Stochastic-Deterministic MMSE STFT Speech Enhancement With General *A Priori* Information

Matthew McCallum, *Student Member, IEEE*, and Bernard Guillemin, *Member, IEEE*

***Abstract*—A wide range of Bayesian short-time spectral amplitude (STSA) speech enhancement algorithms exist, varying in both the statistical model used for speech and the cost functions considered. Current algorithms of this class consistently assume that the distribution of clean speech short time Fourier transform (STFT) samples are either randomly distributed with zero mean or deterministic. No single distribution function has been considered that captures both deterministic and random signal components. In this paper a Bayesian STSA algorithm is proposed under a stochastic-deterministic (SD) speech model that makes provision for the inclusion of *a priori* information by considering a non-zero mean. Analytical expressions are derived for the speech STFT magnitude in the MMSE sense, and phase in the maximum-likelihood sense. Furthermore, a practical method of estimating the *a priori* SD speech model parameters is described based on explicit consideration of harmonically related sinusoidal components in each STFT frame, and variations in both the magnitude and phase of these components between successive STFT frames. Objective tests using the PESQ measure indicate that the proposed algorithm results in superior speech quality when compared to several other speech enhancement algorithms. In particular it is clear that the proposed algorithm has an improved capability to retain low amplitude voiced speech components in low SNR conditions.**

***Index Terms*—Amplitude estimation, Gaussian processes, minimum mean-square error, phase estimation, speech enhancement, stochastic deterministic model.**

I. Introduction

BACKGROUND acoustic noise is a commonly recurring problem in applications involving the recording and processing of real world speech signals, e.g., speech recognition and radio communications. Such applications are reliant on a reasonable signal quality and their performance is often significantly compromised by low signal to noise ratios (SNRs). As a consequence, the problem of speech enhancement, which attempts to mitigate the negative effects of background acoustic noise, has received considerable attention for several decades. In many such cases only a single-channel speech signal is available. Of the available solutions to the single-channel speech enhancement problem, short-time Fourier transform (STFT) based methods achieve relatively good performance and comprise the majority [1]. It is appropriate to further categorize this class of speech enhancement algorithms into the sub-categories of spectral subtraction [2], Wiener filtering [3], and statistical approaches [4]. Of these categories, statistical approaches are perhaps the most sophisticated. Where the Wiener filtering category is restricted by linearity, and the spectral subtraction category involves largely simplified mathematical expressions, statistical approaches are strictly optimal given a set of initial assumptions and optimality criteria. Furthermore, statistical approaches have also been evaluated to be among those with the best performance [5], [6]. Within this class, Ephraim and Malah developed two speech enhancement algorithms based on the perceptual importance of STFT magnitude data, namely the minimum mean-square error (MMSE) [4] and log-MMSE [7] short-time spectral amplitude (STSA) estimators. Since their inception, much research has been undertaken to further understand the mechanism upon which these algorithms operate [8], [9] and further refine the accuracy of the underlying statistical model.

In line with Ephraim and Malah's proposal [4], STFT speech coefficients have often been assumed to be statistically independent, zero mean, complex Gaussian random variables. A number of researchers have focused on alternatives to the Gaussian assumption of the speech STFT coefficients [10], [11], whilst others have investigated how the partially incorrect assumption of statistical independence between STFT coefficients might be amended [12]. However, within the realm of statistical STSA speech enhancement, the assumption that STFT noise and speech coefficients have zero mean has received little attention. It is known that the spectral representation of random signals involving line components does indeed have a non-zero mean at the frequencies where these components exist [13]. Such signals include speech, and also noise in some scenarios. These non-zero mean characteristics are reflected within the STFT of these signals, rendering the zero mean assumption of [4] inaccurate.

In this paper an MMSE STSA speech enhancement algorithm is developed with a non-zero mean speech signal model. This model is referred to as the stochastic-deterministic (SD) signal model and the corresponding algorithm referred to as the SD MMSE STFT estimator. Incorporating this non-zero mean assumption is important as it generalizes the MMSE STSA algorithm to characterize speech simultaneously as both stochastic and deterministic. Therefore signal components that are predictable due to some *a priori* knowledge, may be considered deterministic, and if appropriately exploited, this *a priori* knowledge may be used to augment and improve the estimation of these components. For example, in this paper *a priori* knowledge of periodic components in speech (from previous STFT

Manuscript received September 09, 2012; revised December 16, 2012; accepted February 28, 2013. Date of publication March 15, 2013; date of current version March 29, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. DeLiang Wang.

The authors are with the Department of Electrical and Computer Engineering, The University of Auckland, Auckland 1142, New Zealand (e-mail: m.mccallum@auckland.ac.nz; bj.guillemin@auckland.ac.nz).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2253100

1558-7916/$31.00 © 2013 IEEE

frames) is exploited to improve the estimation of these components in the current frame.

Speech enhancement applied to scenarios involving speech and/or noise signals with deterministic components has been addressed from several perspectives in the literature. A statistical approach for entirely deterministic speech models was developed in [14]. Within the Wiener filtering framework, a speech enhancement algorithm that considers both stochastic and deterministic speech components has been presented in [15]. Hendriks *et al.* considered speech DFT coefficients to be either stochastic or deterministic and derived the MMSE estimate of the DFT coefficients based on both a hard and soft decision between the stochastic and deterministic models [16]. However, the harmonic plus noise model [17], [18] implies that each clean speech DFT coefficient is neither stochastic nor deterministic, but both simultaneously, as may be represented by a unimodal non-zero mean distribution. Such a simultaneous SD noise model has recently been considered [19], however an SD speech model of this form has remained absent from speech enhancement literature. This paper addresses this concept in the context of MMSE STSA speech enhancement.

The entirely stochastic MMSE STSA speech signal framework presented in [4], (and inherited by speech enhancement algorithms based on this literature), largely prevents the exploitation of much of the obvious structure in speech signals when considered as a function of time and frequency. The periodic structure of certain parts of a speech signal has been exploited in the rank deficient covariance matrix assumption in subspace speech enhancement [20], and can also be seen to be exploited in some STFT based speech enhancement algorithms [21]. However, consideration of this periodic structure is absent from most developments in algorithms closely related to the MMSE STSA approach due to the difficulty of incorporating this information into the entirely stochastic signal models here. As will be seen in Section IV this periodic structure is easily exploited with the use of the SD speech model. By explicitly estimating periodic signal parameters, both harmonic structure and broadband components in speech signals can be tracked across time. This allows an increase in the accuracy of estimates of the more predictable harmonic components in the signal, in both amplitude and phase.

The remainder of this paper is organized as follows. Section II establishes the notation used in this paper and discusses the SD speech model. The SD MMSE STFT speech enhancement estimator is derived in Section III giving the MMSE optimal magnitude and maximum likelihood phase of the clean speech STFT. The robust implementation of this estimator is discussed in Section IV. In Section V the developed speech enhancement system is experimentally evaluated and Section VI concludes the findings in this paper.

II. Speech Signal Statistics and Notation

If we denote the digital speech and noise signals as $\mathbf{x}[n]$ and $\mathbf{d}[n]$, respectively, the observed digital noisy speech signal is then given by[1]

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{d}[n]. \tag{1}$$

[1]Note that boldface symbols denote random variables, whilst the corresponding plain font symbols represent the values they take.

Based on the perceptual importance of the STFT magnitude in speech signals, the vast majority of STFT speech enhancement algorithms operate by modifying the magnitude spectrum of short-time observations of the signal $\mathbf{y}[n]$ to better represent that of $\mathbf{x}[n]$. Specifically, observations of $\mathbf{y}[n]$ within a short-time segment of length N, for $mN \leq n < (m+1)N$ are modified in the discrete Fourier transform (DFT) domain, where m corresponds to the window number, and N to the window length, in samples. The STFT representation of $\mathbf{y}[n]$ is obtained through the operation,

$$\mathbf{Y}[k,m] = \sum_{n=0}^{N-1} \mathbf{y}[n+mM]w[n] \exp\left\{\frac{-2\pi ikn}{K}\right\}, \tag{2}$$

for $0 \leq k < K$, where k corresponds to the STFT frequency bin number. The parameters K and $w[n]$ refer to the DFT length and windowing function respectively. M denotes the shift in samples between successive window frames. Due to the linearity of the DFT, the noisy speech signal represented in the STFT domain is a trivial extension of (1),

$$\mathbf{Y}[k,m] = \mathbf{Y}_k[m] = \mathbf{X}_k[m] + \mathbf{D}_k[m], \tag{3}$$

where, for notational simplicity, the dependence on the frame number m will be dropped from the notation where appropriate. Furthermore it will be useful to consider the phase and magnitude of $\mathbf{Y}_k$ and $\mathbf{X}_k$ explicitly as $\mathbf{Y}_k \triangleq \mathbf{B}_k e^{i\boldsymbol{\beta}_k}$ and $\mathbf{X}_k \triangleq \mathbf{A}_k e^{i\boldsymbol{\alpha}_k}$, respectively.

In Bayesian MMSE STSA speech enhancement, the aforementioned modification to observations of $\mathbf{y}[n]$ is posed as finding the best estimate of $\mathbf{A}_k$ given the observation $\mathbf{Y}_k = Y_k$, for all $0 \leq k < K$. The definition of "best" is qualified by minimizing a distortion measure. In the MMSE STSA case [4] this is the mean square error of $\mathbf{A}_k$, although a range of distortion measures have been investigated in the literature [7], [22], [23]. Under the common assumption that each STFT frequency bin is statistically independent [4], this problem may be presented as [24]:

$$\hat{A}_k = E\{\mathbf{A}_k \mid \mathbf{B}_k = B_k, \boldsymbol{\beta}_k = \beta_k\}, \tag{4}$$

where $E\{\cdot\}$ is the expectation operator.

