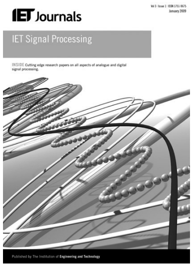
Published in IET Signal Processing
Received on 1st June 2012
Revised on 15th February 2013
Accepted on 24th February 2013
doi: 10.1049/iet-spr.2012.0167



ISSN 1751-9675

Distributed multichannel speech enhancement based on perceptually-motivated Bayesian estimators of the spectral amplitude

Marek B. Trawicki, Michael T. Johnson

Department of Electrical and Computer Engineering, Speech and Signal Processing Laboratory, Marquette University, P.O. Box 1881, Milwaukee, WI 53201-1881, USA
E-mail: marek.trawicki@marquette.edu

Abstract: In this study, the authors propose multichannel weighted Euclidean (WE) and weighted cosh (WCOSH) cost function estimators for speech enhancement in the distributed microphone scenario. The goal of the work is to illustrate the advantages of utilising additional microphones and modified cost functions for improving signal-to-noise ratio (SNR) and segmental SNR (SSNR) along with log-likelihood ratio (LLR) and perceptual evaluation of speech quality (PESQ) objective metrics over the corresponding single-channel baseline estimators. As with their single-channel counterparts, the perceptually-motivated multichannel WE and WCOSH estimators are functions of a weighting law parameter, which influences attention of the noisy spectral amplitude through a spectral gain function, emphasises spectral peak (formant) information, and accounts for auditory masking effects. Based on the simulation results, the multichannel WE and WCOSH cost function estimators produced gains in SSNR improvement, LLR output and PESQ output over the single-channel baseline results and unweighted cost functions with the best improvements occurring with negative values of the weighting law parameter across all input SNR levels and noise types.

1 Introduction

Over the past three decades, research in speech enhancement has concentrated on frequency-domain statistical estimators derived in the minimum mean-square error (MMSE) sense for estimation of the spectral amplitude (SA) [1–3]. Unfortunately, these MMSE estimators that minimise the Bayes risk on a squared-error cost function are not the most subjectively meaningful for three reasons: estimation errors do not necessarily directly relate to speech quality, estimators might not preserve spectral peak (formant) information or account for auditory masking effects and estimation errors do not convey the same perceptual meaning but are treated in the same unweighted fashion [3]. The true goal of speech enhancement is to not necessarily reduce the background noise, which is measured through signal-to-noise ratio (SNR) and segmental SNR (SSNR), but rather to improve both the quality and intelligibility of the noisy signals [4]. To evaluate speech quality in an automated, accurate and reliable way, Hu and Loizou [5] demonstrated that the log-likelihood ratio (LLR) and perceptual evaluation of speech quality (PESQ) objective metrics correlated the best with speech distortion and overall speech quality, which are better indicators for performance evaluation. Through modifications to the statistical prior models or estimator equations, Andrianakis and White [6], Erkelens *et al.* [7], Plourde and Champagne [8] and You *et al.* [9] demonstrated only marginal improvements in LLR and PESQ over the corresponding baseline methods. In order to achieve further gains in performance, these standard and more advanced single-channel estimators can be extended to multiple microphones [10, 11], particularly the relatively new distributed microphone paradigm [12–15].

In this work, the focus is on extending the more advanced and perceptually-relevant cost functions for performing single-channel speech enhancement, namely the weighted Euclidean (WE) and weighted cosh (WCOSH) cost functions developed by Loizou [3], to distributed microphone domain. Specifically, the SA estimation is an extension of the short-time spectral amplitude estimator derived for distributed microphones by Lotter *et al.* [16]. In conjunction with the multichannel SA and spectral phase estimators [12, 17], the goal is to demonstrate that the reconstructed enhanced signal produces increase in not only in SSNR but also in LLR and PESQ performance over the baseline single-channel results with additional microphone channel information. As with their single-channel counterparts, the multichannel WE and WCOSH estimators are functions of a weighting law parameter, which influences attention of the noisy SA through a spectral gain function, emphasises spectral peak (formant) information and accounts for auditory masking effects. Overall, the multichannel SA estimators are now generalisations to the single-channel estimators for improving noise reduction, speech distortion and overall speech quality in a large region with all the available microphone channels.

The remainder of this paper is organised into the following sections: distributed microphone system (Section 2), perceptually-motivated cost functions (Section 3), parameter estimation (Section 4), simulation experiments and results (Section 5) and conclusion (Section 6).

2 Distributed microphone system

Consider an arbitrary array of M microphones, where a particular microphone is represented as $i \in [1, \ldots, M]$. At each microphone i, the source signal $s(t)$ is captured as time-delayed and attenuated coherent clean signals $c_i s(t - \tau_i)$ corrupted by additive and uncorrelated noise $n_i(t)$ with time-invariant attenuation factors c_i and time delays τ_i. Without loss of generality, the first microphone, $i = 1$, is assumed as the reference microphone with $c_1 = 1$. Based on this multichannel scenario, the propagation model in the time domain is given as

$$y_i(t) = c_i s(t) + n_i(t) \tag{1}$$

which can be accurately time-aligned through simple cross-correlation methods [18]. The frequency-domain representation of (1) is expressed as

$$\begin{aligned} Y_i(\lambda, k) &= c_i S(\lambda, k) + N_i(\lambda, k) \\ R_i(\lambda, k)\mathrm{e}^{\mathrm{j}\vartheta_i(\lambda,k)} &= c_i A(\lambda, k)\mathrm{e}^{\mathrm{j}\alpha(\lambda,k)} + N_i(\lambda, k) \end{aligned} \tag{2}$$

where λ and l represent the frame and frequency bin with noisy and clean SAs R_i and A, noisy and clean spectral phases ϑ_i and α and spectral noise N_i for each individual microphone i.

