Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 403–416
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

MMSE estimation of log-filterbank energies for robust speech recognition

Anthony Stark, Kuldip Paliwal *

Signal Processing Laboratory, Griffith University, Nathan Campus, Brisbane QLD 4111, Australia

Received 15 June 2010; received in revised form 27 September 2010; accepted 3 November 2010
Available online 21 December 2010

Abstract

In this paper, we derive a minimum mean square error log-filterbank energy estimator for environment-robust automatic speech recognition. While several such estimators exist within the literature, most involve trade-offs between simplifications of the log-filterbank noise distortion model and analytical tractability. To avoid this limitation, we extend a well known spectral domain noise distortion model for use in the log-filterbank energy domain. To do this, several mathematical transformations are developed to transform spectral domain models into filterbank and log-filterbank energy models. As a result, a new estimator is developed that allows for robust estimation of both log-filterbank energies and subsequent Mel-frequency cepstral coefficients. The proposed estimator is evaluated over the Aurora2, and RM speech recognition tasks, with results showing a significant reduction in word recognition error over both baseline results and several competing estimators.

Keywords: Robust speech recognition; MMSE estimation; Speech enhancement methods

1. Introduction

State-of-the-art automatic speech recognition (ASR) can exhibit impressive recognition performance under laboratory conditions. Unfortunately, performance tends to degrade substantially when ASR is used in real world environments. This degradation is caused by acoustic model mismatch. Here, we use the term mismatch to describe any difference between the acoustic environment the ASR system was trained on, and the acoustic environment the ASR system is actually deployed in. Mismatch can include additive background noise, echoes, transmission channel effects, inter-speaker variability and intra-speaker variability. Taken together, such effects can rapidly reduce recognition accuracy to unacceptably low levels.

For this paper, we address the problem of additive background noise robustness. In the literature, several approaches have previously been proposed. Most fall under the following general categories: robust feature selection and extraction (Davis and Mermelstein, 1990; Hermansky, 1990), speech enhancement (Lathoud et al., 2005; Ephraim and Trees, 1991; Gemello et al., 2006; Hermus et al., 2007; Fujimoto and Ariki, 2000), model adaptation (Gales, 1995; Acero et al., 2000), model-based feature enhancement (Stouten, 2006; Moreno, 1996), missing feature theory (Raj and Stern, 2005; Cooke et al., 2000; Barker et al., 2000) and multistyle training (Deng et al., 2000).

In this paper, we focus on the problem of estimating robust Mel-frequency cepstral coefficients (MFCCs). In particular, we investigate the stochastic estimation of a clean speech MFCC vector from speech that has been corrupted with additive noise. Under the additive noise assumption, a noisy speech signal $y(n)$ is given by

$$y(n) = x(n) + d(n), \quad 0 \leqslant n < N, \tag{1}$$

* Corresponding author.
E-mail addresses: a.stark@griffith.edu.au (A. Stark), k.paliwal@griffith.edu.au (K. Paliwal).
URL: http://maxwell.me.gu.edu.au/spl/ (K. Paliwal).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.11.004

where $x(n)$ and $d(n)$ are the clean speech and noise signal, respectively. Since speech is often assumed to be quasi-stationary over short-time (20–40 ms) intervals, it is typically decomposed with framing. Here, the mth noisy speech frame can be given as

$$\boldsymbol{y}_m = \boldsymbol{x}_m + \boldsymbol{n}_m, \tag{2}$$

where $\boldsymbol{y}_m = [y(mS), y(mS+1), \ldots, y(mS+L-1)]^T$, L is the analysis frame length and S is the analysis frame shift. After discrete short-time Fourier transform (DSTFT) analysis (Rabiner and Schafer, 1978) of (2), we then have the following relationship

$$\boldsymbol{Y}_m = \boldsymbol{X}_m + \boldsymbol{D}_m, \tag{3}$$

where $\boldsymbol{Y}_m, \boldsymbol{X}_m, \boldsymbol{D}_m \in \mathbb{C}^{K\times 1}$ are the noisy speech, clean speech and noise spectral domain vectors (for the mth DSTFT analysis frame), respectively. For notational convenience, we drop the frame index m and dependence on this subscript is implicitly assumed henceforth.

Given the observed noisy speech vector, the goal of a minimum-mean-square error (MMSE) MFCC estimator is the determination of estimate $\hat{\boldsymbol{c}}$, where

$$\hat{\boldsymbol{c}} = E[\boldsymbol{c}|\boldsymbol{Y}], \tag{4}$$

where $\boldsymbol{c}$ is the clean speech MFCC vector, $E[\cdot]$ is the expectation operator and $\hat{\boldsymbol{c}}$ is the estimate that minimizes the mean-square-error to the true clean speech MFCC vector $\boldsymbol{c}$.

While the spectral domain noise distortion model (3) is straightforward, the estimation (4) is not. This is due to the highly non-linear relationship between spectral-domain speech and the MFCC vector. Given a spectral domain speech vector $\boldsymbol{Y}$, several intermediate variables must first be calculated: $\boldsymbol{e}$ – spectral energies, $\boldsymbol{E}$ – filterbank energies and $\boldsymbol{L}$ – log-filterbank energies. Fig. 1 shows the operations required for converting spectral-domain speech into an MFCC vector.

Instead of directly estimating the MFCC vector, we may focus our attention on one of the intermediate feature sets – namely log-filterbank energies. The MMSE log-filterbank estimate $\widehat{\boldsymbol{L}}$ is given by

$$\widehat{\boldsymbol{L}} = E[\boldsymbol{L}|\boldsymbol{Y}]. \tag{5}$$

where $\boldsymbol{L}$ is the clean speech log-filterbank energy vector. Since MFCCs and log-filterbank energies are linearly related, given $\widehat{\boldsymbol{L}}$ it is easy to find the MMSE MFCC estimate $\hat{\boldsymbol{c}}$

$$\hat{\boldsymbol{c}} = \boldsymbol{C}\widehat{\boldsymbol{L}}, \tag{6}$$

where $\boldsymbol{C}$ is the discrete cosine transform matrix. Thus, the core MFCC estimation problem now becomes a log-filterbank energy estimation problem. Unfortunately, a highly non-linear relationship persists between the spectral domain speech and its corresponding log-filterbank energy vector. Several strategies have been adopted in past literature to address this issue, including forced linearization of the noise model (Moreno, 1996 ; Stouten, 2006), numerical integration (of an analytically intractable model) (Erell and Weintraub, 1993) and development of simpler (and more tractable) noise distortion models (Yu et al., 2008; Indrebo et al., 2008). Each of the aforementioned methods is suboptimal in some manner, typically offering a trade-off between noise model simplification and computational tractability.

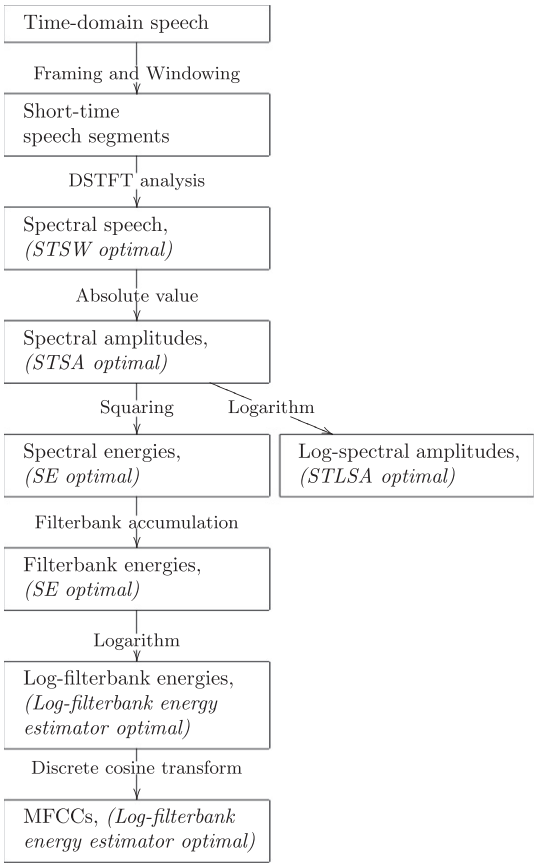

Fig. 1. Overview of the computation required for converting a time-domain frame of speech into an MFCC vector. MMSE optimality is achieved by several common spectral estimators at various intermediate stages of the MFCC derivation.

In many cases, a speech enhancement algorithm from the human listening domain is carried over to the machine recognition domain. Methods such as the short-time spectral amplitude (STSA) estimator and the short-time log-spectral amplitude estimator (STLSA) have commonly been used to reduce the effects of noise from MFCC features. However, the mathematical optimality of these estimators do not provide an exact match with the objectives of ASR – that is, reduction of error within the MFCC/log-filterbank domain. Fig. 1 highlights feature stages where the STSA, STLSA, short-time spectral Wiener (STSW) and short-time spectral energy (SE) estimators are optimal (in the MMSE sense). While none of the aforementioned estimators is strictly optimal (in the log-filterbank MMSE sense), they are all closely related. Because of this, we examine this class of estimators in greater detail, examining their relationship to an MMSE log-filterbank energy estimator.

In this paper, we have two objectives: (1) the extension of the spectral estimation framework to derive an MMSE log-filterbank energy estimator, and (2) quantify its relationship to the other spectral estimators and determine whether significant improvements in robustness may be gained with its use in ASR.

The rest of this paper is organized as follows. In Section 2, we provide a brief review of the short-time spectral noise distortion framework as used by the STSA and STLSA estimators. This framework can be used to derive the posterior probability density function (PDF) $p(\boldsymbol{e}|\boldsymbol{Y})$ for spectral energies $\boldsymbol{e}$. In Section 3, we extend the spectral framework – detailing the mathematical transformations required for converting the spectral domain models into log-filterbank energy domain models. In Section 4, we evaluate the MMSE log-filterbank energy estimator on the Aurora2 and RM recognition tasks. Lastly, in Section 5 we conclude the paper.

