IDIAP RESEARCH REPORT

INTUITIVE RECIPES FOR UNCERTAINTY DECODING WITH SNR FEATURES FOR NOISE ROBUST ASR

Georgios Skoumas Philip N. Garner

Idiap-RR-23-2011

JULY 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Intuitive Recipes for Uncertainty Decoding with SNR Features for Noise Robust ASR

Georgios Skoumas and Philip N. Garner

Idiap Research Institute, Martigny Switzerland

July 12, 2011

Abstract

In this work, uncertainty decoding in automatic speech recognition is investigated in the context of robustness to additive and convolutional noise. In Garner (2009) and in Garner (2011), explicit calculation of a Signal to Noise Ratio (SNR) cepstrum by means of a noise estimate is shown to have theoritical and practiacal adavantages over the usual energy based spectrum, especially when they are combined with traditional feature enhancement techniques, such as Cepstral Mean Normalization (CMN) and Cestral Variance Normalization (CVN). This is the reason we try to investigate how these features behave when we are trying to handle the uncertainty which is introduced by environmental or convolutional noise. We compute uncertainty, which can be translated as a variance estimate, in the normal SNR front end and pass this to the decoder. Unfortunatelly, this approximation suffers from theoritical and practical problems, epsecially in low SNR conditions. The issue described and the performance of the uncertainty decoding schemes we used, are examined with the aurora 2 digit recognition task.

Introduction

Speech recognition in noise has been an area of active research for many years. Powerful model based compensation schemes, such as Parallel Model Combination (PMC) in Gales (1996), and Vector Taylor Series (VTS), achieve good performance but are computationally expensive. Recently interest has grown in a compromise between model-based and front-end schemes. Uncertainty decoding, so called because a measure of the uncertainty introduced by the background acoustic noise is propagated into the recognition process. For front-end uncertainty schemes, this uncertainty is computed mainly from the features. Despite front-end uncertainty decoding achieving good performance for a range of acoustic environments, a fundamental problem arises in our case. By passing a single uncertainty-variance value to the decoder per frame, when the SNR is low, can cause all the model variances to move to higher values. When this occurs, the recogniser can no longer discriminate in these areas. This technical report examines if SNR features can handle this fundamental issue in front-end uncertainty decoding. Experiments show that this scheme is not efficient yet, especially in noisy conditions, but it could be a good start point for further research on uncertainty decoding as it produces better accuracy in same clean condition cases.

Marginalization over Variance

As this report is a continuation of the work that is introduced in Garner (2009) and Garner (2011) we suggest that the reader should study these works first. In the *"Marginalization over Variance"* point of the aforementioned papers, it is shown that if we assume that an estimate, $\hat{\mathbf{v}}$ of the noise variance is always available, the form of equation (11) of Garner (2009) or (13) of Garner (2011), however, with multiplicative instead of additive terms in the denominators, allows marginalization over the noise variance.

If we assume that we have N frames of noise, $\mathbf{n}_{\mathrm{N}} = \{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_{\mathrm{N}}\}$, that are observed in isolation, we can write that:

$$p(\nu_f \mid \{\mathbf{n}\}_N) = \frac{\prod_{i=1}^{N} p(\mathbf{n}_{i,f}\nu_f)(\nu_f)}{\int_0^\infty \prod_{i=1}^{N} p(\mathbf{n}_{i,f}|\nu'_f)p(\nu'_f)d\nu} \tag{1}$$

where the products are over the likelihood terms, not the priors. Again, hereafter we drop subscritps for simplicity. The likelihood terms are exactly the form of equation (5) of Garner (2009) or (6) of Garner (2011), and we arbitrarily choose an non-informative prior $p(\nu) \propto \nu^{-1}$. Equation (1) can then be reduced to the inverse gamma distribution

$$p(\nu \mid \{\mathbf{n}\}_N) = \frac{B^A}{\Gamma(A)}\nu^{-A-1}\exp\left(-\frac{B}{\nu}\right) \tag{2}$$

where

$$A = N, \quad B = \sum_{i=1}^{N} |\mathbf{n}_{i,f}|^2 \tag{3}$$

The MAP solution, $\hat{\nu}$ of ν would be

$$\hat{\nu} = \frac{B}{A+1} \tag{4}$$

However, we can use the distibution to marginalise over ν. So, the posterior becomes in terms of ξ as follows.

$$p(\xi \mid t) \propto p(\xi)\int_0^\infty p(t|\xi,\nu)(\nu|\{n\}_N)d\nu \tag{5}$$

By substututing equation (11) from Garner (2009) or (13) from Garner (2011) and equation (2) into (5), the forms are conjugate and the integral is just the normalizing term from the inverse gamma distribution

$$p(\xi \mid t) \propto p(\xi) \times \frac{B^A}{\Gamma(A)}\frac{\Gamma(A+1)}{\xi+1}\left(\frac{|t|^2+(\xi+1)B}{\xi+1}\right)^{-(A+1)} \tag{6}$$

This distribution can be further simplified as follows

$$p(\xi \mid t) \propto p(\xi) \times \frac{B^A}{\Gamma(A)}\frac{\Gamma(A+1)}{\xi+1}\left(\frac{|t|^2+(\xi+1)B}{\xi+1}\right)^{-(A+1)} \tag{7}$$

$$\propto p(\xi) \times \frac{AB^A}{\xi+1}\left(\frac{|t|^2+(\xi+1)B}{\xi+1}\right)^{-(A+1)} \tag{8}$$

If we assume a conjugate prior $f(\xi) = \frac{1}{(\xi+1)^\gamma}$ the probability density function can be written as follows

$$p(\xi \mid t) \propto \frac{AB^A}{(\xi+1)^{\gamma+1}}\left(\frac{|t|^2+(\xi+1)B}{\xi+1}\right)^{-(A+1)} \tag{9}$$

Which can be simplified as

$$p(\xi \mid t) \propto AB^A\frac{1}{(\xi+1)^{\gamma-A}}\left(\frac{|t|^2+(\xi+1)B}{\xi+1}\right)^{-(A+1)} \tag{10}$$

$$\propto AB^A(\xi+1)^{A-\gamma}\left(|t|^2+(\xi+1)B\right)^{-(A+1)} \tag{11}$$

$$\propto \frac{A}{B}(\xi+1)^{A-\gamma}\left(\frac{|t|^2}{B}+\xi+1\right)^{-(A+1)} \tag{12}$$

By substituting $\xi + 1 = y \Leftrightarrow \xi = y - 1$ and by differentiating we have that $d\xi = dy$. So, equation (12) can be written as

$$p(y \mid t) \propto \frac{A}{B} y^{A-\gamma} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)} \tag{13}$$

evaluating the normalising constant this may be written as

$$p(y \mid t) = \frac{y^{A-\gamma} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)}}{\int_1^\infty y^{A-\gamma} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)}} \tag{14}$$

$$= \frac{\gamma}{{}_2F_1(A+1, \gamma; \gamma+1; -\frac{|t|^2}{B})} y^{A-\gamma} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)} \tag{15}$$

where ${}_2F_1(\cdot)$ is a Gauss hypergeometric function and $\gamma > 0$, $\frac{|t|^2}{B} > 0$.