3 Perceptually-motivated cost functions

Given the multichannel Bayes risk that is represented by the average cost as

$$\begin{aligned} \Re_{\mathrm{B}} = E\big[d(A, \hat{A})\big] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} d(A, \hat{A}) \\ &\quad \times p(A, Y_1, \ldots, Y_M)\,\mathrm{d}A\,\mathrm{d}Y_1 \ldots \mathrm{d}Y_M \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} d(A, \hat{A})\,p(A|Y_1, \ldots, Y_M)\,\mathrm{d}A\right] \\ &\quad \times p(Y_1, \ldots, Y_M)\,\mathrm{d}Y_1 \ldots \mathrm{d}Y_M \end{aligned} \tag{3}$$

the minimisation of (3) with respect to $\hat{A}$ results in different estimators for a particular cost function. For the SA cost function [1]

$$d_{\mathrm{SA}}(A, \hat{A}) = (A - \hat{A})^2 \tag{4}$$

and log-spectral amplitude (LSA) cost function [2]

$$d_{\mathrm{LSA}}(A, \hat{A}) = (\log A - \log \hat{A})^2 \tag{5}$$

the resulting estimators are

$$\hat{A}_{\mathrm{SA}} = E\big[A\big|Y_1, \ldots, Y_M\big] \tag{6}$$

and

$$\hat{A}_{\mathrm{LSA}} = \exp\big(E\big[\ln(A)\big|Y_1, \ldots, Y_M\big]\big) \tag{7}$$

By modifying the cost function $d(A, \hat{A})$ in (3), the consequence is that there are many alternative estimators to the common methods of (6) and (7) for estimating the SA of the clean source signal. Since only the LSA cost function in (5) deals with a more perceptual relevant criterion and has produced higher SNR/SSNR improvements in speech quality for single-channel speech enhancement [2] than the SA cost function in (4), Loizou [3] developed several perceptually significant estimators that outperformed both the SA and LSA cost functions. From the work, the best results occurred with the WE cost function

$$d_{\mathrm{WE}}(A, \hat{A}) = A^p(A - \hat{A})^2 \tag{8}$$

and WCOSH cost function

$$\begin{aligned} d_{\mathrm{W\,COSH}}(A, \hat{A}) &= \left[\frac{1}{2}\left(\frac{A}{\hat{A}} + \frac{\hat{A}}{A}\right) - 1\right]A^p \\ &= \left[\cosh\left(\ln\left(\frac{A}{\hat{A}}\right)\right) - 1\right] \\ A^p &= \big[\cosh\big(\ln(A) - \ln(\hat{A})\big) - 1\big]A^p \end{aligned} \tag{9}$$

where p is the weighting law parameter that influences whether the corresponding estimator produces larger or smaller attenuation of the noisy SA through its spectral gain function or focuses on the spectral peaks ($p > 0$) or spectral valleys ($p < 0$).

Through the minimisation of (3) with the WE and WCOSH cost functions in (8) and (9), the subsequent true source SA estimators for distributed multichannel speech enhancement are given as

$$\hat{A}_{\mathrm{WE}} = \frac{\int_0^{\infty}\int_0^{2\pi} A^{p+1} p(Y_1, \ldots, Y_M|A, \alpha)p(A, \alpha)\,\mathrm{d}\alpha\,\mathrm{d}A}{\int_0^{\infty}\int_0^{2\pi} A^p p(Y_1, \ldots, Y_M|A, \alpha)p(A, \alpha)\mathrm{d}\alpha\,\mathrm{d}A} \tag{10}$$

and

$$\hat{A}^2_{\mathrm{W\,COSH}} = \frac{\int_0^{\infty}\int_0^{2\pi} A^{p+1} p(Y_1, \ldots, Y_M|A, \alpha)p(A, \alpha)\,\mathrm{d}\alpha\,\mathrm{d}A}{\int_0^{\infty}\int_0^{2\pi} A^{p-1} p(Y_1, \ldots, Y_M|A, \alpha)p(A, \alpha)\,\mathrm{d}\alpha\,\mathrm{d}A} \tag{11}$$

which are valid for the parameters $p_{\mathrm{WE}} > -2$ and $p_{\mathrm{WCOSH}} > -1$ and exactly equivalent to the single-channel estimators in [3] for $c_1 = 1$ and $M = 1$.

3.1 Statistical models

Based on the form of the distributions given in [19], Gaussian models are assumed for both the speech prior likelihood

$$p(A, \alpha) = \frac{A}{\pi\sigma_s^2}\exp\left(-\frac{A^2}{\sigma_s^2}\right) \tag{12}$$

doi: 10.1049/iet-spr.2012.0167

and noise likelihood

$$p\left(Y_i|A,\,\alpha\right) = \frac{1}{\pi\sigma^2_{N_i}}\exp\left(-\frac{\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2}{\sigma^2_{N_i}}\right) \qquad (13)$$

where σ^2_S and $\sigma^2_{N_i}$ are the speech and noise spectral variances. Since the WE and WCOSH estimators in (10) and (11) consists of a noise likelihood with M noisy microphone observations $\{Y_1,\,Y_2,\ldots,\,Y_M\}$ conditioned on the true SA A and true spectral phase α, the noise likelihood in (13) must account for all the available information, not simply at the ith microphone. Under the assumption of a diffuse noise field [16], the correlation of the noise between the various microphones is approximately low for high frequencies with relatively large microphone distances according to the magnitude-squared coherence (MSC). Therefore the noises are assumed uncorrelated at each of the microphones, which results in the conditional joint distribution of the noisy spectral observations $\{Y_1,\ldots,\,Y_M\}$ given the SA and spectral phase written as

$$\begin{aligned} p\left(Y_1,\,\ldots,\,Y_M|A,\,\alpha\right) &= \prod_{i=1}^{M} p\left(Y_i|A,\,\alpha\right) \\ &= \prod_{i=1}^{M}\frac{1}{\pi\sigma^2_{N_i}}\exp\left(-\sum_{i=1}^{M}\frac{\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2}{\sigma^2_{N_i}}\right) \end{aligned} \qquad (14)$$

For the distributed microphone WE and WCOSH estimators that will be derived from (10) and (11), the relationship in (14) allows for the estimation of the noise statistics at each of the corresponding microphones.