In order to evaluate this expectation, probability density functions (pdfs) for both $\mathbf{X}_k$ and $\mathbf{D}_k$ must be assumed. It is the definition of these pdfs that is a very debatable step in deriving the solution to (4). As mentioned in Section I, the seminal paper in MMSE STSA speech enhancement [4] assumed both $\mathbf{X}_k$ and $\mathbf{D}_k$ to be zero mean complex Gaussian variables. Under these assumptions, a closed form expression for (4) may be derived [4]. These assumptions make a good compromise between established spectral estimation theory and mathematical tractability. However, more recently much literature has been published on alternative statistical models for both speech and noise, reporting promising results [11], [12], [19]. Despite these investigations, alternatives to the assumption that $\mathbf{X}_k$ has zero mean have not yet been investigated within the MMSE STSA context. It is known in the theory of harmonic analysis that a non-zero mean in signal spectra is characteristic of signals with periodic components [13]. Hence, considering $\mathbf{X}_k$ distributed

with non-zero mean can lead to the explicit consideration of periodic components in speech. The knowledge of periodic components in speech signals has long been applied in the areas of speech analysis and synthesis [17], and even in certain areas of speech enhancement [15], [16].

To incorporate explicit consideration of periodic speech components into the MMSE STSA speech enhancement framework, here clean speech signals (i.e., prior to additive noise corruption) of the following form are considered,

$$\mathbf{x}[n] = \sum_{l=0}^{L-1} r_l \cos\{2\pi f_l n + \varphi_l\} + \mathbf{u}[n]. \tag{5}$$

The signal $\mathbf{u}[n]$ is an arbitrary zero-mean stochastic process representing signal features such as unvoiced speech and highly non-stationary harmonic content. The sinusoid magnitude r_l, phase φ_l, fundamental frequency f_l, and the power spectral density (PSD) of $\mathbf{u}[n]$ may all vary in time, provided that these values vary slowly enough to be considered constant within the window length N without introducing excessive error. This definition covers a wide range of acoustic signals, and may represent the signal models defined in [4], [16], [17], [19], [25] as special conditions. For example, the signal model in [25] is represented in the case $\mathbf{u}[n] = 0$ for all n, and the signal model in [4] is represented in the case $r_l = 0$ for all l. The signal models in [16], [17], [19] may be seen as somewhere between these two extremes.

Applying the STFT to (5) it is clear that the resulting data will consist of two distinct components representing $\sum_{l=0}^{L-1} r_l \cos\{2\pi f_l n + \varphi_l\}$ and $\mathbf{u}[n]$, respectively, as follows,

$$\mathbf{X}[k,m] = \underbrace{\sum_{l=0}^{L-1} \frac{r_l}{2}\left(e^{i\varphi_l} W_{f_l}[k] + e^{-i\varphi_l} W_{1-f_l}[k]\right)}_{V[k,m]} + \mathbf{U}[k,m], \tag{6}$$

where, $W_{f_l}[k] = \sum_{n=0}^{N-1} w[n] \exp\{2\pi i n (f_l - k/K)\}$, is the discrete Fourier transform (DFT) of the windowing function $w[n]$, modulated by a complex exponential. Here by definition, for each k, $\mathbf{U}[k,m]$ is a zero-mean complex Gaussian distributed variable, and $V[k,m]$ may be considered a deterministic quantity. Therefore, under the signal model of (5), an appropriate definition of the probability density function (pdf) for $\mathbf{X}_k$ is,[2]

$$p(A_k, \alpha_k) = \frac{1}{\pi \lambda_{x,k}} \exp\left\{\frac{-\left|A_k e^{i\alpha_k} - \mu_k e^{i\theta_k}\right|^2}{\lambda_{x,k}}\right\}. \tag{7}$$

Here, $\mu_k e^{-i\theta_k}$ represents a complex non-zero mean, with amplitude μ_k, and phase θ_k. The distribution scale $\lambda_{x,k}$ is defined as $E\left\{\left|\mathbf{X}_k - \mu_k e^{-i\theta_k}\right|^2\right\}$ and is twice the variance of $\mathbf{U}[k,m]$. The statistical characterization in (7) will be referred to as the SD speech model and is the statistical model of speech that is employed throughout this paper.[3] Consideration of noise signals in this paper is restricted to zero-mean complex Gaussian distributed STFT samples, i.e.,

$$p(D_k) = \frac{1}{\pi \lambda_{d,k}} \exp\left\{\frac{-|D_k|^2}{\lambda_{d,k}}\right\}, \tag{8}$$

where $\lambda_{x,k} = E\left\{|\mathbf{D}_k|^2\right\}$.

Observations of non-zero mean DFT variables were experimentally demonstrated in [16] for synthetic signals consisting of a sinusoid embedded in white Gaussian noise (WGN). However it is interesting to see if similar observations can be made for real-world speech signals. With the use of pitch contours generated via the RAPT algorithm [26], the phase-normalized STFT observations, $\widetilde{S}[f_m, m]$, (see Appendix A) of voiced speech segments were evaluated at the frequencies f_m corresponding to the fundamental frequency and its harmonics. In making such observations over many vowels in the TIMIT database, it was seen that for a given vowel utterance and harmonic, very often these observations tend to move throughout the complex plane as shown in Fig. 1(a). Here it may be noticed that the observations shown do not appear to be drawn independently from a zero-mean complex Gaussian distribution. That is, it is clear from the figure that for each STFT frame the phase-normalized observations from past and future frames can hold much information on both the current observation's magnitude and phase. To further demonstrate this idea, observations normalized by phase from the previous frame (i.e., $\widetilde{S}[f_m, m] \exp\left\{-2\pi i M \angle(\widetilde{S}[f_{m-1}, m-1])\right\}$)[4] were analyzed.[5] The density of these observations in the complex plane are plotted in Fig. 1(b) and (c) for the vowel /æ/ (as in the word "had") and white noise, respectively. It is clear that while for the case of white noise this density appears to fit a zero-mean complex Gaussian distribution quite well (albeit not perfectly due to the correlation between frames due to windowing overlap), this is not the case for the speech segments of the vowel analyzed. Detail on how information from neighboring frames is used to estimate the mean of the current frame STFT observations is discussed in Section IV-A.

III. The SD MMSE STFT Estimator

If the variables, $\mathbf{X}_k$ and $\mathbf{D}_k$ are assumed independent for all k, then the MMSE STSA estimation problem stated in (4) may be written as,[6]

$$\hat{A}_k = \frac{\int_0^\infty A_k^2 \int_0^{2\pi} p(B_k, \beta_k \mid A_k, \alpha_k) p(A_k, \alpha_k) d\alpha_k dA_k}{\int_0^\infty A_k \int_0^{2\pi} p(B_k, \beta_k \mid A_k, \alpha_k) p(A_k, \alpha_k) d\alpha_k dA_k}, \tag{9}$$

[2]All pdfs in this paper are expressed in Cartesian coordinates with polar arguments. For notational simplicity, no subscripts will be used to identify the random variable that each pdf describes. This should be obvious from the arguments of the pdf, which will be either estimates or observations of this random variable.

[3]For $\mu_k = 0$, $p(A_k, \alpha_k)$ is identical to the complex Gaussian models employed previously in MMSE STSA speech enhancement. Also in the limit, $\lambda_{x,k} \to 0$, $p(A_k, \alpha_k) = \delta(A_k e^{i\alpha_k} - \mu_k e^{i\theta_k})$, where $\delta(\cdot)$ denotes the Dirac delta function (i.e., in this case $\mathbf{X}_k$ becomes a deterministic quantity).

[4]$\angle(\cdot)$ denotes the argument of a complex number in this paper.

[5]To demonstrate the intended point, the observations normalized by phase from the previous frame are required as the absolute phase is arbitrary, due to the arbitrary alignment of STFT windows with the signal.

[6]Equation (9) may be obtained by evaluating the expected value of the regular Bayesian integral formula, [24, (11.6)], using the identity for the integral of Cartesian functions in curvilinear coordinates (polar coordinates in this case), stated in (9) of [27, Chapter 6].

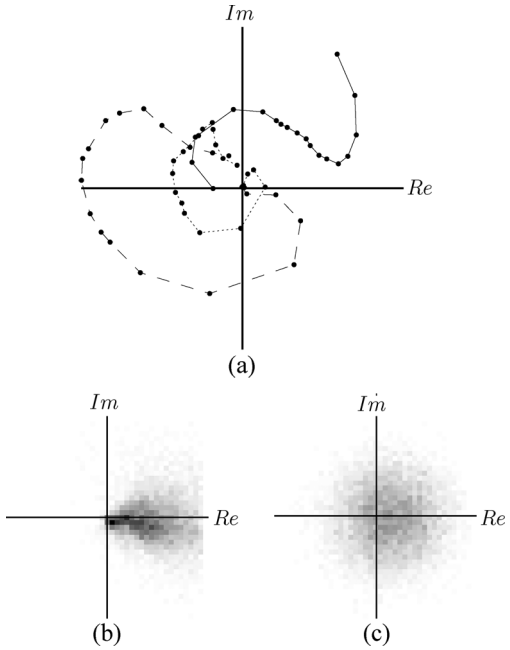

Fig. 1. (a) A set of phase normalized observations plotted on the complex plane for an utterance of the vowel /æ/. Observations are indicated with dots, and lines are drawn between consecutive observations. The dotted line, dashed line and solid line, correspond to the second, fourth and fifth harmonics respectively. These harmonics were chosen as their observations are reasonably separated in the complex plane here, (i.e., chosen for clarity). (b) and (c) refer to the density of phase normalized observations in the complex plane, (b) at the fundamental frequency of the 709 utterances of the word "had" in the TIMIT database, and (c) at 150 Hz for an equivalent length of white noise. All figures use a sampling rate of 8 kHz and STFT parameters $N = 240$, $M = 120$ and $w[n]$ was a Hamming window.

