Speech Enhancement based on Compressive Sensing Algorithm

To cite this article: Amart Sulong et al 2013 IOP Conf. Ser.: Mater. Sci. Eng. 53 012076

View the article online for updates and enhancements.

Related content
- Speech enhancement on smartphone voice recording
  Bagus Tris Atmaja, Mifta Nur Farid and Dhany Arifianto
- An optimization method for speech enhancement based on deep neural network
  Haixia Sun and Sikun Li
- Multi-energy CT based on a prior rank, intensity and sparsity model (PRISM)
  Hao Gao, Hengyong Yu, Stanley Osher et al.
Speech Enhancement based on Compressive Sensing Algorithm

Amart Sulong¹, Teddy S. Gunawan¹, Othman O. Khalifa¹, Jalel Chebil¹
¹ ECE Department, Faculty of Engineering International Islamic University Malaysia
Kuala Lumpur, Malaysia

E-mail: tsgunawan@iium.edu.my, amartuia@gmail.com

Abstract. There are various methods, in performance of speech enhancement, have been proposed over the years. The accurate method for the speech enhancement design mainly focuses on quality and intelligibility. The method proposed with high performance level. A novel speech enhancement by using compressive sensing (CS) is a new paradigm of acquiring signals, fundamentally different from uniform rate digitization followed by compression, often used for transmission or storage. Using CS can reduce the number of degrees of freedom of a sparse/compressible signal by permitting only certain configurations of the large and zero/small coefficients, and structured sparsity models. Therefore, CS is significantly provides a way of reconstructing a compressed version of the speech in the original signal by taking only a small amount of linear and non-adaptive measurement. The performance of overall algorithms will be evaluated based on the speech quality by optimising using informal listening test and Perceptual Evaluation of Speech Quality (PESQ). Experimental results show that the CS algorithm performs very well in a wide range of speech test and being significantly given good performance for speech enhancement method with better noise suppression ability over conventional approaches without obvious degradation of speech quality.

1. Introduction

Technological advances, such as the development of telephone and radio, enhance our speech communication capability even more by enabling direct communication across larger distance, broader audiences, and more challenging circumstances. Speech is a special subclass of audio signals. It is produced by exciting the acoustic cavity, the vocal tract, by pulses of air released through the vocal cords for voiced sounds, or by turbulence for unvoiced sounds. Speech can be explained as the verbal form of human communication. Speech system, a dynamic, information-bearing signal, is also called the acoustic waveform [1]. These waves are produced due the sound pressure generated in the mouth of the speaker as a result of some sequence of coordinated movements of a series of structures in the human vocal system. Speech signal processing is purposed enhancing speech communication systems at a point in the communication link so that the conveyed message can be recovered with minimum loss on the receiver’s side. Some examples of speech signal processing are speech coding, background noise reduction, time-scale or pitch modification, and speech recognition. Speech communication systems therefore are subject to more stringent requirements than music systems.

¹ Corresponding author: Amart Sulong, amartuia@gmail.com
Commonly, in most applications, the aim of speech enhancement is to improve the quality and intelligibility of degraded speech [2]. The motivation behind the Speech enhancement is concerned with improving some perceptual aspect of speech that has been degraded by additive noise that corrupted speech. However, there is always a tradeoff between noise reduction and signal distortion – more noise reduction is always accompanied by more signal distortion [12]. The improvement in quality is highly desirable as it can reduce listener fatigue [2], particularly in situations in which the listener is exposed to high level of noise for long period of time (e.g., manufacturing). In other word, Speech enhancement algorithms reduce or suppress the background noise to some degree and are sometime referred to as noise suppression algorithm. Introducing the speech distortion from noise reduction of the background noise may impair the speech intelligibility. Hence, the main challenge in speech enhancements is to design effective algorithm to suppress the noise without introducing any perceptible distortion in the signal.

Compressive sensing (CS) is a new type of sampling theory, which predicts that sparse signals can be reconstructed from what was previously believed to be incomplete information [3]. This method is considered as main feature to efficient algorithms such as $l_1$-minimization which can be used for recovery. The traditional approach of reconstructing signals from measured data follows the well-known Shannon sampling theorem [4]. It states that the sampling rate must be twice the highest frequency. Similarly, the fundamental theorem of linear algebra suggests that the number of collected samples (measurements) of a discrete finite-dimension signal should be at least as large as its length (its dimension) in order to ensure reconstruction. This principle underlies most devices of current technology, such as analog-to-digital conversion or audio and video electronics.

