Speech Communication 102 (2018) 87–99

Contents lists available at ScienceDirect

Speech Communication

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

Compressive speech enhancement in the modulation domain☆

Siow Yong Low

School of Electronics and Computer Science, University of Southampton, Malaysia Campus (USMC), Iskandar Puteri, Johor, Malaysia

ARTICLE INFO

Keywords:
Modulation domain
Modulation spectrum
Compressed sensing
Sparsity
Compressibility
Speech enhancement

ABSTRACT

Compressive speech enhancement (CSE) has gained popularity in recent years as it bypasses the need for noise estimation. Parallel to that, modulation domain has been widely studied in speech applications as it offers a more compact representation and is closely associated with speech intelligibility enhancement. Motivated by the development in modulation domain and CSE, this paper seeks to explore the suitability of modulation domain based sparse reconstruction for use in CSE. The main idea is to study if the increased sparsity in the modulation domain would benefit sparse reconstruction in CSE. The findings reveal that modulation transformation is sparser and offers a stronger restricted isometry property (RIP) compared to the frequency transformation, which is essential for sparse recovery with a high probability. The results are then extended to show that the sparse reconstruction error in the modulation domain is upper bounded by the frequency domain. Experimental results in a CSE setting concur with the theoretical derivations, with modulation domain CSE outperforming the frequency domain CSE through different speech quality measures.

1. Introduction

Dudley in his landmark paper concluded that speech signals in general are low frequency modulators, which modulate high frequency carriers very much like the amplitude modulation (AM) process (Dudley, 1939; 1940). Speech information can thus be viewed as a composite of modulations at various slow changing rates, on a fast changing carrier signal (Gallun and Souza, 2008). Further physiological studies corroborate with Dudley as they observe mammalian auditory system has specialized sensitivity to amplitude modulation of narrowband acoustic signals (Atlas and Shamma, 2003). Following this development, various studies ranging from speech perception to psychoacoustics point to the fact that speech quality and intelligibility mainly reside in the slow changing modulation information (Atlas and Shamma, 2003; Schimmel, 2007; Schimmel and Atlas, 2005). From the viewpoint of speech enhancement, these findings indicate that the slow changing envelope (modulator) of the carrier frequency is the key component in preserving speech intelligibility.

Paliwal et al. were the first to extend the short time Fourier transform (STFT) analysis-modification-synthesis (AMS) framework to the modulation domain (Paliwal et al., 2010). In the AMS framework, the modulation spectrum is given by the STFT of the envelope of the short time frequency bin, which carries short time information of the envelope as a function of time, frequency and modulation frequency. The short time spectrum represents the short time spectral content of the speech signal akin to the shape of the vocal tract (Paliwal et al., 2010; Wu et al., 2011b). The short time modulation spectrum on the other hand captures the temporal cues, which describes the time evolution of the vocal tract. As mentioned, it is precisely this temporal information that relates the most to speech intelligibility (Atlas and Shamma, 2003; Schimmel, 2007; Vinton and Atlas, 2001). Clearly, the modulation domain processing compactly represents the evolution of spectral-temporal information of speech. Favourable results have been reported in speech enhancement applications via the AMS framework (Wojcicki and Loizou, 2012; Paliwal et al., 2012; Schwerin and Paliwal, 2014; Wang and Brookes, 2018). Parallel developments in automatic speech recognition research show a clear distinction between speech and noise features in the modulation domain (Greenberg and Kingsbury, 1997; Hermansky, 2011). These findings led to further development in modulation based automatic speech recognition (ASR) system (You and Alwan, 2009; Sun and Lee, 2012; Moritz et al., 2011). The usefulness of modulation spectrum has also been extended to speech emotion recognition as modulation spectrum carries the signals long-term temporal patterns, a perceptual cue used by listeners themselves (Wu et al., 2011b).

