Adapting a HMM-based Recogniser for Noisy Speech Enhanced by Spectral Subtraction

J.A. Nolazco Flores & S.J. Young

April 1993

CUED/F-INFENG/TR.123

Abstract

Training HMMs on the same conditions as in recognition makes models learn not only the features of the speech, but also those of the environment. Training in the same conditions allows the recognition system to obtain better recognition performance, but trying to have models for all possible environments is impractical. Therefore, one way to solve this problem is to compensate models trained on clean speech to give "artificially" adapted models. The goal of these noise adaptation techniques is to reach the same recognition performance as would be obtained by training in the noisy conditions. Parallel Model Combination (PMC) [5] is one adaptation technique which has been successful in adapting a clean speech model to noise by automatically generating "noisy speech models".

However, even training in noise can only achieve limited recognition performance because the high variance at low SNR makes the features begin to overlap making the discrimination problem more difficult. The problem is even worse when the vocabulary grows; for example, some experiments have shown that recognition performance is below 80% for 0 dB, see [7], even when training and testing were in the same environment. Therefore, in very noisy environments, or when the vocabulary grows, even training in noise is not enough to obtain good recognition performance. In order to improve recognition performance in very noisy environments, some sort of enhancement technique may be useful. An enhancement scheme could improve the SNR, or minimise the variance, or emphasise the main features of the interesting signal. However, all of these improvements are usually at the expense of signal distortion. Minimising both signal distortion and noise, a signal with better features and lower variability is obtained. However, if we want to exploit the good features of the noise adaptation techniques and the good features of the enhancement techniques, then we need to compensate the speech models to the distorted signal. In other words, we need to adapt the models to the enhanced signal.

In this work, we study how to adapt clean speech models for a signal enhanced by Spectral Subtraction(SS). This scheme improves the SNR but at the expense of signal distortion. Nevertheless, this scheme has been successful for signal enhancement [4] [3], and for speech recognition [10] for noisy environment. Here, the distorted signal is compensated to make SS able to deal with very noisy environments. It will be shown that the signal distortion can be represented in the linear domain by a correction term, Δ. PMC transforms the noise and speech model parameters from the cepstral domain to the linear domain, adds these parameters, and then creates an adapted model by returning to the cepstral domain. Therefore, PMC can be modified to compensate an SS distorted signal in the linear domain by including the correction term, Δ. This modified version of PMC will be called the SS-PMC method.

The results obtained by the SS-PMC technique are very encouraging, showing that it is very effective to use adaptation techniques to compensate for the signal distortion which is a side effect of an SS-based enhancement scheme.

Contents

1 Introduction

The practical application of speech recognition in the real world must deal with the serious problem of environmental noise. This is why a lot of research effort has been given to this subject. Many approaches attempt to clean the noisy signal or improve the features of the speech signal before it reaches the recogniser. SS is a simple scheme, see Fig. 1, which has been successful for speech enhancement [4] [2] and speech recognition [14] [16] [10] [11]. This enhancement scheme has been successful even though it assumes a zero-noise variance. In order to deal with high noise variance, SS sets a minimum value and subtracts an over-estimation of the noise. This over-estimation significantly reduces the noise (ie it increases the SNR) and probably reduces also the variance, but it also distorts the speech signal. Over-estimation can be seen as a way to make a trade-off between reducing the noise and distorting the speech, and the optimum is obtained when both the noise and the distortion are minimised. Experimental results show that an over-estimation factor of about two on the magnitude spectrum is optimum. Another example of this trade-off is Single Value Decomposition (SVD), [15], the key idea of this scheme is that a little distortion can be often introduced in exchange for a large reduction in variance.

In parallel with the above enhancement work, the effects of noise on the performance of a HMM based recogniser have been studied [7], The conclusion is that by training and testing, see Fig. 2, in the same conditions, the HMM recogniser achieves the best performance. This is because the parameters of the HMMs "learn" the features of both the speech and the noise. However, trying to build a database with models for all conditions is impractical. Therefore, one way to deal with this practical problem is to automatically adapt the clean speech models to the noise. The aim of any noise adaptation technique is to reach the recognition performance obtained when the HMMs are trained in the same environment.

In general, the adaptation technique can be described as follows

$$\hat{m} = T_a\{m\} \tag{1}$$

where

m are the parameters of the clean speech model

$\hat{m}$ are the model parameters adapted to the noisy conditions, and

T_a is the transform from the clean speech model to the noisy speech model.

In past work, this transformation has been implemented in different ways. For example, when *cepstral mean compensation* [2] [1] is used, this transformation is the addition in the cepstral domain of the mean of the clean speech and a compensator value, which could be a function of the noise and the actual speech. When *noise masking* [12] is used this transformation is non-linear, for example, in the scheme by Varga & Ponting [18] the frequency band mean of the actual speech is replaced by the noise estimator if the noise is higher than the speech mean. When *Parallel Model Combination* (PMC) [5] is used, we have a non-linear transform in the cepstral domain. In this case, the speech and noise parameters are transformed from the cepstral domain to the linear domain, the parameters are added, and finally they are returned to the cepstral domain. This adaptation method has been very successful in adapting clean models to noisy environments and very good results have been achieved, for example 93% word correct for Lynx Helicopter noise on a digit database at 0 dB, [6], outperforming the results obtained by training and testing in the same noisy conditions.

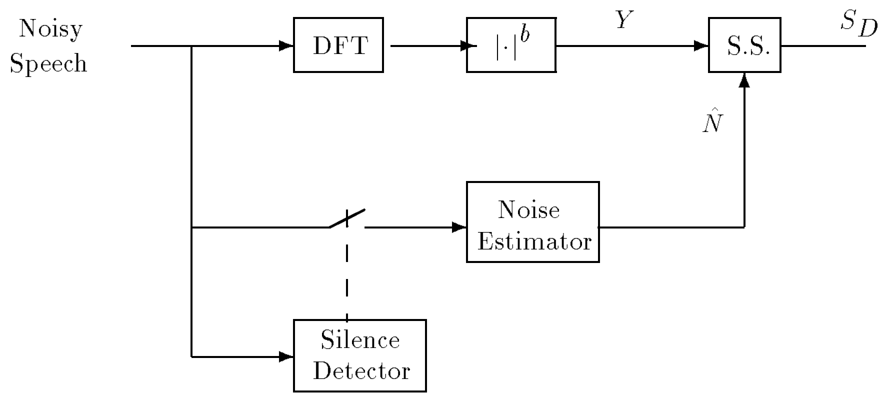

Figure 1: Spectral Subtraction, $b = 1$ gives magnitude subtraction, $b = 2$ gives power subtraction.

Eventually these noise adaptation techniques must reach a limit since recognition performance in very noisy environments is not completely solved by training in noise. Eventually, the variance becomes so large that feature overlapping begins to affect discrimination, and the problem becomes worse when the vocabulary grows. For example, it has been shown, [7] that the recognition performance can drop to 80% at 0 dB even when training and testing was on the same noisy conditions. Therefore, in very noisy environments, or when the vocabulary grows, even training in noise is not enough to obtain good recognition performance. In order to improve recognition performance in very noisy environments, enhancement techniques are needed. These may attempt to improve the SNR, minimise the variance, or emphasis the main features of the interesting signal. However, all of these improvements are usually at the expense of signal distortion. If both signal distortion and noise are minimised, then a signal with better features and lower variability is obtained. However, if the good features of these enhancement techniques are to be exploited, then the speech models need to be compensated to the distorted signal.

Adapting the models to the enhanced signal should raise the achievable limit of recognition performance. In order to determine this limit, a test similar to training and testing on the same condition can be made. Training and testing on the enhanced speech signal is shown in Fig. 3. In this case, the enhancement algorithm affects both the training and testing database. Although, trying to create a model database for all environments is impractical this configuration gives an indication of the maximum performance which can be achieved with an enhancement scheme. As shown in section 5, this maximum appears to be considerably higher than existing compensation schemes thus providing the motivation for seeking automatic methods to adapt clean speech models to the enhanced speech signal.

If Y is the noisy signal and Y is processed by an enhancer, $e(Y)$, a distorted signal, S_D, is obtained that is

$$S_D = e(Y) \tag{2}$$

In this work a HMM based recogniser with Gaussian state distributions is used. The Gaussian probabilistic density function, *pdf*, is completely modelled by the mean and the variance. Hence, in order to adapt the HMM, the mean and variance of each state have to be adapted. Therefore,

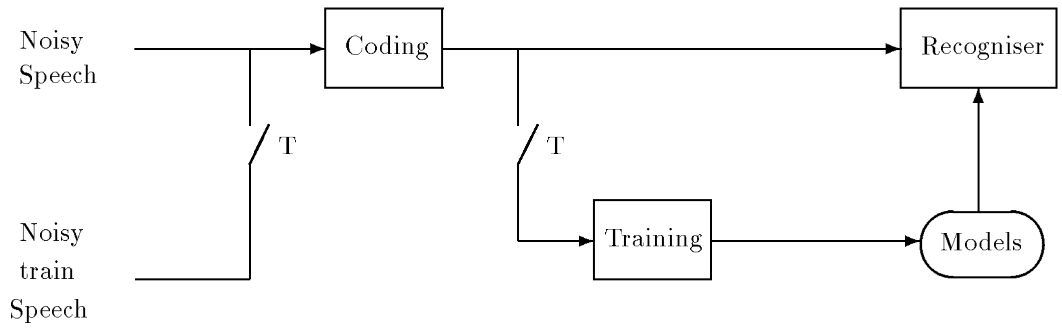

Figure 2: Training in Noise.

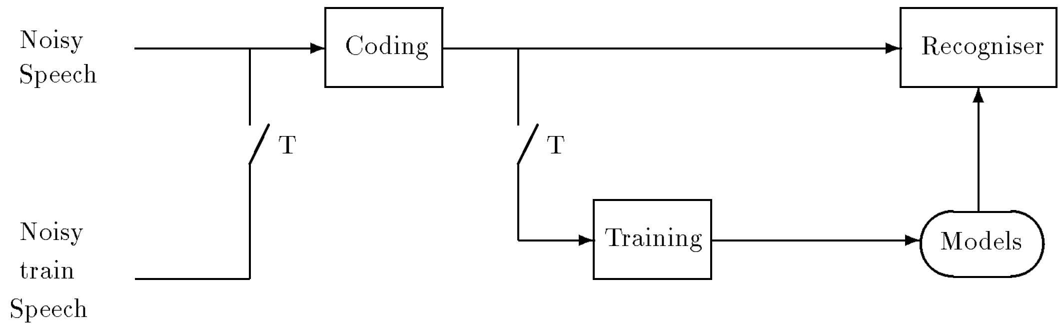

Figure 3: Training models on the enhanced speech.

eq. 1 can be rewritten for HMMs as follows

$$\hat{\mu} = T_{\hat{\mu}}\{\mu\} \tag{3}$$

$$\hat{\Sigma} = T_{\hat{\Sigma}}\{\Sigma\} \tag{4}$$

where

$T_{\hat{\mu}}\{.\}$ is the transformation from the mean of the clean speech model to the enhanced speech model, and

$T_{\hat{\Sigma}}\{.\}$ is the transformation from the variance of the clean speech model to the enhanced speech model.

In this work, we study the integration of Spectral Subtraction as enhancement technique, and PMC as adaptation technique. Because PMC only adapts the HMMs to the noise, it has to be extended to include the effects of the distortion. It is shown in Sec. 2, that the enhanced speech spectrum can be represented in the linear domain by the spectrum of the noisy speech plus a correction term, Δ. Therefore, this extension turns out to be relatively straightforward.

