Proceedings of the Institute of Acoustics

SPEECH RECOGNITION IN NOISE: EXPERIMENTS USING HIDDEN MARKOV MODELS

V.L. Beattie & S.J. Young

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ

ABSTRACT

Implementing speech recognition systems in practical applications requires strategies for handling the inevitable environmental noise. Although some noise cancellation can be accomplished during the preprocessing of the signal, it is also necessary to consider how the recognition component of the system is affected by different noise levels, and how it can be enhanced for greater noise robustness.

This paper describes a new algorithm for improving the performance of a Hidden Markov Model (HMM) recognition system in noisy environments. The method of state-based smoothing exploits the properties of Hidden Markov Models in order to reduce the effects of noise during recognition. The algorithm developed consists of an adaptation of the standard Hidden Markov Model Viterbi scoring procedure. In this adaptation, similar vectors are averaged together, thus smoothing out effects due to noise.

1. INTRODUCTION

Our initial studies of recognition performance in noisy conditions indicate that Hidden Markov Model-based systems are sensitive to even low levels of noise. This is the case even when the front-end processor is relatively noise robust (e.g. a filterbank). Attempts to combat the problem of noise are usually undertaken in the preprocessing stage of speech recognition. However, our work indicates that noise cancellation may profitably be incorporated into the Hidden Markov Model recognition phase itself. Algorithms employed in the recognition phase can take advantage of the specific information about the speech signal embodied in the HMM parameters.

Many speech recognition environments of interest, for example office, automobile, or factory, involve some background noise sources whose characteristics are changing slowly relative to the speech signal itself. The problem in this case is not in obtaining information about the noise, which may be gathered during periods of "silence" in the input, but rather in applying this information.

One option is to implement noise cancelling filters during the initial processing of the signal [1,2]; however, at this stage little specific information about the speech signal is available. The filters can therefore only apply general knowledge about the characteristics of the desired signal, such as short term coherence for voiced speech. Subtraction of the noise energy levels from the noise contaminated speech can also be applied for spectral parametrisations, but this simply shifts the problem to one of zero-mean noise. For the case of random noise with a high variance, this shift to zero-mean will not, by itself, offer much performance improvement.

A significant amount of work has also been done in the area of noise compensation within the recognition phase [3,4,5], with considerable success. These approaches take the presence of noise into account by modelling an independent noise source as well as the speech source. Some relationship is assumed between the two sources, either an additive relationship or (more frequently) a "masking" relationship, where either one source or the other is assumed dominant. Recognition scoring is then changed to reflect the combined model of the unknown which takes into account both speech and noise. Since the same noise model or noise mask is applied for all of the speech HMMs, however, these approaches weaken the discriminant capabilities of the HMM-based system.

This research forms part of Esprit II Project p2101, "Adverse-environment Recognition of Speech"

The method described here, in contrast, attempts to eliminate noise from the input speech. In this way it is more closely related to work such as that by Ephraim et al [6], which addresses the problem of iteratively cleaning noisy speech. Like that work, our method uses the capacity of Markov Models to segment speech into quasi-stationary segments. However Ephraim et al. describe a method for iteratively cleaning the speech waveform, whereas the research discussed here is concerned with recognition phase reduction of noise in sequences of parameterised speech vectors.

2. ALGORITHM

We are interested in improvements to the recognition phase of the HMM system. The recognition task consists of matching an unidentified sequence of observation vectors

$$O = (O_1, O_2, \ldots O_T)$$

against each of the models available. The models are defined by the following parameters:

N	number of states
a[i,j]	probability of transition from state i to state j
$b_i(O_t)$	probability of state i emitting output vector O_t (output probability density function)

For simplicity left-to-right models are assumed where at the start of processing we are assumed to be in the first state, i.e. $S_0 = 1$.

Ideally the matching process for each model finds the state sequence having the highest probability of producing the observed sequence of speech vectors. If we know this optimal state sequence,

$$S = (S_1, S_2, \ldots S_T)$$

then the probability P(O|M) of the model producing the observed sequence can be obtained as follows:

$$P(O|M) = \Phi_T \qquad (2.1)$$

where

$$\Phi_0 = 1$$

and

$$\Phi_t = \Phi_{t-1}\, a[S_{t-1}, S_t] b_{S_t}(O_t) \quad t=1, .T \qquad (2.2)$$

The model M which yields the highest P(O|M) value over the whole observation sequence is the recognition system output, i.e. a guess at the identity of the unknown speech.

Recognition phase decoding of the unknown is aimed at finding the optimal state sequence S. In particular, the Viterbi algorithm [7] is in widespread use, and it is an adaptation to the Viterbi algorithm which is discussed here. However, the method of state-based smoothing is equally applicable to other decoding algorithms.

