The State Space and "Ideal Input" Representations of Recurrent Networks

Tony Robinson
Cambridge University Engineering Department,
Trumpington Street, Cambridge, England.
ajr@eng.cam.ac.uk

Abstract

This paper looks at the data representations used in recurrent networks for two of the supplied sentences for the workshop. One sentence from the database on which the network was trained (timit) is used to illustrate the input, state and output representations for clean speech. Another sentence (clean) is used to illustrate the degradation that results from different recording conditions. Gradient descent in the input space is used on the second sentence so as to make the output better conform to the assumed pronunciation.

1 Introduction

This paper takes the opportunity to present different levels of representation in a recurrent network phone recogniser (Robinson and Fallside, 1991) on standard sentences. This recogniser has been shown to give good performance on a standard database (timit), and so one sentence from this database is analysed. A sentence recorded under different conditions (clean) is also analysed, and recognition is seen to be considerably worse. Analysis for the other two sentences (dirty and spont) has been carried out, but is not presented here as there is sufficient difference in quality between the two analysed sentences.

The first three sections detail the input, internal and output data representations of the network, along with the associated processing. In the next two sections the data representations are given for the sentence from the training database, including a plot of the state space variables. This is followed by two sections analysing the other sentence and computing the input representation which yields an output closer to the assumed transcription of the sentence.

2 The preprocessor

The preprocessor used in this system is fairly conventional. A hamming window of width 256 samples is applied the speech waveform every 16ms. From this window the following features are extracted:

- The log power.
- An estimate of the fundamental frequency from the position of the highest peak in the autocorrelation function.
- An estimate of the degree of voicing from the relative amplitude of the highest peak in the autocorrelation function.
- A power spectrum using a FFT which power normalised then grouped into 20 mel scale bins and cube rooted.

After the preprocessor, all channels are normalised and scaled to fit into a byte using a monotonically increasing function such that every value is eqi-probable. On presentation to the network, these values are expanded into a Gaussian distribution with zero mean and unit variance.

This data will be presented at the top of every diagram, in the order above reading from the bottom up. Thus the top twenty channels on the page form a spectrogram.

3 The recurrent net

Recurrent networks have been shown to give good estimates of the class conditional probabilities needed for Markov model based speech recognition systems (Morgan and Bourlard, 1990; Robinson, 1991). The network architecture used is shown below:

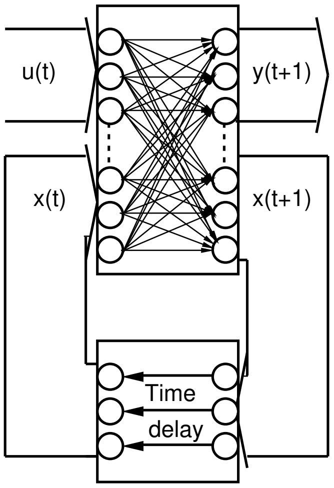

Figure 1: The Recurrent Error Propagation Network

Inputs from the preprocessor, $u(t)$, is presented to the network along with the state vector, $x(t)$. A single layer network computes the output, $y(t+1)$, and the next state vector, $x(t+1)$. The internal state units are considered to be independent estimators of unknown features.

4 Why use a state space?

The use of a recurrent network for phone probability estimation differs significantly from other approaches (e.g. the standard discrete, continuous or semi-continuous HMM estimators, or the use of non-recurrent networks) in the use of an internal state. Hence, conventional approaches are memoryless and contextual variation is explicitly modelled using mechanisms such as triphones or gender dependent networks. In contrast, the recurrent network employs an internal state trained by gradient descent for contextual modelling.

The acoustic realisation of a phoneme is known to depend on coarticulation effects from the immediate phonetic context. The triphone approach to modelling this variation assumes that a window including the preceding and following phoneme is sufficient to capture the contextual variation, and so builds a model for every triplet of phonemes which occur in the language. As well as requiring considerable storage for the models, good smoothing techniques across models are needed so that the parameters may be reliably estimated. Connectionist probability estimation normally takes a data-driven approach to this problem by effectively enlarging the window on the observations to encompass the relevant information. However, a simple finite window is both inefficient in terms of the increase in the number of free parameters with window size, and in terms of run time computation. This limits the window size that can be used although tied weights, such as employed in the time-delay neural network, can reduce these problems. However, a finite window size will always limit the temporal scope of the contextual information. Use of feedback within the network allows the accumulation of information over an arbitrary long period and avoids the duplication of processing associated with presenting multiple observations as a single input.

