SOFT DECISIONS IN MISSING DATA TECHNIQUES FOR ROBUST AUTOMATIC SPEECH RECOGNITION.

Jon Barker, Ljubomir Josifovski, Martin Cooke and Phil Green

Department of Computer Science, University of Sheffield
Sheffield S1 4DP, UK
j.barker, l.josifovski, m.cooke, p.green@dcs.shef.ac.uk

ABSTRACT

In previous work we have developed the theory and demonstrated the promise of the Missing Data approach to robust Automatic Speech Recognition. This technique is based on *hard* decisions as to whether each time-frequency "pixel" is either reliable or unreliable. In this paper we replace these discrete decisions with *soft* estimates of the probability that each "pixel" is reliable. We adapt the probability calculation to use these estimates as weighting factors for the complementary reliable/unreliable interpretations for each feature vector component. Experiments using the TIDigits connected digit recognition task demonstrate that this technique affords significant performance improvements at low SNRs.

1. INTRODUCTION

In previous work [2, 5, 6] we have developed the theory and demonstrated the promise of the Missing Data approach to robust Automatic Speech Recognition. In this technique, spectral-temporal regions uncontaminated by noise are identified and CDHMM recognition methods are adapted to make use of this partial information. For a given acoustic energy vector, we compute the joint probability that (1) the reliable components would be generated from their marginal distribution and (2) that the true values of the unreliable components are between zero and the observed value in the speech-and-noise mixture.

The missing data approach works well as long as the regions uncontaminated by noise can be reliably identified. Indeed, if we cheat and use knowledge of the clean signal to identify these regions precisely, then we can achieve ASR with near human performance. Unfortunately locating the reliable regions is itself a difficult problem and mistakes made at this stage are passed onto the speech decoder and result in suboptimal recognition performance.

The conventional missing data approach has the drawback of forcing a hard decision about whether a spectro-temporal region is speech or background at an early stage of processing (i.e. in ignorance of the speech models). In the technique described in this paper this decision is softened – we replace the discrete decision that a time-frequency pixel is reliable or unreliable with an estimate of the probability that the data is reliable. We then adapt the probability calculation to use this estimate as weighting factors for term (1) and term (2) for each vector component.

This work was supported by a Motorola Partnerships Research Grant, the EU SPHEAR TMR network and EU LTR project RESPITE.

Section 2 of this paper reviews and reformulates a solution to the problem of classification with incomplete data based on averaging the state emission p.d.f. This section also includes description of some minor incremental improvements to our existing MDASR system; the use of temporal derivatives and the inclusion of word insertion penalties. Section 3 describes how the missing data approach is adapted to employ soft decisions (i.e. fuzzy missing data masks). The results of comparative experiments carried out on the TIDigits corpus [7] with added NOISEX noise [9] with missing data with both hard and soft decisions are presented in section 4.

2. ROBUST ASR WITH MISSING DATA

Assuming that the input data vector x has been partitioned into a reliable part x_r and an unreliable part x_u in the previous stage of processing, two approaches to the problem of handling the partial feature vectors in the context of CDHMM recogniser have been developed so far: *bounded marginalisation* (BMG) and *bounded state-based data imputation* (BSDI). In this paper we concentrate on the former of these two approaches.

2.1. Bounded marginalisation

In BMG the emission probability (likelihood of observing the data x when we are in state S) $f(x|S)$[1] is averaged with respect to the data p.d.f. $p(x)$:

$$\overline{f(x|S)} = \int f(x'|S)p(x')dx' \quad (1)$$

In a CDHMM system $f(x|S)$ is a mixture of M multivariate Gaussians with diagonal covariance matrices:

$$f(x|S) = \sum_{k=1}^{M} P(k|S)f(x_r|k,S)f(x_u|k,S), \quad (2)$$

where $P(k|S)$ are the mixing coefficients. For the average likelihood we have:

$$\overline{f(x|S)} = \sum_{k=1}^{M} P(k|S) \int f(x'_r|k,S)f(x'_u|k,S)p(x'_r,x'_u)dx'_r dx'_u. \quad (3)$$

