A NOVEL SINGLE CHANNEL SPEECH ENHANCEMENT APPROACH BY COMBINING WIENER FILTER AND DICTIONARY LEARNING

Hung-Wei Tseng[1], *Srikanth Vishnubhotla*[2], *Mingyi Hong*[1], *Jinjun Xiao*[2], *Zhi-Quan Luo*[1], *Tao Zhang*[2]

[1] University of Minnesota, Minneapolis, MN 55455, USA
[2] Starkey Hearing Technologies, Eden Prairie, MN 55344, USA

ABSTRACT

In this paper, a novel algorithm named Sparsity-based Wiener plus Dictionary Learning (SWDL) is proposed for single channel speech enhancement. SWDL combines both Wiener filter and dictionary learning technique. The Wiener filter is used to ensure the enhanced speech is statistically optimal, while the dictionary learning technique is used to improve the enhanced speech quality and intelligibility by utilizing speech-specific information. Such information is incorporated in the pre-trained speech dictionary that can sparsely represent the clean speech spectra. When applied to the TIMIT database, SWDL outperforms the Log Mean Square-Error Short-Time Spectra Amplitude estimator (LSTSA) according to four different objective metrics measuring speech quality and intelligibility. Subjective tests also show that SWDL produces better speech quality and intelligibility than LSTSA.

Index Terms— Speech Enhancement, Dictionary Learning, Nonnegative Matrix Factorization, Wiener Filtering

1. INTRODUCTION

Single-channel speech enhancement algorithms aim to improve the quality and intelligibility of noisy speech signals. Several approaches have been proposed towards this problem (see [1] for an overview), a large number of which are statistical approaches. For example, [2, 3] and a perceptually motivated variation [4] estimate the amplitude of the short-time-Fourier transform (STFT) of the speech signal by assuming a Gaussian prior distribution in a Bayesian framework. [5] estimates the speech STFT using non-Gaussian speech priors in a maximum *a posteriori* framework. The common aspect of these statistical methods is that their performance heavily relies on the estimation of the signal-to-noise ratio (SNR), which is directly estimated from the noisy speech. When the SNR is estimated correctly, statistical methods typically produce satisfactory results. However, in moderate to low SNR, these methods inaccurately estimate the SNR and hence produce noticeable enhancement artifacts that adversely impact both intelligibility and quality. Furthermore, statistical approaches do not explicitly model or capture characteristics of speech signals. One potential extension of these approaches therefore is to incorporate speech-specific information into their formulation.

Dictionary Learning is an increasingly popular machine learning approach that can capture properties of speech signals. Atoms of a dictionary serve as the building blocks for a speech database, and each observation (e.g., the magnitude spectrum of a speech signal) is modeled as a linear combination of these atoms. The dictionary atoms are learned during a training phase that ensures optimal reconstruction of the database [6, 7, 8]. Since the speech observation can be reconstructed by the atoms, these atoms necessarily capture the salient properties of speech. Dictionary learning was first applied in audio processing for monaural sound separation [9, 10], wherein each sound source was represented by a separate dictionary. Recently, various works [11, 12, 13, 14] have applied dictionary learning to speech enhancement by viewing speech and noise as two sound sources, and training a dictionary for each source.

This paper describes a novel algorithm, called Sparsity-based Dictionary Learning (SWDL), that combines a statistical enhancement technique, i.e., the Wiener filter, with dictionary learning to improve the intelligibility and quality of the enhanced speech. Similar to the statistical methods, SWDL relies on signal statistics and SNR estimation to remove noise in the processed speech. Additionally, dictionary learning is applied to preserve the speech characteristics in the processed signal, thereby constraining the artifacts caused by inaccurate SNR estimation. Compared with existing dictionary learning works [11, 12, 13, 14], this work does not take a source separation perspective and uses only a speech dictionary.

This paper assumes the linear combination weight in dictionary learning is sparse, as assumed in [11, 14]. However, they used a fixed, pre-determined parameter to control the sparsity of the weight. This paper proposes an adaptive parameter to control the sparsity, which is computed automatically from the input noisy speech.

In the remainder of this paper, we will first describe the proposed speech enhancement algorithm and the method for selecting the sparsity parameter. We will then present the performance of the algorithm using objective and subjective evaluations, and draw conclusions and discuss future research directions.