In experiments where the gender of the speaker was explicitly presented to the network, no change in recognition rate was observed. Similarly, training two networks for female and male speech, and testing on the appropriate set yielded a drop in performance, apparently because there was not enough data to make a good estimation of the parameters for the female model. In conventional systems, reasonable increases in performance are gained by having gender dependent models, the lack of similar behaviour with recurrent networks suggests that this information is being successfully held in the internal state.

5 A clean analysis: TIMIT

For the TIMIT speech, the recurrent net is used to produce the activations in fig 2. The display is in two parts: the top part shows the output of the preprocessor as the input to the network; the bottom part shows the output of the network as phone occurrence probabilities. Target outputs, as provided by the hand labellers, are shown by the shaded regions. It can be seen that the recurrent net closely models the task of the hand labeller. When the maximum likelihood symbol string is extracted with a Markov model, there are the following errors:

- Two substitution errors: The word "in" is labelled as `[ix ng]` and recognised as `[iy ng]`, and the word "year" is labelled as `[y ih axr]` and recognised as `[y ux axr]`.
- One insertion error: `[v]` is recognised between "wash" and "water".
- One deletion error: `[y]` in "had your" is lost.

Four errors in 41 symbols is an unusually low error rate for this classifier. On the full TIMIT test set the error rate is 30.3% (3.7% insertions, 20.5% substitutions, 6.1% deletions).

6 A look at the state space

The dimensionality of the state space for the recurrent net is 176. For display reasons, only the first 32 outputs are shown at the top of figure 5. Immediately obvious is a large degree of temporal correlation. The R.M.S. difference between successive state vectors is shown at the bottom of the figure, under the dashed line. If each element in the state vector were subjected to additive noise, then the maximum information would be transmitted if the state values were randomly distributed between 0 and 1. Under this assumption the expected R.M.S. difference is $\sqrt{1/6}$ or about 0.41. The observed R.M.S. difference is 0.07. However, in practice there is no explicit noise added to the state units, so the underlying correlations in the acoustic data and the limited processing capabilities of the (effective) single layer network give a correlated structure.

Looking at the temporal structure, it can be seen that while there are large changes in the network output, these are not on the whole reflected in all the state units. Thus the R.M.S. difference does not give phone boundary information. However, this is to be expected if the contextual information is fully utilised, as the time scale of information to pass through the network is of the order of the mean phone duration, so changes to the state vector will be blurred over the extent of the phone.

7 A dirty analysis: CLEAN

In contrast to the TIMIT sentence, the same network is used for recognition on the "clean" sentence. This sentence has been recorded through a low pass filter with a cut off of 2.8kHz. In this case a transcription is not available, so the pronunciations given in table 1 are assumed.

SENTENCE-END	h#
Fred	f r eh dcl
can	k ae n
go	gcl g ow
Susan	s uw z en
can+t	kcl k ae n tcl
go	g ow
and	ae n dcl d
Linda	l ih n dcl d ax
is	ih z
uncertain	ah n s er tcl t en

Table 1: Pronunciations used

These pronunciations were concatenated and a viterbi alignment performed with the Markov model. The results are shown in figure 3, again the target phones are shaded.

In this case it can be seen that the network output vastly disagrees with the assigned labels, in fact only 35% of the frames agree in labelling. The loss of the high frequency part of the spectra causes the occurrences of `[s]` to be recognised as `[f]`, as well as many other errors. However, it can be seen that the boundaries given by the viterbi algorithm are often aligned with sharp acoustic changes and with a change of recognised symbol.

8 The "ideal input" representation

The "ideal input" representation is that input which minimises the cost function of the network over the whole sentence. This can be achieved by gradient descent in the input space (Linden and Kindermann, 1989), and this method has been proposed for speech synthesis by recognition (Fallside, 1990). After three hundred iterations of gradient decent through the fixed weights no further reduction in the cost funtion occured, and the resulting input and output space is as shown in figure 4. The frame by frame error rate has been halved from 65% to 32%, although many segments are still incorrect, especially the vowels. Even though considerable change in the input space had occured, the final output is still worse than the TIMIT sentence, so it may be concluded that a better input exists than was found with gradient descent. Interestingly, the "ideal input" representation contains many non-speech features in that the normal constraints of smoothness in the time and frequency domains have been relaxed. For instance, in the first `[s]` there is one frame where a single channel has zero energy.