The remainder of this paper is organised as follows. Section 2 shows that the distorted speech (output) can be modelled as the addition in the linear domain of the noisy speech (input) plus a correction value. Section 3 shows how to modify the PMC algorithm to adapt the models to the enhanced speech. Section 4 presents experiments and results on the Noisex-92 database. Finally, Section 5 gives some comments and conclusions.

2 Analysis of Spectral Subtraction Distortion

In this section, Spectral Subtraction(SS) is described and the effect on the expected value and variance of a signal processed by this scheme. SS is attractive because in practice it has been shown to be successful for both signal enhancement is calculated [4] [3] and speech recognition in noise [14] [16] [10] [11]. Since it assumes zero-variance noise and real noise is highly variable, especially for low SNR, SS sets a minimum positive value on the enhanced signal and subtracts an over-estimation of the noise. By using SS, we obtain a signal with better features and lower variability. However, over-estimation and flooring makes the compensator non-linear and hence the noise level is reduced at the expense of introducing distortion into the speech signal. Over-estimation can be seen as a way to make a trade-off between reducing the noise and distorting the speech, the optimum over-estimation value is when both the noise and the distortion is minimum. Experimental results show that the optimum over-estimation factor is about two for the magnitude spectrum.

There are different spectral subtraction schemes, here the following scheme[9] is used, see Fig. 1,

$$D(Y) = Y - \alpha\hat{N}$$

$$Y_D(Y) = \begin{cases} D(Y) & D(Y) > \beta Y \\ \beta Y & otherwise \end{cases} \tag{5}$$

where

$Y_D(Y)$ is the higher SNR noisy signal or the distorted estimation of the clean speech S

Y is either the power or the magnitude spectrum of the noisy speech

$\hat{N}$ is an estimate of either the power or magnitude noise spectrum

α is an over-estimation factor, and

β is the spectral flooring parameter.

When Y is the magnitude spectrum, we will call it a Magnitude Spectral Subtraction (MSS) scheme, and when Y is the power spectrum we will call it a Power Spectral Subtraction (PSS) scheme. The distinction is important because the random variable transformation will lead to a different *pdf*, P(Y), for this two cases.

At the SS output an enhanced signal is obtained but it is distorted. To determine the impact of this distortion, its effect on the mean and variance of the signal need to be calculated. For simplicity it is assumed that each frequency channel is statistically independent, hence it is only necessary to develop the theory for the one-dimensional case. The expected value of the enhanced speech, Y_D is

$$E[Y_D] = \int_{-\infty}^{\infty} Y_D(Y)P(Y)dY$$

by using the SS scheme defined in eq. 5, we obtain

$$E[Y_D] = \int_{a(\alpha,\beta,\hat{N})}^{\infty} (Y - \alpha\hat{N})P(Y)dY + \int_{-\infty}^{a(\alpha,\beta,\hat{N})} \beta Y P(Y)dY \tag{6}$$

where $a(\alpha, \beta, \hat{N}) = \frac{\alpha\hat{N}}{1-\beta}$. If this equation is to be expressed as the linear combination of the noisy speech and the effect of the distortion, then the first integral on the right part of the last equation is completed as follows

$$E[Y_D] = \int_{-\infty}^{\infty} (Y - \alpha\hat{N})P(Y)dY + \int_{-\infty}^{a(\alpha,\beta,\hat{N})} \beta Y P(Y)dY - \int_{-\infty}^{a(\alpha,\beta,\hat{N})} (Y - \alpha\hat{N})P(Y)dY$$

$$E[Y_D] = \int_{-\infty}^{\infty} YP(Y)dY - \alpha\hat{N}\int_{-\infty}^{\infty} P(Y)dY + \int_{-\infty}^{a(\alpha,\beta,\hat{N})} \beta Y P(Y)dY - \int_{-\infty}^{a(\alpha,\beta,\hat{N})} (Y - \alpha\hat{N})P(Y)dY$$

then, the first integral is the expected value of the random variable Y, $E[Y]$, and the value of the second integral is unity since $P(Y)$ is a *pdf*. Hence, we obtain

$$E[Y_D] = E[Y] - \alpha\hat{N} + \int_{-\infty}^{a(\alpha,\beta,\hat{N})} \beta Y P(Y)dY - \int_{-\infty}^{a(\alpha,\beta,\hat{N})} (Y - \alpha\hat{N})P(Y)dY$$

$$E[Y_D] = E[Y] - \alpha\hat{N} + (\beta - 1)\int_{-\infty}^{a(\alpha,\beta,\hat{N})} YP(Y)dY + \alpha\hat{N}\int_{-\infty}^{a(\alpha,\beta,\hat{N})} P(Y)dY. \tag{7}$$

Let us define

$$A(\alpha,\beta,\hat{N},P(Y)) = \int_{-\infty}^{a(\alpha,\beta,\hat{N})} P(Y)dY \tag{8}$$

$$B(\alpha,\beta,\hat{N},P(Y)) = \int_{-\infty}^{a(\alpha,\beta,\hat{N})} YP(Y)dY \tag{9}$$

$A(\alpha,\beta,\hat{N},P(Y))$ is a cumulative distribution of the *pdf* $P(Y)$. $B(\alpha,\beta,\hat{N},P(Y))$ is not a cumulative distribution, but when a goes to infinity this correspond to the expected value of Y. By substituting eq. 8 and 9 in eq. 7, we obtain

$$E[Y_D] = E[Y] - \alpha\hat{N} + (\beta - 1)B(\alpha,\beta,\hat{N},P(Y)) + \alpha\hat{N}A(\alpha,\beta,\hat{N},P(Y)).$$

Now, defining

$$\Delta_\mu(\alpha,\beta,\hat{N},P(Y)) = -\alpha\hat{N} + (\beta - 1)B(\alpha,\beta,\hat{N},P(Y)) + \alpha\hat{N}A(\alpha,\beta,\hat{N},P(Y)) \tag{10}$$

we obtain

$$E[Y_D] = E[Y] + \Delta_\mu(\alpha,\beta,\hat{N},P(Y)) \tag{11}$$

From this equation, it can be observed that the effect on the distorted signal can be expressed as the addition of the noisy speech and a correction Δ_μ in the SS domain, and this function depends on the spectral subtraction parameters, α and β, the noise estimator $\hat{N}$ and the *pdf* $P(Y)$.

By comparing eq. 11 with eq. 3, we see that in this case $T_{\hat{\mu}}\{\hat{\mu}\}$ is the addition in the SS domain of the expected value of the noisy signal, $E[Y]$, plus a correction constant, Δ_μ. This is an interesting result, because it shows that, in the SS domain, we have the freedom to use any noisy adaptation algorithm and the correction Δ_μ can be compensated independently. For example, the PMC [5] adaptation technique assumes that the addition of the expected value of S, $E[S]$, plus the expected value of N, $E[N]$, is equal to the expected value of $E[Y]$.

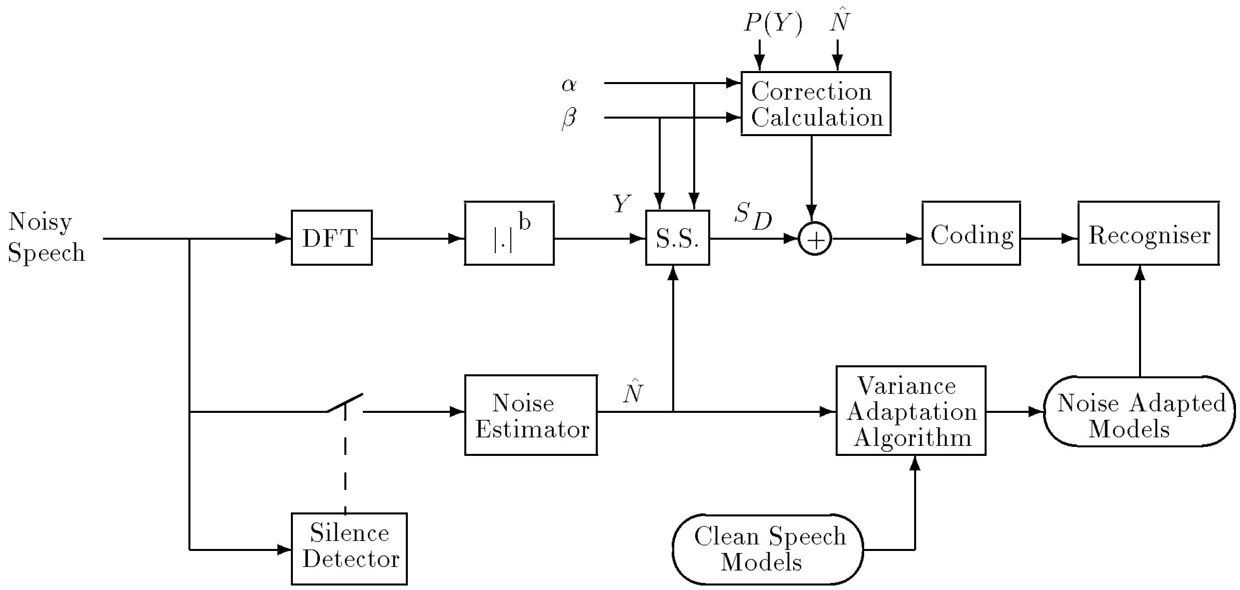

Figure 4: Mean compensation in the signal processing and variance compensation in the adaptation algorithm.

This is an approximation which can be used in eq. 11, but we are not restricted to using this approximation. If more accurate methods of estimating $E[Y]$ are developed, then they can be used instead.

In theory there are two ways in which mean compensation might be applied. In Fig. 4, it is applied as a correction in the signal processing stage. At first sight, this is attractive in that the correction only needs to be applied once to the signal and all of the models means remain unchanged. However, in practice, it is not directly implementable since Δ_μ depends on the spectrum of the underlying clean speech signal which is not known. The alternative scheme is to include the mean compensation within the model adaptation as shown in Fig. 5. In this case, the model means themselves provide the required estimates of the clean speech spectrum.

Moreover, assuming

$$E[Y] = E[S] + E[N],$$

as PMC does, we obtain

$$E[Y_D] = E[S] + E[N] + \Delta_\mu(\alpha, \beta, \hat{N}, P(Y))$$

From this equation it can be seen how the enhanced speech is distorted, and that this distortion can be compensated by simply adding $E[N]$ and Δ_μ to the clean speech model means.