Knowing the reliable features means that $p(x'_r) = \delta(x'_r - x_r)$, so $p(x'_r,x'_u) = \delta(x'_r - x_r)p(x'_u|x'_r)$. Assuming an "additive noise" model and filterbank energy features, the unreliable features are bounded below by 0 and above by

[1] In this paper, $f(\cdot)$ denotes likelihood, $p(\cdot)$ denotes probability density and $P(\cdot)$ denotes probability.

the value of the feature in the noisy speech mixture x_u. Without additional knowledge, we have to assume that they are distributed uniformly in $[0, x_u]$:

$$\overline{f(x|S)} = \sum_{k=1}^{M} P(k|S) f(x_r|k,S) \frac{1}{x_u} \int_0^{x_u} f(x'_u|k,S) dx'_u. \quad (4)$$

In the case of Gaussian distributions, the integral can be evaluated using the standard error function. It models the "counterevidence" [3] against a particular state. For example, for a low x_u, the quieter states will score better then more energetic ones.

If there exists no knowledge about the unreliable features, then $p(x'_r, x'_u) = \delta(x'_r - x_r)$ and (3) simplifies to the marginal distribution:

$$\overline{f(x|S)} = \sum_{k=1}^{M} P(k|S) f(x_r|k,S). \quad (5)$$

The marginal distribution $f(x_r|k,S)$ is diagonal Gaussian itself and is readily computable. The computational cost of (4) is comparable to the full feature vector likelihood calculation, while the cost of evaluating (5) is lower (proportionally to the number of missing features).

2.2. Using temporal constraints in MDASR

The work reported here introduces the use of "velocities" to the MDASR approach.

In traditional ASR it is common place to add "velocities" – approximations to temporal derivatives – to the acoustic vector [4]. If the features are present, their temporal derivatives can be estimated using the standard expression:

$$\Delta x(t) = \frac{\sum_{i=-N}^{N} i \cdot x(t+i)}{\sum_{i=-N}^{N} i^2} \quad (6)$$

The problem with MDASR is that some of the $x(t+i)$ for $i = -N \ldots N$ features may be unreliable and thus missing. One solution to the problem is to treat the derivative $\Delta x(t)$ as missing if any of the features $x(t+i)$, $i = -N \ldots N$ needed to compute $\Delta x(t)$ are missing (the *strict mask*), as depicted in Figure 1. If the missing mask pattern was random, this would create a very sparse mask for the derivatives. However, in the experiments with speech and noise this is not the case. The reliable features tend to be clustered into time-frequency blocks, so the sparsity of the derivative mask is not much greater than that of the feature mask.

2.3. Word insertion penalties

Missing data systems employed in previous work have been noted to be prone to word insertion errors. These insertions are believed to be due to the difficulties of balancing the HMM 'match' and 'transition' scores in the missing data framework. The word insertion problem can be ameliorated by introducing a word boundary penalty during the Viterbi token passing. This penalty is implemented by simply deducting a fixed cost from the score of a token as it passes out of a word-final state. It is possible to set the size of the penalty so that insertions are reduced without generating a corresponding increase in the number of deletions.

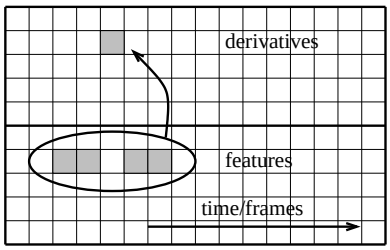

Figure 1: Computing the strict mask for the derivatives

3. MDASR WITH SOFT DECISIONS

In previous work the missing data approach has used a discrete missing data mask. The mask consists of 0's and 1's, with 0 meaning the data at the indicated spectro-temporal point is missing (i.e. masked by the background) and with 1 indicating the point is dominated by the speech source. Due to the large dynamic range of speech signals and the compressive representations that this dynamic range demands, this is a reasonable segmentation. Most regions are dominated by one source, and in regions where one source dominates the noisy representation is very close to the clean representation of that source. However, in practise, the MDASR system has an imperfect estimate of the noise source. If we accept error in our noise estimate, we accept error into the judgement of whether it is the speech or the noise that dominates a particular spectro-temporal regions. These errors are made concrete and irreversible when using a discrete mask.

