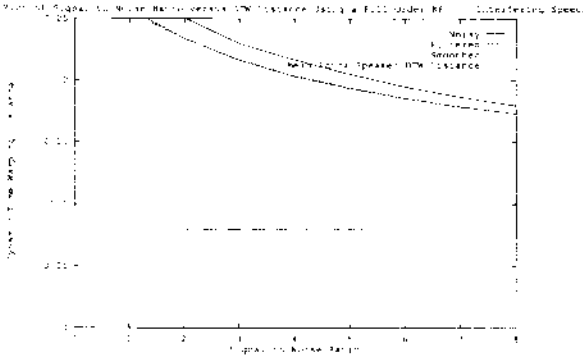

Figure 4. SNR versus DTW distance for the full order Kalman filter/smoother (Order = 10) with interfering speech.

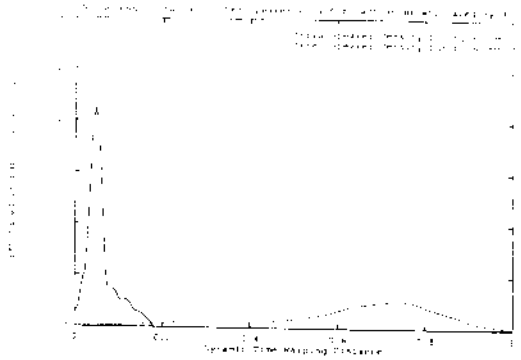

Figure 5: Noiseless intra- and inter speaker probability density functions using the dynamic time warping distance measure

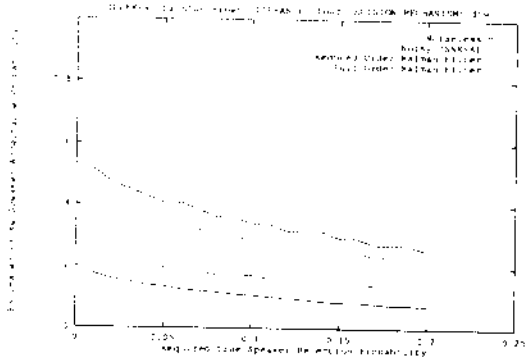

Figure 6: Example 1: Required True Speaker Rejection Rate Versus Estimated False Acceptance Rate — Dynamic Time Warping

Verification Algorithm Noise Performance Comparison

Extensive testing using a small database was carried out. In order to be consistent with previous tests carried out by the group, the only feature sets used for verification were the LPC cepstral coefficients and their first and second differences.

To compare the verification algorithms, we estimate the false-speaker acceptance error rate for various true-speaker acceptance rates. This means that the intraspeaker variability (an example of which is plotted as the solid line in Figure 5) and interspeaker variability (plotted as the dashed line in Figure 5) must first be estimated. From these, the false-speaker acceptance rate for any given true speaker rejection rate may be obtained.

Dynamic Time-Warping

The graph plotted here shows an example of dynamic time-warping noise performance working in conjunction with Kalman filtering (Figure 6).

Remark 5: The Kalman filtering performance of the dynamic time-warping technique (and, in fact, all techniques) is greatly altered depending on the quality of the reference template used. □

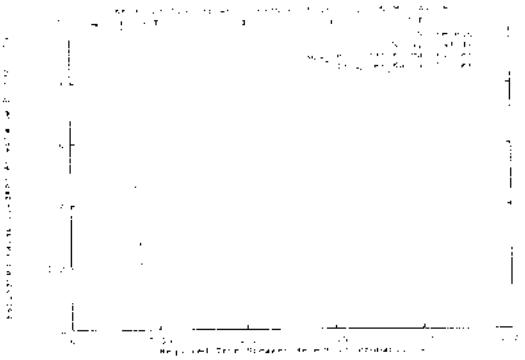

Figure 7 Example 1 Required True Speaker Rejection Rate Versus Estimated False Acceptance Rate — Recurrent Neural Network

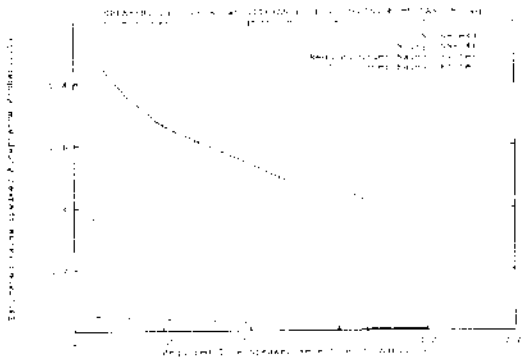

Figure 8 Example 1 Required True Speaker Rejection Rate Versus Estimated False Acceptance Rate - Vector Quantization

Recurrent Neural Network

For the technical details of this approach see [11]. A network of 82 neurons was used in the results presented here. The recurrent neural network appears to combine some good and some bad performance characteristics of the other algorithms.

