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.

4. Quality measures

4.1. Background

In a big data setting such as speech signals, this paper seeks to subsume the affective design by hyper-parameterizing λ via the Gini index and the model selectors, Akaike information criterion (AIC) and Bayesian information criterion (BIC). The set of solutions is then evaluated with respect to PESQ. In particular, λ is to be optimized in such a way that the sparsest solution yields the one with the best quality in terms of noise suppression and target distortion. In this case, the noise suppression and speech distortion can be viewed as the engineering requirement and the affective design attribute, respectively. The idea is to incorporate affective design via the influence of the key design parameter on the aforementioned PESQ measure. By doing so, the parameter can be translated to consumer reactions (via the PESQ measure).

We propose a two-level optimization strategy to optimize $\lambda(\omega)$ to affective measure. In the inner level, the big data is first compressed via the sensing matrix, $\mathbf{\Phi}(\omega)$. In the outer level or the sparse reconstruction stage, the hyperparameter is optimally chosen to incorporate the overall signal quality. Quality measures such as the AIC, BIC and Gini index are used to optimize the value of the hyperparameter. These measures are explicitly used to determine the relationship between key design parameters with the consumer reactions from the processed signal. The following sections explain each of the chosen optimization criteria, namely Gini index, the AIC and BIC model selection methods.

4.2. Gini index

4.2.1. The Gini coefficients – a sparsity measure

As mentioned, the actual sparsity of the signal affects the performance sparse recovery. As an effective measure of sparsity, Zonoobi [27] concluded that the Gini index can induce a significantly improved performance in reconstruction from compressive samples. A signal is considered most sparse if a signal can be represented by only one non-zero coefficient with the rest being zero [26]. Similarly, if a signal has only one high value non-zero coefficient amidst a low non-zero coefficients, then the signal can be said to be most sparse. In essence, sparsity is a measure of disparity, i.e., the relative distribution of the coefficients of a signal is. A non-sparse signal on the other hand is described as having a uniform non-zero coefficients throughout. Of the many sparsity measures, it has been shown that Gini index remains the most consistent and fulfil all of the desirable sparsity criteria [26,27].

Consider a M long ordered vector, $\mathbf{w} = [w_1,...,w_M]$ such that $w_M \geqslant w_{M-1},\cdots,w_2 \geqslant w_1$, then the Gini coefficient is defined as

$$\mathrm{GI}(\mathbf{w}) = 1-2\sum_{m=1}^{M}\frac{w_m}{\|\mathbf{w}\|_1}\left(\frac{M-m+0.5}{M}\right). \tag{9}$$

A zero-valued Gini represents perfect equality whilst a close to unity value shows the opposite. In sparsity terms, a larger Gini coefficient shows a sparser signal. As such, Gini coefficient can be used as a measure to ascertain if a signal is sparse. Table 1 tabulates the Gini coefficients for three types of noise, speech and the noisy speech at different SNR levels. The coefficients show speech indeed is the sparsest signal in comparison with the other noise signals. Note that, of all the noise signals, babble noise has the highest Gini coefficient, owing to its speech-like nature. For the case of noisy speech signals, it can be seen that as the SNR increases, the Gini coefficient approaches unity. As the SNR decreases, the value of the Gini coefficient drops accordingly. This simple example demonstrates that a sparser signal tends to have a higher SNR and as the signal becomes more noisy, sparsity reduces. Fig. 1 shows that the sparsity of different speech signal varies as a function of frequency. It can be seen that the level of sparsity not only varies in the time–frequency domain but also changes with different speaker.

As speech is highly non-stationary across time and frequency, its sparsity level would vary accordingly. Clearly, different speech signal would have a varying Gini index as speech profile changes. From Fig. 1, the Gini index for the speech signal varies as a function of frequency, thus λ will need to be re-estimated every N samples. By properly optimizing $\lambda(\omega)$ based on the Gini coefficient, the sparse reconstruction could potentially lead to better SNR improvement, as appropriate tuning parameter can be set according to the sparsity of the signal in question.

4.2.2. Selection of $\lambda(\omega)$ based on Gini

Consider an N-length signal, $\mathbf{x}(\omega,k)$, then from Eq. (8), its sparse reconstruction is given as

$$\hat{\mathbf{x}}(\omega,k) = \arg\min_{\mathbf{x}(\omega,k)} \|\mathbf{y}(\omega,k)-\mathbf{\Phi}(\omega)\mathbf{x}(\omega,k)\|_2^2 + \lambda(\omega)\|\mathbf{x}(\omega,k)\|_1. \tag{10}$$

