IDIAP RESEARCH REPORT

Noise PDF Transformation in Secondary Feature Processing

Andrew C. Morris

IDIAP-RR 02-29

December 2002

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP-RR 02-29

NOISE PDF TRANSFORMATION IN SECONDARY FEATURE PROCESSING

Andrew C. Morris

December 2002

Abstract

Motivated by the human ability to maintain a high level of speech recognition when large parts of the spectrogram are masked (i.e. dominated) by noise, the original "missing data" (MD) approach to noise robust speech recognition was based on the paradigm whereby models are trained on clean speech and during recognition parts of the spectrogram identified as being dominated by noise are ignored by marginalisation over the clean data pdf. However, the implied rule that each spectral data value should be treated as either as 100% clean or completely missing is inaccurate. The performance of MD based recognition has been steadily improving over the last few years with each increase in the accuracy of the modelling of clean-data uncertainty. Another assumption of the MD approach, which is more reasonable, is that it is often relatively easy to obtain an accurate estimate of the local noise spectrum. In this report we present an analysis of the way in which uncertainty in the noise spectrum is transformed into uncertainty in the clean speech spectrum. The take up of this approach will depend on the existence of closed form and computationally feasible solutions to the equations here presented. This is a preliminary study and no empirical tests are included. It is intended as a theoretical foundation from which practical solutions may be developed in future.

Keywords: noise robust ASR, noise masking, missing data, data uncertainty, pdf transformation

Acknowledgements: Some of the ideas presented arose from discussions with Nestor Becerra Yoma. This work was carried out within the EC/OFES RESPITE (REcognition of Speech by Partial Information TEchniques) project.

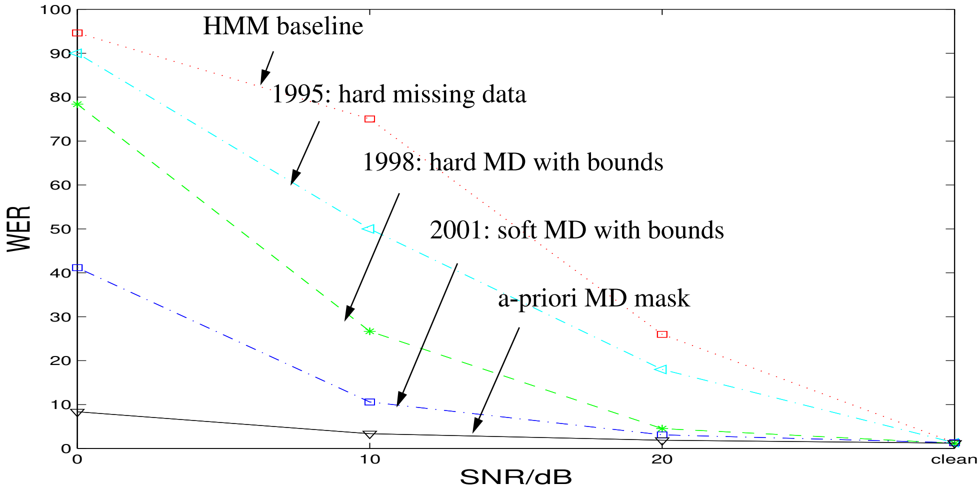

Figure 1. Comparison of WER vs. SNR for baseline HMM system and four missing data based systems (last using a-priori missing-data masks). Task is Aurora 2.0 connected digits, test (a). Results averaged over all 4 noise types.

1. Introduction

Humans are able to maintain a high level of speech recognition when large parts of the spectrogram are masked (i.e. dominated) by noise [9]. In the original "missing data" (MD) approach to noise robust speech recognition [6] was based on the paradigm whereby models are trained on clean speech and during recognition parts of the spectrogram identified as being dominated by noise are ignored by marginalisation over the clean data pdf. However, the implied rule that each spectral data value should be treated as either as 100% clean or completely missing is inaccurate. The performance of MD based recognition has been steadily improving over the last few years with each increase in the accuracy of the modelling of clean-data uncertainty (See Fig. 1). The first main improvement was to take into account the fact that (on the assumption that speech and noise energy are simply additive) the unknown clean speech value is bounded above by its observed value, and below by zero. This model [12] has been referred to as "bounded marginalisation". The second main improvement was to aknowledge that the choice of labelling each spectral coefficient as either clean or noisy was itself probabilistic. In this "soft missing data" model [2] the solutions for the "clean" and "missing" cases are combined in a weighted sum, where the weight w applied to the clean case is the estimated probability that the coefficient is clean (the other weight is $(1-w)$).

In [11] it was shown that the "soft missing data" model can be viewed as principled implementation of the Bayes decision rule for MAP decoding in the presence of data uncertainty, where the data uncertainty is expressed as a separate pdf for each spectral data coefficient, and this pdf is modelled as a mixture pdf comprising a dirac pdf for the clean data, and a uniform pdf (over the interval [0,observed value]) for the masked data. Both the "soft missing data" model, and related models for "uncertainty decoding" introduced around the same date [3], exploit the fact that it is often relatively easy to obtain an accurate, albeit probabilistic, estimate of the local noise spectrum. From this one can obtain an accurate estimate of the underlying clean data pdf. In this representation data coefficients estimated as clean have highly peaked pdfs, while data values deemed uncertain have very flat pdfs.

For observed noisy channel output Y, let S denote the clean input signal, and N denote additive noise. For the linear noise model Y = A.S + N, with N Gaussian, there are a number of established "blind deconvolution and equalisation" techniques which permit s to be recovered from x [13]. The accuracy with which s can be recovered depends on the noise process and the estimation technique used, as well as the number, type and position of microphones used, and the sampling frequency. In the absence of a better noise variance estimate, the variance in the noise power estimate can often be taken as proportional to the square of the estimated noise power [13]. While the details of the particular noise pdf estimation technique used are beyond the scope of this report, it is therefore not difficult to obtain the mean and variance for the additive noise power associated with each noisy observed spectral power coefficient. In this report we assume additive distortion only (convolutional distortion converted if necessary to additive distortion through transformation of the channel output into the log spectral domain).

In Section 2 we recollect the equations for MAP decoding with uncertain data from [11]. In Section 3 we present an analysis of the way in which uncertainty in the noise spectrum is transformed into uncertainty in the clean speech spectrum. In Section 5 we discuss various implementation issues.

2. Theoretical basis for recognition with uncertain data

We do not address the issue of classifier training with uncertain data. It is assumed here that models are trained with clean speech data. This assumption can be inaccurate and an analysis of optimal training with uncertain data would be a natural counterpart to the present study. Let the clean spectral power data for a given utterance (after binning on the Mel scale) be denoted by S. Let the compressed clean spectral data, as is commonly used in model training and in recognition, be denoted by $X = C(S)$ (e.g. logarithm or cube root compression). Let the noise spectral power data be denoted N and the noisy spectral data, assuming additive noise, be denoted $Y = S + N$. Let the compressed noisy spectral data be denoted $Z = C(Y)$.