9 Conclusion

This paper has presented an analysis of the recurrent network on two sentences. Whilst good recognition results are possible on a sentence taken from the training set, this performance shows considerable degradation on sentences recorded under other conditions. This illustrates the need for an acoustic analysis robust to such variations. The "ideal input" has been calculated for one sentence, and found to have several non-speech features. Hence there is considerable scope for improvement of this recognition framework to increase the robustness to recording conditions and to make more accurate models of the speech dynamics.

Acknowledgements

The author wishes to acknowledge the ESPRIT BRA project Auditory / Connectionist Techniques for Speech for funding up to October 1991, and the UK Science and Engineering Research Council for subsequently supporting this work.

References

Fallside, F. (1990). SYNFREQ, speech synthesis from recognition. In *Proc. ESCA Conference on Speech Synthesis*, pages 237–240. Also CUED/F-INFENG/TR.54.

Linden, A. and Kindermann, J. (1989). Inversion of multilayer nets. In *International Joint Conference on Neural Networks*, volume II, pages 425–430, Washington.

Morgan, N. and Bourlard, H. (1990). Continuous speech recognition using multilayer perceptrons with hidden Markov models. In *Proc. ICASSP*, pages 413–416.

Robinson, T. (1991). Several improvements to a recurrent error propagation network phone recognition system. Technical Report CUED/F-INFENG/TR.82, Cambridge University Engineering Department.

Robinson, T. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5(3):259–274.

