Compressively sampled speech: How good is the recovery?

Kenneth V. Domingo* and Maricor N. Soriano

National Institute of Physics, University of the Philippines, Diliman, Quezon City, Philippines
*Corresponding author: kdomingo@nip.upd.edu.ph

Abstract

Modern signal acquisition technologies are made possible by the Nyquist-Shannon sampling theorem (NST). However, this paradigm is extremely wasteful as the signal is compressed before storing it by systematically discarding the imperceptible information. Compressive sensing (CS) aims to directly sense the relevant information. Current literature focus either on formulating more computationally-efficient algorithms, or methods which improve the reconstruction quality. In this paper, we quantify the reconstruction quality of compressively sampled speech with a perceptually intuitive metric—the Perceptual Evaluation of Speech Quality (PESQ)—and with the standard average segmental SNR (SNR_{seg}). It is shown that the quality of recovery of compressively sampled speech depends on the compression ratio and number of subbands used to represent the signal in the spectrogram domain.

Keywords: compressive sensing, signal processing, spectrogram

1 Introduction

Conventional sensing devices are based on the Nyquist-Shannon sampling theorem (NST), which states that given a signal's bandwidth, one can capture all the pertinent information about that signal if it is sampled at a rate

$$f_S \geq 2f_B \tag{1}$$

where f_B is the signal's highest frequency component, and $2f_B$ is known as the Nyquist rate. After this sampling process, the information is compressed by exploiting the signal's natural compressibility in some transform domain. For standard consumer-grade and even commercial-grade applications, this tried-and-true method of sampling and systematically discarding the imperceptible information works without a hitch. However, in certain situations when transmission bandwidth and/or storage comes at a premium, this process is highly wasteful. Candés et al. [1] and Donoho et al. [2] independently pioneered compressive sensing (CS), which could directly sample the portions of a signal that would otherwise survive the compression stage in conventional sampling. In this new sampling paradigm, we consider the linear model of signal acquisition

$$y_k = \mathbf{x} \cdot \mathbf{a}_k \tag{2}$$

In other words, the signal vector $\mathbf{x} \in \mathbb{R}^n$ is correlated with the basis waveforms $\mathbf{a}_k$ to yield a compressed information vector $\mathbf{y} \in \mathbb{R}^m$, where $m \ll n$. This causes the task of recovering $\mathbf{x}$ from $\mathbf{y}$ impossible since there exist an infinite number of candidate solutions $\hat{\mathbf{x}}$ which satisfy (2). This can be circumvented by enforcing constraints based on models of natural signals, as well as those based on optimization techniques. In particular, reconstruction can be achieved by enforcing sparsity and incoherence constraints [3]. From this, the process of reconstruction is then recast into a general minimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \tag{3}$$

where $\|\mathbf{x}\|_1$ denotes the ℓ_1 norm of $\mathbf{x}$. This problem can be solved in a number of ways with various algorithms.

Romero et al. [4, 5] performed compressive sensing of images in the Fourier domain and showed that it could be used to increase the signal-to-noise ratio of a signal. They also showed that such a sampling method was robust against noise. In the realm of audio signals, [6] constructed sensing matrices using a Gaussian-Logistic map; this was one of the first studies which use chaotic maps instead of random sequences, and is a prelude to the use of CS as an encryption algorithm. Through evolutionary algorithms such as differential evolutions, direct reconstruction of CS signals could also be attained in the temporal domain [7]. With audio recordings containing speech, however, the process is not as straightforward. Such signals typically use sampling rates of $\sim 10^3$ Hz, so it is necessary to process them in slices, in

the same way that large images can be processed in patches. Low et al. [8] performed this process by representing a signal in the spectrogram/modulation domain and applying the desired CS algorithm per sampling window.