In addition, CS theory was growing very rapidly in the past few years and undergoes significant advances on an almost daily basis [3]. Its theory provides a fundamentally new approach to data acquisition and asserts that one can recover certain signal from far fewer samples or measurements than traditional methods use [5]. CS relies on the empirical observation that many types of signals can be well approximated by sparse expansion in terms of suitable basis. However, it is by only a small number of nonzero coefficients compare with the key to efficiency of many lossy compression techniques such as JPEG, MP3, etc. A compression is obtained by simply storing only the largest basis coefficients. The non-stored coefficients are simply set to zero while reconstructing the signal. This is certainly a reasonable strategy when full information of the signal is available.

However, when the signal first has to be acquired by a somewhat costly, lengthy, or otherwise difficult measurement (sensing) procedure, this seems to be a waste of resources. Clearly, the measurements have to be suitably designed. It is a remarkable fact that all provably good measurement matrices designed so far are random matrices. It is for this reason that the theory of CS uses a lot of tools from probability. Quite surprisingly, CS provides nevertheless a way of reconstructing a compressed version of the original signal by taking only a small amount of linear and non-adaptive measurements. The precise number of required measurements is comparable to the compressed size of the signal.

2. Speech enhancement

2.1. Understanding the Enemy: Noise

Making meaningful performance of speech enhancement algorithms is concerned to prior of design algorithms to combat additive noise. Understanding of additive noise in various types of application is crucial behavior. It focuses on difference of the noise sources in term of temporal and spectral characteristic and the rang of noise levels that may encounter in real life [2, 5].

The study of the noise can be described in two categories stationary and non-stationary [2]. The stationary of the noise remains unchanged over the time, such as the fan noise coming from PCs. On the other hand, the non-stationary of the noise constantly changes the spectral (and temporal) characteristics, such as restaurant noise. Clearly, the task of suppressing noise that is constantly changing (non-stationary) is more difficult than that of suppressing stationary noise. The shape of their
spectrum, particularly the distribution of the noise energy in the frequency domain, is another
distinctive feature of the various types of noise to the frequency range. The knowledge of the range of
speech and noise intensity levels in real-world scenarios are critical to the design of speech
enhancement algorithm [2]. These terms can estimate the range of signal-to-noise ratio (SNR) levels
encountered in realistic environments. This concern is addressed as important method to speech
enhancement algorithms. Its algorithms will be affected by suppressing noise and improved speech
quality within that range of SNR levels [6].

2.2. Principle of Speech Enhancement
This section presents review and discussion on the different speech enhancement methods used to date
[5,8]. Speech production model is shown in Figure 1 and Figure 2.

![Figure 1. The basic diagram of speech enhancement [5].](image1)

![Figure 2. General speech enhancement algorithm [9].](image2)

Figure 1 and Figure 2 show the speech enhancement method. The speech enhancement
problem can be simplified by assuming the following additive model:

\[ s(n) = \hat{s}(n) + v(n) \]  

(1)

where \( s(n) \), \( \hat{s}(n) \), and \( v(n) \) represent digitized noisy speech, clean speech, and noise, respectively. The
noise \( s(n) \) is assumed to be additive noise. The speech enhancement is a method of extracting clean
speech \( \hat{s}(n) \) from noisy speech signal \( s(n) \) under certain constraints as shown in Figure 1. With
reference [5] was to simplify the fundamental of speech enhancement. Most of the speech
enhancement methods can be described in three steps as shown in Figure 2 [7]; Step 1: decomposing
the speech signal into a transformed domain; Step 2: estimating the clean channel signals in the
transformed domain; Step 3: synthesizing the speech from the estimated channel signals. Different
transformation techniques and estimation algorithms are used by different speech enhancement
methods. Speech enhancement in general is purposed to the main of objectives of enhancement that
come as follow; 1) to improve the overall quality; 2) to increase intelligibility, and; 3) to reduce
listener’s fatigue.

In speech processing to model speech, assumptions are made. The more the assumptions are made
more will be the sensitivity of the enhancement system to the deviations from these assumptions. The
assumption is depended on the requirements and objective different aspect of speech which need to be enhanced. Sometimes there is need to enhance quality at the cost of intelligibility and other times intelligibility needs to be enhanced sacrificing some quality. Another important consideration in speech enhancement systems is that ultimately the resulting enhancement is related to human listener. Thus speech enhancement must inevitably take into account aspects of human perception.