A Brute Force Solution

Now we can estimate the mode and the first and second moments of the $\xi+1$ probability density function. For the mode estimation we will ignore any term that does not depend on variable y.

$$\mu = \frac{d}{dy} \left(y^{A-\gamma} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)} \right) = 0 \tag{16}$$

$$\mu = \max \left(\frac{A-\gamma}{\gamma+1} \times \frac{|t|^2}{B}, 1 \right) \tag{17}$$

$$\mu^1_{str} = \frac{\gamma}{{}_2F_1(A+1, \gamma; \gamma+1; -\frac{|t|^2}{B})} \int_1^\infty y y^{A-\gamma} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)} \tag{18}$$

$$= \frac{\gamma}{{}_2F_1(A+1, \gamma; \gamma+1; -\frac{|t|^2}{B})} \int_1^\infty y^{A-\gamma+1} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)} \tag{19}$$

$$= \frac{\gamma}{\gamma-1} \times \frac{{}_2F_1(A+1, \gamma-1; \gamma; -\frac{|t|^2}{B})}{{}_2F_1(A+1, \gamma; \gamma+1; -\frac{|t|^2}{B})} \tag{20}$$

$$\mu^2_{str} = \frac{\gamma}{{}_2F_1(A+1, \gamma; \gamma+1; -\frac{|t|^2}{B})} \int_1^\infty y^2 y^{A-\gamma} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)} \tag{21}$$

$$= \frac{\gamma}{{}_2F_1(A+1, \gamma; \gamma+1; -\frac{|t|^2}{B})} \int_1^\infty y^{A-\gamma+2} \left(\frac{|t|^2}{B} + y \right)^{-(A+1)} \tag{22}$$

$$= \frac{\gamma}{\gamma-2} \times \frac{{}_2F_1(A+1, \gamma-2; \gamma-1; -\frac{|t|^2}{B})}{{}_2F_1(A+1, \gamma; \gamma+1; -\frac{|t|^2}{B})} \tag{23}$$

Here, we will assume that the SNR $(\xi + 1)$ is log-normally distributed in the Linear Spectral domain. So, we may equate the moments of the SNR distribution with the moments of a Log-Normal Distribution. The first two moments of the SNR distribution can be estimated as follows.

$$E[\xi + 1] = E[Y] \tag{24}$$

$$E[(\xi+1)^2] = E[Y^2] \tag{25}$$

By equating with the moments of a Log Normal distribution we have that

$$E^{ln}[Y] = E[Y] \tag{26}$$

$$e^{\mu_{str}+\frac{\sigma^2_{str}}{2}} = \frac{\gamma}{\gamma-1} \times \frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})} \tag{27}$$

$$\mu_{str} + \frac{\sigma^2_{str}}{2} = \log\left(\frac{\gamma}{\gamma-1} \times \frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right) \tag{28}$$

$$\sigma^2_{str} = 2\log\left(\frac{\gamma}{\gamma-1} \times \frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right) - 2\mu_{str} \tag{29}$$

$$E^{ln}[Y^2] = E[Y^2] \tag{30}$$

$$e^{2\mu_{str}+2\sigma^2_{str}} = \frac{\gamma}{\gamma-2} \times \frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})} \tag{31}$$

$$2\mu_{str} + 2\sigma^2_{str} = \log\left(\frac{\gamma}{\gamma-2} \times \frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right) \tag{32}$$

$$\mu_{str} = \frac{1}{2}\log\left(\frac{\gamma}{\gamma-2} \times \frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right) - \sigma^2_{str} \tag{33}$$

Substituting (33) into (29) we have that

$$\begin{aligned}\mu_{str} &= 2\log\left(\frac{\gamma}{\gamma-1} \times \frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right) \\ &- \frac{1}{2}\log\left(\frac{\gamma}{\gamma-2} \times \frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right)\end{aligned} \tag{34}$$

$$\begin{aligned}\sigma^2_{str} &= -2\log\left(\frac{\gamma}{\gamma-1} \times \frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right) \\ &+ \log\left(\frac{\gamma}{\gamma-2} \times \frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t|^2}{B})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{B})}\right)\end{aligned} \tag{35}$$

where μ_{str} and $\sigma^2 str$ are the mean and variance of a Gaussian distribution in the Log Spectral domain for the brute force case.

A Potentially-Smarter Approximation

Starting again from equation (13) we will assume that for SNR values $0 < \xi < 1$ tha probability density function is flat. Taking this assumption we may change the integral limits from 0 to 1. So, starting from equation (13) and evaluating again the normalization constant we have that

$$p(y|t) = \frac{y^{A-\gamma}\left(\frac{|t|^2}{B}+y\right)^{-(A+1)}}{\int_0^\infty y^{A-\gamma}\left(\frac{|t|^2}{B}+y\right)^{-(A+1)}} \tag{36}$$

$$= \frac{\Gamma(A+1)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\left(\frac{|t|^2}{B}\right)^{\gamma} y^{A-\gamma}\left(\frac{|t|^2}{B}+y\right)^{-(A+1)} \tag{37}$$

The first two moments of the above distibution can be estimated as follows [1]

$$\mu^1_{apr} = \frac{\Gamma(A+1)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\left(\frac{|t|^2}{B}\right)^{\gamma}\int_0^\infty y y^{A-\gamma}\left(\frac{|t|^2}{B}+y\right)^{-(A+1)} \tag{38}$$

$$= \frac{\Gamma(A+1)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\left(\frac{|t|^2}{B}\right)^{\gamma}\int_0^\infty y^{A-\gamma+1}\left(\frac{|t|^2}{B}+y\right)^{-(A+1)} \tag{39}$$

$$= \frac{\Gamma(A+1)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\frac{\Gamma(\gamma-1)\Gamma(A-\gamma+2)}{\Gamma(A+1)}\left(\frac{|t|^2}{B}\right)^{\gamma}\left(\frac{|t|^2}{B}\right)^{1-\gamma} \tag{40}$$

$$= \frac{\Gamma(\gamma-1)\Gamma(A-\gamma+2)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\frac{|t|^2}{B} \tag{41}$$

