MISSING DATA THEORY, SPECTRAL SUBTRACTION AND SIGNAL-TO-NOISE ESTIMATION FOR ROBUST ASR: AN INTEGRATED STUDY

A. Vizinho, P. Green, M. Cooke and L. Josifovski

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello St, Sheffield S1 4DP,UK
{A.Vizinho,P.Green,M.Cooke, L.Josifovski }@dcs.shef.ac.uk

ABSTRACT

In the missing data approach to robust Automatic Speech Recognition (ASR), time-frequency regions which carry reliable speech information are identified. Recognition is then based on these regions alone. In this paper, we address the problem of identifying reliable regions and propose two criteria to solve this based on *negative energy* ($\hat{s} < 0$) and *SNR* ($\hat{s}^2 < \frac{1}{2}|s+n|^2$). These criteria are evaluated on the TIDigits corpus for several noise sources and compared with spectral subtraction. We show that in this task the missing data method performs considerably better than spectral subtraction and the combination of the two techniques outperforms either technique used alone. We report robust performance at 0dB SNR for car noise and 10dB SNR for factory noise.

1. INTRODUCTION

In the missing data approach to robust ASR, two problems arise: the identification of the non-reliable time-frequency regions of the speech and recognition techniques to deal with the incomplete data set. We concentrate here on the first problem, with the aim of finding an automatic procedure for masking unreliable regions. A more detailed treatment of both problems and literature review is available in Cooke et al [2].

Previous work by Drygajlo and El-Maliki [3] combined missing data with spectral subtraction by treating the negative components after subtraction as missing. This is termed the *negative energy criterion* below. This criterion alone results in a modest improvement in recognition performance [2] but essentially corrects errors made by spectral subtraction rather than addressing the challenge of reliable data selection. We develop here another masking procedure which uses an additional SNR criterion.

Section 2 summarises the missing data recognition technique we used for this work. The masking criteria are related to noise estimation: several such techniques are described in section 3. Section 4 formalises the masking criteria employed for the experiments. Recognition results using TIDigits in additive noise from the Noisex database are presented in section 5.

2. MISSING DATA THEORY

Two approaches for classification with missing components exist: data imputation and marginalisation. The theory of both techniques is covered in [2] and in the companion paper [5]. For this work, marginalisation was used.

The objective is to compute the HMM (Hidden Markov Model) state output probabilities using a reduced distribution based on reliable components. We assume that HMMs have been trained on clean data and the density in each state C_i can be modelled using mixtures of *M* Gaussians with diagonal covariance structure:

$$f(x|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x|k,C_i) \qquad \textbf{(1)}$$

where *x* is the input data vector and $P(k|C_i)$ are the mixture coefficients.

The marginal is determined by integrating over all missing components:

$$f(x_r|C_i) = \int f(x_u, x_r|C_i) dx_u \qquad \textbf{(2)}$$

where $x = (x_r, x_u)$ has been partitioned into reliable and unreliable parts from the masking procedure. In the following, the time-frequency regions containing data deemed reliable are collectively referred to as the *mask*.

After substitution, using the independence of reliable and unreliable subvectors within each mix and integrating the right part of (2), we obtain

$$f(x_r|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x_r|k,C_i) \int f(x_u|k,C_i) dx_u \quad \textbf{(3)}$$

For an acoustic vector in the spectral energy domain, bounds can be placed on the possible values of the unreliable components: they must lie between zero and the energy in the speech+noise mixture. If unreliable components are bounded by $[x_{low}, x_{high}]$, (3) becomes

$$f\left(x_r|C_i\right)=\sum_{k=1}^{M}P\left(k|C_i\right)f\left(x_r|k,C_i\right)\int_{x_{low}}^{x_{high}}f\left(x_u|k,C_i\right)dx_u \quad \textbf{(4)}$$

Assuming diagonal Gaussians for the components of the mixture, the integral in (4) can be expanded as the difference of two error functions.

3. NOISE ESTIMATION METHODS

In the case of spectral subtraction, the noise estimate is subtracted from the signal and resulting negative amplitudes are set to zero.

