A Single Channel Speech Enhancement Approach by Combining Statistical Criterion and Multi-Frame Sparse Dictionary Learning

Hung-Wei Tseng [1], Srikanth Vishnubhotla[2], Mingyi Hong [1], Xiangfeng Wang [3], Jinjun Xiao[2], Zhi-Quan Luo[1], Tao Zhang[2]

[1]University of Minnesota, Minneapolis, USA
[2]Starkey Hearing Technologies, Eden Prairie, USA
[3]Nanjing University, Nanjing, China

Abstract

In this paper, we consider the single-channel speech enhancement problem, in which a clean speech signal needs to be estimated from a noisy observation. To capture the characteristics of both the noise and speech signals, we combine the well-known Short-Time-Spectrum-Amplitude (STSA) estimator with a machine learning based technique called Multi-frame Sparse Dictionary Learning (MSDL). The former utilizes statistical information for denoising, while the latter helps better preserve speech, especially its temporal structure. The proposed algorithm, named STSA-MSDL, outperforms standard statistical algorithms such as the Wiener filter, STSA estimator, as well as dictionary based algorithms when applied to the TIMIT database, using four different objective metrics that measure speech intelligibility, speech distortion, background noise reduction, and the overall quality.

Index Terms: Speech Enhancement, Dictionary Learning, STSA, ADMM, contextual effects

10.21437/Interspeech.2013-133

1. Introduction

Traditionally, the single-channel speech enhancement problem has been tackled by leveraging the statistical properties of both speech and noise signals in the short-time-Fourier-transform (STFT) domain (see [1] for an overview). For example, the STSA estimator in [2] estimates the spectrum magnitude by assuming a Gaussian prior distribution in the STFT domain. Under non-Gaussian prior assumption, [3] proposes a spectrum amplitude estimator using a maximum *a posteriori* framework. These statistical methods purely view speech and noise as two statistical sources without taking speech-specific information such as formant behavior or temporal properties into account. Also, their performance primarily depends on the accuracy of the signal-to-noise (SNR) estimation. However, in moderate or high noise environments, accurate SNR estimation is very challenging. Therefore, statistical methods alone typically do not increase speech intelligibility. One potential improvement on these methods is to incorporate the ability to learn speech-specific information.

Dictionary Learning (DL) is a machine learning approach that attempts to learn and model speech-specific information. In this approach, speech data (e.g., a spectrum magnitude) can be modeled as a linear combination of dictionary atoms, which are learned beforehand during a training phase. DL related algorithms have been successfully applied in audio processing for monaural sound separation [4, 5] and in speech enhancement [6, 7]. *Sparse* DL is a popular variation of DL, which further requires that the modeled data be represented by only a *small number* of dictionary atoms [8, 9, 10]. Compared with its non-sparse counterpart, the sparse DL approach is shown to be significantly more effective in applications such as image enhancement [11] and speech enhancement [12].

The classical statistical and Sparse DL-based approaches were recently combined in [13] in the so-called Sparsity-based Wiener plus Dictionary Learning (SWDL) algorithm, which outperformed the Log STSA estimator [14] in terms of both objective and subjective evaluations. SWDL demonstrated that statistical methods can benefit through explicit incorporation of speech-modeling, by constraining the statistical estimation as a function of a dictionary-based speech model.

This paper extends the SWDL approach by modeling the *temporal structure* of speech, the preservation of which is important for both quality and intelligibility. Since, in continuous speech, context significantly influences the distribution of spectro-temporal energy (and thus the spectrum to be modeled), modeling a spectrum by accounting for its context is expected to result in better speech modeling and thus better enhancement. Context modeling can also capture certain unique formant transitions (e.g. influence of the phoneme /r/ on its neighbors) or pitch variations. Existing DL approaches [6, 7, 12, 13] model speech-specific information only from a *single* spectrum frame, which typically captures only 20-30 msec of speech. We therefore propose a novel algorithm named STSA-MSDL, an extension of SWDL [13] that explicitly captures temporal structure by combining the STSA estimator with *multi-frame sparse* DL. In multi-frame sparse DL, we concatenate multiple speech spectra into a speech "patch", and model this contextual speech patch using DL. Since learning information from multiple frames requires operating in a significantly increased problem dimension, we also propose a dimensionality reduction method to render the learning problem manageable.