The pdf, $p(A_k, \alpha_k)$, is given in (7). From the noise statistical model defined in (8), $p(B_k, \beta_k \mid A_k, \alpha_k)$ may be derived with the use of (3),

$$p(B_k, \beta_k \mid A_k, \alpha_k) = \frac{1}{\pi \lambda_{d,k}} \exp\left\{ \frac{-\left|B_k e^{i\beta} - A_k e^{i\alpha_k}\right|^2}{\lambda_{d,k}} \right\}. \tag{10}$$

Following the derivation in Appendix B, it is found that,

$$\hat{A}_k = \sqrt{\frac{\xi_k}{\xi_k + 1} \frac{1}{\gamma_k}} \Gamma(1.5) e^{-\nu_k/2} \cdot \left[(1+\nu_k) I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right) \right] B_k, \tag{11}$$

where $\Gamma(\cdot)$ represents the gamma function, and $I_0(\cdot)$ and $I_1(\cdot)$ represent the modified Bessel functions of zero and first order, respectively. Here, ν_k is defined as,

$$\nu_k = \frac{1}{1+\xi_k} \left(\xi_k \gamma_k + \eta_k + 2\sqrt{\eta_k \gamma_k} \cos\left(\theta_k - \beta_k\right) \right), \tag{12}$$

and ξ_k, γ_k and η_k are given by,

$$\xi_k \triangleq \frac{\lambda_{x,k}}{\lambda_{d,k}}, \quad \gamma_k \triangleq \frac{B_k^2}{\lambda_{d,k}}, \quad \eta_k \triangleq \frac{\mu_k^2}{\lambda_{x,k}}. \tag{13}$$

The MMSE STSA estimator described in [4, (7)] may be considered to be a special case of the estimator described in (11) for $\eta_k = 0$. Under this condition, ξ_k and γ_k may be thought of as the *a priori* and *a posteriori* SNRs as they are defined there. However, as $\eta_k \gg 1$, then to think of ξ_k as an *a priori* SNR is no longer appropriate because most of the signal power is then represented in the term μ_k^2. In this case $\lambda_{x,k}$ better represents the uncertainty or randomness of the signal about its more predictable component μ_k. Here it may be more appropriate to refer to η_k as the signal prediction to uncertainty ratio. It may also be better to consider ξ_k as the *a priori* signal uncertainty to noise ratio. Despite these alternative descriptions, it is more meaningful in this paper to consider a different set of parameters.

Considering the discussion in Appendix B, it is clear that the MMSE STSA estimation problem results in finding the expected value of the magnitude of a complex Gaussian variable with a non-zero mean,

$$\zeta_k = \frac{\lambda_{x,k}}{\lambda_{x,k} + \lambda_{d,k}} B_k e^{i\beta_k} + \mu_k e^{i\theta_k} \left(1 - \frac{\lambda_{x,k}}{\lambda_{x,k} + \lambda_{d,k}}\right), \tag{14}$$

and scale parameter,

$$\lambda_k = \frac{\lambda_{x,k} \lambda_{d,k}}{\lambda_{x,k} + \lambda_{d,k}}. \tag{15}$$

The variable ζ_k will be referred to as the *a posteriori* mean, although it may also be thought of as the MMSE *a posteriori* complex signal estimate (i.e., $E\{\mathbf{X}_k \mid \mathbf{Y}_k = Y_k\}$). The variable λ_k will be referred to as the *a posteriori* uncertainty. It is interesting to note here that (14) is a weighted average of the form $gZ + (1-g)Z$ where $Z \in \mathbb{C}$, $g \in \mathbb{R}$ and $0 \leq g < 1$. Further insight into these parameters is largely dependent on the value of μ_k and is described here for two distinct cases: (i) $\mu_k = 0$, where the MMSE STSA estimation problem reduces to that described in [4], and (ii) $\mu_k \neq 0$, which concerns the SD speech model in (7).

A. SD MMSE STFT Estimation for the Case $\mu_k = 0$

Analysis of the SD MMSE STFT estimator for the case where $\mu_k = 0$ has been thoroughly studied in [4], [8], [9], although allowing for the possibility that $\mu_k \neq 0$ requires some further discussion. In the case $\mu_k = 0$, ζ_k is the observation Y_k with some attenuation, $\zeta_k|_{\mu_k=0} = \Omega_k Y_k$, where,

$$\Omega_k = \frac{\xi_k}{1+\xi_k}. \tag{16}$$

Ω_k is equivalent to the Wiener filter gain for DFT frequency bin k, hence in this case ζ_k is the observed STFT processed via a Wiener filter. The way in which the clean speech magnitude estimate $\hat{A}_k$ depends on ζ_k and λ_k is dependent on the ratio of the power between these two quantities (i.e., the *a posteriori* mean to uncertainty ratio),

$$\nu_k = \frac{|\zeta_k|^2}{\lambda_k}. \tag{17}$$

This ratio is expressed in (12) and for the condition here it may be simplified,

$$\nu_k|_{\mu_k=0} = \gamma_k \frac{\xi_k}{1+\xi_k}. \tag{18}$$

Ephraim and Malah [4] deduced that the MMSE STSA estimator approximates the Wiener filter under the condition

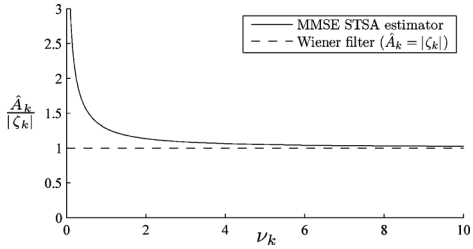

Fig. 2. The gain applied to the magnitude of the *a posteriori* mean to obtain the resulting STFT magnitude estimate $\hat{A}_k$, as a function of ν_k. This gain is also plotted for the Wiener filter for reference. For values of $\nu_k \gtrsim 4$, the difference between the Wiener and MMSE STSA estimators appears insignificant.

$\nu_k \gg 1$ (i.e., for the case $\mu_k = 0$). This is again realized in the context of parameters ζ_k and λ_k when considering that,

$$\hat{A}_k \approx |\zeta_k|, \quad \nu_k \gg 1. \tag{19}$$

However, the MMSE STSA estimator significantly differs from the Wiener filter in the case that $\nu_k \leq 1$. This difference may be observed in Fig. 2, where the relative magnitude of $\hat{A}_k$ to that of ζ_k is shown. This ratio is a function of ν_k only. As $\nu_k \to 0$, the MMSE STSA estimates of A_k become significantly large relative to ζ_k. In fact, in the Wiener filter case, $\hat{A}_k \approx 0$ for $\nu_k \ll 1$, and considering (11), for the MMSE STSA estimator we have,

$$\hat{A}_k \approx \sqrt{\frac{\frac{\pi}{4}}{1+\xi_k}\lambda_{x,k}}, \quad \nu_k \ll 1. \tag{20}$$

This insight into the MMSE STSA estimator may be described as an extreme reliance on the *a priori* information found in $\lambda_{x,k}$ under the condition that the estimate ζ_k is very uninformative in comparison. The idea of compromise between learning from the new information in B_k and relying on the *a priori* information contained within $\lambda_{x,k}$ dependent on the value of $\lambda_{d,k}$ was discussed in [4].

B. SD MMSE STFT Estimation for the Case $\mu_k \neq 0$

Extending to the case where $\mu_k \neq 0$, ζ_k in (14) may be larger or smaller in magnitude than the observation Y_k, dependent on the *a priori* mean $\mu_k e^{i\theta_k}$. The set of possible values for ζ_k in this case is shown in Fig. 3 for a given Y_k and $\mu_k e^{i\theta_k}$. It is seen here that the exact value of ζ_k depends on the Wiener term, Ω_k. In a similar way as for the case $\mu_k = 0$, a value of $\Omega_k \approx 1$ results in a value of ζ_k that is very close to the observation Y_k in the complex plane. A value of $\Omega_k \ll 1$ results in a value of ζ_k with extreme reliance on the *a priori* information, which now resides in both $\mu_k e^{i\theta_k}$ and $\lambda_{x,k}$.

For the case $\mu_k \neq 0$, ν_k is given in (12). Similar insights here may be applied as for the case $\mu_k = 0$, namely (19) and (20) still apply. However, for the case that $\nu_k \gg 1$ the resulting value of ζ_k is now a weighted combination of Y_k and the *a priori* information $\mu_k e^{i\theta_k}$ dependent on Ω_k.

It is worthwhile noting here that, because this weighting between Y_k and $\mu_k e^{i\theta_k}$ is performed in the complex plane, $|\zeta_k|$ is not only a function of the magnitudes B_k and μ_k, but also a function of the phase difference $\phi_k = |\theta_k - \beta_k|$. It is of interest to see what effect ϕ_k may have on $\hat{A}_k$. The effect of increasing

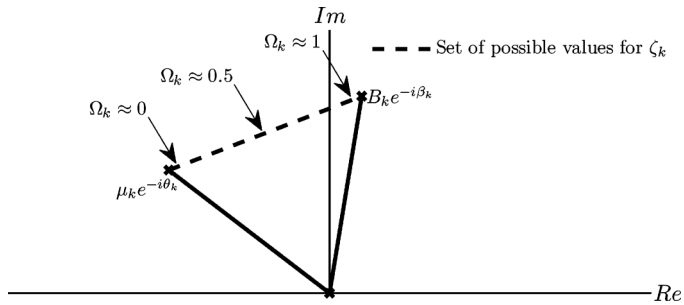

Fig. 3. The set of possible values of ζ_k in the complex plane, for given values of $B_k e^{i\beta_k}$ and $\mu_k e^{i\theta_k}$. Values of ζ_k for approximate values of Ω_k are explicitly indicated.

