Combining Imputation and Classification in a Single Recurrent Neural Network for Robust ASR with Missing Data

S. Parveen[1], P.D. Green[1] and M.R. Khan[2]
1 Speech and Hearing Research Group, Department of Computer Science,
University of Sheffield, Sheffield S14DP, UK
2 AMEC Training and Development Services, 111 Dunsmuir St, Vancouver, BC, V6B 5W3, CANADA
s.parveen@dcs.shef.ac.uk, p.green@dcs.shef.ac.uk, riaz.khan@amec.com

Abstract

Automatic Speech Recognition in the presence of additive background noise is a challenging task. The 'missing data' approach to this problem relies on identifying spectral-temporal regions which are dominated by the speech source. The remaining regions are considered to be 'missing' and generally dealt with either by being ignored or imputed using Hidden Markov Models. In contrast to missing data methods based on HMMs, connectionist approaches open up the possibility of making use of long-term time constraints and making the problems of classification with incomplete data and imputing missing values interact. This paper addresses the problem of combining robust ASR with missing data and pattern completion in a single Recurrent Neural Network. We report isolated digit recognition results on a realistic missing data case, in which the time-frequency regions which are missing are determined by local Signal-to-Noise Ratio estimates.

1. Introduction

Automatic Speech Recognition in the presence of additive background noise is a challenging task because of the mismatch between the acoustic models and incoming data caused by the noise [16]. Conventional techniques for improving recognition robustness (reviewed by Furui [12]) seek to eliminate or reduce the mismatch, for instance by enhancement of the noisy speech, by adapting statistical models for speech units to the noise condition or simply by training in different noise conditions. Success with these techniques has been moderate compared to human performance (see for instance the sessions on Noise Robust Recognition in Eurospeech 2001).

Missing data approaches have the potential to provide highly robust recognition for speech corrupted by high levels of additive noise and make minimal assumptions about the nature of the noise. They are based on identifying uncorrupted, reliable regions in the frequency domain and adapting recognition algorithms so that classification is based on these regions.

Initial processes, based on local signal-to-noise estimates [6], on auditory grouping cues [18], or a combination [4] define a binary 'missing data mask': ones in the mask indicate reliable (or 'present') features and zeros indicate unreliable (or 'missing') features. When test data is obtained by adding noise to clean speech, this a *priori* knowledge can be used to define an *oracle* mask (see Figure 1). Performance on the *oracle* mask is typically robust to noise levels of 0 dB SNR or worse [4]. This proof of concept sets an upper bound on the recognition performance.

Present missing data techniques developed at Sheffield [4, 6] and elsewhere [9, 21] adapt the prevailing technique for ASR based on Continuous Density Hidden Markov Models. In the *marginalisation* approach, missing values are ignored (by integrating over their possible ranges) and recognition is performed with the reduced data vector which is considered reliable. Data *Imputation* is a technique in which missing features are replaced by estimated values to allow the recognition process to proceed in normal way. There are several methods for imputation e.g. *unconditional imputation, conditional imputation* [6] and *feature compensation*