2. Statistical framework for MMSE short-time spectral estimation

Under the assumed noise model, individual DSTFT coefficients of the noisy speech signal can be given as

$$Y_k = X_k + D_k, \tag{7}$$

where X_k and D_k are the DSTFT expansion coefficients for the kth discrete frequency bin of the clean speech signal and noise signals, respectively. Under the statistical framework developed by Ephraim and Malah (1984), Ephraim and Malah (1985), individual DSTFT expansion coefficients X_k and D_k are assumed to be independent complex zero-mean Gaussian random variables (RVs), with expected power $\lambda_{X_k} = E[|X_k|^2]$ and $\lambda_{D_k} = E[|D_k|^2]$. Detailed justification of this statistical assumption may be found in (Ephraim and Malah, 1984, Loizou, 2007). Using this model, several spectral estimators have been derived. These include the short-time spectral Wiener (STSW) (Loizou, 2007), MMSE STSA (Ephraim and Malah, 1984) and MMSE STLSA (Ephraim and Malah, 1985) estimators. However, we may also use this framework to develop models for spectral energies – the intermediate variable required by our models (see Fig. 1). Under this framework, the posterior PDF of individual spectral amplitudes, $p(A_k|\boldsymbol{Y})$ can be given by the following Rice distribution (Ephraim and Malah, 1984)

$$p(A_k|\boldsymbol{Y}) = p(A_k|Y_k) = \frac{A_k \exp\left(\frac{-[A_k]^2}{\lambda_k}\right) I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}A_k\right)}{\int_0^\infty \tau \exp\left(\frac{-\tau^2}{\lambda_k}\right) I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}\tau\right) d\tau}, \tag{8}$$

where $A_k = |X_k|$ is the clean speech spectral amplitude, $I_0(\cdot)$ is the zeroth order modified Bessel function, and

$$\lambda_k = \frac{\lambda_{X_k}\lambda_{D_k}}{\lambda_{X_k} + \lambda_{D_k}}, \tag{9}$$

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k, \tag{10}$$

$$\xi_k = \frac{\lambda_{X_k}}{\lambda_{D_k}}, \tag{11}$$

$$\gamma_k = \frac{|Y_k|^2}{\lambda_{D_k}}, \tag{12}$$

where ξ and γ are interpreted as the *a priori* signal to noise ratio (SNR) and *a posteriori* SNR, respectively. With some algebraic manipulation, (8) may be converted to the posterior spectral energy PDF[1]

$$p(e_k|\boldsymbol{Y}) = \frac{\exp\left(\frac{-e_k}{\lambda_k}\right) I_0\left(2\sqrt{\frac{v_k e_k}{\lambda_k}}\right)}{\lambda_k \exp(v_k)}, \tag{13}$$

where clean speech spectral energy $e_k = |X_k|^2$. To describe the spectral energy variable, it is useful to derive the conditioned spectral energy mean and variance. Using (13), the conditioned expectation of individual spectral energies can be given as[2] (Gradshteyn and Ryzhik, 2007)

$$\begin{aligned} E[e_k|\boldsymbol{Y}] = \hat{e}_k &= \int_0^\infty e_k \cdot p(e_k|\boldsymbol{Y}) de_k \\ &= \left(\frac{\xi_k}{1+\xi_k}\right)^2 \left(1 + \frac{1+\xi_k}{\xi_k \gamma_k}\right) |Y_k|^2. \end{aligned} \tag{14}$$

We may also solve for the conditioned spectral energy variance. Diagonal covariance terms are given by[2]

$$\begin{aligned} E\left[[e_k - \hat{e}_k]^2 | \boldsymbol{Y}\right] &= \Sigma_e(k,k) \\ &= \int_0^\infty [e_k]^2 \cdot p(e_k|\boldsymbol{Y}) de_k - [\hat{e}_k]^2 \\ &= [\hat{e}_k]^2 - \left(\frac{\xi_k}{1+\xi_k}\right)^4 |Y_k|^4. \end{aligned} \tag{15}$$

Since individual Fourier expansion coefficients are assumed to be independent, off-diagonal covariances will be zero; i.e.,

$$\Sigma_e(k,k') = 0, \quad \text{for} \quad k \neq k'. \tag{16}$$

2.1. Estimation of a priori SNR ξ

The practical application of the spectral energy estimator is dependent on the estimation of SNR parameters ξ and γ. While γ requires only the noise power λ_D to be estimated, additional care must be taken when estimating the *a priori* SNR. This is because calculation of spectral energy (14) and variance (15) is particularly sensitive to ξ. For the STLSA and STSA estimators, estimation of ξ is generally performed with the decision-directed framework (Ephraim and Malah, 1984). Here, an estimate of $\xi(m,k)$ for the mth frame and kth frequency bin is given by

$$\xi(m,k) = \rho \frac{\hat{e}(m-1,k)}{\lambda_D(m-1,k)} + (1-\rho)P[\gamma(m,k) - 1], \tag{17}$$

[1] Further detail is given in Appendix A.
[2] Further detail is given in Appendix B.

where mixing constant $\rho \approx 0.98$, and

$$P[x] = \begin{cases} x & \text{if } x \geqslant 0, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The inclusion of estimated value $\hat{e}(m-1,k)$ in (17) introduces positive feedback into the ξ estimation algorithm. This is of particular concern for the spectral energy estimator, as it has a relatively mild spectral amplitude gain (Loizou, 2007) (w.r.t the MMSE STSA and STLSA estimators). As a result, given a poor noise estimate, the spectral energy estimator is especially prone to producing residual noise within the estimate $\hat{e}(m-1,k)$. In the decision-directed approach, residual noise artificially inflates the ξ estimate of following analysis frames, leading to less suppression, and more residual noise. To alleviate this, we may utilize the speech presence uncertainty (SPU) framework (McAulay and Malpass, 1980). Under SPU, an *a posteriori* probability of speech presence φ_k can be given by Ephraim and Malah (1984)

$$\varphi_k = \frac{\Lambda_k}{1+\Lambda_k}, \tag{19}$$

where Λ_k is the generalized speech presence ratio

$$\Lambda_k = \frac{1-q_k}{q_k} \cdot \frac{\exp(v_k)}{1+\xi_k}. \tag{20}$$

The term q_k is a tuned parameter that indicates the *a priori* probability of speech absence. The value of q_k also determines the aggressiveness of the SPU. When $q_k = 0$, the effects of SPU are nullified. When its value is increased ($0 < q_k < 1$), the effect of SPU is also increased.

A number of methods exist for determining q_k. For the STSA estimator, it is common to use a static value of 0.3 (Loizou, 2007). Subsequent proposals geared mostly toward the STLSA estimator are recursive, data driven procedures (Cohen, 2002; Malah et al., 1999; Soon et al., 1999).

Given the *a posteriori* probability of speech presence φ_k, SPU updated estimates for spectral energies can be given as

$$\hat{e}'_k = \varphi_k \hat{e}_k = \varphi_k \lambda_k [1+v_k]. \tag{21}$$

When φ_k is small, it is a good indication that the *a priori* SNR has been overestimated. In our work, we use the SPU modified estimate $\hat{e}'_k$ to derive a more appropriate value for the *a priori* SNR. Using $\hat{e}'_k$ as the desired estimate, the spectral energy estimator (14) can be rearranged to give

$$\xi'_k = -\left(\frac{2|Y_k|^2}{\lambda_{D_k} - \sqrt{[\lambda_{D_k}]^2 + 4|Y_k|^2 \hat{e}'_k}} + 1 \right)^{-1}. \tag{22}$$

When speech is surely present ($\varphi_k = 1$), it can be shown that $\hat{e}'_k = \hat{e}_k$, and thus $\xi'_k = \xi_k$. As the value of φ_k decreases, the value of ξ'_k is reduced to compensate. Using the updated ξ'_k, the SPU updated spectral energy variance can be given as

$$\Sigma_{e'}(k,k) = [\hat{e}'_k]^2 - \left(\frac{\xi'_k}{1+\xi'_k} \right)^4 |Y_k|^4. \tag{23}$$

For further detail on the spectral estimation framework and SPU, the reader is referred to Loizou (2007).

3. Models for filterbank and log-filterbank variables

In this section, we develop models for filterbank and log-filterbank energy variables. Since filterbank energies are linearly related to spectral energies, we may estimate the conditioned filterbank mean and variances as follows:

$$E[E_q|\boldsymbol{Y}] = \widehat{E}_q = \sum_k H(q,k)\hat{e}'_k \tag{24}$$

$$E[(E_q - \widehat{E}_q)^2|\boldsymbol{Y}] = \Sigma_E(q,q) = \sum_k [H(q,k)]^2 \Sigma'_e(k,k), \tag{25}$$

where $H(q,k)$ is the filterbank gain for the kth frequency bin and qth filterbank.

To determine the actual structure of the PDF $p(E_q|\boldsymbol{Y})$, we would ideally convolve individual, filterbank scaled spectral energy PDFs (13) together. Unfortunately such a method leads to a complicated closed form solution. However, the resulting PDF does appear to be well approximated by the gamma distribution. Thus, given estimates for the filterbank mean and variance, the filterbank variable E_q can be described by the following gamma distribution

$$p(E_q|\boldsymbol{Y}) = \frac{[E_q]^{\alpha_q - 1} \exp\left(-\frac{E_q}{\beta_q}\right)}{\beta_q^{\alpha_q} \Gamma(\alpha_q)}, \tag{26}$$

where $\Gamma(\cdot)$ is the gamma function. The shape parameter α_q, and scale parameter β_q can be found using the method of moments:

$$\alpha_q = \frac{[\widehat{E}_q]^2}{\Sigma_E(q,q)}, \tag{27}$$

$$\beta_q = \frac{\Sigma_E(q,q)}{\widehat{E}_q}. \tag{28}$$

Further detail of the gamma PDF approximation is given in Section 3.1. Using (26), we may also define the posterior PDF for the log-filterbank variable[3] $L_q = \log E_q$

$$p(L_q|\boldsymbol{Y}) = \frac{\exp\left(\alpha_q [L_q - \log\beta_q]\right)}{\exp\left(\exp\left(L_q - \log\beta_q\right)\right) \cdot \Gamma(\alpha_q)}. \tag{29}$$

The MMSE log-filterbank energy $\widehat{L}_q$ is given as[4] (Gradshteyn and Ryzhik, 2007)

$$E\left[\log E_q|\boldsymbol{Y}\right] = \widehat{L}_q = \log \widehat{E}_q - \log(\alpha_q) + \Psi_0(\alpha_q), \tag{30}$$

where $\Psi_0(\cdot)$ is the digamma function. We may use an efficient series expansion of the digamma function (Spouge, 1994) for calculating the log-filterbank mean. The MMSE log-filterbank energy can be estimated as

[3] Further detail is given in Appendix A.
[4] Further detail is given in Appendix B.

$$\widehat{L}_q \approx \log \widehat{E}_q - \frac{0.500}{(\alpha_q + 0.045)} - \frac{0.108}{(\alpha_q + 0.045)^2}, \quad \alpha_q \geqslant 1. \tag{31}$$

Despite being only a second order expansion, the above approximation is accurate to within 0.031% relative error over the $1 \leqslant \alpha_q < 10^9$ interval.