The speech enhancement problem covers a broad spectrum of constraints, applications and issues. Environments in which an additive background signal has been introduced are common. The background may be noise-like such as in aircraft, street noise, etc. or may be speech-like such as environment with competing speakers. Therefore, Speech enhancement depends on; 1). Good signal processing technique, 2). Human perceptual factor, 3). Speech quality and intelligibility are dependent on short term spectral and 4). Amplitude and insensitive to spectral phase.

2.3. Classification of Speech Enhancement Techniques
The research [1] mentioned that speech enhancement systems can be classified in a number of ways based on the criteria used or application of the enhancement system as shown in Table 1.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Possible Strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of input channels</td>
<td>One/Two/Multiple</td>
</tr>
<tr>
<td>Domain of processing</td>
<td>Time/Frequency</td>
</tr>
<tr>
<td>Type of algorithm</td>
<td>Non-adaptive/Adaptive</td>
</tr>
<tr>
<td>Additional constraints</td>
<td>Speech production/Perception</td>
</tr>
</tbody>
</table>

2.4. Subjective and Objective Evaluation

Subjective Evaluation: The most accurate method for evaluating speech quality in real-world noise that introduced by four classes of speech enhancement algorithms: Spectral subtraction, subspace, statistical-model based, and Wiener algorithms. The subjective quality ratings were obtained using ITU-T P.835 methodology design to evaluate the quality of enhanced speech along three dimensions: Signal distortion, noise distortion, and overall quality. ITU-T P.835 was designed to lead the listeners to integrate the effects of both signal and background distortion in making their rating of overall quality. It is costly and time consuming to predict speech quality with high correlation. But it is accurate and reliable (i.e., repeatable) provided it is performed under stringiest conditions (e.g., sizeable listener panel, inclusion of condition) [2]. These algorithms were evaluated using a newly developed noisy speech corpus (NOIZEUS) suitable for evaluation of speech enhancement algorithms and available. The sentences were originally sampled at 25 kHz and downsampled to 8 kHz. As shown in Figure 3. The statistical analysis is presented to compare the performance of the algorithms within each of the four classes (subspace, statistical-model, subtractive, and Wiener-type). This comparison was meant to examine whether there were significant differences algorithms within each class as mention in [6].

In Objective evaluation: the measurement matrix be able to assess the quality of the enhanced speech without needing access to the original speech signal by only using formula therefore it is very fast to testing process. it incorporates knowledge from different levels, including low-level processing (e.g., psychoacoustics) and higher-level processing such as prosodics, semantics, linguistics, and pragmatics. It is also predict with high accuracy to the results obtained from subjective listening tests with normal-hearing listeners. Reference [7,8,9,10] explained in their study, the speech enhancement evaluates the performance of several objective measures in terms of predicting the quality of noisy speech enhanced by noise suppression algorithm.
3. Compressive sensing theory

Theoretically, speech enhancement is required in many situations in which signal to be communicated or store. This research field is focused on suppression of additive noise. From the point of view of signal processing, additive noise is easier to deal with than convolution noise or nonlinear disturbances. The ultimate goal of speech enhancement is to eliminate the additive noise present in speech signal and restore the speech signal to its original form. Several methods have been developed in its research efforts [10]. However, it is very difficult to reliably predict the characteristics of the interfering noise signal or the exact characteristics of the speech waveform.

Hence, in effect, the speech enhancement methods are sub-optimal and can only reduce the amount of noise in the signal to some extent. Due to the sub-optimal nature of these methods, some of the speech signal can be distorted during the process. Hence, there is a trade-off between distortions in the processed speech and the amount of suppressed. The effectiveness of the speech enhancement system can therefore be measured based on how it performs in light of this trade-off.

There are two structure design approach can be applied for speech enhancement from classical theory and the novel proposed theory from structure design approach as follow:

3.1 Classical Theory (non-compressive sensing): Transform Coding for Speech Enhancement

One would first acquire a large amount of data in speech signal, compute and an appropriate basis and projections on it, and then transmit these projections and basic used. This is wasteful of resources since many more data points are initially collected than are transmitted. This approach will produce long-established paradigm for digital data acquisition with uniformly sample data at Nyquist rate and compress data as shown in Figure 4. Unfortunately, this sample-then-compress framework suffers from three inherent inefficiencies.