Of late sparse reconstruction methods such as compressed sensing (CS) (Donoho, 2006; Candés and Wakin, 2008) have been applied in speech enhancement. CS theory states that sparse signals with a small set of linear measurements can be reconstructed with an overwhelming probability (Candés et al., 2006; Candés and Tao, 2006). Various CS

☆ This research was supported by the Fundamental Research Grants Scheme (FRGS 2015-1) under the Malaysian Ministry of Higher Education (MOHE).
E-mail address: sy.low@soton.ac.uk.

https://doi.org/10.1016/j.specom.2018.08.003
Received 2 October 2017; Received in revised form 28 June 2018; Accepted 5 August 2018
Available online 07 August 2018

based methods with favorable results have been reported (Low et al., 2013; Sreenivas and Kleijn, 2009; Wu et al., 2011a), demonstrating its popularity for speech enhancement applications. The general idea behind compressive speech enhancement lies in the CS strength to maintain only the sparse components (speech) and its weakness in preserving the non-sparse components such as noise. The main assumption is that whilst speech is fairly compact and dense in the time domain, they are in fact sparse in the time-frequency representations (Pham et al., 2009; Gardner and Magnasco, 2006). This is because speech signal rarely excites all frequency components at any one time and there will be lapses of time-frequency periods where the speech power is negligible compared to the average power (Singh et al., 2018; Davis et al., 2006), which makes it sparse. Unlike speech, background noise is omnipresent and is thus generally non-sparse.

Similar to the time-frequency domain, one direct consequence of the modulation domain is that it tends to increase the sparsity of the signal representation. Modulation domain gives a compact representation of the temporal speech dynamics, which is bounded by the physiological limit of how fast the vocal tract can change. As such, the modulation speech spectrum accentuates the sparsity of speech dynamics as speech excitation can be considered to be a spiky excitation of a quasi-periodic nature (Giacobello et al., 2012). In fact, a compactness study of speech in the modulation domain shows the energy of the modulation coefficients mainly reside in the low modulation bands (Nilsson et al., 2007). Further studies in the modulation spectrum demonstrate that speech and noise have distinct modulation characteristics, which could be exploited in speech discrimination or segregation applications (You and Alwan, 2009; Sephus et al., 2013; Bentsen et al., 2016).

Coupled with the increased sparsity in the modulation spectrum and its importance in speech intelligibility, this paper sets out to investigate the use of modulation domain in sparse reconstruction. The main research question here is to ascertain if the modulation spectrum is indeed more "compressible", giving rise to a sparser representation. If so, will the sparse reconstruction error for a sparser representation be smaller? The paper first examines the sparsity of speech in the modulation domain through the notion of compressibility. The results are then used to show that the sparse reconstruction error in the modulation domain is indeed upper bounded by the reconstruction error in the time-frequency domain. By using the compressive speech enhancement system in Low et al. (2013) as a case example, this paper demonstrates that the modulated approach provides improved performance for compressive speech enhancement in terms of segmental signal to noise ratio (SNR), perceptual evaluation of speech quality (PESQ) (Rix et al., 2001; P.862, 2001) and the short-time intelligibility improvement measure (STOI) (Taal et al., 2011a) for a wide range of SNRs and different types of noise.

2. Compressive speech enhancement

2.1. Introduction

As mentioned, CS states that super-resolved signals and images can be reconstructed from far fewer measurements than the Nyquist sampling (Candés, 2006). This is based on the assumption that the signals involved have a sparse representation in one basis, which can be recovered from a few projections onto another incoherent basis. By sparseness, it means that the majority of the signal measurements concentrate in the neighbourhood of some baseline value. In most literature, this baseline value is set to zero. However, such definition is not always sufficient because a sparse signal may have a baseline value other than zero (Karvanen and Cichocki, 2003). In point of fact, many sparse signals are "compressible" when expressed in the proper basis. This means that CS allows for sampling right at the signal actual intrinsic information rate, with very little redundancy.

2.2. Signal model in the modulation domain

Dudley observed that speech signals are low bandwidth processes, which modulate the higher bandwidth carriers (Dudley, 1939). As such, speech signals can be described as a summation of amplitude modulated narrow frequency bands spanning the full signal bandwidth. The speech signal $s(n)$ can then be represented as

$$s(n) = m(n)c(n) \tag{1}$$

where $m(n)$ is the signal's modulator and $c(n)$ is the signal's carrier. Equivalently, in the short-time frequency domain

