Joint Stochastic-Deterministic Wiener Filtering with Recursive Bayesian Estimation of Deterministic Speech

Matthew McCallum and Bernard Guillemin

Department of Electrical and Computer Engineering
The University of Auckland, Auckland, New Zealand

Abstract

Stochastic-deterministic (SD) speech modelling exploits the predictability of speech components that may be regarded deterministic. This has recently been employed in speech enhancement resulting in an improved recovery of deterministic speech components, although the improvement achieved is largely dependant on how these components are estimated. In this paper we propose a joint SD Wiener filtering scheme that exploits the predictability of sinusoidal components in speech. Estimation of sinusoidal speech components is approached in the recursive Bayesian context, where the linearity of the joint SD Wiener filter and Gaussian assumptions suggest a Kalman filtering scheme for the estimation of sinusoidal components. A further refinement also imposes a restriction of a smooth spectral envelope on sinusoidal magnitude estimates. The resulting joint SD Wiener filtering scheme improves speech quality in terms of the perceptual evaluation of speech quality (PESQ) metric when compared to both the traditional Wiener filter and the proposed Wiener filter based on alternative estimates of deterministic speech components.

Index Terms: Speech enhancement, stochastic deterministic model.

1. Introduction

Background acoustic noise is a commonly recurring problem in mobile voice communication, where speech quality and intelligibility may be significantly compromised in the presence of low signal to noise ratios (SNRs). Accordingly, attempts to mitigate the negative effects of background acoustic noise (commonly referred to as speech enhancement) have been researched for several decades now. Speech enhancement for mobile voice communication is particularly difficult due to the necessity for real-time processing of signals, limited computational resources and often the constraint of a single microphone. This paper considers speech enhancement within this context.

Many single-channel short-time Fourier transform (STFT) speech enhancement methods are applicable for such purposes, including spectral subtraction [1], Wiener filtering [2] and short-time spectral amplitude estimators (STSA) [3]. Much of the focus in improving these algorithms has made use of entirely stochastic models for noise and speech. For example, investigations have been made into alternate stochastic characterisations of STFT coefficients, considering a range of purely stochastic (zero-mean) *a priori* speech distribution shapes [4,5], STFT coefficient correlations [6] and cost functions [7]. Alternative approaches to speech enhancement have focussed instead on deterministic models for speech. McAulay and Malpass [8] considered speech signal STFT coefficients as deterministic quantities corrupted by complex Gaussian noise, and Jensen and Hansen [9] considered speech enhancement where speech is modelled by a set of deterministic sinusoids.

Despite the success of attempts to characterise speech as either stochastic or deterministic, it is known that speech is more accurately described to contain both stochastic and deterministic components simultaneously. Explicitly considering speech to have both stochastic and deterministic components generally leads to an improved estimation and hence enhancement of each component, as observed in [10, 11]. This earlier work considered estimates of stochastic and deterministic components independently which are then combined based on additive or soft-decision procedures. However, only recently has a joint stochastic-deterministic (SD) model been considered [12], leading to further improvements in the enhancement of deterministic speech components. Whilst [12] works in the minimum mean-square error (MMSE) STSA context, this paper endeavours to apply such a joint SD approach to Wiener filtering, with careful attention paid to the estimation of deterministic components.

As in [12], here the harmonic plus noise or SD model commonly employed in the areas of speech coding and speech synthesis [13, 14] is considered. That is, the signal under consideration is assumed to be of the form,[1]

