Speech Recognition in Impulsive Noise

S. V. Vaseghi B. P. Milner

School of Information Systems, University of East Anglia, Norwich, UK.

Abstract

This paper presents experimental results on the use of noise compensation schemes with hidden Markov model (HMM) speech recognition systems operating in the presence of impulsive noise. A measure of signal to impulsive noise ratio is introduced, and the effects of varying the percentage of impulsive noise contamination, and the power of impulsive noise, on speech recognition are investigated. For the modelling of an impulsive noise process, an amplitude-modulated binary sequence model and a binary-state HMM are considered. For impulsive noise compensation a front-end method and a noise-adaptive method are evaluated. Experiments demonstrate that the noise compensation methods achieve a substantial improvement in speech recognition accuracy across a wide range of signal to impulsive noise ratios.

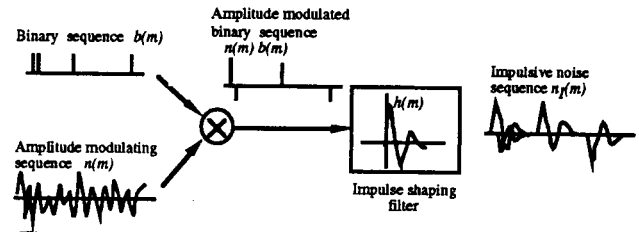
1- Introduction

Speech recognition systems operating in a practical environment, may have to deal with a wide variety of disturbances including impulsive-type noise. The sources of impulsive noise can be electronic or acoustic and include transmission errors, switching noise, adverse channel environments, or click sounds from say a computer keyboard.

The major signal processing stages in an HMM-based speech recognition system are acoustic-feature extraction, segmentation, and model likelihood calculation [1]. Noise affects each stage of the recognition process, and results in increasing deterioration in recognition accuracy, as the signal to noise ratio decreases. The effects of increasing the rate of occurrence of impulsive noise, and/or the power of impulses, on the recognition accuracy are investigated.

Noise filtering methods that assume the noise is a Gaussian, slowly time-varying, random process can not deal effectively with impulsive noise, because impulsive noise is a highly nonstationary and non Gaussian process. For the statistical characterisation of an impulsive noise process, an amplitude-modulated random binary sequence and a 2-state HMM are considered.

For noisy speech recognition the aim is to reduce the noise-induced discrepancy between the noisy signal parameters and those of the clean speech model. This may be achieved in two ways by either removing the impulsive noise from the noisy input signal or by modification of the HMMs to include the effects of the noise [2-6]. In this paper a front-end impulsive noise removal filter, and a noise adaptive model are evaluated and compared.



Figure(1) - An impulsive noise model as the output of a filter excited by an amplitude-modulated binary sequence.

2- Impulsive Noise Models

An impulsive noise sequence $n_I(m)$ can be modelled as an amplitude-modulated, binary-state, random sequence and expressed as

$$n_I(m) = n(m)\ b(m) \tag{1}$$

where $b(m)$ is a binary-valued random sequence of one's and zero's that signals the presence or the absence of a noise pulse, and $n(m)$ is a random noise process. Two statistical processes for the modelling of an amplitude-modulated binary sequence are the Bernoulli-Gaussian process and the Poisson-Gaussian process. The autocorrelation function of an uncorrelated impulsive noise process is also a binary-state process modelled as

$$r_{n_I n_I}(k,m) = \sigma_n^2\ \delta(k,1-b(m)) \tag{2}$$

where $\delta(i,j)$ is the Kronecker delta function. For an uncorrelated noise process the power spectrum of the impulsive noise model of eq(2) is

$$P_{n_I n_I}(f,m) = \sigma_n^2\ \delta(1-b(m)) \tag{3}$$

0-7803-2431-5/95 $4.00

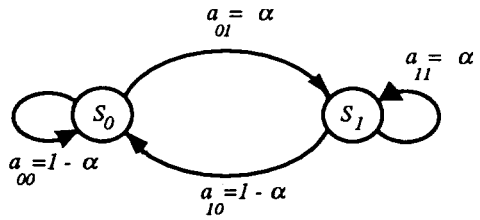
Assuming that the amplitude of a noise pulse is a zero mean Gaussian process with variance σ_n^2, $\mathcal{N}(0,\sigma_n^2)$, the pdf of an impulsive noise process may be defined as

$$f_{N_I}(n) = (1-\alpha)\delta(n) + \alpha\mathcal{N}(0,\sigma_n^2) \qquad (4)$$

