HIDDEN MARKOV MODEL DECOMPOSITION OF SPEECH AND NOISE

A.P. Varga and R.K. Moore

Speech Research Unit, Royal Signals and Radar Establishment,
St Andrew's Road, Malvern, WR14 3PS, Great Britain.

ABSTRACT

This paper addresses the problem of automatic speech recognition in the presence of interfering signals and noise with statistical characteristics ranging from stationary to fast changing and impulsive.

A technique of signal decomposition using hidden Markov models, [1], is described. This is a generalisation of conventional hidden Markov modelling that provides an optimal method of decomposing simultaneous processes. The technique exploits the ability of hidden Markov models to model dynamically varying signals in order to accomodate concurrent processes, including interfering signals as complex as speech.

This form of signal decomposition has wide implications for signal separation in general and improved speech modelling in particular. However, this paper concentrates on the application of decomposition to the problem of recognition of speech contaminated with noise.

1 INTRODUCTION

There are two fundamentally different approaches to the recognition of speech contaminated by noise. The first attempts to derive the best estimate, in some sense, of the speech signal from the contaminated signal. This may be achieved through some pre-processing technique, such as spectral estimation or adaptive filtering. The second approach is not to pre-process the signal but rather to allow for the presence of noise in the recognition process itself. Here the problem of dealing with noise contamination in the recognition phase of a hidden Markov model based recogniser is one of obtaining the best estimate of the likelihood of an input observation conditioned on a particular state of the model and given the knowledge available about the contaminating noise. A previous paper, [2], developed three noise compensation techniques that used the latter approach. The results showed that such an approach works well and in particular that the masking principle of Klatt, [3], gives the best performance.

Signal decomposition using hidden Markov modelling, introduced in this paper, is an optimal theoretical framework for the second approach. It provides a new and significantly enhanced technique for the recognition of speech contaminated with other signals, such as background noise, and can be used to deal with noises that have statistical characteristics ranging from stationary to highly time varying and impulsive. Decomposition also provides a framework from which the three noise compensation techniques can be understood and their relative performances explained.

The advantage of decomposition over the previous approaches is that it provides an optimal method for recognising the speech and the noise simultaneously. Since hidden Markov models can model dynamically varying signals, the technique makes it possible to deal with structured and highly time varying noise, e.g. background talkers, key clicks or a door slam.

2 SIGNAL DECOMPOSITION USING HIDDEN MARKOV MODELS

Signal decomposition using hidden Markov modelling, [1], is a general technique in which concurrent events are recognised simultaneuosly. This is acheived by using parallel hidden Markov models, one set for each of the components into which the signal is to be decomposed.

Consider a signal made up of two separate components added together. The two individual components can be modelled by conventional HMMs, and the signal resulting from the combination of the two components can be modelled as a function of their combined outputs. The observation probability evaluated for the combined effect of the simultaneous HMMs is thus:

$$Observation\ Probability = P(Observation|M1 \otimes M2)$$

where $M1$ and $M2$ are the parallel hidden Markov models of the simultaneous components, and $\otimes$ is any combination operator, e.g. addition, multiplication, convolution etc. Recognition is carried out by extending the normal Viterbi decoding algorithm to a search of the combined state-space of the two models.

In the normal Viterbi process the recurrent relation for evaluating the most likely state sequence is:

$$P_t(i) = \max_u P_{t-1}(u).a_{u,i}.b_i(O_t) \qquad (1)$$

where $P_t(i)$ is the probability of being in state i at time t, $a_{u,i}$ is the transition probability from state u to state i, and $b_i(O_t)$ is the probability of the observation O_t coming from state i.

In the case of decomposition for two simultaneous components the relation becomes:

$$P_t(i,j) = \max_{u,v} P_{t-1}(u,v).a1_{u,i}.a2_{v,j}.b1_i \otimes b2_j(O_t) \qquad (2)$$

where $P_t(i,j)$ is the probability, at time t, of the first component being in state i and the second in state j: $a1_{u,i}$ is the transition probability from state u to state i for the first component; $a2_{v,j}$ is the transition probability from state v to state j for the second component; $b1_i \otimes b2_j(O_t)$ is the observation probability. Evaluation of the observation probability will take the general form

CH2847-2/90/0000-0845 $1.00 © 1990 IEEE

$$b1_i \otimes b2_j(O_t) = \int P(O1_t, O2_t|i,j) \quad (3)$$

where the integration is over all couples $(O1_t, O2_t)$ such that:

$$O_t = O1_t \otimes O2_t$$

Using equation (2) the optimal state sequence for each of the simultaneous models sets can be found, thus carrying out recognition of simultaneous signal components by searching the 3-dimensional lattice of the state-space shown in figure 1. The extension to more than two components is straight forward.