$$\mathbf{x}[n] = v[n] + \mathbf{u}[n], \tag{1}$$

where the deterministic component $v[n]$ is of the form,

$$v[n] = \sum_{l=0}^{L-1} r_l[n] cos(\phi_l[n]), . \tag{2}$$

This component contains L harmonic sinusoids indexed by l, where $\phi_l[n]$ represents a discrete instantaneous phase function, $\phi_l[n] = \sum_{c=1}^{n} 2\pi l f_0[c] + \phi_l[0]$ with $f_0[c]$ the instantaneous fundamental frequency and $\phi_l[0]$ the phase offset at $n = 0$. The stochastic component, $\mathbf{u}[n]$, may be considered the signal residual, i.e., all signal components that can not be represented by $v[n]$, although more specific stochastic-deterministic models restrict $\mathbf{u}[n]$ to a quasi-stationary, zero-mean stochastic process [13,15]. In the context of speech enhancement, the stochastic signal component is necessary to allow for the possibility of aperiodic sounds and unpredictable signal perturbations in speech. The obvious advantage of considering the deterministic component explicitly (i.e., sinusoids in the model of (1)), is that its properties (e.g. the periodicity of sinusoids) may be exploited to allow improved estimation of these components. However, the improvement observed is largely dependent on the accuracy of the estimation of deterministic components. A range of techniques for the estimation of sinusoidal amplitude and phase exist both in a general sense [16, 17] and specific to speech signals [18, 19]. The best method for use in the case of SD speech enhancement is still unclear and largely uninvestigated.

In this paper we derive the joint SD STFT Wiener filter with recursive Bayesian estimation of deterministic speech components. The highly non-stationary nature of speech signals suggests a recursive Bayesian approach is appropriate. Due to the

This research is supported by Capability funding from the MBIE Science and Innovation Group.

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

Gaussian assumptions and linear processing in the Wiener filtering framework, the estimation of deterministic speech components reduces to the design of a complex Kalman filter. This is further combined with exploitation of the smooth spectral contour characteristic of speech [20]. Unlike many other approaches to the estimation of sinusoidal parameters in the STFT framework, this approach exploits correlation in both time and frequency of frequency modulated sinusoidal components.

2. The Joint SD Wiener filter

The common objective of STFT speech enhancement is to estimate a clean speech signal $\mathbf{x}[n]$ from a signal mixture, $\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{d}[n]$. This estimate is obtained via the STFT,

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

for $0 \leq k < K$, where k corresponds to the STFT frequency bin number. The parameters M, K and $w[n]$ refer to the shift in samples between successive window frames, the DFT length and the windowing function, respectively. Due to the linearity of the STFT it is equivalent to estimate $\mathbf{X}[k,m]$ from the mixture $\mathbf{Y}[k,m] = \mathbf{X}[k,m] + \mathbf{D}[k,m]$. Perhaps the most effective STFT speech enhancement methods have been developed in the Bayesian MMSE context [3, 5, 6]. The derivation of such estimators relies on the definition of the statistical properties of both noise and speech. Here we focus on the statistical characterisation of speech and so the noise process, $\mathbf{d}[n]$, is generically assumed to be white Gaussian. Under this assumption it is adequate for the purposes of speech enhancement to assume that $\mathbf{D}[k,m]$ is independently zero-mean complex Gaussian distributed for each k, with scale parameter $\lambda_{d,k} = E\left\{|\mathbf{D}[k,m]|^2\right\}$, where $E\{\cdot\}$ denotes the expectation operator [3]. Applying similar assumptions to $\mathbf{u}[n]$ of (1), $\mathbf{X}[k,m]$ may be assumed to be independently complex Gaussian distributed for each k with complex mean [12],[2]

$$V[k,m] = \sum_{l=0}^{L-1} \frac{r_l[mM]}{2}\left(e^{i\phi_l[mM]}W_{f_l}[k] + e^{-i\phi_l[mM]}W_{1-f_l}[k]\right), \tag{3}$$

and scale parameter $\lambda_{d,k} = E\left\{|\mathbf{X}[k,m] - V[k,m]|^2\right\}$. Here $W_{f_l}[k] = \sum_{n=0}^{N-1} w[n]\exp\{2\pi in\,(f_l - k/K)\}$ is the DFT of the windowing function $w[n]$, modulated by a complex exponential. f_l may be considered the average of $\omega_l[n]/2\pi$ for $mM \leq n < N$.[3] With these statistical preliminaries, the clean speech is estimated within the Bayesian MMSE context as,

$$\hat{X}[k,m] = E\left\{\mathbf{X}[k,m]|\mathbf{Y}[k,m] = Y[k,m]\right\}. \tag{4}$$