$$= \frac{A-\gamma+1}{\gamma-1}\frac{|t|^2}{B} \tag{42}$$

$$\mu^2_{apr} = \frac{\Gamma(A+1)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\left(\frac{|t|^2}{B}\right)^{\gamma}\int_0^\infty y^2 y^{A-\gamma}\left(\frac{|t|^2}{B}+y\right)^{-(A+1)} \tag{43}$$

$$= \frac{\Gamma(A+1)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\left(\frac{|t|^2}{B}\right)^{\gamma}\int_0^\infty y^{A-\gamma+2}\left(\frac{|t|^2}{B}+y\right)^{-(A+1)} \tag{44}$$

$$= \frac{\Gamma(A+1)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\frac{\Gamma(\gamma-2)\Gamma(A-\gamma+3)}{\Gamma(A+1)}\left(\frac{|t|^2}{B}\right)^{\gamma}\left(\frac{|t|^2}{B}\right)^{2-\gamma} \tag{45}$$

$$= \frac{\Gamma(\gamma-2)\Gamma(A-\gamma+3)}{\Gamma(\gamma)\Gamma(A-\gamma+1)}\left(\frac{|t|^2}{B}\right)^{2} \tag{46}$$

$$= \frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{|t|^2}{B}\right)^{2} \tag{47}$$

By equating again the moments of $\xi+1$ distribution with the moments of a Log Normal one we have that

$$E^{ln}[Y] = E[Y] \tag{48}$$

$$e^{\mu_{apr}+\frac{\sigma^2_{apr}}{2}} = \frac{A-\gamma+1}{\gamma-1}\frac{|t|^2}{B} \tag{49}$$

$$\mu_{apr} + \frac{\sigma^2_{apr}}{2} = \log\left(\frac{A-\gamma+1}{\gamma-1}\frac{|t|^2}{B}\right) \tag{50}$$

$$\sigma^2_{apr} = 2\log\left(\frac{A-\gamma+1}{\gamma-1}\frac{|t|^2}{B}\right) - 2\mu_{apr} \tag{51}$$

[1] The mode here is the same as in the previous case.

$$E^{ln}[Y^2] = E[Y^2] \tag{52}$$

$$e^{2\mu_{apr}+2\sigma^2_{apr}} = \frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{|t|^2}{B}\right)^2 \tag{53}$$

$$2\mu_{apr}+2\sigma^2_{apr} = \log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{|t|^2}{B}\right)^2\right) \tag{54}$$

$$\mu_{apr} = \frac{1}{2}\log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{|t|^2}{B}\right)^2\right) - \sigma^2_{apr} \tag{55}$$

Substituting (55) into (51) we have that

$$\mu_{apr} = 2\log\left(\frac{A-\gamma+1}{\gamma-1}\frac{|t|^2}{B}\right) - \frac{1}{2}\log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{|t|^2}{B}\right)^2\right) \tag{56}$$

$$\sigma^2_{apr} = -2\log\left(\frac{A-\gamma+1}{\gamma-1}\frac{|t|^2}{B}\right) + \log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{|t|^2}{B}\right)^2\right) \tag{57}$$

Here, we will introduce a noise mean vector which can be estimated as follows

$$B = \sum_{i=1}^{N} |\mathbf{n}_{i,f}|^2 \tag{58}$$

$$\frac{B}{A} = \frac{\sum_{i=1}^{N} |\mathbf{n}_{i,f}|^2}{A} = \bar{x} \tag{59}$$

From now on we will express B in terms of A and noise mean $\bar{x}$ like $B = A\bar{x}$. Following this formula, we can rewrite equations (34), (35), (56) and (57)

$$\begin{aligned}\mu_{str} &= 2\log\left(\frac{\gamma}{\gamma-1}\times\frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}\right)\\ &- \frac{1}{2}\log\left(\frac{\gamma}{\gamma-2}\times\frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}\right)\end{aligned} \tag{60}$$

$$\begin{aligned}\sigma^2_{str} &= -2\log\left(\frac{\gamma}{\gamma-1}\times\frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}\right)\\ &+ \log\left(\frac{\gamma}{\gamma-2}\times\frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t|^2}{\bar{x}}\frac{1}{A})}\right)\end{aligned} \tag{61}$$

$$\mu_{apr} = 2\log\left(\frac{A-\gamma+1}{\gamma-1}\frac{1}{A}\frac{|t|^2}{\bar{x}}\right) - \frac{1}{2}\log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{1}{A}\frac{|t|^2}{\bar{x}}\right)^2\right) \tag{62}$$

$$\sigma^2_{apr} = -2\log\left(\frac{A-\gamma+1}{\gamma-1}\frac{1}{A}\frac{|t|^2}{\bar{x}}\right) + \log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{(\gamma-1)(\gamma-2)}\left(\frac{1}{A}\frac{|t|^2}{\bar{x}}\right)^2\right) \tag{63}$$

The simulation shows that the assumption of the distibution of $y = \xi + 1$ being flat for $0 < \xi < 1$ is really close to hypergeometric brute force solution even for low raw SNR values. This is summarized in figure 1.

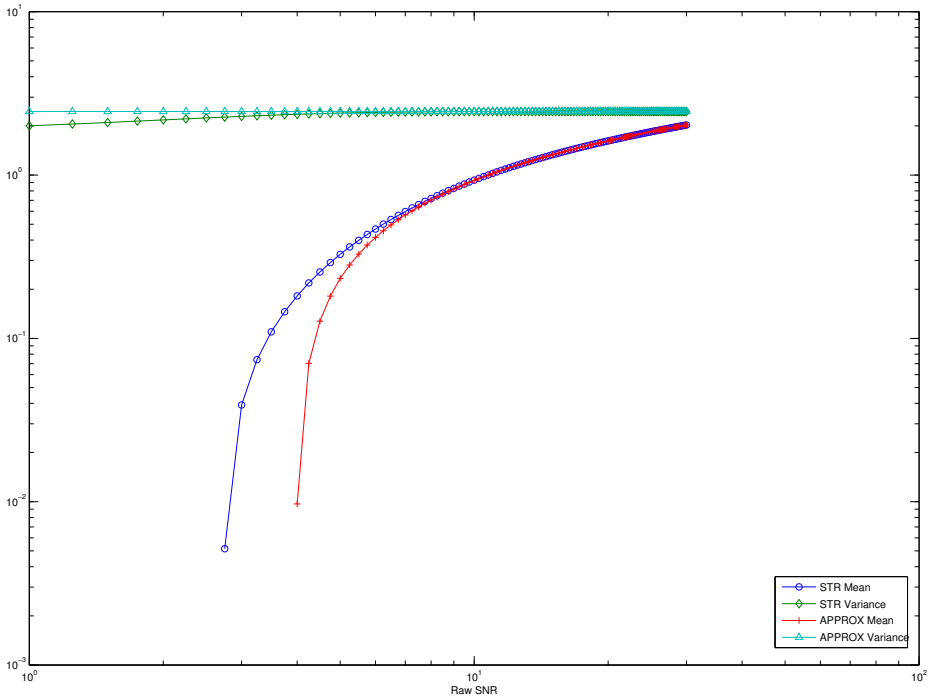


Figure 1: Brute Force Solution vs Approximation

Cepstral Domain and Decoding

As we mentioned before μ and σ^2 are the mean and variance in the Log Spectral space. In terms of vectors and matrices equations (60), (61), (62) and (63) can be written as follows