where α is the probability of occurrence of an impulse.
In a communication system, an impulsive noise originates at some point in time and space and propagates through the channel to the receiver. At the channel output, the impulsive noise is shaped by the characteristics of the channel, and may be considered as the channel impulse response. The duration of an impulsive noise depends on the source of the noise and the channel response and may vary from a few microseconds to a few tens of milliseconds. A real impulsive-type noise sequence, $n_I(m)$, can be modelled, figure(1), as the output of a channel filter excited by an amplitude-modulated random binary sequence as

$$n_I(m) = \sum_{k=0}^{P-1} h(k)n(m-k)b(m-k) \qquad (5)$$

where $h(m)$ is the impulse response of a filter that models the duration and the shape of each impulse.
An alternative model for an impulsive noise sequence is the two-state HMM shown in figure(2). In this binary model, the state S_0 corresponds to the *'off'* condition when impulsive noise is absent. In this state the model 'emits' zero-valued samples. The state S_1 corresponds to the *'on'* condition. In this state the model emits short duration pulses of random amplitude and duration. The probability of transition from state S_i to state S_j is denoted by a_{ij}. The impulsive noise state S_1 can be configured to accommodate a variety of noise pulses of different shapes, duration and pdf, using a codebook of M prototypes of impulsive noise, and their associated pdfs.



Figure(2)- A binary-state HMM of impulsive noise. With the values of the transition probabilities as shown, the likelihood of occurrence an impulsive noise is independent of the state.

3- Signal to Impulsive Noise Ratio

Let $P_{Impulse}$ denote the average power of each impulse, and P_{Signal} the signal power. A 'local' time-varying signal to impulsive noise ratio can be defined as

$$SINR(m) = \frac{P_{\text{Signal}}(m)}{P_{\text{Impulse}}\delta(1-b(m))} \qquad (6)$$

For impulsive noise, the average signal to impulsive noise ratio (averaged over a long noise sequence and including the instances when the impulses are absent), depends on two parameters: (a) the average power of each impulsive noise, and (b) the rate of occurrence of impulsive noise. An average signal to impulsive noise ratio, assuming that α is the fraction of signal samples contaminated by impulsive noise, can be defined as

$$SINR = \frac{P_{\text{Signal}}}{\alpha \,.\, P_{\text{Impulse}}} \qquad (7)$$

Note that for a given signal power, many different values of α and P_{signal} can yield the same average SINR.

4- Impulsive Noise Detection

In this section we consider a front-end method and a model-based method for the detection of an impulsive noise sequence.

4.1- Pulse Detection using Inverse Predictor

An impulsive-type noise introduces uncharacteristic discontinuity in a correlated signal. The detectability of a noise pulse, observed in a high level of correlated signal, can be improved by a decorrelation (spectral whitening) operation, which has the effect of enhancing the amplitude of an impulsive type event relative to the "background" signal. The correlation structure of the signal $x(m)$ may be modelled by a linear predictor, and the noisy signal $y(m)$ can be described as

$$y(m) = \sum_{k=1}^{P} a_k x(m-k) + e(m) + \delta(b(m)-1)n(m) \qquad (8)$$

where a_k is the k^{th} linear predictor coefficient, and $e(m)$ is the speech excitation. The process of de-correlation is performed by an inverse predictor filter[3]. Inverse linear prediction is a differencing operation, it makes the discontinuities in a correlated signal more detectable. An alternative interpretation is that the inverse filtering is equivalent to a spectral whitening operation ; it effects the energy of the coloured signal spectrum whereas the, theoretically flat, spectrum of the impulsive noise is largely unaffected.

4.2- Noise Detection Based on the ML State Sequence

The maximum likelihood (ML) state sequence of a hidden Markov model of impulsive noise, figure(2), can be used