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. Bold face represents complex-valued quantities. x_n represents the n^{th} column of the matrix X, and $x_{k,n}$ represents the corresponding (k, n) entry of X. $\bar{x}_n = [x_{n-2}; x_n; x_{n+2}]$ denotes the concatenation of three columns of X centered at column n. For complex-valued $\mathbf{X}$, X and θ denote its magnitude and phase respectively. $\|X\|_1 = \sum_{k,n} |X_{k,n}|$ denotes the sum of the absolute values of all entries. $\odot$ and $\geq$ denote entry-wise multiplication and entry-wise "greater or equal to" respectively. $I_0(\cdot)$ and $I_1(\cdot)$ represent the modified Bessel function of zero and first order. $\mathrm{blkdiag}(X, Y, Y)$ denotes the block diagonal matrix with X, Y, Z being the diagonal blocks. $\langle \cdot, \cdot \rangle$ represent the standard inner product in the Euclidean space.

2.1. System Model

Consider the single-channel speech enhancement problem, which aims to recover the clean speech $x^c(t)$ from a noisy observation $y(t)$:

$$y(t) = x^c(t) + u(t)$$

where $u(t)$ denotes additive noise and t denotes time. Applying STFT, the equivalent time-frequency model is $\mathbf{y}_{k,n} = \mathbf{x}^c_{k,n} + \mathbf{u}_{k,n}$, where $\mathbf{y}_{k,n}$, $\mathbf{x}^c_{k,n}$, and $\mathbf{u}_{k,n}$ denote, respectively, the complex-valued spectrum of $y(t)$, $x^c(t)$ and $u(t)$, at frequency bin $k \in \{1, 2, \cdots, K\}$ and time frame $n \in \{1, 2, \cdots, N\}$. $\mathbf{x}^c_{k,n}$ and $\mathbf{u}_{k,n}$ are assumed to be independent zero-mean random variables with variance $\varsigma^2_{k,n}$ and $\sigma^2_{k,n}$, respectively. In this paper, $\sigma_{k,n}$ is assumed to be known and $\varsigma_{k,n}$ is estimated from the noisy spectrum using Eq. (10).

2.2. Proposed Formulation

Let $\hat{\mathbf{X}}$ denote the complex-valued enhanced spectrum, with magnitude $\hat{X}$. The popular STSA enhancement method [2] estimates $\hat{X}$ by minimizing the square error:

$$\begin{aligned} \hat{x}_{k,n} &= \arg\min_x \frac{1}{2}\mathbb{E}\left[|x - x^c_{k,n}|^2 | \mathbf{y}_{k,n}, \sigma_{k,n}, \varsigma_{k,n}\right] \\ &= \arg\min_x \frac{1}{2}|x - z_{k,n}|^2 \end{aligned} \quad (1)$$

where

$$\begin{aligned} z_{k,n} &= y_{k,n}\frac{\sqrt{\pi}}{2}\frac{v_{k,n}}{\gamma_{k,n}}e^{-\frac{v_{k,n}}{2}} \\ &\quad \left[(1+v_{k,n})I_0\left(\frac{v_{k,n}}{2}\right) + v_{k,n}I_1\left(\frac{v_{k,n}}{2}\right)\right] \end{aligned} \quad (2)$$

and $v_{k,n} \triangleq \frac{\xi_{k,n}}{1+\xi_{k,n}}\gamma_{k,n}$. Here, $\xi_{k,n} \triangleq \frac{\varsigma^2_{k,n}}{\sigma^2_{k,n}}$ denotes the *a priori* SNR and $\gamma_{k,n} \triangleq \frac{Y^2_{k,n}}{\sigma^2_{k,n}}$ denotes the *a posteriori* SNR.