This problem is simplified by considering the posterior pdf, $p_{\mathbf{X}|\mathbf{Y}}\left(X[k,m]|Y[k,m]\right)$, which is known to be [12],

$$p_{\mathbf{X}|\mathbf{Y}}\left(X[k,m]|Y[k,m]\right) = \frac{1}{\pi\lambda_k}\exp\left\{\frac{-\left|X[k,m] - \zeta_k\right|^2}{\lambda_k}\right\},$$

where,

$$\zeta_k = Y[k,m]\frac{\lambda_{x,k}}{\lambda_{x,k}+\lambda_{d,k}} + V[k,m]\left(1 - \frac{\lambda_{x,k}}{\lambda_{x,k}+\lambda_{d,k}}\right),$$

and $\lambda_k = \lambda_{x,k}\lambda_{d,k}/\left(\lambda_{x,k}+\lambda_{d,k}\right)$. The resulting expectation of (4) is simply the mean of this posterior distribution, ζ_k. Hence the SD STFT Wiener estimate, $\hat{X}[k,m] = \zeta_k$, is seen to be a weighted combination of the observed speech signal $Y[k,m]$ and the *a priori* information contained within the speech mean $V[k,m]$, dependent on the distribution scales $\lambda_{x,k}$ and $\lambda_{d,k}$. The potential improvement in performance of the SD STFT Wiener filter over purely stochastic Wiener filters is heavily dependent on the estimation of the speech mean $V[k,m]$. As will be shown in Section 4, given a very accurate estimate is possible, the SD STFT Wiener filter offers a significant improvement in performance.

[2]This model relies upon the assumption that both $r_l[n]$ and $\omega_l[n]$ are changing slowly enough to be considered constant for N samples.

[3]Whilst f_l is dependent on window index m, for notational simplicity, this will not be stated explicitly (e.g. $f_l[m]$) unless necessary.

3. Estimation of deterministic speech parameters

In this paper $V[k,m]$ is entirely specified by the magnitude and phase of each sinusoid in $v[n]$ (or equivalently their complex amplitude for frame m, $\boldsymbol{\nu}_l[m]$), and the fundamental frequency f_0. The value, f_0, may be accurately estimated using a range of techniques [21,22]. Here the estimation of $\boldsymbol{\nu}_l[m]$ is focussed on, assuming prior knowledge of the fundamental frequency.

The values $\boldsymbol{\nu}_l[m]$ may be considered to be correlated in time and frequency, rendering them predictable. First, in a STFT analysis/synthesis system, the slowly varying nature of the frequency and amplitude of voiced speech (relative to M) mean that sinusoid amplitude and phase in the current frame is predictable from the previous frame. Secondly, it is well known that the spectral envelope of speech is smooth, hence the amplitude of a sinusoid may be predicted from other sinusoids in the same frame. Here a recursive Bayesian estimation approach is used to exploit correlation in time, and a parametric autoregressive (AR) model is used to exploit correlation in frequency.

3.1. Recursive Bayesian estimation of amplitude and phase

Given the STFT observation of the clean signal $[X[m]]_{k=0}^{K-1} = [X[0,m], X[1,m], \ldots, X[K-1,m]]^T$,[4] a measurement of the complex amplitude of a sinusoid with known frequency may be obtained via the maximum likelihood method [23],