In the original “missing data” (MD) approach to noise robust ASR [6], each spectral coefficient was assumed to be either dominated by speech = “clean”, or dominated by noise = “missing”. Having partitioned the noisy observation sequence Z into “present” $X_p = Z_p$ and noisy components $X_m \neq Z_m$, the noisy data Z_m was then treated as “missing” by marginalising the original objective $P(Q|X)$ over X_m to obtain $P(Q|X_p)$.

While the term “clean data” in the context of MD ASR originally referred to speech data in the absence of noise, what we mean here by “clean data” is any data which is from the same data population as that used for model training. However, while models trained with noisy data can provide considerable increases in noise robustness, in this case the “clean” (i.e. matched) data pdf could not be inferred directly from an estimated noise data pdf, because matching data should itself be noisy.

Note also that what we refer to as “Q” here is really (W, Q), where W is the word sequence to be recognised, and Q is the associated HMM state sequence. However, to simplify notation we just write Q. The same equations carry through with (W, Q) if $P(Q)$ is everywhere replaced by $P(W, Q)$, or $P(W)P(Q|W)$.

2.1 Evolution of models for classification with uncertain data

A parametric classifier using model M is usually trained on clean data Z_{tr} to have parameters Θ using

$$\Theta = argmax_{\theta} P(Z_{tr}|Q_{tr}, Z_{tr}, M;\theta) \quad (1)$$

Although ML parameter estimation is theoretically suboptimal, we are not concerned in this report about how the model parameters Θ were obtained.

In general the Bayes optimal class decision Q_{MAP} depends on what is known κ_X about the value of the clean data sequence X when we are given the noisy observed data sequence Z. If κ_X tells us that $X = Z$ (i.e. if we know a-priori that Z is clean), then we have the usual MAP rule (see Fig.1 "HMM baseline"),

$$Q_{MAP} = argmax_Q P(Q|Z, M, Z_{tr}) \cong argmax_Q P(Q|Z;\Theta) \tag{2}$$

It is shown in [12] that in general, if our knowledge of the clean model data X is incomplete or uncertain, then[1]

$$Q_{MAP} = argmax_Q E[P(Q|Z;\Theta)|\kappa_X, \Theta] \tag{3}$$

In the first MD model (1995) [6] κ_X tells us that Z_p is clean and Z_m is completely uninformative (viz. that $P(clean)$ for each coefficient of Z is 1 or 0). In this case (see Fig.1 "hard missing data")

$$Q_{MAP} = argmax_Q P(Q|Z_p, \Theta) \tag{4}$$

This situation could occur in reality in the case of visual masking by opaque objects. In this case the approach to recognition with visual missing data described in [1] gives the same solution. However, that approach does not generalise to the case of uncertain rather than missing data, which instead must be based on Eq.3 from [12].

If $P(clean)$ are still 1 or 0, but we assume that noise is additive $(Y = S + N)$, then (on the assumption of simple additivity of speech and noise energy) *uncertain data is bounded below by zero and above by the observed noisy spectral value* (1998) [12], then, assuming[2] $p(X_m|\kappa_X) = U(X_m;0, Z_m)$,

$$Q_{MAP} = argmax_Q P(Q|Z_p, \Theta)\int_0^{Z_m} p(X_m|Z_p, Q, \Theta)dX_m \tag{5}$$

(See Fig.1 "hard MD with bounds"). If uncertain data are still bounded but $P(clean)$ is anywhere in $[0, 1]$, then it was shown in [2, 11] that, providing the prior pdf $p(X)$ is approximately flat, then (see Fig.1 "soft MD with bounds")

$$Q_{MAP} = argmax_Q P(Q|\Theta)\int_0^{Z} p(X|Q, \Theta)p(X|\kappa_X)dX \tag{6}$$

In the case of Eq.6, during standard HMM/GMM Viterbi decoding each $p(x_i|m_j, q_k, \Theta)$ is replaced by $\int_0^{z_i} p(x_i|m_j, q_k, \Theta)p(x_i|\kappa_i)dx_i$. Equation 6 forms the basis for Bayes optimal decoding when the clean data value is uncertain and is represented by a pdf. In the next section we look at new ways in which the clean-data pdf can be evaluated from a given noise pdf for each spectral component.

1. *Any technique for classification with clean models and uncertain data which differs in any way from this rule, including the general class of data imputation techniques,* $Q_{best} = argmax_Q P(Q|\hat{X}, \Theta)$, *is therefore a-priori sub-optimal.*
2. *This assumption has dubious validity. An alternative model is presented below.*

3. Inferring clean data pdf from estimated noise pdf

In this section we present two different approaches to this problem. The first approach is an incremental advance building on the success of the 2-mix representation of the clean data pdf introduced in [2], where it is assumed that speech and noise energies are strictly additive. The second approach is more ambitious in that it is assumed only that the speech and noise signals are additive, so that their energies may add together or cancel out, according to the phase difference between them (assuming a constant phase difference over each speech sample).

3.1 From 2-mix to 3-mix clean-data pdf, assuming additivity of speech/noise energy

In [2] the clean data pdf $p(x_i|\kappa_i)$ was approximated by a mixture pdf,

$$p(x_i|\kappa_i) = P(snr_i > 0)p(x_i|snr_i > 0) + P(snr_i \le 0)p(x_i|snr_i \le 0) \tag{7}$$

It was also assumed that $(snr_i > 0)$ tells us that x_i is dominated by speech [on the basis that under log compression, $(\log(a) > \log(b)) \Rightarrow (a \gg b)$], so that

$$(x_i|snr_i > 0) = (x_i|\log s_i > \log n_i) = (x_i|s_i \gg n_i) \sim \delta(x_i - z_i) \tag{8}$$

i.e. the probability density is approximately zero everywhere except at the observed value. It was also assumed that $(snr_i \le 0)$ tells us that x_i is dominated by noise, so that we know nothing about the clean value except that it is bounded below by 0 and above by the observed noise data value, i.e. that

$$(x_i|snr_i \le 0) \sim U(x_i;0, z_i) \tag{9}$$

The mixture pdf in Eq.7 makes a number of modelling assumptions which can easily be seen to be highly inaccurate.

1. Eq.8 assumes that when $snr_i > 0$, the observed compressed (i.e. log or cube root) spectral value z_i is "dominated by speech", i.e. that $s_i \gg n_i$. While it is true that the observed value will be dominated by speech for some $snr_i > snr_{hi}$, this assumption would be considerably more accurate if $(snr_{hi} > 0)$.
2. Eq.9 assumes that when the SNR is negative, the only information the observed value gives about the clean value is an upper bound. This assumption would be more accurate for some $snr_i < snr_{lo}$ if $(snr_{lo} < 0)$.
3. Even when z_i is dominated by noise, the assumption that the pdf for the compressed data x_i should be uniform is at least debatable. In this case it would seem more reasonable that the pdf for s_i is uniform, in which case the pdf for z_i will be far from uniform, with most of the probability mass concentrated near to z_i.
4. The point estimate of the local noise power is used in the calculation of the mix weight $P(snr_i > 0)$ in Eq.7, but the shape and spread of the noise power pdf is not used, though this extra information is often available.