3.2 Optimal estimators

From the given statistical models (12) and (14), the SA estimators (10) and (11) are now rewritten as (see (15))

and (see (16))

As in [12], the spectral phase α is integrated out from the inner integrals in both (15) and (16) to produce (see (17))

and (see (18))

where $I_0(\bullet)$ denotes the modified Bessel function of the first kind of the zeroth order and

$$\frac{1}{\lambda} = \frac{1}{\sigma^2_S} + \sum_{i=1}^{M}\frac{c_i^2}{\sigma^2_{N_i}} \qquad (19)$$

By utilising (8.406.3) and (6.631.1) in [20] and [21], the closed-form solutions for (17) and (18) are given in terms of the confluent hypergeometric function ${}_1F_1(\bullet;\bullet;\bullet)$ described by 9.210 in [21] as

$$\hat{A}_{\mathrm{WE}} = \frac{\Gamma((p/2)+(3/2))}{\Gamma(((p/2)+1))}\frac{1}{(1/\lambda)^{1/2}}\frac{{}_1F_1\left(((p+3)/2);\,1;\,z\right)}{{}_1F_1(((p+2)/2);\,1;\,z)} \qquad (20)$$

and

$$\hat{A}^2_{\mathrm{W\,COSH}} = \frac{\Gamma((p/2)+(3/2))}{\Gamma((p/2)+(1/2))}\frac{1}{(1/\lambda)}\frac{{}_1F_1(((p+3)/2);\,1;\,z)}{{}_1F_1(((p+1)/2);\,1;\,z)} \qquad (21)$$

where

$$z = \frac{\left|\sum_{i=1}^{M}\left(c_iY_i/\sigma^2_{N_i}\right)\right|^2}{(1/\lambda)} = \frac{\left|\sum_{i=1}^{M}\sqrt{\xi_i\gamma_i}\mathrm{e}^{\mathrm{j}\vartheta_i}\right|^2}{1+\sum_{i=1}^{M}\xi_i} \qquad (22)$$

with $A_i = c_iA$, $\sigma^2_{S_i} = c_i^2\sigma^2_S$, a priori ξ_i SNR and a posteriori γ_i SNR. Unlike in [12], the ratio of the ${}_1F_1(\bullet;\bullet;\bullet)$ terms in both (20) and (21) cannot be further simplified into a single ${}_1F_1(\bullet;\bullet;\bullet)$ term since the first-term argument changes each time for the corresponding the p parameter. Based on the simplification of the term

$$\frac{1}{(1/\lambda)} = \frac{\sigma^2_S}{1+\sum_{i=1}^{M}\xi_i} \qquad (23)$$

$$\hat{A}_{\mathrm{WE}} = \frac{\int_0^\infty A^{p+2}\exp\left(-\left(A^2/\sigma^2_S\right)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2\right)/\left(\sigma^2_{N_i}\right)\right)\right)\mathrm{d}\alpha\,\mathrm{d}A}{\int_0^\infty A^{p+1}\exp\left(-\left(A^2/\sigma^2_S\right)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2/\left(\sigma^2_{N_i}\right)\right)\right)\mathrm{d}\alpha\,\mathrm{d}A} \qquad (15)$$

$$\hat{A}^2_{\mathrm{WCOSH}} = \frac{\int_0^\infty A^{p+2}\exp\left(-\left(A^2/\sigma^2_S\right)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2\right)/\sigma^2_{N_i}\right)\right)\mathrm{d}\alpha\,\mathrm{d}A}{\int_0^\infty A^{p}\exp\left(-\left(A^2/\sigma^2_S\right)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2\right)/\left(\sigma^2_{N_i}\right)\right)\right)\mathrm{d}\alpha\,\mathrm{d}A} \qquad (16)$$

$$\hat{A}_{\mathrm{WE}} = \frac{\int_0^\infty A^{p+2}\exp\left(-\left(A^2/\sigma^2_S\right)\right)\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\left(\sigma^2_{N_i}\right)\right)\right|\right)\mathrm{d}A}{\int_0^\infty A^{p+1}\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\left(\sigma^2_{N_i}\right)\right)\right|\right)\mathrm{d}A} \qquad (17)$$

$$\hat{A}^2_{\mathrm{W\,COSH}} = \frac{\int_0^\infty A^{p+2}\exp\left(-\left(A^2/\sigma^2_S\right)\right)\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\left(\sigma^2_{N_i}\right)\right)\right|\right)\mathrm{d}A}{\int_0^\infty A^{p}\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\left(\sigma^2_{N_i}\right)\right)\right|\right)\mathrm{d}A} \qquad (18)$$

the final form of the distributed multichannel WE and WCOSH estimators in (20) (equation 3.43 in [12]) and (21) (equation 3.44 in [12]) are written as

$$\hat{A}_{\mathrm{WE}} = \frac{\Gamma((p/2)+(3/2))}{\Gamma((p/2)+1)}\left(\frac{\sigma_S^2}{1+\sum_{i=1}^{M}\xi_i}\right)^{(1/2)} \times \frac{{}_1F_1\big(-((p+1)/2);\,1;\,-z\big)}{{}_1F_1(-(p/2);\,1;\,-z)} \tag{24}$$

and

$$\hat{A}_{\mathrm{WCOSH}} = \sqrt{\frac{\Gamma((p/2)+(3/2)}{\Gamma((p/2)+(1/2)}\frac{\sigma_S^2}{1+\sum_{i=1}^{M}\xi_{i\,s}}\frac{{}_1F_1(-((p+1)/2);\,1;\,-z)}{{}_1F_1(-((p-1)/2);\,1;\,-z)}} \tag{25}$$

which decays to the single-channel perceptually-motivated Bayesian noise reduction filters [3] for the case of $M = 1$. Full derivations of the estimators are presented in the appendices.