$$\hat{\nu}_l[m] = s[f_l]^T F^H \left[X[m]\right]_{k=0}^{K-1}, \tag{5}$$

where $s[f_l] = \left[e^0, e^{-2\pi i l f_0}, e^{-2\pi i 2 l f_0}, \ldots, e^{2\pi i(N-1)lf_0}\right]^T$ is a vector of length N containing complex exponential samples at known frequency $f_l[m]$ and $F \in \mathbb{C}^{K\times K}$ is the DFT matrix. A series of measurements, $\tilde{\nu}_l[m] = \exp\left\{-2\pi i M\sum_{\tau=m}^{q} f_l[\tau]\right\}\hat{\nu}_l[m]$,[5] for a sample of voiced speech are plotted in Figure 1. It may be noted here that these measurements are highly correlated. That is, given p measurements of $\tilde{\nu}_l[m]$ for $q-p \leq m < q$, it is clear that the measurement $\tilde{\nu}_l[q]$ may be estimated with some accuracy based on this history of data and no information from the current frame. Specifically, if this *a priori* estimate, $\tilde{\nu}_{l,prior}[q]$, is assumed to be linear, then it may be obtained from a history of *a posteriori* estimates $[\tilde{\nu}_{l,post}]_{m=q-p}^{q-1}$ via linear prediction with an appropriate modification allowing for the shift in phase between successive STFT windows,

$$[\tilde{\nu}_{l,prior}]_{m=q-p+1}^{q} = e^{-2\pi i f[q]M} A\,[\tilde{\nu}_{l,post}]_{m=q-p}^{q-1}, \tag{6}$$

where,

$$A = \begin{bmatrix} \alpha_{p-1} & a_p \\ I_{p-1} & \Theta_{p-1} \end{bmatrix}.$$

[4]In this paper $(\cdot)^T$ represents the vector transpose operation, and $(\cdot)^H$ the Hermitian transpose. Throughout the paper vectors are constructed in the way shown here, i.e., a column vector of a signal $z[n]$ containing elements $0 \leq n \leq N-1$ will be denoted $[z]_{n=0}^{N-1}$.

[5]The exponential term is included here to account for the expected phase shift between samples of $\hat{\nu}_l[m]$ in successive frames [12].

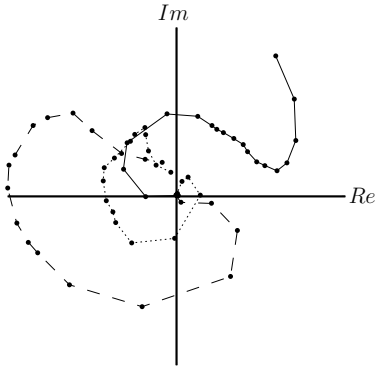

Figure 1: *Observations of $\boldsymbol{\nu}_l[m]$ plotted on the complex plane for an utterance of the vowel /æ/. Observations are indicated with dots, and lines are drawn between consecutive observations. The dotted line, dashed line and solid line, correspond to the second, fourth and fifth harmonics respectively. A sampling rate of 8kHz was used and STFT parameters $N = 240$, $M = 120$. $w[n]$ was a Hamming window.*

Here $I_p \in \mathbb{R}^{p\times p}$ denotes the identity matrix and $\Theta_p \in \mathbb{R}^{p\times 1}$ is the zero vector. The linear prediction coefficients $\alpha_p = [a_1, a_2, \ldots, a_p]$ may be estimated from a history of previous estimates (e.g. from $[\tilde{\nu}_{l,post}]_{m=q-p}^{q-1}$) via the Yule-Walker equations [20]. The linear restriction here simplifies the Bayesian estimation to be applied in this section by preserving the Gaussian assumptions of $\mathbf{D}[k,m]$ and $\mathbf{X}[k,m]$.