![Figure 4](image)

Figure 4. The compression of the signal in traditional methods [11].
In his research [11] mentioned the fact that compressible signals are well approximated by $K$-sparse representations forms the foundation of transform coding. In data acquisition systems, transform coding plays a central role: the $N$-sample signal $x$ is acquired; the complete set of transform coefficients $\{s_i\}$ is computed via $s = \Phi x$; the $K$ largest coefficients are located and the $(N-K)$ smallest coefficients are discarded; and the $K$ values and locations of the largest coefficients are encoded. Unfortunately, this sample-then-compress framework suffers from three inherent inefficiencies. First, the initial number of samples $N$ may be large even if the desired $K$ is small. Second, the set of all $N$ transform coefficients $\{s_i\}$ must be compute even though all but $K$ of them will discarded. Third, the locations of the large coefficients must encoded, thus introducing overhead.

3.2 Compressive Sensing Theory

CS [11] referred in his research as a new approach, compressive sensing address these inefficiencies by directly acquiring a compressed signal representation without going through the intermediate state of acquiring $N$-sample. Consider a general linear measurement process that computes $M < N$ inner products between $x$ and a collection of vectors $\{\phi_j\}$ as in $y_j = \langle x, \phi_j \rangle$. Arrange the measurements $y_j$ in an $M \times 1$ vector $y$ and measurement vector $\phi_j$ as rows in an $M \times N$ matrix $\Phi$. Then, by substituting $\Psi$ from (2), $y$ can be written as

$$y = \Phi x = \Phi \psi s = \Theta s$$

where $\Theta = \Phi \psi$ is $M \times N$ matrix. The measurement process is not adaptive, meaning that $\Phi$ is fixed and does not depend on the signal $x$. The problem consists of designing $\Phi$ a stable measurement matrix $\Phi$ such that the salient information in any $K$-sparse or compressible signal is not damaged by the dimensionality reduction from $x \in \mathbb{R}^N$ to $y \in \mathbb{R}^M$ and $\Phi$ a reconstruction algorithm to recover $x$ from only measurements $y$ (or about as many measurements as the number of coefficients recorded by a traditional transform coder). As shown in Figure 5.

![Figure 5](image)

**Figure 5.** The method shows the new approach of speech enhancement algorithm based on compressive sensing (CS).

4. Proposed method for speech enhancement using compressive sensing Algorithm

Compressive sensing (CS) is an alternative to Shannon/Nyquist sampling for the acquisition of sparse or compressible signal that can be well approximated by just $K << N$ elements from an $N$-dimensional basis. Instead of taking periodic samples, CS measures inner products with $M < N$ random vectors and then recovers the signal via a sparsity-seeking optimization or greedy algorithm. Standard CS dictates that robust signal recovery is possible from $M = O(\log(N/K))$ measurement. A model-based CS theory provides concrete algorithms with provable performance guarantees. This model is designed as a new
class of the structured compressible signals along with a new sufficient condition for robust structured compressible signal recovery.

The research in CS has focused primarily both on reducing the number of measurements $M$ (as a function of $N$ and $M$) and on increasing the robustness and reducing the computational complexity of recovery algorithm. The CS system can robustly recover $K$-sparse and compressible signals from just $M = O(\log(N/K))$ noisy measurements using polynomial-time optimization solvers or greedy algorithms.

4.1 Proposed algorithm
The proposed speech enhancement algorithm using compressive sensing is illustrated in Figure 6. Analysis filter bank used gammatone filter bank to utilize due to its resemblance to the shape of human auditory filters [13]. Of the various transforms available to sparsify speech signal, Discrete Cosine Transform (DCT) will be chosen due to its simplicity and it good decorrelation property [14].

![Figure 6. Proposed speech enhancement based on compressive sensing algorithm](image)

Then by using subband modification to produce of subband signal coefficients for analysis the speech signal. On the synthesis part, for solving convex optimization of compressive sensing, the gradient projection for sparse reconstruction (GPSR) algorithm [15] was utilized due to it reconstruction quality can be traded with the available processing power at the synthesis side. The higher the processing power or the longer available to solve convex optimization problem, the higher the
synthesis signal quality. In other word, invert DCT (IDCT) and delay compensation were applied to the compressed signal. The amount of filter delay accumulated by each subband is different and without compensating for this delay, the synthesis of subband signal will lead to an incoherent output signal, i.e. lower quality signal.

4.2 Performance results
The performance of the proposed algorithm was evaluated using PESQ score. As mention in the previous section, the Table 2, shows the objective evaluation using PESQ. The results produce excellence score for PESQ measurement which can be considered as acceptable quality for everyday communication.