2. METHOD

Notation: An upper case letter denotes a matrix, and a lower case letter denotes either a vector or a scalar depending on the context. $X_{k,n}$ represents the k^{th} row and n^{th} column of the matrix X. The n^{th} column of a matrix X is denoted by X_n. Bold face represents complex-valued quantities. For complex-valued $\mathbf{X}_{k,n}$, $X_{k,n}$ and $\theta_{k,n}$ denote its magnitude and phase respectively. $\|X\|_1 = \sum_{k,n} |X_{k,n}|$ denotes the sum of the absolute values of all entries. $\|x\|_u^2 = \sum_i u_i |x_i|^2$ represented the weighted L_2-norm. $R_+^{K\times M}$ denotes the set of $K \times M$ matrices with nonnegative entries. $\bigodot$ and $\geq$ denote entry-wise multiplication and entry-wise "greater or equal to" respectively.

2.1. System Model

Let $x(t)$, $y(t)$ and $v(t)$ denote, respectively, the clean speech, the observed noisy speech, and the noise, at time index t. In the single channel additive model, we assume $y(t) = x(t) + v(t)$. Applying STFT, the equivalent time-frequency model is $\mathbf{Y}_{k,n} = \mathbf{X}_{k,n} + \mathbf{V}_{k,n}$, where $\mathbf{Y}_{k,n}$, $\mathbf{X}_{k,n}$, and $\mathbf{V}_{k,n}$ denote, respectively, the complex-valued spectrum of $y(t)$, $x(t)$ and $v(t)$, at frequency bin $k \in \{1, 2, \cdots, K\}$ and at time frame $n \in \{1, 2, \cdots, N\}$. $\mathbf{X}_{k,n}$ and $\mathbf{V}_{k,n}$ are assumed to be independent zero-mean random variables with variance $\varsigma_{k,n}^2$ and $\sigma_{k,n}^2$, respectively.

2.2. Proposed Formulation

The enhanced spectrum $\hat{\mathbf{X}}_{k,n}$ is obtained by multiplying the observed noisy spectrum $\mathbf{Y}_{k,n}$ with a potentially complex-valued gain coefficient $\mathbf{h}_{k,n}$. The Wiener filter can be used to find the optimal gain coefficients that minimize the mean square error between the enhanced speech and the clean speech according to the following:

$$\begin{aligned}\mathbf{h}_{k,n} &= \arg\min_{\mathbf{h}} \frac{1}{2}\mathbb{E}\left[\|\mathbf{h}\mathbf{Y}_{k,n} - \mathbf{X}_{k,n}\|^2\right] \\ &= \arg\min_{\mathbf{h}} \frac{1}{2}\left(\varsigma_{k,n}^2 + \sigma_{k,n}^2\right)\left(\mathbf{h} - h_{k,n}^w\right)^2 \end{aligned} \quad (1)$$

where $h_{k,n}^w \triangleq \frac{\varsigma_{k,n}^2}{\varsigma_{k,n}^2+\sigma_{k,n}^2}$ is known as the Wiener gain coefficient. Because the Wiener gain coefficient is real and non-negative, the phase of the enhanced spectrum is the same as that of the noisy spectrum, and the Wiener gain only scales the magnitude of the noisy spectrum. Therefore, in the remainder of this paper, we consider only the spectrum magnitude and refer to it as the "spectrum", unless otherwise specified. Using a change of variable $h = \frac{x}{Y_{k,n}}$, we can reformulate the Wiener filter (1) into the following

$$\begin{aligned}\hat{X}_{k,n} &= \arg\min_{x} \frac{1}{2}\left(\varsigma_{k,n}^2 + \sigma_{k,n}^2\right)\left(\frac{x}{Y_{k,n}} - h_{k,n}^w\right)^2 \\ &= \arg\min_{x} \frac{1}{2}\frac{1+\xi_{k,n}}{\gamma_{k,n}}\left(x - X_{k,n}^w\right)^2 \end{aligned} \quad (2)$$

where $X_{k,n}^w \triangleq h_{k,n}^w \times Y_{k,n}$, $\xi_{k,n} \triangleq \frac{\varsigma_{k,n}^2}{\sigma_{k,n}^2}$ denotes the *a priori* SNR, and $\gamma_{k,n} \triangleq \frac{Y_{k,n}^2}{\sigma_{k,n}^2}$ denotes the *a posteriori* SNR. The Wiener filter (2) is now cast as an unconstrained optimization problem over all possible spectra.