For each given value of $\lambda(\omega)$ value, an estimation of $\hat{\mathbf{x}}(\omega,k)$ is denoted as $\hat{\mathbf{x}}_{\lambda(\omega)}(\omega,k)$. The Gini coefficient of $\hat{\mathbf{x}}_{\lambda(\omega)}(\omega,k)$ is then defined as

$$\mathrm{GI}(\hat{\mathbf{x}}_{\lambda}(\omega,k)) = 1-2\sum_{n=1}^{N}\frac{\hat{x}_{\lambda(\omega)}(\omega,k,n)}{\|\hat{\mathbf{x}}_{\lambda(\omega)}(\omega,k)\|_1}\left(\frac{N-n+0.5}{N}\right). \tag{11}$$

where $\hat{x}_{\lambda(\omega)}(\omega,k,n)$ is the n-th ordered value of vector $\hat{\mathbf{x}}_{\lambda(\omega)}(\omega,k)$ in a descending order. The corresponding optimization problem of maximizing the GI coefficients can be written as

$$\lambda_{\mathrm{maxGini}}(\omega) = \arg\max_{\lambda(\omega)} \mathrm{GI}(\hat{\mathbf{x}}_{\lambda(\omega)}(\omega,k)) \tag{12}$$

where $\mathrm{GI}(\hat{\mathbf{x}}_{\lambda(\omega)}(\omega,k))$ is given in Eq. (11). Equivalently, the optimization formulation for finding $\lambda(\omega)$ for the minimum Gini index is

$$\lambda_{\mathrm{minGini}}(\omega) = \arg\min_{\lambda(\omega)} \mathrm{GI}(\hat{\mathbf{x}}_{\lambda(\omega)}(\omega,k)). \tag{13}$$

Eqs. (12) and (13) can be viewed as the extreme ends of compressive speech enhancement, as Eq. (12) recovers the sparsest signal it could possibly tuned and vice versa for Eq. (13). In the numerical experiments to follow, we will show that both the optimization above behaves very differently for the PESQ and segmental SNR measures, with Eq. (12) leaning towards noise suppression and Eq. (13) acting towards more on speech preservation.

4.3. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Whilst the tuning parameter selection based on Gini criterion is intuitive, it is by no means the only approach. For any regularization method, finding the best regularization parameter is essential. As explained by Dicker et al. [28], the estimators are typically found to correspond to a range of tuning parameter values, which is referred to

Table 1
The Gini coefficients for speech and different types of noise and at different SNRs.

Signal	Gini coefficient	SNR	Speech + Babble	Speech + White	Speech + Destroyerops
Speech	0.9266	0	0.7522	0.7382	0.7372
Babble	0.6634	5	0.8302	0.8234	0.8239
White	0.6352	10	0.8848	0.8823	0.8828
Destroyerops	0.6243	15	0.9108	0.9099	0.9104

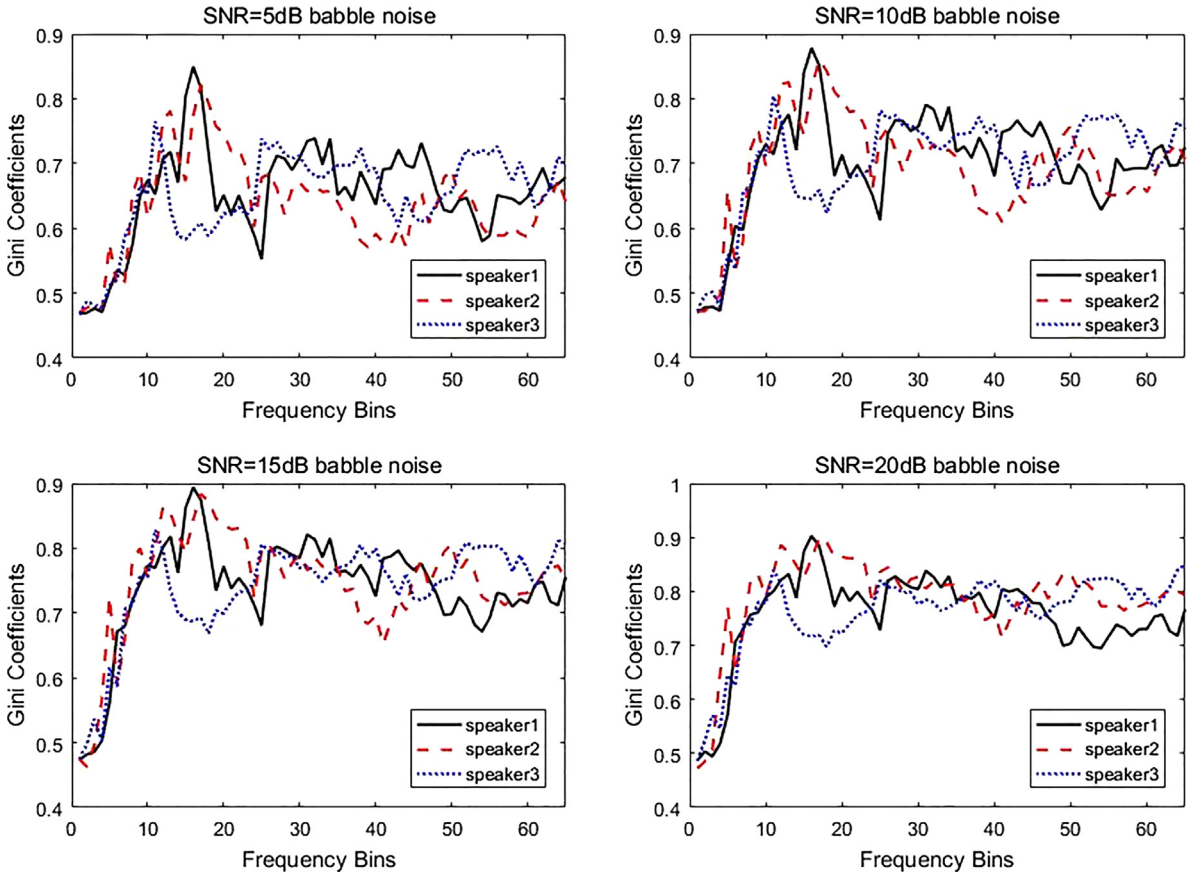

Fig. 1. The evolution of the Gini coefficients with different speakers and varying SNRs.