Assuming that an estimate for the spectral noise power pdf as available (here assuming that noise is uncorrelated across frequency so that we can model a separate univariate pdf for each spectral coefficient) we now show how this noise pdf can be used to estimate a pdf for each clean spectral coefficient. This clean data pdf can then be used to overcome the modelling inaccuracies described in points 1-4 above[1].

1. *If only a point estimate of the noise is available, the noise power variance can often be approximated by the square of the noise power estimate [13].*

In repeating here the MD tests using the 2-mix pdf given by Eqns.7-9 we noted that a number of minor adjustments to these equations which we tested resulted in a dramatic loss in performance. In what follows we are going to give an alternative derivation for the clean data pdf which does not require a mixture pdf. However, rather than abolish the mix pdf in favour of a single function, we are going to stay with the idea of a mix pdf (which seem to work very well), but change the pdf details as follows:

1. replace the two mix conditions $(snr < 0)$ and $(0 < snr)$ by the three conditions $(snr < snr_{lo})$, $(snr_{lo} < snr < snr_{hi})$ and $(snr_{hi} < snr)$.
2. for $(snr < snr_{lo})$ we will assume not that the noisy compressed data has a uniform pdf, but that the uncompressed clean spectral power has a uniform pdf.
3. for $(snr_{lo} < snr < snr_{hi})$ we will assume not that the noisy compressed data has a uniform pdf, but that the uncompressed clean spectral power has a uniform pdf.

We now consider appropriate pdf mix components for each of these three conditions.

For the conditions $(snr_{hi} < snr)$ and $(snr < snr_{lo})$ we can make use of the Dirac and Uniform pdfs used before with the conditions $(0 < snr)$ and $(snr < 0)$ respectively, though now with more attention to detail.

For the condition $(snr_{lo} < snr < snr_{hi})$, clean data x is bounded by the interval $[x_{lo}, x_{hi}]$. For any given x_i these bounds are determined by given snr limits (snr_{lo}, snr_{hi}) and the observed noisy value z_i as follows:

$$snr = \log\left(\frac{s}{n}\right) = \log\left(\frac{s}{y-s}\right), \text{ so that } s = y/(1+e^{-snr}) \tag{10}$$

and $x = C(s)$ and $z = C(y)$, so $x_{lo} = C[C^{-1}(z_i)/(1+e^{-snr_{lo}})]$ and $x_{hi} = C[C^{-1}(z_i)/(1+e^{-snr_{hi}})]$ (11)

Assuming that an estimated pdf is available for the spectral noise power n as $p_N(n)$, over $[0, C^{-1}(z)]$, the clean data pdf $p_X(x)$ can be obtained as follows. For $x > x_{hi}$ assume $x \sim \delta(x - C^{-1}(z))$. Otherwise we will need to make use of the following identity (where the function B is the inverse of C) (see Appendix B for derivation).

$$p_X(x) = p_N(B(z) - B(x))B'(x) \tag{12}$$

For $x < x_{lo}$ assume that $p_N(n) \sim U(n;0,y)$.

For $C(s) = s^{1/3}$, $p_X(x) = 3x^2/z$ for $x \in [C^{-1}(0), z] = [0, z]$

For $C(s) = \log(s)$, $p_X(x) = e^x/(z-1)$ for $x \in [C^{-1}(0), z] = [-\infty, z]$

Note here that whichever compression function is used, the effect of this compression of the clean data pdf is to make clean-speech power values closer to the observed noisy value much more probable than lower values, even when the noise power pdf is uniform.

If we have obtained only a mean and variance for the noise, then a candidate noise pdf for a continuous random variable bounded by $[0, y]$, is the following form of beta pdf[1] [5],

---

1. *Although this pdf is quite complicated, both it and its product with a Gaussian pdf can be integrated in closed form - which is vital for implementation purposes. However, for practical purposes it may be sufficient to approximate* $f_{N(n)}$ *by a Gaussian.*

$$f_N(n;v,w) = \frac{n^{v-1}(y-n)^{w-1}}{y^{v+w+1}\beta(v,w)} \quad (13)$$

where $u = \mu_n(1-\mu_n)/\sigma_n^2 - 1$, $v = \mu_n u$, $w = (1-\mu_n)u$. (14)

4. Generic clean-data pdf, assuming only additivity of speech/noise signal

In this section we assume only that the speech and noise signal, not their separate energies, are additive. They may reinforce or cancel, depending on the phase difference between them (which we will assume constant). In this case the energy of the combined speech and noise signal, at each centre frequency f, is the average value of the square of the sum of the speech and noise signals. If we can assume that the signal sample contains several pitch periods, then this average energy rate x^2 is given in terms of the speech signal $s(t) = a_s \sin(t)$ (with energy rate $s^2 = a_s^2/2$) and noise signal $n(t) = a_n \sin(t+\phi)$ (with energy rate $n^2 = a_n^2/2$) by

$$x^2 = a_x^2/2 = f\int_0^{1/f}(s(t)+n(t))^2 dt = f\int_0^{1/f}(a_s\sin(t)+a_n\sin(t+\phi))^2 dt \quad (15)$$

$$= (a_s^2 + a_n^2 + 2a_s a_n Cos(\phi))/2 = s^2 + n^2 + 2snCos(\phi) \quad (16)$$

The clean speech power s^2 is now a function of the observed noisy observed power x^2 and the noise power n^2 as follows:

$$s^2 = (-nCos(\phi) + (x^2 - n^2 Sin(\phi))^{1/2})^2 \quad (17)$$

We can now use Eq.12 together with Eq.17 to obtain the clean-speech power pdf directly from a given the noise power pdf $p_{N^2}(n^2)$.

First case: assume $\phi = 0$, so that $x = s + n$. In this case we have

$$s = A(n) = x - n \Rightarrow p_S(s) = p_N(B(s))B'(s) = p_N(x-s) \quad (18)$$

Then we have $y = A(s) = C(s)$, $z = C(x)$, $x = B(z)$

$$\Rightarrow \quad p_Y(y) = p_S(B(y))B'(y) = p_N(x - B(y))B'(y) = p_N(B(z) - B(y))B'(y) \quad (19)$$

Two commonly used compression functions are $z = \log(x)$ and $z = x^{1/3}$.

$$p_X(x) = p_N(B(z) - B(x))B'(x) \text{ for } x \in [C(0), z], \text{ else } p_X(x) = 0$$