4 Parameter estimation

Based on an arbitrary reference microphone $m = 1$, the perceptually-motivated SA estimators are rewritten as

$$\hat{A}_{\mathrm{WE}} = \frac{1}{\gamma_1}\frac{\Gamma(p/2)+(3/2)}{\Gamma((p/2)+1)}\left(\frac{\xi_1\gamma_1}{1+\sum_{i=1}^{M}\xi_i}\right)^{(1/2)} \times \frac{{}_1F_1(-((p+1)/2);\,1;\,-z)}{{}_1F_1(-(p/2);\,1;\,-z)}R_1 \tag{26}$$

and (see (27))

From (26) and (27), the fundamental components are the a priori SNR ξ_i and a posteriori SNR γ_i and attenuation factors c_i. In order to fully estimate the true source signal $\hat{s}$, the SA estimators $\hat{A}$ require the estimate of the spectral phase α.

4.1 A priori and a posterior SNR

The decision-directed [1] smoothing approach is utilised to recursively estimate the a priori SNR as

$$\hat{\xi}_i = \frac{\sigma_{S_i}^2}{\sigma_{N_i}^2} = \frac{c_i^2\cdot\sigma_S^2}{\sigma_{N_i}^2} = \alpha_{\mathrm{SNR}}\cdot\hat{c}_i^2\cdot\frac{\hat{A}^2(\lambda-1)}{\sigma_{N_i}^2} + \big(1-\alpha_{\mathrm{SNR}}\big)\cdot P\big[\gamma_i(\lambda)-1\big] \tag{28}$$

and the a posteriori SNR is calculated as

$$\gamma_i = \frac{R_i^2}{\sigma_{N_i}^2} \tag{29}$$

for each channel as with $\alpha_{\mathrm{SNR}} = 0.98$ using thresholds of $\xi_{i,\,\min} = 10^{-25/10}$ and $\gamma_{i,\,\min} = 40$ (implemented as a floor on $\sigma_{N_i}^2$). By utilising the perceptually-motivated SA estimators of (24) and (25) and spectral phase [12] for distributed multichannel speech enhancement, the clean source signal is reconstructed using the overlap-add technique.

4.2 Attenuation factors

The attenuation factors c_i must be accurately estimated for estimation of the a priori SNR in (28). Based on the discussion in [12], the attenuation factors are estimated from the signal powers of the noisy observations y_i under the assumed independence of the speech s and noise n as

$$\hat{c}_i = \sqrt{\sigma_{y_i}^2 - \sigma_{n_i}^2}/\sigma_s = \sqrt{\sigma_{y_i}^2 - \sigma_{n_i}^2}/\sqrt{\sigma_{y_1}^2 - \sigma_{n_1}^2} \tag{30}$$

where $c_1 = 1$ serves as the reference microphone defined as $m = 1$. From (30), the estimated attenuation factors are simply a relative SNR ratio of the particular microphone i to a reference microphone. Thus, the value of attenuation factors can be determined by assuming a known c_i at any arbitrary reference microphone.

4.3 Spectral phase

To estimate the spectral phase α for both the single-channel and multichannel WE and WCOSH estimators, the multichannel MMSE spectral phase estimator [12, 17] is used as

$$\hat{\alpha} = \tan^{-1}\left(\frac{\sum_{i=1}^{M}\Big(\big(\sqrt{\xi_i}\big)/\sigma_{N_i}\Big)\mathrm{Im}\big(Y_i\big)}{\sum_{i=1}^{M}\Big(\big(\sqrt{\xi_i}\big)/\sigma_{N_i}\Big)\mathrm{Re}\big(Y_i\big)}\right) \tag{31}$$

which is an a priori SNR weighted sum of the noisy microphone observations. For the single-channel case with $M = 1$, the multichannel MMSE estimator in (31) simplifies to the single-channel noisy spectral phase estimator [1].

5 Simulation experiments and results

To evaluate the proposed optimal multichannel WE and WCOSH estimators derived in (24) and (25), distributed multiple microphone noisy signals were simulated using the TIMIT [22] and NOISEX [23] corpora. The noisy signals were sampled at 16 kHz and created according to (1) with equal number of uncorrelated noises as microphones. Although the signals were assumed to be perfectly synchronised without any time misalignment, previous work has illustrated that cross-correlation methods can accurately estimate time delays in the signals and effectively time align signals without any significant degradation in the enhancement results [24]. To demonstrate the best-case results, constant attenuation factors ($c_i = 1$), which represent the equal amplitude reduction between the original acoustic clean source signal and recorded noisy signals, were estimated at each of the microphones using the signal powers of the noisy signals across an entire utterance [24]. At each of the non-reference

$$\hat{A}_{\mathrm{W\,COSH}} = \frac{1}{\gamma_1}\sqrt{\frac{\Gamma((p/2)+(3/2))}{\Gamma((p/2)+(1/2))}\left(\frac{\xi_1\gamma_1}{1+\sum_{i=1}^{M}\xi_i}\right)\frac{{}_1F_1(-((p+1)/2);\,1;\,-z)}{{}_1F_1(-((p-1)/2);\,1;\,-z}}R_1 \tag{27}$$

doi: 10.1049/iet-spr.2012.0167

microphones, the noises were scaled according to the noise at the reference microphone and added to each of the attenuated clean signals at an input SNR of 0 dB. The noisy signals were truncated to produce an equal number of samples in each frame. Analysis conditions consisted of frames of 256 samples (16 ms) with 50% overlap using Hanning windows. Noise estimation was performed on five initial silence frames without any subsequent updating of the spectrum. Objective measures of SSNR [25], LLR [26] and PESQ [27] were utilised to measure the noise reduction, speech distortion and overall quality [5] averaged over ten enhanced signals, which were reconstructed using the overlap-add technique. At input SNRs ranging from −10 to +10 dB at increments of +5 dB, the input LLR and input PESQ were 1.71, 1.69, 1.64, 1.55 and 1.36 and 1.16, 1.37, 1.64, 1.94 and 2.29.