$$S(\omega, k) = \mathcal{M}(\omega, k)*C(\omega, k) \tag{2}$$

where * denotes the convolution operator, $\mathcal{M}(\omega, k)$ and $C(\omega, k)$ are the frequency representations of the modulator and carrier at frequency ω and time instant k, respectively. From Eq. (2), $\mathcal{M}(\omega, k)$ is a slowly varying temporal modulation spectrum, which modulates the carrier signal, $C(\omega, k)$. Studies show $C(\omega, k)$ characterizes the fine structure of the signal, whilst $\mathcal{M}(\omega, k)$ carry information involving both segmental and suprasegmental, which contribute to the overall speech intelligibility (Paliwal et al., 2010). Clearly, the amplitude or the envelope of the temporal modulation spectrum holds the modulation frequency components, which have been well linked to the perception of speech quality and speech intelligibility. The envelope $m(n)$ is given as

$$m(n) = \mathcal{D}_{\mathsf{ENV}}\{s(n)\} \tag{3}$$

where $\mathcal{D}_{\mathsf{ENV}}\{\cdot\}$ represents the envelope detector operator. Correspondingly, the $\mathcal{N}$-point short time Fourier transform (STFT) representation of the envelope at time instant k and frequency ω is

$$\begin{aligned}\mathcal{M}(\omega, k) &= \mathcal{D}_{\mathsf{ENV}}\{S(\omega, k)\} \\ &= \mathcal{D}_{\mathsf{ENV}}\left\{\sum_{n=0}^{N-1} s(n)w(n - kR_1)e^{-j\omega n}\right\}.\end{aligned} \tag{4}$$

The time-limited window $w(n - kR_1)$ is with a hop size of R_1, length $\mathcal{N}$, $\omega \in \omega_0, ..., \omega_{\mathcal{N}-1}$ and k is the time index in the short-time frequency domain.

Hence, the short-time modulation spectrum at the lth time instant and ν modulation frequency of Eq. (1) for acoustic frequency ω is

$$\begin{aligned}S_{\mathsf{MOD}}(\omega, \nu, l) &= \mathfrak{M}\{s(n)\} \\ &= \mathfrak{F}\{\mathcal{M}(\omega, k)\} \\ &= \mathfrak{F}\{\mathcal{D}_{\mathsf{ENV}}\{S(\omega, k)\}\} \\ &= \sum_{k=0}^{\mathcal{K}-1} |S(\omega, k)|w(k - lR_2)e^{-j\nu l}\end{aligned} \tag{5}$$

where $\mathfrak{M}\{\cdot\}$ and $\mathfrak{F}\{\cdot\}$ denotes the modulation transform and Fourier transform operators, respectively and $|\cdot|$ is the absolute value operator. The time-limited window $w(n - kR)$ is now with a hop size of R_2, length $\mathcal{K}$ and the modulation frequency $\nu \in \nu_0, ..., \nu_{\mathcal{K}-1}$.

Similarly, let the noisy signal, $x(n)$ be

$$x(n) = s(n) + v(n), \tag{6}$$

where $s(n)$ and $v(n)$ are the speech and noise signals, respectively. Then, the short-time modulation representation of the noisy signal is

$$\begin{aligned}x_{\mathsf{MOD}}(\omega, \nu, l) &= \mathfrak{M}\{x(n)\} \\ &= \mathfrak{F}\{\mathcal{D}_{\mathsf{ENV}}\{X(\omega, k)\}\} \\ &= \sum_{k=0}^{\mathcal{K}-1} |X(\omega, k)|w(k - lR_2)e^{-j\nu l}.\end{aligned} \tag{7}$$

Eqs. (5) and (7) show that the modulation representation is equivalently defined as computing the STFT of the envelopes of a signal's frequency representation. In other words, the modulation information can be obtained by taking a STFT on the envelope of the signal's spectrum. Studies have shown modulation envelope frequencies between 1 − 16Hz carry the most speech information as they reflect the

syllabic rate of speech (Wojcicki and Loizou, 2012).

2.3. Modulated compressive sensing

Consider a $N \times N$ matrix $\mathbf{\Psi}(\omega, \nu)$ whose columns form an orthonormal basis. A $\mathcal{P}$-sparse signal in the modulation domain, $\mathbf{x}(\omega, \nu, l) \in \mathbb{R}^N$ can then be defined as

$$\mathbf{x}_{\text{MOD}}(\omega, \nu, l) = \Psi(\omega, \nu)\theta(\omega, \nu, l), \tag{8}$$

where the N-length envelope vector is defined as

$$\mathbf{x}_{\text{MOD}}(\omega, \nu, l) = [|x(\omega, \nu, l)|, ..., |x(\omega, \nu, l - N + 1)|]^T, \tag{9}$$