$$\begin{aligned}\mu^{i}_{str} &= 2\log\left(\frac{\gamma}{\gamma-1}\times\frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}\right)\\ &- \frac{1}{2}\log\left(\frac{\gamma}{\gamma-2}\times\frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}\right)\end{aligned} \tag{64}$$

$$\begin{aligned}\Sigma^{ii}_{str} &= -2\log\left(\frac{\gamma}{\gamma-1}\times\frac{{}_2F_1(A+1,\gamma-1;\gamma;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}\right)\\ &+ \log\left(\frac{\gamma}{\gamma-2}\times\frac{{}_2F_1(A+1,\gamma-2;\gamma-1;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}{{}_2F_1(A+1,\gamma;\gamma+1;-\frac{|t_i|^2}{\bar{x}}\frac{1}{A})}\right)\end{aligned} \tag{65}$$

$$\mu^{i}_{apr} = 2\log\left(\frac{A-\gamma+1}{A(\gamma-1)}\frac{|t_i|^2}{\overline{x}}\right) - \frac{1}{2}\log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{A^2(\gamma-1)(\gamma-2)}\left(\frac{|t_i|^2}{\overline{x}}\right)^2\right) \tag{66}$$

$$\Sigma^{ii}_{apr} = -2\log\left(\frac{A-\gamma+1}{A(\gamma-1)}\frac{|t_i|^2}{\overline{x}}\right) - \log\left(\frac{(A-\gamma+1)(A-\gamma+2)}{A^2(\gamma-1)(\gamma-2)}\left(\frac{|t_i|^2}{\overline{x}}\right)^2\right) \tag{67}$$

Here the mean vectors and diagonal covariance matrices are in the log spectral domain. Moving to the cepstral domain we will use a descrite cosine transform (DCT). The DCT will be represented as a matrix C. In a general case of mean vectors μ^l, μ^c and covariance matrices Σ^l, Σ^c in the log spectral and cepstral domains respectively, we can write as in Gales (1996) that

$$\mu^c = C\mu^l \tag{68}$$

$$\Sigma^c = C\Sigma^l C^T \tag{69}$$

Mapping from the Log spectral to the cepstral domain will lead to full covariance matrices. Using delta and delta-delta (acceleration) parameters the mean vectors and covariance matices will have the following form in the cepstral domain.

$$\mu^c = \begin{bmatrix}(C\mu^l)^T & (C\Delta\mu^l)^T & (C\Delta^2\mu^l)^T\end{bmatrix}^T \tag{70}$$

where $\Delta\mu^l$ and $\Delta^2\mu^l$ are the delta and delta-delta mean vectors in the log spectral domain.

For the covariance matrix we have that

$$\Sigma^c = \begin{bmatrix} C\Sigma^l C^T & C\delta^{sd}\Sigma^l C^T & C\delta^{s-dd}\Sigma^l C^T \\ C(\delta^{sd}\Sigma^l)^T C^T & C\Delta\Sigma^l C^T & C\delta^{d-dd}\Sigma^l C^T \\ C(\delta^{s-dd}\Sigma^l)^T C^T & C(\delta^{d-dd}\Sigma^l)^T C^T & C\Delta^2\Sigma^l C^T \end{bmatrix} \tag{71}$$

where $\Delta\Sigma^l$ and $\Delta^2\Sigma^l$ are the covariance matrices of the delta and delta-delta parameters and $\delta^{sd}\Sigma^l$, $\delta^{d-dd}\Sigma^l$ and $\delta^{s-dd}\Sigma^l$ are the covariance matrices which represent the correlation between static and delta, delta and accelaration, and static and accelaration parameters respectively. In our case we assume diagonal covariance matrices which means that the cross correlation terms between static, delta and delta-delta parameters are considered equal to zero. We keep only the the diagonals of the full static, delta and delta-delta covariance matrices. Therefore, the final covariance matrices in the cepstral domain will have the following form

$$\Sigma^c = \begin{bmatrix} \sigma^2_{s_{11}} & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & \dots & \dots & \dots & \dots & \vdots \\ \vdots & 0 & \sigma^2_{s_{nn}} & 0 & \dots & \dots & \dots & \dots & \vdots \\ \vdots & \dots & 0 & \sigma^2_{d_{11}} & 0 & \dots & \dots & \dots & \vdots \\ \vdots & \dots & \dots & 0 & \ddots & 0 & \dots & \dots & \vdots \\ \vdots & \dots & \dots & \dots & 0 & \sigma^2_{d_{nn}} & 0 & \dots & \vdots \\ \vdots & \dots & \dots & \dots & \dots & 0 & \sigma^2_{dd_{11}} & 0 & \vdots \\ \vdots & \dots & \dots & \dots & \dots & \dots & 0 & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 & \sigma^2_{dd_{nn}} \end{bmatrix} \tag{72}$$

where σ^2_s, σ^2_d and σ^2_{dd} are the diagonal elements of the covariance matrices of the static, delta and delta-delta parameters respectively and n the dimensionality.

For decoding noise uncertainty, the Gaussian likelihood estimation in the Viterbi decoder will be estimated as follows

$$p(\mu^c|M) = \frac{1}{\sqrt{(2\pi)^n}|\Sigma^m + \Sigma^c|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\mu^c - \mu^m)^T(\Sigma^c + \Sigma^m)^{-1}(\mu^c - \mu^m)) \quad (73)$$

where μ^m and Σ^m are the mean vector and covariance matrix of the already trained acoustic model M. This translates into that during the decoding stage we add the estimated variance features with the trained acoustic model's variance parameters.

Experiments

As we mentioned before, we evaluated our method with the aurora 2 database which is a digit recognition task. The experimental setup is the same as in Garner (2009) and Garner (2011). We tried four basic uncertainty decoding schemes. Each scheme was combined with cepstral mean normalization being applied on both observations (SNR estimates) and variances. The basic Front-Ends we used are shown in figure (2). Feature extracrion sequence with and without the use of CMN are shown in block diagrams (*a*) and (*b*) respectively. In these experiments we consider as baseline case, the case where SNR estimates plus one are passed to the decoder. In our baseline case, variance features are again calcuted with equation (61). In the second scenario, observations are replaced by the mode of the $SNR + 1$ probability density function, as it is described in the *"Brute Force Solution"* part. In this case, variance features are estimated using equation (61). In the third case, SNR estimates are replaced by calculating a mean vector using equation (60) with variance features estimated again with equation (61). In the final case scenario, we use the mode of the SNR probability density function and we calculate variance features by using equation (63). The values of A and γ, used for the estimation of the observations and variance features are shown in the following tables.