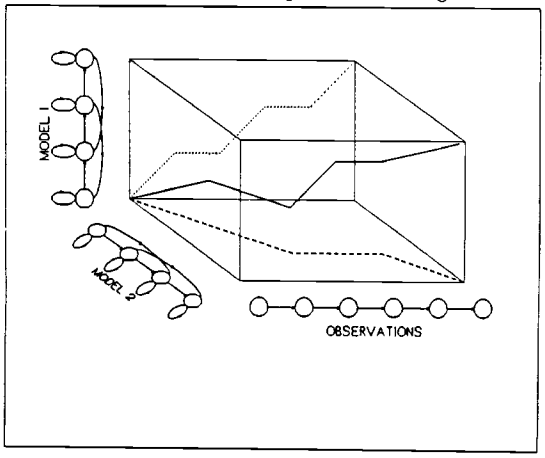

Figure 1: *Decomposition of 3-dimensional state-sequence into two 2-dimensional projections in the M1 and M2 state spaces.*

3 DECOMPOSITION FOR RECOGNITION OF SPEECH IN NOISE

In the case of speech contaminated with noise, the components of the decomposition are the usual speech models of whole words (for example) and a second concurrent set of noise models. The observation probabilities are evaluated on the basis of the output of a speech model combined with the output of a noise model. Recognition of the speech is carried out with one set of models while the noise is "recognised" simultaneously by the other set of models.

In general an observed signal will consist of various component signals combined together in some way, i.e. by means of some combination operator or set of operators. In the examples used in this paper the observed signal consists of speech with noise added to it at various signal-to-noise ratios. This is passed through a filter bank front end which generates log energy levels in each channel, i.e.

$$O_t = \log(O'1_t + O'2_t). \quad (4)$$

where $O'1_t$ & $O'2_t$ are the energy levels of the two components, the speech and the noise. To a first approximation it is possible to write:

$$O_t = \log(O'1_t + O'2_t) \approx \max(O1_t, O2_t) \quad (5)$$

where $O1_t = \log(O'1_t)$ and $O2_t = \log(O'2_t)$. Thus it is possible to approximate the required integration, in equation (3), for evaluation of the observation probability as follows:

$$b1_i \otimes b2_j(O_t) = P(\max(O1_t, O2_t)|i,j) = \quad (6)$$
$$C(O1_t, \mu1_i, \sigma1_i^2)\mathcal{N}(O2_t, \mu2_j, \sigma2_j^2) +$$
$$C(O2_t, \mu2_j, \sigma2_j^2)\mathcal{N}(O1_t, \mu1_i, \sigma1_i^2)$$

where $C(O_t, \mu, \sigma^2)$ is the cumulative probability of all observation levels less than O_t coming from a Normal distribution with mean μ and variance σ^2. Similarly $\mathcal{N}(O_t, \mu, \sigma^2)$ is the probability of observation O_t coming from a Normal distribution with mean μ and variance σ^2.

This combination operator exploits the fact that in a filter bank the effect of noise in a particular frequency band or channel has only a limited effect (due to passband overlap) across the rest of the spectrum.

In a recent paper, [4], Nádas *et al.* also recognised equation (6).

4 RELATIONSHIP WITH KLATT MASKING

In the Klatt noise masking algorithm, [2], both the input observation and the template are "masked" with a noise threshold that depends on the composite of the estimate of the noise during training and the estimate of the input noise. Re-phrasing this in hidden Markov model terms; if either the model mean or the observation is below the composite noise mask then it is replaced with the noise mask, otherwise the distance is calculated as usual. The observation probability is thus evaluated as:

$$P(max(O,N)|i:\mu_i = max(\mu_i, N)) =$$
$$\mathcal{N}(max(O,N), max(\mu_i, N), \sigma^2)$$

Where O is the observed energy level, N is the current noise estimate, μ is the model mean and σ^2 is the model variance. Thus the observation probability of the masked observation is evaluated as coming from a Normal distribution with masked model mean; the variance is that of the model. It can be seen that this turns out to be an interesting approximation to equation (6).