Similar to the STSA, this work will also focus on only estimating the spectrum magnitude of the enhanced speech. The phase of the enhanced spectrum will be assumed to be the same as that of the noisy spectrum. Therefore, in the remainder of this paper, the term "spectrum" will refer to the spectrum *magnitude* alone, unless otherwise specified.

Since x is *unconstrained* in (1), the optimal value of the estimated spectrum magnitude is $z_{k,n}$, the minimum mean square estimate given the noisy observation $\mathbf{y}_{k,n}$. Our proposed formulation STSA-MSDL is a *constrained* version of the STSA, where the constraints exploit the temporal structure of speech.

To preserve temporal dynamics, we propose using multi-frame sparse dictionary learning. Specifically, we learn a dictionary that can simultaneously represent three consecutive frames x_{n-2}, x_n, and x_{n+2}, thereby capturing the temporal relationship between them. In this setting, each *spectrum patch* $\bar{x}_n \triangleq [x_{n-2}; x_n; x_{n+2}]$ covers a long contextual window, instead of a traditional frame. We collect patches $\{\bar{x}_n\}^{3K\times 1}$ from the training sentences, and learn a dictionary $D \in \mathcal{R}^{3K\times M}$ with M dictionary atoms each of dimension $3K$ that can *sparsely represent* all the training patches: as Eq. (3):

$$\bar{x}_n \approx Dg_n, \quad \text{and } g_n \text{ is a sparse vector} \quad (3)$$

Typically, $M \gg 3K$ to allow overcompleteness. Eq. (3) means that any speech-like temporal dynamic pattern can be reconstructed by using a sparse linear combination of atoms of D, since the pattern has been modeled by D. Conversely, any patch not exhibiting the temporal dynamics of typical speech will not be sparsely represented by D. Thus, the sparsity of the representation is the key to capturing speech temporal dynamics, and furthermore to separate speech-like signals from non-speech.

Due to the concatenation of frames into a patch, the dictionary dimensionality is large $(3K \times M)$, making the training process difficult due to the curse of dimensionality [15]. It is therefore desirable to learn the sparse dictionary in a reduced dimension, that still contains most of the information in the original signal $\bar{x}_n$. One possible way to do this is by *linear compressing* the training data to a lower dimensionality $3d\,(\ll 3K)$:

$$\begin{aligned} \bar{x}_n &\approx RW\bar{x}_n \\ W\bar{x}_n &= Dg_n, \quad \text{and } g_n \text{ is a sparse vector} \end{aligned} \quad (4)$$

where $W \in \mathcal{R}^{3d\times 3K}$ and $R \in \mathcal{R}^{3K\times 3d}$ is a pair of linear compression and decompression matrices. This reduces the dictionary dimension from $(3K \times M)$ to $(3d \times M)$, thus facilitating fast training. In section 2.3, we will present an efficient way to learn $\{W, R\}$ that preserves most of the perceptually relevant spectral and temporal information of the original signal $\bar{x}_n$. It should be noted that larger patch sizes (context window) are also possible; the choice of 3 frames (corresponding to 90 msec) was motivated by a compromise between capturing adequate context and limiting training complexity.

To combine the above sparse dictionary with the statistical criterion, we constrain that the reconstructed signal x_n is close to the STSA estimator while at the same time being sparsely representable by the multi-frame sparse dictionary. Specifically, we estimate the enhanced speech spectrum by solving the following optimization problem:

$$\begin{aligned} [\hat{X}, \hat{G}] &= \arg\min_{X,G} \sum_{n=1}^{N} \frac{1}{2}\|x_n - z_n\|^2 + \lambda_s\|x_n\|_1 + \lambda_g\|g_n\|_1 \\ &\quad \text{s.t.} \quad \bar{x}_n = RDg_n, \;\; x_n \geq 0 \;\; \forall n = 1, \cdots N \end{aligned} \quad (5)$$

In (5), we require that the enhanced spectrum $\hat{X}$ has small statistical mean square error $\sum_{n=1}^{N}\frac{1}{2}\|\hat{x}_n - z_n\|^2$. We further require the enhanced spectrum patch $\hat{\bar{x}}_n$ to be sparsely represented by the dictionary, by: *i)* constraining the enhanced speech to satisfy $\hat{\bar{x}}_n = RD\hat{g}_n$, and *ii)* penalizing the sparsity-inducing L_1 norm of the coefficient vector $\hat{g}_n$. We also penalize the L_1 norm of the enhanced speech $\hat{x}_n$ itself, since the clean spectrum typically has sparse columns.