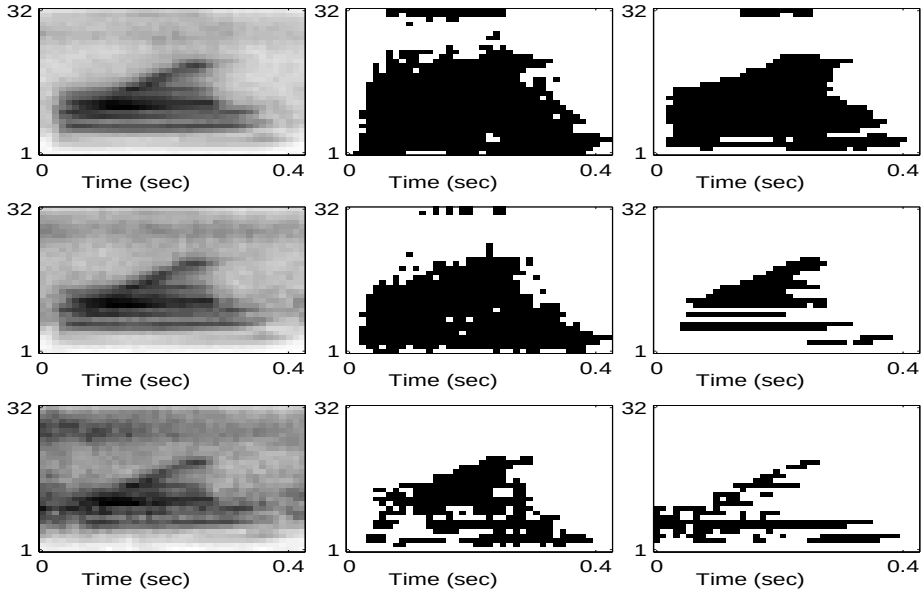


Figure 1: Auditory spectrograms (column 1), *oracle* reliable data mask (column 2) and SNR mask (column 3) for digit 'One' with subway noise at various SNRs (top to bottom: 20 dB, 10 dB, 0 dB)

[21]. In the latter, missing data is reconstructed using statistical information derived from the clean speech and neighbouring feature vector components.

For the multivariate mixture Gaussian distributions used in CDHMMs, marginalisation and conditional imputation can be formulated analytically [6]. For missing data ASR, improvements in both techniques follow from using the knowledge that for spectral energy features the true value of the speech in unreliable data is bounded between zero and the energy in the speech+noise mixture [23], [15]. These techniques, which limit the range of integration over unreliable feature values, are referred to as *bounded marginalisation* and *bounded imputation*. Missing data techniques coupled with a 'soft' reliable/unreliable decision produce good performance gains on a standard connected-digits-in-noise recognition task [4]. Marginalisation is a good approach if we are only concerned with recognition but imputation can be a suitable solution for cases where an existing recogniser is to be used unchanged [21]. Imputation can also be used as a way of enhancing the speech.

In this paper, we consider a connectionist alternative to imputation-based missing data techniques. One motivation is that CDHMMs are generative models which do not give direct estimates of posterior probabilities of the classes given the acoustics. Neural Networks, unlike HMMs, are discriminative models which do give direct estimates of posterior probabilities and have been used with success in hybrid ANN/HMM speech recognition systems [5].

In our previous exploration of this theme [19], classification and imputation was combined in a single RNN for robust ASR with missing data. Results were reported on an isolated digit recognition task for randomly deleted time/frequency regions in the speech spectrum. In this paper, we move on to a more realistic and demanding missing data case [7], in which the time-frequency regions which are 'missing' are determined by local Signal-to-Noise Ratio estimates. We also introduce additional output units performing imputation. This allows us to train our networks on clean speech with added noise, using the true values of corrupted features as training targets for the imputation units.

Use of actual targets for missing values has been reported by [22] but the RNN architecture in the latter work supports only pattern completion. We propose a combined RNN architecture using missing features both to recover the noise corrupted regions in the spectrogram and perform a word recognition task.

2. Recurrent neural networks for missing data robust ASR

2.1. Classification with Missing Data

Several neural net architectures have been proposed to

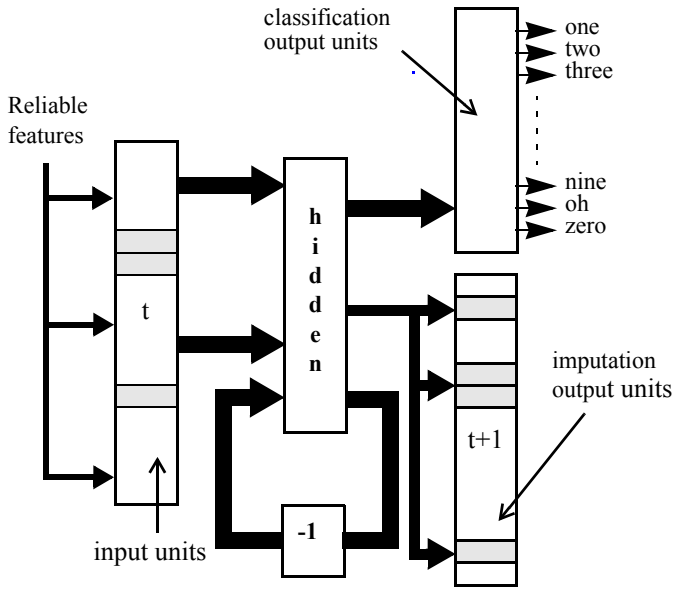


Figure 2: RNN architecture for robust ASR with missing data technique. Broad arrows show full forward and recurrent connections between two layers. Shaded blocks in the input and output layer indicate missing inputs at time *t* and the estimated values at time *t*+1 respectively, which change at every time step.