We will now leverage dictionary learning to capture speech characteristics in the spectrum domain, and use it to transform (2) into a constrained optimization to reduce artifacts and improve intelligibility. In particular, we exploit the fact that a clean speech spectrum can be written as a non-negative *sparse* linear combination of a properly designed dictionary. To formalize ideas, let $x \in \mathcal{R}_+^{K\times 1}$ denote the clean speech spectrum of a particular time frame, $D \in \mathcal{R}_+^{K\times M}$ denote the dictionary, and $g \in \mathcal{R}_+^{M\times 1}$ denote the gain coefficient, where typically $M >> K$ to allow overcomplete representation. Then

$$x \approx D \times g, \quad \text{and g is a sparse vector} \quad (3)$$

Columns of D, called atoms, serve as the basic building blocks that can approximate any clean speech spectrum. Equation (3) implies that the clean speech spectrum can be written as a weighted sum of only a few atoms; however, different speech spectra can use different sets of atoms. Provided D is chosen properly, the sparsity of the gain coefficient plays an important role in capturing the speech characteristics. Specifically, the sparsity assumption implies that for any $\hat{x}$ that can be expressed as $D \times g$, only those $\hat{x}$ for which g is sparse will typically approximate clean speech. In other words, one way to ensure $\hat{x}$ approximates clean speech is to search for a sparse g that can represent $\hat{x}$ using D.

We will now illustrate how such a sparse gain coefficient g can be estimated given the dictionary D, and defer the training of the sparsity-inducing dictionary D to the next section. Assuming $X_{k,n}^w$ is known, the proposed SWDL formulation is shown in (4):

$$\begin{aligned}\hat{X}_n &= D \times \hat{g} \\ \hat{g} &= \arg\min_{g\geq 0} \sum_{k=1}^{K}\left\{\frac{1}{2}\frac{1+\xi_{k,n}}{\gamma_{k,n}}\left((Dg)_{k,n} - X_{k,n}^w\right)^2\right\} + \beta\|g\|_1 \\ &= \arg\min_{g\geq 0} \frac{1}{2}\|Dg - X_n^w\|_{u_n}^2 + \beta\|g\|_1 \end{aligned} \quad (4)$$

where $u_n = [\frac{1+\xi_{1,n}}{\gamma_{1,n}}, \cdots, \frac{1+\xi_{K,n}}{\gamma_{K,n}}]^T$

We constrain the enhanced spectrum $\hat{X}_n$ (optimization variable of (2)) to be represented by the properly-chosen D and a gain coefficient $\hat{g}$, and use the L_1-norm regularization term to promote the sparsity of $\hat{g}$. The problem is now converted into a constrained optimization over gain, with the parameter β controlling the sparsity level. Thus, SWDL still uses the mean square error as the objective, but additionally requires that the enhanced spectrum be sparsely represented by the dictionary. Since the enhanced speech spectrum has a sparse representation under D, our constrained formulation is expected to result in better performance than the unconstrained one.

The overall SWDL algorithm is summarized in Algorithm 1. The calculation of the required signal statistics for speech enhancement occurs in step 4. We use a decision directed approach for *a priori* SNR estimation, using an adaptive weight $\alpha_{k,n}$ as proposed in [15]. For the quadratic minimization with a nonnegative constraint (4), we use the optimal first order method proposed by Nesterov [16] due to its simplicity and efficiency.

Algorithm 1 SWDL

Require: : noisy speech $y(t)$, dictionary D, sparsity parameter β

1: $\mathbf{Y} = Y \bigodot \exp(j\theta) = \text{STFT}(y(t))$

2: **for** $n = 1 \rightarrow N$ **do**

3: Estimating noise power $\sigma_{k,n}^2$

4: Estimating $\xi_{k,n}$ and $\gamma_{k,n}$

$$\begin{aligned}\gamma_{k,n} &= \frac{Y_{k,n}^2}{\sigma_{k,n}^2} \\ \xi_{k,n} &= \alpha_{k,n}\frac{\hat{X}_{k,n-1}^2}{\sigma_{k,n}^2} + (1-\alpha_{k,n})\bar{\gamma}_{k,n} \quad (5) \\ \alpha_{k,n} &= \frac{1}{1+\left(\frac{\bar{\gamma}_{k,n}-\xi_{k,n-1}}{1+\bar{\gamma}_{k,n}}\right)^2} \quad (6) \\ \bar{\gamma}_{k,n} &= \max[\gamma_{k,n}-1, 0] \end{aligned}$$