It can be shown that the maximum *a posteriori* (MAP) estimate for log-filterbank energies has an even simpler solution[4]

$$\widehat{L}_{q-MAP} = \arg\max_{L_q} \left[p(L_q|biY)\right] = \log\left(\alpha_q \beta_q\right) = \log \widehat{E}_q. \tag{32}$$

From (30) and (32), we can see that the MMSE and MAP estimates are closely related. Here, the two estimators differ by a term Δ_q, given by

$$\Delta_q = \widehat{L}_{q-MAP} - \widehat{L}_q = \log \alpha_q - \Psi_0(\alpha_q). \tag{33}$$

Aside from having a simple analytic formula, the MAP estimate (32) is of interest because it is equivalent to the filterbank energy estimator; that is, both are MMSE optimal in the filterbank energy domains. This means the MMSE spectral energy estimator (see Section 2) also happens to the MAP log-filterbank estimator. We should point out that MAP optimality (unlike MMSE optimality) does not carry through to the MFCC domain. The MAP estimator itself is also related to several other common estimators. Firstly, the MAP estimate is obtained if we implicitly ignore filterbank variance (assume it to be zero). Secondly, the MAP estimate is equivalent to the estimate obtained via vector Taylor series expansion (Moreno, 1996) (for zeroth and first order expansions of $\log E_q$ pivoted on $\widehat{E}_q$).

The effect of α on the difference term is shown in Fig. 2. Here we can see that the MAP estimate is always larger than the MMSE estimate. Such a result is consistent with Jensen's inequality: i.e., since the logarithm is a concave operator, we have the following relationship between the MMSE filterbank and MMSE log-filterbank estimators

$$\log E[\log E_q|\boldsymbol{Y}] \geqslant E[\log E_q|\boldsymbol{Y}]. \tag{34}$$

When $\alpha \gg 1$, the difference term (33) tends toward zero. This suggests the MAP and MMSE estimators should be equivalent at higher values of α. We may further note that α cannot take values below 1.[5] As a result, substituting $\alpha_q = 1$ into (33) we find that the maximum difference is given as $\Delta_{\max} \approx 0.577$ (the Euler–Mascheroni constant). Further investigation into the behavior of α is given in the Section 3.1.

3.1. Empirical analysis of the filterbank approximations

In this subsection, we first examine the use of the gamma PDF for modeling the filterbank energy variable. We then evaluate the performance of the resulting MMSE log filterbank energy estimator with respect to other estimators on synthetic data. Finally, we compare these estimators using real speech data.

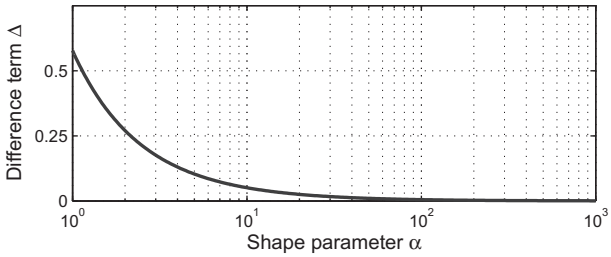

Fig. 2. Effect of shape parameter α on the log-filterbank energy MAP estimates.

In order to examine the use of the gamma PDF for modeling the filterbank energy variable, we use synthetic data obtained by simulating a single filterbank, consisting of a few spectral bins with known λ_{X_k} and λ_{D_k}. Here, the statistical framework underpinning earlier spectral estimators (e.g., STSA, STLSA) are assumed to be correct. Using the λ_{X_k} and λ_{D_k} values, we generate a parameter set of X_k and D_k, (where X_k and D_k are realizations of complex zero-mean Gaussian RVs, as dictated by the framework detailed earlier). Using (13), we can then determine the posterior PDF $p(e_k|Y_k)$ for each bin. Without loss of generality, we assume the filterbank gain for each bin is unity. This means the true filterbank PDF can be estimated as the discrete convolution of individual (discretized) spectral energy PDFs. To determine how close this PDF is to the gamma PDF approximation, we must calculate gamma PDF parameters α and β. Filterbank energy mean and variance can be given as the summation of spectral energy means (14) and variances (15), respectively. Using these estimates, gamma PDF parameters α and β, is calculated from (27) and (28), respectively. For completeness, we also tried fitting Gaussian, log-Gaussian/normal and chi-square distributions to the filterbank PDF. These PDFs are fitted with an equivalent method of moments.

To determine the quality of fit, we use a chi-square statistic,

$$\chi^2 = \sum_j \frac{(p_E[i] - p_T[i])^2}{p_T[x]}, \tag{35}$$

where $p_E[i]$ is the discretized form of the approximating PDF and $p_T[i]$ is the discretized form of the true PDF. To calculate the χ^2 statistic, we used discretized bins where $p_T[i] > 0.0001$.

Table 1 lists the PDF fitting results for six parameter sets. For simulations A1–A3, we simulate a 3-bin filterbank at −10 dB, 0 dB and 10 dB SNR, respectively. For simulations B1–B3, we simulate a 6-bin filterbank at −10 dB, 0 dB and 10 dB SNR, respectively. We can see from Table 1 that the gamma PDF gives a much better fit than the other PDFs for all of the simulations. While the quality of fit changes from simulation to simulation, we notice that the quality of the gamma PDF fit (over multiple fittings) is consistently very good for larger values of α ($\alpha > 10$). Large

[5] See Appendix C.

Table 1
Analysis of the filterbank energy probability density function shape.

Set	Simulation parameters							PDF χ^2 fit			
	$\boldsymbol{\lambda}_X$	$\boldsymbol{\lambda}_D$	$\boldsymbol{X}$	$\boldsymbol{D}$	E_y	α	β	Gamma	Gaussian	Log-Norm.	Chi-square
A1	$\begin{bmatrix}9\\25\\16\end{bmatrix}$	$\begin{bmatrix}1\\3\\1\end{bmatrix}$	$\begin{bmatrix}0.604,-2.547i\\-1.783,-3.284i\\3.647,-0.651i\end{bmatrix}$	$\begin{bmatrix}0.665,0.717i\\-0.208,-0.044i\\0.875,0.027i\end{bmatrix}$	40.836	13.208	2.952	0.003	1.083	0.221	1.150
A2	$\begin{bmatrix}9\\25\\16\end{bmatrix}$	$\begin{bmatrix}10\\30\\10\end{bmatrix}$	$\begin{bmatrix}0.604,-2.547i\\-1.783,-3.284i\\3.647,-0.651i\end{bmatrix}$	$\begin{bmatrix}2.103,2.266i\\-0.658,-0.139i\\2.769,0.086i\end{bmatrix}$	66.553	3.739	12.183	0.010	3.018	0.792	16.748
A3	$\begin{bmatrix}9\\25\\16\end{bmatrix}$	$\begin{bmatrix}100\\300\\100\end{bmatrix}$	$\begin{bmatrix}0.604,-2.547i\\-1.783,-3.284i\\3.647,-0.651i\end{bmatrix}$	$\begin{bmatrix}6.650,7.166i\\-2.080,-0.439i\\8.755,0.271i\end{bmatrix}$	256.698	2.674	18.222	0.034	3.894	0.828	24.515
B1	$\begin{bmatrix}4\\13\\33\\45\\15\\10\end{bmatrix}$	$\begin{bmatrix}1\\2\\4\\2\\2\\1\end{bmatrix}$	$\begin{bmatrix}-1.310,1.438i\\-1.416,1.451i\\6.933,-0.548i\\-6.879,1.481i\\4.957,-0.368i\\0.920,-2.304i\end{bmatrix}$	$\begin{bmatrix}-0.349,1.346i\\-1.642,-0.097i\\2.763,1.383i\\0.696,1.364i\\-0.830,0.860i\\-0.987,-0.755i\end{bmatrix}$	189.362	32.897	5.010	0.006	0.243	0.114	5.050
B2	$\begin{bmatrix}4\\13\\33\\45\\15\\10\end{bmatrix}$	$\begin{bmatrix}10\\20\\40\\20\\20\\10\end{bmatrix}$	$\begin{bmatrix}-1.310,1.438i\\-1.416,1.451i\\6.933,-0.548i\\-6.879,1.481i\\4.957,-0.368i\\0.920,-2.304i\end{bmatrix}$	$\begin{bmatrix}-1.102,4.257i\\-5.194,-0.307i\\8.739,4.374i\\2.202,4.313i\\-2.625,2.720i\\-3.120,-2.388i\end{bmatrix}$	436.753	6.703	23.098	0.002	1.269	0.430	29.729
B3	$\begin{bmatrix}4\\13\\33\\45\\15\\10\end{bmatrix}$	$\begin{bmatrix}100\\200\\400\\200\\200\\100\end{bmatrix}$	$\begin{bmatrix}-1.310,1.438i\\-1.416,1.451i\\6.933,-0.548i\\-6.879,1.481i\\4.957,-0.368i\\0.920,-2.304i\end{bmatrix}$	$\begin{bmatrix}-3.486,13.460i\\-16.424,-0.971i\\27.634,13.832i\\6.963,13.639i\\-8.302,8.602i\\-9.866,-7.552i\end{bmatrix}$	2419.570	4.093	30.634	0.135	2.833	0.314	36.699

α filterbank values typically occur under two circumstances: (1) high SNR parameters are used, or (2) a large number of spectral bins are summed into the filterbank. In the opposite circumstances (low SNR, low spectral bin count), the gamma PDF fit is less consistent, sometimes having a suboptimal fit. One such suboptimal filterbank realization is shown in the B3 simulation (3-bin, -10 dB SNR). To show this visually, we plot the gamma PDF fitting of simulations A1 through A3 in Fig. 3(a)–(c), respectively. The simulations A1–A3 are all similar, only differing by the amount of noise present.