When $p_N(n) = U(n;0,B(z))$ and $B(z) = z^3$,

then $p_X(x) = 3x^2/z^3$ for $x \in [0, z]$

For given noise pdf $p_N(n)$, and compression function C, with inverse B, the pdf for clean compressed data is in general given by

$$p_X(x) = p_N(B(z) - B(x))B'(x) \text{ for } x \in [C(0), z] \quad (20)$$

e.g. for estimated noise pdf uniform, with cube root compression, the clean compressed data pdf is

$$p_X(x) = 3x^2 / z^3 \tag{21}$$

5. Discussion

We have described the steps by which the original paradigm of recognition with missing spectral data led to a more accurate model for recognition with uncertain data. On the understanding that it is not difficult to obtain an accurate estimate of the noise pdf for each spectral data coefficient, we then described a number of ways in which the clean-data pdf can be modelled in terms of the noise pdf. The first and simplest of these models extends the 2-mixture pdf previously used in the "soft missing data" model, to a 3-mixture pdf. The second more ambitious model aims to increase the accuracy of the clean-data pdf model still further by taking into account the speech/noise phase difference. While the framework for the necessary calculations was introduced, this calculation was only implemented so far for the relatively trivial case in which $\phi = 0$. It would be instructive in future to follow through the necessary calculations for a number of further cases, according to whether

1. noise power is a given exact value or a pdf
2. logarithm or cube-root compression is in use
3. phase difference is zero, a fixed known value, or has a uniform pdf over $[-\pi, \pi]$

In the most general case, the noise power will be represented by a pdf and the phase difference will be unknown (so having a uniform pdf). It has been verified that in this case there is a closed form solution for the case of cube-root compression, but not for log compression.

The take up of this approach will depend on the existence of closed form and computationally feasible solutions to the equations here presented. This is a preliminary study and no empirical tests are included. It is intended as a theoretical foundation from which practical solutions may be developed in future.

Appendix A: Nomenclature and commonly used abbreviations

ASR	Automatic Speech Recognition
SNR	Signal to Noise Ratio
WER	Word Error Rate
$P(x)$	probability of "event x" occurring
$p(x)$	probability density at x of a continuous value
$P(x;\Theta)$	function with given model M (not indicated) with parameters Θ used to estimate $P(x)$
$\hat{P}(x)$	function (model and parameters unspecified) used to estimate $P(x)$
q_k	speech unit whose presence is being estimated, or event that data x is from this class
x, x^n	data window vector at time step n
d	number of spectral sub-bands
x clean	x is from data population used in model training (i.e. x has no data mismatch)
κ_x	any knowledge which can be used to fix or constrain the value of x
$\Phi(x)$	cumulative density function for the standard Gaussian pdf
$\delta(x-a)$	the Dirac delta pdf, centred at (a). $\delta(x) = 0$ for $x \neq 0$, and integrates to 1
$U(x;a,b)$	the Uniform pdf over the interval $[a, b]$
s	unknown uncompressed clean speech
$x = C(s)$	unknown compressed clean speech
$y = s + n$	uncompressed noisy speech, $n \in [0, y]$
$z = C(y)$	observed compressed noisy speech $(y = C^{-1}(z) = B(z))$
$p_N(n)$	given (estimated) noise pdf
$p_X(x)$	required compressed clean data pdf

Appendix B: Results used

If y is a 1-1 increasing function $A(x)$ of variable x, with $B(x) = A^{-1}(x)$, and $x \sim p_X(x)$, then[1]

$$p_Y(y) = p_X(B(y))B'(y) \tag{1}$$

If y is a 1-1 decreasing function of x, then

$$p_Y(y) = -p_X B(y)B'(y) \tag{2}$$

so in general, if y is a 1-1 function of x, then

$$p_Y(y) = -p_X B(y)|B'(y)| \tag{3}$$

If $A(x)$ is increasing then $B(y)$ is increasing, so $x \in [x_1, x_2]$ => $B(y) \in [x_1, x_2]$

$$=> y \in [A(x_1), A(x_2)] \tag{4}$$

If $A(x)$ is decreasing then $B(y)$ is decreasing, so $x \in [x_1, x_2]$ => $B(y) \in [x_1, x_2]$

$$=> y \in [A(x_2), A(x_1)]\,. \tag{5}$$

Equation relating (x, s, n)

$$x^2 = s^2 + n^2 + 2ns\cos\theta \tag{6}$$

Solving (6) for s, with $x^2 = a^2n^2$, s real and $x = s + n$ when $\theta = 0$, gives $a^2 \geq 1$ and

$$s = -nCos(\theta) + n(a^2 - Sin^2(\theta))^{1/2} \tag{7}$$

1. *This can be shown easily as follows.* $F_Y(y) = P(Y < y)$. *If* $y = A(x)$ *is increasing, then applying* B *(the inverse of* A *) to both sides of the inequality gives* $F_Y(y) = P(X < B(y)) = F_X(B(y))$. *Differentiating both sides with respect to* y *then gives* $f_Y(y) = f_X(B(y))|B'(y)|$.

References

[1] Ahmed, S. & Tresp, V. (1993) “Some solutions to the missing feature problem in vision”, in Advances in Neural Information Processing Systems 5, Morgan Kauffman, San Mateo, pp. 393-400.

[2] Barker, J., Josifovski, L., Cooke, M.P. & Green, P.D. (2000) “Soft decisions in missing data techniques for robust automatic speech recognition”, Proc. ICSLP-2000, pp.373-376.

[3] Droppo, J., Acero, A. & Deng, L. (2002) “Uncertainty decoding with SPLICE for noise robust speech recognition”, Proc. ICASSP’02, Vol.1, pp. 57-60.

[4] El-Maliki, M. (2000) “Speaker verification with missing features in noisy environments”, PhD thesis, Dept. d’Electricité, Ecole Polytechnique Fédérate de Lausanne.

[5] Evans, M., Hastings, N. & Peacock, B. (2000) **Statistical Distributions**, Wiley series in probability and statistics.

[6] Green, P.D., Cooke, M.P. & Crawford, M.D. (1995), “Auditory scene analysis and HMM recognition of speech in noise”, Proc. ICASSP’95, pp.401-404.

[7] Hirsch, H. G. and C. Ehrlicher (1995) “Noise estimation techniques for robust speech recognition”, ICASSP95 , 1995, pp. 153-156.

[8] Kristjansson, T.T. & Frey, B.J. (2002) “Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition”, Proc. ICASSP’02, Vol.1, pp.61-64.

[9] Lippmann, R. P. & Carlson, B. A. (1997) “Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise”, Proc. Eurospeech’97, pp. 37-40

[10] McCowan, I.A., Morris, A.C. & Bourlard, B. (2002) “Improving speech recognition performance of small microsphone arrays using missing data techniques”, Proc ICSLP’02 (in press).

[11] Morris, A. C. (2001) “Data utility modelling for mismatch reduction”, Proc. CRAC workshop on Consistent & Reliable Acoustic Cues for sound analysis.

[12] Morris, A. C., Cooke, M. & Green, P. (1998) “Some solutions to the missing feature problem in data classification, with application to noise robust ASR”, Proc. ICASSP'98, pp.737-740.

[13] Vasegi, S. (1998) Advanced signal processing and digital noise reduction”, Wily Neubner.