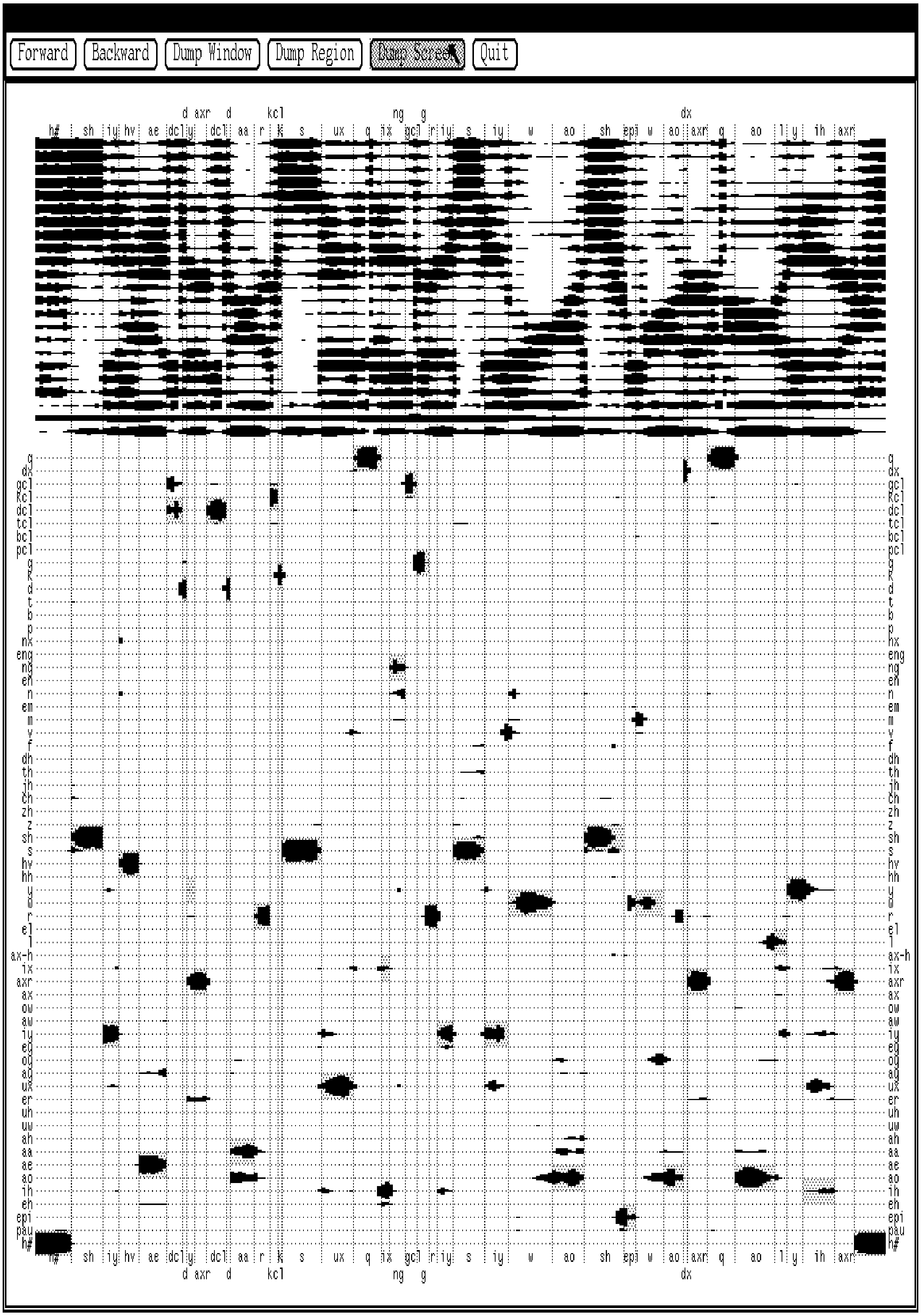

Figure 2: Actual network input and output for the TIMIT sentence

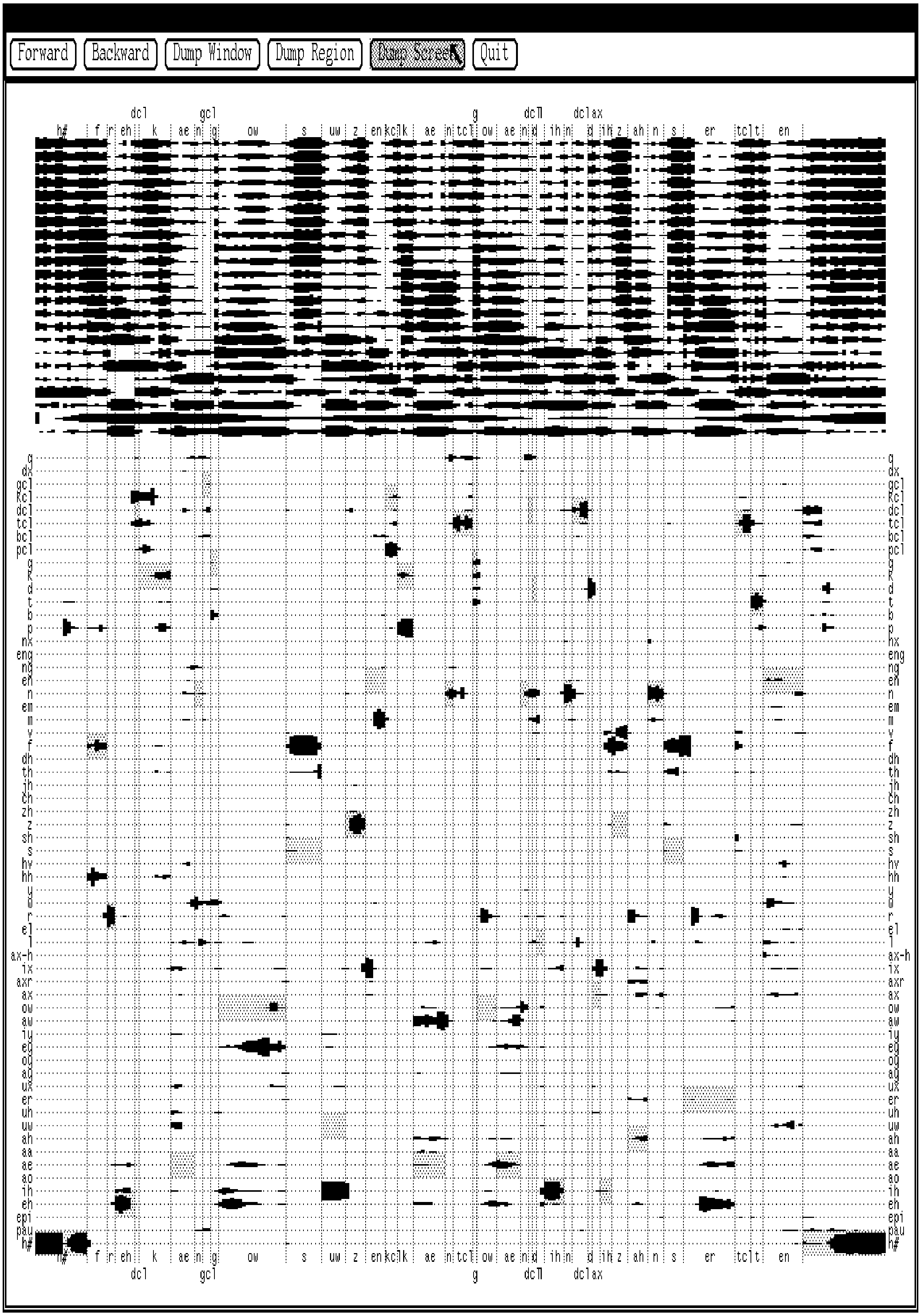

