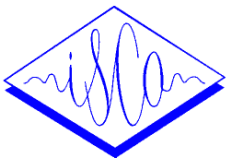

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

SIMULTANEOUS RECOGNITION OF CONCURRENT SPEECH SIGNALS USING HIDDEN MARKOV MODEL DECOMPOSITION

A.P. Varga and R.K. Moore

Speech Research Unit, Defence Research Agency, Electronics Division, R.S.R.E., Malvern, Great Britain.

ABSTRACT

This paper addresses the problem of automatic recognition of two simultaneous speech signals, that is the recognition of speech in the presence of speech from interfering talkers. The approach used is that of hidden Markov model based signal decomposition in which two or more concurrent signals are recognised simultaneously. In order to accommodate concurrent dynamically varying processes the decomposition technique uses conventional hidden Markov modelling of speech, but a generalisation of the conventional recognition algorithm. The performance of a decomposition based recogniser is compared with that of a conventional recogniser; the results demonstrate the ability of the new technique to decompose and recognise interfering signals as structurally complex as speech.

Keywords: Speech recognition, hidden Markov model based decomposition, recognition of simultaneous speech signals.

10.21437/Eurospeech.1991-168

1 INTRODUCTION

Recent work, [1] [2], described the technique of signal decomposition using hidden Markov models. This is a generalisation of conventional hidden Markov model based recognition that provides a basis for the optimal decomposition of simultaneous processes (that is, the recognition of interfering signals). The technique exploits the ability of hidden Markov models to model dynamically varying signals and extends/generalises the conventional approach to recognition in order to accommodate *concurrent* processes. This approach makes it possible to deal with structured and highly time varying concurrent signals, e.g. speech in the presence of non-stationary background noise. It does this by recognising simultaneously both the required signal (e.g. the speech) and the concurrent signal (e.g. interfering background noise)

There is, as yet, little experience in the use of decomposition, see only [2]. In order to examine the degree of interfering signal complexity that can be accommodate one of the most complex structured interfering signals, i.e. speech itself, has been investigated. This paper describes an experiment in which decomposition based recognition was used for simultaneously recognising concurrent speech signals.

2 SIGNAL DECOMPOSITION USING HIDDEN MARKOV MODELS

Signal decomposition using hidden Markov modelling, [1], is a general technique in which concurrent events are recognised simultaneously. This is achieved by using parallel or concurrent sets of hidden Markov models (HMMs), one set for each of the streams (or components) into which the signal is to be decomposed. Recognition is carried out using an extension and generalisation of the normal recognition process in which a multi–dimensional state space is searched in order to explain the observations; each dimension of this space corresponds to one of the sets of hidden Markov models and thus one of the streams into which the signal is to be decomposed. This can be compared with the normal recognition process in which a single set of HMMs, or single dimensional state space is searched in order to explain the observations.

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 can be 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:

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

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 described above. Evaluation of this observation probability will take the general form:

$$b1_i \otimes b2_j(O_t) = \oint_{C_t} P(O1_t, O2_t|i,j) \quad (3)$$

where the integration is along the contour C_t of fixed observation value, i.e. over all couples $(O1_t, O2_t)$ such that:

$$O_t = O1_t \otimes O2_t$$

Figure 1 shows this contour for various combination functions.

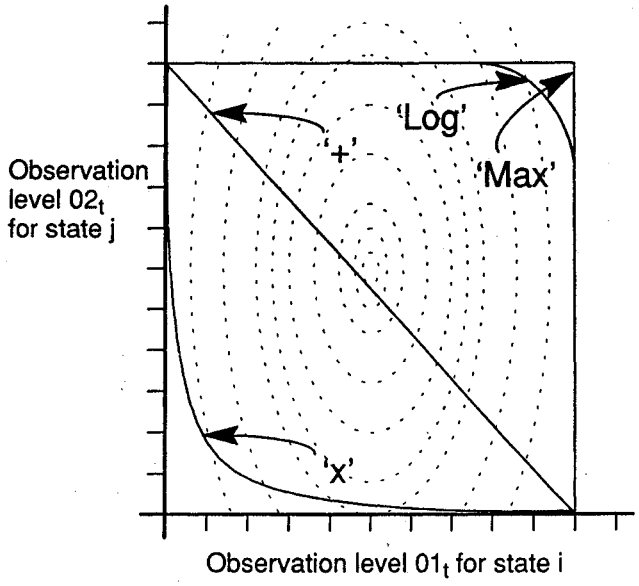

Figure 1: *Example contour plot of the joint output pdf for states i & j. Superimposed are various example integration contours C_t for; addition, multiplication, log and max combination operators.*

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 2. The extension to more than two components is straight forward.

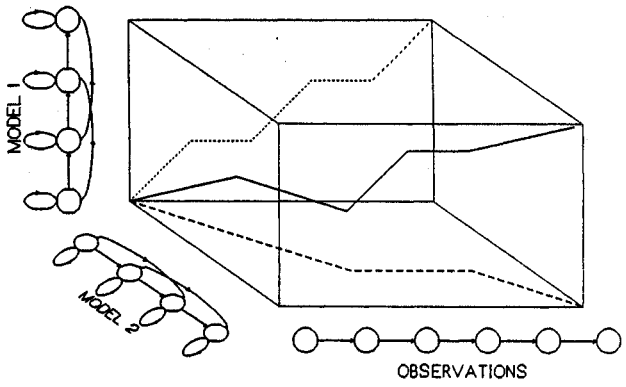

Figure 2: *Decomposition of 3-dimensional state-sequence into two 2-dimensional projections in the M1 and M2 state spaces.*

2.1 Decomposition for recognition of simultaneous speech signals using a filter bank front end

In the case of simultaneous speech from two talkers the components of the decomposition are two sets of speech models; in the work reported here two sets of whole word models were used.