deal with the missing data case [2], [13], [14]. The problem is to compute the output of a unit when some of its input values are unavailable. For marginalisation, this involves finding a way of integrating over the range of the missing values. A robust ASR system to deal with missing data using marginalisation in radial basis function neural networks has recently been proposed by Morris et al. [17].

We use RNNs to estimate missing features in the input vector. RNNs have the potential to capture long-term contextual effects over time, and hence to use temporal context to compensate for missing data, which CDHMM based missing data techniques do not do: the estimated likelihood of the data at a particular time in a CDHMM system depends on the observed acoustics and the state distribution. Our architecture also allows a single net to perform both imputation and classification, with the potential of combining these processes to mutual benefit.

2.2. RNN architecture trained using true targets for the missing inputs

The RNN architecture we described in [19] was a modified version of that reported by Gingras & Bengio [14]. The assumption was that input features were missing at random, and the technique worked well both for classification and pattern completion with random deletions in the training and the test data. However, this performance was not maintained for realistic deletions based on estimates of local SNR. Performance was improved by preventing the network from imputing values outside the bounds (see section 1), but a more satisfactory solution is to allow the net to learn about bounds during training. We can do this by using the true values of corrupted features as targets during training.

We now make use of a revised RNN architecture, illustrated in Figure 2. Here, there are additional output units for imputation. The network is capable of doing one step ahead imputation efficiently without any help of bounds during training. It is basically an Elman RNN [11], where there are fully connected recurrent links from the past hidden layer to the present hidden layer.

The number of input units depends on the size of feature vector, i.e. the number of spectral channels (32 channels in the experiments reported). The number of hidden units is determined by experimentation (120 in our experiments). There are output units for each pattern class and extra units for pattern completion. In our case the classes are taken to be whole words, so in the isolated digit recognition experiments we report, there are eleven output units, for '1' - '9', 'zero' and 'oh', with additional 32 output units corresponding to the length of feature vector.

In training, missing inputs are initialised with their unconditional means. The RNN is then allowed to impute missing values for the next frame as the weighted sum of hidden activation.

$$X_{(m,t)} = f\left(\sum_{i=1}^{H} w_{jm} f(hid_{(j,t-1)}) \right)$$

Where $X_{(m,t)}$ is the missing feature at time t, $hid_{(j,t-1)}$ is the activation of hidden unit j at time t-1, $f(\dots)$ is the hyperbolic tangent activation function, w_{jm} indicates forward links from a hidden unit to the missing input.

After all the frames of a training example have been forwarded through the net, the error for both output classes and the 'missing' features is estimated as the sum squared error between the correct targets (one of n for the classification units and the clean values for the imputation units) and the RNN output for each frame. The error for the reliable or present features is set to zero. These errors are used to update RNN weights using back-propagation through time [24].

The recognition phase consists of a forward pass to produce RNN output for unseen data and imputation of missing features at each time step. The highest value in the averaged output vector is taken as the recognised class.

3. Isolated word recognition experiments

Continuous pattern classification experiments were performed using data from 55 male speakers in the isolated digits section of the AURORA database [20]. This database contains about 1200 isolated digits from 55 male speakers, where each speaker spoke 2 examples of the 11 word vocabulary (the digits 1-9, 'oh' and 'zero'). All speech data in the Aurora database is in turn obtained from the TIDigit database after downsampling to 8 KHz and filtering with a G712 characteristic.

1000 examples were chosen for training. Recognition was performed on the isolated digit examples from the male speakers in Aurora test set A. A validation set of 110 examples was used to control the stopping condition in training.