Figure 3: Actual network input and output for the "clean" sentence

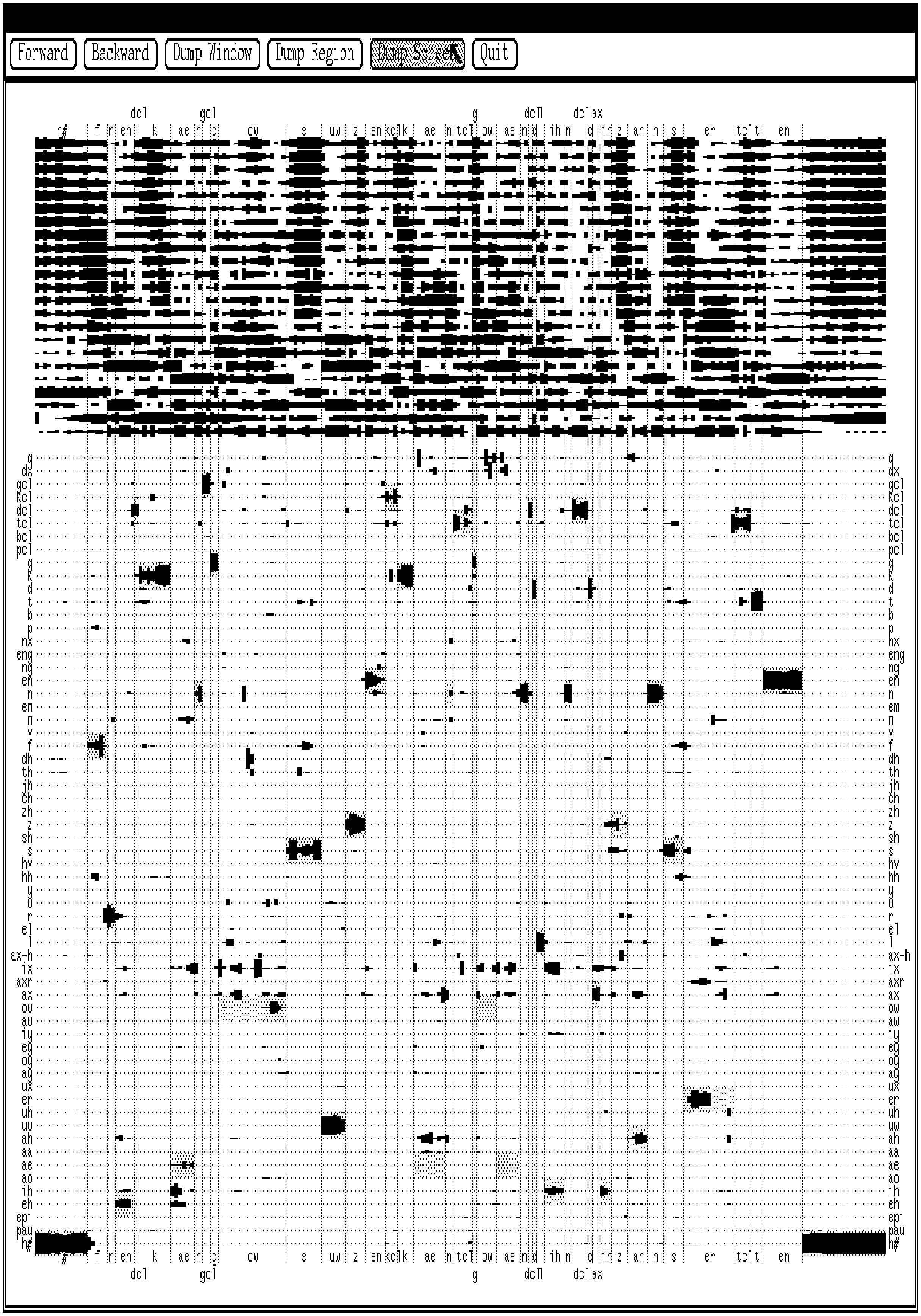

Figure 4: “Ideal” network input