IDIAP RESEARCH REPORT

NOISE PDF TRANSFORMATION IN SECONDARY FEATURE PROCESSING

Andrew C. Morris

IDIAP-RR 02-29

December 2002

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 - 27 - 721 77 11
fax +41 - 27 - 721 77 12
email secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP-RR 02-29

NOISE PDF TRANSFORMATION IN SECONDARY FEATURE PROCESSING

Andrew C. Morris

December 2002

Abstract

Motivated by the human ability to maintain a high level of speech recognition when large parts of the spectrogram are masked (i.e. dominated) by noise, the original "missing data" (MD) approach to noise robust speech recognition was based on the paradigm whereby models are trained on clean speech and during recognition parts of the spectrogram identified as being dominated by noise are ignored by marginalisation over the clean data pdf. However, the implied rule that each spectral data value should be treated as either as 100% clean or completely missing is inaccurate. The performance of MD based recognition has been steadily improving over the last few years with each increase in the accuracy of the modelling of clean-data uncertainty. Another assumption of the MD approach, which is more reasonable, is that it is often relatively easy to obtain an accurate estimate of the local noise spectrum. In this report we present an analysis of the way in which uncertainty in the noise spectrum is transformed into uncertainty in the clean speech spectrum. The take up of this approach will depend on the existence of closed form and computationally feasible solutions to the equations here presented. This is a preliminary study and no empirical tests are included. It is intended as a theoretical foundation from which practical solutions may be developed in future.

Keywords: noise robust ASR, noise masking, missing data, data uncertainty, pdf transformation

Acknowledgements: Some of the ideas presented arose from discussions with Nestor Becerra Yoma. This work was carried out within the EC/OFES RESPITE (REcognition of Speech by Partial Information TEchniques) project.

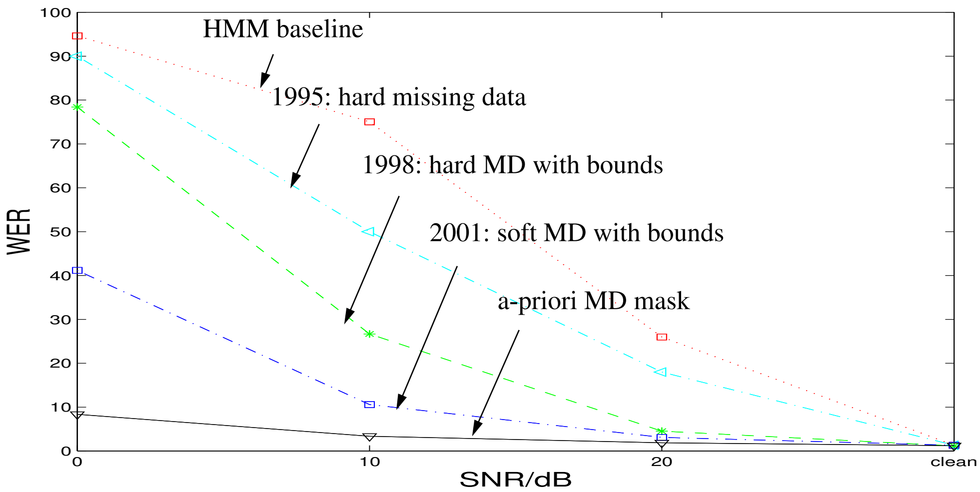

Figure 1. Comparison of WER vs. SNR for baseline HMM system and four missing data based systems (last using a-priori missing-data masks). Task is Aurora 2.0 connected digits, test (a). Results averaged over all 4 noise types.

1. Introduction

Humans are able to maintain a high level of speech recognition when large parts of the spectrogram are masked (i.e. dominated) by noise [9]. In the original "missing data" (MD) approach to noise robust speech recognition [6] was based on the paradigm whereby models are trained on clean speech and during recognition parts of the spectrogram identified as being dominated by noise are ignored by marginalisation over the clean data pdf. However, the implied rule that each spectral data value should be treated as either as 100% clean or completely missing is inaccurate. The performance of MD based recognition has been steadily improving over the last few years with each increase in the accuracy of the modelling of clean-data uncertainty (See Fig. 1). The first main improvement was to take into account the fact that (on the assumption that speech and noise energy are simply additive) the unknown clean speech value is bounded above by its observed value, and below by zero. This model [12] has been referred to as "bounded marginalisation". The second main improvement was to aknowledge that the choice of labelling each spectral coefficient as either clean or noisy was itself probabilistic. In this "soft missing data" model [2] the solutions for the "clean" and "missing" cases are combined in a weighted sum, where the weight w applied to the clean case is the estimated probability that the coefficient is clean (the other weight is $(1-w)$).

In [11] it was shown that the "soft missing data" model can be viewed as principled implementation of the Bayes decision rule for MAP decoding in the presence of data uncertainty, where the data uncertainty is expressed as a separate pdf for each spectral data coefficient, and this pdf is modelled as a mixture pdf comprising a dirac pdf for the clean data, and a uniform pdf (over the interval [0,observed value]) for the masked data. Both the "soft missing data" model, and related models for "uncertainty decoding" introduced around the same date [3], exploit the fact that it is often relatively easy to obtain an accurate, albeit probabilistic, estimate of the local noise spectrum. From this one can obtain an accurate estimate of the underlying clean data pdf. In this representation data coefficients estimated as clean have highly peaked pdfs, while data values deemed uncertain have very flat pdfs.

For observed noisy channel output Y, let S denote the clean input signal, and N denote additive noise. For the linear noise model Y = A.S + N, with N Gaussian, there are a number of established "blind deconvolution and equalisation" techniques which permit s to be recovered from x [13]. The accuracy with which s can be recovered depends on the noise process and the estimation technique used, as well as the number, type and position of microphones used, and the sampling frequency. In the absence of a better noise variance estimate, the variance in the noise power estimate can often be taken as proportional to the square of the estimated noise power [13]. While the details of the particular noise pdf estimation technique used are beyond the scope of this report, it is therefore not difficult to obtain the mean and variance for the additive noise power associated with each noisy observed spectral power coefficient. In this report we assume additive distortion only (convolutional distortion converted if necessary to additive distortion through transformation of the channel output into the log spectral domain).

In Section 2 we recollect the equations for MAP decoding with uncertain data from [11]. In Section 3 we present an analysis of the way in which uncertainty in the noise spectrum is transformed into uncertainty in the clean speech spectrum. In Section 5 we discuss various implementation issues.

2. Theoretical basis for recognition with uncertain data

We do not address the issue of classifier training with uncertain data. It is assumed here that models are trained with clean speech data. This assumption can be inaccurate and an analysis of optimal training with uncertain data would be a natural counterpart to the present study. Let the clean spectral power data for a given utterance (after binning on the Mel scale) be denoted by S. Let the compressed clean spectral data, as is commonly used in model training and in recognition, be denoted by $X = C(S)$ (e.g. logarithm or cube root compression). Let the noise spectral power data be denoted N and the noisy spectral data, assuming additive noise, be denoted $Y = S + N$. Let the compressed noisy spectral data be denoted $Z = C(Y)$.