Acoustic vectors were obtained from a 32 channel auditory filter bank [8] with centre frequencies spaced linearly in ERB-rate from 50 to 3750 Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a first order filter with an 8 ms time constant, and sampled at a frame rate of 10 ms. Finally a cube root compression was applied to the frame of energy values.

In the experiments we report, the missing data masks in training were formed by deleting spectral energy features at random and the nets were trained on clean speech with deleted features initialised by unconditional mean. For training, 1/3rd of the training examples had 0% deletions, 1/3rd had 25% deletions and 1/3rd had 50% deletions.

For comparison purpose, we have also trained the RNN on *oracle* missing data masks (section 1). The *oracle* mask provided ideal conditions for training. Isolated digits with added noises (the Aurora noise types subway, car, babble and exhibition hall) were used for this part of training.

Recognition performance was evaluated on that part of the isolated digit section of Aurora test set A which has subway noise added. SNR and *oracle* masks were obtained for the noisy speech at SNRs from 20 dB to -5 dB at 5 dB intervals.

4. Results

We chose to train the RNN to impute missing inputs only with the RNN activation (described in the section 2.2). The baseline systems for the experiments in this work were:

- HMM:MARG - is a CDHMM system using marginalisation based missing data recognition. This system consisted of eleven whole word HMMs ('1' - '9', 'oh', 'zero'), each with 16 states and 2 mixtures per state, and was trained on clean isolated digits using HTK [25].

- RNN:NP & RNN:SS - Baseline RNN systems trained on clean speech for classification only; and tested on noisy speech before and after spectral subtraction respectively.

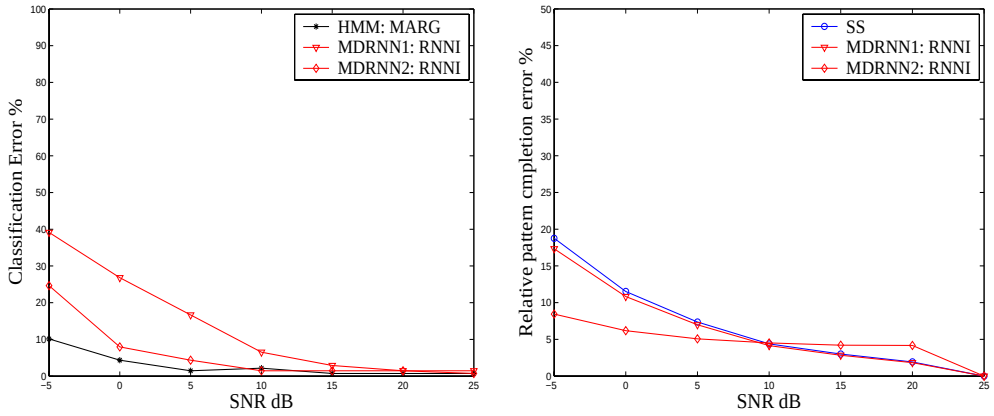
Two RNNs were trained with missing data in order to perform recognition with realistic missing data, i.e. generated by SNR mask criterion.

- MDRNN1:RNNI - This net used the clean speech and a random mask during training. Allowing the net to impute a step ahead randomly deleted features is an equivalent to the noise injection method [10] in which white noise is added to training examples in order to improve generalisation performance.

- MDRNN2:RNNI - The RNN in this approach was trained on ratemap features extracted from the isolated digit portion of multicondition Aurora database which had noisy examples with subway, babble, car and exhibition noises added with speech at SNRs 20 to 5 dB with a difference of 5 dB. *Oracle* masks were used during training to decide which channels have to be imputed. This scheme was chosen to set a matched training and test condition compared to 'MDRNN1:RNNI', where the training data was clean with random deletions and test data was noisy with realistic SNR based deletions.