where $[\cdot]^T$ is the transposition operator and $\theta(\omega, \nu, l) \in \mathbb{R}^N$ has $\mathcal{P}$ non-zero entries, hence the term $\mathcal{P}$-sparse. Simply, Eq. (8) assumes that given the basis function $\mathbf{\Psi}(\omega, \nu)$, the signal $\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$ can be decomposed in terms of the coefficient vector, $\theta(\omega, \nu, l)$ and the basis function as $\mathbf{x}_{\text{MOD}}(\omega, \nu, l) = \Psi(\omega, \nu)\theta(\omega, \nu, l)$. The compressed measurement vector or the CS measurements in the modulation domain is described by

$$\mathbf{y}_{\text{MOD}}(\omega, \nu, l) = \Phi(\omega, \nu)\mathbf{x}_{\text{MOD}}(\omega, \nu, l), \tag{10}$$

where $\mathbf{\Phi}(\omega, \nu)$ is a $M \times N$ sensing matrix at frequency, ω and modulation band, ν. The sensing matrix, $\mathbf{\Phi}(\omega, \nu)$ compresses the signal's envelope for each modulation frequency ν at frequency ω by making the measurement as incoherent as possible to reduce redundancy, hence the term CS measurements. Since modulation domain is sparse, Eq. (10) allows sparse signals to be readily sampled close to their intrinsic information rate as redundancy is expected to be less. Here, the sensing matrix, $\mathbf{\Phi}(\omega, \nu)$ is said to satisfy RIP of order $\mathcal{P}$ for all $\mathcal{P}$-sparse signal, $\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$, if there exists a constant, $\delta_{\mathcal{P}} \in (0, 1)$ such that

$$\begin{aligned}(1 - \delta_{\mathcal{P}})\|\mathbf{x}_{\text{MOD}}(\omega, \nu, l)\|_2^2 &\le \|\Phi(\omega, \nu)\mathbf{x}_{\text{MOD}}(\omega, \nu, l)\|_2^2 \\ &\le (1 - \delta_{\mathcal{P}})\|\mathbf{x}(\omega, \nu, l)\|_2^2,\end{aligned} \tag{11}$$

where $\|\cdot\|_2$ denotes ℓ_2 norm. Note that $\delta_{\mathcal{P}}$ is the smallest number from the set (0,1), which satisfies Eq. (11). The RIP condition in Eq. (11) must also be satisfied by all the submatrices of $\mathbf{\Phi}(\omega, \nu)$. Candes et al. have shown that the tractable recovery of $\mathcal{P}$-sparse signal, $\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$ from the measurements, $\mathbf{y}_{\text{MOD}}(\omega, \nu, l)$ requires the sensing matrix, $\mathbf{\Phi}(\omega, \nu)$ to obey the restricted isometry property (RIP) (Candés and Tao, 2006).

Property 1. *Define signals $\mathbf{x}_{\mathcal{P}_1}$ and $\mathbf{x}_{\mathcal{P}_2}$ with sparsity $\mathcal{P}_1$ and $\mathcal{P}_2$, where $\mathcal{P}_1 < \mathcal{P}_2$, respectively. If a matrix $\mathbf{\Phi}$ satisfies the RIP condition (11) for all $\mathbf{x}_{\mathcal{P}_2}$ with a small constant $\delta_{\mathcal{P}_2}$, then the matrix $\mathbf{\Phi}$ would also hold the RIP property for $\mathbf{x}_{\mathcal{P}_1}$.*

Proof of Property 1... Since $\mathbf{\Phi}$ satisfies the $\mathcal{P}_2$-RIP condition for all $\mathbf{x}_{\mathcal{P}_2}$ with a small constant, $\delta_{\mathcal{P}_2}$, then all its submatrices $\Phi_{\mathcal{P}_2}$ are well conditioned. Given that $\mathcal{P}_1 < \mathcal{P}_2$, then $\Phi_{\mathcal{P}_1} \subset \Phi_{\mathcal{P}_2}$, thus $\mathbf{\Phi}$ holds for $\mathcal{P}_1$-RIP. □

A $\mathcal{P}$-RIP condition ensures all submatrices of $\mathbf{\Phi}$ are close to an isometry and thus information preserving (Baraniuk et al., 2010). Property 1 suggests that the sparser the signal is, the more its distance is preserved. In fact, most practical sparse recovery algorithms require a stronger RIP condition to preserve the distance. Conversely, if the same signal can be transformed into a sparser representation, then it is akin to making the RIP condition stronger as shown in Property 1.