Parameter	Mode Case	Straight Solution Case	Maximum Likelihood case
A - Noise Vectors	20	20	20
γ - Prior	0.0	2.1	-

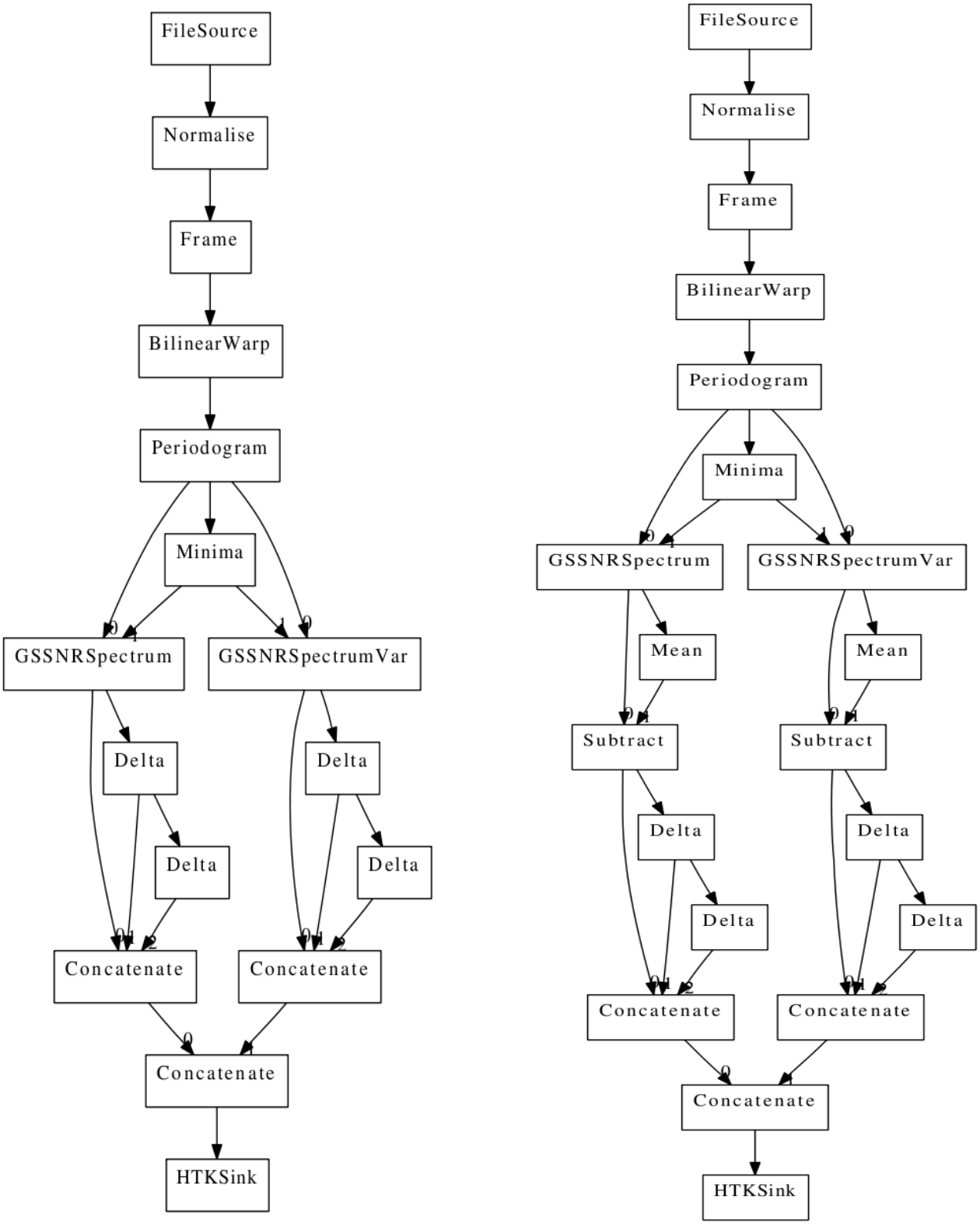
Table 1: Calculation of Observations - Simple and all Uncertainty Cases

Parameter	Mode Case	Straight Solution Case	Maximum Likelihood case
A - Noise Vectors	10	10	10
γ - Prior	2.1	2.1	2.1

Table 2: Calculation of Variance Features - Hypegeometric Case

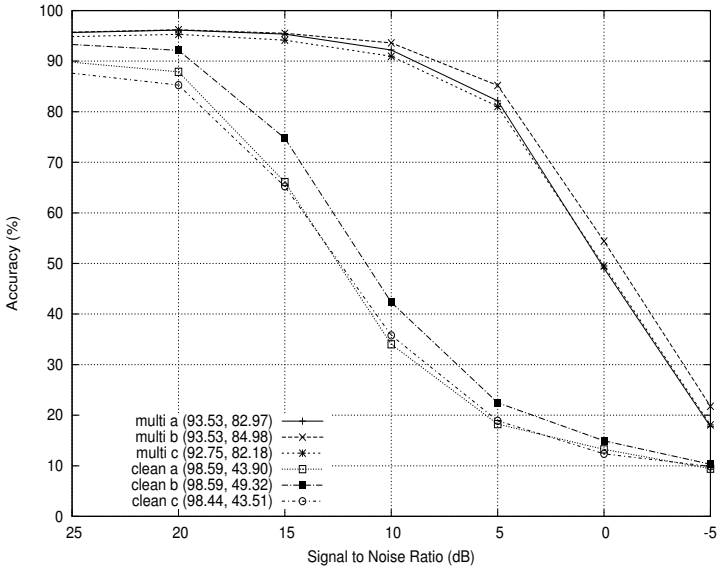
Parameter	Mode Case
A - Noise Vectors	20
γ - Prior	2.1