![Original signal](image1.png)

a). Original signal

![Signal after using Compressive sensing method](image2.png)

b). Signal after using Compressive sensing method

![The cooperation of the constructed and original signal after using CS algorithm](image3.png)

c). The cooperation of the constructed and original signal after using CS algorithm

Figure 6. The method shows the new approach of speech enhancement based on compressive sensing (CS) algorithm
The speech enhancement by using compressive sensing, in **Figure 6**, are constructed the signal almost as the same as original signal of the speech and it produces high performance score as shown in **Table 2**. The speech sound tested from **Table 2.**, used the datasets with 22 speech files of the non-trained files. The quality of the speech enhancement is evaluated using Perceptual Evaluation of Speech Quality (PESQ), Signal-to-Noise Ratio (SNR). The results show that the speech enhancements have achieved the excellent result. The average of PESQ score is 4.478 which is a good quality as listening tests confirms.

**Table 2.** The performance results by evaluate using PESQ for the speech sound tested from the datasets with 22 speech files of the non-trained files.

<table>
<thead>
<tr>
<th>Speech Signal</th>
<th>PESQ</th>
<th>Elapsed time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp01.wav</td>
<td>3.748</td>
<td>9.757911</td>
</tr>
<tr>
<td>sp02.wav</td>
<td>3.727</td>
<td>9.298481</td>
</tr>
<tr>
<td>sp03.wav</td>
<td>3.866</td>
<td>8.929454</td>
</tr>
<tr>
<td>sp04.wav</td>
<td>3.681</td>
<td>8.377577</td>
</tr>
<tr>
<td>sp05.wav</td>
<td>3.769</td>
<td>9.068006</td>
</tr>
<tr>
<td>sp06.wav</td>
<td>3.985</td>
<td>10.250037</td>
</tr>
<tr>
<td>sp07.wav</td>
<td>3.598</td>
<td>8.828342</td>
</tr>
<tr>
<td>sp08.wav</td>
<td>3.883</td>
<td>9.526404</td>
</tr>
<tr>
<td>sp09.wav</td>
<td>3.741</td>
<td>9.757164</td>
</tr>
<tr>
<td>sp10.wav</td>
<td>3.877</td>
<td>10.821275</td>
</tr>
<tr>
<td>sp11.wav</td>
<td>3.439</td>
<td>9.611207</td>
</tr>
<tr>
<td>sp12.wav</td>
<td>3.376</td>
<td>9.293070</td>
</tr>
<tr>
<td>sp13.wav</td>
<td>3.335</td>
<td>8.686577</td>
</tr>
<tr>
<td>sp14.wav</td>
<td>3.536</td>
<td>15.548540</td>
</tr>
<tr>
<td>sp15.wav</td>
<td>3.323</td>
<td>9.605843</td>
</tr>
<tr>
<td>sp16.wav</td>
<td>3.317</td>
<td>8.624775</td>
</tr>
<tr>
<td>sp17.wav</td>
<td>3.363</td>
<td>8.723107</td>
</tr>
<tr>
<td>sp18.wav</td>
<td>3.198</td>
<td>8.356940</td>
</tr>
<tr>
<td>sp19.wav</td>
<td>3.402</td>
<td>9.317810</td>
</tr>
<tr>
<td>sp20.wav</td>
<td>3.561</td>
<td>9.439650</td>
</tr>
<tr>
<td>sp21.wav</td>
<td>3.904</td>
<td>10.839918</td>
</tr>
<tr>
<td>sp22.wav</td>
<td>3.819</td>
<td>9.321426</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>3.6113</strong></td>
<td><strong>9.6356143</strong></td>
</tr>
</tbody>
</table>

5. Conclusions

In Speech Enhancement, the contention in this research is that it is possible to do even better by more fully leveraging concepts from state-of-the-art signal compression and processing algorithms. The key ingredient is a more realistic **structured sparsity** model that goes beyond simple sparsity by codifying the interdependency **structure** among the signal coefficients. The further evaluation for speech enhancement using compressive sensing has been presented. The proposed algorithm used gammatone filter banks and Discrete Cosine Transform (DCT) to sparsify the speech signal as the synthesis speech using compressive sensing will work best on sparse signals. The object evaluation using PESQ score revealed that the average **PESQ score** was around 3.6113 and **processing time** 9.6356 which is acceptable and given the excellent result.

6. References