Most of the algorithms for speech recognition in noise prefer not to compensate the variance, and it is common to use fixed variances. Fixed variance techniques replace the state variances with a single global variance which is obtained from all the words in the training database. Although, this is not a very elegant way to tackle the problem some good results have been obtained using this approach [13]. Parallel Model Combination(PMC), see Sec. 3, combines in the linear domain the speech and noise parameters, and transforms them back to the cepstral domain to obtain both mean and variance compensated models. In [6], Gales & Young show that this technique is more successful than the fixed variance technique. Therefore, we can either

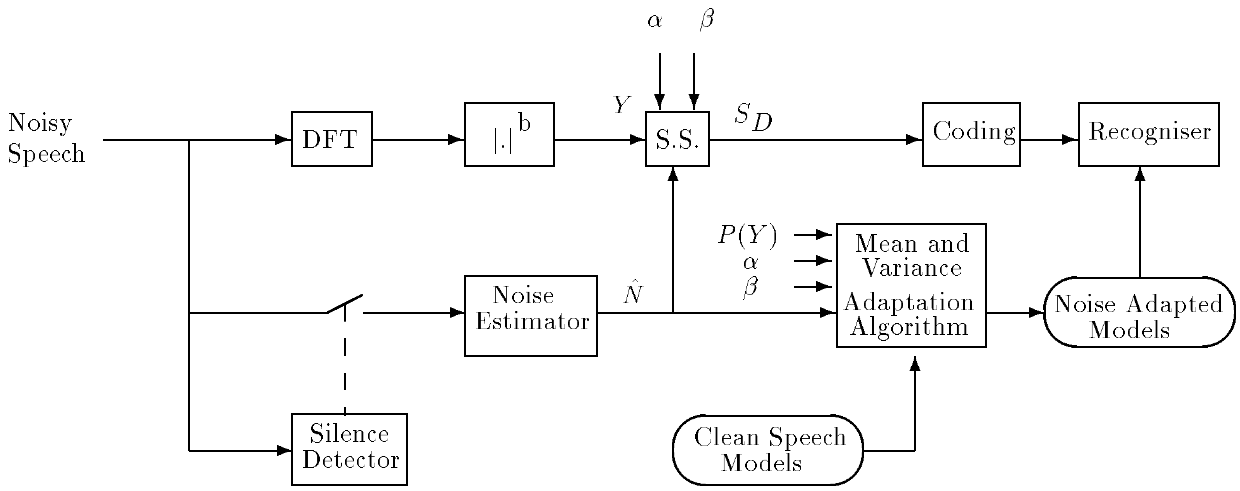

Figure 5: Mean and variance compensation in the adaptation algorithm.

replace the state variance of the clean speech by a the fixed variance or use the combination of the clean speech and noise variances generated by PMC.

In the latter case, we can also extend the equations to compensate for the effect of the distortion on the variances. The effect of SS on the variance can be calculated as follows

$$V[Y_D] = E[Y_D^2] - \{E[Y_D]\}^2 \tag{12}$$

$E[Y_D]$ is given by eq. 11, for $E[Y_D^2]$ we have

$$E[Y_D^2] = \int_{a(\alpha,\beta,\hat{N})}^{\infty} (Y - \alpha\hat{N})^2 P(Y)dY + \int_{-\infty}^{a(\alpha,\beta,\hat{N})} \beta^2 Y^2 P(Y)dY$$

where $a = \frac{\alpha\hat{N}}{1-\beta}$. Proceeding as before,

$$E[Y_D^2] = \int_{-\infty}^{\infty} (Y - \alpha\hat{N})^2 P(Y)dY + \int_{-\infty}^{a(\alpha,\beta,\hat{N})} \beta^2 Y^2 P(Y)dY - \int_{-\infty}^{a(\alpha,\beta,\hat{N})} (Y - \alpha\hat{N})^2 P(Y)dY$$

$$\begin{aligned} E[Y_D^2] &= E[Y^2] - 2\alpha\hat{N}E[Y] + \alpha^2\hat{N}^2 - \int_{-\infty}^{a(\alpha,\beta,\hat{N})} Y^2 P(Y)dY + 2\alpha\hat{N}\int_{-\infty}^{a(\alpha,\beta,\hat{N})} YP(Y)dY \\ &\quad -\alpha^2\hat{N}^2 \int_{-\infty}^{a(\alpha,\beta,\hat{N})} P(Y)dY + \beta^2 \int_{-\infty}^{a(\alpha,\beta,\hat{N})} Y^2 P(Y)dY \end{aligned}$$

by using the definitions of $A(\alpha, \beta, \hat{N}, P(Y))$ and $B(\alpha, \beta, \hat{N}, P(Y))$, eq. 8 and 9, and by defining

$$C(\alpha, \beta, \hat{N}, P(Y)) = \int_{-\infty}^{a(\alpha,\beta,\hat{N})} Y^2 P(Y)dY$$

we obtain

$$\begin{aligned} E[Y_D^2] &= E[Y^2] - 2\alpha\hat{N}E[Y] + \alpha^2\hat{N}^2 + 2\alpha\hat{N}B(\alpha, \beta, \hat{N}, P(Y)) - \alpha^2\hat{N}^2 A(\alpha, \beta, \hat{N}, P(Y)) \\ &\quad +(\beta^2 - 1)C(\alpha, \beta, \hat{N}, P(Y)). \end{aligned} \tag{13}$$

When a goes to infinity, $C(a, P(Y))$ is the second moment of the random variable Y with *pdf* $P(Y)$. By replacing eq. 13 in eq. 12, we obtain

$$\begin{aligned} V[Y_D] &= E[Y^2] - 2\alpha\hat{N}E[Y] + \alpha^2\hat{N}^2 + 2\alpha\hat{N}B(\alpha,\beta,\hat{N},P(Y)) - \alpha^2\hat{N}^2A(\alpha,\beta,\hat{N},P(Y)) \\ &\quad +(\beta^2-1)C(\alpha,\beta,\hat{N},P(Y)) - \{E[Y_D]\}^2. \end{aligned}$$

From this, it seems that an estimator for the second moment of the noisy speech is needed, $E[Y^2]$. However, since

$$V[Y] = E[Y^2] - \{E[Y]\}^2.$$

then

$$\begin{aligned} V[Y_D] &= V[Y] + \{E[Y]\}^2 - 2\alpha\hat{N}E[Y] + \alpha^2\hat{N}^2 + 2\alpha\hat{N}B(\alpha,\beta,\hat{N},P(Y)) \\ &\quad -\alpha^2N^2A(\alpha,\beta,\hat{N},P(Y)) + (\beta^2-1)C(\alpha,\beta,\hat{N},P(Y)) - \{E[Y_D]\}^2. \end{aligned}$$

Hence, defining

$$\begin{aligned} \Delta_\Sigma(\alpha,\beta,\hat{N},P(Y)) &= \{E[Y]\}^2 - 2\alpha\hat{N}E[Y] + \alpha^2\hat{N}^2 + 2\alpha\hat{N}B(\alpha,\beta,\hat{N},P(Y)) \\ &\quad -\alpha^2\hat{N}^2A(\alpha,\beta,\hat{N},P(Y)) + (\beta^2-1)C(a,P(Y)) - \{E[Y_D]\}^2 \quad (14) \end{aligned}$$

we can re-express the variance equation as the linear combination of the noisy variance and Δ_Σ to obtain

$$V[Y_D] = V[Y] + \Delta_\Sigma(\alpha,\beta,\hat{N},P(Y)). \quad (15)$$

It can be observed that Δ_Σ depends on the parameters of SS, α and β, the noise estimator $\hat{N}$ and the *pdf* $P(Y)$.

By comparing eq. 15 with eq. 4, it can be seen that the transform $T_\Sigma\{V[Y_D]\}$ is the addition of the variance of the noisy signal, $V[Y]$, plus a correction in the SS domain, $\Delta_\Sigma(\alpha,\beta,\hat{N},P(Y))$. As before, the practical way to compensate the variance is by modifying the adaptation algorithm as shown in Fig. 5.

A potential problem of this scheme is the solution of the integrals $A(\alpha,\beta,\hat{N},P(Y))$, $B(\alpha,\beta,\hat{N},P(Y))$ and $C(\alpha,\beta,\hat{N},P(Y))$ since the transformation could yield an intractable *pdf* P(Y). However, for this work it is assumed that the distributions are log normal in the linear domain and solutions exist for this case.

Assume that Y has a log normal *pdf*, defined as follows

$$P(Y) = \frac{1}{Y\sqrt{2\pi}\psi}e^{-\frac{1}{2\psi^2}(ln(Y)-\xi)^2}$$

P(Y) is completely defined by ξ and ψ, hence the general notation $A(a, P(Y))$ may be written as $A(a,\xi,\psi)$ where

$$A(a,\xi,\psi) = \int_{-\infty}^{a(\alpha,\beta,\hat{N})} \frac{1}{Y\sqrt{2\pi}\psi}e^{-\frac{1}{2\psi^2}(ln(Y)-\xi)^2}dY$$

this integral can be solved by the variable substitution $z = ln(Y)$ to give

$$A(\alpha, \beta, \hat{N}, \xi, \psi) = \int_{-\infty}^{ln(a(\alpha,\beta,\hat{N}))} \frac{1}{\sqrt{2\pi}\psi} e^{-\frac{1}{2\psi^2}(z-\xi)^2} dz.$$

The integrand is a Normal distribution and hence given the Normal cumulative density function

$$\mathcal{C}(b, \mu, \sigma^2) = \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

$A(a, \xi, \psi)$ can be expressed as

$$A(\alpha, \beta, \hat{N}, \xi, \psi) = \mathcal{C}(ln(a(\alpha, \beta, \hat{N})), \xi, \psi^2)$$

Similarly, see the appendix, $B(a, \xi, \psi)$ and $C(a, \xi, \psi)$ for log normal *pdf*, P(Y), can be reduced to

$$B(\alpha, \beta, \hat{N}, \xi, \psi) = E[Y]\mathcal{C}(ln(a(\alpha, \beta, \hat{N})), \xi + \psi^2, \psi^2)$$

$$C(\alpha, \beta, \hat{N}, \xi, \psi) = E[Y^2]\mathcal{C}(ln(a(\alpha, \beta, \hat{N})), \xi + 2\psi^2, \psi^2)$$

where
$E[Y] = e^{\xi+\psi^2/2}$
$E[Y^2] = e^{2(\xi+\psi^2)}$
To reduce calculation, we can define the normalized Normal cumulative function

$$G\left(\frac{b-\mu}{\sigma}\right) = \mathcal{C}(b, \mu, \sigma^2)$$

Hence, $A(\alpha, \beta, \hat{N}, \xi, \psi)$, $B(\alpha, \beta, \hat{N}, \xi, \psi)$ and $C(\alpha, \beta, \hat{N}, \xi, \psi)$ can be re-expressed as follows

$$A(\alpha, \beta, \hat{N}, \xi, \psi) = G\left(\frac{ln(a(\alpha, \beta, \hat{N})) - \xi}{\psi}\right) \tag{16}$$

$$B(\alpha, \beta, \hat{N}, \xi, \psi) = E[Y]G\left(\frac{ln(a(\alpha, \beta, \hat{N})) - (\xi + \psi^2)}{\psi}\right) \tag{17}$$

$$C(\alpha, \beta, \hat{N}, \xi, \psi) = E[Y^2]G\left(\frac{ln(a(\alpha, \beta, \hat{N})) - (\xi + 2\psi^2)}{\psi}\right) \tag{18}$$

The cumulative function $G(x)$ does not have an exact solution but for small values, it can be approximated by a table look-up and for large values, it can can be approximated by $G(x) = \frac{1}{x}e^{-x^2}$ or by $G(x) = (\frac{1}{x} + \frac{1}{x^3})e^{-x^2}$. Moreover, $\mathcal{N}(0, 1)$ is a symmetric function, hence the size of the table look-up can be reduced by half using the following equation [8]

$$G(x) = 1 - G(-x).$$

The above expressions for $A(a,\xi,\psi)$, $B(a,\xi,\psi)$ and $C(a,\xi,\psi)$ allow the required correction factors Δ_μ and Δ_Σ to be calculated in terms of the expected values of Y and Y^2 and the parameters ξ and ϵ of the log normal distirbution $P(Y)$. As shown in the appendix,

$$E[Y] = e^{\xi+\psi^2/2} \tag{19}$$

and

$$E[Y^2] = e^{2(\xi+\psi^2)}$$

from which it is straighforward to show that

$$\psi^2 = ln(V[Y] + \{E[Y]\}^2) - 2ln(\{E[Y]\}) \tag{20}$$

and

$$\xi = ln(E[Y]) - \psi^2/2 \tag{21}$$

replacing eq. 20 in eq. 21 we obtain

$$\xi = -0.5ln(V[Y] + \{E[Y]\}^2) + 2ln(\{E[Y]\}^2) \tag{22}$$

3 SS-PMC

In the previous section, it has been shown that assuming the spectral distributions in the linear domain are log normal then it is possible to calculate correction values Δ_μ and Δ_Σ for the SS distortion. The assumptions of log normality in the linear domain cover the main representations used for speech recognition. The SS may be applied to the magnitude spectrum (MSS) or to the power spectrum (PSS) and the same equations apply. Recognition may use either the log filter bank parameters or a cepstral representation. However, these are linked by a linear transformation and both may be assumed to be log normal.