To demonstrate the preservation of Gaussian assumptions, the Gaussian assumption of the variables $\mathbf{X}[k,m]$ combined with the linear processing of (5), mean that the measurements $\tilde{\nu}_l[m]$ are also Gaussian distributed with mean $\mu_{\boldsymbol{\nu}}[m] = s[f]^T F^H \left[V[m]\right]_{k=0}^{K-1}$, and covariance,

$$\lambda_{\tilde{\boldsymbol{\nu}},l}[m] = s[f]^T F^H C_{\mathbf{XX}}[m] F \left(s[f]^T\right)^H, \qquad (7)$$

where $C_{\mathbf{XX}}[m] \in \mathbb{C}^{K\times K}$ is the covariance matrix characterising $\left[\mathbf{X}[m]\right]_{k=0}^{K-1}$. A similar argument applies to both observations made from the noisy STFT $\mathbf{Y}[k,m]$, and the prediction of (6) which is a recursive linear combination of Gaussian distributed variables. Hence applying recursive Bayesian estimation here reduces to the design of a Kalman filter.

3.2. Kalman filter prediction

For each m and l, a prediction of a complex sinusoid amplitude may be obtained via (6). The covariance matrix of this *a priori* estimate is given by,

$$P_{prior}[q] = AP_{post}[q-1]A^H + \Psi, \qquad (8)$$

where $\Psi = \text{diag}\left\{\lambda_{\tilde{\boldsymbol{\nu}},l}[q], 0, \ldots, 0\right\}$, and $\lambda_{\tilde{\boldsymbol{\nu}},l}[q]$ is the covariance of the complex amplitude of sinusoidal component l in frame q. $P_{post}[q-1]$ is the covariance of the *a posteriori* estimate $[\tilde{\nu}_{l,post}]_{m=q-p}^{q-1}$ at frame $q-1$, as will be described in Section 3.3. Under the Gaussian and independence assumptions of $\mathbf{X}[k,m]$ for all k, the covariance matrix $C_{\mathbf{XX}}[m]$ is diagonal and is often estimated in STFT speech enhancement via the decision-directed method [3] or in the case of SD STFT speech enhancement, via the modified decision-directed method of [12].

3.3. Kalman filter correction

Observations, $\tilde{\gamma}_l[m]$, of $\boldsymbol{\nu}_l[m]$ may be made from the noisy STFT signal via,

$$\tilde{\gamma}_l[m] = s[f_l]^T F^H \left[Y[m]\right]_{k=0}^{K-1}. \qquad (9)$$

Under the independence and Gaussian assumptions of $\mathbf{Y}[k,m]$ made in Section 2, observations $\tilde{\gamma}_l[m]$ have covariance,

$$\lambda_{\boldsymbol{\gamma},l}[m] = s[f]^T F^H C_{\mathbf{YY}}[m] F \left(s[f]^T\right)^H,$$

where $C_{\mathbf{YY}}[m] = C_{\mathbf{DD}}[m] + C_{\mathbf{XX}}[m]$. For a variety of noise types in speech enhancement $C_{\mathbf{DD}}[m]$ is also often assumed diagonal and may be estimated with a range of techniques, such as that of [24]. Considering the prediction of (6) and the observation made via (9), an *a posteriori* estimate may be obtained,

$$[\tilde{\nu}_{l,post}]_{m=p-q+1}^{q} = K[q]\tilde{\gamma}_l[q] + (I_p - K[q])\,[\tilde{\nu}_{l,prior}]_{m=q-p+1}^{q}. \qquad (10)$$

The Kalman gain $K[q] = \text{diag}\left\{k[q], 0, \ldots, 0\right\}$ is specified by,

$$k[q] = \frac{\lambda_{\tilde{\boldsymbol{\nu}},l}[q]}{\lambda_{\tilde{\boldsymbol{\nu}},l}[q] + \lambda_{\tilde{\gamma},l}[q]}.$$

Finally the *a posteriori* estimate covariance is calculated as,

$$P_{post}[q] = (I - K_m)P_{prior}[q]. \qquad (11)$$

Equations (6), (8), (10) and (11) define the Kalman filter prediction, *a priori* covariance, correction and *a posteriori* covariance, respectively, for the estimation of deterministic components of the type (2) in noisy speech. Unlike previous applications of Kalman filtering in speech enhancement [25, 26], the approach here focuses on the estimation of deterministic components (i.e., sinusoids) that exist across several STFT frames. Furthermore, it incorporates explicit consideration of frequency modulation and phase shift between successive frames.