To efficiently solve Eq. (5), we present an inexact variant of the popular Alternating Direction Method of Multiplier (ADMM [16]) in Algorithm 1, which admits closed form updates and is guaranteed to converge. Let $L(X, G, U)$ denote the augmented Lagrangian function, where U is the dual variable and $\rho > 0$ denotes the constraint violation parameter:

$$\begin{aligned} L(X, G, U) &= \sum_n \frac{1}{2}\|x_n - z_n\|^2 + \lambda_s\|x_n\|_1 + \lambda_g\|g_n\|_1 \\ &\quad + \langle u_n, RDg_n - \bar{x}_n\rangle + \frac{\rho}{2}\|RDg_n - \bar{x}_n\|^2 \end{aligned}$$

The three steps of classical ADMM are primal updates in Eq. (6) and dual update in Eq. (8).

$$G^{r+1} = \arg\min_G L(X^r, G, U^r) \quad (6)$$

$$X^{r+1} = \arg\min_{X\geq 0} L(X, G^{r+1}, U^r) \quad (7)$$

$$U^{t+1} = U^t + \rho\left(RDG^{r+1} - \bar{X}^{r+1}\right) \quad (8)$$

Because of the structure of $L(X, G, U)$, the update for X admits a closed form, but not the update for G. Therefore, we instead solve the G update *inexactly* by minimizing another local upper-bound function as shown in Eq. (9), which now admits a closed form update. The inexact ADMM is significantly faster

Algorithm 1 Inexact ADMM for solving Eq. (5)

Require: : STSA estimate Z, decompression matrix R, dictionary D, sparsity parameter λ_s and λ_g

1: **for** iteration r **do**

2: G update: solve a local upper bound minimization

$$\begin{aligned} g_n^{r+1} &= \arg\min_{g_n} \rho\langle -(RD)^T\left(\bar{x}_n^r - \frac{1}{\rho}u_n^r\right), g - g_n^r\rangle \\ &\quad +\rho\frac{L_{RD}}{2}\|g_n - g_n^r\|^2 + \lambda_f\|g_n\|_1 \\ &= Shrink\left(g_n^r - \frac{1}{L_{RD}}e_n^r, \frac{\lambda_f}{\rho}, L_{RD}\right) \end{aligned} \quad (9)$$

where $e_n^r = -(RD)^T\left(\bar{x}_n^r - \frac{1}{\rho}u_n^r\right)$, and L_{RD} is the largest eigenvalue of $(RD)^T(RD)$.

$$Shrink(x,\gamma,\rho) = \begin{cases} 0, & \text{if } |x| \le \frac{\gamma}{\rho} \\ x - \frac{\gamma}{\rho}, & \text{if } x > \frac{\gamma}{\rho} \\ x + \frac{\gamma}{\rho}, & \text{if } x < -\frac{\gamma}{\rho} \end{cases}$$

3: X update: Eq. (7)

4: U update: Eq. (8)

5: **end for**

than the classical ADMM. We have proved that Algorithm 1 (inexact ADMM) indeed converges to the global optimal solution of problem (5). This proof is omitted due to the lack of space.

The overall STSA-MSDL algorithm is summarized in Algorithm 2. A simple Maximum Likelihood Estimator (10) is used to estimate the speech variance $\varsigma_{k,n}^2$.

Algorithm 2 STSA-MSDL

Require: : noisy speech $y(t)$, multi-frame sparse dictionary D, decompression matrix R, sparsity parameter λ_s and λ_g

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

2: Estimate noise variance $\sigma_{k,n}^2$ from $\mathbf{Y}$ using any noise tracking algorithm.