The calculation of Δ_μ and Δ_Σ require both an estimate of the noise N and an estimate of the clean speech S to enable the expectations $E[Y]$ and $Var[Y]$ to be calculated.

As noted earlier, the PMC technique developed by Gales & Young [5] allows the latter expectations to be calculated by assuming that the HMM Gaussian output distributions characterise N and S in the log domain. These parameters are then mapped into the linear domain where additivity is assumed to hold, thereby allowing $E[Y]$ and $Var[Y]$ to be calculated. Given the theory developed in section 2, it is straightforward to extend this PMC approach to the SS case. As shown in Fig. 6, the basic PMC framework is unaltered save for the inclusion of the Δ factors in the linear domain. Thus, the overall steps in the SS-PMC scheme are

1. Transform the cepstral (or log) means and variances back into the linear domain as in standard PMC.

2. Calculate $E[Y]$ and $E[Y^2]$ assuming additivity of the speech and noise.

3. Calculate Δ_μ and Δ_Σ and adjust the means and variances.

4. Transform the compensated means and variances back to the cepstral (or log) domain as in standard PMC.

Fig. 7 shows the block-diagram representation for the Δ_μ and Δ_Σ calculation. In order to make the explanation of the algorithm clear, the problem is divided into three general steps. Each step is completed with intermediate steps. Fig. 7 shows the general steps separated by thicker lines. The calculations are based on the equations developed in Sec. 2, specifically, on eqs. 16, 17, 18, 10, 14, 21 and 20.

The realization of this block diagram in a sequential computer can be done has follows, given the mean and variance of the noisy speech, μ_Y and Σ_Y, and the parameters of the SS, α and β, as input parameters, the general steps to solve are:

- Obtain a, ξ and ψ.

- Calculate $A(a,\xi,\psi)$, $B(a,\xi,\psi)$ and $C(a,\xi,\psi)$

- Calculate Δ_μ and Δ_Σ

The detailed intermediate steps should be clear from Fig. 7.

Once the correction factors have been applied we return to the standard PMC algorithm and transform from the linear domain back to the cepstral (or log) domain.

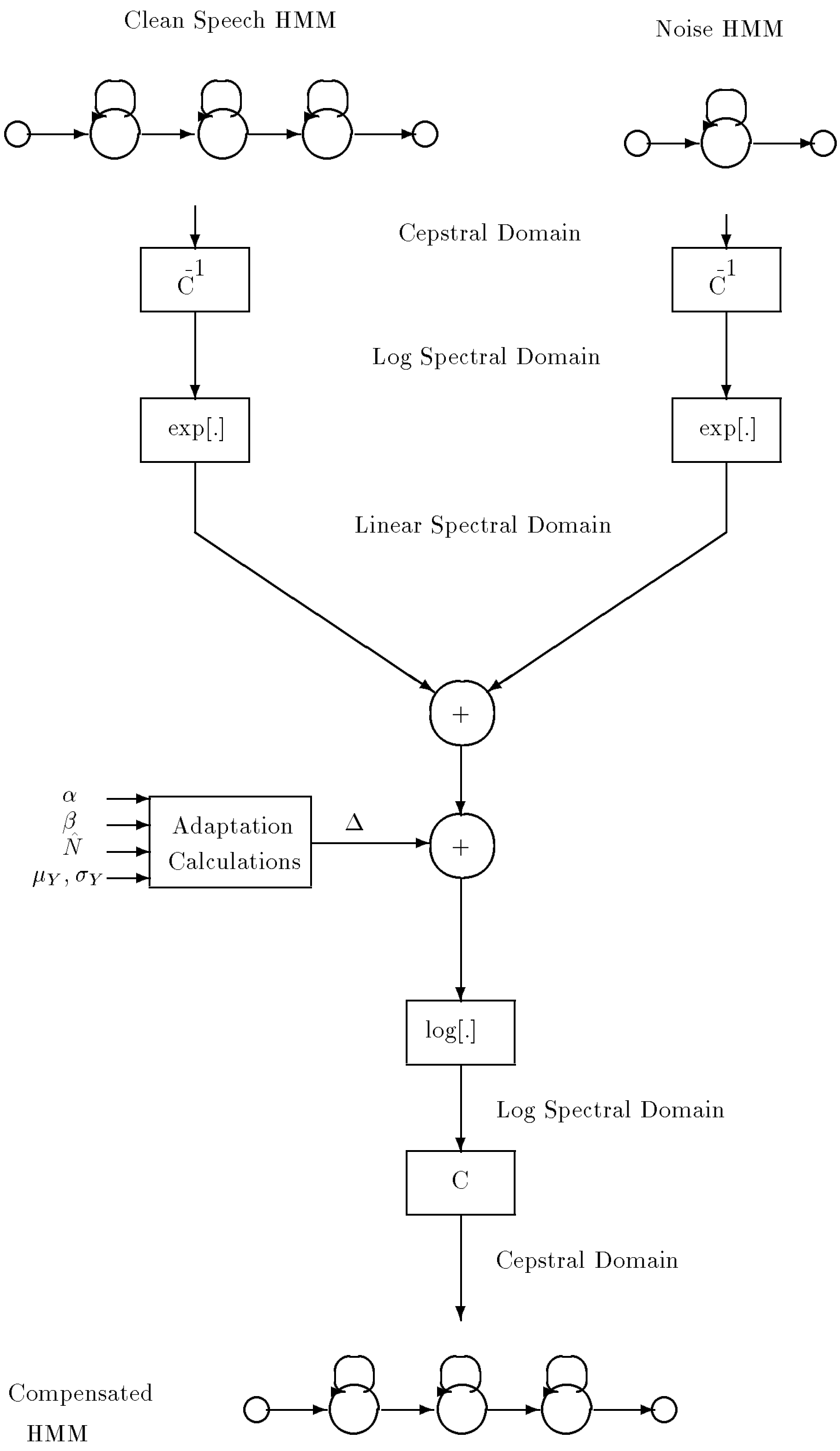

Figure 6: Block-diagram representation of SS-PMC

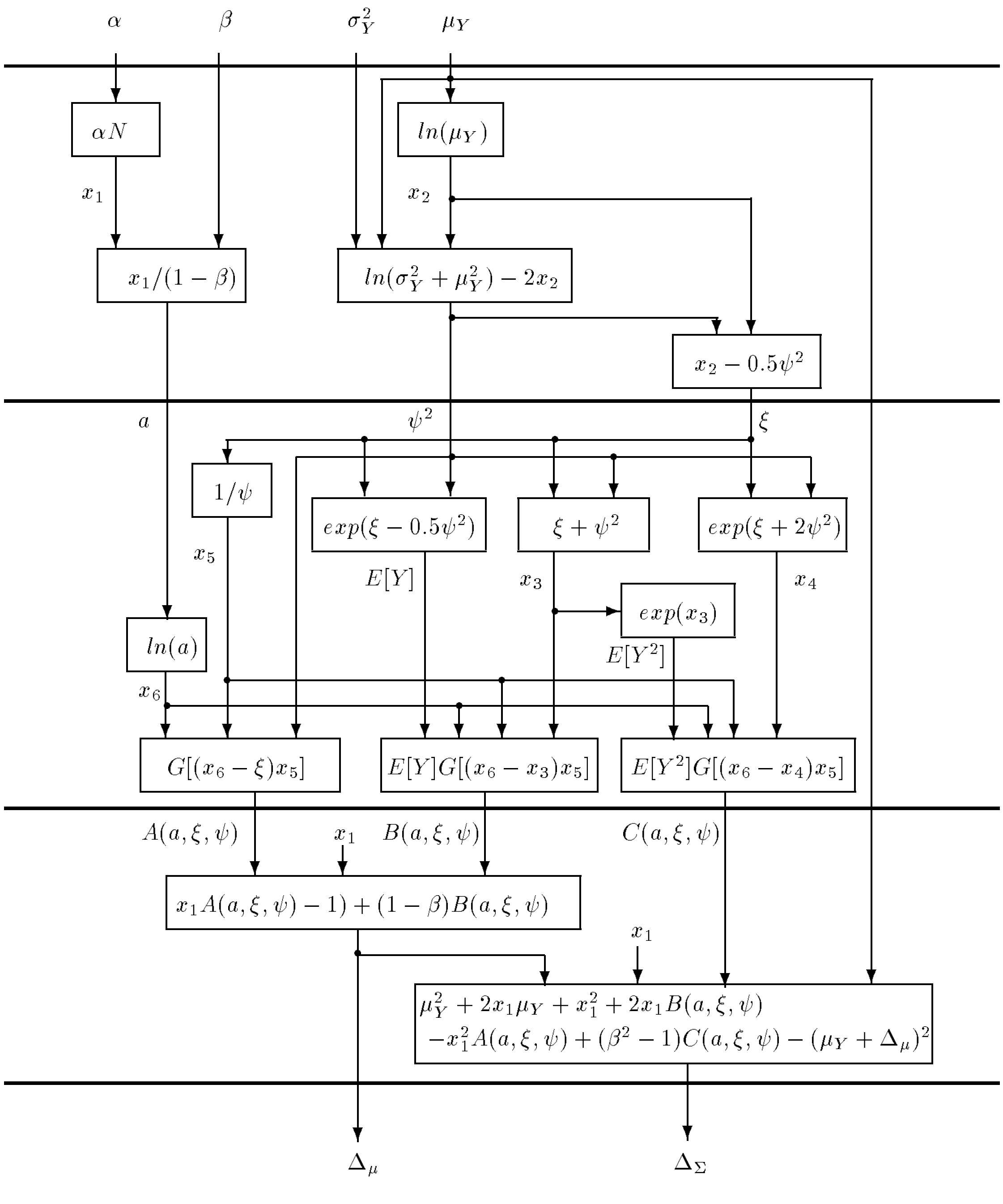

Figure 7: Block-diagram representation for Δ_μ and Δ_Σ

4 Experiments and Results

The SS-PMC technique described in the previous sections has been evaluated using the Noisex-92 database. This database was created by artificially adding noise to clean speech, therefore, it does not exhibit the Lombard effect. This database contains digits, 100 training utterances recorded in silence and 100 testing utterances with different noises, e.g. car and helicopter, at levels in the range +18 dB to -6 dB. Lynx helicopter, car and F16 were the noises used for our experiments. These noises were chosen because the scheme assumes stationary noise, and these are more or less stationary. 0 dB and -6 dB were the levels used to test very noisy conditions. This database has a male speaker and a female speaker, using their native language, English. The male speaker was used for our experiments.