In the original "missing data" (MD) approach to noise robust ASR [6], each spectral coefficient was assumed to be either dominated by speech = "clean", or dominated by noise = "missing". Having partitioned the noisy observation sequence Z into "present" $X_p = Z_p$ and noisy components $X_m \neq Z_m$, the noisy data Z_m was then treated as "missing" by marginalising the original objective $P(Q|X)$ over X_m to obtain $P(Q|X_p)$.

While the term "clean data" in the context of MD ASR originally referred to speech data in the absence of noise, what we mean here by "clean data" is any data which is from the same data population as that used for model training. However, while models trained with noisy data can provide considerable increases in noise robustness, in this case the "clean" (i.e. matched) data pdf could not be inferred directly from an estimated noise data pdf, because matching data should itself be noisy.

Note also that what we refer to as "Q" here is really (W, Q), where W is the word sequence to be recognised, and Q is the associated HMM state sequence. However, to simplify notation we just write Q. The same equations carry through with (W, Q) if $P(Q)$ is everywhere replaced by $P(W, Q)$, or $P(W)P(Q|W)$.

2.1 Evolution of models for classification with uncertain data

A parametric classifier using model M is usually trained on clean data Z_{tr} to have parameters Θ using

$$\Theta = argmax_{\theta} P(Z_{tr}|Q_{tr}, Z_{tr}, M;\theta) \tag{1}$$

Although ML parameter estimation is theoretically suboptimal, we are not concerned in this report about how the model parameters Θ were obtained.

In general the Bayes optimal class decision Q_{MAP} depends on what is known κ_X about the value of the clean data sequence X when we are given the noisy observed data sequence Z. If κ_X tells us that $X = Z$ (i.e. if we know a-priori that Z is clean), then we have the usual MAP rule (see Fig.1 "HMM baseline"),

$$Q_{MAP} = argmax_Q P(Q|Z, M, Z_{tr}) \cong argmax_Q P(Q|Z;\Theta) \tag{2}$$

It is shown in [12] that in general, if our knowledge of the clean model data X is incomplete or uncertain, then[1]

$$Q_{MAP} = argmax_Q E[P(Q|Z;\Theta)|\kappa_X, \Theta] \tag{3}$$

In the first MD model (1995) [6] κ_X tells us that Z_p is clean and Z_m is completely uninformative (viz. that $P(clean)$ for each coefficient of Z is 1 or 0). In this case (see Fig.1 "hard missing data")

$$Q_{MAP} = argmax_Q P(Q|Z_p, \Theta) \tag{4}$$

This situation could occur in reality in the case of visual masking by opaque objects. In this case the approach to recognition with visual missing data described in [1] gives the same solution. However, that approach does not generalise to the case of uncertain rather than missing data, which instead must be based on Eq.3 from [12].

If $P(clean)$ are still 1 or 0, but we assume that noise is additive $(Y = S + N)$, then (on the assumption of simple additivity of speech and noise energy) *uncertain data is bounded below by zero and above by the observed noisy spectral value* (1998) [12], then, assuming[2] $p(X_m|\kappa_X) = U(X_m;0, Z_m)$,

$$Q_{MAP} = argmax_Q P(Q|Z_p, \Theta)\int_0^{Z_m} p(X_m|Z_p, Q, \Theta)dX_m \tag{5}$$

(See Fig.1 "hard MD with bounds"). If uncertain data are still bounded but $P(clean)$ is anywhere in $[0, 1]$, then it was shown in [2, 11] that, providing the prior pdf $p(X)$ is approximately flat, then (see Fig.1 "soft MD with bounds")

$$Q_{MAP} = argmax_Q P(Q|\Theta)\int_0^{Z} p(X|Q, \Theta)p(X|\kappa_X)dX \tag{6}$$

In the case of Eq.6, during standard HMM/GMM Viterbi decoding each $p(x_i|m_j, q_k, \Theta)$ is replaced by $\int_0^{z_i} p(x_i|m_j, q_k, \Theta)p(x_i|\kappa_i)dx_i$. Equation 6 forms the basis for Bayes optimal decoding when the clean data value is uncertain and is represented by a pdf. In the next section we look at new ways in which the clean-data pdf can be evaluated from a given noise pdf for each spectral component.

1. *Any technique for classification with clean models and uncertain data which differs in any way from this rule, including the general class of data imputation techniques,* $Q_{best} = argmax_Q P(Q|\hat{X}, \Theta)$, *is therefore a-priori sub-optimal.*
2. *This assumption has dubious validity. An alternative model is presented below.*

3. Inferring clean data pdf from estimated noise pdf

In this section we present two different approaches to this problem. The first approach is an incremental advance building on the success of the 2-mix representation of the clean data pdf introduced in [2], where it is assumed that speech and noise energies are strictly additive. The second approach is more ambitious in that it is assumed only that the speech and noise signals are additive, so that their energies may add together or cancel out, according to the phase difference between them (assuming a constant phase difference over each speech sample).

3.1 From 2-mix to 3-mix clean-data pdf, assuming additivity of speech/noise energy

In [2] the clean data pdf $p(x_i|\kappa_i)$ was approximated by a mixture pdf,

$$p(x_i|\kappa_i) = P(snr_i > 0)p(x_i|snr_i > 0) + P(snr_i \le 0)p(x_i|snr_i \le 0) \tag{7}$$

It was also assumed that $(snr_i > 0)$ tells us that x_i is dominated by speech [on the basis that under log compression, $(\log(a) > \log(b)) \Rightarrow (a \gg b)$], so that

$$(x_i|snr_i > 0) = (x_i|\log s_i > \log n_i) = (x_i|s_i \gg n_i) \sim \delta(x_i - z_i) \tag{8}$$

i.e. the probability density is approximately zero everywhere except at the observed value. It was also assumed that $(snr_i \le 0)$ tells us that x_i is dominated by noise, so that we know nothing about the clean value except that it is bounded below by 0 and above by the observed noise data value, i.e. that

$$(x_i|snr_i \le 0) \sim U(x_i;0, z_i) \tag{9}$$

The mixture pdf in Eq.7 makes a number of modelling assumptions which can easily be seen to be highly inaccurate.

1. Eq.8 assumes that when $snr_i > 0$, the observed compressed (i.e. log or cube root) spectral value z_i is "dominated by speech", i.e. that $s_i \gg n_i$. While it is true that the observed value will be dominated by speech for some $snr_i > snr_{hi}$, this assumption would be considerably more accurate if $(snr_{hi} > 0)$.
2. Eq.9 assumes that when the SNR is negative, the only information the observed value gives about the clean value is an upper bound. This assumption would be more accurate for some $snr_i < snr_{lo}$ if $(snr_{lo} < 0)$.
3. Even when z_i is dominated by noise, the assumption that the pdf for the compressed data x_i should be uniform is at least debatable. In this case it would seem more reasonable that the pdf for s_i is uniform, in which case the pdf for z_i will be far from uniform, with most of the probability mass concentrated near to z_i.
4. The point estimate of the local noise power is used in the calculation of the mix weight $P(snr_i > 0)$ in Eq.7, but the shape and spread of the noise power pdf is not used, though this extra information is often available.