Figs. 1 and 2 show the SSNR improvement, LLR improvement and PESQ improvement as a function of the number of microphones in the array and weighting law parameter p for the multichannel WE and WCOSH estimators with white noise (pink and babble noises produced similar results), where LLR (lower scores indicate better performance) and PESQ (higher scores indicate better performance) are defined with ranges of 0–2 and 0.5–4.5. Based on the trends as a function of the weighting law parameter p, the multichannel WE and WCOSH estimators produced significant increases in noise reduction, decreases in speech distortion and increases in overall speech quality. In terms of SSNR improvement, the estimators had the largest gains at lower input SNR levels and largest increases over the single-channel baseline at higher input SNR levels. Conversely with both LLR output and PESQ output, the estimators' largest gains and largest increases over the single-channel baseline were both at higher input SNR levels. As the weighting law p was decreased from positive p values (spectral peaks) to $p = 0$ (unweighted estimators) to negative p values (spectral valleys), it is clear that the estimators had more improvements with the negative p values since the estimators focused more on the spectral valleys, where the quantisation noise was not masked by the spectral peaks (formants) and the estimators would produce audible differences between the noisy spectrum and enhancement spectrum. For SSNR improvement, it should be noted that the larger values of the weighting law parameter p, specifically at $p = 2$ for both estimators, produced little separation between the various input SNR levels and much lower gains and smaller increases over the single-channel baseline. In contrast, the LLR and PESQ improvements did not experience the same trend since they do not directly measure noise reduction as with SSNR and focus more on the actual speech quantity. Overall, the multichannel WE and WCOSH estimators performed much better than the unweighted baselines and single-channel baselines, particularly for larger negative values of the weighting law parameter p.

Tables 1 and 2 summarise the impact of the additional microphones and modified multichannel WE and WCOSH cost functions on speech enhancement. Specifically, the results were evaluated using the PESQ output metric to determine the gains in overall speech quality with increments of 0.1 for the weighting law parameter p. From Table 1, the multichannel WE cost function slightly outperformed the multichannel WCOSH cost function over the corresponding single-channel WE and WCOSH cost

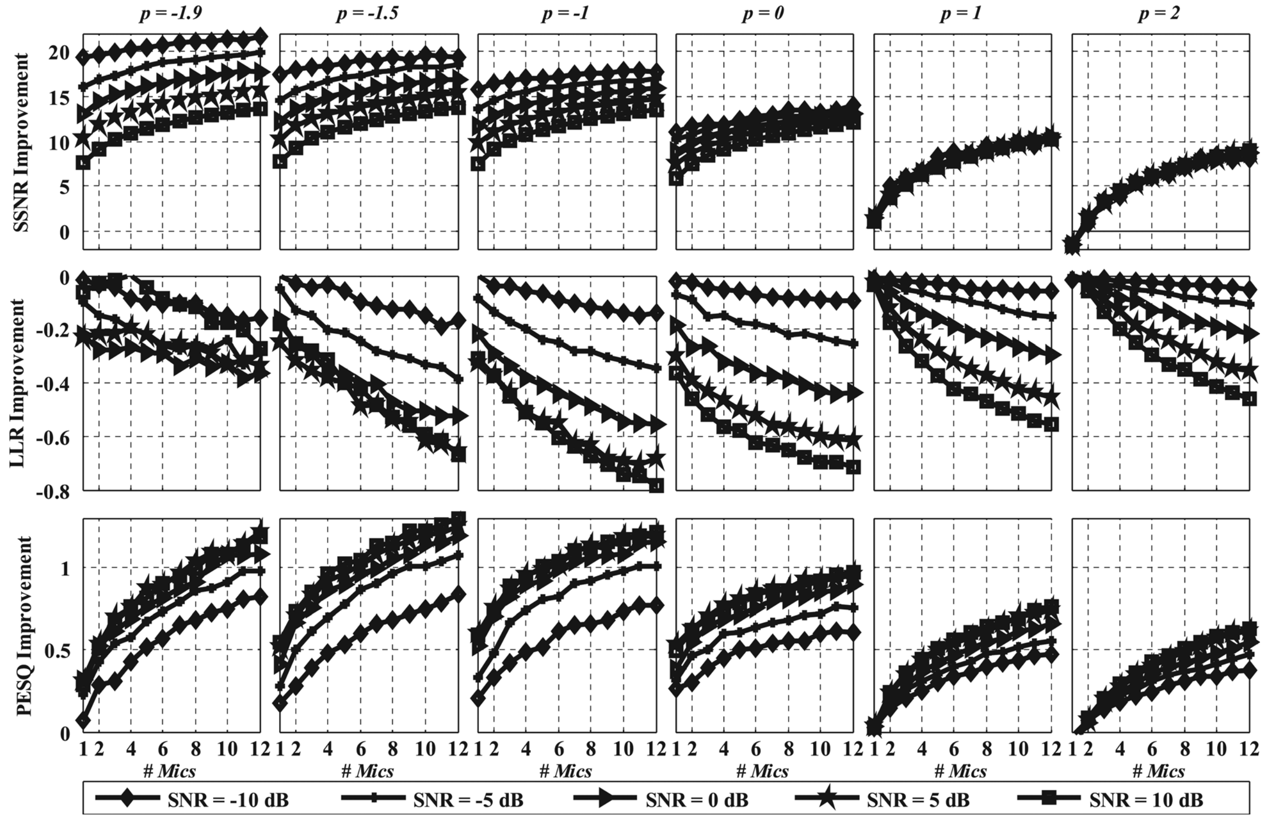


Fig. 1 *SSNR improvement, LLR improvement and PESQ improvement for multichannel WE cost function estimator for input SNR of −10, −5, 0, +5 and +10 dB with input SSNRs (−20, −15, −10, −5 and 0 dB), input LLRs (1.71, 1.69, 1.64, 1.55 and 1.36) and input PESQs (1.36 and 1.16, 1.37, 1.64, 1.94 and 2.29) in white noise*