The baseline recogniser used 10 state word-based HMMs, with 8 emitting states and single Gaussian diagonal covariance matrix output probability distributions. The speech was pre-processed using a 25 ms Hamming window, and then parameterised into the first 14 cepstral coefficients obtained from either magnitude spectrum (MMFCC) or power spectrum (PMFCC). The clean speech model variances for the baseline recogniser were replaced by a fixed-variance. This fixed-variance was obtained from all the training data.

For the SS-PMC scheme, a HMM for each digit was trained using the clean speech data only. The noise model used a single emitting state model and it was trained on all the available noise data. This noise model uses 1 emitting state and single gaussian diagonal covariance output probability distribution. The topology for all models was left-right with no skips and diagonal covariances were assumed throughout. For each frame, a set of 15 MFCC (either MMFCC or PMFCC) coefficients were computed. The zeroth cepstral coefficient is computed and stored since it is needed in the SS-PMC mapping procedure. However, it is subsesquently dropped in the actual recognition process. SS-PMC compensates for the means as discussed in Sec. 3. In order to obtain wider variances, the term Δ_Σ was set to zero. Another way to keep the variance wider is using fixed-variance, but no experiments were tried using fixed-variance.

Recognition used a standard connected word Viterbi decoder. All the training and testing used version 1.4 of the portable HTK HMM toolkit [19], with suitable extensions to perform SS-PMC.

Noisex-92 database is a continuous digit database, but an ideal word detector is assumed since this work is only concerned with determining how well the models have been adapted to the signal enhancement scheme.

The spectral subtraction scheme discribed in section 2 is used for the signal enhancement. The parameter α was varied from 0 to 2.4, and β was fixed at 0.1. Some experiments varying β were performed, and it was found that $\beta = 0.1$ has a good behaviour and values around this have similar recognition performance. The noise estimator is obtained from the average of the twenty spectral samples before the words starts. No attempt was made to apply time-frequency smoothing.

As was discussed in Sec. 1, an upper limit on recognition performance can be estimated by applying the enhancement technique to the training noise data, see Fig. **??**, such that the models "learn" the features of the enhanced signal. Fig. 8, shows the results on the Lynx noise when MSS-MFMCC and PSS-MFPSS schemes are applied for different values of α. Fig. 8 also shows the result when the spectral subtractor is applied before the filterbank, we will refer these latter schemes as the bMSS-MMFCC and bPSS-PMFCC. The theory developed in Sec.

2 assumes that the spectral subtraction is applied after filterbank processing, but assuming that the filterbank is some kind of frequency smoother and that the smoother does not have a significant effect on the subtracted signal, hence the theory of Sec. 2 can also be applied to the bMSS-MMFCC and bPSS-PMFCC cases.

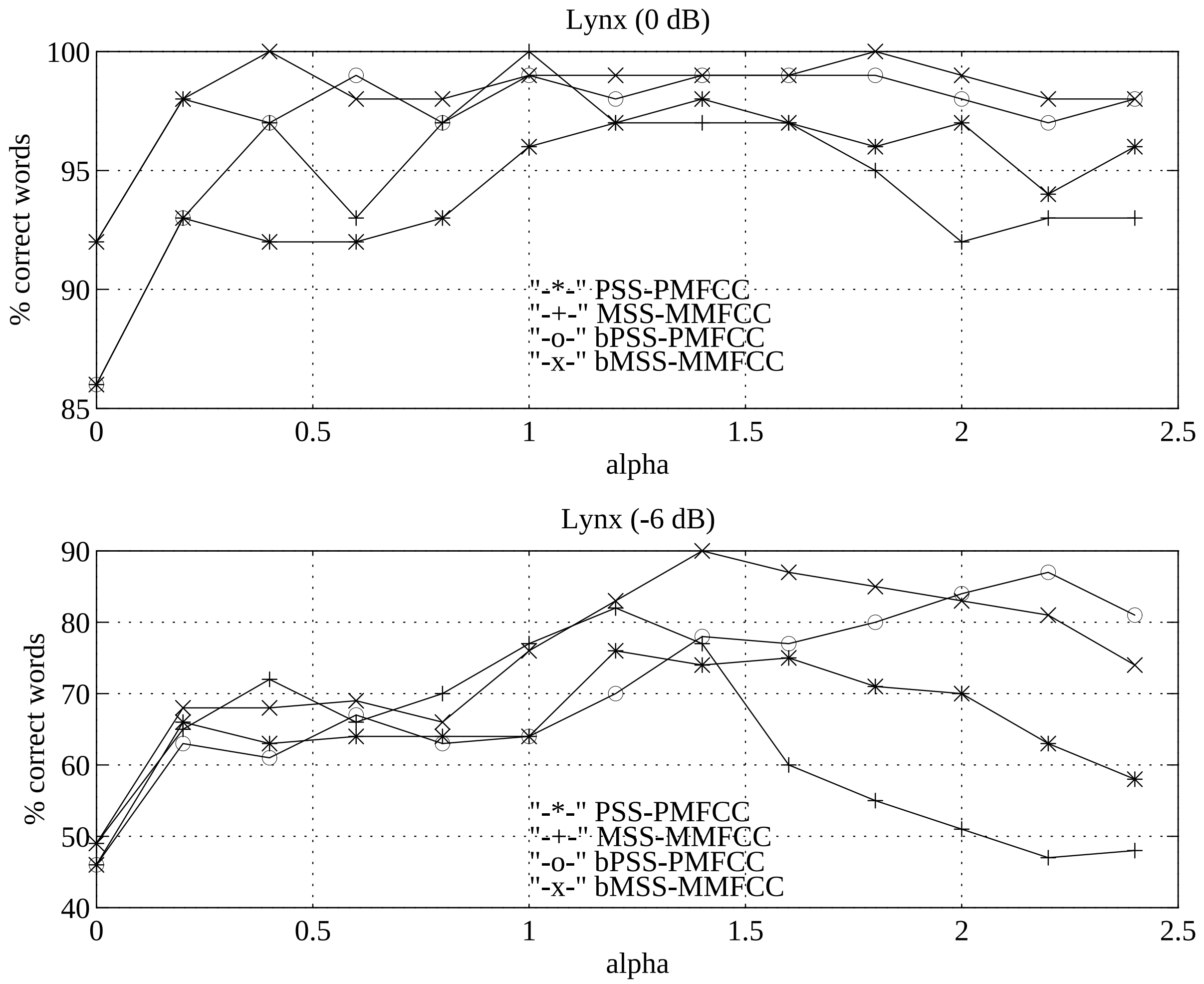

Figure 8: Best recognition performance for Lynx at 0 and -6 dB SNR using four different enhancement techniques

Fig. 8 indicates the best performance of each of the schemes for Lynx noise at 0 and -6 dB SNR. For all cases, the introduction of the speech enhancer improves the recognition performance. The best recognition performance, when training and testing in the same noise environment without any enhancement corresponds to $\alpha = 0$. For example, Fig. 8 shows that when training and testing in the noise conditions for the Lynx noise at -6 dB, 42% recognition performance is obtained for MSS-MMFCC. This represents the best performance that we can expect to obtain with an ideal model adaptation algorithm. However, when the enhacement scheme is included, the best recognition performance goes up to 78% ($\alpha = 1.6$). Hence, by adding the enhancement scheme, the upper limit of recognition performance is increased.

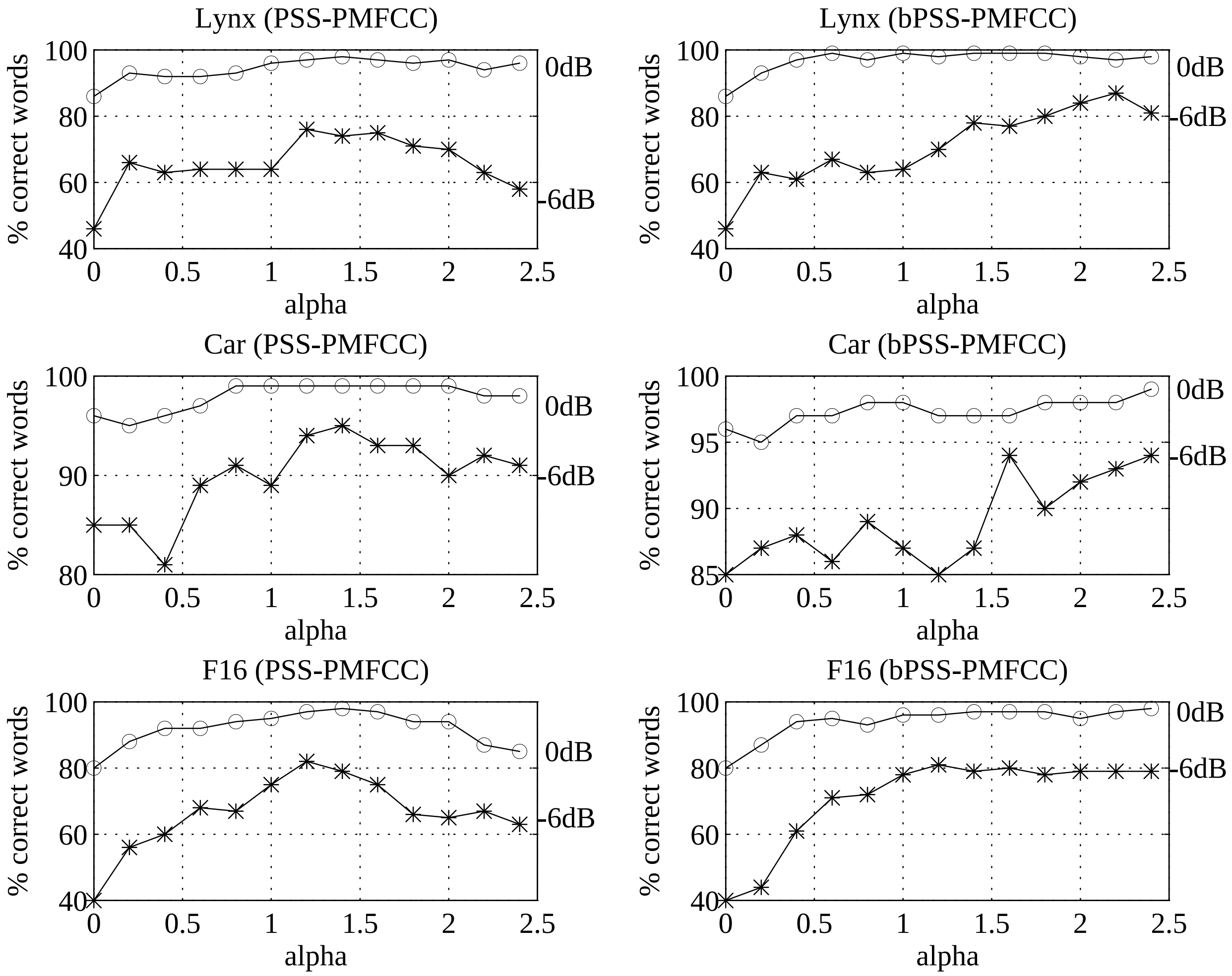

Figure 9: Best recognition performance for Lynx, Car and F16 for PSS-PMFCC and bPSS-PMFCC

