# MINIMUM MEAN SQUARE ERROR FILTERING OF NOISY CEPSTRAL COEFFICIENTS WITH APPLICATIONS TO ASR

*Tor André Myrvoll* and Satoshi Nakamura†*

*Department of Telecommunications, NTNU, Trondheim, Norway
†ATR Spoken Language Translation Research Laboratory, Kyoto, Japan
myrvoll@tele.ntnu.no, nakamura@slt.atr.co.jp

**ABSTRACT**

In our previous work ([1]), we investigated a new approach to robust speech recognition. An exact procedure was developed to filter noisy cepstral coefficients in the mean-square-error sense, and it was shown that this method outperformed the well known Vector Taylor Series (VTS) approach, which in turn is based on linear approximations to the non-linear filtering problem. Unfortunately, the procedure presented involved several integral equations with no known closed form solution. Numerical integration techniques was needed, which in turn led to slow performance, and in some cases, numerical problems. In this work we address this problem by using piecewise approximations to the integrands, which in turn yield closed form solutions. The revised procedure is tested on a subset of the Aurora 2 database, and the results are compared with the original numerical integration based approach, as well as VTS.

## 1. INTRODUCTION

It is a well known fact that mismatch between training and test conditions has a significant impact on the performance of an ASR system. Such mismatches can take the form of unknown channels when various microphones or room impulse responses are encountered, or more general mismatches due to speaker variations like accent, gender and speaking rate. Another type of mismatch, and the topic of this work, is the case of additive noise. This is of special concern when an ASR system is to be deployed outside of a controlled environment, for instance in a moving car, an environment crowded by people or in a military environment.

There are several approaches that can be used to compensate for the additive noise. Many approaches fall into the model adaptation category, aiming to tune the model to the new operating conditions. Any well known adaptation method like MLLR[2] or MAP[3] can in principle be used, but when the mismatch is known to be due to some additive noise it is prudent to try to use this information to improve upon the more general adaptation approaches. One such approach is parallel model combination (PMC), in which a noise model is used to update the model parameters using an approximation to the non-linear mixing function[4].

The approach we will follow in this work, is to try to recover the clean features from the noisy features using some estimate of the speech and noise characteristics. Typically the recovery is done by filtering the speech features in the log-spectral domain so that the average mean-square-error between the estimated features and the true features is minimized. The reason that the log-spectral domain is preferred to the spectral domain, is that the spectral features are strictly positive. This positiveness complicates any statistical modeling of the speech, although use of the log-normal distribution for power-spectrum modeling has been utilized[4].

There are several problems that complicates what in principle is a simple filtering approach. As the noise and speech are non-linearly mixed in the log-spectral domain there are no closed form solutions available as far as maximum likelihood estimation of the noise parameters or MSE-estimation of the clean features are concerned. This intractability has encouraged the use of various approximations, often based on a linearizion of the non-linear mixing function[5, 6]. The non-linearity also causes the distribution of the noisy features to be non-Gaussian, even under the idealized assumption that both the speech and the noise has a normal distribution.

In this work we assume that the noise can be modeled as a multivariate normal distribution in the log-spectral domain. We also restrict ourselves to the case of the noise being stationary. For various treatments of feature denoising under non-stationarity see [7, 8, 9].

In the previous work[1], we derived equations according to the expectation-maximization (EM)[10] formulation to estimate the noise parameters under the above assumptions, and ended up with a set of integrals that had to be solved numerically. Solving the integrals using standard numerical techniques turned out to be non-trivial, and analytical solutions were presented for the parts of the integration domain that proved to be ill-behaved. This still left some mostly well-behaved parts of the integration domain to dealt with using numerical integration, resulting in a very slow filtering procedure.

In next section we briefly review our previous results before we present a piecewise approximation scheme that enables closed form solutions to the integrals. Experiments are conducted on the Aurora 2 database and compared to the VTS approach that uses Taylor series to approximate the non-linear function[5].

## 2. NOISE PARAMETER ESTIMATION

When speech is corrupted by additive noise in the time or spectral domain, the effect in the log-spectral domain is a non-linear mixing of the noise and the speech,

$$z_t = x_t + \log\left(1 + e^{n_t - x_t}\right), \tag{1}$$

Part of this work was done while Dr. Myrvoll was a visiting researcher at the ATR Spoken Language Translation Laboratory.

where $x$ is the speech data, $n$ is the noise and $z$ is the corrupted speech, all in the log-spectral domain and all indexed by the time $t$. We follow common practice and assume that the noise $n$ is Gaussian with unknown mean, $\mu_n$, and variance, $\sigma_n$, while the speech is modeled as a mixture distribution with known parameters. One way to alleviate the effect of the noise is to find the minimum mean-square-error estimate of the clean speech. The optimal mean-square-error (MSE) estimator is given by

$$\hat{x} = E_{X|Z}[x], \tag{2}$$

where $E_{X|Z}$ is the conditional expectation operator. In order to perform this filtering we need to estimate the unknown parameters of the noise distribution.