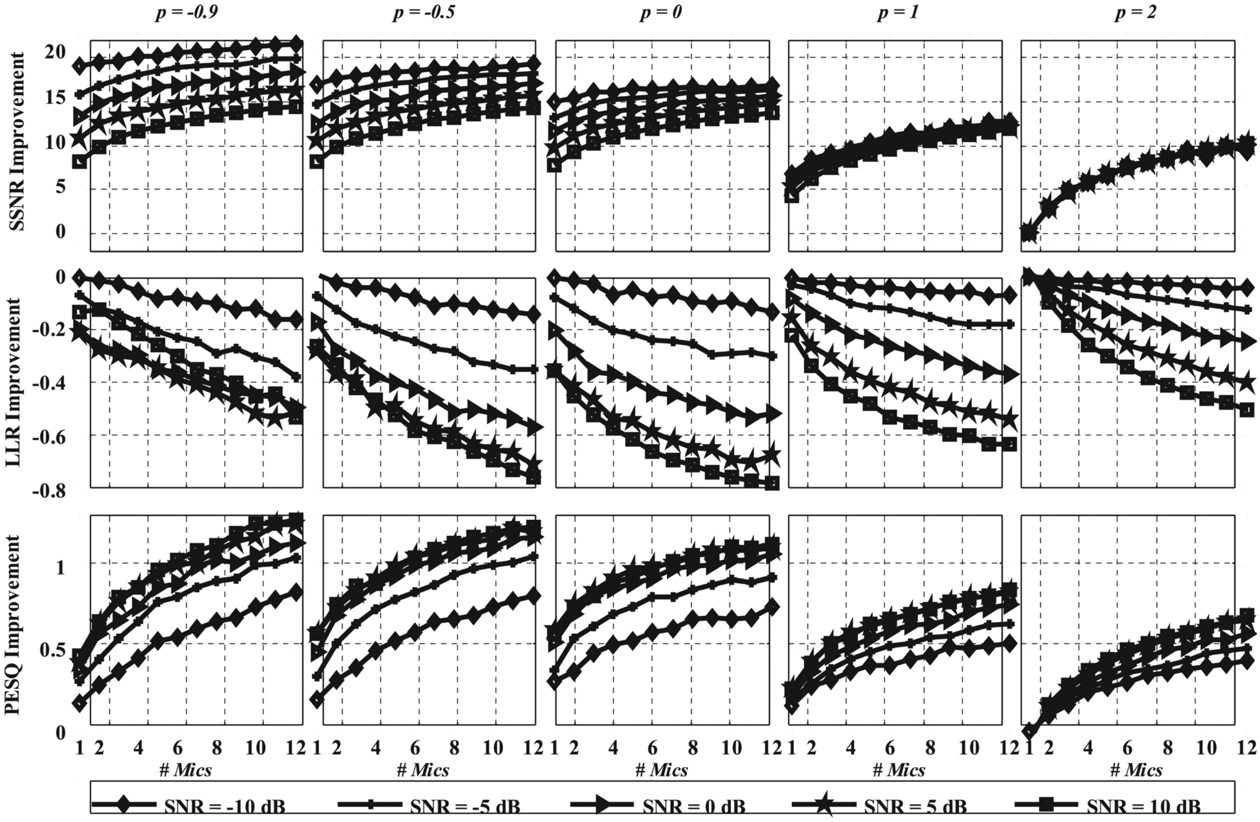


Fig. 2 *SSNR improvement, LLR improvement and PESQ improvement for multichannel WCOSH cost function estimator for input SNR of −10, −5, 0, +5 and +10 dB with input SSNRs (−20, −15, −10, −5 and 0 dB), input LLRs (1.71, 1.69, 1.64, 1.55 and 1.36) and input PESQs (1.36 and 1.16, 1.37, 1.64, 1.94 and 2.29) in white noise*

Table 1 PESQ output improvement of multichannel (12 microphones) WE and WCOSH estimators over single-channel baseline WE and WCOSH estimators for input SNR of −10, −5, 0, +5 and +10 dB with input SSNRs (−20, −15, −10, −5 and 0 dB) and input PESQs (1.36 and 1.16, 1.37, 1.64, 1.94 and 2.29) in white noise

Method	WE		WCOSH	
Input SNR, dB	Best weighting law parameter	PESQ output improvement over single channel	Best weighting law parameter	PESQ output improvement over single channel
−10	−1.6	0.71	−0.6	0.71
−5	−1.7	0.81	−0.8	0.80
0	−1.8	0.86	−0.9	0.80
+5	−1.8	0.87	−0.9	0.87
+10	−1.9	0.89	−0.9	0.83
average	−1.8	0.83	−0.8	0.80

Table 2 PESQ output improvement of multichannel (12 microphones) WE and WCOSH estimators over multichannel baseline unweighted (Euclidean and COSH, where $p = 0$) estimators for input SNR of −10, −5, 0, +5 and +10 dB with input SSNRs (−20, −15, −10, −5 and 0 dB) and input PESQs (1.36 and 1.16, 1.37, 1.64, 1.94 and 2.29) in white noise

Method	WE		WCOSH	
Input SNR, dB	Best weighting law parameter	PESQ output improvement over unweighted Euclidean	Best weighting law parameter	PESQ output improvement over unweighted cosh
−10	−1.6	0.24	−0.6	0.11
−5	−1.2	0.32	−0.9	0.15
0	−1.4	0.32	−0.8	0.16
+5	−1.7	0.30	−0.9	0.17
+10	−1.5	0.36	−0.8	0.17
average	−1.5	0.31	−0.8	0.15

doi: 10.1049/iet-spr.2012.0167

functions across all five input SNR levels for 12 microphones. In general, the additional microphone information produced increases in PESQ output improvement. With both estimators, the range of PESQ output improvement over the single-channel baseline results ranged from ~ 0.7–0.9, which is a considerable amount of gain considering that the PESQ output metric only ranges from 0.5–4.5. The most improvement was again seen with negative values of p since the weighting law parameter embedded in the cost functions concentrated more on the important spectral valleys, not spectral peaks. In terms of Table 2, the PESQ output improvements of the multichannel WE and WCOSH cost functions over the multichannel Euclidean and COSH cost functions (i.e. unweighted estimators with weighting law parameter $p = 0$) produced slightly less gains than including the additional microphone information. The range of PESQ output improvement of the modified cost functions was around 0.1–0.4 with negative values of the weighting law parameter p producing the best results, which were consistent across the different input SNR levels. In the end, the multichannel WE and WCOSH cost function estimators worked better in the PESQ output sense with additional microphones, but still have noteable gains over the unweighted multichannel WE and WCOSH cost function estimators.