A more in depth analysis of α and its relationship to SNR and filterbank bin count is shown in Fig.4. For this experiment, we simulate multiple filterbanks to find the mean and variance of α. To find α values, we first generate a set of $\boldsymbol{\lambda}_X, \boldsymbol{\lambda}_D \in \mathbb{R}^{B\times 1}$, where B is the desired filterbank bin count. Each element of $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_D$ is then assigned a (uniform distributed) random value between the limits $[0, 10]$. The vector $\boldsymbol{\lambda}_D$ can then be scaled to give the desired SNR, where filterbank SNR is given as

$$\text{FBE SNR (dB)} = 10\log_{10}\left(\frac{\sum_k \lambda_{X_k}}{\sum_k \lambda_{D_k}}\right). \tag{36}$$

We then generate realizations of X_k and D_k in a similar manner to the previous experiment, using both to find α values. Values for α used here are averaged over 1000 realizations of $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_D$, each of which is used to generate 1000 realizations of X_k and D_k; i.e., one million filterbank simulations per SNR / bin count setting. In Fig. 4(a), we show the range of α values for a 10-bin filterbank. For Fig. 4(b), we show the average value of α for a 5-bin, 10-bin, 20-bin and 40-bin filterbank. From both plots, we can see a positive correlation between the SNR and α. The relationship between α and the number of filterbanks is even clearer in Fig. 4(b). It is evident there is a linear relationship between filterbank count and α. That is, doubling the filterbank bin count will double the value of α.

In the earlier filterbank chi-square fitting experiment, we performed an analysis of the gamma PDF approximation using several specific realizations of a filterbank energy variable. However, our primary goal is not to determine the quality of fit but to determine whether the MMSE log-filterbank estimator resulting from this gamma PDF assumption is good enough. For this, the following *toy* experiment is carried out on synthetic data. It operates under two main assumptions: (1) the statistical assumptions used by the spectral Wiener, STSA and STLSA estimators are correct, and (2) we have exact estimates of the *a priori* SNR ξ_k and *a posteriori* SNR γ_k. A single experimental simulation consists of the following steps:

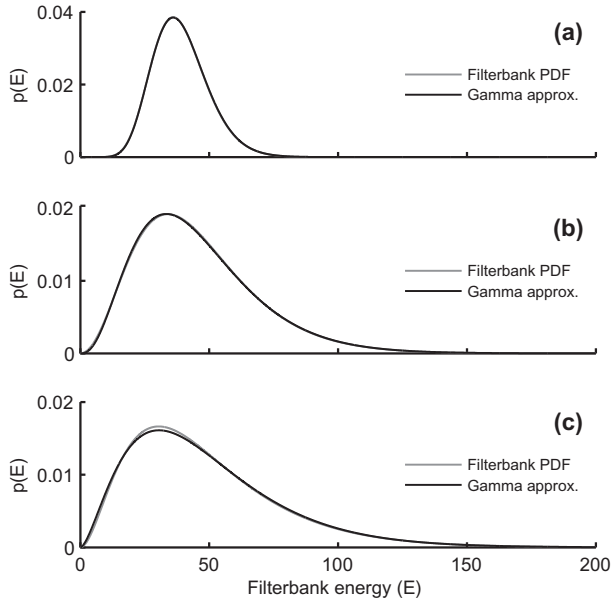

Fig. 3. Using a gamma PDF to approximate the conditioned filterbank variables. The filterbank is the summation of three conditioned spectral energies whose PDFs are given by (13). Subplots: (a) 10 dB SNR parameter set gamma PDF fitting (see Table 1-A1) (b) 0 dB SNR parameter set gamma PDF fitting (see Table 1-A2) and (c) −10 dB SNR parameter set gamma PDF fitting (see Table 1-A3). For the −10 dB simulation, the gamma PDF fitting begins noticeably deviate from the true filterbank PDF.

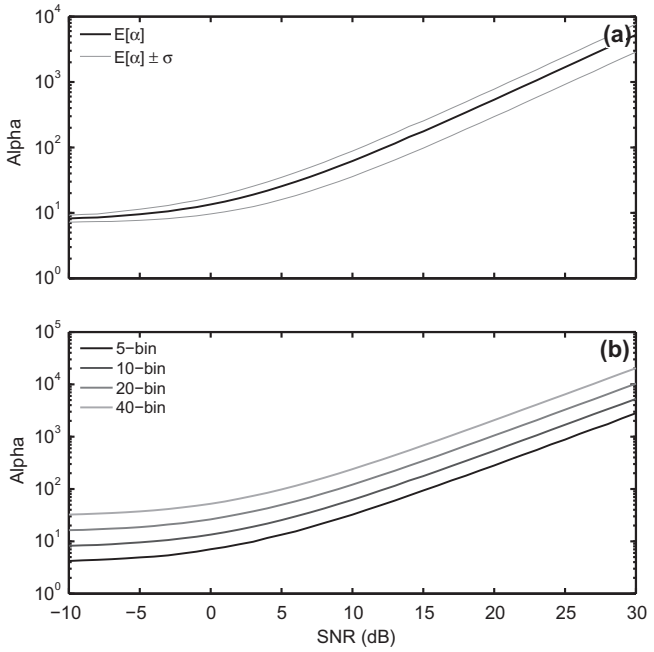

Fig. 4. The effect of SNR and bin count on the filterbank shape parameter α. Subplots: (a) Average α value for a 10-bin filterbank, with ± one standard deviation, and (b) Average α value for 5-bin, 10-bin, 20-bin and 40-bin filterbanks.

1. Generate realizations of X_k and D_k for $k = 0, 1, \ldots, K-1$. X_k and D_k are complex zero-mean Gaussian RV realizations, generated from a known set of λ_{X_k} and λ_{D_k}.
2. Calculate the oracle estimate for clean filterbank energy $L_{\text{oracle}} = \log(\sum_k |X_k|^2)$.
3. Find clean filterbank estimates $\widehat{L}_{\text{est}}$ using a particular estimator.
4. Determine the bias $[\widehat{L}_{\text{est}} - L_{\text{oracle}}]$ and square error $[\widehat{L}_{\text{est}} - L_{\text{oracle}}]^2$ of the estimate.

Bias and root mean-square-error (RMSE) is then calculated for each estimator over 500,000 such simulations. We use the spectral subtraction (SS), short-time spectral Wiener (STSW), short-time log-spectral amplitude (STLSA), short-time spectral amplitude (STSA), log-filterbank energy (LFBE) (Yu et al., 2008), proposed MAP (32) and proposed MMSE (31) estimators to generate log-filterbank estimates. For the LFBE estimator, we derive SNR parameters using filterbank versions of $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_D$ as given in (Yu et al., 2008).

In Table 2, we simulate a 5-bin, 10-bin and 20-bin log-filterbank variable at −10 dB, 0 dB and 10 dB SNRs. For the 5-bin simulations $\boldsymbol{\lambda}_X = [3, 250, 10, 100, 150]$ and $\boldsymbol{\lambda}_D \propto [3, 20, 20, 5, 30]$. For the 10-bin filterbank simulation, $\boldsymbol{\lambda}_X = [3, 3, 100, 250, 250, 100, 150, 50, 10, 4]$ and $\boldsymbol{\lambda}_D \propto [3, 10, 5, 5, 20, 50, 30, 10, 20, 20]$. Lastly, for the 20-bin simulation individual values for $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_D$ are doubled up from the previous experiment, i.e., $\boldsymbol{\lambda}_X = [3, 3, 3, 3, 100, 100, 250, 250, 250, 250 \ldots]$. For all of the simulations, $\boldsymbol{\lambda}_D$ is scaled to give the required filterbank energy SNR (36).

From Table 2, we can see that the MMSE estimator perform better than the conventional spectral estimators (e.g., STSW, STSA and STLSA) in terms of bias and RMSE in estimating the log filterbank energy. It also outperforms the other estimators. We can make a few more observations from Table 2. Firstly, a large positive bias is incurred when no enhancement is undertaken. Secondly, the MAP estimator is also consistently positively biased. The amount of MAP bias varies between the experiments, with low SNRs and low bin counts increasing the bias. This means that for the 10 dB SNR, 20-bin simulation, the MAP and MMSE estimators are virtually identical.

Ideally, we would want the MMSE estimator to be unbiased in all conditions. We can see from Table 2 that this is essentially true for all but the −10 dB 5-bin and −10 dB 10-bin conditions. This corresponds to the suboptimal conditions covered in previous experiments (i.e., low SNR, small number of filterbank bins). However, even under suboptimal conditions such as these the vast majority of bias is successfully removed. Technically, this does make the proposed estimator biased under these conditions. However, this deficiency tends to be eclipsed by the practical estimation of γ_k and ξ_k (which is a very difficult task at SNR equal to −10 dB or lower).

We can also observe from Table 2 that the LFBE estimator (Yu et al., 2008), in general, performs comparatively poorly, especially for the higher SNR/larger filterbank conditions. These cases correspond to filterbanks with high degrees of freedom (which is roughly proportional to α).

Table 2
Analysis of log-filterbank energy estimation using synthetic data from simulated filterbanks.