3: Estimate speech variance $\varsigma_{k,n}^2$ from $\mathbf{Y}$.

$$\varsigma_{k,n}^2 = \max\left[|\mathbf{y}_{k,n}|^2 - \sigma_{k,n}^2, 0\right] \quad (10)$$

4: Estimate STSA result $z_{k,n}$ via solving Eq. (2).

5: Estimate enhanced spectrum $\hat{X}$ by solving Eq. (5) using Algorithm 1

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

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

2.3. Dimensionality Reduction by Linear Compression

A coordinate descent approach is used to learn the linear compression and decompression pair $\{W, R\}$ that meet the requirement in (4). Let $\bar{X}$ denote the collection of all training patches. We find the optimal linear compression and decompression matrix pair $\{W, R\}$ by minimizing the reconstruction error:

$$[W, R] = \arg\min_{W,R} \|\bar{X} - RW\bar{X}\|_F^2 \quad (11)$$

Eq. (11) is solved by alternatively minimizing with respect to W and R, with initialization $W^0 = \text{blkdiag}(W_d, W_d, W_d)$. Here, $W_d \in \mathcal{R}^{d\times K}$ denotes the frequency weighting matrix that maps the K linear spectrum frequency bins to d mel-frequency bins. This minimization is guaranteed to converge to a stationary point [17, Proposition 2.7.1].

Informal listening and formal objective evaluations indicated no perceptually noticeable difference in both the quality and intelligibility between the original signal ($\bar{X}$) and reconstructed signal ($RW\bar{X}$) after compression and decompression. This suggests acceptable perceptual loss due to compression, and justifies learning the dictionary in the compressed domain.

2.4. Sparse Dictionary Learning Algorithm

Given the compression matrix W and the training speech patches $\{\bar{x}_n\}$, we use Eq. (12) to learn the dictionary that meets the sparse dictionary learning assumption (3).

$$\begin{aligned} [D, G] &= \arg\min_{D,G} \sum_{n=1}^{N} \frac{1}{2}\|W\bar{x}_n - Dg_n\|^2 + \lambda_d\|g_n\|_1 \\ &\quad \text{s.t.} \quad \|d_m\| \le 1, \forall m = 1, \cdots, M \end{aligned} \quad (12)$$

We seek a dictionary that well represents the data, by minimizing the square error between $W\bar{x}_n$ and Dg_n. We also encourage sparse representation by penalizing the L_1 norm of g_n. To avoid scaling ambiguities, we constrain the norm of columns of D to be at most 1. We customize the Block Successive Upper-Bound Minimization (BSUM) algorithm [18] to solve Eq. (12). BSUM supports computationally efficient closed form updates, and is guaranteed to converge to the set of stationary points[18, Theorem 2b].

To demonstrate that the learned dictionary D indeed captures the temporal dynamics of speech, we construct an experiment to compare the performance of a specific dictionary in representing different speech transitions. We first collect all speech patches that correspond to the phoneme transition ($/iy/ \rightarrow /r/$) into $\bar{X}_{/iy/\rightarrow/r/}$. We can define $\bar{X}_{/iy/\rightarrow/aa/}$ and $\bar{X}_{/iy/\rightarrow/ae/}$ in a similar fashion. These three speech transitions have different temporal dynamics because the second phoneme is different. For example, as shown in Figure 1, $/r/$ typically demonstrates a low third formant (F_3), while $/aa/$ and $/ae/$ do not. Furthermore, $/r/$ is often characterized by lower energy compared to $/aa/$ and $/ae/$. These characteristic properties of $/r/$ impact the spectro-temporal energy transition from $/iy/$, making the dynamics of /iy/ significantly different from the other two cases. We now train a dictionary that can sparsely represent the ($/iy/ \rightarrow /r/$) transition, by using only data from $\bar{X}_{/iy/\rightarrow/r/}$ to train the dictionary. We finally use this dictionary $D_{/iy/\rightarrow/r/}$ to approximate all three speech transitions that correspond to different temporal dynamics, and calculate the representation error defined as Eq. (13):

$$\text{Error}\ (\bar{x}_n) = \min_{g_n} \frac{\|W\bar{x}_n - D_{/iy/\rightarrow/r/}g_n\|^2}{\|W\bar{x}_n\|^2} \quad (13)$$