In Figure 7, the noiseless performance of the recurrent net is poor — so poor, in fact, that the addition of noise appears to enhance the performance. This may be explained by noting that the addition of a stationary white noise sequence to the signal may well have the effect of reducing the intra-speaker (and inter-speaker) distribution variances, reducing the overlap of the two density functions and hence reducing the error rates.

Once the signal is Kalman filtered and smoothed in this example, and even more marked improvement is noticed.

Remark 6 In the examples shown and others tried, the neural network approach would appear to benefit the most from Kalman smoothing or filtering. □

Vector Quantization

For the technical and other details of this approach see [13] and [7].

Figure 8 shows the noise performance of vector quantization working in conjunction with Kalman filtering.

Clearly here and in other tests completed, the vector quantization approach yielded the best noiseless performance of all the systems examined.

However, as in the dynamic time warping case, since the degradation of the noisy case is not as extreme as using the dynamic time warping or recurrent neural network approaches, the noise performance is relatively good.

Remark 7: The performance of vector quantization in the noiseless case is, most instances, the best of all the systems examined. □

Remark 8: While vector quantization undergoes a marked degradation in performance in the presence of noise, its noise performance is still of the same order of or better than the other two systems. □

CONCLUSIONS

1. Kalman Filtering: Because of the special form of our signal model computational simplifications arise. The order of the model used with the Kalman filter has a marked affect on the computation time.

2. Types of Noise: For the dynamic time-warping distance measure used, the most performance-degrading form of noise is the interfering speech. These harmonic noise results may be somewhat misleading, and a worst-case simulation as discussed above should be made in order to verify these results. Impulsive (but not necessarily repetitive and impulsive) noise does not appear to greatly affect the dynamic time warping distance measures.

3. Verification performance: A satisfactory means of using multiple reference templates to produce a improved "average" reference template must be found. In the examples shown and others tried, the neural network approach would appear to benefit the most from Kalman smoothing or filtering. The performance of vector quantization in the noiseless case is, most instances, the best of all the systems examined. While vector quantization undergoes a marked degradation in performance in the presence of noise, its noise performance is still of the same order of or better than the other two systems.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Prentice-Hall, Inc., 1979.

[2] I. Barbier, C. Mokbel and G. Chollet, "Trainable Noise Subtraction Filter for Speech Enhancement in the Car," Signal Processing V: Theories and Applications, pp. 1111–1114, 1990.

[3] S. Crisafulli, J. D. Milis and R. R. Bitmead, "Kalman Filtering, Prediction and Smoothing in Speech Coding," DRAFT, September 1991.

[4] M. J. F. Gales and S. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise," To Appear in Proceedings of ICASSP 92.

[5] M. Kadirkamanathan and A. P. Varga, "Simultaneous Model Re-estimation From Contaminated Data by Composed Hidden Markov Modelling," *Proceedings of ICASSP 91*, 897-900.

[6] P. J. Kootsookos, "Speech Enhancement for Robust Speaker Verification," Internal Technical Report, Speaker Verification Group, UQ.

[7] P. J. Kootsookos, "Noise Performance of Various Speaker Verification Algorithms," Internal Technical Report, Speaker Verification Group, UQ.

[8] J. D. Markel and A. H. Gray, *Linear prediction of speech*. Springer Verlag, 1976.

[9] K. K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," Proceedings of ICASSP 87, pp. 177–180, 1987.

[10] K. K. Paliwal and M. M. Sondhi, "Recognition of Noisy Speech Using Cumulant Based Linear Prediction Analysis," Proceedings of ICASSP 91, pp. 429–432, Toronto, May, 1991.

[11] B. Watson, N. Common, A. Parr, D. Strimpton and S. Machin, "Report on Neural Implementations of Speaker Verification Modules," Internal Technical Report, Speaker Verification Group, UQ.