	−10 dB SNR		0 dB SNR		10 dB SNR	
	RMSE	Bias	RMSE	Bias	RMSE	Bias
5-bin filterbank						
None	2.565	2.438	0.276	0.105	0.276	0.105
STSW	2.143	−1.962	0.270	−0.088	0.270	−0.088
SS	3.732	0.174	0.300	−0.024	0.300	−0.024
STLSA	0.733	−0.388	0.260	−0.072	0.260	−0.072
STSA	0.625	−0.059	0.246	−0.018	0.246	−0.018
LFBE (Yu et al., 2008)	0.630	0.047	0.266	0.040	0.266	0.040
MAP	0.647	0.177	0.247	0.029	0.247	0.029
MMSE	0.622	−0.009	0.245	−0.000	0.245	−0.000
10-bin filterbank						
None	2.489	2.424	0.822	0.721	0.190	0.103
STSW	1.651	−1.520	0.578	−0.443	0.175	−0.082
SS	1.636	1.242	0.541	0.164	0.169	0.000
STLSA	0.622	−0.443	0.417	−0.259	0.167	−0.068
STSA	0.454	−0.133	0.330	−0.085	0.152	−0.026
LFBE (Yu et al., 2008)	0.467	−0.096	0.386	0.083	0.176	0.075
MAP	0.444	0.091	0.322	0.0494	0.150	0.011
MMSE	0.434	−0.002	0.318	−0.000	0.149	−0.000
20-bin filterbank						
None	2.444	2.411	0.759	0.707	0.146	0.099
STSW	1.552	−1.484	0.500	−0.430	0.130	−0.078
SS	1.533	1.391	0.394	0.202	0.112	0.007
STLSA	0.573	−0.485	0.352	−0.269	0.122	−0.068
STSA	0.351	−0.177	0.243	−0.105	0.105	−0.029
LFBE (Yu et al., 2008)	0.372	−0.191	0.266	0.047	0.134	0.080
MAP	0.307	0.046	0.220	0.024	0.100	0.005
MMSE	0.303	−0.000	0.218	−0.000	0.100	−0.000

Here it appears the Rayleigh distribution (with two degrees of freedom) is too restrictive to adequately model filterbanks of varying size and/or SNR. We should point out that the framework we use here for testing is not the framework used by the original authors to derive the MMSE LFBE estimator (Yu et al., 2008). The filterbank energies were described in (Yu et al., 2008) as being the amplitudes of a hidden, complex zero-mean Gaussian RVs. This was the same model Ephraim and Malah used to model spectral amplitudes (Ephraim and Malah, 1984; Ephraim and Malah, 1985). The motivation for this was the reduction in computational complexity offered by applying the STLSA estimator directly on the filterbank level; i.e., tracking 20–30 filterbank SNR parameters instead of a few hundred spectral SNR parameters. However, there does not appear to be any other justification for this and such a modeling assumption violates the statistical framework used by the original STLSA estimator (Ephraim and Malah, 1985).

So far, we have studied the estimation performance of different estimators on synthetic data. Now, we investigate their performance on real speech data. For this, we compute the bias and RMSE values for log-filterbank energy estimates using real speech data. The results are shown in Table 3. In order to compute bias and RMSE values, a five-point filterbank (from the 500 Hz region of speech data) is constructed. White noise (from which $\boldsymbol{\lambda}_D$ is estimated) is added to the speech stimulus at several SNRs. The parameter $\boldsymbol{\lambda}_X$ is then estimated as a 40 ms moving average of the clean speech power spectrum $|X(m,k)|^2$. We construct filterbanks from the non-silence regions of several Aurora2 (Pearce et al., 2000) corpus sentences, giving roughly 20,000 individual filterbank energies. The bias and RMSE values for log-filterbank energy estimates are calculated in a similar manner to the previous experiment. From Table 3, we can see a similar performance profile for each of the estimators. The proposed MMSE estimator performs better than the conventional estimators (e.g., STSW, STSA and STLSA) for the real speech data in terms of bias and RMSE in estimating the log filterbank energy variable. It reduces the bias considerably for the −10 dB and 0 dB experiments and is better than the other estimators. But, the bias correction is too aggressive for the 10 dB simulation – leading to an increase in the RMSE

Table 3
Analysis of 5-bin log-filterbank energy estimation using real speech data.

	−10 dB SNR		0 dB SNR		10 dB SNR	
	RMSE	Bias	RMSE	Bias	RMSE	Bias
None	1342.733	7.280	1052.966	5.250	801.523	3.676
STSW	466.163	−1.590	441.832	−2.157	479.585	−2.559
SS	744.311	−0.441	744.311	−0.441	744.311	−0.441
STLSA	155.901	−0.424	157.718	−0.484	154.550	−0.542
STSA	142.243	−0.090	141.034	−0.155	134.141	−0.216
LFBE	142.963	0.033	142.661	0.008	137.199	−0.042
MAP	143.549	0.150	139.779	0.082	130.110	0.018
MMSE	141.691	0.020	139.397	−0.048	131.223	−0.112

value. We can also observe from this table that the MAP estimator is again consistently positively biased for the real speech data.

4. Experimental results

4.1. Enhancement system description

For our experiments, we decompose speech utterances into overlapping frames. Each analysis frame is 25 ms in length, and overlaps the previous analysis frame by 15 ms. Each analysis frame has a Hamming window applied before being enhanced with a given regime. To derive the noise estimate $\lambda_D(m,k)$, we use a simple voice activity detector (VAD). An initial noise estimate is generated from the first 125 ms of each speech stimulus, and recursively updated. The recursive update is given as follows:

$$\lambda_D(m,k) = \eta\lambda_D(m-1,k) + (1-\eta)|Y_{m,k}|^2, \tag{37}$$

where $\eta = 0.98$ in the case that a noise-only frame has been detected and $\eta = 1$ otherwise. The *a posteriori* SNR can then be calculated via (12). To calculate the *a priori* SNR ξ, we use the decision-directed approach covered in Section 2.1. For the LFBE estimator, filterbank-levels $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_D$ are estimated as per (Yu et al., 2008), though we use a VAD for the noise estimation.

For the proposed MFCC estimator, the estimation of a single MFCC frame can be summarized as follows:

1. For each spectral bin, estimate spectral energies (21) and variance (23).
2. Determine each filterbank energy (24) and variance (25).
3. For each filterbank, estimate filterbank shape parameter α (27).
4. Calculate log-filterbank estimates (31).
5. Calculate cepstral coefficient vector (6).

4.2. Automatic speech recognition system description

To test ASR performance, we use a standard MFCC feature set in conjunction with the HTK recognition framework (Young et al., 2000). We accumulate 26 log-filterbank energies, and retain the first 12 cepstral coefficients (excluding the zeroth). In place of the zeroth cepstral coefficient, the total log energy of each frame is used. Once this is done, we append delta and acceleration coefficients to give a 39 dimensional feature vector. Training is provided by clean, unaltered utterances. We give results for the MMSE short-time spectral amplitude (STSA), MMSE short-time log-spectral amplitude (STLSA), vector Taylor series (Moreno et al., 1996), ETSI advanced front end and log-filterbank energy (LFBE) (Yu et al., 2008) estimators. For the VTS estimator, we use a 16 mixture diagonal covariance log-filterbank GMM (built from the clean speech training corpus) for the speech prior. For the MAP estimator, we build cepstral vectors from the log-filterbank MAP estimates (32). We conduct experiments over the following two speech tasks:

- *Aurora2 digits*. Continuous word-based recognition, small vocabulary with no language model.
- *Resource management*. Continuous triphone-based recognition, medium vocabulary with structured language model.

4.3. Aurora2 digit recognition

Aurora2 is a speaker independent database for connected digit recognition (Pearce et al., 2000). Unlike the RM database, Aurora2 lacks a language model, though its acoustic models are relatively sparse. Spoken digits in the database consist of zero through nine as well as 'oh', giving a vocabulary size of 11. Testing and training utterances were down-sampled to 8 kHz and filtered with G712 characteristics. For training models, we use clean condition stimuli. We use test utterances with noise artificially added at several SNRs. CMS is applied as a standard post-processor. The recognizer uses word-level HMMs, each with 16 states and 3 Gaussian mixtures per state. For the proposed estimators, we use an SPU of $q_k = 0.05$. For the STSA and STLSA estimators, SPU severely degraded recognition accuracy and was thus omitted. ASR WER scores are given in Tables 4 and 5 for recognition tasks A and B, respectively.

4.4. Resource management word recognition

A speaker independent section of the DARPA resource management (RM) database is used for medium-vocabulary recognition (Price et al., 1988). The database was recorded in clean conditions (sample rate of 16 kHz) and has a vocabulary of approximately 1000 words. For training, there are 3990 sentences spoken by 109 speakers. For testing, we use the February '89 test set which has 300 sentences spoken by 10 different speakers. White, Volvo and babble noises are artificially added at several SNRs. For recognition, we train triphone-level HMMs (from clean condition stimulus), having three states with eight Gaussian mixtures each. Cepstral mean subtraction (CMS) is applied as a standard post-processor. For the proposed estimators, we use an SPU of $q_k = 0.3$. For the STSA and STLSA estimators, SPU did not improve recognition accuracy and was thus omitted. ASR word error rate (WER) scores are given in Table 6.

4.5. Discussion

For both the RM and Aurora2 tasks, the proposed MMSE estimator has superior performance compared with the other MMSE spectral estimators (STSA, STLSA and STSW). However, performance of all spectral MMSE

Table 4
Aurora2A ASR word error rates.