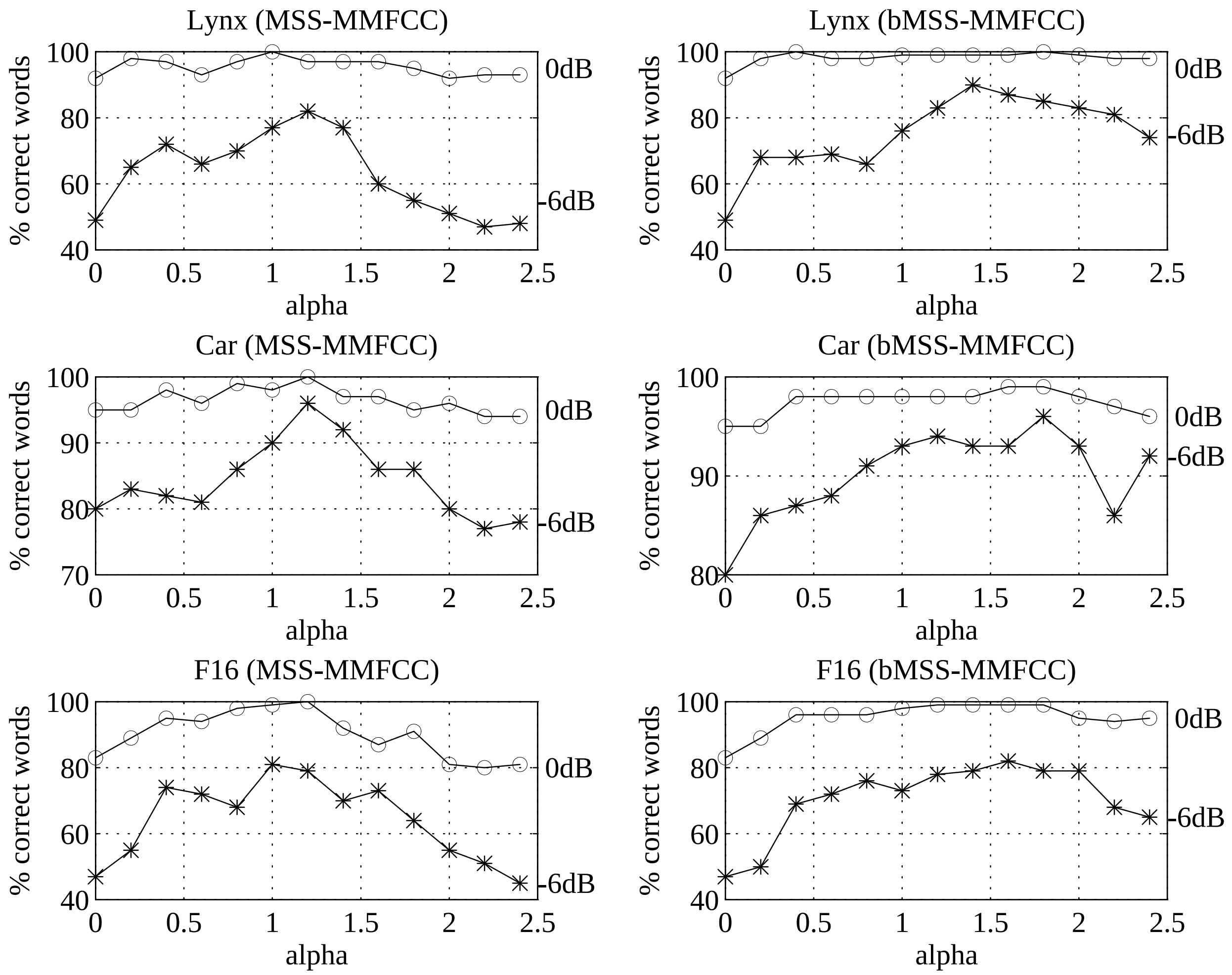

Figure 10: Best recognition performance for Lynx, Car and F16 for MSS-MMFCC and bMSS-MMFCC

Therefore, compensating for the distortion of the SS scheme, in the best case it should be possible to reach this improved recognition performance. It can be observed that a similar improvement is obtained by the other schemes. Finally, Fig. 8 shows that for magnitude spectrum schemes a sharp curve is observed around the optimal value of $\alpha = 1.4$. The curve for the power spectrum schemes is smoother suggesting that they are more robust to the precise setting of α.

In general, for Lynx noise at 0 and -6 dB SNR, for any value of α, the schemes with SS before the filterbank obtain better results. Although this results suggest the use of SS before the filterbank preprocessing, it is useful to see how well each of the enhancement schemes work with our compensation algorithm.

It can be observed that these schemes reach maximum recognition performance for some values of α (e.g. maximum recognition performance for bMSS-MFCC is at $\alpha = 1.4$ for Lynx noise at -6 dB). These values of α represent the values at which maximum noise reduction and maximum discrimination is reached. It is expected that these values will be dependent on the SNR, the variance and the vocabulary size.

Fig. 9 shows the best performance for the Lynx noise, Car noise and F16 noise for 0 dB and -6 dB for PSS-PMFCC and bPSS-PMFCC, and Fig. 10 shows the best performance for the Lynx noise, Car noise and F16 noise for 0 dB and -6 dB for MSS-MMFCC andb bMSS-MMFCC. Again, for Car noise and F16 noise the best performance results using any of the signal enhancement techniques outperform the recognition performance without using them. This means that these signal enhancement techniques also deal well with Car noise and F16 noise. In fact, Figs. 9 and 10 show that SS deals very well with car noise, and the best recognition performance for an ideal model adaptation technique goes over 94% using any of the signal enhancers tested in this work. These figures also show that the best expected results for F16 noise were less successful with the best recognition performance around 80% for $-6dB$.

To test our adaption scheme, the models were trained with clean speech and compensated by using the SS-PMC decribed in Sec. 2. For pre-processing either MSS or PSS were used, before and after the filter bank, and as discussed above, in order to keep the variance wider, we only compensated for the means. In order to compensate for all the assumptions and approximations, Δ_μ is weighted by γ. The experiments show that values of γ between 0.8 and 0.7 are good for 0 dB SNR, and 0.7 and 0.5 for -6 dB SNR.

Fig. 11 shows the results for Lynx noise, Car noise and F16 noise at 0 and -6 dB SNR, when PSS-PMFCC is used as the signal enhancer. This graph shows the best performance using an ideal adaptation technique (i.e. the models were trained on the enhanced signal), and the results when the Δ correction were applied for two values of γ. From the point of view of recognition performance the results are very good for any value of α, and from the point of view of how close the results are to best, these are not very satisfactory for $\alpha < 1$, but they improve when $\alpha > 1$, and they reach comparable and some times better results than the best expected results. The maximum performance is reached for $1.5 \leq \alpha \leq 2.0$.

Fig. 12 shows the experimental results when bPSS-PMFCC is used as enhancer. In this case, most of the time for 0 dB when $\alpha < 1$ the results are comparable to the best expected results and for $\alpha > 1$ the results most of the time outperform the best expected results. In this case, the recognition performance is not very sensitive to small changes in α, hence a smoother graph is obtained. Again, the best performance is obtained for $1.5 \leq \alpha \leq 2.0$.

Fig. 13 shows the results for MSS-MMFCC, the results for this scheme are very poor, and much of the time, the recognition performance without SS ($\alpha = 0.0$) is better than when compensating the distortion of the enhanced speech. The best performance is obtained for $1.4 \leq \alpha \leq 1.8$.

Fig. 14 shows the results for bMSS-MMFCC, the results of this scheme are better than MSS-MMFCC, and we can see that compensating the distortion of the enhanced speech is effective. Here, again the best performance is obtained for $1.0 \leq \alpha \leq 1.4$. Even though good results are obtained with this magnitude scheme, the schemes which use the power spectrum outperform it.

Table 1 sumarises the results for a standard fixed variance HMM based recognition system, MSS, bMSS, PSS and bPSS. These results were obtained for $0.0 \leq \alpha \leq 3.0$ and $\beta = 0.1$. As noted earlier, the value of β was fixed at 0.1 from the outset. This may not therefore represent an optimum setting. Table 2 sumarises the performance obtained using the various SS-PMC schemes. From this table, it appears that bPSS-PMFCC is the best scheme for 0 dB and PSS-PMFCC is the best for -6 dB, the difference is not great however and looking at Fig. 12 and 13 it can be observed that bPSS-PMFCC sometimes presents a smoother curve and is therefore less sensitive to small variations of α.

	SNR(dB)	Lynx (%)	Car (%)	F16 (%)
Std. HMM	0	32	42	42
	-6	20	16	12
PSS-PMFCC	0	70	81	60
	-6	46	56	43
bPSS-PMFCC	0	78	83	64
	-6	50	53	46
MSS-MMFCC	0	77	84	69
	-6	54	61	47
bMSS-MMFCC	0	84	88	69
	-6	56	59	46

Table 1: Correct words baseline results ($0 \leq \alpha \leq 3.0$ and $\beta = 0.1$).

Method	SNR (dB	Lynx (%)	Car (%)	F16 (%)
PSS-PMFCC	0	97($\alpha = 2.0$)	99 ($\alpha = 1.6; \alpha = 2.0$)	99 ($\alpha = 1.4; \alpha = 1.6$)
	-6	82($\alpha = 1.4$)	92 ($\alpha = 1.8$)	85($\alpha = 1.4$)
bPSS-PMFCC	0	100($1.6 \leq \alpha \leq 1.8$)	100($1.4 \leq \alpha \leq 1.8$)	99($1.8 \leq \alpha \leq 2.0$)
	-6	81($1.4 \leq \alpha \leq 1.6$)	92 ($\alpha = 1.8$)	81 ($1.8 \leq \alpha \leq 2.0$)
MSS-MMFCC	0	67($\alpha = 1.4$)	92($\alpha = 1.8$;$\alpha = 2.0$)	79($\alpha = 1.6$)
	-6	37($\alpha = 0.0$)	68($1.0 \leq \alpha \leq 1.4$)	62($\alpha = 1.4$)
bMSS-MMFCC	0	83($\alpha = 1.6$)	97($\alpha = 1.8$)	100($\alpha = 1.6$)
	-6	47($\alpha = 1.6$)	86($1.0 \leq \alpha \leq 1.8$)	79($\alpha = 1.8$)

Table 2: Best results for our adaptive schemes.