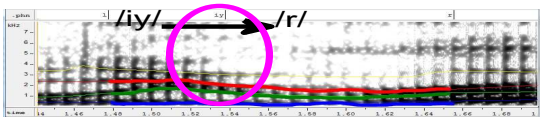
where $\bar{x}_n$ denotes the speech patch that corresponds to any of the three temporal dynamics. A histogram of the approximation errors for the three speech transitions is shown in Figure 2. Figure 2 shows that using $D_{/iy/\rightarrow/r/}$ to approximate $\bar{X}_{/iy/\rightarrow/aa/}$ and $\bar{X}_{/iy/\rightarrow/ae/}$ results in a much higher error than for $\bar{X}_{/iy/\rightarrow/r/}$. This supports that the dictionary $D_{/iy/\rightarrow/r/}$ has specifically captured the temporal dynamics of ($/iy/ \rightarrow /r/$), but not that of the other transitions.

3. Performance Evaluation

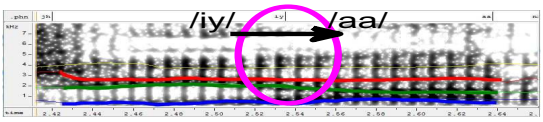
The quality and intelligibility of the speech processed by STSA-MSDL was compared with that of unprocessed speech and speech processed by STSA [2], SWDL [13] and Wiener filter.

3.1. Experiment Setup

The TIMIT database [19] was used for evaluation due to the availability of a large database for dictionary training. One



(a) $/iy/ \rightarrow /r/$



(b) $/iy/ \rightarrow /aa/$

Figure 1: Temporal formant patterns in different phoneme transitions. The first, second, and the third formant are drawn in bold blue, green, and red respectively.

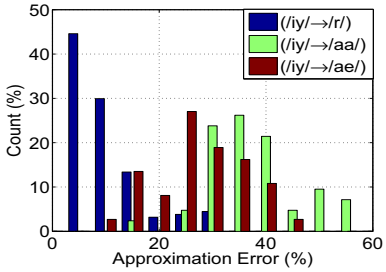


Figure 2: Histogram of representation error for using $D_{/iy/\rightarrow/r/}$ to represent $\bar{X}_{/iy/\rightarrow/r/}$, $\bar{X}_{/iy/\rightarrow/aa/}$, and $\bar{X}_{/iy/\rightarrow/r/}$

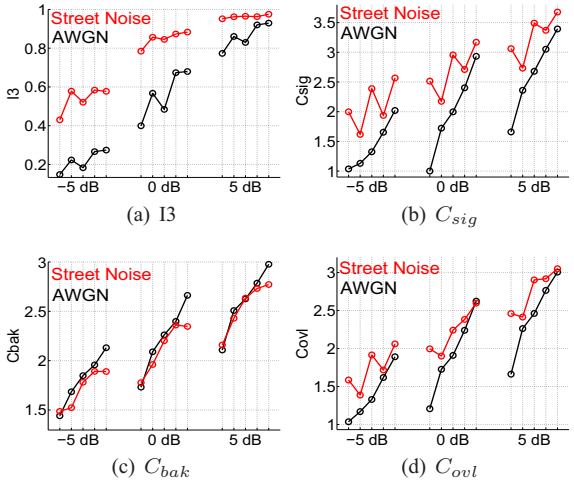
universal dictionary for both genders was trained using six hours of speech selected randomly form the "train" subset. For enhancement, 320 male sentences and 160 female sentences were selected randomly from the "test" subset. Computed-generated additive Gaussian noise (AWGN) and a real world noise (Street) 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 [20], 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 4 algorithms (Wiener filter, STSA, SWDL, STSA-MSDL), and the noise variance was assumed to be known for all algorithms.