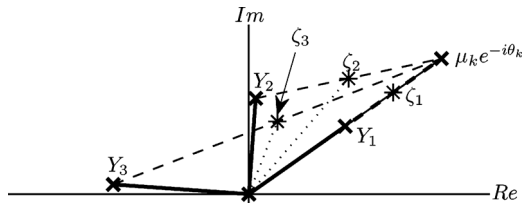

Fig. 4. Three values of *a posteriori* means ζ_1, ζ_2 and ζ_3 corresponding to three noisy spectral observations Y_1, Y_2 and Y_3, respectively, plotted in the complex plane for $\Omega_k = 0.5$ and $\mu_k e^{i\theta_k} = 2Y_1$. Here $B_1 = B_2 = B_3$, $\beta_1 = \pi/4$, and $\beta_1 < \beta_2 < \beta_3 < \pi$. The decrease in $|\zeta_k|$ is clear as $\phi_k \to \pi$.

ϕ_k for given values of Y_k, $\mu_k e^{i\theta_k}$ and Ω_k is shown in Fig. 4. Here, it appears that any increase in ϕ_k up to $\phi_k = \pi$ will result in a decrease in $|\zeta_k|$. After this conjecture it is reasonable to consider the difference of (squared) magnitudes of ζ_k for the case $\phi_k = 0$ and the case $\phi_k \neq 0$. This quantity may be considered the *a posteriori* mean reduction due to phase error and it is given by,

$$|\zeta_k|^2_{\phi_k=0} - |\zeta_k|^2_{\phi_k\neq 0} = \Omega_k B_k \mu_k (1-\Omega_k)\left[1-\cos\phi_k\right]. \tag{21}$$

From this expression it is clear that the closer ϕ_k modulo 2π is to 0 or 2π, the larger $|\zeta_k|$. Any movement further away from these values will result in a decrease in $|\zeta_k|$, with a minimum $|\zeta_k|$ occurring at $\phi_k = \pi$.

C. Phase Estimation

It may be noticed in Fig. 4 that in the SD MMSE STFT estimator for the case $\mu_k \neq 0$, the phase of ζ_k does not coincide with either θ_k or β_k, unless $\theta_k = \beta_k$ or $\theta_k = \beta_k \pm \pi$. In observing this, it may be conjectured that the best value for $\hat{\alpha}_k$ is not β_k but another value, perhaps dependent on ζ_k. Given that (11) is used to estimate $|\mathbf{X}_k|$, then a good estimate of $\boldsymbol{\alpha}_k$ may be given by that which gives the maximum likelihood under the condition $\mathbf{A}_k = \hat{A}_k$. In maximum likelihood estimation theory [24], it is well known that this value is given by a solution to,

$$\frac{\partial\left\{\log p(\hat{A}_k, \alpha_k \mid B_k, \beta_k)\right\}}{\partial \alpha_k} = 0, \tag{22}$$

where $p(\hat{A}_k, \alpha_k \mid B_k, \beta_k)$ is given in (40). In solving this equation it is found that the maximum and minimum likelihood values of $\boldsymbol{\alpha}_k$ under the condition $\mathbf{A}_k = \hat{A}_k$, are the set of values that satisfy the equation $\sin(\alpha_k - \angle(\zeta_k)) = 0$, i.e., the values $\alpha_k = \angle(\zeta_k)$ and $\alpha_k = \angle(\zeta_k) + \pi$. Upon again

differentiating (22), the value of this second derivative at these potential maxima/minima shows that the value,

$$\hat{\alpha}_k = \angle\left(\zeta_k\right), \tag{23}$$

does indeed provide a function maxima, and hence it is the maximum likelihood estimate of $\boldsymbol{\alpha}_k$ under these conditions. This provides a method of clean signal phase estimation that is potentially more accurate than the noisy phase alone. Given more accurate estimates of phase are possible, then an improvement in speech quality is likely [28]. The resulting accuracy of the estimated phase is dependent on ζ_k and hence the estimation of $\mu_k e^{i\theta_k}$, $\lambda_{x,k}$ and $\lambda_{d,k}$. For the experiments in this paper the estimation of $\mu_k e^{i\theta_k}$ and $\lambda_{x,k}$ is described in Section IV.

D. Interpretation of the SD MMSE STFT Estimator

The (11) and (23) completely describe the estimation of $\mathbf{X}_k$ for the SD MMSE STFT estimator in terms of the values Y_k, $\mu_k e^{i\theta_k}$, $\lambda_{x,k}$ and $\lambda_{d,k}$. Following the discussion here, it is perhaps more meaningful to consider that the SD MMSE STFT estimator not as the solutions to (11) and (23) but as an algorithm comprised of the following set of steps:

1) Find $|\zeta_k|$ for the case $\theta_k = \beta_k$ with the use of (14)
2) Reduce the mean magnitude according to phase difference ϕ_k with the use of (21)
3) Calculate the estimate $\hat{A}_k$ by applying the gain function plotted in Fig. 2 to $|\zeta_k|$
4) Calculate the phase estimate $\hat{\alpha}_k$ with the use of (23)
5) Use the data $\hat{A}_k e^{i\hat{\alpha}_k}$ to re-synthesize the estimated clean signal via the inverse DFT and overlap-add method [1].

With the series of steps above, the values of ζ_k, ϕ_k and ν_k are explicitly featured in steps 1, 2 and 3, respectively.

IV. Robust Implementation

A. Speech Parameter Estimation

The SD MMSE STFT speech enhancement estimator derived in Section III and requires the estimation of $\mu_k e^{i\theta_k}$ and $\lambda_{x,k}$. If the speech signal is considered to be of the form in (5) then the estimation of each of these statistical parameters reduces to the estimation of f_l, r_l, φ_l and the PSD of $\mathbf{u}[n]$.

If it is considered that the values f_l are harmonically related for all l, then estimation of the parameters f_l, r_l, φ_l may be reduced to the estimation of f_0 and r_l, φ_l for $0 \leq l < L$ where $L = \left\lfloor f_0^{-1} \right\rfloor$ and $\lfloor\cdot\rfloor$ denotes the floor operator. The harmonic relationship of f_l is a typical assumption made in the research of pitch estimation algorithms [29].

For the purposes of the investigations in this paper, only a single speech source is assumed present. With regards to the estimation of the pitch (f_0) of a single talker in the presence of a non-harmonic noise source, the maximum-likelihood pitch estimation technique is known to perform well in low SNRs of approximately 0 dB [30], and it is the method employed in this paper. Specifically $\hat{f}_0[m]$ will denote the maximum-likelihood estimate of f_0 at frame m.

There are many methods for estimating r_l and φ_l in the literature [13], [30]. Recently, several publications have shown that the use of the methods in [13] are effective for spectral estimation in STFT based speech enhancement [31], [32]. In the experiments conducted in this research, the results confirm that this provides an effective method for estimating spectral amplitude and phase information for speech signals over the typical sample lengths of the STFT. Hence, in this paper the method of harmonic analysis in [13, Chapter 13] is used to provide estimates for r_l and φ_l. Specifically the estimate of $r_l e^{-i\varphi_l}$ at frequency f_l obtained from an arbitrary data source, $s[n]$, will be denoted $\rho\left\{f_l, s[n]\right\}$ and is obtained with the use of [13, equation (13.6)].

Acknowledging that during a vowel segment of speech, the complex mean of periodic components in speech is likely to be slowly varying, given a sufficiently short frame length and a sufficiently large frame overlap. It is then possible to further reduce the variance of the estimator $\rho\left\{f_l, s[n]\right\}$ with the use of information from previous frames. In the experiments in this paper, the final estimate, $\rho_{av}[l,m]$, of $r_l e^{i\varphi_l}$ is obtained as,

$$\begin{aligned}\rho_{av}[l,m] = {} & \beta e^{2\pi i \hat{f}_0[m-1]M} \\ & \cdot \rho\left\{ l\hat{f}_0[m-1], \frac{\mathcal{F}^{-1}\left\{\hat{X}[k,m-1]\right\}}{W(0)} \right\} \\ & + (1-\beta)\rho\left\{ l\hat{f}_0[m], y[n+mM] \right\},\end{aligned} \tag{24}$$

for $0 \leq n < N$ and $0 \leq l < L$. Here $\mathcal{F}^{-1}\{\cdot\}$ denotes the inverse DFT operation, β is a mean estimate smoothing parameter and $\hat{X}[k,-1] = 0.1$. The term $e^{2\pi i \hat{f}_0[m-1]M}$ is used to normalize phase for the expected shift in time between successive windows (see Appendix A). A value of $\beta \approx 0.9$ has provided good results in the experiments in this paper. Finally the estimate, $\hat{\mu}_k e^{i\hat{\theta}_k}$, of $\mu_k e^{i\theta_k}$ at frame m may be obtained by applying the STFT operation (i.e., that indicated in (2)) to the signal in (5) with $\mathbf{u}[n] = 0$ and estimated parameters $r_l = |\rho_{av}[l,m]|$, $\varphi_l = \angle(\rho_{av}[l,m])$ and $f_l = l\hat{f}_0[m]$.

The "decision-directed" approach to the estimation of ξ_k or equivalently $\lambda_{x,k}$, presented in [4, (51)], has been identified as a crucial feature of the MMSE STSA estimator in providing a high level of signal quality [9]. To maintain the established advantages of the decision-directed approach, it is important that the method of estimating $\lambda_{x,k}$ used here is equivalent to this approach for the case $\hat{\mu}_k = 0$. Considering the definition of $\lambda_{x,k}$ as $E\left\{\left|\mathbf{X}_k - \mu_k e^{i\theta_k}\right|^2\right\}$, the extension of the decision-directed approach for the estimate of $\lambda_{x,k}$ in the case $\hat{\mu}_k \neq 0$ is trivial. If we denote $\hat{\mu}_k e^{i\hat{\theta}_k}$ for frame m as $\varrho_k[m]$, then the estimation of $\lambda_{x,k}$ may be described as,

$$\begin{aligned}\hat{\lambda}_{x,k}[m] = {} & \alpha\left|\hat{X}_k[m-1] - \varrho_k[m-1]\right|^2 \\ & + (1-\alpha)\left|Y_k[m] - \varrho_k[m]\right|^2,\end{aligned} \tag{25}$$

where α may be seen as a smoothing parameter. For the case $\mu_k[m] = 0$ for frames m and $m-1$, $\hat{\lambda}_{x,k}[m]/\hat{\lambda}_{d,k}[m]$ does indeed represent the decision-directed estimator of [4] provided that $\hat{\lambda}_{d,k}[m] = \hat{\lambda}_{d,k}[m-1]$. For the experiments in this paper, (25) is used to estimate $\lambda_{x,k}[m]$.