6 Conclusion

In this paper, the multichannel perceptually-motivated WE and WCOSH cost functions were derived for multichannel speech enhancement using distributed microphones. The focus was to demonstrate the benefits of utilising additional microphones and modified cost functions for providing gains in noise reduction, speech distortion and overall speech quality, which were measured by the SSNR, LLR and PESQ objective metrics. From the simulation results, the multichannel WE and WCOSH cost function estimators showed significant gains in SSNR improvement, LLR improvement and PESQ improvement over both the corresponding single-channel WE and WCOSH cost functions and multichannel unweighted cost functions baseline results. From scenarios that require distributed microphones, the recommendation from this work is to employ negative values of the weighting law parameter p across all input SNR levels and noise types, emphasising spectral valleys.

7 Acknowledgments

The authors thank Dr. Philip Loizou from the University of Texas at Dallas for graciously providing further insight regarding the single-channel WE and WCOSH perceptually-motivated cost functions and LLR and PESQ objective evaluation metrics.

8 References

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

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

3 Loizou, P.C.: 'Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum', *IEEE Trans. Acoust. Speech Signal Process.*, 2005, **13**, pp. 857–869

4 Loizou, P.C.: 'Speech enhancement theory and practice' (CRC Press, 2007)

5 Hu, Y., Loizou, P.: 'Evaluation of objective quality measures for speech enhancement', *IEEE Trans. Audio Speech Lang. Process.*, 2008, **16**, pp. 229–238

6 Andrianakis, I., White, P.R.: 'Speech spectral amplitude estimators using optimally shaped gamma and Chi priors', *Speech Commun.*, 2009, **51**, (1), pp. 1–14

7 Erkelens, J.S., Hendriks, R.C., Heusdens, R., Jensen, J.: 'Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, pp. 1741–1752

8 Plourde, E., Champagne, B.: 'Auditory-based spectral amplitude estimators for speech enhancement', *IEEE Trans. Audio Speech Lang. Process.*, 2008, **16**, pp. 1614–1623

9 You, C.H., Koh, S.N., Rahardja, S.: 'Beta-order MMSE spectral amplitude estimation for speech enhancement', *IEEE Trans. Speech Audio Process.*, 2005, **13**, pp. 475–486

10 Polastre, J., Szewczyk, R., Mainwaring, A.: 'Chapter 18: analysis of wireless sensor networks for habitat monitoring', in Raghavendra, C. S., Sivalingam, K.M., Zruti, T. (Ed.): 'Wireless sensor networks' (Kluwer Academic Publishers, Norwell, MA, USA, 2004)

11 Hendriks, R.C., Heusdens, R., Kjerns, U., Jensen, J.: 'On optimal multichannel mean-squared error estimators for speech enhancement', *IEEE Signal Process. Lett.*, 2009, **16**, pp. 885–888

12 Trawicki, M.B.: 'Distributed multichannel processing for signal enhancement'. Electrical and Computer Engineering, Marquette University, Milwaukee, Dissertation, 2009, pp. 228

13 Himawan, I., McCowan, I., Sridharan, S.: 'Clustered blind beamforming from ad-hoc microphone arrays', *IEEE Trans. Audio Speech Lang. Process.*, 2001, **19**, pp. 661–676

14 Milani, A.A., Kannan, G., Panahi, I.M.S., Briggs, R.: 'A multichannel speech enhancement method for functional MRI systems using a distributed microphone array'. Annual Int. Conf. IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 2009

15 Bertrand, A., Callebaut, J., Moonen, M.: 'Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks'. Int. Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv, Israel, 2010

16 Lotter, T., Benien, C., Vary, P.: 'Multichannel direction-independent speech enhancement using spectral amplitude estimation', *EURASIP J. Appl. Signal Process.*, 2003, **2003**, (1), pp. 1147–1156

17 Trawicki, M.B., Johnson, M.T.: 'Optimal distributed microphone phase estimation'. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Taipei, Taiwan, R.O.C., 2009

18 Knapp, C.H., Carter, G.C.: 'The generalized correlation method for estimation of time delay', *IEEE Trans. Acoust. Speech Signal Process.*, 1976, **ASSP-24**, pp. 320–327

19 Martin, R.: 'Speech enhancement based on minimum mean-square error estimation and supergaussian priors', *IEEE Trans. Acoust. Speech Signal Process.*, 2005, **13**, pp. 845–856

20 Gradshteyn, I.S., Ryzhik, Z.M.: 'Table of integrals, series, and products' (Academic, New York City, NY, USA, 1980)

21 Gradshteyn, I.S., Ryzhik, Z.M.: 'Table of integrals, series, and products' (Academic, New York, 1994, 5th edn.)

22 Garofolo, J., Lamel, L., Fisher, W.: 'TIMIT acoustic-phonetic continuous speech corpus' (Linguistic Data Consortium, 1993)

23 Varga, A., Steeneken, H.J.M.: 'Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems', *Speech Commun.*, 1993, **12**, pp. 247–251

24 Trawicki, M.B., Johnson, M.T.: 'Distributed multichannel speech enhancement with minimum mean-square error short-time spectral amplitude, log-spectral amplitude, and spectral phase estimation', *Signal Process.*, 2012, **92**, pp. 345–356