In this work we soften the missing data mask. Rather than use either 0 or 1, we use a continuous value in the range [0.0, 1.0] which is interpreted in the missing data probability calculation as "the probability that the point is dominated by the speech signal". So, 0 and 1 still have the same meaning as in the discrete version, but we may now also have intermediate values indicating our degree of confidence whether or not the point is masked.

How do we generate the fuzzy mask?

If we were to assume the error in our noise estimate has a Gaussian distribution (as measured in the units of the representation we are using, e.g. log energy, cube root energy) then we can work out the ideal fuzzy mask based on the variance, σ, of this distribution. If we are using a compressive representation for which the approximation, $signal+noise = max(signal, noise)$ is reasonable, then the ideal mask can be approximated by the difference between the estimated local noise and the estimated local signal, which we shall call x, compressed by the function:

$$f_\sigma(x) = \int_{-\inf}^{x} p_{0,\sigma}(x) dx. \quad (7)$$

where $p_{0,\sigma}(x)$ is the p.d.f. of the Gaussian distribution with 0 mean and variance σ. In general, the distribution of the noise estimation error will be channel dependent, and a

channel dependent compression function should be used, i.e. $f_{\sigma_i}(x)$.

In practice the noise estimation error is only likely to be Gaussian if we have a good model of the noise. The missing data approach however attempts to avoid employing noise dependent models. In the current work we employ a simple stationary noise estimate for all noise types. For non-stationary noises the error in the estimate is likely to have a non-Gaussian distribution. Accepting this, we have not attempted to compute ideal fuzzy masks, but have instead generated a mask of values between 0 and 1 by compressing x with a simple sigmoid function (as is illustrated in figure 2) with empirically derived parameters. The mapping is of the form:

$$f(x) = \frac{1}{1 + \exp(-\alpha(x - \beta))} \tag{8}$$

where α is the sigmoid slope, and β is the sigmoid center. Appropriate values for these parameters are found via a series of tuning experiments. The valid range for α is $[0, \text{inf})$. For large values of α the sigmoid becomes steep and the resultant fuzzy mask approximates a traditional discrete missing data mask. In this case we are implicitly assuming a small variance in the noise estimation error. At the other extreme as the value of α tends to 0, we approach a mask where all values are 0.5. If $\alpha = 0$, we are assuming no knowledge of the noise and admitting maximum uncertainty into the mask.

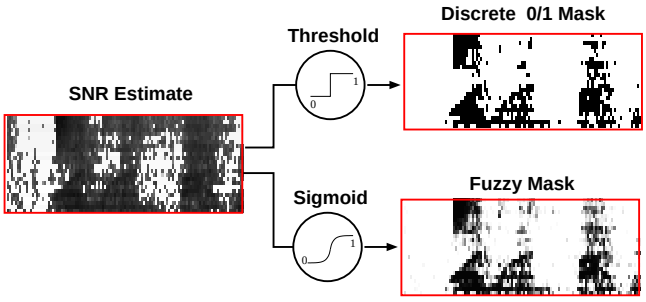

Figure 2: Illustration of the difference between discrete and fuzzy missing data masks.

Our use of fuzzy masks has parallels with the work of Renevey and Drygajlo in which a fuzzy mask is used in conjunction with missing data imputation[8].