B. Robustness to Estimation Errors

After the discussion in Sections III-B and IV-A, it is clear that spurious values of $\mu_k e^{i\theta_k}$ can result in correspondingly spurious

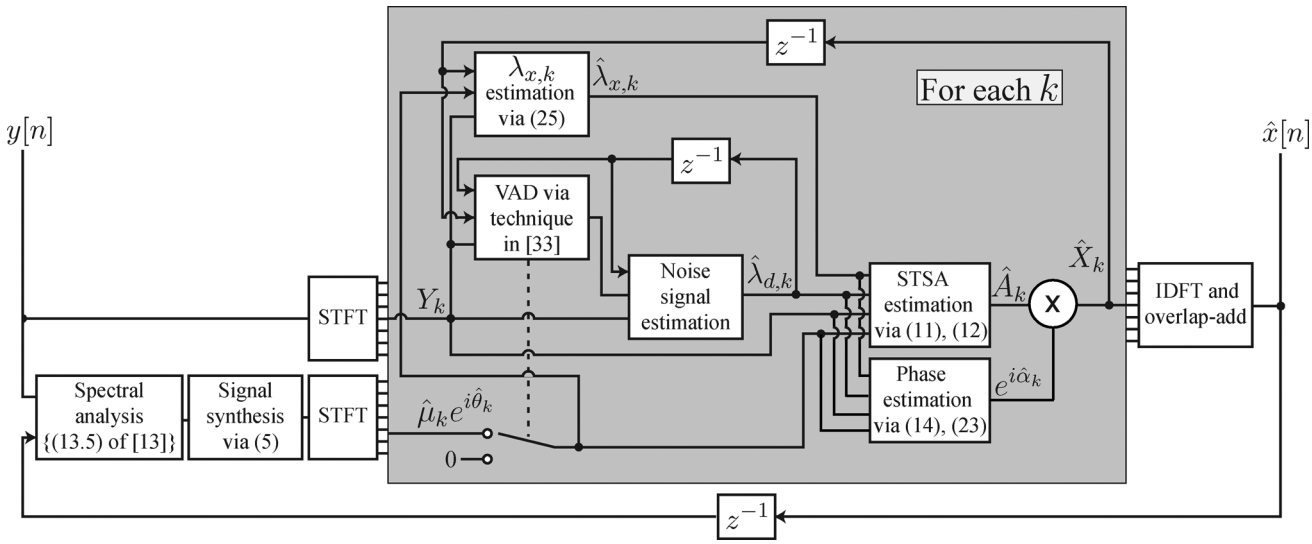

Fig. 5. A system diagram highlighting the major functions within the SD MMSE STFT speech enhancement algorithm. Signal flow is from left to right, except where indicated otherwise via the use of arrow heads. Usage of the z^{-1} term indicates a single frame delay (i.e., it occurs where data from the previous window/frame is used). The switch shown represents the use of a non-zero complex mean only in frames where voice is detected present, as discussed in Section IV-B. The method of noise signal estimation used in this paper is described in Section V-B1.

values of ζ_k. Hence the resulting value of $\hat{x}[n]$ may include large spurious spectral peaks depending on the estimation of $\mu_k e^{i\theta_k}$. Due to the recursive estimation methods in (24) and (25) such peaks will likely cause further errors in $\hat{\mu}_k e^{i\hat{\theta}_k}$ and $\hat{\lambda}_{x,k}[m]$. Here two important methods of ensuring robust parameter estimation are discussed.

STFT analysis of an arbitrary speech signal is expected to present a number of STFT frames that do not contain any periodic components, for example, those representing speech pauses and a range of unvoiced consonants. Hence, setting $\hat{\mu}_k \neq 0$ in these frames would be theoretically inaccurate. Furthermore a high proportion of these STFT frames that are lacking in periodic components are reliably detectable in relatively low SNRs with a simple voice activity detection (VAD) algorithm. In the experiments in this paper, the values $\hat{\mu}_k = 0$ for all k were imposed where voice was detected absent via [33]. Where background noise is sufficiently stationary, a VAD algorithm will be sufficient to estimate $\lambda_{d,k}$ also, hence such an algorithm may be considered of almost no additional computational expense.

To further improve the speech enhancement process under poor estimation of $\mu_k e^{i\theta_k}$ in voiced speech segments, a method of reducing the amplitude of spurious estimates must be considered. Accounting for speech presence uncertainty as described in [4] is well known to notably improve the performance of the MMSE STSA estimator. Conceptually, this method may be thought to pull estimates of $\mathbf{X}_k$ towards zero when the corresponding observation Y_k is more likely to be noise than speech. Therefore, applying similar principles here may increase the robustness of the SD MMSE STFT estimator. It is known for a given set of noise and speech pdfs, the MMSE STSA estimator under speech presence uncertainty is [4],

$$\hat{A}_k = \frac{\Lambda(Y_k, q_k)}{1 + \Lambda(Y_k, q_k)} E\left\{A_k | \mathbf{Y}_k = Y_k, H_k^1\right\}, \qquad (26)$$

where H_k^1 represents the hypothesis that speech is present in the observation Y_k. For the SD MMSE STFT estimator, $E\left\{A_k | \mathbf{Y}_k = Y_k, H_k^1\right\}$ is given in (11). Under the statistical models assumed in (8) and (7), the generalized likelihood ratio, $\Lambda(Y_k, q_k) \triangleq c_k p\left(Y_k | H_k^1\right) / p\left(Y_k | H_k^0\right)$, must be derived. Here $c_k \triangleq (1 - q_k)/q_k$, and q_k is the probability of speech absence in Y_k. Applying the pdfs of (7) and (8), $\Lambda(Y_k, q_k)$ is given by,

$$\Lambda(Y_k, q_k) = \frac{c_k}{1 + \xi_k} \cdot \exp\left\{\frac{\xi_k}{1 + \xi_k} \cdot \left(\gamma_k - \eta_k + \sqrt{\frac{\gamma_k \eta_k}{\xi_k}} \cos\left(\theta_k - \beta_k\right)\right)\right\} \qquad (27)$$

It should be noted here that with the mean and variance estimation described in (24) and (25), the consideration of signal presence uncertainty mentioned here was required to provide any improvement in speech quality. Without it, erroneous estimation of μ_k and θ_k caused too many artefacts for the proposed speech enhancement algorithm to prove effective. In particular, the erroneous estimation of μ_k and θ_k begins to consistently become audible with $q_k < 0.5$. However, the consideration of speech presence uncertainty presented here is very effective in attenuating spurious estimates of $\mu_k e^{i\theta_k}$ due to spurious estimates of either f_0, r_l, or φ_l. This was observed to the extent that the algorithm continued to perform well in the case where large errors were deliberately forced on these parameters, as will be seen in Section V-B3. Whether the proposed algorithm can perform without the consideration of speech presence uncertainty under more reliable estimation of $\mu_k e^{i\theta_k}$ is left for further research.

Algorithm Overview

The implementation of the complete system is shown in Fig. 5. Here the dependence of each of the system functions on others is clearly indicated.

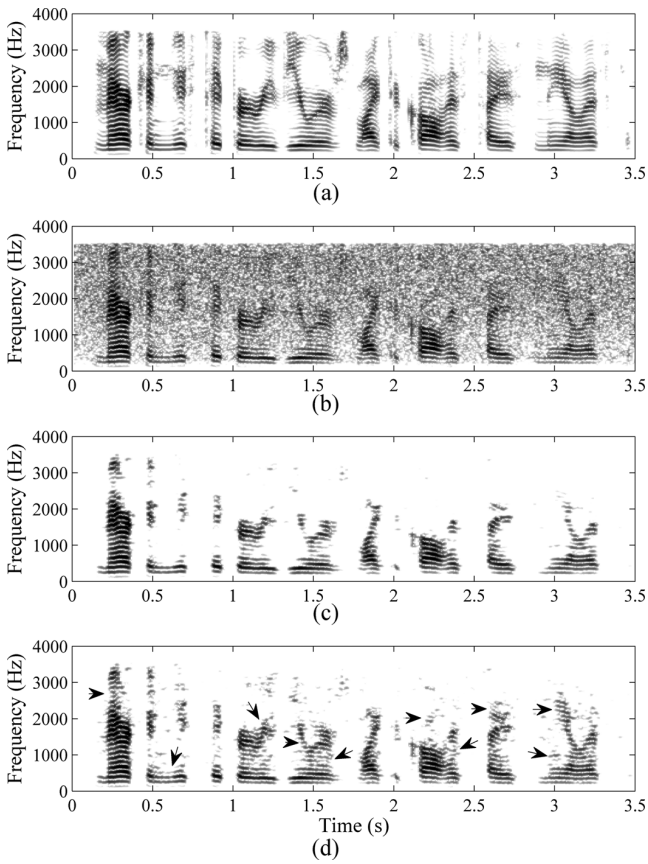

Fig. 6. A series of spectrograms indicative of the performance of the MMSE STSA and SD MMSE STFT algorithms. (a) A clean speech signal of the utterance "American newspaper reviewers like to call his plays nihilistic," (b) the clean speech signal in (a) corrupted by WGN at 5 dB SNR and bandwidth limited according to [35], (c) the corrupted speech signal in (b) processed by the MMSE STSA algorithm, (d) the corrupted speech signal in (b) processed by the SD MMSE STFT algorithm. Notable speech components present in (d) but not (c) are explicitly indicated.