Table 3: Calculation of Variance Features - Assumption Case

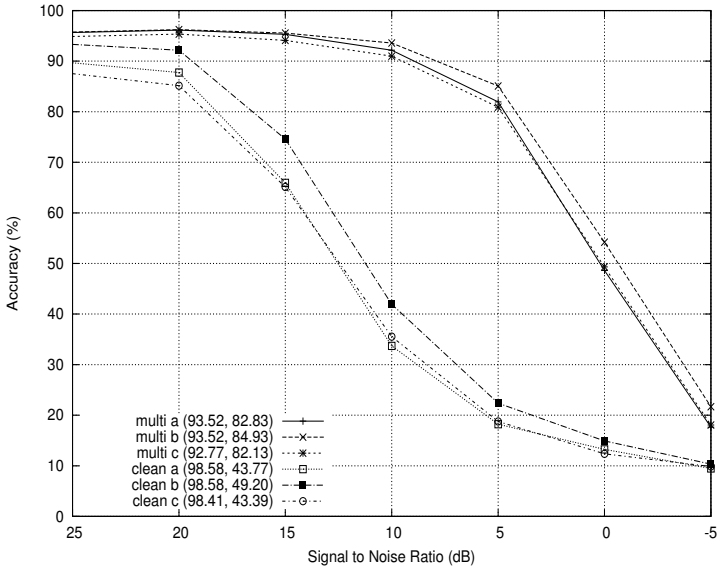


(a) (b)

Figure 2: a) Front-End without CMN b) Front-End with CMN

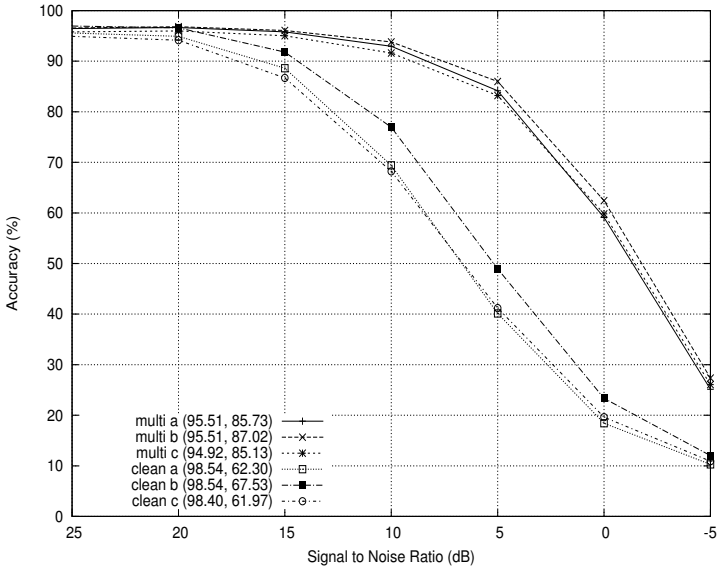


(a)

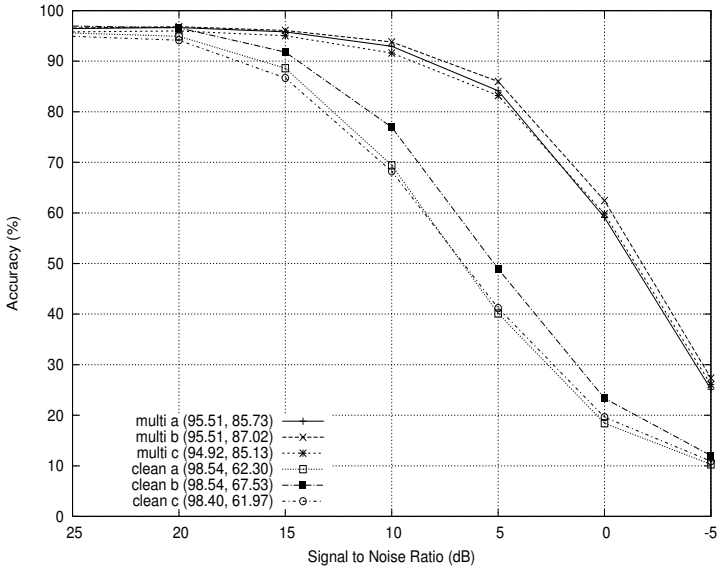


(b)

Figure 3: Baseline Maximum Likelihood Mean Case vs Maximum Likelihood Mean with Uncertainty Decoding - Hypergeometric Variance Case

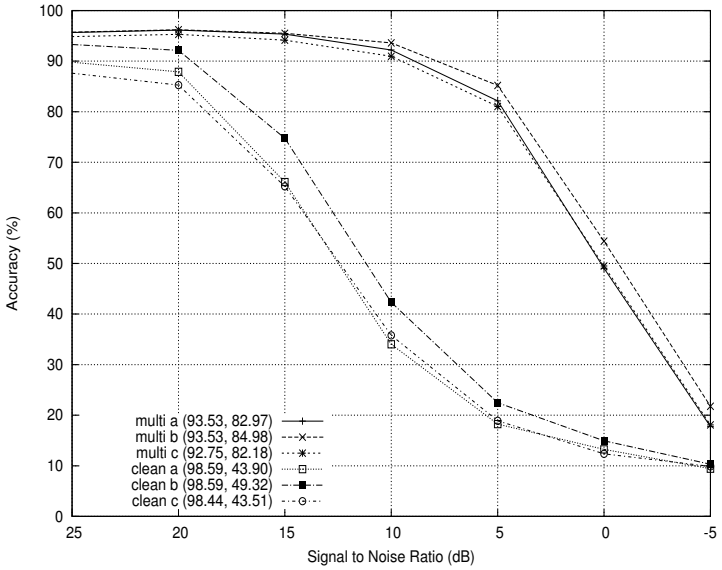


(a)

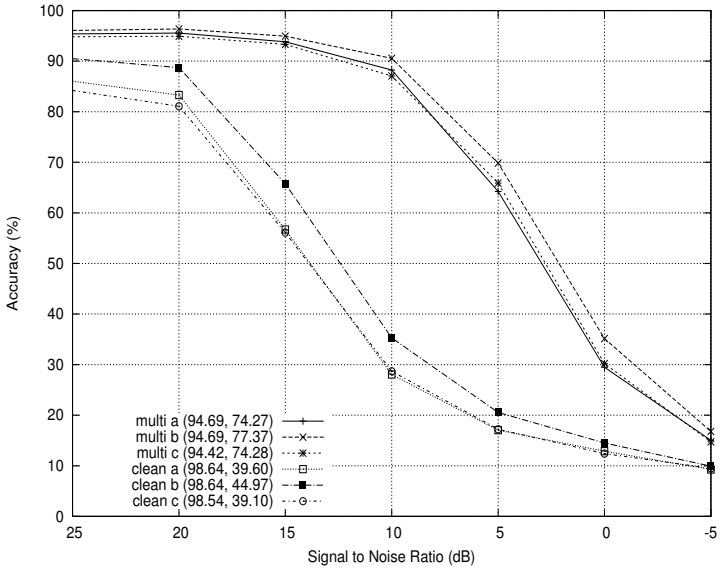


(b)

Figure 4: Maximum Likelihood Mean with CMN Case vs Maximum Likelihood Mean with CMN and Uncertainty Decoding - Hypergeometric Variance Case. Here, CMN masks the impact of uncertainty decoding in all cases. This is why the plots are identical.