as a solution path. Subsequently, the preferred estimator is identified along the solution path as the estimator, which fits the optimization criteria. In the same vein, this paper considers the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) based approach for the selection of the tuning parameter, $\lambda(\omega)$ [29]. It is well known that AIC and BIC are popular model selection criteria. As shown in Zou [30], AIC and BIC possess different asymptotic optimality. AIC converges at the minimax optimal rate to the true regression mode, whereas BIC is consistent in selecting the true model. In this case, we ascertain the heuristics usefulness of both the AIC and BIC in tuning $\lambda(\omega)$, for compressive speech enhancement. The major difference between AIC and BIC is that they possess different asymptotic optimality [30]. For AIC (Akaike, 1973), it seeks the model with the least average squared error irrespective of whether the true model is in the candidate list. BIC, on the other hand, guarantees in selecting the true model, should the true model be selectable. Readers may refer to [28,30,31] for in-depth view of the two approaches.

Let us define the residual sum of squares (RSS) as

$$\mathrm{RSS} = \|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k)\|_2^2. \tag{14}$$

From [30], given an estimator $\hat{\mathbf{x}}$, the number of nonzero entries of an estimator $\hat{\boldsymbol{x}}$ is an unbiased estimate of the degree of freedom (df), that is

$$\mathrm{df} = \text{number of nonzero entries of } \hat{\mathbf{x}}(\omega,k). \tag{15}$$

AIC and BIC are usually used to make model selection and predict models, both of them could be represented as a combination of a likelihood term and a penalty term. Thus from Eqs. (14) and (15), the corresponding AIC and BIC can be formulated as

$$\mathrm{AIC} = \ell\log(\mathrm{RSS}/\ell) + 2\mathrm{df}, \tag{16}$$

$$\mathrm{BIC} = \ell\log(\mathrm{RSS}/\ell) + \mathrm{df}\cdot\log(\ell), \tag{17}$$

where ℓ is the length of estimator $\hat{\mathbf{x}}$. The tuning parameter selection procedure can be reduced to the minimization of AIC or BIC, and as discussed previously, AIC is comparatively more conservative in its variable selection. Inserting Eq. (14) into (16) and (17), respectively, yields the $\lambda(\boldsymbol{\omega})$ selection as follows:

$$\lambda_{\mathrm{AIC}}(\omega) = \underset{\lambda(\omega)}{\operatorname{argmin}}\, n\log(\|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_\lambda(\omega,k)\|_2^2/n) + 2\mathrm{df}, \tag{18}$$

$$\lambda_{\mathrm{BIC}}(\omega) = \underset{\lambda(\omega)}{\operatorname{argmin}}\, n\log(\|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_\lambda(\omega,k)\|_2^2/n) + \mathrm{df}\log(n). \tag{19}$$

5. Perceptual evaluation of speech quality

Broadly, the assessment of speech quality can be classified as subjective and objective evaluation. As the name implies, subjective evaluation involves subjective listening test by some listeners. Objective evaluation on the hand, measures the numerical distance between the reference signal and the processed signals [32]. One established method of evaluating how good the enhancement process is via the use perceptual evaluation of speech quality (PESQ). PESQ is an automated computation algorithm developed by the International Telecommunications Union (ITU) to replace human subjects in the evaluation of the mean opinion score (MOS). The PESQ model considers how human perceive speech and it has been used widely in the evaluation of speech quality [33]. PESQ is defined mathematically as [34]

$$\mathrm{PESQ} = a_0 + a_1 d_{sym} + a_2 d_{asym} \tag{20}$$

where $a_0 = 4.5, a_1 = -0.1$ and $a_2 = -0.0309$. The variables d_{sym} and d_{asym} are the average disturbance values for the symmetrical and asymmetrical components. The former measures the distortion due to noise and the latter describes the omission of the actual speech.

PESQ bypasses the need for human subjects to take part in the

evaluation process and can be used as part of the affective design process. Numerous studies have shown that PESQ consistently rated to be the most reliable objective measure for speech quality assessment [35,36]. In fact, PESQ has also been shown to be consistent in measuring speech intelligibility [37]. As PESQ gives the overall speech quality score, consequently, it is regarded as an affective indicator as to how 'pleased' the consumers are with the processed speech.