and output for the “clean” sentence

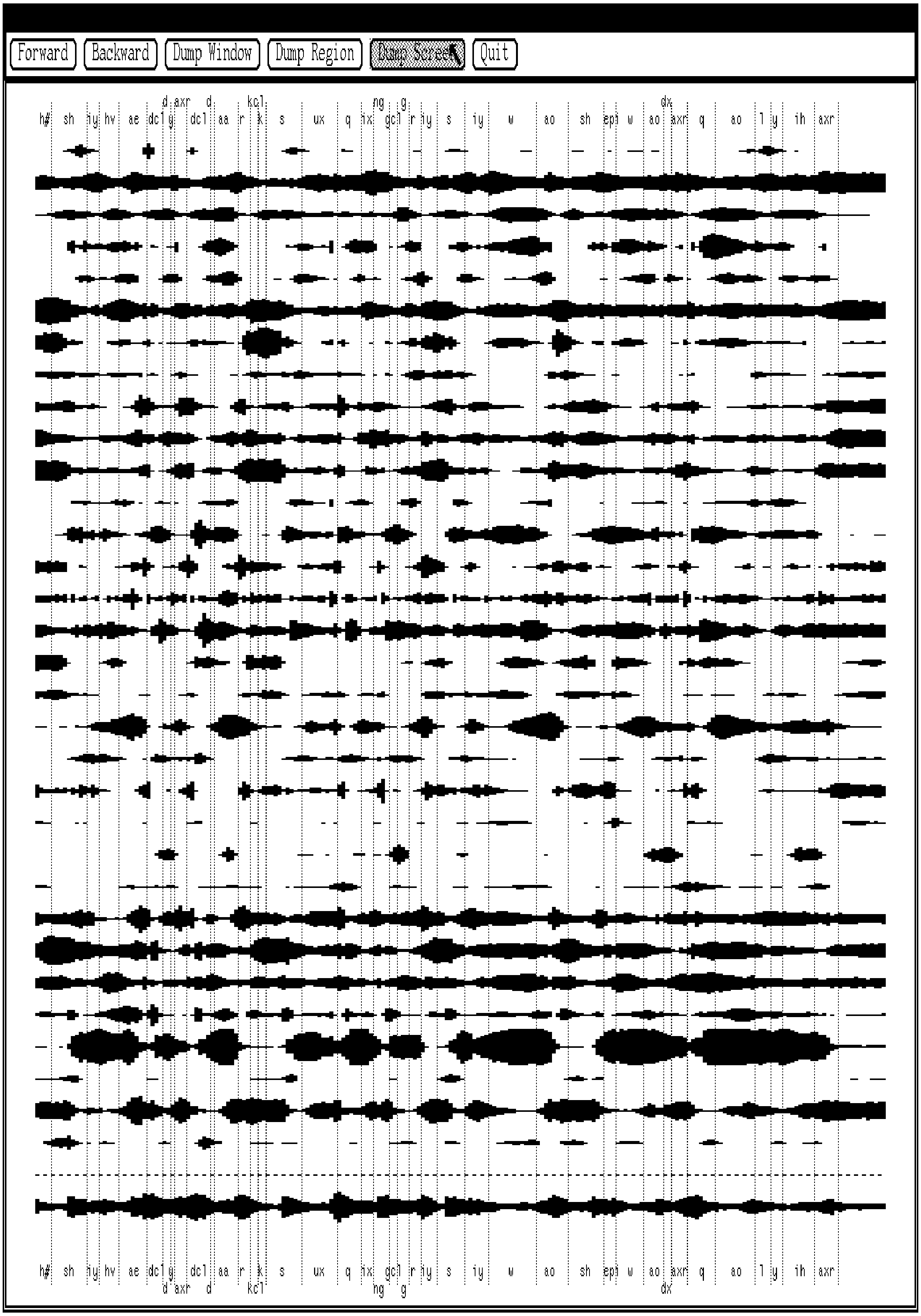

Figure 5: The first 32 state activations for the TIMIT sentence