5 EXPERIMENTS AND RESULTS

The objective of the experiments was to compare the performance of decomposition with that of a conventional baseline recogniser and the Klatt masking technique.

5.1 Experimental data

The speech data used were isolated digits, extracted from the NATO RSG-10 single digit database, [8]. These consist of several continuous tables each of 100 digits spoken in isolation. One table was used to train the models, one table was used for parameter optimisation and the remaining three tables for tests. The noise data were taken from the NATO RSG-10 noise database [9]. The noises used in the experiments reported here were stationary pink noise (equal energy in 1/3 octaves) and machine-gun noise (an example of highly impulsive irregularly time varying noise).

The speech and the noise signals were recorded separately and added together digitally at seven different signal-to-noise ratios: +21, +15, +9, +3, −3, −9 and $-15dB$. The signal-to-noise ratio was calculated on the basis of signal level measurements made using the British Telecom SV6 speech voltmeter. The SV6 conforms to the CCITT standard, [10], for speech level measurement.

The corresponding segmental signal-to-noise ratio, [2], and global signal-to-noise ratio were calculated for comparison: segmental SNR is roughly $7-8dB$ lower than the SNR values given above; the global SNR, calculated only over speech regions, is roughly $2dB$ lower than the values given above.

5.2 Experimental setup

The recognisers all had a single microphone input with no extra sensors. The one-pass continuous speech recognition algorithm, [5], was used. The observation vectors were the log energy levels of a 27 channel filter bank analyser, [6]. The channels of the filter bank are roughly critical band spaced and overlapping, based on a successful channel vocoder design, [7]; the frequency range covered is 0 – 10kHz. The channel energies were quantised with 8 bits in $0.5dB$ steps; the filter bank analysis was carried out at a rate of 100 frames per second.

Ten-state left-to-right speaker-dependent whole-word hidden Markov models were used and the output distribution for each state was multi-variate Normal with diagonal covariance matrix. The speech models were all trained under noise-free conditions. Separate training, optimisation and test sets were used.

In the case of speech with pink noise the baseline recogniser and the Klatt recogniser both used a noise tracking silence model to recognise the non-speech periods, details of this can be found in [2]. A true silence model, i.e. zero means, was used together with the word models in the speech component for the decomposition test, a single state pink noise model, described in section 5.3, was used as the second (noise) component.

In the case of speech with machine gun noise the baseline recogniser used the five-state machine gun model described in section 5.3 together with a noise tracking silence model and the digit models. The results for the Klatt algorithm were obtained without the use of the machine gun model, simply a noise tracking silence model. It was found that the use of a machine gun model together with Klatt masking worked very poorly (no investigation of the reasons for this were made). The decomposition results were obtained as with the pink noise, i.e. a true silence model together with the digits in the speech component and the five-state machine gun model as the second (noise) component.

5.3 Noise models

Models of stationary pink noise and machine gun noise were built using a standard Baum-Welch re-estimation algorithm. The pink noise model was a single state model, consisting simply of the means and variances of the noise in each channel of the filter bank front end. The machine gun was modelled with a five state non left-right model. A minimum threshold on the magnitude of the model variances was set. The magnitude of this threshold was found empirically using the optimisation data set.

5.4 Results

Tables 1 and 2 give the result for recognition of speech with pink noise. Table 1 shows the number of digits misclassified out of a test set of 300, table 2 gives the corresponding number of insertions. It can be seen that decomposition gave the best overall performance.