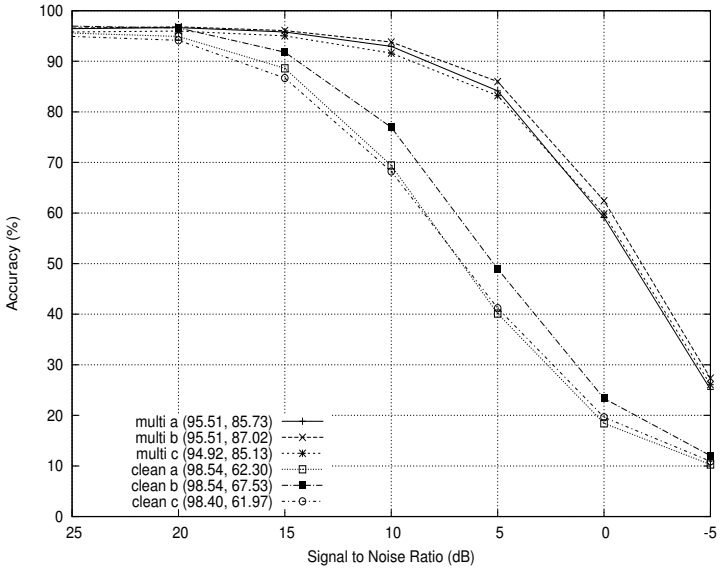


(a)

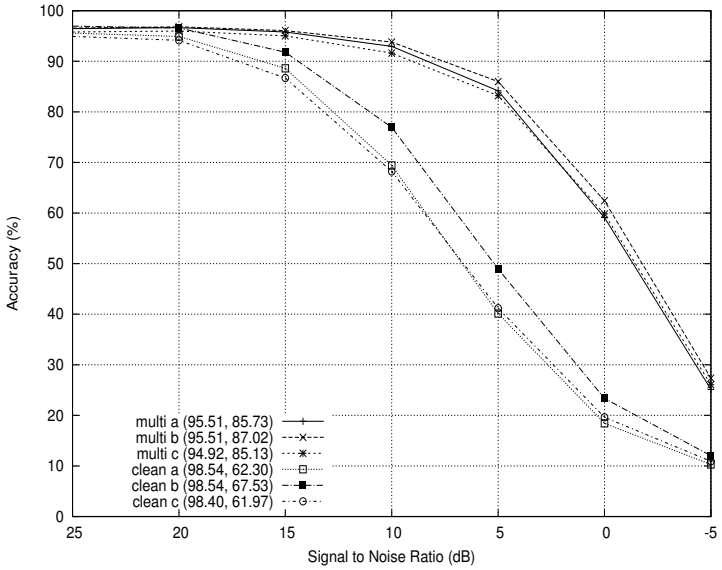


(b)

Figure 5: Mode Case vs Mode with Uncertainty Decoding - Hypergeometric Variance Case

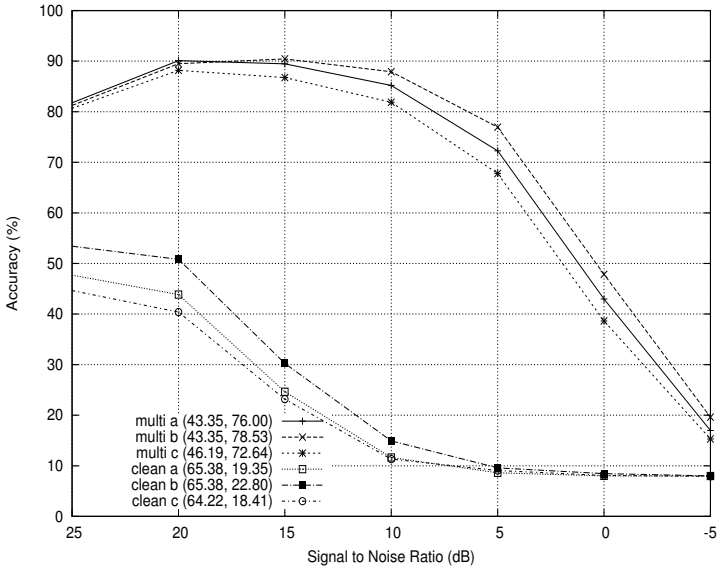


(a)

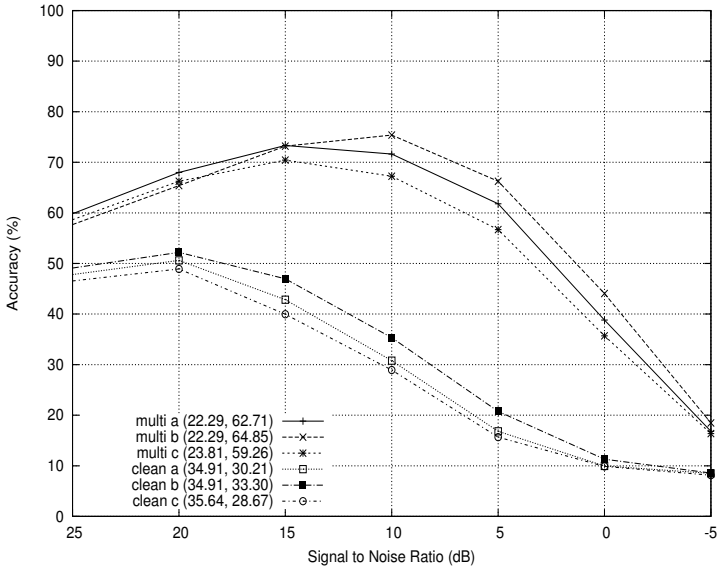


(b)

Figure 6: Mode with CMN Case vs Mode with CMN and Uncertainty Decoding - Hypergeometric Variance Case. Here, CMN masks again the impact of uncertainty decoding in all cases. This is why the plots are identical.