We have previously shown that the noise parameter estimation problem can be cast as a missing data problem, where the corrupted speech $\{z_t\}$ is the incomplete data, and $\{z_t, x_t\}$ are the complete data. This motivates the use of the EM-algorithm [10] to find the noise parameter estimates. We form the auxiliary function

$$\begin{aligned} &Q(\Lambda', \Lambda_i) \\ =&E_{X|Z}\left[\log p_{X,Z}\left(\{x_t, z_t\}_{t=1}^T|\Lambda'\right)\Big|\Lambda_i, \{z_t\}_{t=1}^T\right] \\ =&\int \log p_{X,Z}\left(\{x_t, z_t\}_{t=1}^T|\Lambda'\right) dP_{X|Z}\left(\{x_t\}_{t=1}^T|\{z_t\}_{t=1}^T, \Lambda_i\right), \end{aligned} \tag{3}$$

where $\Lambda = \{\mu_n, \Sigma_n\}$ are the parameters of the noise model.

In [1] it is shown that the maximum of (3) with respect to the noise parameters is obtained using,

$$\hat{\mu}_n = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{z_t} n(z_t, x_t) p_{X|Z}(x_t|z_t, \Lambda_i) dx_t \tag{4}$$

$$\hat{\sigma}_n^2 = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{z_t} \left(n(z_t, x_t) - \hat{\mu}_n\right)^2 p_{X|Z}(x_t|z_t, \Lambda_i) dx_t, \tag{5}$$

where

$$\begin{aligned} p_{X|Z}(x_t|z_t, \Lambda) =& \frac{p_{Z|X}(z_t|x_t, \Lambda) p_X(x_t)}{p_Z(z_t|\Lambda)} \\ =& \frac{\partial n(z_t, x_t)}{\partial z_t} p_N(n(z_t, x_t)|\Lambda) \frac{p_X(x_t)}{p_Z(z_t|\Lambda)}. \end{aligned} \tag{6}$$

and

$$n(z_t, x_t) = \log\left(1 - e^{x_t - z_t}\right) + z_t, \tag{7}$$

There is no known closed form of the probability density function $p_Z(z_t|\Lambda)$, but it can be calculated numerically using the integral

$$p_Z(z_t|\Lambda) = \int_{-\infty}^{z_t} p_{Z|X}(z_t|x_t, \Lambda) p_X(x_t) dx_t. \tag{8}$$

The procedure now goes as follows: Given an initial noise model estimate $\Lambda_0$, calculate the new model estimates $\hat{\mu}_n$ and $\hat{\sigma}_n$. Let $\Lambda_1 = \{\hat{\mu}_n, \hat{\sigma}_n\}$, and maximize the auxiliary function (3) based on this new estimate. Repeat the procedure until convergence is achieved.

## 3. APPROXIMATIONS

### 3.1. Previous Work

To estimate the new model estimates in each new EM iteration, a series of numerical integrals corresponding to equations (4), (5) and (8) has to be calculated. In [1] we showed that this was non-trivial, as the integrands could be very ill-behaved around $x_t = z_t$, and the fact that the integral was performed on a semi-infinite domain. The solution to this problem was to split the integral into three parts – an infinite tail, $(-\infty, x_t^l]$, an $\varepsilon$-ball around $z_t$, $(z_t - \varepsilon, z_t]$, and the rest, $(x_t^l, z_t - \varepsilon]$. If $x_t^l \ll z_t$ and $\varepsilon$ is sufficiently small, good closed-form approximations can be found for the infinite tail, as well as the $\varepsilon$-ball. The details can be found in our previous paper.

The rest of the integral was considered smooth enough to solve numerically. The problem with this approach is the computational complexity of the numerical integration routines that we utilized. We also noticed that some integrals still proved problematic, although the impact on the final estimates seemed negligible.