A. Simple Estimation

A simple way to estimate the noise spectrum is to use the periods of non-speech activity. The average of the first 10 frames is taken as the noise spectrum.

B. Weighted Average Method

This method, introduced by Hirsch and Ehrlicher [4], attempts to adapt its estimate to changes in the noise. It is based on a first order recursion to estimate the level of noise and uses an adaptive threshold to stop the recursion when the speech is most likely to be present. For each frequency band, an estimate of the noise magnitude in frame i is obtained by the first order recursion:

$$\text{if } X(i)\le \beta N(i-1),\ \text{then } N(i)=\alpha N(i-1)+(1-\alpha)X(i)$$
$$\text{else } N(i)=N(i-1) \quad \textbf{(5)}$$

where X is the spectral magnitude, N the noise magnitude estimation, $\beta\approx 2$ and $\alpha\approx 0.98$. The initialisation is based on technique A.

C. Second Order Method

Use of equation (5) tends to overestimate the noise in the frequency channels where the signal-to-noise ratio is low. In order to reduce this error, a second order recursion is added:

$$E(N^2)^i=\alpha E(N^2)^{i-1'}+(1-\alpha)X^2(i) \quad \textbf{(6)}$$

Equations (5) and (6) estimate the noise mean and variance respectively. Both B and C are applied only when the frame is considered noisy:

$$\left\|X-N\right\|\le k\sigma_b \quad \textbf{(7)}$$

where σ_b is the noise standard deviation [7].

D. Histogram Method

This technique was also developed by Hirsch and Ehrlicher [4]. It is an improved version of the weighted average method (technique B). Past spectral values below the adaptive threshold (equation 5) over a given duration window are used to produce sub-band energy histograms. Because the speech distribution has been roughly eliminated from such a histogram, the noise level is estimated as its maximum. To obtain a more accurate estimation of the noise spectrum, a Gaussian distribution can be fitted to the energy values.

4. INTEGRATING SPECTRAL SUBTRACTION AND MISSING DATA THEORY

Identification of unreliable data is based on the *negative energy* and *SNR* criteria.

The *negative energy criterion* was introduced by Drygajlo and El-Maliki [3]. If the observed magnitude in any frame is denoted by $|s+n|$ and the estimated noise spectrum by $\hat{n}$, then the *negative energy criterion* drops spectral regions from the mask if:

$$|s+n|-\hat{n}<0 \quad \textbf{(8)}$$

$\hat{s}=|s+n|-\hat{n}$ is the resulting "cleaned" speech.

In the context of spectral subtraction, the estimate of the noise spectrum can be used to identify those regions dominated by noise. This *SNR criterion* treats data as unreliable when the estimated *SNR* is negative:

$$log\left(\frac{\hat{s}^2}{\hat{n}^2}\right)<0 \text{ or } \hat{s}^2<\hat{n}^2 \quad \textbf{(9)}$$

By adding $\hat{s}^2$ in both sides of (9) and using the Cauchy-Schwartz inequality, we obtain the *SNR criterion* which treats data as missing if

$$\hat{s}^2<\tfrac{1}{2}|s+n|^2 \quad \textbf{(10)}$$

5. ASSESSING THE MASKING CRITERIA

In order to compare the masks obtained by applying the criteria of section 4 with the `right answer' we form `a-priori masks' using the clean speech signal in addition to the noisy speech. The a-priori mask is formed from those regions where the energy in the noisy speech is within 1dB of the energy in the clean speech: i.e. the regions which are speech-dominated.

Figure 1 shows the a-priori mask and the masks resulting from the *negative energy criterion* alone and from both criteria (the *joint mask*) for a spoken digit sequence "439" added to factory noise at 10dB. Technique A was used for both criteria. Unless the spectral subtraction is near perfect, it is clear that the *negative energy criterion* is not good enough on its own but in conjunction with the *SNR criterion* we obtain a good approximation to the a priori mask. Because we are using a constant noise estimate rather than the real (varying) noise level, the joint mask has a cleaner

appearance than the a-priori mask. More examples of such masks can be found in [2]. How well does this technique stand up in recognition experiments?