5: Estimating $X_{k,n}^w$: $X_{k,n}^w = \frac{\xi_{k,n}}{1+\xi_{k,n}}Y_{k,n}$

6: Estimating $\hat{X}_{k,n}$

$$\begin{aligned}\hat{X}_n &= D \times \hat{g} \\ \hat{g} &= \text{SWDL (4)} \end{aligned} \quad (7)$$

7: **end for**

8: Enhanced STFT: $\hat{\mathbf{X}} = \hat{X} \bigodot \exp(j\theta)$

9: **return** Enhanced speech: $\hat{x}(t) = \text{IFFT}(\hat{\mathbf{X}})$

2.3. Dictionary Training

The SWDL assumes that there exists a dictionary D such that clean speech spectra can be represented sparsely. In practice, such a sparsity-inducing dictionary need to be properly trained during a training stage. In the training stage, we use sparse Nonnegative Matrix Factorization (NMF) [8] to arrive at such a dictionary. Specifically, let X denote the collection of speech spectra from the training sentences. Then the desired dictionary is obtained by solving problem (8), which enforces the dictionary to well-represent the speech using a sparse gain matrix:

$$[D, G] = \arg\min_{D\geq 0, G\geq 0} \frac{1}{2}\|X - DG\|_F^2 + \beta_t\|G\|_1 \quad (8)$$

Even though the general NMF problem has been shown to be NP-hard [17], we use an efficient algorithm proposed in [8] to approximately solve (8).

2.4. Automatic β Selection for SWDL

The sparsity coefficient β is a key parameter: the higher the β, the sparser the gain coefficient $\hat{g}$. In this section, we will present a method for selecting β from the noisy speech automatically. Given a noisy observation $\mathbf{Y}$, let $\hat{\mathbf{X}}_\beta$ be the enhanced signal using the parameter β. Intuitively, if $\hat{\mathbf{X}}_\beta$ is a good estimate of the clean speech spectrum, then the power of the estimated noise, i.e., $||\hat{\mathbf{X}}_\beta - \mathbf{Y}||^2$, should be close to the true noise power:

$$E(\beta) \equiv \|\hat{\mathbf{X}}_\beta - \mathbf{Y}\|^2 \approx \|\mathbf{X} - \mathbf{Y}\|^2 = \sum_{k,n} \sigma_{k,n}^2 \qquad (9)$$

Thus, the optimal β^* is the β that makes (9) true. The noise power $\sigma_{k,n}^2$ is typically estimated from the noisy speech directly. Figure 1 illustrates how $E(\beta)$ varies with β. The monotonicity of $E(\beta)$ can be explained by the fact that the larger the β, the sparser the G, hence the smaller the $\hat{\mathbf{X}}_\beta$ and the larger the value of $E(\beta)$. This was also confirmed empirically in our experiments across different noise types and SNR values. Such monotonicity suggests that a simple bisection scheme can be used to find the optimal β^*.

Interestingly, the optimal β^* found in this way is also perceptually meaningful. Figure 1 also shows the variation of two perceptually motivated objective metrics, Perceptual Evaluation of Speech Quality (PESQ, [18]), and Intelligibility Index (I3, [19]), as a function of β. PESQ is an objective metric that assesses the quality of processed speech using the loudness difference between clean and processed speech, and has been shown to correlate highly with the quality of speech processed by noise reduction algorithms. PESQ ranges from 1 (bad quality) to 4.5 (excellent quality). I3 is an objective metric similar to the Speech Intelligibility Index (SII) [19] that assesses the intelligibility of the processed speech, and ranges from 0 to 1. The PESQ and I3 curves of Figure 1 were obtained by enhancing the noisy speech using SWDL with different values of β, and then evaluating the enhanced speech for each β using these two metrics. For the ease of comparison, we normalize PESQ and I3 by their maximum values, and plot the normalized metrics in Figure 1. Figure 1 shows that the PESQ and I3 scores using β^* is close to the optimal PESQ and I3 scores.