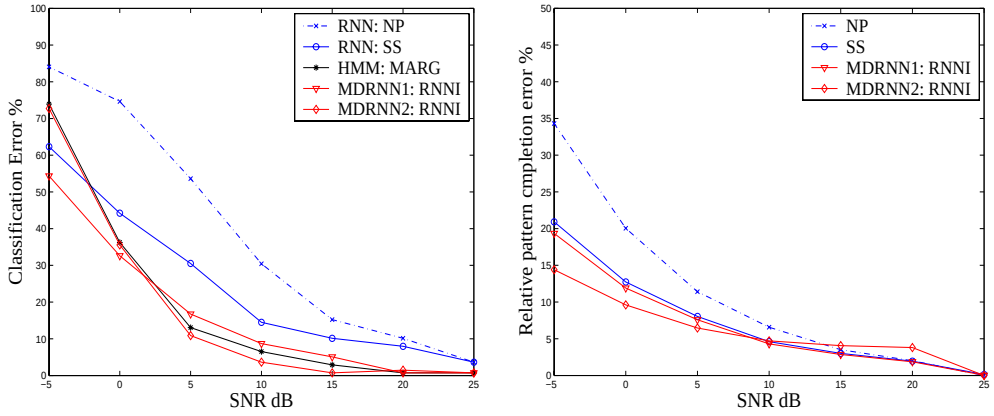
4.1. Classification Performance

Classification performance with *oracle* and SNR masks



(a) (b)

Figure 3: Performance with oracle masks: (a) classification, RNN compared to HMM, (b) imputation, RNN compared to spectral subtraction.



(a) (b)

Figure 4: Performance with SNR masks: (a) classification, RNN compared to HMM, (b) imputation, RNN compared to spectral subtraction.

is shown in Figure 3 (a) and 4 (a) respectively for 'MDRNN1:RNNI', 'MDRNN2:RNNI' and 'HMM:MARG'. As expected, results with *oracle* masks are superior to those for SNR masks in equivalent conditions. Training with *oracle* mask ('MDRNN2:RNNI') yields equivalent recognition performance to that of 'HMM:MARG' and is superior to both the 'MDRNN1:RNNI' and the standard RNN with spectral subtraction ('RNN:SS').

The average word error rates (WER) over all SNRs for missing data RNNs 'MDRNN1:RNNI' and 'MDRNN2:RNNI' were 16.99% and 17.96% respectively compared to 19.16% with marginalisation.

4.2. Pattern Completion Performance

In a similar way, Figure 3 (b) and 4 (b) show relative pattern completion error (calculated only for the features outside the mask) with oracle and SNR masks respectively for the missing data RNN and spectral subtraction. There is clear advantage for imputation by missing data RNN trained on noisy speech in equivalent conditions. It is also evident that training with the noisy speech to impute/predict clean values of the missing features gives the RNN a better chance to learn imputation.

5. Conclusion & future work

The RNN architecture described in this paper can deal with both realistic and random missing patterns during training and recognition. Training with random deletions may be more useful to deal with various types of unseen noises. The RNN also allows the imputed vectors to be used with standard ASR systems.

Our preliminary experiments suggest that combining

classification with the imputation of the same frame (supplied at the input at time *t*) rather than the frame at time *t*+1 results in better imputation performance. This confirms similar observations in other domains, [1, 3], that combining highly-correlated cues, in a neural network result in better generalisation of the desired task. This idea needs further investigations.

Another extension is to upgrade this recognition system for connected digit recognition with missing data, following the Aurora standard for robust ASR. This will provide a direct comparison with HMM-based missing data recognition [4]. In this case we will need to introduce 'silence' as an additional recognition class, and the training targets will be obtained by forced-alignment on clean speech with an existing recogniser.

6. Acknowledgement

This work was supported by Nokia Mobile Phones, Denmark and the UK Overseas Research Studentship scheme.

7. References

[1] Abu Mostafa. Y. (1995). HINTS. *Neural Computation, 7:639-671, July 1995*.

[2] Ahmed, S. & Tresp, V. (1993). Some solutions to the missing feature problem in vision. *Advances in Neural Information Processing Systems 5 (S.J.Hanson, J.D.Cowan & C.L.Giles, eds.), Morgan Kaufmann, San Mateo, CA, p.393-400*.

[3] Allen, J. & Christiansen, M.H. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the 18th Annual Cognitive Science Society Conference*.