3.4. Calculation of deterministic spectra

The Kalman filtering scheme described in Sections 3.2 and 3.3 yields an estimate of the complex amplitude $\boldsymbol{\nu}_l[m]$ of sinusoidal component l, independently for $0 < l < L$. These estimates, combined with knowledge of f_0, completely describe the mean spectrum of (3). That is, $r_l[mM] = |\tilde{\nu}_{l,post}[m]|$, $\phi_l[mM] = \angle(\tilde{\nu}_{l,post}[m])$ and $f_l = lf_0$. Hence, $V[k,m]$ may be synthesised according to (3).

Whilst the estimates $\tilde{\nu}_{l,post}[m]$ exploit the correlation in time of deterministic speech components well, because they are estimated independently, the correlation in frequency, i.e., between $\tilde{\nu}_{l,post}[m]$ for various l, remains unexploited. AR models are widely employed in speech signal processing and enhancement to exploit the smooth spectral envelope seen in speech [20]. For the deterministic speech component estimates here, an estimate of AR model parameters, b_c for $0 < c \leq \rho$, may be obtained from the inverse DFT of $V[k,m]$ via the Yule-Walker equations. The AR spectrum,

$$B[f,m] = \frac{\sigma_y}{1 + \sum_{c=1}^{\rho} b_c e^{-2\pi ifc}}$$

where $\sigma_y^2 = 1/K \sum_{k=0}^{K-1} \left|V[k,m]\left(1 + \sum_{c=1}^{\rho} b_c e^{-2\pi ick/K}\right)\right|^2$, may then be sampled at $f = f_l$, to obtain a refined estimate of the magnitudes, $|\tilde{\nu}_{l,post}[m]|$, which is smoothly evolving in frequency. In the experiments in this paper this refinement was found to further improve both the accuracy of deterministic speech component estimation, and the PESQ measure at the output of the proposed SD Wiener filter.

4. Experimental Evaluation

In informal experimental observations over a range of male and female speakers from the TIMIT database, it was commonly observed that the deterministic speech component estimation method of Section 3 was notably more accurate than the methods used recently in SD-based speech enhancement [11, 12]. A typical comparison of the magnitude of deterministic speech

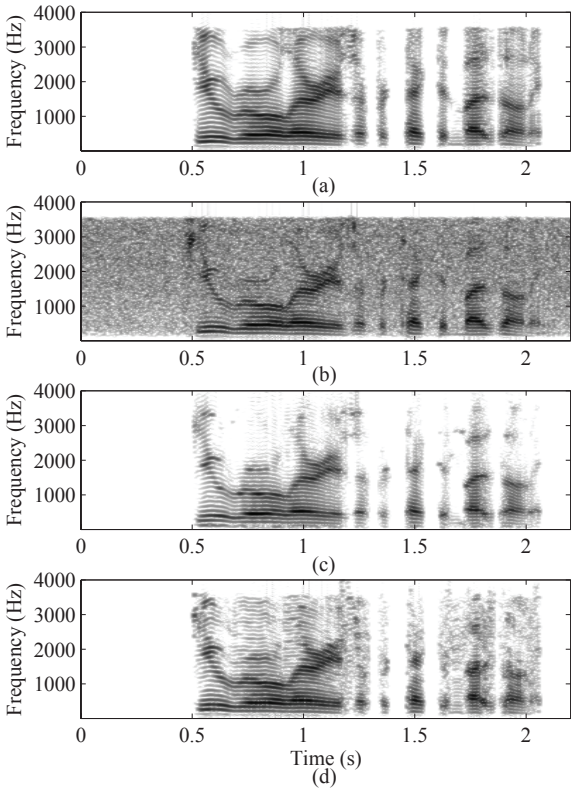

Figure 2: *A log-magnitude spectrogram demonstration of the recovery of deterministic speech components using the utterance "Few rural areas are protected by zoning". (a) The deterministic speech component obtained from the clean speech signal (i.e., via (5)). (b) The speech signal corrupted by WGN at an SNR of 5dB. (c) An estimate of the deterministic speech component using the proposed method of Section 3. (d) An estimate of the deterministic speech component via the method of [12]. The STFT parameters* $N = 240$*,* $M = 60$ *and* $K = 480$ *were used.* $w[n]$ *was a Hamming window.*