V. Experimental Evaluation

A. Demonstration With Specific Signals

At the output of the SD MMSE STFT algorithm, weaker deterministic components in speech are retained when compared to the output of the MMSE STSA algorithm. This observation is demonstrated in the spectrograms seen in Fig. 6, where some areas in which weak speech components are retained in the SD MMSE STFT algorithm, but not in the MMSE STSA algorithm, are explicitly indicated. In hypothesizing why this may be, there are two obvious advantages the SD MMSE STFT algorithm has over the MMSE STSA algorithm. Firstly, with the use of (24), information specific to a periodic component is tracked across changes in frequency in the proposed algorithm, whereas the MMSE STSA algorithm considers each DFT coefficient entirely independently. Secondly, a lower averaging coefficient of the mean spectrum ($\beta = 0.9$ in (24)) compared to that of the stochastic spectrum ($\alpha = 0.98$ in (25)) means that the SD MMSE STFT algorithm is able to more closely track likely speech components (i.e., at the harmonics of the fundamental frequency).

To demonstrate the ability of the SD MMSE STFT algorithm to better retain weak speech components, another experiment

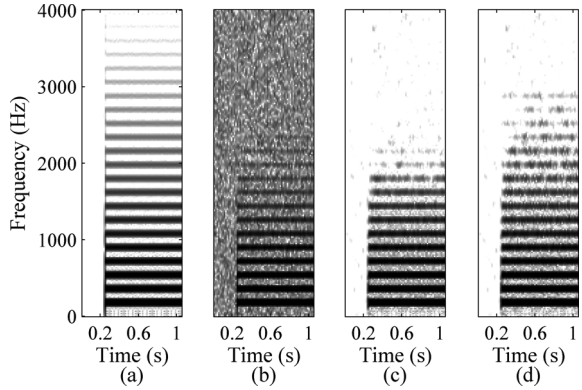

Fig. 7. A set of spectrograms indicating the ability of the MMSE STSA and SD MMSE STFT algorithms to enhance a signal with known properties. (a) The clean synthetic harmonic signal, (b) the clean signal in (a) corrupted by WGN to give a minimum sinusoidal component SNR of -40 dB and a maximum of 10 dB, (c) the corrupted signal in (b) enhanced by the MMSE STSA algorithm, (d) the corrupted signal in (b) enhanced by the SD MMSE STFT algorithm.

was conducted. Here the clean signal was comprised of 0.25 seconds of silence followed by 0.815 seconds of a harmonic signal. The harmonic signal included 22 sinusoids harmonically related with a fundamental frequency of 180 Hz. The SNR of these sinusoids was varied at equal intervals on a dB scale from 10 dB at 180 Hz to -40 dB at 3.96 kHz, and each sinusoid had a constant phase offset randomly selected from a uniform distribution between 0 and 2π. This clean signal was corrupted by WGN and enhanced by both the MMSE STSA algorithm, and the SD MMSE STFT algorithm. Both algorithms were given perfect knowledge of $\lambda_{d,k}$ for all k, and the SD MMSE STFT algorithm was given perfect knowledge of the fundamental frequency of the clean harmonic signal. The results from this experiment are displayed in Fig. 7. When observing the harmonic components at frequencies of approximately 2500 Hz in Fig. 7(c) and (d) (for example the 12th harmonic at 2340 Hz and an SNR of -18.57 dB) it is clear that the SD MMSE STFT algorithm is more capable of retaining these components than the MMSE STSA algorithm.

Fig. 8 highlights another point of difference between the MMSE STSA and SD MMSE STFT algorithms seen in the same experiment. Here, observing the spectral amplitude of a frequency bin centered on a particularly low SNR sinusoidal component the SD MMSE STFT algorithm is able to respond to the onset of this component in a smaller number of frames than the MMSE STSA algorithm. The preservation of transient speech components is known to be important for speech intelligibility [34].

B. Objective Evaluation Experiments

Here the proposed estimator is evaluated under a range of conditions. Specifically four alternative algorithms are chosen for comparison: the MMSE STSA algorithm with speech presence uncertainty (M-SPU) [4], the log-MMSE algorithm (M-LOG) [7] and a Wiener based algorithm that incorporates both speech presence uncertainty and stochastic and deterministic speech components (W-SD), i.e., the algorithm in [16] that relies on a Gaussian model for stochastic speech components. The proposed algorithm will be referred to as M-SD.

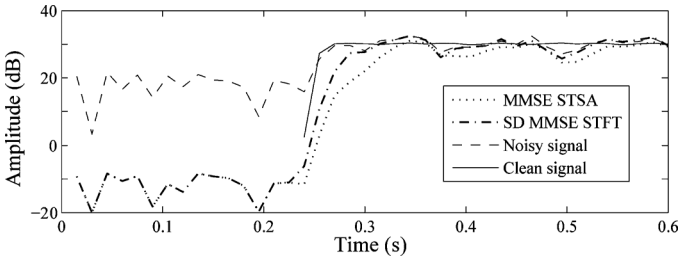

Fig. 8. Time magnitude plots of a specific bin of the spectrograms in Fig. 7 (i.e., the bin corresponding to the peak of the 8th harmonic of the clean signal). MMSE STSA, SD MMSE STFT, Noisy signal and Clean signal correspond to Fig. 7(d), (c), (b) and (a), respectively.

The M-SPU and M-LOG algorithms work as broad reference algorithms as they are very common in experimental comparisons in speech enhancement literature including many of the algorithms developed with alternative statistical speech models and distortion measures considered in [11], [12], [22]. These algorithms are also known to provide speech enhancement results that perform well with regards to subjective measures of speech quality [5]. In addition, the M-SPU algorithm represents a special case of the estimator developed in this paper, where $\hat{\mu}_k = 0$ for all k, m. The W-SD algorithm provides performance that is both indicative of Wiener filtering algorithms and of current speech enhancement theory that considers deterministic speech components.

The M-SD algorithm will be evaluated with both the pitch track obtained from the noisy speech signal as described in Section IV-A, the pitch track obtained from the clean speech signal also via maximum-likelihood techniques [30] (with speaker dependent frequency limits), and a randomly generated pitch track where each frequency sample was randomly and uniformly distributed between 70 Hz and 350 Hz. These methods are denoted in the results here as M-SD-N, M-SD-C, and M-SD-R, respectively. M-SD-C demonstrates the algorithm's potential performance given more accurate estimates of f_0 may be possible, and M-SD-R demonstrates the effectiveness of the robustness considerations in Section IV-B, in the case of erroneous pitch estimation.

The W-SD algorithm requires estimates of the frequency of sinusoids in speech. The estimates provided here (in oscillations per sample) are $[f_0, 2f_0, \ldots, Lf_0]$ where $L < f_0^{-1}$. This differs from the methods in [16] in that there is a harmonic restriction on the relationship of sinusoids. This ensures the algorithm's consistency with the proposed estimator. To eliminate any handicap from imperfect pitch estimation, the results here only show the W-SD algorithm with the more ideal case of f_0 estimated from the clean speech signal.

Algorithm Configuration: For all algorithms, $w[n]$ was a Hamming window, $N = 240$, $M = 120$ and $K = 480$. Each of the algorithms tested make use of the decision-directed estimate of ξ_k. With regard to this estimate, $\alpha = 0.98$ and a minimum value of $\xi_{\min} = -25$ dB was imposed. These values were chosen for consistency across all algorithms and to follow the recommendations in the literature, however, in informal tests it was observed that M-SD performed better with a slightly decreased value of $\alpha \approx 0.97$. All tested algorithms operate under speech presence uncertainty (i.e., according to (26)). For the M-SPU and M-SD algorithms $q_k = 0.65$. This value was empirically determined to provide effective results in both algorithms. The W-SD algorithm was configured as suggested in [16].

For the proposed algorithm, β of (24) was set to $\beta = 0.9$. To prevent artifacts due to inaccurate frequency estimates at high frequencies and estimates outside of the signal bandwidth (approximately 300 Hz to 3.6 kHz for radio communications), deterministic component estimates obtained via (24) were only estimated for frequencies less than 3 kHz. Finally, estimates of $\lambda_{d,k}$ were obtained by averaging over the magnitude spectra in the first 495 ms of data (i.e., for $m < 33$), where during this segment $x[n] = 0$. This noise only segment was removed from the signal prior to objective performance evaluation.

1) Data: All objective results presented in this paper were averaged over 300 speech utterances from the TIMIT database. Specifically, these utterances consisted of the "phonetically-diverse" sentences from 100 different speakers (equal parts male and female) belonging to the TIMIT test set. Three types of noise were used, including WGN, factory noise and babble noise. WGN was chosen as it closely fits the assumptions made for the noise statistics here [4]. Factory noise and babble noise represent less and less stationary noise sources, and were chosen to test the algorithm's robustness to variable noise conditions. Babble noise is a particularly problematic case as it may contain many spurious deterministic components.

Each speech utterance was combined with a randomly selected noise segment from each noise source at SNRs of -5 dB, 0 dB, 5 dB, 10 dB and 15 dB, according to the recommendations in [36]. Perceptual evaluation of speech quality (PESQ) [37], was used as an objective quality measure here. This measure is well known to correlate highly with mean opinion subjective test scores [38], and is a measure adopted in much of the speech enhancement literature.