[4] Barker, J., Cooke, M. and Green, P. (2001). Robust ASR based on clean speech models: An evacuation of missing data techniques for connected digit recognition in noise, *Eurospeech 2001, Aalborg, Denmark*.

[5] Bourlard, H. and N. Morgan (1998). Hybrid HMM/ ANN systems for speech recognition: Overview and new research directions. In *C. L.Giles and M. Gori (Eds.), Adaptive Processing of Sequences and Data Structures*.

[6] Cooke, M., Green, P., Josifovski, L. and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication, vol. 34, no. 3, p.267-285*.

[7] Cooke, M.P., Morris, A. & Green, P.D. (1996). Recognising occluded speech. *ESCA Tutorial & Workshop on the Auditory Basis of Speech Perception, Keele University, July 15-19*.

[8] Cooke, M.P. (1991). Modelling auditory processing and organisation. *PhD thesis, Department of Computer Science, University of Sheffield*.

[9] Drygajlo, A. & El-Maliki, M. (1998). Speaker verification in noisy environment with combined spectral subtraction and missing data theory. *Proc ICASSP-98, vol. I, p. 121-124*.

[10] Dupont, S. & Ris, C. (1999). Assessing local noise level estimation methods. *in Robust Methods for Speech Recognition in Adverse Conditions, Tempere, 1999*

[11] Elman, J.L. (1990). Finding structure in time. *Cognitive Science, vol. 14, p. 179-211*.

[12] Furui, S. (1997). Recent advances in robust speech recognition. *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, France, p. 11-20*.

[13] Ghahramani, Z. & Jordan, M.I. (1994). Supervised learning from incomplete data via an EM approach. *Advances in Neural Information Processing Systems 6 (J.D. Cowan, G. Tesauro & J. Alspector, eds.), Morgan Kaufmann, San Mateo, CA, p. 120-129*.

[14] Gingras, F. and Bengio, Y. (1998). Handling Asynchronous or Missing Data with Recurrent Networks. *International Journal of Computational Intelligence and Organizations, vol. 1, no. 3, p. 154-163*.

[15] Josifovski, L., Cooke, M., Green, P. and Vizinho, A. (1999). State based imputation of missing data for robust speech recognition and speech enhancement. *Proc. Eurospeech'99, Budapest, vol. 6, p. 2837-2840*.

[16] Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication, vol. 22, no. 1, p. 1-15*.

[17] Morris, A., Josifovski, L., Bourlard, H., Cooke, M.P. and Green, P.D. (2000). A neural network for classification with incomplete data: application to robust ASR. *ICSLP 2000, Beijing, China*.

[18] Palomäki K. J., Brown G. J. and Barker J. (2002). Missing data speech recognition in reverberant conditions. *ICASSP 2002, Orlando, Florida, USA, May 13-17, 2002*.

[19] Parveen, S. and Green, P. (2001). Speech Recognition with Missing data techniques using Recurrent Neural Networks. *Advances in Neural Information Processing Systems 14, (T.G.Dietterich, S. Becker and Z. Ghahramani eds.), MIT Press*.

[20] Pearce, D. and Hirsch, H.G. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP 2000, vol. IV, p. 29-32, Beijing, China*.

[21] Raj, B., Seltzer, M., & Stern, R. (2000). Reconstruction of damaged spectrographic features for robust speech recognition. *Proc. ICSLP 2000, Beijing, China*.

[22] Seung, H. S. (1997). Learning continuous attractors in Recurrent Networks. *Advances in Neural Information Processing Systems 10, (Michael I. Jordan and Sara A. Solla eds.), MIT press, p. 654-660*.

[23] Vizinho, A., Green, P., Cooke, M. and Josifovski, L. (1999). Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study. *Proc. Eurospeech'99, Budapest, Sep. 1999, vol. 5, p. 2407-2410*.

[24] Werbos. P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE, vol. 78, no. 10, p. 1550-1560*.

[25] Young, S. J. and Woodland P.C. (1993). HTK version 1.5: User, reference and programmer manual, *Cambridge University Engineering Department, Speech Group, 1993*.