Treatment	SNR (dB)						
	∞	20	15	10	5	0	AVG
Subway noise							
None	0.68	2.95	5.59	12.34	30.49	70.56	24.39
SA	0.89	3.07	4.54	9.12	17.68	39.70	14.82
LSA	1.04	3.90	5.68	11.30	20.94	42.95	16.95
VTS	0.68	2.67	5.37	11.79	25.76	53.15	19.75
ETSI	0.77	2.03	3.38	7.92	15.44	35.31	12.82
LFBE	0.74	3.81	5.83	13.72	35.40	70.62	25.88
MAP	0.74	3.04	4.76	9.30	19.10	43.41	15.92
MMSE	0.77	2.92	4.64	9.15	18.21	42.03	15.39
Babble noise							
None	0.76	2.03	4.63	16.51	53.02	90.84	33.41
SA	0.82	3.42	7.56	17.23	38.33	68.80	27.07
LSA	0.88	8.86	15.57	26.57	45.77	70.95	33.54
VTS	0.68	2.24	4.20	12.06	34.67	75.18	25.67
ETSI	0.77	1.90	3.57	8.22	19.74	47.19	16.12
LFBE	0.74	4.20	10.94	26.57	54.93	85.25	36.38
MAP	0.74	2.00	4.11	11.79	31.71	68.20	23.56
MMSE	0.77	2.09	4.53	12.70	32.29	66.20	23.56
Car noise							
None	0.78	2.09	4.38	11.39	34.39	81.12	26.67
SA	0.81	1.58	2.80	5.31	14.08	34.30	11.61
LSA	0.92	1.70	3.40	6.53	16.10	38.98	13.34
VTS	0.68	2.27	4.00	6.89	14.94	52.16	16.05
ETSI	0.77	1.25	2.24	4.98	10.53	30.06	9.81
LFBE	0.74	1.97	3.58	9.78	34.09	74.98	24.88
MAP	0.74	1.79	2.86	6.32	16.10	43.93	14.20
MMSE	0.77	1.73	2.80	5.91	14.88	40.41	13.15
Exhibition noise							
None	0.74	3.73	6.97	15.21	38.88	82.14	29.39
SA	0.80	3.70	6.17	13.45	28.51	51.96	20.76
LSA	1.02	6.76	9.97	18.82	35.51	57.76	25.76
VTS	0.68	3.12	4.44	9.04	19.84	44.62	16.21
LFBE	0.74	3.86	8.36	19.25	48.19	81.61	32.25
ETSI	0.77	2.41	3.61	7.47	16.04	36.66	13.24
MAP	0.74	3.33	5.62	12.19	27.24	54.86	20.65
MMSE	0.77	3.21	5.65	11.54	25.58	51.71	19.54
Set A averages							
None	0.74	2.70	5.39	13.86	39.20	81.17	28.47
SA	0.83	2.94	5.27	11.28	24.65	48.69	18.57
LSA	0.97	5.31	8.66	15.81	29.58	52.66	22.40
VTS	0.68	2.58	4.50	9.95	23.80	56.28	19.42
ETSI	0.77	1.90	3.20	7.15	15.44	37.30	13.00
LFBE	0.74	3.46	7.18	17.33	43.15	78.11	29.85
MAP	0.74	2.54	4.34	9.90	23.58	52.60	18.58
MMSE	0.77	2.49	4.41	9.83	22.74	50.09	17.91

WER average is computed from 0 dB to 20 dB SNR.

Table 5
Aurora2B ASR word error rates.

Treatment	SNR (dB)						
	∞	20	15	10	5	0	AVG
Restaurant noise							
None	0.68	2.33	4.70	16.18	45.96	78.97	29.63
SA	0.89	5.13	11.30	23.64	42.03	67.58	29.94
LSA	1.04	11.45	18.02	31.01	47.80	71.45	35.95
VTS	0.68	2.00	4.82	12.93	33.74	67.18	24.13
ETSI	0.77	1.93	4.76	9.49	22.44	47.25	17.17
LFBE	0.74	5.22	12.62	28.22	52.07	79.71	35.57
MAP	0.74	3.04	6.29	15.20	36.26	66.17	25.39
MMSE	0.77	2.98	6.45	15.84	36.66	66.26	25.64
Street noise							
None	0.76	2.60	5.11	13.45	37.64	75.09	26.78
SA	0.82	3.33	5.20	9.95	22.61	47.85	17.79
LSA	0.88	4.41	7.65	13.48	28.96	54.38	21.78
VTS	0.68	3.23	4.87	10.82	24.21	55.41	19.71
ETSI	0.77	2.15	3.54	7.68	16.44	37.58	13.48
LFBE	0.74	3.20	6.41	16.93	42.62	75.97	29.03
MAP	0.74	2.60	4.59	8.62	23.61	52.66	18.42
MMSE	0.77	2.48	4.41	8.43	22.10	49.64	17.41
Airport noise							
None	0.78	1.79	4.09	11.72	37.16	73.90	25.73
SA	0.81	2.54	5.91	13.09	29.76	53.80	21.02
LSA	0.92	5.13	9.54	18.70	35.67	58.75	25.56
VTS	0.68	1.37	3.34	7.84	22.25	53.71	17.70
ETSI	0.77	1.40	2.86	6.59	15.54	36.24	12.53
LFBE	0.74	2.71	6.50	18.52	46.47	75.63	29.97
MAP	0.74	1.82	3.55	9.48	25.77	54.88	19.10
MMSE	0.77	1.88	3.64	10.02	25.77	53.06	18.87
Train noise							
None	0.74	1.54	4.32	11.66	34.74	80.31	26.51
SA	0.80	2.93	4.75	8.05	20.98	40.76	15.49
LSA	1.02	4.17	6.39	11.08	24.07	44.99	18.14
VTS	0.68	1.60	4.35	9.26	21.14	54.24	18.12
ETSI	0.77	1.64	3.58	6.11	15.15	35.51	12.40
LFBE	0.74	2.90	5.12	12.13	40.82	74.85	27.16
MAP	0.74	1.94	3.86	7.22	20.12	48.35	16.30
MMSE	0.77	2.16	3.76	7.03	19.44	44.80	15.44
Set B averages							
None	0.74	2.07	4.56	13.25	38.88	77.07	27.16
SA	0.83	3.48	6.79	13.68	28.85	52.50	21.06
LSA	0.97	6.29	10.40	18.57	34.13	57.39	25.36
VTS	0.68	2.05	4.35	10.21	25.34	57.64	19.92
ETSI	0.77	1.78	3.69	7.47	17.39	39.14	13.89
LFBE	0.74	3.51	7.66	18.95	45.50	76.54	30.43
MAP	0.74	2.35	4.57	10.13	26.44	55.52	19.80
MMSE	0.77	2.38	4.57	10.33	25.99	53.44	19.34

WER average is computed from 0 dB to 20 dB SNR.

estimators (including the proposed approach) fell short of the ETSI front-end on the Aurora2 task. For the Aurora2 recognition task, there is a fairly consistent reduction in WER when moving from the MAP estimator to the MMSE estimator. Not surprisingly, the two estimators only diverge at lower SNRs.

For the RM recognition task, differences between the MAP and MMSE estimators are much smaller. Here, the benefits of removing bias seemed to be offset by increased speech degradation at higher SNRs. Nonetheless, the proposed estimators have an advantage over the other common estimators (such as the STSA, STLSA and STSW).

An absolute improvement of 1.95% and 1.75% for the Aurora2 and RM tasks, respectively is required to meet statistical significance tests (for $p = 0.05$) (Gillick and Cox, 1989). While the proposed MMSE estimator demonstrates significant gains over the baseline (referred to as treatment 'none' in Tables 4–6) and STLSA estimator, the differences between the STSA, MAP and MMSE estimators does not meet the requirements for significance.

Table 6
RM ASR word error rates.

Treatment	SNR (dB)					
	∞	30	20	10	0	AVG
White noise						
None	4.30	5.48	11.89	47.13	95.89	17.20
SA	4.26	5.04	7.43	27.02	79.55	10.94
LSA	4.42	5.36	7.55	23.39	76.85	10.18
VTS	4.18	4.77	8.96	47.01	96.36	16.23
ETSI	4.61	4.65	7.98	25.60	80.05	10.46
LFBE	4.18	5.28	8.25	29.06	90.61	11.69
MAP	4.50	5.01	7.00	23.03	75.87	9.89
MMSE	4.54	4.97	6.92	23.19	76.46	9.91
Babble noise						
None	4.30	4.73	8.21	38.87	94.68	14.03
SA	4.26	4.89	7.78	28.90	89.87	11.46
LSA	4.42	5.08	8.56	32.58	90.61	12.66
VTS	4.18	4.81	6.92	31.13	99.26	11.76
ETSI	4.61	5.01	7.43	27.93	80.69	9.67
LFBE	4.18	4.77	9.93	43.76	100.31	15.66
MAP	4.50	4.93	7.74	30.00	90.30	11.79
MMSE	4.54	5.04	7.90	30.31	89.71	11.95
Volvo noise						
None	4.30	4.03	5.01	7.86	23.03	5.30
SA	4.26	4.42	4.34	4.73	10.09	4.44
LSA	4.42	4.61	4.46	4.93	8.76	4.61
VTS	4.18	4.42	5.71	9.07	17.21	5.84
ETSI	5.44	4.58	4.77	5.93	9.27	5.10
LFBE	4.18	4.18	4.89	8.45	21.20	5.42
MAP	4.50	4.69	4.42	4.93	8.96	4.64
MMSE	4.54	4.69	4.38	4.89	8.88	4.63
Average						
None	4.30	4.75	8.37	31.29	71.20	12.18
SA	4.26	4.78	6.52	20.22	59.84	8.94
LSA	4.42	5.02	6.86	20.30	58.74	9.15
VTS	4.18	4.67	7.20	29.07	70.94	11.28
ETSI	4.89	4.75	6.73	19.82	56.67	9.05
LFBE	4.18	4.74	7.69	27.09	70.71	10.93
MAP	4.50	4.88	6.39	19.32	58.38	8.77
MMSE	4.54	4.90	6.40	19.46	58.35	8.83

WER average is computed from 10 dB to ∞ dB SNR.

Table 7
Effect of SPU parameter q_k on RM ASR word error rates.

Treatment	SNR (dB)					
	∞	30	20	10	0	AVG
White noise						
SPU ($q_k = 0.3$)	4.54	4.97	6.92	23.19	76.46	9.91
No SPU ($q_k = 0$)	4.15	5.20	9.11	34.81	91.08	13.32
Babble noise						
SPU ($q_k = 0.3$)	4.54	5.04	7.90	30.31	89.71	11.95
No SPU ($q_k = 0$)	4.15	4.85	6.69	28.63	89.48	11.08
Volvo noise						
SPU ($q_k = 0.3$)	4.54	4.69	4.38	4.89	8.88	4.63
No SPU ($q_k = 0$)	4.15	4.18	4.73	5.87	13.65	4.73
Average						
SPU ($q_k = 0.3$)	4.54	4.90	6.40	19.46	58.35	8.83
No SPU ($q_k = 0$)	4.15	4.74	6.84	23.10	64.74	9.71

WER average is computed from 10 dB to ∞ dB SNR.

4.6. Effect of SPU on the MMSE estimator

In Table 7, ASR comparisons with SPU ($q_k = 0.3$) and without SPU ($q_k = 0$) are shown for the RM noise task. For the white noise task, SPU can be seen to give a large increase in robustness, especially at lower SNRs. For this task, the majority of noise energy falls in non-speech regions, allowing SPU to be used to great effect. However, SPU has much less effect on the babble noise task – typically degrading robustness. While the introduction of SPU increases robustness overall for the white and Volvo noise tasks, it comes with the trade-off of increased speech degradation at higher SNRs.