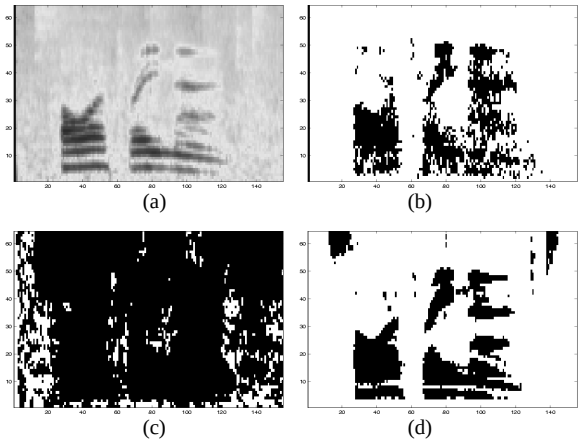

Figure 1: Masks obtained from a noisy signal (speech "439" mixed with the factory noise at 10dB), (a) spectrogram, (b) a-priori mask, (c) negative energy mask and (d) joint mask.

6. RECOGNITION EXPERIMENTS

A. Experimental Details

The Tidigits corpus (American English digit sequences) was used to test the approaches outlined in the previous sections. Acoustic vectors were obtained via a 64 channel auditory filter bank [1] 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 8ms time constant, and sampled at a frame-rate of 10ms. The training section of the corpus was used to estimate the parameters of 12 HMMs (1-9,'oh,'zero' and a silence model), each with 8 emitting states. Observations in each state were modelled with a 10-component mixture. Testing was performed on a 240 utterance subset of the TIDigits test portion. HTK [8] was used for training and a local MATLAB Viterbi decoder adapted for missing data was used for all recognition tests.

Three noise signals (car, lynx helicopter and factory) from the Noisex corpus were added at a range of SNRs. Because of limited space, we present only the results obtained with factory noise here. This is the most difficult for recognition, because it is the least stationary, has energy peaks in the formant region and additional impulsive energetic regions (hammer blows). Similar conclusions to those presented below were obtained for the different noise sources, albeit from a higher performance baseline. For a full account, see [2].

B. Results

1) Spectral Subtraction comparisons

Baseline performance, with no noise estimation technique, and performance with spectral subtraction alone is shown in figure 2 as a function of SNR with added factory noise. Spectral subtraction uses the different noise estimation techniques described in section 3. As can be seen, all these techniques improve recognition accuracy but the weighted average and histogram methods give the best results.

2) Missing Data results

Figure 3 shows the performance with the missing data method only, by using the joint criterion to define a mask for recognition on the noisy speech. For convenience, only plots for simple noise estimation and weighted average estimation are shown. Results are given for missing data recognition with and without the bounds constraint (equations 3 and 4). Performance improves dramatically with the use of bounds. This can be explained by the following argument:

- In the joint-constraint masks (see figure 1) there are time segments in which all or nearly all of the data is deemed unreliable.
- In these regions, the marginalisation expression (3) provides little or no discrimination.
- The counter-evidence provided by the bounds constraint (equation 4) recovers some discrimination.

The missing data results represent a significant improvement upon those obtained from spectral subtraction. Hirsch's weighted average technique produces a small advantage at low SNRs.

3) Combination results

Table 1 shows the recognition results when spectral subtraction and missing data (i.e. marginalisation with bounds) are combined. The results show a further small improvement from missing data alone. Useful performance on factory noise is obtained up to around 10dB SNR For the more benign car noise, a level of 0dB can be tolerated. For lynx helicopter noise the corresponding level is around 5dB [2].