In a standard Hidden Markov Model recognition system, Viterbi decoding of an unknown observation sequence proceeds as follows:

$$P(O|M) = \max_{\text{over } i} \{\Phi_T(i)\} \text{ for all states } i=1..N \qquad (2.3)$$

where

$$\Phi_0(i) = \begin{cases} 1 & i=1 \\ 0 & \text{otherwise} \end{cases} \qquad (2.4)$$

and

$$\Phi_t(j) = \max_{\text{over } i} \{\Phi_{t-1}(i) a[i,j]\} b_j(O_t) \quad \text{for } i,j=1..N \text{ and } t=1..T \qquad (2.5)$$

This decoding effectively divides a sequence of observations into segments corresponding to the states of the HMMs. For the left-to-right models which are employed for modelling speech, each state is visited only once. Word models typically contain from 3 to 8 states, and word duration ranges from 0.5 to 1.5 seconds, or 50 to 150 observations if a new parameter vector is produced every 10 milliseconds. Thus on average we expect to remain in one state for several contiguous observations.

The aim of adapting the Viterbi algorithm is to improve recognition performance by allowing HMM recognition scores to be based on an average of several input vectors, rather than on the single most recent vector. The method is based on the assumption that Hidden Markov models segment speech into quasi-stationary segments corresponding to the states of the model. Within a segment we expect the parameter vectors of speech (irrespective of any noise distortion) to remain fairly constant. Therefore, the vectors in a segment may be averaged together without a significant loss of information about the speech. Averaging will, however, have the desired effect of eliminating random variations in the input due to noise.

The state-based smoothing algorithm is outlined below. We use the following definitions:

- $n_{i,t}$ the length of time spent in state i up to and including time t
- $\hat{O}_{i,t}$ the average of all observation vectors assigned to state i up to and including time t
- $\Phi_{init}(j)$ the initial probability score for state i, assigned upon entering that state

Immediately before the calculation of $\Phi_t(j)$, the following parameter updates take place:

$$n_{i,t} = n_{i,t-1} + 1 \qquad (2.6)$$

$$\hat{O}_{i,t} = \frac{(n_{i,t}-1)\,\hat{O}_{i,t-1} + O_t}{n_{i,t}} \qquad (2.7)$$

Then $\Phi_t(j)$ is calculated according to the following equations, which replace equation 2.5 of the standard Viterbi algorithm:

$$\Phi_t(j) = \max_{\text{over } i} \{\psi_t(i,j)\} \qquad \text{for } i,j=1..N \text{ and } t=1..T \qquad (2.8)$$

where

$$\psi_t(i,j) = \begin{cases} \Phi_{t-1}(i)\, a[i,j]\, b_j(O_t) & i \neq j \qquad (2.9a) \\ \Phi_{init}(i)\, a[i,j]^{n_{i,t}-1}\, b_j(\hat{O}_{i,t})^{n_{i,t}} & i=j \qquad (2.9b) \end{cases}$$

If the maximum value yielded by equation 2.8 corresponds to the case i≠j, then a transition from state i into state j has occured. In this case the parameters for state j must be re-initialised:

$$\hat{O}_{j,t} = O_t \quad (2.10)$$

$$n_{j,t} = 1 \quad (2.11)$$

$$\Phi_{init(j)} = \Phi_{t-1}(i)\, a[i,j] \quad (2.12)$$

The noise-cancelling effect of the algorithm lies in the application of equation 2.9b for the self-transition case. Equation 2.9a, on the other hand, is identical to the standard Viterbi algorithm. Equation 2.9b can be interpreted as follows: the score for self-transition is calculated as if the average vector had been the observed input for the entire time spent thus far in state i. In other words, assuming we entered state i at time τ and remained in it up to and including time t, the score is identical to that which would have been calculated via normal Viterbi if the input observation sequence for time ($\tau, \tau+1, \ldots t$) had been:

$$(\hat{O}_{i,t}, \hat{O}_{i,t}, \ldots \hat{O}_{i,t})$$

Thus for self-transitions (remaining in one state of a model) the probability score is calculated based on an average of all the vectors assigned thus far to a given HMM state. For transition from one state to another the assumption of quasi-stationarity cannot be made since we are assumed to be in transition between one "stable" segment and the next. Therefore only the most recent noisy vector is used to update the probability score.

The efficacy of state-based smoothing depends on zero-mean noise at the input to the recognition phase. For an energy level based preprocessor such as a filterbank, additive noise in the parameter vectors will not be zero mean. As noted previously, however, general parameters of the noise such as mean energy levels and variances may be obtained from periods of silence if the noise source is changing slowly. The noise mean may then be subtracted at the output of the filterbank, before further processing takes place, to achieve the desired zero-mean noise condition.

3. EXPERIMENTAL SETUP

The recognition system used to test the state-based smoothing algorithm used 3-state continuous output probability density Hidden Markov Models. The output pdf for each state was a single multivariate Gaussian with a diagonal covariance matrix. Model parameters were determined using well-established statistical training methods [8,9].

The experimental data consisted of the ten digits recorded in the car environment. The clean portion of the database was recorded in anechoic conditions; noisy data was obtained in a vehicle driving along the motorway at 100 km/h. At these high motorway speeds the background noise is dominated by aerodynamic noise whose spectral distribution is roughly flat over the range 0.5 to 4.0 KHz, the range which also contains the most useful information about the speech signal.