component estimation is shown in Figure 2. Here it can be seen that the proposed estimation method in Figure 2(c) more accurately recovers the true formant structure seen in Figure 2(b), when compared to the reference method of Figure 2(d). It can be seen that the latter method has a tendency to overestimate the magnitude of deterministic speech components (for example, in the region of 0.6s to 1s). It was noticed in these regions of notable magnitude overestimation, that the output contains significant unnatural and harsh sounding artefacts.

The evaluation of the proposed joint SD Wiener filter was conducted in a range of configurations to demonstrate the relationship between performance and the method of deterministic speech component estimation used. The evaluation considers the proposed method of Section 3 (named KAL), the window averaging method of [11] (named AV), the recursive averaging method of [12] (named REC), and the maximum likelihood method [23] taken from the current frame of clean speech data (named CLEAN). Also included for reference are the results from the unmodified Wiener filter [2] (named WIEN), and the noisy speech signal itself (named NOISY). All configurations were tested over 30 "phonetically diverse" utterances from 10 speakers of the TIMIT database, equal parts male and female. Each utterance was combined with white noise at SNRs from -5dB (typically unintelligible) to 15dB (typically entirely intelligible) according to the methods of [27]. All resulting input data was sampled at 8kHz and filtered according to [28] to better simulate mobile speech communication. For all algorithm configurations, $N = 240$, $K = 480$, $M = 60$ and the modified decision-directed method of [12] was used to estimate $\lambda_{x,k}$ with a smoothing parameter of 0.98. For the proposed algorithm $p = 3$ and $\rho = 10$. Noise estimates (i.e., estimates of $\lambda_{d,k}$) were obtained from a 0.48s noise-only segment at the start of each utterance (removed prior to objective evaluation). Estimates of f_0 were obtained via the maximum likelihood method [16] from the noise corrupted data.

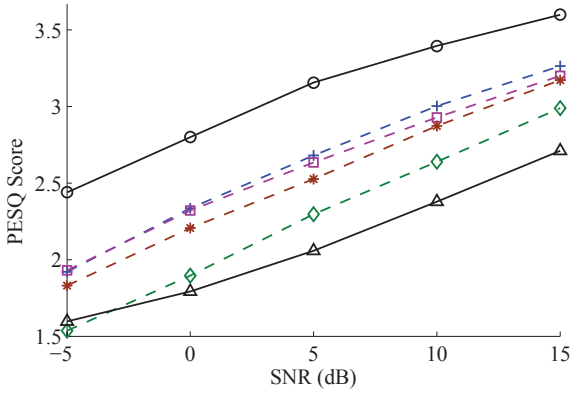

Figure 3: *PESQ measures obtained with the proposed and reference algorithms. From top to bottom, (◯) CLEAN, (+) KAL, (□) REC, (∗) WIEN, (◇) AV, (△) NOISY.*

The results of the aforementioned evaluation are shown in Figure 3 as measured via the PESQ metric [29] averaged over all utterances. The PESQ metric is a common tool for the evaluation of speech enhancement algorithms [5, 7, 11, 12, 25, 30] and is in the range of 1-4.5, where a higher score indicates an increase in perceptual speech quality. When compared to unmodified Wiener filtering, it can be seen that the proposed KAL algorithm increases the average PESQ score by approximately 0.1-0.16, where the best performance is seen at an SNR of 5dB (an increase of 0.16 in PESQ score). On individual utterances increases of up to 0.3 were observed. The results for the AV and the CLEAN algorithms demonstrate how heavily the output of the joint SD Wiener algorithm depends on the method of deterministic component estimation. The excellent results for the CLEAN algorithm are promising in that they indicate even more significant improvements may be possible given further research into deterministic component estimation is successful.

