Applied Acoustics 138 (2018) 72–79

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Hyper-parameterization of sparse reconstruction for speech enhancement

Yue Shi[a], Siow Yong Low[b], Ka Fai Cedric Yiu[a,*]

[a] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hunghom, Hong Kong, China
[b] School of Electronics and Computer Science, University of Southampton, Malaysia Campus, Iskandar Puteri, Johor, Malaysia

ARTICLE INFO

Keywords:
Speech enhancement
Compressed sensing
Regularized least squares

ABSTRACT

The regularized least squares for sparse reconstruction is gaining popularity as it has the ability to reconstruct speech signal from a noisy observation. The reconstruction relies on the sparsity of speech, which provides the demarcation from noise. However, there is no measure incorporated in the sparse reconstruction to optimize on the overall speech quality. This paper proposes a two-level optimization strategy to incorporate the quality design attributes in the sparse solution in compressive speech enhancement by hyper-parameterizing the tuning parameter. The first level involves the compression of the big data and the second level optimizes the tuning parameter by using different optimization criteria (such as Gini index, the Akaike information criterion (AIC) and Bayesian information criterion (BIC)). The set of solutions can then be measured against the desired design attributes to achieve the best trade-off between suppression and distortion. Numerical results show the proposed approach can effectively fuse the trade-offs in the solutions for different noise profile in a wide range of signal to noise ratios (SNR).

1. Introduction

The ever growing demand for mobile electronic devices, e.g., smart phones, has made voice interfaces ubiquitous. Given the mobility of these electronic devices, the input speech signal will suffer from the various environmental noise. Clearly, delivering a clean speech signal in the communication system is an important aspect of the product requirement. The objective of speech enhancement is to estimate the desired speech signal from the noisy observation, which consists of both speech and noise signals [1,2]. The two key performance measures for speech enhancement are usually measured in terms of noise suppression and speech distortion [3,4]. Interestingly, these two measures can be viewed as engineering design and quality design requirements, respectively [5–7]. In terms of engineering design, the enhancement must yield the highest signal to noise ratio (SNR) possible, which translates to noise suppression capability. In order to satisfy its quality design, the enhancement process must also maintain the perceptual features, i.e., minimizes speech quality degradation. Indeed, it is a challenge to optimize the overall noisy speech as the engineering and quality requirements are at times conflicting as maximizing SNR tend to result in speech degradation, resulting in a natural trade-off [8].

Given its volume, speech signal is considered to be a big data. Additionally, speech is highly non-stationarity across the time and frequency domains. The varying nature of speech adds to the challenge as the data is not just 'big' but also changing as a function of time and frequency. There is a wealth of literature examining the characteristics of speech to reveal its patterns and trends, which are useful in application such as speech recognition, speech enhancement and computational auditory scene analysis. Of late, one important characteristics of speech is its sparsity. Speech sparsity has gained popularity as it may hold the key to making the 'big' speech data, 'small'. Whilst speech is fairly compact and dense in the time domain, speech signals are in fact sparse in the time-frequency representations [9,10]. This is because speech is highly non-stationary and there will be lapses of time-frequency periods where the speech power is negligible compared to the average power [11]. On average, a speech signal consists of approximately ten to fifteen phonemes per second and each of these phonemes has a varying spectral rate [12].

The notion of sparsity has led to sparse reconstruction methods such as compressed sensing (CS) [13,14]. CS theory states that sparse signals with a small set of linear measurements can be reconstructed with an overwhelming probability [15,16]. Potentially, CS has the capability to compress big data such as speech signal. In speech enhancement, CS exploits the sparsity of speech and non-sparse nature of environmental noise in its reconstruction. Low et al. [17] demonstrated the use of CS as a speech enhancer by relying upon the strength of CS to maintain only the sparse components (speech) and its weakness in preserving the non-sparse components (noise). Various CS based methods with favorable results have been reported [17–19], demonstrating its efficacy for speech enhancement applications. A very popular technique for sparse

* Corresponding author.
E-mail addresses: yue.shi@connect.polyu.hk (Y. Shi), sy.low@soton.ac.uk (S.Y. Low), cedric.yiu@polyu.edu.hk (K.F. Cedric Yiu).

https://doi.org/10.1016/j.apacoust.2018.03.020
Received 30 September 2017; Received in revised form 15 March 2018; Accepted 19 March 2018

signal reconstruction is the regularized ℓ_1-norm least squares [20]. This is because ℓ_1 regularized least squares yields a sparser solution since the solution tends to have a fewer nonzero coefficients compared to the ℓ_2 based Tikhonov regularization [20]. One important parameter in solving the regularized sparse solution is the tuning parameter or the penalty constant, λ. The regularization parameter, λ holds significance as a heavier weighting would penalize the Tikhonov regularization. In other words, the tuning parameter holds the key in determining how sparse a solution is reconstructed.