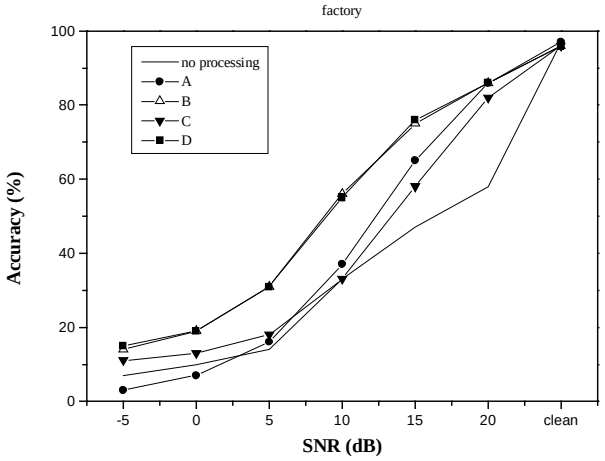

Figure 2: Recognition accuracy against SNR (factory noise) for spectral subtraction alone, with the different techniques described in section 3.

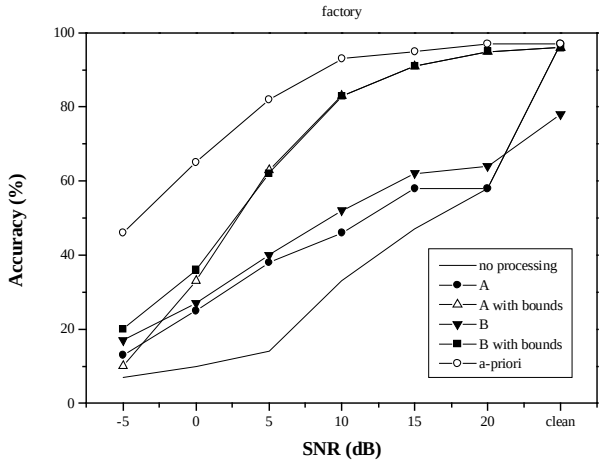

Figure 3: Recognition accuracy against SNR (factory noise) for missing data alone (without and with bounds), with two techniques from section 3. The accuracy obtained from the a-priori information is also added for comparison.

SNR	0	10	20
no processing	10	33	58
technique A	34	81	94
technique B	38	83	95
a-priori	65	93	97

Table 1: Recognition accuracy (%) on factory noise for combination spectral subtraction/missing data with bounds, with two noise estimation techniques (section 3) and for the a-priori information.

6. CONCLUSION

For the application of missing data theory, it is crucial to identify the reliable data. Here, we proposed a joint criterion for detecting the unreliable information containing in the signal. The recognition results obtained from the marginalisation with bounds are very promising.

The a-priori mask results indicate that reliable information is present: performance depends on how good the noise estimation technique is.

ACKOWLEDGEMENTS

Thanks to Andrew Morris and Miguel Carreira-Perpinan for useful discussions.

This work is supported within the SPHEAR Transfer and Mobility of Researchers network and the EU LTR Project RESPITE.

REFERENCES

[1] M.P. Cooke, 1993, "Modelling Auditory Processing and Organisation", Cambridge University Press.

[2] M.P. Cooke, P. Green, L. Josifovski and A. Vizinho, 1999, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data", Research Memorandum CS-99-05, Dept. of Computer Science, University of Sheffield.

[3] A. Drygajlo and M. El-Maliki, 1998, "Speaker Verification in Noisy Environment with Combined Spectral Subtraction and Missing Data Theory", Proc. ICASSP-98, Vol 1, p 121-124.

[4] H.G. Hirsch and C. Ehrlicher, 1995, "Noise Estimation Techniques for Robust Speech Recognition", ICASSP-95, p 153-156.

[5] L. Josifovski, M.P. Cooke , P. Green and A. Vizinho, 1999, "State Based Imputation of Missing Data for Robust Speech Recognition and Speech Enhancement", Eurospeech-99.

[6] R.J. McAulay and M.L. Malpass, 1980, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 2.

[7] C. Mokbel, 1992, "Reconnaissance de la Parole dans le bruit: Bruitage/Débruitage", Ecole Nationale Supérieure des Télécommunications.

[8] S.J. Young and P.C. Woodland, 1993, "HTK Version 1.5: User, Reference and Programmer Manual", Cambridge University Engineering Department, Speech group.