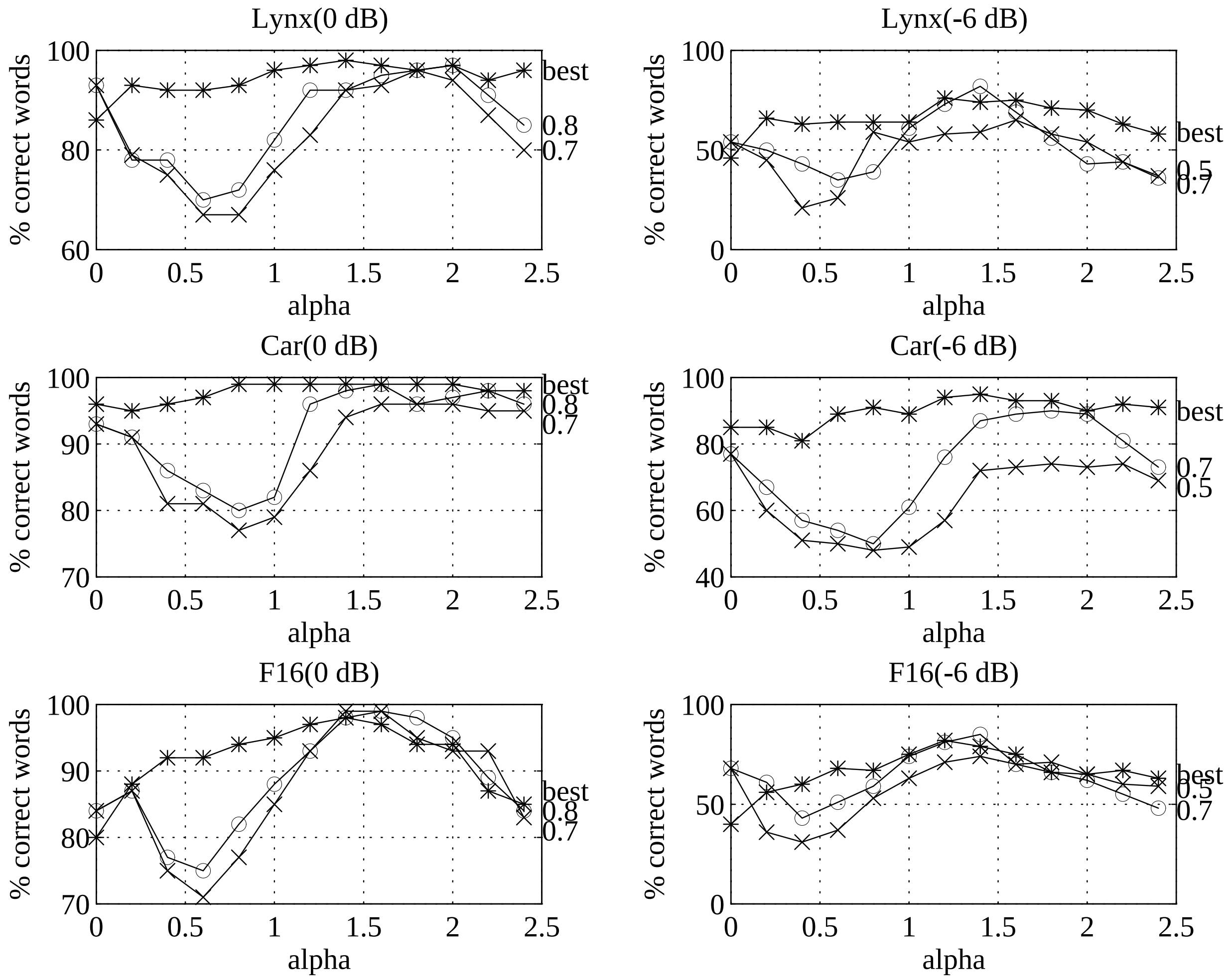

Figure 11: SS-PMS results for PSS-PMFCC at 0 and -6 dB SNR

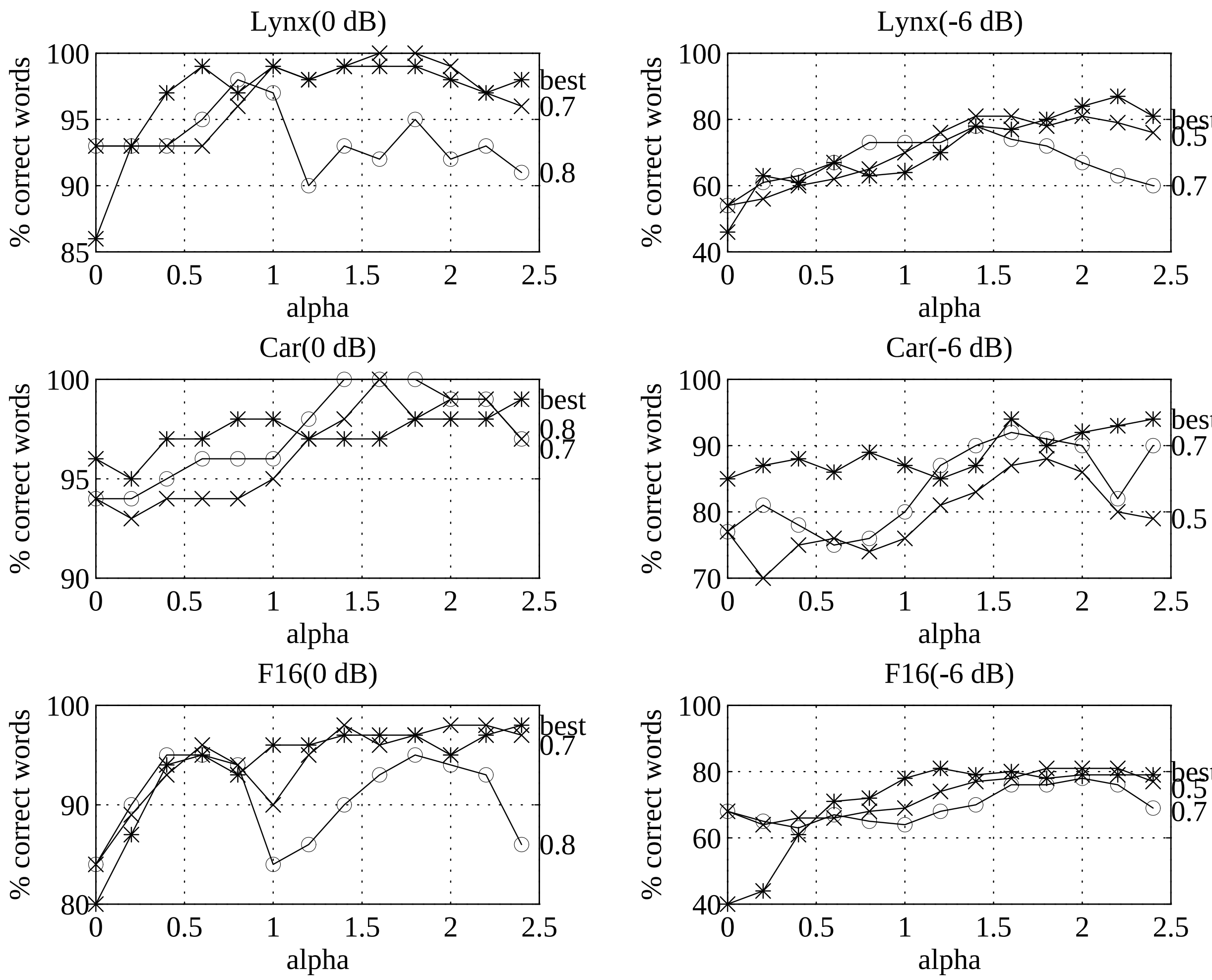

Figure 12: SS-PMS results for bPSS-PMFCC at 0 and -6 dB SNR

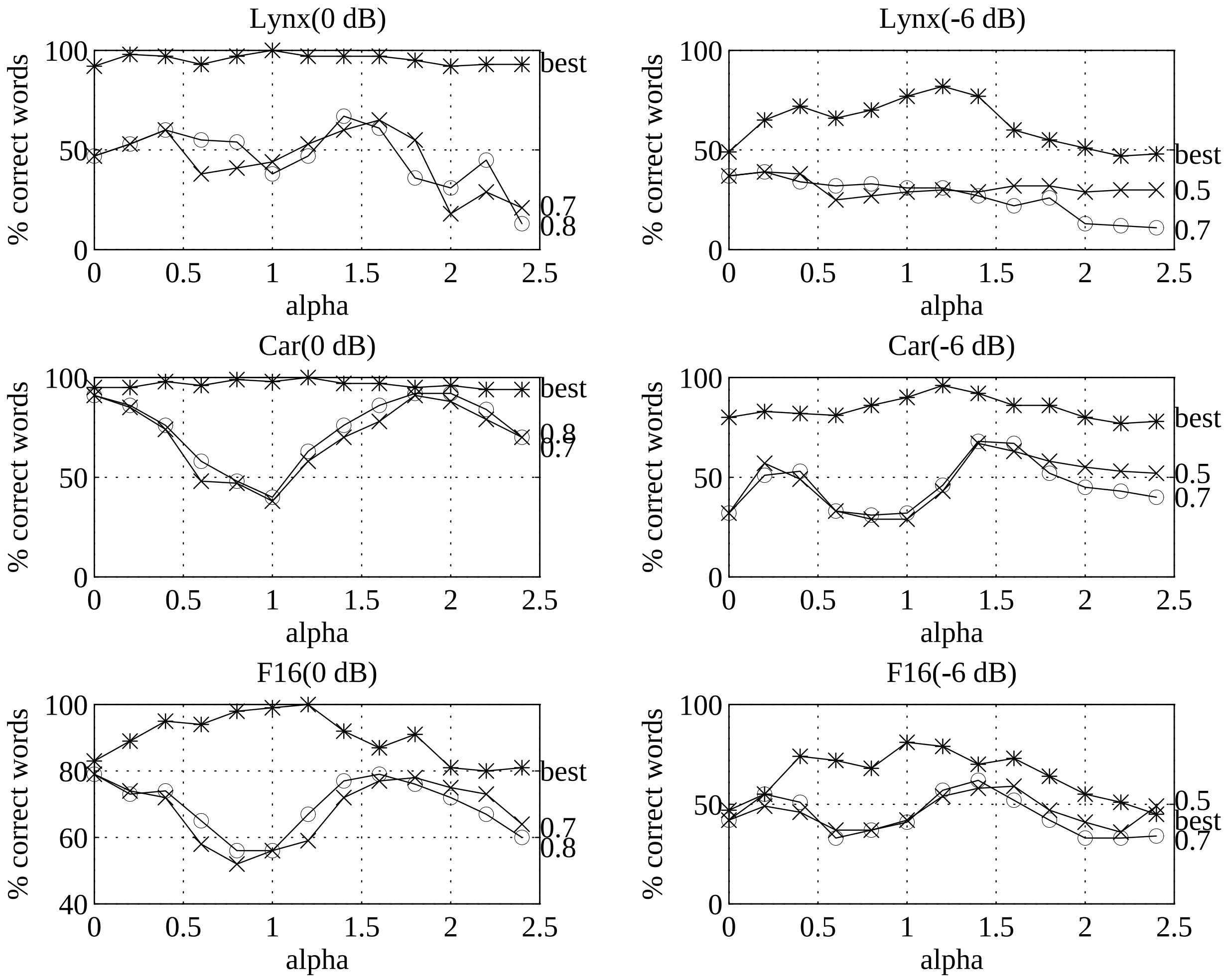

Figure 13: SS-PMS results for MSS-MMFCC at 0 and -6 dB SNR

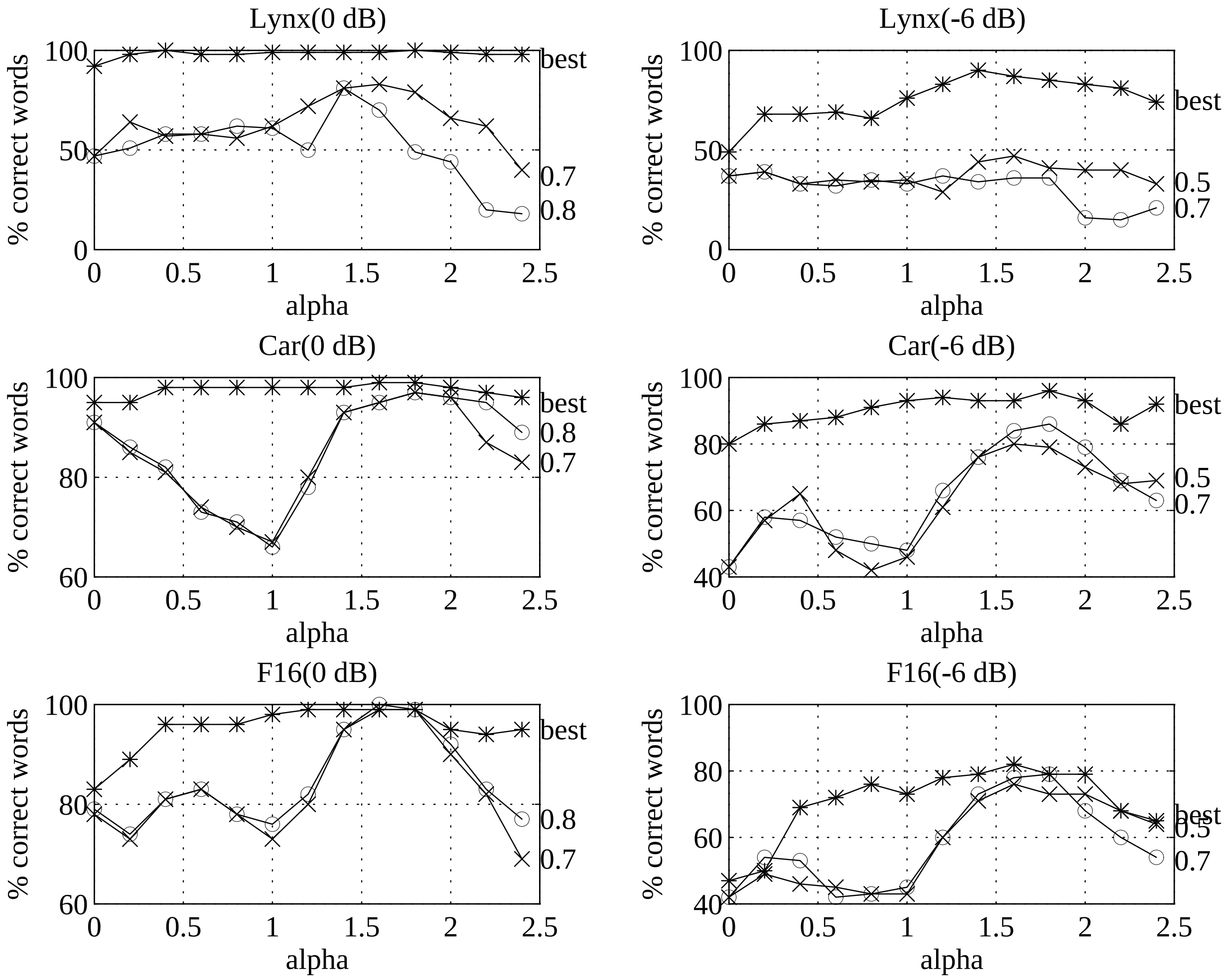

Figure 14: SS-PMS results for bMSS-MMFCC at 0 and -6 dB SNR

5 Comments and Conclusions

By enhancing noisy speech using spectral subtraction, a signal with better features and less variability is obtained, at the expense of signal distortion. In this work, we have proposed the use of adaptation techniques to compensate clean speech models to match the distorted speech produced by a signal enhancer.

First, an ideal signal enhanced adaptation algorithm was simulated by training and testing directly the enhanced speech. The experimental results showed that adapting the models to the enhanced signal can significantly improves recognition on the NOISEX-92 database.

Furthermore, these results suggests that better results should be possible than can be obtained with a noise adaptation scheme alone (such as PMC).

The MSS-MMFCC scheme was defined for when the magnitude spectrum is used for Spectral Subtraction and coding uses the Magnitude spectrum. Similarly, the PSS-PMFCC scheme is when the power spectrum is used for Spectral Subtraction and coding uses the power spectrum. The prefix *b* was used to indicate that SS was performed before the filterbank.

The necessary compensation for the basic MSS-MFCC scheme was developed and it was tested on the NOISEX-92 database. It was shown that the same expressions are valid for both the variants defined above. First, it was shown that the effect of SS on noisy speech can be separated into the effect of the noise on the speech and the distortion caused by the SS. Therefore, compensating the distortion is easily achieved in the linear domain, and the PMC adaptation technique was modified to compensate for the enhanced signal. We refered to this modified approach as the SS-PMC scheme.

The theory behind this adaptation algorithm is based on several assumptions and approximations. In order to compensate for these approximations, the correction term, Δ_μ was weighted by a γ factor, and the best performance was obtained when this factor was larger than 0.7 and smaller than 0.9 for 0 dB SNR, and larger than 0.5 and smaller than 0.8 for -6 dB SNR. The experimental results for the MSS-MMFCC and bMSS-MMFCC schemes were not very satisfactory, but the results for the PSS-PMFCC and bPSS-PMFCC schemes were comparable or better than the results of an ideal adaptation algorithm.

The results obtained for the SS-PMC scheme open the possibility of building a practical recogniser for very noisy environments.

The computational time required for the SS-PMC is practically the same as for the computational time of the spectral subtractor, because, PMC computational time is minimal and the compensation Δ for the distorted speech does not represent a significant computational load.

As already noticed, the Noisex-2 database artificially adds real noise to clean speech, hence there is no Lombard effect. If SS-PMC is applied to real noisy speech, then it may be necessary to compensate for this effect also.

6 Acknowledgment

Thanks to CONACyT for the financial support for this research and to Mark Gales for many fruitful discussions on this work.

A Appendix.

By definition,

$$B(a(\alpha,\beta,\hat{N}),\xi,\psi) = \int_{-\infty}^{a(\alpha,\beta,\hat{N})} Y \frac{1}{Y\sqrt{2\pi}\psi} e^{-\frac{1}{2\psi^2}(ln(Y)-\xi)^2} dY$$

this integral can be solved by the variable substitution $z = ln(Y)$ to give

$$B(a(\alpha,\beta,\hat{N}),\xi,\psi) = \int_{-\infty}^{ln(a(\alpha,\beta,\hat{N}))} \frac{e^z}{\sqrt{2\pi}\psi} e^{-\frac{1}{2\psi^2}(z-\xi)^2} dz$$