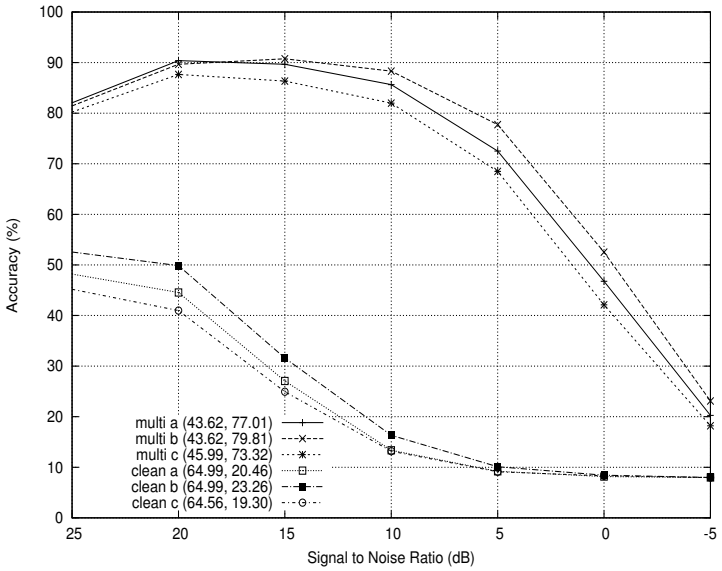


(a)

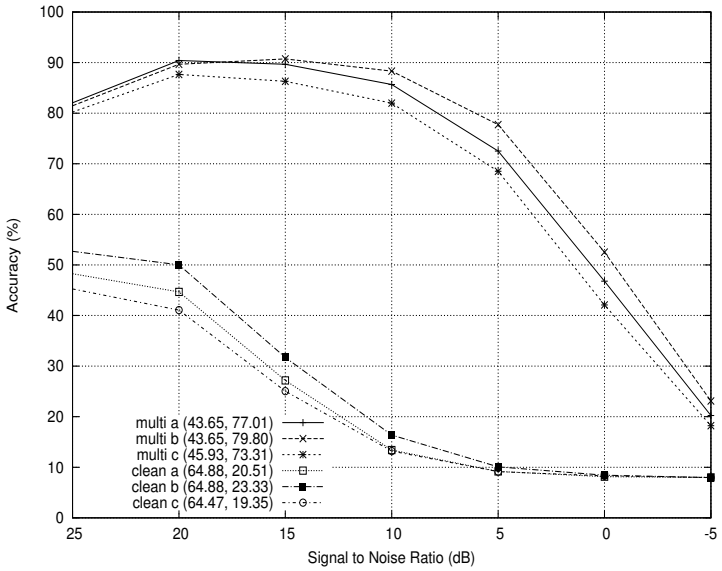


(b)

Figure 7: Hypergeometric Mean Case vs Hypergeometric Mean with Uncertainty Decoding - Hypergeometric Variance Case



(a)



(b)

Figure 8: Hypergeometric Mean with CMN Case vs Hypergeometric Mean with CMN and Uncertainty Decoding - Hypergeometric Variance Case. Here, CMN masks the impact of uncertainty decoding in almost all cases. This is why the plots are almost identical.

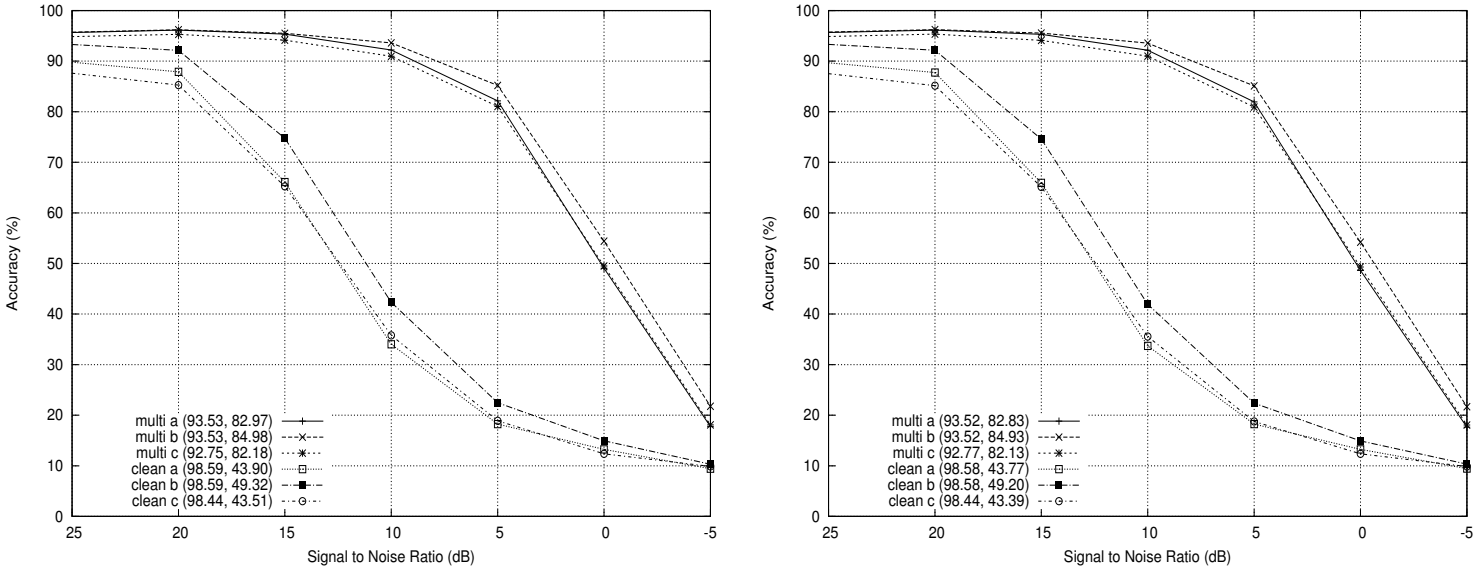


Figure 9: Mode Case vs Mode with Uncertainty Decoding - Assumption Variance Case

Conclusions

Besides the fact that SNR features are shown to work very well in noisy conditions, we don't have any evidence that variance features associated with SNR ones can increase the recognition performance in noisy environments. As it is depicted in the graphs, we have an improvement in the final case scenario, but this improvement is masked when we are using basic feature enhancement techniques such as CMN, in almost all cases. This means that the proposed method faces some theoretical and/or practical problems. The main theoritical problem could be the log-normal assumption in the spectral domain. What could be a practical problem is the fact that we use variance estimates only for testing and not for training. Overall, we believe that the proposed method has potential but it needs more investigation in the aforementioned possible causes of problems. Here, we should make a reference on Ephraim and Rahim (1999) and Ephraim and Roberts (2005), as in these papers another alternative method of estimating variance features in the cepstral domain is introduced. The reader is advised to study these papers, too.

References

Y. Ephraim and M. Rahim. On second order statistics and linear estimation of cepstral coefficients. *IEEE Transactions on Speech and Audio Processing*, 7(2):162–176, March 1999.

Y. Ephraim and W. J. Roberts. On second-order statistics of log-periodogram with correlated components. *IEEE Signal Processing Letters*, 12(9):625–628, September 2005.

M. Gales. *Model-Based Techniques for Noise Robust Speech Recognition*. PhD thesis, University of Cambridge, 1996.

P. N. Garner. SNR features for automatic speech recognition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009.

P. N. Garner. Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition. *Speech Communication*, 53(8):991–1001, October 2011.