Whilst a sparse solution indicates the existence of a sparse component such as speech, there is no measure incorporated in the CS reconstruction to optimize on the overall speech quality. The idea is to establish the relationships between sparsity and quality. Since the tuning parameter has influence over the sparsity of the solution, then a quality measure should be factored into link the two. More specifically, this paper sets out to find the tuning parameter that best suits the sparsity profile of the corresponding frequency data in question. This paper proposes to formulate the solution in compressive speech enhancement by hyper-parameterizing the tuning parameter.

For the sparsity model to hold for sparse reconstruction, the data is decomposed in the frequency domain. As mentioned, the focus here is to ascertain if properly optimized tuning parameter would increase the overall PESQ. Since the PESQ is formulated in fullband, each combination of the tuning parameter in each frequency point would need to be computed and then reconstructed into fullband representation for PESQ evaluation. Thus, optimizing $\lambda(\omega)$ directly based on PESQ would be computationally prohibitive as the number of combinations would be to the order of the number of frequency points. To bypass that, the tuning parameter is then optimized in each frequency bin by using a different optimization criterion (such as Gini index, the Akaike information criterion (AIC) and Bayesian information criterion (BIC)) to achieve the sparsest set of solutions. The set of sparsest solutions is then evaluated against the perceptual evaluation speech quality (PESQ) improvement as a quality measure for speech [21]. Experimental results show that both the Gini index and the model selectors help to select the tuning parameters, which improve the PESQ, thus directly parameterizing the performance of compressive speech enhancement with the tuning parameter.

2. Signal model

Let the noisy signal be

$$x(n) = s(n) + v(n) \tag{1}$$

where $s(n)$ and $v(n)$ are the speech and noise signals, respectively. Its corresponding L-point STFT is given as

$$X(\omega,k) = \sum_{n=0}^{L-1} x(n)w(n-kR)e^{-j\omega n} = S(\omega,k) + V(\omega,k) \tag{2}$$

where $w(n-kR)$ is a time-limited window function with a hop size of R and length L,$\omega \in \omega_0,\ldots,\omega_{L-1}$ and k is the time index. The k-th instant data envelope of (2) is $|X(\omega,k)|$, where $|\cdot|$ denotes the absolute value operator.

Consider a $N \times N$ matrix $\mathbf{\Psi}$ whose columns form an orthonormal basis. The K-sparse signal, $\mathbf{x}(\omega,k) \in \mathbb{R}^N$ can then be given as

$$\mathbf{x}(\omega,k) = \mathbf{\Psi}(\omega)\theta(\omega,k) \tag{3}$$

where the N-length envelope vector $\mathbf{x}(\omega,k) = [|X(\omega,k)|,|X(\omega,k-1)|,\cdots,|X(\omega,k-N+2)|,|X(\omega,k-N+1)|]^T$, the symbol $[\cdot]^T$ is the transposition operator and $\theta(\omega,k) \in \mathbb{R}^N$ has K non-zero entries. The compressed measurement vector is given as

$$\mathbf{y}(\omega,k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k) \tag{4}$$

where $\mathbf{\Phi}(\omega)$ is a $M \times N$ sensing matrix/linear mapping matrix. In this instant, the sensing matrix compresses the signal's envelope for each frequency ω. Since $M \ll N$, this means that the dimension of $\mathbf{y}(\omega,k)$ is considerably smaller than $\mathbf{x}(\omega,k)$, hence the term "compressed". Eq. (4) represents an alternative sampling procedure, which samples sparse signals close to their intrinsic information rate rather than their Nyquist rate. It has been shown that the tractable recovery of K-sparse signal, $\mathbf{x}(\omega,k)$ from the measurements, $\mathbf{y}(\omega,k)$ requires the sensing matrix, $\mathbf{\Phi}(\omega)$ to obey the restricted isometry property (RIP) [16]. Here, a sensing matrix, $\mathbf{\Phi}(\omega)$ is said to satisfy RIP of order K for all K-sparse signal, $\mathbf{x}(\omega,k)$, if there exists a constant, $\delta_K \in (0,1)$ such that

$$(1-\delta_K)\|\mathbf{x}(\omega,k)\|_2^2 \leqslant \|\mathbf{\Phi}(\omega)\mathbf{x}(\omega,k)\|_2^2 \leqslant (1-\delta_K)\|\mathbf{x}(\omega,k)\|_2^2 \tag{5}$$

where $\|\cdot\|_2$ denotes ℓ_2 norm.