4. EXPERIMENTS WITH FUZZY MASKS

The TIDigits corpus of digit sequences was used. Acoustic vectors were obtained via a 32 channel auditory filter bank (Cooke, 1993) with centre frequencies spaced linearly in ERB-rate from 50 to 8000 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 (this is the same representation as employed in [2] except here 32 channels are being used rather than 64). Finally, a cube root compression was applied to the frame of energy values. HTK [10] was used for training, and an in-house C++ decoder for recognition. Twelve models ('1'-'9', 'oh', 'zero' and 'silence') consisting of 8 no-skip, straight-through states with observations modeled with a 10 component diagonal Gaussian mixture were trained on clean speech. An additional 1-state silence model was used to model the brief inter-digit pauses that may occur during long digit strings. Factory noise and Lynx helicopter noise from NOISEX were added (with random start points) at SNRs from +20dB to 0dB to a subset of the TIDigits test set consisting of 240 digit strings. For bounded marginalisation, the lower bound was set to 0 and the upper bound to the value of the noisy speech mixture at each time-frequency point.

Two types of masks were used for the filter bank features: discrete (i.e. 0/1) missing data masks, and continuous fuzzy missing data masks. Both these are based on estimates of the local difference between cube root signal and cube root noise energy, $\sqrt[3]{\hat{s}} - \sqrt[3]{\hat{n}}$, which we will denote X.[2] An average of the first 10 frames in each utterance is used as a simple noise estimate, $\hat{n}$. The signal estimate, $\hat{s}$, is then obtained by assuming the noise is additive in the energy domain.

The discrete mask is obtained by estimating X at each time-frequency point, and treating as reliable those points where $X > \beta$. The fuzzy mask is obtained by passing X through the sigmoid function given in equation 8.

For the derivatives, the same discrete "strict mask" (as explained in Section 2.2) was used for both the discrete and the fuzzy cases. No knowledge about the bounds on the derivatives was used (5).

A preliminary series of tuning experiments were run to find appropriate values for the parameters α and β of the sigmoid function employed in the fuzzy mask technique. Informal tests were conducted to find a sensible range of values, and then tests with a set of 240 utterances were run over a grid of α and β values in these ranges. These tuning experiments were run using both the Lynx helicopter noise and the factory noise corrupted data, and at a range of SNRs. It was found that the optimal α and β values were largely independent of SNR but were dependent on noise type – as discussed below.

Finally, a third (discrete) mask based on a priori knowledge (APR) was used to establish the limits of the MDASR techniques. This mask assumes correct knowledge about the positions of the reliable features in the time-frequency plane, obtained by comparing the noisy and the clean speech and treating points as reliable if the difference is less than a tuned acceptance threshold.

5. RESULTS AND DISCUSSION

Figures 3 and 4 show results for both factory and Lynx helicopter noise at the range of SNRs tested. Consider first the factory noise results, Figure 3. The lower line ("FBank32") is the baseline for the discrete mask without delta features or word penalties. Adding delta features ("+delta") produces a consistent enhancement of around 3% or 4%. Adding word penalties leads to a similar improvement. Moving from a discrete mask to a fuzzy mask ("Fuzzy") leads to further performance improvement with the largest gains made at the lower SNRs (the 0 dB result increases from 46% to 60%). The fuzzy mask results here use a sigmoid that has been tuned using the factory noise corrupted data.

Figure 4 shows a similar set of results for the Lynx helicopter noise. Here we see a roughly similar pattern of improvements though starting from a higher baseline (the

[2]In previous work in which log compressed representations have been employed, thresholded local SNR has formed the basis of the discrete missing data mask.

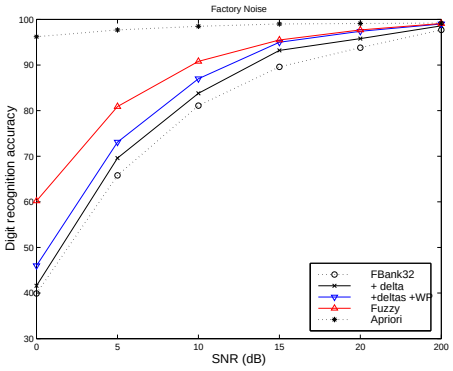

Figure 3: Results for factory noise.