Assuming that an estimate for the spectral noise power pdf as available (here assuming that noise is uncorrelated across frequency so that we can model a separate univariate pdf for each spectral coefficient) we now show how this noise pdf can be used to estimate a pdf for each clean spectral coefficient. This clean data pdf can then be used to overcome the modelling inaccuracies described in points 1-4 above[1].

1. *If only a point estimate of the noise is available, the noise power variance can often be approximated by the square of the noise power estimate [13].*

In repeating here the MD tests using the 2-mix pdf given by Eqns.7-9 we noted that a number of minor adjustments to these equations which we tested resulted in a dramatic loss in performance. In what follows we are going to give an alternative derivation for the clean data pdf which does not require a mixture pdf. However, rather than abolish the mix pdf in favour of a single function, we are going to stay with the idea of a mix pdf (which seem to work very well), but change the pdf details as follows:

1. replace the two mix conditions $(snr < 0)$ and $(0 < snr)$ by the three conditions $(snr < snr_{lo})$, $(snr_{lo} < snr < snr_{hi})$ and $(snr_{hi} < snr)$.
2. for $(snr < snr_{lo})$ we will assume not that the noisy compressed data has a uniform pdf, but that the uncompressed clean spectral power has a uniform pdf.
3. for $(snr_{lo} < snr < snr_{hi})$ we will assume not that the noisy compressed data has a uniform pdf, but that the uncompressed clean spectral power has a uniform pdf.

We now consider appropriate pdf mix components for each of these three conditions.

For the conditions $(snr_{hi} < snr)$ and $(snr < snr_{lo})$ we can make use of the Dirac and Uniform pdfs used before with the conditions $(0 < snr)$ and $(snr < 0)$ respectively, though now with more attention to detail.

For the condition $(snr_{lo} < snr < snr_{hi})$, clean data x is bounded by the interval $[x_{lo}, x_{hi}]$. For any given x_i these bounds are determined by given snr limits (snr_{lo}, snr_{hi}) and the observed noisy value z_i as follows:

$$snr = \log\left(\frac{s}{n}\right) = \log\left(\frac{s}{y-s}\right), \text{ so that } s = y/(1+e^{-snr}) \tag{10}$$

and $$x = C(s) \text{ and } z = C(y), \text{ so } x_{lo} = C[C^{-1}(z_i)/(1+e^{-snr_{lo}})] \text{ and } x_{hi} = C[C^{-1}(z_i)/(1+e^{-snr_{hi}})] \tag{11}$$

Assuming that an estimated pdf is available for the spectral noise power n as $p_N(n)$, over $[0, C^{-1}(z)]$, the clean data pdf $p_X(x)$ can be obtained as follows. For $x > x_{hi}$ assume $x \sim \delta(x - C^{-1}(z))$. Otherwise we will need to make use of the following identity (where the function B is the inverse of C) (see Appendix B for derivation).

$$p_X(x) = p_N(B(z) - B(x))B'(x) \tag{12}$$

For $x < x_{lo}$ assume that $p_N(n) \sim U(n;0,y)$.

For $C(s) = s^{1/3}$, $p_X(x) = 3x^2/z$ for $x \in [C^{-1}(0), z] = [0, z]$

For $C(s) = \log(s)$, $p_X(x) = e^x/(z-1)$ for $x \in [C^{-1}(0), z] = [-\infty, z]$

Note here that whichever compression function is used, the effect of this compression of the clean data pdf is to make clean-speech power values closer to the observed noisy value much more probable than lower values, even when the noise power pdf is uniform.

If we have obtained only a mean and variance for the noise, then a candidate noise pdf for a continuous random variable bounded by $[0, y]$, is the following form of beta pdf[1] [5],

1. *Although this pdf is quite complicated, both it and its product with a Gaussian pdf can be integrated in closed form - which is vital for implementation purposes. However, for practical purposes it may be sufficient to approximate $f_{N(n)}$ by a Gaussian.*

$$f_N(n;v,w) = \frac{n^{v-1}(y-n)^{w-1}}{y^{v+w+1}\beta(v,w)} \tag{13}$$

where $u = \mu_n(1-\mu_n)/\sigma_n^2 - 1$, $v = \mu_n u$, $w = (1-\mu_n)u$. (14)

4. Generic clean-data pdf, assuming only additivity of speech/noise signal

In this section we assume only that the speech and noise signal, not their separate energies, are additive. They may reinforce or cancel, depending on the phase difference between them (which we will assume constant). In this case the energy of the combined speech and noise signal, at each centre frequency f, is the average value of the square of the sum of the speech and noise signals. If we can assume that the signal sample contains several pitch periods, then this average energy rate x^2 is given in terms of the speech signal $s(t) = a_s \sin(t)$ (with energy rate $s^2 = a_s^2/2$) and noise signal $n(t) = a_n \sin(t+\phi)$ (with energy rate $n^2 = a_n^2/2$) by

$$x^2 = a_x^2/2 = f\int_0^{1/f}(s(t)+n(t))^2 dt = f\int_0^{1/f}(a_s \sin(t) + a_n \sin(t+\phi))^2 dt \tag{15}$$

$$= (a_s^2 + a_n^2 + 2a_s a_n Cos(\phi))/2 = s^2 + n^2 + 2snCos(\phi) \tag{16}$$

The clean speech power s^2 is now a function of the observed noisy observed power x^2 and the noise power n^2 as follows:

$$s^2 = (-nCos(\phi) + (x^2 - n^2 Sin(\phi))^{1/2})^2 \tag{17}$$

We can now use Eq.12 together with Eq.17 to obtain the clean-speech power pdf directly from a given the noise power pdf $p_{N^2}(n^2)$.