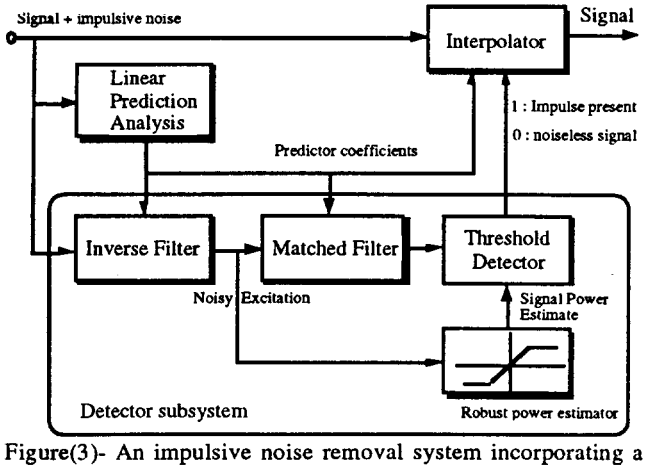
as a detector of the presence or the absence of impulsive noise. For a given observation sequence $y = [y(0), y(1), ..., y(N-1)]$, the maximum likelihood state sequence $s = [s(0), s(1), ..., s(N-1)]$, of an HMM λ is obtained as

$$s_{ML} = \arg\max_{s} f_{Y|S,\Lambda}(y|s,\lambda) \qquad (9)$$

where $f_{Y|S,\Lambda}(y|s,\lambda)$ is the pdf of the observation sequence y along state sequence s of model λ. The ML state sequence is derived using the Viterbi algorithm.
A problem in using HMMs for the detection of impulsive noise is the sensitivity of the accuracy of the ML state sequence to the presence of a background signal. A solution is to decorrelate the speech using an inverse linear predictor, and to train the HMMs on impulsive noise observed in a decorrelated, white, background signal. An alternative solution is to use HMMs that combine the speech and the impulsive noise states [5,6].

5- Front-End Impulsive Noise Removal

A typical impulsive noise sequence leaves a large fraction of speech samples unaffected. Thus it is advantageous to locate the individual noise pulses, and correct only those samples which are distorted. The front-end impulsive noise removal system evaluated in this paper is composed of a detector and an interpolator as shown in figure(3), [3,4]. The detector locates the position of the individual impulses and the interpolator replaces the distorted samples using speech samples on both sides of the impulsive noise. The output of the detector is a binary switch which controls the interpolator. A detector output of '0' signals the absence of impulsive noise and the interpolator is bypassed, a detector output of '1' signals the presence of impulsive noise and the interpolator is activated to replace the samples obliterated by noise.



Figure(3)- An impulsive noise removal system incorporating a detection and an interpolation subsystems.

Speech samples distorted by impulsive noise are discarded and replaced using a least mean squared error interpolator [3]. The interpolator is based on a linear prediction model of speech, and makes effective use of the undistorted samples on both sides of the discarded speech samples. The interpolator works well for the replacement of missing speech segments of upto 50 samples at a 10 kHz sampling rate.

6- Speech and Noise Model Combination

An alternative method to the filtering of noisy speech signal is to modify the speech models to include the statistical effects of the noise on the speech parameters. In parallel model combination an HMM of a speech signal is combined with an HMM of noise to produce an HMM for the noisy speech observation signal. For each state of the clean speech HMM there are two states in the combined model, corresponding to whether impulsive noise is present or absent [5,6].

7- Experimental Results

The noise compensation techniques were evaluated using the NOISEX spoken digit database, with machine gun noise and simulated short duration impulsive noise. The digits were modelled using an 8 state, single mode per mixture HMM, with a diagonal covariance matrix. To generate features, the speech was Hamming windowed every 16ms with a window width of 32ms. Each signal window was transformed into a feature vector of 25 mel-spaced filterbank channels. This was then converted to 14-dimensional MFCC features.

Effects of Impulsive Noise on Speech Recognition- Experimental results on the effects of varying the frequency of occurrence and the amplitude of impulsive noise on the performance of HMM-based speech recognition systems are tabulated in table-1, which shows the recognition performance for speech contaminated by simulated impulsive noise. In this experiment both the percentage of speech samples contaminated by impulsive noise, and the overall signal to impulsive noise ratio (SINR), have been varied.

SINR(dB) / %	30	20	10	0	-10
1	100	100	100	96	88
5	100	93	68	32	20
10	100	90	51	26	17
20	100	90	44	23	13

Table(1) - Recognition performance for speech contaminated by impulsive noise.

Table-1 shows that as more speech samples are contaminated by impulsive noise, or as the impulsive noise power increases, the performance of the recogniser deteriorates. Note that, for a given signal to impulsive noise ratio, increasing the percentage of samples corrupted

by impulsive noise means that the average impulse amplitude is decreased. From the columns of table-1, at a given SNR, as the frequency of occurrence of impulses increase the recognition performance deteriorates. As expected, a few large impulses have a lesser degrading effect than a large number of small amplitude impulses .