6. Proposed two-level optimization process

This section details the proposed two-level optimization strategy to optimize $\lambda(\omega)$ with respect to the quality measures. In the first level optimization, the big data is first compressed via the sensing matrix, $\mathbf{\Phi}(\omega)$. The second level then optimizes the hyperparameter through the quality measures, which then improves the overall signal affective's quality.

6.1. First level optimization: compressive Sensing

The first step entails the compressive sensing matrix selection. The data compression from Eq. (4) is reproduced here for convenience

$$\mathbf{y}(\omega,k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k) \tag{21}$$

where $\mathbf{y}(\omega,k) \in \mathbb{R}^M$,$\mathbf{x}(\omega,k) \in \mathbb{R}^N$, and $\mathbf{\Phi}(\omega) \in \mathbb{R}^{M\times N}$ is the compressive sensing matrix, which compresses the signal dimension by projecting the signal from $\mathbb{R}^N$ into $\mathbb{R}^M$, where $M \ll N$. The sensing matrix is typically generated by using a random Gaussian matrix or a partial DCT matrix [17].

Under the Restricted Isometry Property condition (5), the solution to (21) can be solved by using the popular basis pursuit as follows [20]

$$\hat{\mathbf{x}}(\omega,k) = \arg\min_{\mathbf{x}(\omega,k)} \|\mathbf{x}(\omega,k)\|_1 \text{s. t. } \mathbf{y}(\omega,k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k). \tag{22}$$

Alternatively, Eq. (22) can be viewed as a linear regression

$$\mathbf{y}(\omega,k) = \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k) + \varepsilon, \text{s. t. } \|\mathbf{x}(\omega,k)\|_1 \leqslant \nu, \tag{23}$$

where ν is a constant relating to the sparsity constraint and $\varepsilon \in \mathbb{R}^M$ is the intercept or error. Thus Eq. (22) can be reposed as the following

$$\min_{\mathbf{x}(\omega,k)} \|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\mathbf{x}(\omega,k)\|_2^2 + \lambda(\omega)\|\mathbf{x}(\omega,k)\|_1 \tag{24}$$

where $\lambda(\omega)$ is the tuning hyperparameter. The solution to Eq. (24) is the key to finding the best affective solution to the problem in question. Here, the $\lambda(\omega)$ plays a key role in mapping the solution to the affective measures. The following section explains how the solution to (24) is optimized with respect to the affective measures as discussed in the previous section.

6.2. Second level optimization: hyperparameter selection

To solve model (24), we implement the interior point method for large-scale l_1 regularized least squares algorithm in [20] with the following properties:

(i) When $\lambda(\omega) \to 0$, the estimator has the limiting behavior with (24), satisfying $\mathbf{\Phi}(\omega)^T[\mathbf{\Phi}(\omega)\mathbf{x}(\omega,k) - \mathbf{y}(\omega,k)] = 0$.

(ii) As $\lambda(\omega) \to \infty$, the estimator shrinks to the zero vector, $\mathbf{0}$. The convergence occurs for a finite value of $\lambda(\omega)$, i.e., $\lambda(\omega) \geqslant \lambda_{max}(\omega) = \|2\mathbf{\Phi}(\omega)^T\mathbf{y}(\omega,k)\|_\infty$, where $\|\mathbf{x}\|_\infty = \max_i|x_i|$ is the l_∞ norm of vector $\mathbf{x}$. However, for $\lambda(\omega) > \lambda_{max}(\omega)$, the optimal solution of (24) is trivial, i.e., $\mathbf{0}$.

(iii) As λ varies across $(0,\infty)$, the solution path of $\mathbf{x}$ is piecewise linear. That is, with tuning parameters satisfy $0 = \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_k = \lambda_{max}$, the regularization path of $\mathbf{x}$ is a piecewise linear curve on $\mathbb{R}^N$:

$$\mathbf{x} = \frac{\lambda_{i+1}-\lambda}{\lambda_{i+1}-\lambda_i}\mathbf{x}^{(i)} + \frac{\lambda-\lambda_i}{\lambda_{i+1}-\lambda_i}\mathbf{x}^{(i)}, \lambda_i \leqslant \lambda \leqslant \lambda_{i+1}, i = 1,2,\ldots,k-1.$$

(iv) Clearly as a general rule, with properly chosen $\lambda(\omega)$, Eq. (24) will result in a sparser solution.

(v) The Computational complexity of this algorithm is determined by the product of the total number of Preconditioned Conjugate Gradient (PCG) steps during all iterations and the cost of a PCG step. As noted in [20], extensive testing suggest that the total number of PCG steps vary from a few tens to several hundreds to compute a solution. The computational complexity of a PCG step is $\mathcal{O}(NM)$, where M,N are the dimensions of sensing matrix $\mathbf{\Phi}(\omega)$. Then the total computational complexity is at most $\mathcal{O}(cNM)$, where c is the number of iterations in the order of hundreds.