2.4. CS Recovery

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

$$\begin{aligned}&\hat{\mathbf{x}}_{\text{MOD}}(\omega, \nu, l) = \arg\min_{\mathbf{x}(\omega,\nu,l)} \|\mathbf{x}_{\text{MOD}}(\omega, \nu, l)\|_0 \\ &\text{s.t.} \quad \mathbf{y}_{\text{MOD}}(\omega, \nu, l) = \Phi(\omega)\mathbf{x}_{\text{MOD}}(\omega, \nu, l)\end{aligned} \tag{12}$$

where $\|\mathbf{x}_{\text{MOD}}(\omega, \nu, l)\|_0$ is the number of non-zero components of $\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$. However, solving (12) requires a combinatorial search, which is NP-hard (Gill et al., 2011). A computational tractable formulation, which allows for a trade-off between the exact congruence of $\mathbf{y}(\omega, k) = \Phi(\omega, \nu)\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$ and a sparser $\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$ is the popular basis pursuit denoising (Kim et al., 2007) given as

$$\begin{aligned}\hat{\mathbf{x}}_{\text{MOD}}(\omega, \nu, l) = \arg\min_{\mathbf{x}_{\text{MOD}}(\omega,\nu,l)} &\|\mathbf{y}_{\text{MOD}}(\omega, \nu, l) - \Phi(\omega, \nu)\mathbf{x}_{\text{MOD}}(\omega, \nu, l)\|_2^2 \\ &+ \lambda(\omega, \nu)\|\mathbf{x}_{\text{MOD}}(\omega, \nu, l)\|_1\end{aligned} \tag{13}$$

where $\|\cdot\|_2$ is the l_2-norm and $\lambda(\omega, \nu)$ is the regularization parameter. The formulation in (13) is a simple least-squares minimization process with a l_1-norm penalizer and the dictionary matrix $\mathbf{\Phi}(\omega, \nu)$. It is worth noting that since l_1-norm is non-differentiable, the optimization leads to a decomposition which is sparser (Chen et al., 2001). However, it remains not so straightforward to set the regularization parameter $\lambda(\omega, \nu)$. Note that the optimal solution tends to trivial as $\lambda(\omega, \nu) \to \infty$ (Kim et al., 2007). Typically, $\lambda(\omega, \nu)$ is fixed and not optimized. The regularizer, $\lambda(\omega, \nu)$ penalizes the sum of the observed signal. A higher value of $\lambda(\omega, \nu)$ would generally result in a sparser solution since the l_1-norm is being penalized more heavily. Nevertheless, a sparser solution may not guarantee an improved speech quality. Moreover, the sparsity of the signal varies as a function of frequency and the regularizer should vary according to the signal's profile.

Recall that CS speech enhancement makes use of its sparse recovery to retrieve only the sparse components. Since speech is sparse in nature, CS automatically acts as a denoiser in the sense that only sparse components such as speech is recovered in its output. It has been shown in Low et al. (2013) that the SNR in the compressed domain is greater or equal than the uncompressed domain, which demonstrates the CS noise suppression capability, i.e., compressive speech enhancement (CSE). In CSE, each of the signal's frequency envelope representation is in a time-scale dictionary. The task at hand is to obtain the sparsest set of envelope representation for each frequency ω, which will be synthesized via inverse STFT for the fullband representation. The next section investigates the sparsity of the modulation and the frequency domains via the compressibility measure.

3. Compressibility in the modulation domain

3.1. Compressibility versus sparsity

A signal is considered to be $\mathcal{P}$-sparse if $\mathcal{P}$ out of I coefficients of a signal x are nonzero where $\mathcal{P} \ll I$ (Hurley and Rickard, 2009). Real-world signals may not strictly fit into the definition of sparsity but can be approximated via its compressibility (Duarte et al., 2009). In this instance, the notion of signal compressibility arises as an approximation to signal sparsity. Candes et al. (Candes et al., 2006; Candes, 2008) showed that for compressible signals with its $\mathcal{P}$ largest coefficients chosen ($\mathcal{P}$-sparse), the reconstruction error is almost the same as if the full information about the signal is available. The findings led to the interchangeability of compressibility and sparsity, which paves way for sparse signal reconstruction in a wide range of real-world settings, such as speech signals.