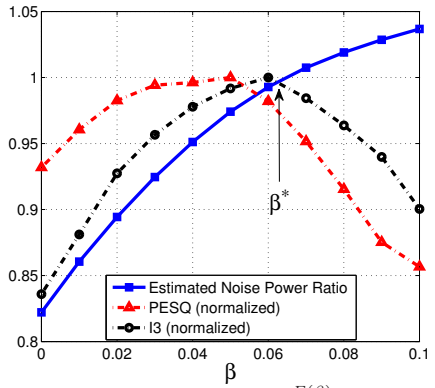


Fig. 1. Estimated Noise Power Ratio ($\frac{E(\beta)}{\sum_{k,n}\sigma_v^2(k,n)}$) and normalized PESQ and I3 (both normalized by their maximum values) versus different β value in -5dB AWGN noise. β^* is the selected β

3. PERFORMANCE EVALUATION

The quality and intelligibility of the speech processed by SWDL was compared with that processed by the Wiener filter, LSTSA [2] and unprocessed speech through objective evaluations. Since the Wiener filter was outperformed by the LSTSA in most scenarios of this evaluation, subjective comparison was performed only with the LSTSA, and we report only comparisons with the LSTSA in this paper.

The TIMIT database [20] was used for evaluation due to the availability of a large database for dictionary training. One universal dictionary was trained for each gender of speakers, each using one hour of speech selected randomly form the "train" subset. For enhancement (test), 320 male sentences and 160 female sentences were selected randomly from the "test" subset. Four different stationary noises (White Gaussian Noise, Speech Shape Noise, Vacuum, and Airplane) were added to each test sentence at three different SNRs ($-5, 0, 5$ dB). The active speech level of the clean speech signal was first determined using the method B of ITU-T P.56 [21], and the noise sample was then appropriately scaled and added to the clean speech to obtain the desired SNR. All noisy signals were processed by each of the 2 algorithms (SWDL, LSTSA). For the objective evaluation, all processed samples were evaluated. For the subjective evaluation, only a subset of the processed samples were used, as described below.

All sentences were sampled at 8 kHz, and segmented into 30-ms duration frames using a Hamming window with 50% overlap. A 512 point FFT/IFFT was used for the time-frequency analysis and synthesis operations. To eliminate the effect of inaccurate noise power estimation, we assumed the true noise power was known for SWDL and LSTSA. In LSTSA, we fixed $\alpha_{k,n}$ for the *a priori* SNR estimation (5) to 0.98 [2]. For SWDL dictionary training, the dictionary size (M) was set to 512 and the sparsity constant (β_t) to 0.001.

3.1. Objective Evaluation

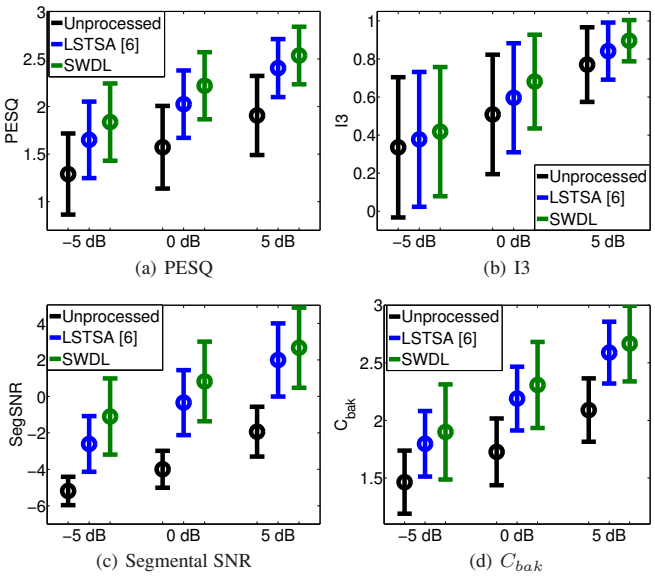
We used PESQ [18] to evaluate the quality and I3 [19] to assess the intelligibility of the processed signals (see section 2.4). Segmental SNR [22] was used to measure the residual distortions in the processed speech compared to the clean speech. The background noise reduction was measured using C_{bak} [23]. For all objective measures, higher values indicate better performance. Figure 2 shows standard error plots for the performance of the three *treatments* (SWDL, LSTSA, unprocessed) using each of the four metrics. Each error plot shows the standard deviation of the metric scores (whiskers) about their mean values (circle), for a given SNR and treatment and averaged across all noise types.