We propose a grid search tuning parameter selection based on minimizing/maximizing the AIC, the BIC and the Gini index. Here, a set of $\lambda(\omega)$ is set as in interval length of 0.01 as $\lambda(\omega) = \{\lambda_1(\omega),\lambda_2(\omega),\ldots,\lambda_{100}(\omega)\}$ where $\lambda_1(\omega) = 0.01,\lambda_2(\omega) = 0.02,\cdots,\lambda_{100}(\omega) = 1$. For each fixed $\lambda_i(\omega)$, we can obtain $\hat{\mathbf{x}}_{\lambda_i(\omega)}$ by optimizing (24). Note that for a high-dimensional least squares Lasso problem, it is computationally expensive to implement through the Newton system. In order to balance between computation and convergence rate we propose to use the iterative method to solve the Newton system by using the truncated Newton method combined with interior point method [20]. From Eqs. (11), (18) and (19), we have

$$\mathrm{AIC}(\lambda_i(\omega)) = \ell\log(\|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_{\lambda_i(\omega)}(\omega,k)\|_2^2/\ell) + 2\mathrm{df} \tag{25}$$

$$\mathrm{BIC}(\lambda_i(\omega)) = \ell\log(\|\mathbf{y}(\omega,k) - \mathbf{\Phi}(\omega)\hat{\mathbf{x}}_{\lambda_i(\omega)}(\omega,k)\|_2^2/\ell) + \mathrm{df}\cdot\log(\ell) \tag{26}$$

$$\mathrm{GI}(\lambda_i(\omega)) = 1 - 2\sum_{n=1}^{N} \frac{\hat{x}_{\lambda_i(\omega)}(\omega,k,n)}{\|\hat{\mathbf{x}}_{\lambda_i(\omega)}(\omega,k)\|_1}\left(\frac{N-n+0.5}{N}\right) \tag{27}$$

From the above, each optimized parameter can be found as $\lambda_i(\omega) \in \lambda(\omega)$ as follows

$$\lambda_{\mathrm{MinAIC}} = \arg\min_{\lambda(\omega)} \mathrm{AIC}\{\lambda(\omega)\} \tag{28}$$

$$\lambda_{\mathrm{MinBIC}} = \arg\min_{\lambda(\omega)} \mathrm{BIC}\{\lambda(\omega)\} \tag{29}$$

$$\lambda_{\mathrm{MinGI}} = \arg\min_{\lambda(\omega)} \mathrm{GI}\{\lambda(\omega)\} \tag{30}$$

$$\lambda_{\mathrm{MaxGI}} = \arg\max_{\lambda(\omega)} \mathrm{GI}\{\lambda(\omega)\} \tag{31}$$

Finally, the corresponding optimal estimators are obtained as

$$\hat{\mathbf{x}}(\lambda(\omega)_{\mathrm{MinAIC}}),\hat{\mathbf{x}}(\lambda(\omega)_{\mathrm{MinBIC}}),\hat{\mathbf{x}}(\lambda(\omega)_{\mathrm{MinGI}}),\hat{\mathbf{x}}(\lambda(\omega)_{\mathrm{MaxGI}}). \tag{32}$$

Each optimal estimator is then evaluated against the affective measures, i.e., PESQ and segSNR. As mentioned the proposed approach is a grid based ratio selection method to optimize $\lambda(\omega)$. Here, the optimized $\lambda(\omega)$ is chosen based on the optimization of either on the Gini index, AIC and BIC criterion as shown above. In the following numerical study, we investigate the influence of hyperparameterizing $\lambda(\omega)$ on the results of compressive speech enhancement in terms of perceptual evaluation of speech quality (PESQ) and the segmental SNR (segSNR). Generally speaking, PESQ measures the overall improvement in the perceptibility of the speech signal, whereas segmental SNR rests more heavily on the suppression of noise in the observation.

7. Numerical experiments

7.1. Experiment settings

Four different types of noise sources from the NOISEX database, namely, babble, subway, destroyer and car noise were tested over a wide range of SNR, from 0 dB to 20 dB, with similar SNR setting as in [17]. The noise types were chosen to represent the different degree of non-stationarity noise encountered in the real world. Five female and

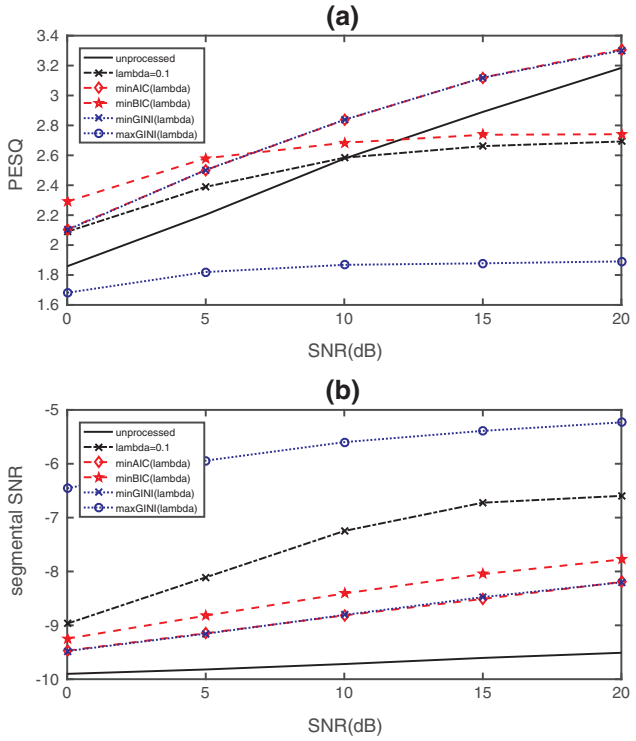

Fig. 2. The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for babble noise.