which can be re-expressed as follows

$$B(a(\alpha,\beta,\hat{N}),\xi,\psi) = \int_{-\infty}^{ln(a)} \frac{1}{\sqrt{2\pi}\psi} e^{-\frac{1}{2\psi^2}(z^2-2\xi z+\xi^2)+z} dz$$

completing the square gives

$$B(a(\alpha,\beta,\hat{N}),\xi,\psi) = e^{\xi+\psi^2/2} \int_{-\infty}^{ln(a(\alpha,\beta,\hat{N}))} \frac{1}{\sqrt{2\pi}\psi} e^{-\frac{1}{2\psi^2}(z-(\xi+\psi^2))^2} dz$$

hence

$$B(a,\xi,\psi) = e^{\xi+\psi^2/2} G\left(\frac{ln(a(\alpha,\beta,\hat{N}))-(\xi+\psi^2)}{\psi}\right)$$

and if $ln(a(\alpha,\beta,\hat{N})) = \infty$ we obtain

$$E[Y] = e^{\xi+\psi^2/2} \tag{23}$$

Similarly,

$$C(\alpha,\beta,\hat{N},\xi,\psi) = \int_{-\infty}^{a(\alpha,\beta,\hat{N})} Y^2 \frac{1}{Y\sqrt{2\pi}\psi} e^{-\frac{1}{2\psi^2}(ln(Y)-\xi)^2} dY$$

again using the variable substitution $z = ln(Y)$ gives

$$C(\alpha,\beta,\hat{N},\xi,\psi) = \int_{-\infty}^{ln(a(\alpha,\beta,\hat{N}))} \frac{e^{2z}}{\sqrt{2\pi}\psi} e^{-\frac{1}{2\psi^2}(z-\xi)^2} dz$$

and by completing the square we obtain

$$C(\alpha,\beta,\hat{N},\xi,\psi) = e^{2(\xi+\psi^2)} G\left(\frac{ln(a(\alpha,\beta,\hat{N}))-(\xi+2\psi^2)}{\psi}\right)$$

If $ln(a(\alpha,\beta,\hat{N})) = \infty$ we obtain

$$E[Y^2] = e^{2(\xi+\psi^2)}.$$

By using eq. 23, $E[Y^2]$ can also be expressed as follows

$$E[Y^2] = E[Y]e^{\psi^2}$$

References

[1] Beattie, V. & Young, S.J., *Hidden Markov Model State-Based Cepstral Noise Compensation*, Proc. ICSLP, pp. 519-522, Banff, Canada, 1992.

[2] Berstain, A. & Shallom, I., *An hypothesised Wigner Filtering Approach to Noisy Speech Recognition*, Proc. ICASSP, S14.9, Toronto, 1991.

[3] Berouti, M. Schwartz, R., Makhoul, J. *Enhancement of Speech Corrupted by Acoustic Noise*, ICASSP, pp208-211, 1979.

[4] Boll, S.F., *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Trans. on ASSP, Vol. ASSP-27, No. 2, 1979.

[5] Gales, M.J. & Young, S.J., *An improved approach to the Hidden Markov Model Decomposition of Speech and Noise*, ICASSP, 1992.

[6] Gales, M.J. & Young, S.J., *Cepstral Parameter Compensation for HMM Recognition in Noise*, to appear in Speech Communications.

[7] Juang., B.H., *Speech Recognition in Adverse Environments*, Computer Speech and Language, Vol. 5, pp. 275-294, 1991.

[8] Lass, H. & Gottlieb, P., *Probability and Statistics*, Addison Wesley, 1971.

[9] Lim, J.S., *Enhancement and Bandwidth Compression of Noisy Speech*, IEEE Trans. on ASSP, Vol. ASSP-67, No. 12, pp.1586-1604, 1979.

[10] Lockwood, P. & Boudy, J. *Experiments with a Non-linear Spectral Subtractor (NSS), Hidden Markov Models and the Projection, for Robust Speech Recognition in Cars*, Eurospeech, pp. 79-82, 1991.

[11] Lockwood, P., Baillargeat, C., Gillot, J.M., Boudy, J. & Faucon, G., *Noise Reduction for speech enhancement in cars: Non-linear Spectral Subtraction / Kalman Filtering*, Eurospeech, pp. 83-86, 1991.

[12] Mellor, B.A. & Varga, A.R., *Noise Masking in a Transform Domain*, ICASSP, 1993.

[13] Paul, D.B., *A Speaker-stress Resistant HMM Isolated Word Recogniser*, Proceedings ICASSP, pp. 713-716, 1987.

[14] Ruehl, H.W., Dobler, S., Weith, J., Meyer, P., Noll, A., Hamer, H.H. & Piotrowski, H., *Speech Recognition in the Noise Car Environment*, Speech Comm, Vol. 3, pp. 151-167, 1989.

[15] Sharf, L.L. *The SVD and reduced rank signal processing*, Signal Processing, Vol. 25, pp. 113-133, 1991.

[16] Van Campernolle, D., *Noise Adaptation in a Hidden Markov Model Speech Recognition System*, Computer, Speech and Language, Vol. 3, pp.151-167, 1989.

[17] Varga, A.R. & Moore, R.K. *Hidden Markov Model Decomposition of Speech and Noise*, ICASSP, pp. 845-848, 1990.

[18] Varga, A.P. & Pointing, K.M., *Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous Word Recogniser*, ESCA Proc. EUROSPEECH, 1989.

[19] Young, S. J.,*HTK Version 1.4: Reference and User Manual. Cambridge University Engineering Dept., Speech Group, August, 1991.*