SNR in dB	∞	+21	+15	+9	+3	−3	−9	−15
Baseline	0	67	202	282	*	*	*	*
Klatt	0	0	0	2	6	68	248	*
Decomp	0	0	0	0	0	9	163	*

Table 1: *Words not correctly recognised out of 300 digits spoken in isolation for each algorithm at various signal-to-noise ratios with added pink noise. (* indicates that no words were recognised correctly).*

SNR in dB	∞	+21	+15	+9	+3	−3	−9	−15
Baseline	0	29	1	0	*	*	*	*
Klatt	0	1	3	2	7	1	0	*
Decomp	0	3	4	66	3	3	0	*

Table 2: *Insertions corresponding to results in table 1.*

Tables 3 and 4 give corresponding results for speech and machine gun noise. The results for decomposition here are preliminary, it is expected that further development of the machine gun model will improve performance. However, it can be seen that decomposition gives improved recognition performance, in particular significantly reducing the number of insertions.

SNR in dB	+21	+15	+9	+3	−3	−9	−15
Baseline	16	16	27	52	96	125	169
Klatt	22	24	33	30	45	70	94
Decomp	0	3	11	27	75	35	31

Table 3: *Words not correctly recognised out of 300 digits spoken in isolation for each algorithm at various signal-to-noise ratios with added machine gun. (* indicates that no words were recognised correctly).*

SNR in dB:	+21	+15	+9	+3	−3	−9	−15
Baseline	238	290	350	639	932	775	1316
Klatt	410	527	804	887	978	871	603
Decomp	6	14	31	60	214	327	449

Table 4: *Insertions corresponding to results in table 2.*

6 CONCLUSIONS

The results of the experiments showed that signal decomposition using hidden Markov modelling provides significant improvements in speech recognition performance for both stationary and highly non-stationary noise. In stationary pink noise good performance is obtained down to a signal-to-noise ratio of $-3dB$. The technique also successfully deals with impulsive background noise.

Finally it is important to note the artificial nature of these experiments, the speech and the noise were recorded independently and *added* together for these experiments. Many other effects must be taken into account for real speech in noise, e.g. the Lombard effect, microphone characteristics, etc. Also models of the noise were created *off-line* and were somewhat artificial in nature. Despite this, it is believed that decomposition is a very important technique in the speech recognition armoury having wide application in problems other than recognition of speech in noise, see [1].

References

[1] R.K. Moore, *"Signal Decomposition Using Markov Modelling Techniques."*, Royal Signals & Radar Establishment memo no. 3931, July 1986.

[2] A.P. Varga and K.M. Ponting, *"Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous Word Recognition."*, ESCA Proc. Eurospeech'89, Paris, Sept. 1989.

[3] D.H.Klatt, *"A digital filter bank for spectral matching."*, IEEE Proc. Int. Conf. Acoust. Speech & Signal Process., ICASSP'79, pp.573-576, 1976.

[4] A. Nádas, D. Nahamoo and M.A. Picheny, *" Speech Recognition Using Noise-Adaptive Prototypes."*, IEE Trans. Acoust. Speech & Sig. Proc.,vol ASSP-37, no.10, pp1495-1503 Oct. 1989.

[5] J.S.Bridle, M.D.Brown, and R.M.Chamberlain, *"An algorithm for connected word recognition."*, IEEE Proc. Int. Conf. Acoust. Speech & Signal Process., ICASSP'82, pp. 899-902, May 1982.

[6] *Specification of the filter bank analyser and design programs can be obtained from: The Head of the Speech Research Unit, Royal Signals and Radar Establishment, Malvern, Great Britain.*

[7] J.N.Holmes, *"The JSRU channel vocoder"*, IEE Proc. F, vol.127, no.1, pp. 53–60, Feb. 1980.

[8] R.S. Vonusa, J.T. Nelson, S.E. Smith, and J.G. Parker, *"NATO AC/243 (Panel 111 RSG-10) Language database"*, Proc. US National Bureau of Standards workshop on *"Standards for Speech I/O Technology"*, pp.223–228, 1982.

[9] H.J.M.Steeneken, F.W.M.Geurtsen, *"Description of the RSG10 noise database"*, TNO Institute for perception, report no IZF 1988-3, 1988.

[10] The International Telegraph and Telephone Consultative Committee, CCITT, *"Objective Measurement of Active Speech Level."* Suppl. no. 8, Red Book, vol. V, $VIII^{th}$ Plenary Assembly, pp242-247, Malaga, Oct. 1984.