2) Results: The results of the objective tests conducted here are shown in Fig. 9. For clarity of presentation, only the M-SD-N algorithm is shown in Fig. 9(a)–(c), whilst the relative scores of M-SD-C, M-SD-R and M-SPU are compared to M-SD-N in Fig. 9(d)–(f). Noting that M-SPU represents the algorithm in this paper without any consideration of deterministic components, it is clear that under all conditions, use of the SD signal model improves the PESQ results regardless of frequency estimation. Specifically the most substantial performance improvement is observed at mid to low SNRs (-5 dB to 5 dB). In the case of white noise, the averaged PESQ difference between M-SPU and M-SD-C is between 0.13 and 0.16 for these SNRs. A more significant difference was observed when considering only male speakers. The maximum PESQ improvement observed over the M-SPU algorithm was under the 0 dB white noise condition, with a PESQ difference of 0.44.

Fig. 9(d)–(f) demonstrate that the accuracy of frequency estimation does affect the performance of the algorithm. However, it is interesting to note here that the M-SD-R algorithm has performance comparable to the M-SPU algorithm. This is a testament to how effective the consideration of speech presence uncertainty is in reducing the effect of spurious DFT coefficient estimates due to erroneous pitch estimation. That is, in the experiments conducted, it was observed that given an erroneous pitch

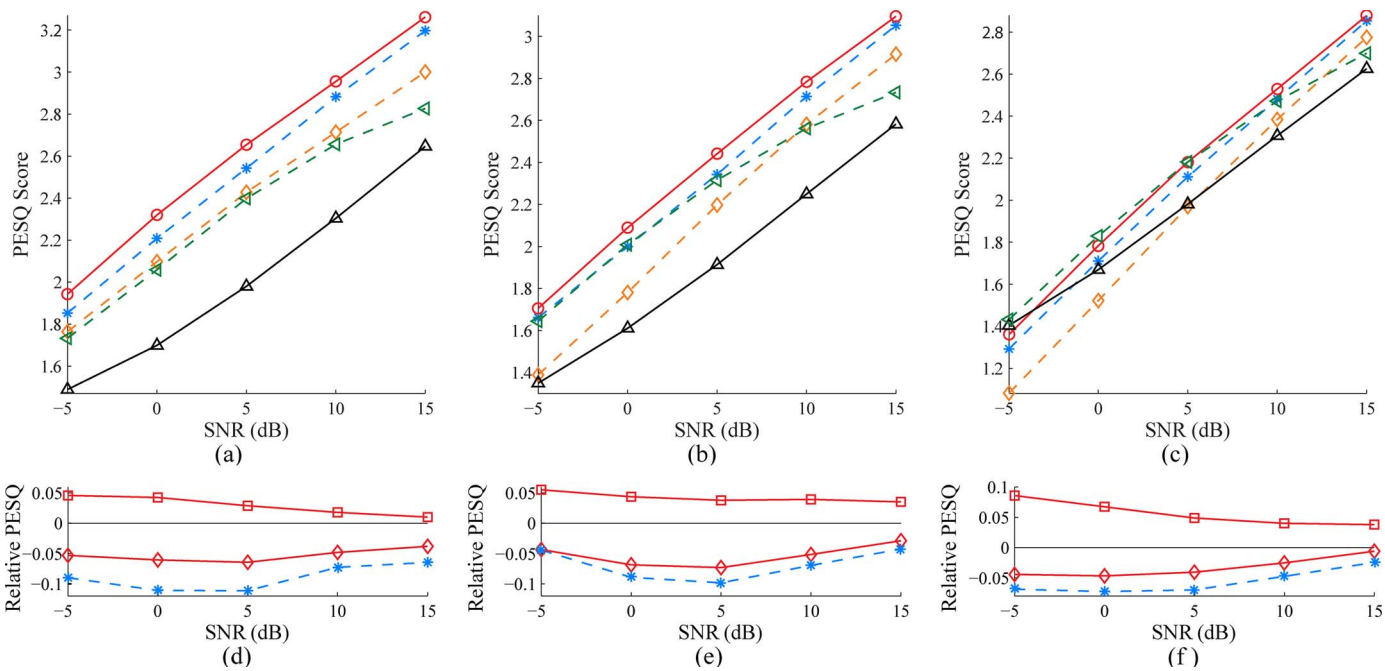

Fig. 9. PESQ testing results performed under three noise types (a, d) WGN, (b, e) factory floor noise, (c, f) babble noise. Figures (a–c) present absolute PESQ results for the proposed and reference algorithms, and Figures (d–f) present the results of M-SD-C, M-SD-R and M-SPU relative to the PESQ score for M-SD-N. For Figures (a–c), (○) M-SD-N, (∗) M-SPU, (◊) W-SD, (◁) M-LOG, (△) Noisy Signal. For Figures (d–f), (□) M-SD-C, (◊) M-SD-R, (∗) M-SPU.

estimate, negligible degradation was observed at the output of the algorithm, however a reasonably accurate pitch estimate improved recovery of the clean signal STFT.

3) Subjective Quality of Results: It was observed in informal listening tests that the M-SD algorithm had a clearer and less muffled quality when compared to many other algorithms, particularly the M-SPU algorithm which may be thought of as the most closely related to M-SD. In particular the recovered voiced speech of the M-SD algorithm sounded louder and crisper (i.e., appeared to contain more high frequency content) when compared to the M-SPU algorithm, where the processed audio of the M-SPU algorithm at low SNRs sounded less natural and reminiscent of talking through a pillow or other low-pass environment. Such a result is consistent with the observations made in Section V-A. The differences between the M-SD and other algorithms were most audible with male speakers, which is likely attributable to the denser harmonic content of male speech compared to that of female speech.

This improved clarity of voiced speech was not observed in the W-SD case, which appeared to emphasize lower frequency voiced speech content. The residual noise of the W-SD algorithm also had a musical quality that was not present in the M-SD and M-SPU algorithms. This is a well known difference between these two types of algorithms [9]. In the white noise case, the residual noise of the M-SPU and M-SD algorithms was only clearly audible at an SNR of -5 dB. Here it had a broadband characteristic that might be subjectively described as slightly watery.

Results from the M-SD-R tests sounded almost identical to those from the M-SPU algorithm, although had a slight synthetic character. That is, faint artefacts were observed in voiced speech regions that sounded less natural than the M-SPU algorithm. Such artefacts were rare in the M-SD-N tests and almost non-existent in the M-SD-C tests.

VI. Conclusions

In this paper a non-zero mean MMSE STSA speech enhancement algorithm has been proposed. The non-zero mean concept is implied by the harmonic plus noise speech model and supported by STFT observations of speech signals. Under the non-zero mean assumption, analytical expressions were obtained for estimates of the clean speech STFT amplitude in the MMSE sense, and clean speech STFT phase in the maximum-likelihood sense. Furthermore, estimation methods were proposed for both the *a priori* mean and uncertainty parameters of signals with harmonically related sinusoidal components. As a final step towards a robust implementation, the SD MMSE STFT estimator was considered under speech presence uncertainty, which was necessary to counter the negative effects of spurious estimates of the *a priori* mean.

The proposed estimator recovers sinusoidal components at lower local SNRs, and is able to respond to low SNR sinusoidal components in a shorter time than the MMSE STSA estimator. Objective testing performed with the PESQ measure indicated that the proposed estimator is capable of outperforming some major and fundamental Gaussian STFT-based speech enhancement algorithms under a variety of noise conditions.

The spectral mean estimation described in Section IV-A exploits the predictability of periodic components in speech with the use of the spectral estimation techniques described in [13]. Given other more general forms of *a priori* information are able to increase the accuracy of this spectral mean estimation (for example, with the use of alternate spectral estimation techniques, a more sophisticated speech model, or the additional channels in

a multichannel signal), then the proposed algorithm may be further improved. Furthermore, as the SD MMSE STFT estimator considers a fundamental alteration to the statistical assumptions made for the MMSE STSA estimator in [4], many of the improvements that have been recently made to the MMSE STSA estimator may yet be considered with respect to the SD statistical model. In particular, speech enhancement under the SD speech model in (7) may be considered under alternative cost functions [22], [23], with super-Gaussian distributions [11], and with correlated spectral components [12].

Appendix A
Phase Normalized Signal Observations

Consider a digital signal consisting of a complex exponential with varying frequency, $f[n]$ (in cycles per sample, where $0 \leq f[n] < 1$) and amplitude $z[n]$,

$$s[n] = z[n] \exp\left\{2\pi i \left(\sum_{p=0}^{n} f[p]\right) + i\vartheta\right\}. \tag{28}$$

where ϑ is the phase of $s[n]$ at $n = 0$ in radians. If we wish to analyze this signal via the STFT, we consider the signal over a window $mM \leq n < mM + N - 1$. Separating the summation into a set of terms within the window to be analyzed, and a set of terms prior to the window via the substitution $p = l + mM$,

$$s[n] = z[n] \exp\left\{2\pi i \left(\sum_{l=0}^{N-1} f[l+mM] + \sum_{l=-mM}^{-1} f[l+mM]\right) + i\vartheta\right\}. \tag{29}$$

If $f[n]$ is slowly varying so that it may be considered constant within the window length N, the expression may be simplified,

$$s[n] \approx z[n] \exp\{2\pi i f_m n + i\vartheta\} \exp\left\{2\pi i M \sum_{q=-m}^{-1} f_q\right\}, \tag{30}$$

where f_m is the approximate frequency for $mM \leq n < mM + N - 1$. If analyzed by the STFT (2) at frequency f_m the equation becomes,

$$S[f_m, m] \approx \widetilde{S}[f_m, m] \exp\left\{2\pi i M \sum_{q=-m}^{-1} f_q\right\}. \tag{31}$$

Here the signal is represented by a term consisting of samples at the frequency within the window, $\widetilde{S}[f_m, m]$, and a history of frequencies, $\exp\left\{2\pi i M \sum_{q=-m}^{-1} f_q\right\}$, that adjust the signal's phase for that window. $\widetilde{S}[f_m, m]$ is referred to as the phase-normalized STFT observation of the deterministic component at frequency f_m.