### 3.2. Piecewise Approximations

We now want to replace the numerical integration with an approximations scheme that is flexible with respect to complexity and accuracy. In the derivations that follow we have stripped down the complexity of the equations somewhat with respect to the actual parameters used. This is done to keep the presentation readable and simple to follow, as the full level of detail will only confuse the issue. We will focus on the integral in equation (8), with extensions to the mean and variance estimates presented later. Writing out the complete expression we have,

$$\begin{aligned} &p_Z(z_t|\Lambda) \\ =&\int_{-\infty}^{z_t} p_{Z|X}(z_t|x_t, \Lambda) p_X(x_t) dx_t \\ =&\int_{-\infty}^{z_t} \frac{\partial n(z_t, x_t)}{\partial z_t} p_N(n(z_t, x_t)|\Lambda) p_X(x_t) dx_t \\ =&\int_{-\infty}^{z_t} \frac{1}{\sqrt{2\pi\sigma_n^2}} \frac{e^{-\frac{1}{2}\left(\frac{\log\left(1-e^{x_t-z_t}\right)+z_t-\mu_n}{\sigma_n}\right)^2}}{1 - e^{x_t - z_t}} p_X(x_t) dx_t \end{aligned} \tag{9}$$

To make the expression more manageable we do the variable substitution $t = 1 - e^{x_t - z_t}$, which in turn yields,

$$\begin{aligned} &p_Z(z_t|\Lambda) \\ =&\int_0^1 \frac{1}{\sqrt{2\pi\sigma_n^2}} \frac{e^{-\frac{1}{2}\left(\frac{\log(t)+z_t-\mu_n}{\sigma_n}\right)^2}}{t} \frac{p_X(\log(1-t)+z_t)}{1-t} dt. \end{aligned} \tag{10}$$

The expression can be simplified even further using the fact that $\frac{1}{t} = e^{-\log(t)}$. This enables us to include the denominator $t$ in the exponential, which in turn can be written,

$$\begin{aligned} &\frac{e^{-\frac{1}{2}\left(\frac{\log(t)+z_t-\mu_n}{\sigma_n}\right)^2}}{t} \\ =&e^{-\frac{1}{2}\left(\frac{\log(t)+z_t-\mu_n+\sigma_n^2}{\sigma_n}\right)^2} e^{-\mu_n+z_t+\frac{\sigma^2}{2}}. \end{aligned} \tag{11}$$

The same reasoning is used to include the term $\frac{1}{1-t}$ in every mixture component in $p_X(\log(1-t)+z_t)$.

The integral is still to our knowledge unsolvable, and in this respect our manipulations haven't gained us much. On the other hand, for every mixture component of $p_X(\log(1-t)+z_t)$, we now have a product of two Gaussians. The key step now is to do piecewise linear approximations of the two functions $\log t$ and $\log(1-t)$. Both functions are smooth over the majority of the $[0,1]$ domain, with the exceptions of $t=0,1$, where $\log t$ and $\log(1-t)$ goes to minus infinity. This is fortunately not a problem, as the approximations referred to in section 3.1 gives good results over small neighborhoods of these critical points.

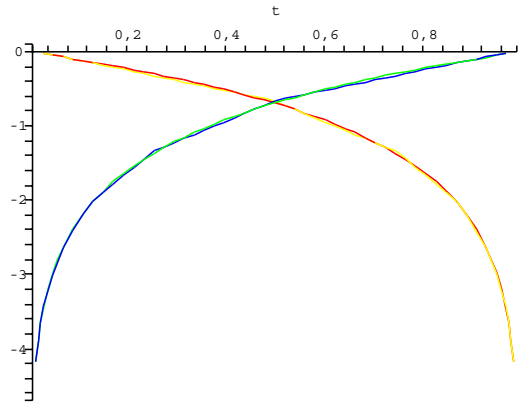


**Fig. 1**. Piecewise linear approximations to the two functions, $\log t$ and $\log(1-t)$. Five non-uniform line segments are used for each functions.

In figure 1 we show piecewise linear approximations to the two functions $\log t$ and $\log(1-t)$, using five non-uniform line segments per function. The lengths of the line segments are powers of $1/2$, from $1/2$ to $1/32$, and the segments becomes shorter as the functions goes towards minus infinity. We see that the logarithms are well approximated using only a few line segments.