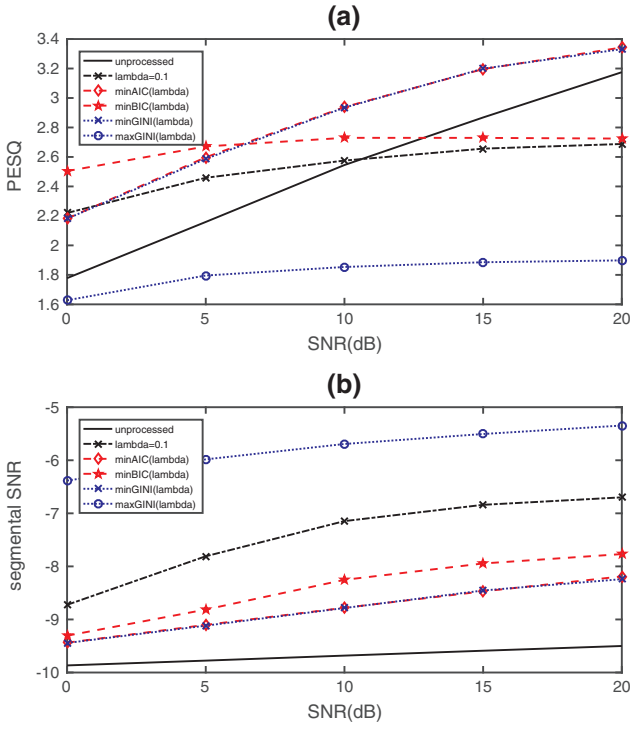

Fig. 3. The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for car noise.

five male speech signals from the TIMIT database were used as stimuli. The performance was evaluated by using the segmental SNR and the PESQ measure with a total of five female and five male speech signals from the TIMIT database. As mentioned in the introduction, PESQ measure is an automated evaluation process, which in this case a key measure for the inclusion of affective design. The PESQ score reveals how good or bad the perceptual quality of the audio signal to a human listener. This paper also includes the objective measure segmental SNR as a comparison. The number of frequency points was fixed at 256 with 50% oversampling and the compressive ratio, M/N was set to 0.9.

7.2. Hyperparameterizing λ based on Gini, AIC and BIC criterion

Four criteria based on Eqs. (12), (13), (18) and (19) were used to examine the influence of $\lambda(\omega)$ on compressive speech enhancement for a range of SNRs. In this case, each of the criteria is evaluated in each frequency band via grid search. We take fixed $\lambda(\omega) = 0.1$ for comparison purposes as the same implementation in [17]. Figs. 2–5 show the PESQ and segmental SNR performance of the four model selection criterion for babble noise, car noise, subway noise and destroyer noise, respectively. Evidently, the role of $\lambda(\omega)$ is crucial as its variation results in a very different performance across the SNRs.

In terms of PESQ, the minimization of the Gini and AIC criterion provide a consistent performance across the SNR range for the different types of noise. Both the criterion achieves higher PESQ values over the performance of having a fixed value of $\lambda(\omega)$ e.g., $\lambda(\omega) = 0.1$ (see [17]) and the unprocessed observation. Note that minimization of the Gini index results in the most non-sparse solution in the set of sparse solution. This means that the recovery process emphasizes on maintaining the speech signal as opposed to the reduction of noise (via a sparser solution). Interestingly, the minimization of BIC does not provide much improvement when the SNR>10 dB. Also, when compared to the AIC criterion, BIC obtains lower PESQ improvement but a higher segmental SNR improvement. This corroborates with the fact that in general, BIC tends to choose a parsimonious model compared to AIC. Hence for

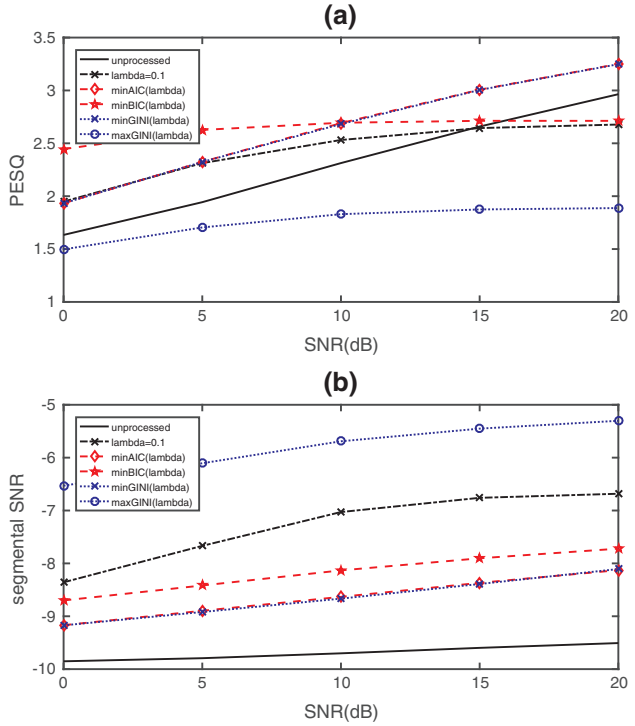

Fig. 4. The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for subway noise.

compressive speech enhancement, AIC is more inclined to select a model with less sparsity. This explains why AIC criterion results in a higher PESQ score but a lower segmental SNR compared to the BIC criterion.