5. Conclusion

In this paper, we have investigated a family of spectral estimators for use in robust ASR. While several estimators (such as the short-time spectral amplitude estimator and short-time log-spectral amplitude estimator) are commonly used for robust ASR, they are sub-optimal for this task. In this paper, we have extended the statistical framework used by these estimators to derive an MMSE log-filterbank energy estimator. To make this framework suitable for MFCC estimation, several mathematical transformation were studied to covert spectral domain models into log-filterbank domain models. The proposed estimator gave significant improvements to robustness over the baseline ASR system. While performance gains over wider spectral estimator family were demonstrated, gains in some cases were quite small. Here results indicated similar ASR robustness among the proposed MMSE, STSA, and MAP (or SE) estimators.

Appendix A. Derivation of probability density functions

This appendix provides a step-by-step derivation of the spectral energy and log-filterbank energy PDFs. Under the assumed statistical framework given in Section 2, the PDF $p(A_k)$ is given by a Rayleigh distribution

$$p(A_k) = \frac{2A_k}{\lambda_{X_k}} \exp\left(-\frac{[A_k]^2}{\lambda_{X_k}}\right). \tag{A.1}$$

The Rayleigh distribution describes a variable $y = \sqrt{x_a^2 + x_b^2}$, where x_a and x_b are independent and identically distributed zero-mean Gaussian variables. For our purposes, it is used to describe the amplitudes of spectral variables (which were previously assumed Gaussian distributed along both the real and imaginary axis). The conditional PDF $p(Y_k|A_k,\theta_k)$ is given as

$$\begin{aligned} p(Y_k|A_k,\theta_k) &= \frac{1}{\pi\lambda_{D_k}} \exp\left(-\frac{|D_k|^2}{\lambda_{D_k}}\right) \\ &= \frac{1}{\pi\lambda_{D_k}} \exp\left(-\frac{|Y_k - A_k \exp\quad(j\theta_k)|^2}{\lambda_{D_k}}\right), \end{aligned} \tag{A.2}$$

where θ_k is the spectral phase of clean spectral value X_k. Here we have assumed that spectral value Y_k is only related to A_k and not any other spectral bins. If we additionally assume θ_k to be uniformly distributed over the $[-\pi,\pi]$ interval, it may be integrated out of (A.2) to give (Gradshteyn and Ryzhik, 2007): {3.339}

$$\begin{aligned} p(Y_k|A_k) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{\pi\lambda_{D_k}}\exp\left(-\frac{|Y_k - A_k\exp[j\theta_k]|^2}{\lambda_{D_k}}\right)d\theta_k \\ &= \frac{1}{2\pi^2\lambda_{D_k}}\exp\left(\frac{-|Y_k|^2-[A_k]^2}{\lambda_{D_k}}\right) \\ &\quad\cdot\int_{-\pi}^{\pi}\exp\left(\frac{2|Y_k|A_k\cos\theta_k}{\lambda_{D_k}}\right)d\theta_k \\ &= \frac{1}{\pi\lambda_{D_k}}\exp\left(\frac{-|Y_k|^2-[A_k]^2}{\lambda_{D_k}}\right)I_0\left(\frac{2|Y_k|A_k}{\lambda_{D_k}}\right) \\ &= \frac{1}{\pi\lambda_{D_k}}\exp\left(\frac{-|Y_k|^2-[A_k]^2}{\lambda_{D_k}}\right)I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}A_k\right), \end{aligned} \tag{A.3}$$

where $I_0(\cdot)$ is the zeroth order modified Bessel function. Using Bayes rule, (A.1) and (A.3) can be combined to give the conditioned spectral amplitude PDF $p(A_k|Y_k)$

$$\begin{aligned} p(A_k|\boldsymbol{Y}) &= \frac{p(A_k)p(Y_k|A_k)}{\int_0^\infty p(A_k)p(Y_k|A_k)dA_k} \\ &= \frac{A_k\exp\left(\frac{-[A_k]^2}{\lambda_{X_k}+\lambda_{D_k}}\right)I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}A_k\right)}{\int_0^\infty \tau\exp\left(\frac{-\tau^2}{\lambda_{X_k}+\lambda_{D_k}}\right)I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}\tau\right)d\tau} \\ &= \frac{A_k\exp\left(\frac{-[A_k]^2}{\lambda_k}\right)I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}A_k\right)}{\int_0^\infty \tau\exp\left(\frac{-\tau^2}{\lambda_k}\right)I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}\tau\right)d\tau}. \end{aligned} \tag{A.4}$$

Original derivation of the spectral amplitude estimator and its corresponding PDF can be found in (Ephraim and Malah, 1984). We may derive the spectral energy PDF with a few additional algebraic manipulations. Firstly, the integral in the denominator of (A.4) can be solved[6] (Gradshteyn and Ryzhik, 2007): {6.631−7, 8.406−1, 8.464−1, 8.464−2, 9.210−1},

$$p(A_k|\boldsymbol{Y}) = \frac{2A_k\exp\left(\frac{-[A_k]^2}{\lambda_k}\right)I_0\left(2\sqrt{\frac{v_k}{\lambda_k}}A_k\right)}{\lambda_k\exp(v_k)}. \tag{A.5}$$

There is a one to one mapping between e_k and A_k over the $[0,\infty]$ interval. If we equate the cumulative density functions (CDFs) for each variable, then differentiate both w.r.t e_k, we get

$$p(e_k|\boldsymbol{Y}) = p(A_k|\boldsymbol{Y})\cdot\left|\frac{dA_k}{de_k}\right| = \frac{p(A_k|\boldsymbol{Y})}{2\sqrt{e_k}}. \tag{A.6}$$

[6] Detail of a similar integration is given in Appendix B.

Substituting (A.5) into (A.6) and using the substitution $A_k = \sqrt{e_k}$ yields the conditioned spectral energy PDF

$$p(e_k|Y_k) = \frac{\exp\left(\frac{-e_k}{\lambda_k}\right)I_0\left(2\sqrt{\frac{v_k e_k}{\lambda_k}}\right)}{\lambda_k\exp(v_k)}. \tag{A.7}$$

A similar approach may be used for converting the filterbank energy PDF (26) to a log-filterbank energy PDF (29). The main difference is that the logarithm (in comparison to the squaring operator) is a one to one mapping from the $[0,\infty]$ to $[-\infty,\infty]$ intervals. Assuming a gamma PDF for the conditioned filterbank energy variable, the PDF for the conditioned log-filterbank energies can be given as

$$\begin{aligned} p(L_x(q)|\boldsymbol{Y}) &= p(E_x(q)|\boldsymbol{Y})\cdot\left|\frac{dE_x(q)}{dL_x(q)}\right| \\ &= \frac{[E_x(q)]^{\alpha_q-1}\exp\left(-\frac{E_x(q)}{\beta_q}\right)}{\beta_q^{\alpha_q}\Gamma(\alpha_q)}\cdot\exp(L_x(q)) \\ &= \frac{\exp\left(\alpha_q[L_q-\log\beta_q]\right)}{\exp\left(\exp\left(L_q-\log\beta_q\right)\right)\cdot\Gamma(\alpha_q)}. \end{aligned} \tag{A.8}$$

Appendix B. Derivation of spectral energy and log-filterbank estimates

This appendix provides a step by step derivation of the spectral energy and log-filterbank estimation. Given the conditioned spectral energy PDF $p(e_k|\boldsymbol{Y})$ (13), the first raw moment (mean) of spectral energy is given by

$$\begin{aligned} E[e_k|\boldsymbol{Y}] = \hat{e}_k &= \int_0^\infty e_k\cdot p(e_k|\boldsymbol{Y})de_k \\ &= \frac{\int_0^\infty e_k\exp\left(\frac{-e_k}{\lambda_k}\right)I_0\left(2\sqrt{\frac{v_k e_k}{\lambda_k}}\right)de_k}{\lambda_k\exp(v_k)}. \end{aligned} \tag{B.1}$$

The above equation may be solved and simplified with (Gradshteyn and Ryzhik, 2007): {6.643−1, 9.220−2, 9.212−1, 9.210−1},

$$\begin{aligned} \hat{e}_k &= \frac{[v_k]^{-0.5}\cdot[\lambda_k]^2\exp\left(\frac{v_k}{2}\right)M_{-1.5,0}(v_k)}{\lambda_k\exp(v_k)} \\ &= [v_k]^{-0.5}\cdot\lambda_k\exp\left(-\frac{v_k}{2}\right)\cdot M_{-1.5,0}(v_k) \\ &= [v_k]^{-0.5}\cdot\lambda_k\exp\left(-\frac{v_k}{2}\right)\cdot[v_k]^{0.5}\exp\left(-\frac{v_k}{2}\right)\Phi_{2,1}(v_k) \\ &= [v_k]^{-0.5}\cdot\lambda_k\exp\left(-\frac{v_k}{2}\right).[v_k]^{0.5}\exp\left(\frac{v_k}{2}\right)\Phi_{-1,1}(-v_k) \\ &= [v_k]^{-0.5}\cdot\lambda_k\exp\left(-\frac{v_k}{2}\right)\cdot[v_k]^{0.5}\exp\left(\frac{v_k}{2}\right)(1+v_k) \\ &= \lambda_k(1+v_k), \end{aligned} \tag{B.2}$$

where, $M(\cdot)$ is the Whittaker function and $\Phi(\cdot)$ is the confluent hypergeometric function. The second central moment (variance) of the spectral energy is given as

$$E\left[(e_k-\hat{e}_k)^2|\boldsymbol{Y}\right]=\Sigma_{e_x}(k,k)=\frac{\int_0^\infty [e_k]^2\exp\left(\frac{-e_k}{\lambda_k}\right)I_0\left(2\sqrt{\frac{v_k e_k}{\lambda_k}}\right)de_k}{\lambda_k\exp(v_k)}-[\hat{e}_k]^2. \tag{B.3}$$