Using these approximations we can split the integral into eight non-uniform parts indexed by $i=1..8$, where $\log t$ and $\log(1-t)$ are replaced by $a_i t+b_i$ and $c_i t+d_i$, respectively. In general we can replace any Gaussian $\mathcal{N}(ax+b;\mu,\sigma)$ by $(1/a)\mathcal{N}(x;(\mu-b)/a,\sigma/a)$, so now equation (10) is replaced by the product of two Gaussian distributions. The product of two Gaussian distributions $\mathcal{N}(x;\mu,\sigma)\mathcal{N}(x;\phi,\tau)$ is also a Gaussian distribution scaled by $\alpha$, with mean $\tilde{\mu}$ and variance $\tilde{\sigma}$ equal to,

$$\tilde{\mu} = \frac{\tau^2}{\sigma^2+\tau^2}\mu+\frac{\sigma^2}{\sigma^2+\tau^2}\phi, \quad (12)$$

$$\tilde{\sigma} = \frac{\sigma\tau}{\sqrt{\sigma^2+\tau^2}}. \quad (13)$$

and the scale $\alpha$ equal to

$$\alpha = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}e^{-\frac{1}{2}\frac{(\mu-\phi)^2}{\sigma^2+\tau^2}} \quad (14)$$

This means that we have approximated the highly non-linear integral by a sum of Gaussian integrals which has well-known solutions in the form of error functions. The approximations presented here can be used as-is for the mean and variance integrals as well. Equations (4) and (5) both involves a multiplicative term containing $n(z_t,x_t)$, which corresponds to $\log t+z_t$ after the variable substitution $t=1-e^{x_t-z_t}$. This means that the multiplicative term will be a polynomial in $t$ when we use the piecewise linear approximation to $\log t$. Closed form solutions to the integrals exists in these cases too, if somewhat more analytically involved.

Finally we need to consider the MMSE-filter itself, which has the analytical form

$$\hat{x}_t = \int_{-\infty}^{z_t} x_t p_{X|Z}(x_t|z_t,\Lambda)dx_t. \quad (15)$$

The same variable substitution that we did earlier gives $x_t = log(1-t)+z_t$, which in turn can be exchanged by a piecewise linear approximation. Again we have analytical solutions to the integral available.

The key points of this section can be summed up as follows: Through a change of variables the integrals in equations (4), (5) and (8) can be written as simple products of Gaussian distributions and a polynomial term, all of which are functions of $\log t$ and $\log(1-t)$. Using piecewise linear approximations of $\log t$ and $\log(1-t)$, we can write the integrals as sums of approximative closed form solutions. In the next section we present a set of experiments on the Aurora 2 database.

## 4. EXPERIMENTS

In this section we presents some results that were obtained on a subset of the Aurora 2 database. The results are compared with an implementation of the well-known Vector Taylor Series (VTS) approach[5].

### 4.1. Experimental Setup

In the experiments that follow we used the Aurora 2 database, which consists of spoken digit strings that has been processed to add noise and channel variations. In this work we are only interested in the effect of noisy speech resulting from additive noise only, and so we will only report results on a subset of test set A.

The clean data Hidden Markov Model used for recognition was built with HTK using the standard scripts provided with the Aurora 2 database. We used 12 cepstral coefficients, $c_1$-$c_{12}$, together with $c_0$ as a replacement for energy. Together with the $\Delta$- and $\Delta^2$-features this makes for a 39-dimensional feature vector.

We used a Gaussian Mixture Model (GMM) with 64 components as our clean speech model. The GMM was trained on the same training data as the HMM, but with the silence parts excluded. Also, the model was trained on the 13 dimensional cepstral data, and the model was then transformed to the log-spectral domain using an inverse cosine transformation.

The two logarithmic functions $\log t$ and $\log(1-t)$ were approximated by piecewise linear functions on a non-uniform segmentation of the $[0,1]$ domain. The linear approximation in each segment was estimated in terms of first order Chebyshev polynomials using the Maple software package[11]. The linear segments are tabulated in table 1, although the full accuracy used in the experiments has been trimmed to better fit the table format.

### 4.2. Results

The VTS approach and the optimal approach were both tested on the speech corrupted with subway noise from Aurora 2 test set A.

| Domain | $\log t$ | $\log(1-t)$ |
|---|---|---|
| $[\frac{1}{32}, \frac{1}{16}]$ | $-4.118 + 21.961t$ | $0.02645 - 1.372t$ |
| $[\frac{1}{16}, \frac{1}{8}]$ | $-3.425 + 10.980t$ | $0.02645 - 1.372t$ |
| $[\frac{1}{8}, \frac{1}{4}]$ | $-2.732 + 5.4903t$ | $0.02645 - 1.372t$ |
| $[\frac{1}{4}, \frac{1}{2}]$ | $-2.039 + 2.7451t$ | $0.02645 - 1.372t$ |
| $[\frac{1}{2}, \frac{3}{4}]$ | $-1.346 + 1.3725t$ | $.7058 - 2.745t$ |
| $[\frac{3}{4}, \frac{7}{8}]$ | $-1.346 + 1.3725t$ | $2.757 - 5.490t$ |
| $[\frac{7}{8}, \frac{15}{16}]$ | $-1.346 + 1.3725t$ | $7.555 - 10.98t$ |
| $[\frac{15}{16}, \frac{31}{32}]$ | $-1.346 + 1.3725t$ | $17.84 - 21.96t$ |