25 Papamichalis, P.E.: 'Practical approaches to speech coding' (Prentice-Hall, New York, NY, USA, 1987)

26 Quackenbush, S.R., Barnwell, I.T.P., Clements, M.A.: 'Objective measures of speech quality' (Prentice-Hall, New York, 1998)

27 ITU-T: 'Recommendation P.862: perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs'. 2001.

9 Appendix

9.1 Appendix 1

In this appendix, the perceptually-motivated WE SA estimator is derived for distributed multichannel signals. By substitution of the statistical models in (12) and (14), (10) is written as (see (32))

As in Appendix D of [12], the spectral phase α is integrated out from both of the inner integrals as

$$\hat{A}_{\mathrm{WE}} = \frac{\int_0^\infty A^{p+2}\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\sigma_{N_i}^2\right)\right|\right)\mathrm{d}A}{\int_0^\infty A^{p+1}\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\sigma_{N_i}^2\right)\right|\right)\mathrm{d}A} \tag{33}$$

where $(1/\lambda)$ is defined in (19). By utilising (8.406.3) and (6.631.1) in [20] and [21], (33) is given in terms of the gamma function $\Gamma(\bullet)$ and confluent hypergeometric function ${}_1F_1(\bullet;\bullet;\bullet)$ described by 9.210 in [21] as

$$\hat{A}_{\mathrm{WE}} = \frac{\Gamma((p/2)+(3/2))}{\Gamma((p/2)+1)}\frac{1}{(1/\lambda)^{(1/2)}}\frac{{}_1F_1(((p+3)/2);\,1;\,z)}{{}_1F_1(((p+2)/2);\,1;\,z)} \tag{34}$$

where

$$\frac{1}{(1/\lambda)^{(1/2)}} = \left(\frac{\sigma_S^2}{1+\sum_{i=1}^{M}\xi_i}\right)^{(1/2)} \tag{35}$$

with $\sigma_{S_i}^2 = c_i^2\sigma_S^2$. From (34) and (35), the final closed-form solution $\hat{A}_{\mathrm{WE}}$ is given in (24) as

$$\hat{A}_{\mathrm{WE}} = \frac{\Gamma((p/2)+(3/2))}{\Gamma((p/2)+1)}\left(\frac{\sigma_S^2}{1+\sum_{i=1}^{M}\xi_i}\right)^{(1/2)} \times\frac{{}_1F_1(((p+1)/2);\,1;\,z)}{{}_1F_1((p/2);\,1;\,z)} \tag{36}$$

with free parameter $p_{\mathrm{WE}} > 2$.

9.2 Appendix 2

In this appendix, the perceptually-motivated WCOSH SA estimator is derived for distributed multichannel signals. From substitution of the statistical models in (12) and (14), (11) is written as (see (37))

After integrating out the spectral phase α from the both of the inner integrals as in Appendix D of [12], (37) is given as

$$\hat{A}_{\mathrm{WCOSH}}^2 = \frac{\int_0^\infty A^{p+2}\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\sigma_{N_i}^2\right)\right|\right)\mathrm{d}A}{\int_0^\infty A^{p}\exp\left(-A^2(1/\lambda)\right)I_0\left(2A\left|\sum_{i=1}^{M}\left(c_iY_i/\sigma_{N_i}^2\right)\right|\right)\mathrm{d}A} \tag{38}$$

where $(1/\lambda)$ is defined in (19). Through (8.406.3) and (6.631.1) in [20] and [21], (38) is given in terms of the gamma function $\Gamma(\bullet)$ and confluent hypergeometric function ${}_1F_1(\bullet;\bullet;\bullet)$ described by 9.210 in [21] as

$$\hat{A}_{\mathrm{W\,COSH}}^2 = \frac{\Gamma((p/2)+(3/2))}{\Gamma((p/2)+(1/2))}\frac{1}{\frac{1}{\lambda}}\frac{{}_1F_1(((p+3)/2);\,1;\,z)}{{}_1F_1(((p+2)/2);\,1;\,z)} \tag{39}$$

where

$$\frac{1}{\frac{1}{\lambda}} = \frac{\sigma_S^2}{1+\sum_{i=1}^{M}\xi_i} \tag{40}$$

using $\sigma_{S_i}^2 = c_i^2\sigma_S^2$. As a result of (39) and (40), the closed-form solution of $\hat{A}_{\mathrm{W\,COSH}}$ is given in (25) as (see (41))

with free parameter $p_{\mathrm{WCOSH}} > -1$.

$$\hat{A}_{\mathrm{WE}} = \frac{\int_0^\infty A^{p+2}\exp\left(-(A^2/\sigma_S^2)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2/\sigma_{N_i}^2\right)\right)\mathrm{d}\alpha\,\mathrm{d}A}{\int_0^\infty A^{p+1}\exp\left(-(A^2/\sigma_S^2)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2/\sigma_{N_i}^2\right)\right)\mathrm{d}\alpha\,\mathrm{d}A} \tag{32}$$

$$\hat{A}_{\mathrm{W\,COSH}}^2 = \frac{\int_0^\infty A^{p+2}\exp\left(-(A^2/\sigma_S^2)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2/\sigma_{N_i}^2\right)\right)\mathrm{d}\alpha\,\mathrm{d}A}{\int_0^\infty A^{p}\exp\left(-(A^2/\sigma_S^2)\right)\int_0^{2\pi}\exp\left(-\sum_{i=1}^{M}\left(\left|Y_i - c_iA\mathrm{e}^{\mathrm{j}\alpha}\right|^2/\sigma_{N_i}^2\right)\right)\mathrm{d}\alpha\,\mathrm{d}A} \tag{37}$$

$$\hat{A}_{\mathrm{W\,COSH}} = \sqrt{\frac{\Gamma((p/2)+(3/2))}{\Gamma((p/2)+(1/2))}\left(\frac{\sigma_S^2}{1+\sum_{i=1}^{M}\xi_i}\right)\frac{{}_1F_1(-((p+1)/2);\,1;\,z))}{{}_1F_1(-((p-1)/2);\,1;\,-z))}} \tag{41}$$