3. CS recovery

One solution to ensure sparse recovery is to solve the following:

$$\hat{\mathbf{x}}(\omega,k) = \arg\min_{\mathbf{x}(\omega,k)} \|\mathbf{x}(\omega,k)\|_0 \quad \text{s. t.} \quad \mathbf{y}(\omega,k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k) \tag{6}$$

where $\|\mathbf{x}(\omega,k)\|_0$ is the number of non-zero components of $\mathbf{x}(\omega,k)$. However, solving (6) requires a combinatorial search, which is NP-hard [22]. A computational tractable solution to (6) is the widely known basis pursuit method as follows

$$\hat{\mathbf{x}}(\omega,k) = \arg\min_{\mathbf{x}(\omega,k)} \|\mathbf{x}(\omega,k)\|_1 \quad \text{s. t.} \quad \mathbf{y}(\omega,k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k) \tag{7}$$

where $\|\cdot\|_1$ is the ℓ_1 norm. Whilst the basis pursuit is a weaker formulation compared to (6), it allows efficient solution via linear programming techniques [22,20]. A more flexible formulation, which allows for a trade-off between the exact congruence of $\mathbf{y}(\omega,k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k)$ and a sparser $\mathbf{x}(\omega,k)$ is the popular basis pursuit denoising [20] given as

$$\hat{\mathbf{x}}(\omega,k) = \arg\min_{\mathbf{x}(\omega,k)} \|\mathbf{y}(\omega,k)-\mathbf{\Phi}(\omega)\mathbf{x}(\omega,k)\|_2^2 + \lambda(\omega)\|\mathbf{x}(\omega,k)\|_1 \tag{8}$$

where $\|\cdot\|_2$ is the $L2$-norm and $\lambda(\omega)$ is the regularization parameter. The formulation in (6) is a simple least-squares minimization process with a $L1$-norm penalizer and the dictionary matrix $\mathbf{\Phi}(\omega)$. It is worth noting that since $L1$-norm is non-differentiable, the optimization then leads to a decomposition which is sparser [23]. Simply, the first term in Eq. (8) is to reduce the mean square area whilst the regulator seeks a sparser solution.

Note that the optimal solution tends to trivial as $\lambda(\omega) \to \infty$ [20]. A higher value of $\lambda(\omega)$ would generally result in a sparser solution since the ℓ_1-norm is being penalized more heavily. This means that the regularizer, $\lambda(\omega)$, penalizes the sum of the observed signal. In other words, the solution to (8) is indeed a function of $\lambda(\omega)$, i.e., fixing $\lambda(\omega)$ is equivalent to setting it to a particular subset of sparse solution for the least squares to be performed on [24]. Simply, the optimization problem is a trade-off between a quadratic misfit error (mean square error) against the sparsity of the data, i.e., ℓ_1-norm [25]. Clearly, if the incoming signal is already sparse, then $\lambda(\omega)$ can be relaxed and vice versa. Since the sparsity of the signal varies as a function of frequency, the regularizer should ideally vary according to the signal's profile.

A good choice of $\lambda(\omega)$ should provide a reasonable trade-off between the smoothness of the reconstructed signal and similarity to the original signal [17]. Nevertheless, it remains not so straightforward to set the regularization parameter $\lambda(\omega)$ and thus far, $\lambda(\omega)$ has been empirically determined. In practice, $\lambda(\omega)$, should be set according to the sparsity of the actual signal as $\lambda(\omega)$ controls the amount of regularization that can be imposed. It is precisely this quality control that this paper seeks to establish, i.e., by linking sparsity to quality. Since a larger value of $\lambda(\omega)$ yields a sparser solution, then more noise would be suppressed. However, how much can $\lambda(\omega)$ be set before the signal quality is compromised.