**Table 1**. Linear approximations of the logarithmic functions $\log t$ and $\log(1-t)$.

The data has seven different signal-to-noise ratios from clean to -5 dB. The results for this noise condition is presented in table 2. The original numerical integration approach is referred to as "NI" in these experiments, and the new approximation is referred to as "PLA".

We see that the optimal filtering outperforms both the baseline, and the improvement that VTS gives, for the most serious noise conditions. The improvement of PLA over NI indicates that the problem of numerical stability was graver than previously thought. The speed was also significantly improved, sometimes by a factor of three. For higher SNRs VTS outperforms our proposed method, and one conjecture is that the VTS formulation more closely resembles a standard adaptation approach due to the approximations made, and in that case VTS is able to compensate for other variations than just the additive noise. The slight performance degradation as compared with the baseline for clean speech and 20 dB SNR can be explained by the approximation made around $t = 0$. Using more line segments and thereby shrinking the area around this extreme point is will improve the approximation and is likely to improve the performance. Clearly further investigation into this matter is needed.

| *Recording inside a subway* | | | | |
|---|---|---|---|---|
| SNR | Baseline | VTS | NI | PLA |
| -5dB | 10,72 | 11,61 | 18,30 | 20,08 |
| 0dB | 20,94 | 29,44 | 35,40 | 43,78 |
| 5dB | 45,26 | 59,75 | 68,28 | 72,34 |
| 10dB | 73,87 | 84,34 | 86,92 | 88,49 |
| 15dB | 92,08 | 95,30 | 93,83 | 94,14 |
| 20dB | 96,90 | 97,05 | 96,44 | 96,53 |
| $\infty$dB | 99,11 | 99,08 | 98,53 | 98,50 |
| Average | 62,70 | 68,08 | 71,10 | 73,41 |

**Table 2**. Recognition results for the filtered subway recordings. The numbers reflect the number of correct words, or accuracy, of the recognizer. Infinite dB refers to clean speech.

## 5. CONCLUSIONS

We have in this paper presented work done on optimal filtering of cepstral coefficients using an approximation to the numerical integration techniques introduced earlier. We have shown that this approach clearly outperforms both the original numerical integration method, as well as VTS, a method that relies on approximating the filtering problem by linearizing the non-linear mixing function. The performance that is achieved is promising, and we believe that it warrants further investigation into the use of more accurate modeling of the cepstral filtering problem. Further research into non-stationary noise and online estimation techniques is a natural extension of this work.

## 6. ACKNOWLEDGMENTS

This research was supported in part by the Telecommunications Advancement Organization of Japan. Dr. Myrvoll was also partly supported by the Department of Telecommunications, Norwegian University of Science and Technology, and the Norwegian Research Council through the BRAGE program.
(http://www.tele.ntnu.no/projects/brage/index.php).

## 7. REFERENCES

[1] Tor André Myrvoll and Satoshi Nakamura, "Optimal filtering of noisy ceptral coefficients for robust ASR," St. Thomas, U.S. Virgin Islands, Nov.-Dec. 2003, IEEE.

[2] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[3] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, April 1994.

[4] M. F. J. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communications*, vol. 12, no. 3, pp. 231–239, July 1993.

[5] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE ICASSP-96*, Atlanta,Georgia, May 1996, vol. 2, pp. 733–736.

[6] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proc. IEEE ICASSP-97*, Munich, Germany, Apr. 1997, vol. 2, pp. 835–838.

[7] Nam Soo Kim, "Nonstationary environment compensation based on sequential estimation," *IEEE Signal Processing Letters*, vol. 5, no. 3, pp. 57–59, March 1998.

[8] K. Yao and S. Nakamura, "Sequential noise compensation by sequential Monte Carlo method," in *Neural Information Processing Systems*, Vancouver, British Columbia, Canada, Dec. 2001.

[9] B. J. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *EuroSpeech '01*, Aalborg, Denmark, Sep. 2001, pp. 901–904.

[10] A. J. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, 1977.

[11] Andre Heck, *Introduction to Maple*, Springer-Verlag Berlin and Heidelberg GmbH & Co. KG, 2003.