The observation probabilities are evaluated on the basis of the output from a model from one set combined with the output from a model from the other set. 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 two speech signals added together at various level ratios. This is passed through a filter bank front end which generates log energy levels in each channel, thus:

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

where $O'1_t$ & $O'2_t$ are the energy levels of the two components.

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:

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

where $\mathcal{C}(O_t, \mu, \sigma^2)$ 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 a signal in a particular frequency band or channel has only a limited effect (due to passband overlap) across the rest of the spectrum.

3 EXPERIMENTS AND RESULTS

The objective of the experiments was to examine the performance of decomposition based recognition when recognising speech in the presence of a second contaminating speech signal.

3.1 Experimental data

Two different speech signals, spoken by different speakers, were combined together at various relative levels. The first speech signal consisted of isolated digits extracted from the NATO RSG-10 single digit database, [6]. These are in the form of several continuous tables each of 100 digits spoken in isolation. One table was used to train the models. The second signal was a series of 60 repetitions of the word *"monosyllabic"* spoken in isolation. This data was split in two, half used for training and half for testing.

The test data was generated by adding the two sets of speech signals together at three different signal level ratios (SLR): 0, -6 and $-12dB$ (the level of the *monosyllabic* data was successively reduced). The speech signal levels were measured using the British Telecom SV6 speech voltmeter (the SV6 conforms to the CCITT standard, [7], for speech level measurement). The 30 test repetitions of *"monosyllabic"* were used repetitively in order to generate an interfering signal of duration equal to that of the digit data. The data corresponding to one of the digit tables was used as an optimisation set. Optimisation was carried out by running decomposition recognition over the speech–on–speech data comparing the performance of various styles of modelling, e.g. varying the number of states etc. These experiments showed that word models with good resolution (i.e. at least as many states as phonemes) and no state skip transitions gave the best performance. Such models allow relatively little "smearing" across states and so provide detailed descriptions of the signals in each of the decomposition components. A detailed description, such as this, gives the least ambiguous decomposition interpretation of the observed data.

3.2 Experimental setup

The recogniser has a single microphone input with no extra sensors. The one-pass continuous speech recognition algorithm, [3], was used. The observation vectors were the log energy levels of a 27 channel filter bank analyser, [4]. The channels of the filter bank are roughly critical band spaced and overlapping, based on a successful channel vocoder design, [5]; the frequency range covered is $0 - 10$kHz. The channel energies were quantized with 8 bits in $0.5dB$ steps; the filter bank analysis was carried out at a rate of 100 frames per second.

The models used were speaker-dependent and the output distributions were multi-variate Normal with diagonal covariance matrix. The speech models were all trained under noise-free conditions. The first set of models (i.e. those corresponding to the source in the first stream) consisted of 10 whole word digits models; the second set consisted of a single whole word model of the word *monosyllabic*. Both sets also had two special models to account for the periods when there was no speech; a silence model and a tracking background noise model, details of these can be found in [2].

3.3 Results

Table 2 shows the performance of the decomposition based recogniser for the set of models corresponding to the first source or stream, i.e. the isolated digits. Also, for comparison table 1 shows the performance of a baseline recogniser and a recogniser which used Klatt's noise masking [2]. These latter recognisers were both continuous whole word recogniser using the 10 digit models and the tracking background model used in the decomposition recogniser. The baseline recogniser had no special features to cope with interfering noise, while the Klatt was implemented by extending the baseline recogniser to carry out noise masking as described in [2].

It can be seen that decomposition gave the best overall performance by a very large margin.

SLR in dB	Baseline		Klatt	
	% correct	% accuracy	% correct	% accuracy
0	32	-71	35.3	-155.4
-6	45	-22.7	49.3	-82.7
-12	63	-16.7	61	-56.7

Table 1: *Performance in terms of words correct and word accuracy for 300 digits spoken in isolation for the baseline and Klatt noise masking recognisers at various signal–level–ratios (SLR).*

SLR in dB	Decomosition	
	% correct	% accuracy
0	98.5	47.75
-6	99	78
-12	100	92.25

Table 2: *Performance in terms of words correct and word accuracy for 300 digits spoken in isolation for decomposition at various signal–level–ratios (SLR).*

The decomposition recogniser gave 100% word accuracy at all signal level ratios for the second decomposition stream, i.e. the monosyllabics.

4 CONCLUSIONS

The experiments reported here represent a first attempt at simultaneous recognition of concurrent speech signals, further work is required to develop the technique. It can be seen that the word accuracy performance is degraded by a high number of insertions when the signals are at similar levels. Also, the interfering signal used in this experiment is constrained; more general sets of models would be required for practical systems. However, the experiments demonstrate that signal decomposition using hidden Markov modelling can be an effective technique for the recognition of either simultaneous speech signals or speech in the presence of interfering speech. Further, it is believed that decomposition is a very important technique in the speech recognition armoury having wide application in problems other than those described to date.

References

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

[2] A.P. Varga and R.K. Moore, *"Hidden Markov Model Decomposition of Speech And Noise."*, IEEE Proc. Int. Conf. Acoust. Speech & Sig. Proc., ICASSP'90, Albuquerque, April. 1990.

[3] 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.

[4] *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.*

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

[6] 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.

[7] 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.