The proposed KAL algorithm also outperforms the REC algorithm with regards to the PESQ measure, particularly at higher SNRs. It is interesting to note that whilst this improvement is minimal at lower SNRs, the most audible differences between the output of KAL and REC were observed under these conditions. That is, the REC algorithm consistently had quite audible artefacts (mentioned earlier in this section) during voiced speech segments, these being more prominent at low SNRs. In the case of the KAL algorithm these artefacts were non-existent.

5. Conclusion

In this paper the joint SD Wiener filter was introduced with specific attention to the estimation of deterministic speech components. A recursive Bayesian approach was used for the estimation of these components which reduced to the design of a Kalman filter. In addition, it was found advantageous to limit the spectral envelope of deterministic components to an AR spectrum, imposing smoothness of the envelope. The resulting algorithm was found to better estimate deterministic speech components than other recent methods employed, resulting in an improved PESQ score, and a mitigation of unnatural sounding artefacts at the algorithm output.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[2] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 845–856, 2005.

[5] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1741–1752, 2007.

[6] E. Plourde and B. Champagne, "Multidimensional STSA estimators for speech enhancement with correlated spectral components," *Signal Processing, IEEE Transactions on*, vol. 59, no. 7, pp. 3013–3024, 2011.

[7] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 475–486, 2005.

[8] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 2, pp. 137–145, 1980.

[9] J. Jensen and J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 731–740, 2001.

[10] J. Hardwick, C. Yoo, and J. Lim, "Speech enhancement using the dual excitation speech model," in *Acoustics, Speech, and Signal Processing, International Conference on*, vol. 2, 1993, pp. 367–370.

[11] R. Hendriks, R. Heusdens, and J. Jensen, "An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 406–415, 2007.

[12] M. McCallum and B. Guillemin, "Stochastic-deterministic MMSE STFT speech enhancement with general a priori information," *Acoustics, Speech and Language Processing. IEEE Transactions on*, vol. 21, no. 7, pp. 1445–1457, 2013.

[13] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *Acoustics, Speech, and Signal Processing, International Conference on*, vol. 2, 1993, pp. 550–553.

[14] Y. Stylianou, "On the harmonic analysis of speech," in *Circuits and Systems. International Symposium on*, vol. 5, 1998, pp. 5–8.

[15] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[16] S. M. Kay, *Modern Spectral Estimation: Theory & Application*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[17] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[18] D. B. Paul, "The spectral envelope estimation vocoder," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 4, pp. 786–794, 1981.

[19] M. Kuropatwinski and W. Kleijn, "Estimation of the short-term predictor parameters of speech under noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1645–1655, 2006.

[20] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2002.

[21] D. Talkin, *Speech Coding and Synthesis*. Elsevier Science B.V., 1995, ch. 14: A Robust Algorithm for Pitch Tracking (RAPT), pp. 495–518.

[22] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 229–241, 2003.

[23] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall, 1993, vol. 1.

[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, 2001.

[25] V. Grancharov, J. Samuelsson, and W. Kleijn, "Improved Kalman filtering for speech enhancement," in *Acoustics, Speech, and Signal Processing. International Conference on*, vol. 1, 2005, pp. 1109–1112.

[26] T. Esch and P. Vary, "Model-based speech enhancement using SNR dependent MMSE estimation," in *Acoustics, Speech and Signal Processing. International Conference on*, 2011, pp. 4652–4655.

[27] ITU-T, *Objective measurement of active speech level*, ITU-T Recommendation P.56, 2011.

[28] ——, *Subjective performance assessment of telephone-band and wideband digital codecs*, ITU-T Recommendation P.830, 1996.

[29] ——, *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 Recommendation P.862, 2001.

[30] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, 2008.