The compressibility of a signal is closely related to the decay rate of its coefficients via the power laws (Candés et al., 2006; Duarte et al., 2009). Consider a signal ϑ, whose I coefficients are sorted in the order of decreasing magnitude. According to the power law, ϑ would then decay as

$$|\vartheta_{\mathcal{I}}(i)| \le C_1 i^{-\tau} \tag{14}$$

where $i = 1, ..., I$, $\mathcal{I}(i)$ indexes the sorted coefficients and C_1 is a constant, which depends on τ, with $\tau \ge 1$. Clearly, the larger τ is, the more compressible the signal is. This is because most of its coefficients would decay rapidly to zero and the signal can then be represented by only its $\mathcal{P}$ largest coefficients, where $\mathcal{P} \ll I$. The error of the approximation can

be measured as follows (Duarte et al., 2009; Candes et al., 2006)

$$\sigma_{\mathcal{P}}(\vartheta) := \arg\min_{\overline{\vartheta}\in\Sigma_{\mathcal{P}}} \left\|\vartheta - \overline{\vartheta}\right\|_2 = \|\vartheta - \vartheta_{\mathcal{P}}\|_2 \tag{15}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm. The error bound for (15) is then given as (Candes et al., 2006; Baraniuk et al., 2011)

$$\sigma_{\mathcal{P}}(\vartheta) \le C_2 \mathcal{P}^{-r} \tag{16}$$

where C_2 is a constant and $r > 0$, which are dependent on C_1 and τ. It has been shown that $\sigma_{\mathcal{P}}(\vartheta)$ will decay as $\mathcal{P}^{-r}$ if and only if the sorted coefficients $|\vartheta_I(i)|$ decay as $i^{-r+\frac{1}{2}}$ (Baraniuk et al., 2011). Plainly, the more compressible a signal is, the sparser the signal is.

Most stable sparse reconstruction algorithms would rely on the restricted isometry property (RIP) (Candes et al., 2006) and thus require at least the following compressive measurements

$$M = O(\mathcal{P}\log(N/\mathcal{P})) \tag{17}$$

where $\mathcal{P} << N$ is the length of the original signal observations. The relationship suggests, the more compressible a signal is (i.e., sparser), the less the measurement samples are required for its reconstruction.

Property 2. *Define* $M_1 = O(\mathcal{P}_1\log(N/\mathcal{P}_1))$ *and* $M_2 = O(\mathcal{P}_2\log(N/\mathcal{P}_2))$. *If* $\mathcal{P}_1 < \mathcal{P}_2$, *where* $\mathcal{P}_2 < \arg\max_{\mathcal{P}_2}\left\{O\left(\mathcal{P}_2\log\frac{N}{\mathcal{P}_2}\right)\right\}$, *then*

$$M_1 < M_2. \tag{18}$$

Proof of Property 2... Eq. (17) is an increasing function, hence if $\mathcal{P}_1 < \mathcal{P}_2$ then $M_1 = O(\mathcal{P}_1\log(N/\mathcal{P}_1)) < M_2 = O(\mathcal{P}_2\log(N/\mathcal{P}_2))$ until (17) hits the maximum turning point at $\arg\max_{\mathcal{P}_2}\left\{O\left(\mathcal{P}_2\log\frac{N}{\mathcal{P}_2}\right)\right\}$. Given that $\mathcal{P}_2 << N$, then it is reasonable to assume $\mathcal{P}_2 < \arg\max_{\mathcal{P}_2}\left\{O\left(\mathcal{P}_2\log\frac{N}{\mathcal{P}_2}\right)\right\}$. □

Property 2shows that a sparser signal would naturally require less compressive measurements for its subsequent reconstruction. In other words, the sparser the signal is, the better it satisfies the RIP requirement for sparse reconstruction.

3.2. Modulation - A more compressible domain?

Whilst speech is compact in its time representation, they are sparse when transformed into the time-frequency representations (Pham et al., 2009; Gardner and Magnasco, 2006). This is because speech signal rarely excites all frequency at the same time and there will be lapses of time-frequency periods where the speech power is negligible compared to the average power. Owing to sparsity, sparse speech representation is naturally discriminative (Panagakis et al., 2009; Wright et al., 2009). This suggests the sparser the signal is, the more discriminative the signal becomes. The sparsity assumption in the time-frequency has been extensively used in various speech applications with much success, ranging from localization (Panagakis et al., 2009; Wright et al., 2009), blind speech separation (Zhang and Zhao, 2013) to compressive speech enhancement (Low et al., 2013).