Analysis of variance (ANOVA) performed individually for each SNR indicated a highly significant effect of speech enhancement on all four metrics ($p < 10^{-5}$ for all SNRs and metrics). Multiple comparison tests according to Tukey's HSD test [24] indicated that SWDL consistently produced a significant improvement ($p < 10^{-5}$) over LSTSA at all SNRs, and across all four metrics.

Since the speakers and the sentences used in dictionary training and evaluation were different, it is fair to say that the SWDL generalized well to unseen data. In a preliminary study, however, we found the performance of SWDL improved when the dictionary was customized to each individual speaker. In this work, we present only the results of using a universal dictionary, and defer a comprehensive investigation of the performance using a customized dictionary to the future.

3.2. Subjective Evaluation

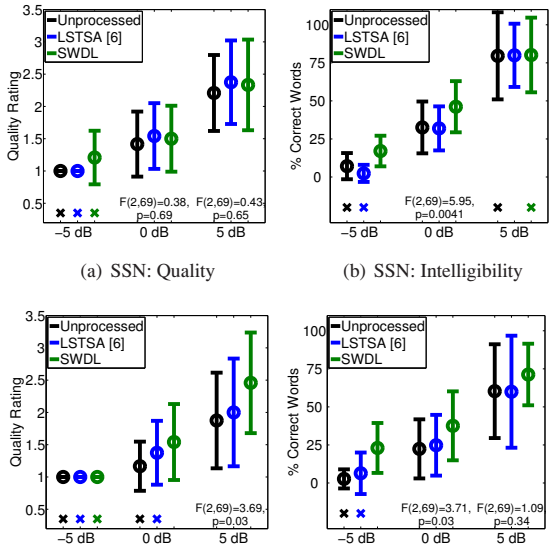
The SWDL algorithm was compared to the LSTSA and unprocessed speech for quality and intelligibility using subjective experiments. Only the Speech-Shaped Noise (SSN) and Vacuum noise at SNRs ($-5, 0, 5$ dB) were evaluated, giving a total of six *conditions* over the three *treatments* (SWDL, LSTSA, unprocessed) for each experiment. For subjective evaluation, only a subset of the sentences used for the objective evaluation was used: low-context sentences between 4



(a) PESQ (b) I3

(c) Segmental SNR (d) C_{bak}

Fig. 2. Error plots showing the performance of unprocessed speech, SWDL and LSTSA at different SNRs based on four different objective measures. The result is averaged over four noise types (AWGN, SSN, Vacuum, Airplane) and all 480 test sentences.



(a) SSN: Quality (b) SSN: Intelligibility

(c) Vacuum: Quality (d) Vacuum: Intelligibility

Fig. 3. Error plots for the subject responses in the Quality (left) and Intelligibility (right) tests for SSN (top) and vacuum noise (bottom) and for the three SNRs. A cross ("x") under an error plot indicates that the data was not normally distributed and showed a flooring or ceiling effect for that treatment. The text under a set of error plots indicates the F-statistic and p-value for ANOVA using that set of treatments.

and 10 words long, from male speakers. For clarity, the term "stimulus" refers to a test sentence under a given treatment.

Nine native American English speakers with normal hearing (all employees of Starkey Hearing Technologies) participated in the study. Each subject participated in one Intelligibility and one Quality session, and the sessions were spaced one week apart. There was no overlap of stimuli between the two sessions for any subject. Subjects were presented four stimuli per condition per treatment for the Intelligibility and Quality sessions. The test data was staggered across subjects such that each stimulus received an equal number of responses. The presentation order of the stimuli was randomized for each session. The stimuli were scaled such that the active speech level [21] of the clean speech (i.e., before noise addition and reduction) was 79 dB Sound Pressure Level (SPL). The stimuli were presented diotically via Sennheiser HD600 supra-aural headphones.

For the intelligibility session, subjects were asked to repeat all the words they heard, and the percentage of correctly identified words was calculated. For the quality session, subjects were asked to rate the overall quality of the stimulus on a scale of 1 to 5, taking into account both the speech distortions and background intrusions, as recommended in ITU-T P.835 [25]. A score of 1 indicated *Bad* quality, and a score of 5 indicated *Excellent* quality.