In this paper, we perform compressive sensing of audio signals containing speech, and quantify the reconstruction quality using the International Telecommunication Standardization Sector (ITU-T) recommendation P.862, otherwise known as the Perceptual Evaluation of Speech Quality (PESQ). This metric is a full-reference, perceptually intuitive scoring system which models the obsolete mean opinion scores (MOS). Additionally, the average segmental signal-to-noise ratio (SNR_{seg}), which is commonly used in evaluating audio reconstruction quality, was also used. However, unlike the former, this is non-intuitive since its values are unbounded and, for a constant signal, varies with the number of sampling windows. We adopt the processing workflow of Low et al. while varying the length and number of sampling windows, as well as the compression ratio. In contrast with Mathew et al., we stick with the usual sensing matrices derived from i.i.d. uniform random samples. In an earlier work [9], we compared common algorithms in terms of computation time and reconstruction quality, and showed that LASSO strikes a balance between these two. We use this algorithm in this paper to perform the CS reconstruction.

2 Methodology

2.1 Obtaining & preprocessing the test signals

The test signals used in this paper were obtained from the TIMIT Acoustic-Phonetic Continuous Speech Corpus [10], which contains speech recordings, in WAV format, of English speakers separated by region, sex, and spoken sentence. All recordings have a sampling rate of 16 kHz and are, on average, 3 seconds long. This was then downsampled to 8 kHz to reduce the number of sampling windows needed, and also because the PESQ algorithm only applies to 8 kHz signals.

2.2 Transforming to a sparse domain

The first requirement for CS is that a signal should be transformed to some domain where it can be represented sparsely. In the case of recorded speech, the most appropriate domain would be the modulation domain, otherwise known as its spectrogram. First, we define the length of the sampling window and the percent overlap between adjacent windows, and divide the signal into frames using this sliding window. We then multiply each frame by a window function, typically a Hann window, defined as

$$w[n] = \sin^2\left(\frac{\pi n}{N}\right), \quad 0 \leq n \leq N \tag{4}$$

where $N + 1$ is the length of the window. Finally, we take the Fourier transform of each frame. The entire process is also called a short-time Fourier transform and can be summarized as

$$X(\omega, p) = \sum_{p=0}^{P-1} x[p]w[p - kR]e^{-i\omega p} \tag{5}$$

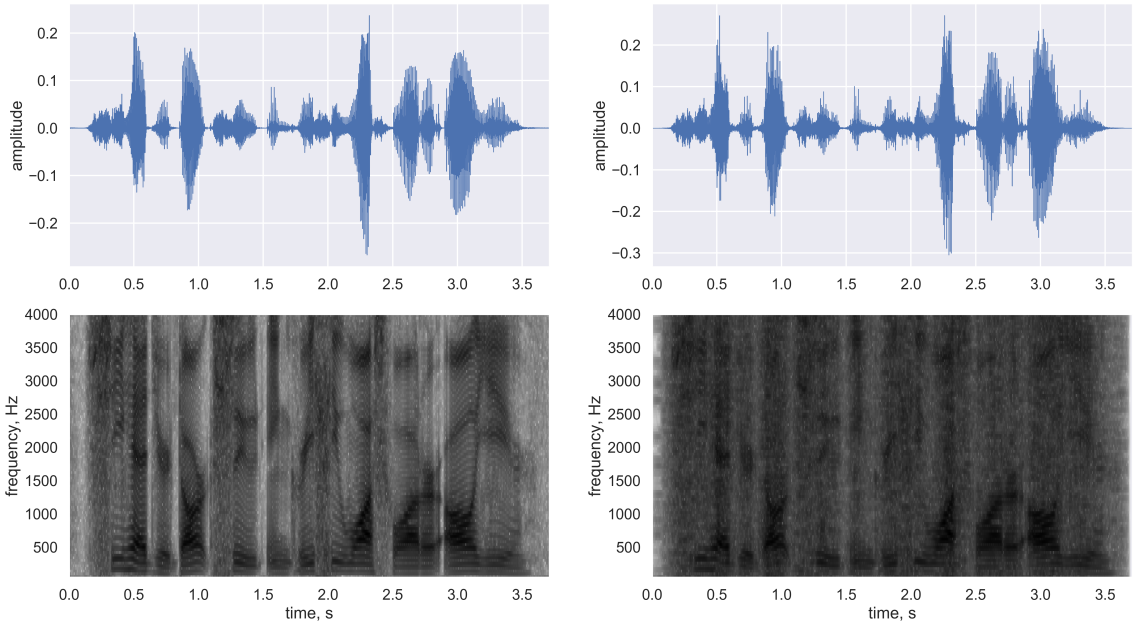
where $x[p]$ indicates the pth signal frame, $w[n - k]$ is the window function with hop size R, k is the time index, and ω is the frequency. Figure 1a shows a test signal in the temporal domain (top) and its corresponding spectrogram (bottom) with a 25 ms sampling window and 75% frame overlap.