In terms of segmental SNR improvement, the maximization of the Gini index gains the highest improvement with an approximately 4 dB

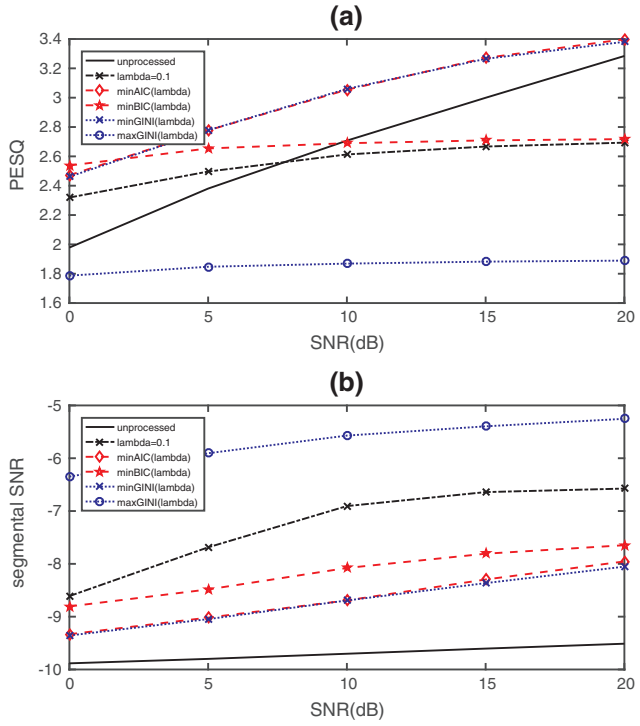

Fig. 5. The (a) PESQ and (b) segmental SNR of the different hyperparameter optimization methods as a function of SNRs for destroyer noise.

gain over the range of SNRs and the different types of noise. This is because the maximization of the Gini index results in the sparsest representation, which as shown in Section 4.2 is often the ones with the highest SNR. However, having an SNR improvement does not necessarily translate to overall speech intelligibility improvement. This is shown by the corresponding results in terms of the PESQ, where the maximization of Gini index attains the lowest PESQ improvement. This indicates that maxGINI maximally suppresses noise at the expense of the perceptual aspects of the output. This may be suitable for applications such as speech recognition where noise is the main issue. However, for hearing instruments such as assistive listening devices, SNR may not be the primary factor as improving SNR does not necessarily improve the perceptual part of speech as measured by PESQ. The proposed method allows such tuning by choosing the different criterion for the application in question. In a way it effectively parameterizes the sparse reconstruction through $\lambda(\omega)$ to allow for an engineering trade-off between noise suppression and perceptual preservation. Informal listening test confirms the improvement with respect to the different criteria used.

8. Conclusions

This paper presents a two-level optimization approach to incorporate quality measures in a speech application such as compressive speech enhancement. The results show that quality measures can be factored in the solutions by hyperparameterizing the tuning parameter in the sparse reconstruction. By doing so, the solutions are effectively tailored to the desired design attributes by a single parameter. The two-level approach first compresses the big data and subsequently optimizes the sparse the solution via the AIC, BIC model selection and the Gini performance index. The set of solutions is then measured against the quality measures for the desired solution. Comprehensive numerical experiments in a range of real-world noise with varying SNRs show that proper tuning of the hyperparameter can effectively trade-off between speech distortion and noise suppression. For future work, the optimization of the tuning parameter will be extended to the use of heuristics methods such as the particle swarm optimization (PSO).

Acknowledgement

This paper is partially supported by RGC Grant PolyU. 152200/14E and PolyU 4-ZZGS. S.Y. Low would like to acknowledge the Malaysian Ministry of Higher Education (MOHE) Fundamental Research Grant Scheme (FRGS 2015-1) for the support of the research.