The above equation can be solved in a similar manner to (B.1) using (Gradshteyn and Ryzhik, 2007): {6.643−1, 9.220−2, 9.212−1, 2.9210−1},

$$\begin{aligned}\Sigma_{e_x}(k,k)&=\frac{2[v_k]^{-0.5}\cdot[\lambda_k]^3\exp\left(\frac{v_k}{2}\right)M_{-2.5,0}(v_k)}{\lambda_k\exp(v_k)}-[\lambda_k(1+v_k)]^2\\&=[2v_k]^{-0.5}\cdot[\lambda_k]^2[v_k]^{0.5}\exp(-v_k)\Phi_{3,1}(v_k)-[\lambda_k(1+v_k)]^2\\&=[2v_k]^{-0.5}\cdot[\lambda_k]^2\cdot[v_k]^{0.5}\Phi_{-2,1}(-v_k)-[\lambda_k(1+v_k)]^2\\&=2[\lambda_k]^2\left(1+2v_k+\frac{[v_k]^2}{2}\right)-[\lambda_k(1+v_k)]^2\\&=[\lambda_k]^2(1+2v_k).\end{aligned} \tag{B.4}$$

Given the conditioned filterbank energy PDF $p(E_x(q)|\boldsymbol{Y})$ (26), the first raw moment (mean) of the log-filterbank energy is given by (Gradshteyn and Ryzhik, 2007): {4.352−1},

$$\begin{aligned}E[L_x(q)|\boldsymbol{Y}]=\widehat{L}_x(q)&=\int_0^\infty \log E_x(q)\cdot p(E_x(q)|\boldsymbol{Y})dE_x(q)\\&=\int_0^\infty \log E_x(q)\cdot\frac{[E_x(q)]^{\alpha_q-1}\exp\left(-\frac{E_x(q)}{\beta_q}\right)}{\beta_q^{\alpha_q}\Gamma(\alpha_q)}dE_x(q)\\&=\frac{\beta_q^{\alpha_q}\Gamma(\alpha_q)}{\beta_q^{\alpha_q}\Gamma(\alpha_q)}\left(\Psi_0(\alpha_q)-\log\left(\frac{1}{\beta_q}\right)\right)\\&=\log(\alpha_q\beta_q)-\log(\alpha_q)+\Psi_0(\alpha_q)\\&=\log\widehat{E}_x(q)-\left[\log(\alpha_q)-\Psi_0(\alpha_q)\right].\end{aligned} \tag{B.5}$$

To find the MAP log-filterbank estimate, we are interested in finding the value of L_q that maximizes $p(L_q|\boldsymbol{Y})$; i.e., the location of the PDF peak

$$\begin{aligned}\widehat{L}_{q-MAP}&=\arg\max_{L_q}\left[p(L_q|\boldsymbol{Y})\right]\\&=\arg\max_{L_q}\left[\log p(L_q|\boldsymbol{Y})\right]=\arg\max_{L_q} f(L_q),\end{aligned} \tag{B.6}$$

where

$$f(L_q)=\left[\alpha_q(L_q-\log\beta_q)-\exp(L_q-\log\beta_q)\right]. \tag{B.7}$$

To find the maxima, we first find the derivative of (B.7) w.r.t L_q,

$$\begin{aligned}&\frac{d}{dL_q}\alpha_q(L_q-\log\beta_q)-\exp(L_q-\log\beta_q)\cdot dL_q\\&\quad=\alpha_q-\exp(L_q-\log\beta_q).\end{aligned} \tag{B.8}$$

then, setting the derivative (B.8) at $L_q=L_{q-MAP}$ to zero

$$\begin{aligned}&\alpha_q-\exp(L_{q-MAP}-\log\beta_q)=0,\\&L_{q-MAP}=\log\alpha_q+\log\beta_q,\\&L_{q-MAP}=\log\widehat{E}_q.\end{aligned} \tag{B.9}$$

Appendix C. Derivation of lower limit for the filterbank energy parameter

In this section we show that the gamma PDF shape parameter α_q cannot take values below 1 when used to model filterbank energies under the assumed noise model. In order to actually model the filterbank, we first assume filterbanks have non-zero energy; i.e., $\hat{e}_k>0$. Substituting (24) and (25) into (27), we have

$$\begin{aligned}\alpha_q&=\frac{\left[\sum_k H(k,q)\hat{e}_k\right]^2}{\sum_k[H(k,q)]^2\Sigma_{e_x}(k,k)}\\&=\frac{\sum_k[H(k,q)]^2[\hat{e}_k]^2+\sum_k\sum_{\substack{i\\i\neq k}}H(k,q)H(i,q)\hat{e}_k\hat{e}_i}{\sum_k[H(k,q)]^2\Sigma_{e_x}(k,k)}.\end{aligned} \tag{C.1}$$

Substituting (15) into (C.1), gives

$$\begin{aligned}\alpha_q&=\frac{\sum_k[H(k,q)]^2[\hat{e}_k]^2+\sum_k\sum_{\substack{i\\i\neq k}}H(k,q)H(i,q)\hat{e}_k\hat{e}_i}{\sum_k[H(k,q)]^2\hat{e}_k]^2-\sum_k[H(k,q)]^2\left(\frac{\xi_k}{1+\xi_k}\right)^4|Y_k|^4}\\&=\frac{T_1(q)+T_2(q)}{T_1(q)-T_3(q)},\end{aligned} \tag{C.2}$$

where terms

$$T_1(q)=\sum_k[H(k,q)]^2[\hat{e}_k]^2, \tag{C.3}$$

$$T_2(q)=\sum_k\sum_{\substack{i\\i\neq k}}H(k,q)H(i,q)\hat{e}_k\hat{e}_i, \tag{C.4}$$

$$T_3(q)=\sum_k[H(k,q)]^2\left(\frac{\xi_k}{1+\xi_k}\right)^4|Y_k|^4. \tag{C.5}$$

We first note that terms $T_1(q)$, $T_2(q)$ and $T_3(q)$ are non-negative. This can be reasoned by using the fact that ξ_k, $H(k,q)$, and $\hat{e}_k$ are all non-negative. As a result, the numerator of (C.2) is greater than, or equal to the denominator of (C.2). Secondly, we note that the denominator of (C.2) is also non-negative. This is because the denominator is the filterbank variance – which again is strictly non-negative. From both of these observations, it can be inferred that the value of α_q cannot fall below 1.

References

Acero, A., Deng, L., Kristjansson, T., Wang, J., 2000. HMM adaptation using vector taylor series for noisy speech recognition. In: Proc. Interspeech.

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. ICSLP, pp. 373–376.

Cohen, I., 2002. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. Signal Process. Lett., IEEE 9, 113–116.

Cooke, M., Green, P., Josifofski, L., Vizinho, A., 2000. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Comm. 34, 267–285.

Davis, S., Mermelstein, P., 1990. Readings in Speech Recognition: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Deng, L., Acero, A., Huang, X., 2000. Large vocabulary speech recognition under adverse acoustic environments. In: Proc. ICSLP.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32, 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33, 443–445.

Ephraim, Y., Trees, H.V., 1991. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. 39, 795–805.

Erell, A., Weintraub, M., 1993. Energy conditioned spectral estimation for recognition of noisy speech. IEEE Trans. Speech Audio Process. 1, 84–89.

Fujimoto, M., Ariki, Y., 2000. Noisy speech recognition using noise reduction method based on Kalman filter. IEEE Trans. Acoust. Speech Signal Process. 3, 1727–1730.

Gales, L.,1995. Model-based techniques for robust speech recognition. Ph.D. thesis, University of Cambridge, UK.

Gemello, R., Mana, F., Mori, R., 2006. Automatic speech recognition with a modified Ephraim–Malah rule. IEEE Signal Process. Lett. 13, 56–59.

Gillick, L., Cox, S., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: IEEE Internat. Conf. ICASSP Acoustics, Speech, and Signal Processing, pp. 532–535.

Gradshteyn, I., Ryzhik, I., 2007. Table of Integrals Series and Products. Elsevier.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. Acoust. Soc. Amer. J. 87, 1738–1752.

Hermus, K., Wambacq, P., Hamme, H.V., 2007. A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP J. Appl. Signal Process. 2007, 195–197.

Indrebo, K., Povinelli, R., Johnson, M., 2008. Minimum mean-squared error estimation of mel-frequency cepstral coefficients using a novel distortion model. IEEE Trans. Audio Speech Lang. Process. 16, 1654–1661.

Lathoud, G., Magimai-Doss, M., Mesot, B., Bourlard, H., 2005. Unsupervised Spectral Subtraction for Noise-Robust ASR. In: Proc. 2005 IEEE ASRU Workshop, pp. 343–348.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press.

Malah, D., Cox, R.V., Accardi, A.J., 1999. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. Proc. Acoustics, Speech, and Signal Processing, IEEE Internat. Conf. ICASSP. IEEE Computer Society, pp. 789–792.

McAulay, R., Malpass, M., 1980. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Process. 28, 137–145.

Moreno, P., 1996. Speech recognition in noisy environments. Ph.D. thesis, Carnegie Mellon University.

Moreno, P., Raj, B., Stern, R., 1996. A vector Taylor series approach for environment-independent speech recognition. In: IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, pp. 733–736.

Pearce, D., Hirsch, H.G., Gmbh, E.E.D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ISCA ITRW ASR2000, pp. 29–32.

Price, P., Fisher, W., Bernstein, J., Pallett, D., 1988. The darpa 1000-word resource management database for continuous speech recognition. In: IEEE Internat. Conf. on ICASSP Acoustics, Speech, and Signal Processing, Vol. 1, pp. 651–654.

Rabiner, L., Schafer, R., 1978. Digital Processing of Speech Signals. Prentice Hall.

Raj, B., Stern, R., 2005. Missing-feature approaches in speech recognition. Signal Process. Mag., IEEE 22, 101–116.

Soon, I., Koh, S., Yeo, C., 1999. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. Signal Process. 75, 151–159.

Spouge, J., 1994. Computation of the gamma, digamma, and trigamma functions. SIAM J. Numer. Anal. 31, 931–944.

Stouten, V., 2006. Robust speech recognition in time-varying environments. Ph.D. thesis, Katholieke Universiteit Leuven.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2000. The HTK Book Version 3.0. Cambridge University Press.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., Acero, A., 2008. Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor. IEEE Trans. Audio Speech Lang. Process. 16, 1061–1070.