Matched and Un-matched Conditions

Un-matched conditions form the worst case, due to the mismatch between the models trained on clean speech and tested on noisy speech. With matched conditions the models are trained and tested with speech contaminated under similar noise conditions, which should indicate the best performance the system can achieve.

Impulsive Noise Compensation

Figure(4) shows the performance of speech recognition in the presence of a machine gun noise. Machine gun noise can be considered as the impulse response of the machine gun and the acoustic environment. As such they have a relatively long and well defined shape. The method of parallel model combination works well for the longer duration impulses with a relatively high amplitude, with performance approaching that of the matched conditions. Figure(5) shows that for shorter duration impulses the front-end combination of noise detection and interpolation improves the performance considerably at a signal to noise ratio as low as 0 dB.

8- Conclusion

A number of nonstationary stochastic processes for the modelling of an impulsive noise sequence have been considered. For speech recognition in impulsive noise, a front-end method and the model combination method are evaluated. The model combination method achieves good performance for longer duration noise pulses such as a machine gun noise. Experimental results indicate that the front-end noise removal method is more effective in compensating for short duration impulses. This may be due to the utilisation of the distinct and localised character of an impulsive noise in the time domain by the front-end system. Whereas, the model combination method, using the same frequency-based features as that of the speech model, compensates for the effects of the noise in the frequency spectrum.

References

[1] Deller J., Discrete-time processing of Speech Signals, Macmillan 1993.

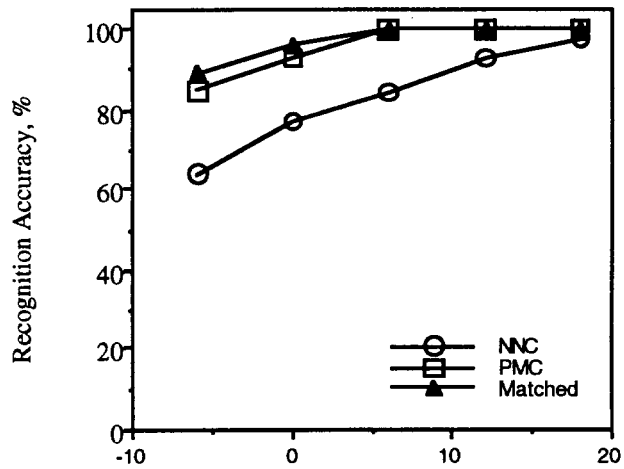
[2] Juang B.H., "Speech recognition in adverse environments," Computer Speech and Language, pp. 275-294, 5 1991.

[3] Vaseghi S. and Rayner P.,"Detection and suppression of impulsive noise in speech communications systems.", IEE Proc-I Communications Speech and Vision, pages 38-46, Feb 1990.

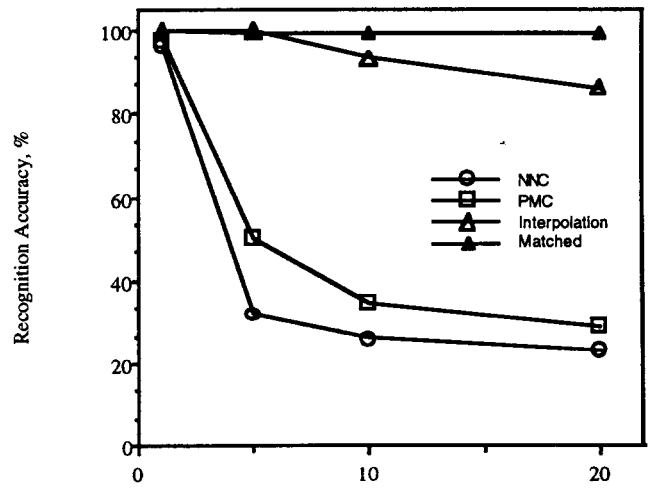
[4]Godsill S.J., "The restoration of degraded audio signals", PhD Thesis, Cambridge University, UK 1993.,

[5] Gales,M., Young S.,"An improved approach to the hidden Markov model decomposition of speech and noise", IEEE Proc., ICASSP-92, pages I-223--226.

[6]Varga A., Moore R.K. , "Hidden Markov model decomposition of speech and noise", Proc. IEEE Int., Conf. on ICASSP-1990, pages 845-848.



Figure(4) - Recognition performance for machine gun noise. Key : NNC=No Noise Compensation, PMC=Parallel Model Combination, Matched = matched conditions.



Figure(5) - Recognition performance for short duration impulsive noise, with SINR = 0 dB.