2.3 Compressive sensing

The second requirement of CS is that the sensing matrix should be incoherent with the sparse basis. Typically, random samples are drawn that are distributed uniformly to simulate compressive measurements. We define an index sequence ξ corresponding to a random subset of m samples from the signal. The compressed signal vector can be defined as $\mathbf{y} = \mathbf{x}_\xi$, and the sensing matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ can be constructed by stacking the columns of a discrete cosine transform (DCT) matrix indexed by ξ. This sensing matrix can then be used for all the frames. After multiplying the frame with a window function as in Section 2.2, instead of directly taking the frame's Fourier transform, we can now perform reconstruction of the signal from the compressive measurements. The objective of LASSO [11] is

$$\min_{\mathbf{x}} \frac{1}{2m}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \alpha\|\mathbf{x}\|_1 \tag{6}$$

where m is the number of samples, and α is the ℓ_1 regularization parameter.



(a) Original (b) Recovered

Figure 1: Test speech signal obtained from the TIMIT speech corpus with the (a) original signal, and (b) reconstructed signal from compressive measurements. The top row is the representation in the time domain, while the bottom row is the spectrogram representation (25 ms sampling window with 75% overlap). The signal reads "She had your dark suit in greasy wash water all year".

2.4 Reconstruction evaluation

Reconstruction quality was evaluated using PESQ, which performs a series of standardized tests modeled after qualitative metrics. This algorithm analyzes and compares the reconstructed signal with the original and returns a value from 1 (bad) to 5 (excellent) [12]. The SNR_{seg} was also used, which is defined as [13]

$$\text{SNR}_{\text{seg}} = \frac{10}{B}\sum_{b=0}^{B-1}\log_{10}\frac{\sum_{i=Nb}^{Nb+N-1}x_i^2}{\sum_{i=Nb}^{Nb+N-1}(x_i-\hat{x}_i)^2} \tag{7}$$

where N is the frame length, B is the number of frames, x_i is the original signal, and $\hat{x}_i$ is the reconstructed signal.

3 Results and Discussion

A test signal was chosen at random from the TIMIT corpus, specifically the `DR8/MJLN0/SA1.wav` file. This indicates that the speaker was from dialect region 8 (nomadic), was male with speaker code `JLN0`, and spoke sentence code `SA1`, which reads "She had your dark suit in greasy wash water all year". After downsampling, the signal was compressively sampled with a compression ratio of 40% using 1024 frames and 75% frame overlap. The optimal regularization α was determined automatically via 5-fold cross validation, and the reconstruction result is shown in Fig. 1b. Qualitative comparison in the temporal domain shows that the waveforms of the original and reconstructed are similar; in the modulation domain, the dynamic range appears to have decreased, but the dominant frequencies can still be observed. Evaluating the PESQ and SNR_{seg} yields values of 2.50 and 0.07, respectively. At face value, we can immediately tell from the PESQ that the reconstructed signal quality is slightly below average. However, this distinction cannot be made for the SNR_{seg} since its bounds are not clearly defined.

Next, the error maps for the signal were generated by compressively sensing the signal and evaluating the metrics for varying compression ratios $m/n \in [0.1, 0.9]$ (in increments of 0.1) and number of frames/subbands $\in \{128, 256, 512, 1024\}$, while keeping the frame overlap constant. Figure 2 shows the PESQ map (left) and the SNR_{seg} map (right). The former exhibits sensitivity to the compression ratio and achieves a value of 4.0 (good) at around 70% compression. The latter tells a different story: it shows sensitivity towards the number of frames/subbands (as it is an *average* metric) with some additional degradation at lower compression ratios, and achieves a maximum value of 0.08 at around 1024 frames.