Figure 3 shows standard error plots for the quality ratings and word recognition rates from the subjects. The data for the two noise types are shown in separate plots. Each error plot shows the standard deviation of the subject responses (whiskers) about their mean response (circle), for a given condition and treatment. For some of these conditions and treatments, the subject responses were not normally distributed (Lilliefors test [26] at the $p < 0.005$ significance level) due to flooring or ceiling effects. The flooring effect was likely due to the difficulty of the task at low SNR, and the ceiling effect likely due to the ease of the task at high SNR. For the sake of consistency, these responses were also represented using error plots, but a cross mark ("x") was used to identify such data.

Analysis of variance (ANOVA) was performed only for the conditions wherein all three treatments were normally distributed. The F-statistic and p-value for each ANOVA analysis are given underneath the corresponding error plot in Figure 3. ANOVA indicated a significant effect of speech enhancement on intelligibility at 0 dB SNR for both noise types. Furthermore, multiple comparison tests according to Tukey's HSD test [24] indicated that SWDL produced a significant improvement ($p < 0.05$) in intelligibility over LSTSA at 0 dB SNR for both noise types. This is in line with predictions by the objective evaluation (Figure 2(b)). ANOVA also indicated a significant effect ($p < 0.05$) of speech enhancement on speech quality for vacuum noise at 5 dB SNR, in which case the SWDL showed a significant improvement over LSTSA. In all other conditions, the SWDL and LSTSA showed insignificant difference.

These objective and subjective results indicate that the SWDL outperforms or matches the LSTSA in quality and intelligibility benefits. This superior performance emphasizes the advantage that the SWDL offers over the LSTSA, in terms of capturing speech characteristics by using a dictionary.

4. CONCLUSION AND FUTURE WORK

In this paper we describe SWDL, a novel algorithm combining the Wiener filter with dictionary learning for speech enhancement, and demonstrate its superior performance through objective evaluation and subjective evaluation. SWDL currently captures speech characteristics in the frequency domain but not in the time domain. Since speech exhibits unique spectro-temporal characteristics, we plan to extend SWDL to jointly capture the spectro-temporal information. We also plan to customize the dictionary towards individual phonemes to capture their distinct characteristics.

Acknowledgements:

The authors would like to thank Dr. Drew Dundas for helping with the subject experiment design and execution.

5. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, CRC, 1 edition, June 2007.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443 – 445, apr 1985.

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

[4] P.C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857 – 869, sept. 2005.

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

[6] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311 –4322, nov. 2006.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[8] J. Eggert and E. Korner, "Sparse coding and nmf," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, july 2004, vol. 4, pp. 2529 – 2533 vol.4.

[9] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," *Interspeech"06, Int. Conf. Spoken Lang. Process.*, 2006.

[10] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 –1074, march 2007.

[11] M.N. Schmidt, J. Larsen, and Fu-Tien Hsiao, "Wind noise reduction using non-negative sparse coding," in *IEEE Workshop on Machine Learning for Signal Processing, 2007*, aug. 2007, pp. 431 –436.

[12] K.W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.*, 31 2008-april 4 2008, pp. 4029 –4032.

[13] M.N. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*, oct. 2008, pp. 486 –491.

[14] C.D. Sigg, T. Dikk, and J.M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1698 –1712, aug. 2012.

[15] M.K. Hasan, S. Salahuddin, and M.R. Khan, "A modified a priori snr for speech enhancement using spectral subtraction rules," *Signal Processing Letters, IEEE*, vol. 11, no. 4, pp. 450 – 453, april 2004.

[16] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer, 2003.

[17] S.A. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1364–1377, 2009.

[18] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings.*, 2001, vol. 2, pp. 749 –752 vol.2.

[19] J.M. Kates and K.H. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, pp. 2224, 2005.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993.

[21] ITU-T P.56, "Objective measurement of active speech level ITU-T recommendation p.56.," 1993.

[22] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," *Proc. Int. Conf. Spoken Lang. Process.*, vol. 17, 1998.

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

[24] W. L. Hays, *Statistics*, Holt, Rinehart and Winston, 4 edition, 1988.

[25] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm ITU-T recommendation p.835.," 2003.

[26] H. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, 1967.