The models were trained and tested on the log energy levels of a 19 channel filterbank, sampled at 10 millisecond intervals. The filter bandwidths and center frequencies are approximately mel-frequency spaced, and are based on the vocoder design described in [10]. Clean data was used for training the models; data recorded in a noisy environment, or synthesised by mixing recorded noise with clean data, was used for testing.

Mixing was accomplished by combining periods of noise recorded on the motorway with clean speech repetitions. The waveforms were first scaled so that the overall waveform energy ratio corresponded to the desired signal-to-noise ratio, and then added together and processed via the filterbank. All of the training and testing data was manually endpointed. This and the synthetic mixing of speech and noise introduce a certain degree of artificiality into the

system. However it enables us to evaluate the algorithm over a wide range of noise conditions, as well as independently of other system factors such as the accuracy of automatic endpointing.

4. RESULTS

The graph below indicates the results for noisy speech recorded on the motorway, as well as for clean data mixed with motorway noise at several different signal-to-noise ratios.

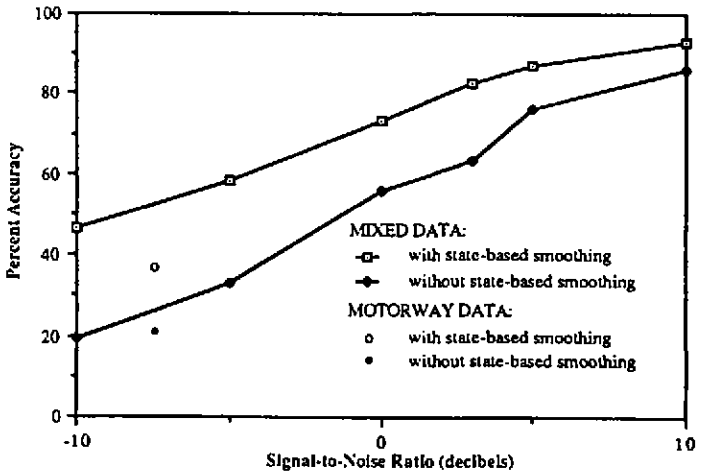

Figure 4.1 - Performance of State-Based Smoothing Algorithm

The performance gains indicate significant improvement over a wide range of signal-to-noise ratios. Particularly large improvements occur for low SNRs, in the range -10 to +3 dB. Thus the algorithm may be well suited for use in environments characterised by high levels of background noise. The weaker performance of the algorithm on high signal-to-noise ratios may be due to the approximate nature of the assumptions made. At these SNRs performance without the algorithm is already fairly high, and the amount of speech information lost through the averaging of several input vectors may be significant relative to the noise cancellation obtained. Moreover the adaptation of the Viterbi equation is biased towards self-transition. Score updates for self-transition can take advantage of the smoothing effect, whereas transitions from one state to the next are based on a single noisy vector alone. Thus it is possible that the noise cancelling method may introduce segmentation errors of its own.

The general applicability of the method described here remains to be determined through testing in a variety of noise conditions, recognition styles, and vocabulary sizes. However the results obtained here suggest that this straightforward method may be effective in enhancing the noise robustness of HMM recognition systems. Perhaps even more importantly, the success of this initial work suggests further research directions, particularly in applying more sophisticated filtering methods to the smoothing process. As an extension of this work, research is currently underway into the development of HMM state-based Wiener filtering.

REFERENCES

[1] D. Mansour & B.-H. Juang, "The Short-Time Modified Coherence Representation and Noisy Speech Recognition," IEEE Trans. on ASSP, ASSP-37(6), pp. 795-803, 1989.

[2] M. Feder & A.V. Oppenheim, E. Weinstein, "Maximum Likelihood Noise Cancellation Using the EM Algorithm," IEEE Trans. on ASSP, ASSP-37(2), pp. 204-216, 1989.

[3] A.P. Varga & K.M. Ponting, "Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous Word Recognition," ESCA Proc. Eurospeech'89, pp. 167-170, Paris, Sept. 1989.

[4] A.P. Varga & R.K. Moore, "Spectral Decomposition of Speech and Noise," IEEE Proc. Int. Conf. Acoust. Speech & Signal Process., ICASSP'90 pp. 845-848, 1990.

[5] A. Nadas, D. Nahamoo & M.A. Picheny, "Speech Recognition Using Noise-Adaptive Prototypes," IEEE Proc. Int. Conf. Acoust. Speech & Signal Process., ICASSP'88, pp. 517-520, 1988.

[6] Y. Ephraim, D. Malah & B.-H. Juang, "Speech Enhancement Based on Hidden Markov Modeling," IEEE Proc. Int. Conf. Acoust. Speech & Signal Process., ICASSP'89, pp. 353-356, 1989.

[7] A.J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm, IEE Trans. Info. Theory, IT-13(4), 1969.

[8] L.A. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," IEEE Trans. Info. Theory, IT-28(9), 1982

[9] M.J. Russell, "Experiments in Speaker Dependent Isolated Digit Recognition Using Hidden Markov Models," Proc. Conf. Acoustics 1986

[10] J.N. Holmes, "The JSRU Channel Vocoder," IEE proc. F, vol.127, no.1, pp. 53-60, Feb. 1980.