All sentences were sampled at 8 kHz, and segmented into 30-ms duration frames using a Hamming window with 50% overlap (thus, a contextual window of 90 msec). A 512 point FFT/IFFT was used for the time-frequency analysis and synthesis operations (thus, $K = 257$). This choice of parameters were motivated by the values proposed in [13]. For the pilot study we report in this paper, the reduced dimensionality ($3d$) was set to 120, while the original patch dimension ($3K$) is 771. The dictionary size (M) was fixed to 240, and the sparsity parameter (λ_d) was set to 0.3. The best sparsity parameters (λ_s, λ_g) in Eq. (5) were found using a grid search on 16 randomly selected sentences from the train subset, and were retained for the enhancement phase.

Four perceptually oriented objective metrics were used to measure four different aspects of the enhancement performance. The intelligibility of the processed speech was measured using the Intelligibility Index [21] (I3, ranging from 0 to 1). The other three metrics were C_{sig}, C_{bak}, and C_{ovl} [22] (all ranging from 0 to 5) that measure the speech signal distortion, background noise reduction, and the overall speech quality respectively. For all metrics, lower values indicate worse performance and higher values indicate better performance along that metric. These objective metrics were chosen because of their high correlation with perceptual responses (see [21, 22] for details).

3.2. Objective Evaluation Results and Discussions

Figure 3 compares the performance of various algorithms for different noise types at different SNRs. The five circles on each SNR in each figure denote the average metric value of, from left to right, Unprocessed speech, Wiener filter, STSA, SWDL, and STSA-MSDL. Under AWGN, both the dictionary-based algorithms (SWDL, STSA-MSDL) outperform the traditional statistical estimators (Wiener, STSA) in all of the four objective metrics; this is because both the SWDL and STSA-MSDL exploit the speech-specific information by leveraging a dictionary. Furthermore, STSA-MSDL achieves a noticeable improvement over SWDL in signal distortion, background noise reduction, and overall quality while not sacrificing the intelligibility. This superior performance of the STSA-MSDL can be explained by the exploitation of speech temporal dynamics in a longer contextual window using multi-frame sparse DL, as opposed to SWDL where only a single frame spectrum information is exploited. Under street noise, the relative improvement of STSA-MSDL over SWDL, Wiener and STSA is not as large as in the AWGN case. This may suggest that the performance of STSA-MSDL depends on the noise variance distribution over frequency. The performance of the proposed algorithm under other real-world noises will be evaluated and analyzed in the future.



(a) I3 (b) C_{sig}

(c) C_{bak} (d) C_{ovl}

Figure 3: Performance at different SNRs under AWGN and Street noise. The five circles in each SNR denote average metric of, from left to right, Unprocessed speech, Wiener filter, STSA, SWDL, and STSA-MSDL. The result is averaged over all 480 test sentences.

4. Conclusions and Future Work

In this paper, we present a novel approach for single-channel speech enhancement (STSA-MSDL) that combines the STSA estimator with multi-frame sparse DL. Due to the exploitation of speech temporal dynamics, STSA-MSDL achieves a superior performance over traditional statistical algorithms, as evidenced by objective evaluation in our preliminary study. In the future, we plan to extend the STSA-MSDL from a batch algorithm to an online algorithm aiming for real-time processing. To better capture the temporal dynamics of individual speakers, we also plan to develop online updates to the dictionary. Another interesting future direction is to use multi-frame sparse dictionary to also capture the temporal dynamics of highly non-stationary noise, and thus improve enhancement in difficult non-stationary scenarios like a cocktail party.

5. References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] K. 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.

[7] M. 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.

[8] 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.

[9] 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.

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

[11] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec.

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

[13] H.-W. Tseng, S. Vishnubhotla, M. Hong, J. Xiao, Z.-Q. Luo, and T. Zhang, "A novel single channel speech enhancement approach by combining wiener filter and dictionary learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*, 2013.

[14] 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.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, Oct. 2007. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0387310738

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1561/2200000016

[17] D. P. Bertsekas, "Nonlinear programming," 1999.

[18] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *arXiv preprint arXiv:1209.2385*, 2012.

[19] 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. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

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

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

[22] Y. Hu and P. 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.