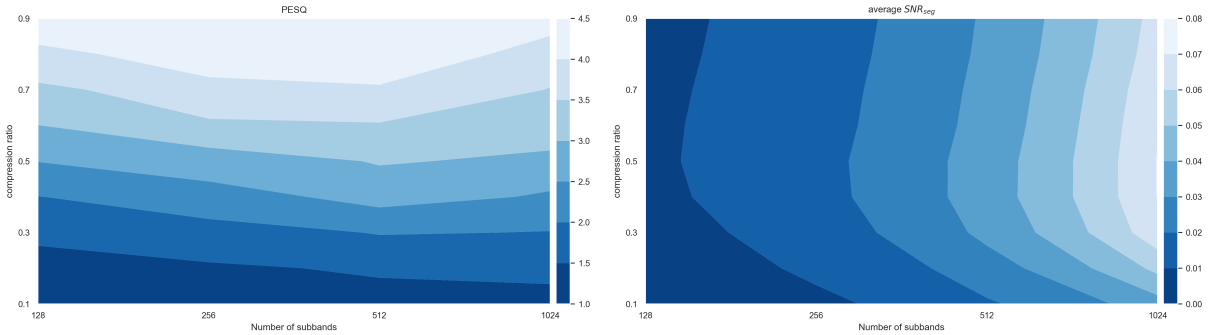


Figure 2: Two most commonly used metrics in evaluating quality of reconstructed speech recordings. PESQ is sensitive to the signal compression (left), while the $\mathrm{SNR}_{\mathrm{seg}}$ is sensitive to the number of subbands (right).

4 Conclusions

The reconstruction quality of compressively sampled audio signals containing speech recordings was shown to be dependent on two distinct factors, as well as the metric being used. Evaluation of a perceptually intuitive metric such as PESQ shows that it is sensitive to changes in compression, wherein a compression ratio of 60% is sufficient to yield an average (3.0) PESQ score. On the other hand, evaluation of a statistically-based metric such as $\mathrm{SNR}_{\mathrm{seg}}$ shows that it is sensitive to changes in the number of frames used to represent the signal in the spectrogram domain, where the use of more frames yields a higher $\mathrm{SNR}_{\mathrm{seg}}$ score.

References

[1] E. J. Candès, J. K. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Commun. Pur. Appl. Math.* **59**, 1207 (2006).

[2] D. Donoho, M. Elad, and V. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE T. Inform. Theory* **52**, 6 (2006).

[3] E. J. Candes and M. B. Wakin, An introduction to compressive sampling: A sensing/sampling paradigm that goes against the common knowledge in data acquisition, *IEEE Signal Proc. Mag.* **25**, 21 (2008).

[4] R. A. Romero, G. A. Tapang, and C. A. Saloma, Compressive sensing on the Fourier domain as a method for increasing image signal-to-noise ratio, in *Proceedings of the Samahang Pisika ng Pilipinas Physics Conference* (University of the Philippines Visayas, Iloilo City, Philippines, 2016), vol. 34, SPP-2016-PA-14.

[5] R. A. Romero, G. A. Tapang, and C. A. Saloma, Robustness of compressive Fourier-domain sampling against rounding-off errors and noise, in *Proceedings of the Samahang Pisika ng Pilipinas Physics Conference* (Puerto Princesa City, Philippines, 2018), vol. 36, SPP-2018-PB-21.

[6] M. R. Mathew and B. Premanand, Sub-Nyquist Sampling of Acoustic Signals Based on Chaotic Compressed Sensing, *Proc. Tech.* **24**, 941 (2016).

[7] I. András, P. Dolinský, L. Michaeli, and J. Šaliga, A time domain reconstruction method of randomly sampled frequency sparse signal, *Measurement* **127**, 68 (2018).

[8] S. Y. Low, Compressive speech enhancement in the modulation domain, *Speech Commun.* **102**, 87 (2018).

[9] K. V. Domingo and M. N. Soriano, Frequency domain reconstruction of stochastically sampled signals based on compressive sensing, in *Proceedings of the Samahang Pisika ng Pilipinas Physics Conference* (2019), vol. 37, SPP-2019-PB-38.

[10] J. S. Garafolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1* (Linguistic Data Consortium, 1993).

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**, 2825 (2011).

[12] Telecommunication Standardization Sector of ITU, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, *ITU-T Recommendation P.862 (02/01)* (2001).

[13] P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, 2013), 2nd ed.