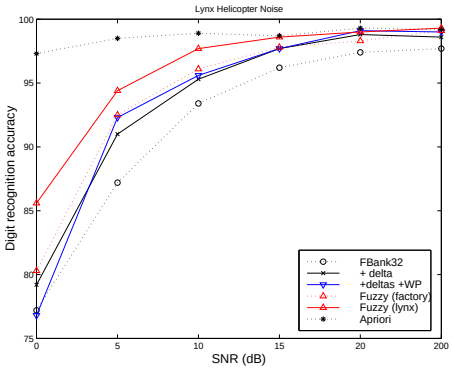

Figure 4: Results for Lynx helicopter noise.

Lynx helicopter noise is more stationary and hence the baseline missing data system performs better). This figure also compares results for the fuzzy mask in which the sigmoid has either uses the same parameter as tuned for the factory noise data, or parameters tuned specifically for the Lynx noise. It can be seen that although the fuzzy mask performs no worse than the discrete mask when tuned on mis-matched data, it only demonstrates a clear performance improvement when tuned and tested on matched data.

The Lynx helicopter noise is more stationary than factory noise and our noise estimates will be more reliable. This means we expect less uncertainty in the mask (i.e. less fuzziness) and the appropriate sigmoid will have a steeper slope. Although different noise types have different optimal sigmoid parameters, we can compromise using parameters that do well for the highly non-stationary factory noise, and do no worse that discrete masks for the more stationary Lynx noise. In future work we hope to have develop adaptive sigmoid parameters based on estimates of the variance in the noise estimation error.

Also shown in the figures are the results obtained using the *a priori* mask. These results demonstrate near human levels of performance and far exceed the best fuzzy mask results. Although fuzzy masks ameliorate the problems of discrete masks and poor noise estimates, they are not the full solution to the robust recognition problem. Fuzzy masks soften the mistakes but ideally we would like the mask construction to be informed by top-down speech knowledge from the speech models themselves. We are investigating this idea in parallel work in which we exploit coherent fragments of the data - i.e. spectro temporal regions which we are confident belong to the same source. The task of finding the subset of these fragments that describes the speech source is then integrated with the task of decoding the speech [1].

6. CONCLUSIONS

We have presented an approach for replacing the hard present/missing decisions employed in previous missing data with a soft probabilistic mask estimated using a sigmoid function. This approach has been shown to afford a significant performance improvement when tested on a connected digit recognition task with considerable gains at low SNR. Best results are obtained with a noise-specific tuning of the sigmoid parameters. In future work we hope to use estimates of the noise estimation error variance to find suitable sigmoid parameters directly.

7. REFERENCES

[1] J.P. Barker, M.P. Cooke, and D.P.W. Ellis. Decoding speech in the presence of other sound sources. In *Proc. ICSLP '00*, Beijing, China, October 2000. to appear.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, jun 1999. Accepted for publication.

[3] S. Cunningham and M. Cooke. The role of evidence and counter–evidence in speech perception. In *ICPhS'99*, pages 215–218, 1999.

[4] S. Furui. Speaker–independent isolated word recognition using dynamic features of the speech spectrum. *IEEE Transactions on acoustics, speech, and signal processing*, ASSP–34(1):52–59, feb 1986.

[5] P. D. Green, M. P. Cooke, and M. D. Crawford. Auditory scene analysis and Hidden Markov Model recognition of speech in noise. In *ICASSP'95*, pages 401–404, 1995.

[6] L. Josifovski, M. Cooke, P. Green, and A. Vizinho. State based imputation of missing data for robust speech recognition and speech enhancement. In *Eurospeech'99*, volume 6, pages 2837–2840, sep 1999.

[7] R.G. Leonard. A database for speaker-independent digit recognition. In *Proc. ICASSP '84*, pages 111–114, 1984.

[8] P. Renevey and A. Drygajlo. Introduction of a reliability measure in missing data approach for robust speech recogntion. In *Proc. EUSPICO'2000*, 2000.

[9] A.P. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, Speech Research Unit, Defence Research Agency, Malvern, U.K., 1992.

[10] S. J. Young and P. C. Woodland. *HTK Version 1.5: User, reference and programmer manual.* CUED, Speech Group, 1993.