First case: assume $\phi = 0$, so that $x = s + n$. In this case we have

$$s = A(n) = x - n \Rightarrow p_S(s) = p_N(B(s))B'(s) = p_N(x-s) \tag{18}$$

Then we have $y = A(s) = C(s)$, $z = C(x)$, $x = B(z)$

$$\Rightarrow \quad p_Y(y) = p_S(B(y))B'(y) = p_N(x - B(y))B'(y) = p_N(B(z) - B(y))B'(y) \tag{19}$$

Two commonly used compression functions are $z = \log(x)$ and $z = x^{1/3}$.

$$p_X(x) = p_N(B(z) - B(x))B'(x) \text{ for } x \in [C(0), z] \text{, else } p_X(x) = 0$$

When $p_N(n) = U(n;0,B(z))$ and $B(z) = z^3$,

then $p_X(x) = 3x^2/z^3$ for $x \in [0, z]$

For given noise pdf $p_N(n)$, and compression function C, with inverse B, the pdf for clean compressed data is in general given by

$$p_X(x) = p_N(B(z) - B(x))B'(x) \text{ for } x \in [C(0), z] \tag{20}$$

e.g. for estimated noise pdf uniform, with cube root compression, the clean compressed data pdf is

$$p_X(x) = 3x^2/z^3 \tag{21}$$

5. Discussion

We have described the steps by which the original paradigm of recognition with missing spectral data led to a more accurate model for recognition with uncertain data. On the understanding that it is not difficult to obtain an accurate estimate of the noise pdf for each spectral data coefficient, we then described a number of ways in which the clean-data pdf can be modelled in terms of the noise pdf. The first and simplest of these models extends the 2-mixture pdf previously used in the "soft missing data" model, to a 3-mixture pdf. The second more ambitious model aims to increase the accuracy of the clean-data pdf model still further by taking into account the speech/noise phase difference. While the framework for the necessary calculations was introduced, this calculation was only implemented so far for the relatively trivial case in which $\phi = 0$. It would be instructive in future to follow through the necessary calculations for a number of further cases, according to whether

1. noise power is a given exact value or a pdf
2. logarithm or cube-root compression is in use
3. phase difference is zero, a fixed known value, or has a uniform pdf over $[-\pi, \pi]$

In the most general case, the noise power will be represented by a pdf and the phase difference will be unknown (so having a uniform pdf). It has been verified that in this case there is a closed form solution for the case of cube-root compression, but not for log compression.

The take up of this approach will depend on the existence of closed form and computationally feasible solutions to the equations here presented. This is a preliminary study and no empirical tests are included. It is intended as a theoretical foundation from which practical solutions may be developed in future.

Appendix A: Nomenclature and commonly used abbreviations

ASR	Automatic Speech Recognition
SNR	Signal to Noise Ratio
WER	Word Error Rate
$P(x)$	probability of "event x" occurring
$p(x)$	probability density at x of a continuous value
$P(x;\Theta)$	function with given model M (not indicated) with parameters Θ used to estimate $P(x)$
$\hat{P}(x)$	function (model and parameters unspecified) used to estimate $P(x)$
q_k	speech unit whose presence is being estimated, or event that data x is from this class
x, x^n	data window vector at time step n
d	number of spectral sub-bands
x clean	x is from data population used in model training (i.e. x has no data mismatch)
κ_x	any knowledge which can be used to fix or constrain the value of x
$\Phi(x)$	cumulative density function for the standard Gaussian pdf
$\delta(x-a)$	the Dirac delta pdf, centred at (a). $\delta(x) = 0$ for $x \neq 0$, and integrates to 1
$U(x;a, b)$	the Uniform pdf over the interval $[a, b]$
s	unknown uncompressed clean speech
$x = C(s)$	unknown compressed clean speech
$y = s + n$	uncompressed noisy speech, $n \in [0, y]$
$z = C(y)$	observed compressed noisy speech $(y = C^{-1}(z) = B(z))$
$p_N(n)$	given (estimated) noise pdf
$p_X(x)$	required compressed clean data pdf

Appendix B: Results used

If y is a 1-1 increasing function $A(x)$ of variable x, with $B(x) = A^{-1}(x)$, and $x \sim p_X(x)$, then[1]

$$p_Y(y) = p_X(B(y))B'(y) \tag{1}$$

If y is a 1-1 decreasing function of x, then

$$p_Y(y) = -p_X B(y) B'(y) \tag{2}$$

so in general, if y is a 1-1 function of x, then

$$p_Y(y) = -p_X B(y) |B'(y)| \tag{3}$$

If $A(x)$ is increasing then $B(y)$ is increasing, so $x \in [x_1, x_2]$ => $B(y) \in [x_1, x_2]$

$$\Rightarrow y \in [A(x_1), A(x_2)] \tag{4}$$

If $A(x)$ is decreasing then $B(y)$ is decreasing, so $x \in [x_1, x_2]$ => $B(y) \in [x_1, x_2]$

$$\Rightarrow y \in [A(x_2), A(x_1)]. \tag{5}$$

Equation relating (x, s, n)

$$x^2 = s^2 + n^2 + 2ns\cos\theta \tag{6}$$

Solving (6) for s, with $x^2 = a^2n^2$, s real and $x = s + n$ when $\theta = 0$, gives $a^2 \geq 1$ and

$$s = -nCos(\theta) + n(a^2 - Sin^2(\theta))^{1/2} \tag{7}$$

1. *This can be shown easily as follows.* $F_Y(y) = P(Y < y)$. *If* $y = A(x)$ *is increasing, then applying* B *(the inverse of* A *) to both sides of the inequality gives* $F_Y(y) = P(X < B(y)) = F_X(B(y))$. *Differentiating both sides with respect to* y *then gives* $f_Y(y) = f_X(B(y))|B'(y)|$.

References

[1] Ahmed, S. & Tresp, V. (1993) "Some solutions to the missing feature problem in vision", in Advances in Neural Information Processing Systems 5, Morgan Kauffman, San Mateo, pp. 393-400.

[2] Barker, J., Josifovski, L., Cooke, M.P. & Green, P.D. (2000) "Soft decisions in missing data techniques for robust automatic speech recognition", Proc. ICSLP-2000, pp.373-376.

[3] Droppo, J., Acero, A. & Deng, L. (2002) "Uncertainty decoding with SPLICE for noise robust speech recognition", Proc. ICASSP'02, Vol.1, pp. 57-60.

[4] El-Maliki, M. (2000) "Speaker verification with missing features in noisy environments", PhD thesis, Dept. d'Electricité, Ecole Polytechnique Fédérate de Lausanne.

[5] Evans, M., Hastings, N. & Peacock, B. (2000) **Statistical Distributions**, Wiley series in probability and statistics.

[6] Green, P.D., Cooke, M.P. & Crawford, M.D. (1995), "Auditory scene analysis and HMM recognition of speech in noise", Proc. ICASSP'95, pp.401-404.

[7] Hirsch, H. G. and C. Ehrlicher (1995) "Noise estimation techniques for robust speech recognition", ICASSP95 , 1995, pp. 153-156.

[8] Kristjansson, T.T. & Frey, B.J. (2002) "Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition", Proc. ICASSP'02, Vol.1, pp.61-64.

[9] Lippmann, R. P. & Carlson, B. A. (1997) "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", Proc. Eurospeech'97, pp. 37-40

[10] McCowan, I.A., Morris, A.C. & Bourlard, B. (2002) "Improving speech recognition performance of small microsphone arrays using missing data techniques", Proc ICSLP'02 (in press).

[11] Morris, A. C. (2001) "Data utility modelling for mismatch reduction", Proc. CRAC workshop on Consistent & Reliable Acoustic Cues for sound analysis.

[12] Morris, A. C., Cooke, M. & Green, P. (1998) "Some solutions to the missing feature problem in data classification, with application to noise robust ASR", Proc. ICASSP'98, pp.737-740.

[13] Vasegi, S. (1998) Advanced signal processing and digital noise reduction", Wily Neubner.