References

[1] Chan KY, Nordholm S, Yiu KFC, Togneri R. Speech enhancement strategy for speech recognition microcontroller under noisy environments. Neurocomputing 2013;118:279–88288.
[2] Chan KY, Low SY, Nordholm S, Yiu KFC, Togneri R. A decision-directed adaptive gain equalizer for assistive hearing instruments. IEEE Trans Instrum Meas 2014;63(8):1886–95.
[3] Yiu KFC, Chan KY, Low SY, Nordholm S. A multi-filter system for speech enhancement under low signal-to-noise ratios. J Indus Manage Optim 2009;5(3):671–82.
[4] Low SY, Grbic N, Nordholm S. Robust microphone array using subband adaptive beamformer and spectral subtraction. IEEE Int Conf Commun Syst 2002;2:1020–4.
[5] Jiang H, Kwong C, Liu Y, Ip WH. A methodology of integrating affective design with defining engineering specifications for product design. Int J Prod Res 2015;8(53):2472–88.
[6] Kwong CK, Jiang H, Luo XG. Ai-based methodology of integrating affective design, engineering, and marketing for defining design specifications of new products. Eng Appl Artif Intell 2016;47(C):49–60.
[7] Dahlgaard JJ, Schütte S, Ayas E, Dahlgaard?Park SM. Kansei/affective engineering design: a methodology for profound affection and attractive quality creation. TQM J 2008;20(4):299–311.
[8] Low SY, Nordholm S, Teo KL. Use of efficient frontier in microphone arrays electronics letters. IEEE Electron Lett 2006;20(42):1186–7.
[9] Pham DT, El-Chami Z, Guérin A, Servière C. Modeling the short time fourier transform ratio and application to underdetermined audio source separation, ser. Lecture Notes in Computer Science. Berlin: Springer; March 2009, 5441/2009.
[10] Gardner TJ, Magnasco MO. Sparse time-frequency representations. Proc Natl Acad Sci 2006;103:6094–9.
[11] Davis A, Low SY, Nordholm S. A multi-decision sub-band voice activity detector. In: European signal processing conference; 2006. p. 1–5 [September].
[12] Ghosh PK, Tsiartas A, Narayanan S. Robust voice activity detection using long-term signal variability. IEEE Trans Audio, Speech Lang Process 2011;19(3):600–13.
[13] Donoho DL. Compressed sensing. IEEE Trans Inform Theory 2006;52(4):1289–306.
[14] Candès EJ, Wakin MB. An introduction to compressive sampling. IEEE Signal Process Magaz 2008:21–30.
[15] Candès E, Romberg J, Tan T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inform Theory 2006;52(2):489–509.
[16] Candès EJ, Tao T. Near-optimal signal recovery from random projections: universal encoding strategies?. IEEE Trans Inform Theory 2006;52(12):5406–25.
[17] Low SY, Pham DS, Venkatesh S. Compressive speech enhancement. Speech Commun 2013;55(6):757–68.
[18] Sreenivas TV, Kleijn WB. Compressive sensing for sparsely excited speech signals. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing; April 2009. p. 4125–8.
[19] Wu D, Zhu WP, Swamy MNS. A compressive sensing method for noise reduction of speech and audio signals. In: IEEE 54th international Midwest symposium on circuits and systems (MWSCAS)); August 2011. p. 1–4.
[20] Kim SJ, Koh K, Lustig M, Boyd S, Gorinevsky D. An interior-point method for large-scale l1-regularized least squares. IEEE J Select Top Signal Process 2007;1(4):606–17.
[21] Rix A, Beerends J, Hollier M, Hekstra A. Perceptual evaluation of speech quality (pesq) – a new method for speech quality assessment of telephone networks and codecs. IEEE Int Conf Acoust, Speech Signal Process 2001;2:749–52.
[22] Gill PR, Wang A, Molnar A. The in-crowd algorithm for fast basis pursuit denoising. IEEE Trans Signal Process 2011;59(10):4595–605.
[23] Chen S, Donoho D, Suanders M. Atomic decomposition by basis pursuit. SIAM Rev 2001;43(1):129–59.
[24] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol) 1996;58(1):267–88.
[25] Chen S, Donoho D, Saunders M. Atomic decomposition by basis pursuit. SIAM J Scient Comput 1998;20(1):33–61.
[26] Hurley N, Rickard S. Comparing measures of sparsity. IEEE Trans Inform Theory 2009;55(10):4723–41.
[27] Zonoobi D, Kassim AA, Venkatesh YV. Gini index as sparsity measure for signal reconstruction from compressive samples. IEEE J Select Top Signal Process 2011;5(5):927–32.
[28] Dicker L, Huang B, Lin X. Variable selection and estimation with the seamless l_0 penalty. Stat Sin 2013;23:929–62.
[29] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6(2):461–4.
[30] Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. Ann Stat 2007;35(5):2173–92.

[31] Mariani A, Giorgetti A, Chiani M. Model order selection based on information theoretic criteria: design of the penalty. IEEE Trans Signal Process 2015;63(11):2779–89.
[32] Loizou P. Speech quality assessment. Multimedia analysis, processing and communications, Vol. 346. Springer-Verlag; 2011. p. 623–54.
[33] Benesty J, Makino S, Chen J. Speech enhancement, ser. signals and communication technology. Berlin: Springer-Verlag; 2005.
[34] Loizou P. Speech enhancement: theory and practice. Boca Raton, Florida, USA: CRC Press, Taylor and Francis; 2007.
[35] Liu WM, Jellyman KA, Mason JSD, Evans NWD. Assessment of objective quality measures for speech intelligibility estimation. In: IEEE International conference on acoustics speech and signal processing; 2006. p. I.
[36] Hu Y, Loizou PC. Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio, Speech, Lang Process 2008;16(1):229–38.
[37] Ma J, Hu Y, Loizou PC. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J Acoust Soc Am 2009;125(5):3387–405.