Appendix B
Derivation of the SD MMSE STFT Amplitude Estimator

In solving (9) it is informative to consider the form of the *a posteriori* pdf, $p_{post} = p(A_k, \alpha_k \mid B_k, \beta_k)$,

$$p_{post} = \frac{p(B_k, \beta_k \mid A_k, \alpha_k) p(A_k, \alpha_k)}{\int_0^\infty \int_0^{2\pi} A_k p(B_k, \beta_k \mid A_k, \alpha_k) p(A_k, \alpha_k) d\alpha_k dA_k}. \tag{32}$$

The joint distribution $p(B_k, \beta_k \mid A_k, \alpha_k) p(A_k, \alpha_k)$, denoted as p_{joint}, is first algebraically manipulated. By making the substitution $Q_k = \left(A_k e^{-i\alpha_k} - \mu_k e^{-i\theta_k}\right) e^{i\angle(\Delta_k)}$, where $\Delta_k \triangleq B_k e^{-i\beta_k} - \mu_k e^{-i\theta_k}$, this distribution may be simplified. Starting with the equations defined in (7) and (10),

$$\begin{aligned} p_{joint} &= \Upsilon_k \exp\left\{\frac{-\left|B_k e^{-i\beta_k} - A_k e^{-i\alpha_k}\right|^2}{\lambda_{d,k}} - \frac{\left|A_k e^{-i\alpha_k} - \mu_k e^{-i\theta_k}\right|^2}{\lambda_{x,k}}\right\} \\ &= \Upsilon_k \exp\left\{-\frac{|Q_k - |\Delta_k||^2}{\lambda_{d,k}} - \frac{|Q_k|^2}{\lambda_{x,k}}\right\} \\ &= \Upsilon_k \exp\left\{-\frac{|Q_k - \Omega_k |\Delta_k||^2}{\lambda_k} - \frac{|\Delta_k|^2}{\lambda_{x,k} + \lambda_{d,k}}\right\}, \end{aligned} \tag{33}$$

where Ω_k and λ_k are given in (16) and (15), respectively, and $\Upsilon_k = 1/(\pi^2 \lambda_{d,k} \lambda_{x,k})$. By substituting values for Q_k and Δ_k it can be found that as a function of $A_k e^{-i\alpha_k}$, (33) takes the form of a complex Gaussian function with mean ζ_k given in (14) and scale λ_k. Substituting (33) back into (32), the expression concerned here may be simplified to,

$$p_{post} = \frac{\exp\left\{\frac{-\left|A_k e^{-i\alpha_k} - \zeta_k\right|^2}{\lambda_k}\right\}}{\int_0^\infty A_k \int_0^{2\pi} \exp\left\{\frac{-\left|A_k e^{-i\alpha_k} - \zeta_k\right|^2}{\lambda_k}\right\} d\alpha_k dA_k}. \tag{34}$$

The complex exponentials $B_k e^{-i\beta}$ and $\mu_k e^{-i\theta}$ that feature in ζ_k may then be combined into a single magnitude term, C_k, and phase term, Θ_k,

$$p_{post} = \frac{\exp\left\{\frac{-\left|A_k e^{-i\alpha_k} - \zeta_k\right|^2}{\lambda_k}\right\}}{\int_0^\infty A_k \int_0^{2\pi} \exp\left\{\frac{-\left|A_k e^{-i\alpha_k} - C_k e^{-i(\beta_k + \Theta_k)}\right|^2}{\lambda_k}\right\} d\alpha_k dA_k}, \tag{35}$$

where,

$$\Theta_k = \arctan\left\{\frac{(1-\Omega_k)\mu_k \sin\theta_k + \Omega_k B_k \sin\beta_k}{(1-\Omega_k)\mu_k \cos\theta_k + \Omega_k B_k \cos\beta_k}\right\}, \tag{36}$$

and,

$$C_k = \Big((1-\Omega_k)^2 \mu_k^2 + \Omega_k^2 B_k^2 + 2\Omega_k(1-\Omega_k)\mu_k B_k \cos(\theta_k - \beta_k)\Big)^{1/2} \tag{37}$$

Furthermore, by evaluating the magnitude operator in the exponent in the denominator of (35) in terms of real and imaginary components, it may be seen that (see equation (38) at the top of

$$p_{post} = \frac{\exp\left\{\frac{-\left|A_k e^{-i\alpha_k} - \zeta_k\right|^2}{\lambda_k}\right\}}{\int_0^\infty A_k \exp\left\{-\frac{A_k^2 + C_k^2}{\lambda_k}\right\} \int_0^{2\pi} \exp\left\{\frac{2A_k C_k \cos(\alpha_k - \beta_k - \Theta_k)}{\lambda_k}\right\} d\alpha_k dA_k} \tag{38}$$

the page). Noting that the term within the inner integral of the denominator here is periodic in α_k with period 2π, and using the integral form of the modified Bessel function for order n,

$$I_n(z) = \frac{1}{2\pi} \int_0^{2\pi} \cos \alpha n \ \exp\left(z \cos \alpha\right) d\alpha, \tag{39}$$

we may simplify (38) to,

$$p_{post} = \frac{\frac{1}{\pi\lambda_k} \exp\left\{\frac{-\left|A_k e^{-i\alpha_k} - \zeta_k\right|^2}{\lambda_k}\right\}}{\int_0^\infty \frac{2A_k}{\lambda_k} \exp\left\{-\frac{A_k^2 + C_k^2}{\lambda_k}\right\} I_0\left(\frac{2A_k C_k}{\lambda_k}\right) dA_k}. \tag{40}$$

Here it can now be seen that the term in the denominator is the integral of a Rice distribution across its entire domain, therefore it is equal to 1. This concludes the proof that the *a posteriori* distribution in (32) is in fact a complex Gaussian distribution with mean ζ_k and scale parameter λ_k. The expected value of the magnitude of such a distribution (i.e., the expression in (9)) is known to be the expected value of a Rice distribution with centrality parameter C_k and scale parameter $\lambda_k/2$ [39]. This expected value has the closed form expression described in (11).

ACKNOWLEDGMENT

The authors would like to thank their colleague Dr M. Andrews for several useful discussions on the mathematical notation used. They would also like to thank the reviewers of the paper for their conscientious endeavour resulting in many very helpful comments and suggestions.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC, 2007.
[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[3] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.
[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[5] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7, pp. 588–601, 2007.
[6] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Amer.*, vol. 122, no. 3, pp. 1777–1786, 2007.
[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[8] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
[9] C. Breithaupt and R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.
[10] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.
[11] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.
[12] E. Plourde and B. Champagne, "Multidimensional STSA estimators for speech enhancement with correlated spectral components," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3013–3024, Jul. 2011.
[13] D. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE*, vol. 70, no. 9, pp. 1055–1096, Sep. 1982.
[14] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
[15] J. Hardwick, C. Yoo, and J. Lim, "Speech enhancement using the dual excitation speech model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1993, vol. 2, pp. 367–370.
[16] R. Hendriks, R. Heusdens, and J. Jensen, "An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 406–415, Feb. 2007.
[17] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1993, vol. 2, pp. 550–553.
[18] Y. Stylianou, "On the harmonic analysis of speech," in *Proc. Circuits Syst.. Int. Symp.*, 1998, vol. 5, pp. 5–8.
[19] M. C. McCallum and B. J. Guillemin, "Accounting for deterministic noise components in a MMSE STSA speech enhancement framework," in *Proc. 12th Int. Symp. Commun. Inf. Technol.*, 2012, pp. 174–179.
[20] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
[21] J. Jensen and J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 731–740, Oct. 2001.
[22] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.
[23] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.
[24] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993, vol. 1.
[25] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.
[26] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. Amsterdam, The Netherlands: Elsevier, 1995, ch. 14, pp. 495–518.
[27] M. Spiegel, *Advanced Mathematics for Engineers and Scientists*. New York, NY, USA: McGraw-Hill, 1980.
[28] K. K. Paliwal, K. Wojcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.
[29] Z. Jin and D. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091–1102, Jul. 2011.
[30] S. M. Kay, *Modern Spectral Estimation: Theory & Application*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[31] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 59–67, Jan. 2004.
[32] W. Charoenruengkit and N. Erdol, "The effect of spectral estimation on speech enhancement performance," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1170–1179, Jul. 2011.
[33] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.
[34] S. Yoo, J. Boston, J. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C. Li, "Relative energy and intelligibility of transient speech information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 1, pp. 69–72.
[35] "Subjective performance assessment of telephone-band and wideband digital codecs," ITU-T, 1996, Rec. P.830.
[36] "Objective measurement of active speech level," ITU-T, 2011, Rec. P.56.
[37] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T, 2001, Rec. P.862.
[38] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[39] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell Syst. Tech. J.*, vol. 27, pp. 109–157, 1948.

Matthew McCallum (S'07) received the B.E. degree in 2010 from The University of Auckland. He is currently pursuing the Ph.D. degree in the Department of Electrical and Electronic Engineering, The University of Auckland, New Zealand. His research interests include statistical signal processing, audio analysis and processing, active noise cancellation and Bayesian approaches to speech enhancement.

Bernard Guillemin (M'02) received the PhD from the University of Auckland, New Zealand, in 1986. His thesis investigated vocal tract shape determination from a linear prediction analysis of the speech signal. He is currently a Senior Lecturer in the Department of Electrical & Computer Engineering, University of Auckland. His research interests include speech analysis, synthesis and recognition, speech enhancement, active noise cancellation, as well as forensic voice comparison. In respect to the latter, he is regularly engaged as an expert witness in forensic speech science in the courts in New Zealand, and is a member of the Forensic Speech Science Committee within the Australasian Speech Science Technology Association (ASSTA).