Whilst modulation domain has been well documented for speech intelligibility improvement, its sparsity property is only marginally explored. Yi et al. (Zhang and Zhao, 2013) were the first to demonstrate that modulation spectrum yields a higher overall sparsity score compared to the time-frequency domain. The increased sparsity is attributed to the fact that modulation domain accentuates the speech dynamics as speech excitation can be viewed as a spiky excitation of a quasi-periodic nature (Zhang and Zhao, 2013). The very different dynamics exhibit by speech across the modulation spectrum clearly suggests a sparser domain compared to the time-frequency.

In the same vein, this paper seeks to ascertain the compressibility of modulation spectrum as signal compressibility is a more practical assumption to sparsity. A simple experiment was conducted to measure the compressibility of the speech signals in the time domain, frequency domain and the modulation domain. Sixty female and male speech utterances from the TIMIT database were used to measure the average compressibility. The coefficients in the respective domain were ordered in decreasing magnitude and the compressibility index measures the percentage of coefficients which has decayed down to a threshold, e.g., $\epsilon = 0.01$. The larger the decay rate, the greater the compressibility and thus, the sparser a signal is.

Table 1
The compressibility level of speech in time, frequency and modulation domains across the different SNRs (babble).

	Percentage ofdecayed coefficients			
SNR	0dB	10dB	20dB	Clean
Time domain	10.8%	26.3%	51.3%	59.2%
Frequency domain	9.2%	38.6%	60.9%	64.3%
Modulation domain	52.3%	75.7%	78.3%	78.6%

Table 1shows the percentage of decayed coefficients for the female and male utterances in all the three domains for different SNRs with babble noise. Across the SNRs, it is evident that the compressibility level of the modulation domain is the highest followed by the frequency domain, with time domain the least compressible. However, when the SNR drops to 0dB, the compressibility level of both the frequency and time domains become similar at around 10%. Modulation domain still maintains its compressibility level with more than 50% of its coefficients decayed down to the threshold.

It is interesting to note that whilst the time domain and frequency domain coefficients only have 10% of decayed coefficients, the modulation domain decayed more than half of its coefficients for SNR of 0dB. This suggests, modulation still maintains part of its sparsity property even in noisy situations. The results conclusively show that signals in the modulation are better approximated by their $\mathcal{P}$-sparse signals compared to the time and frequency domains, owing to the rapid decay of their coefficients. In other words, modulation domain is indeed sparser compared to its frequency domain counterpart.

Property 3. *The RIP constant for the modulation,* δ_{MOD} *is upper bounded by the RIP constants of the frequency domain* δ_{FRE},

$$\delta_{\mathrm{MOD}} < \delta_{\mathrm{FRE}} \tag{19}$$

Proof of Property 3... Let the sparsity of the modulation and frequency domains be $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. The monotonicity property of the RIP constant states that for any two integers such that $A \le A'$, then their respective RIP constants satisfy $\delta_A \le \delta_{A'}$ (Dai and Milenkovic, 2009). From Property 1, $\Phi_{\mathcal{P}_1} \subset \Phi_{\mathcal{P}_2}$, thus $\mathcal{P}_1 < \mathcal{P}_2$. Then it follows from the monotonicity property that their respective RIP constants satisfy $\delta_{\mathcal{P}_1} < \delta_{\mathcal{P}_2}$. The property shows that the RIP constant for the modulation is upper bounded by the RIP constant of the frequency domain, which is a direct consequence of its increased sparsity. □

Properties 1, 2 and 3 point to the fact that modulation domain gives a stronger RIP condition for sparse speech recovery with high probability compared to the frequency domain. In particular, Properties 2 and 3 clearly show that a smaller value of RIP constant in the modulation requires less samples to guarantee a stable sparse recovery. Equivalently, given the same signal in a sparser transformation domain, the better the sparse recovery will be. This is analogous to preconditioning the signal via an appropriate representation, i.e., a more compact form, which offers better performance for processing.

4. Performance bound of the modulation sparse recovery

4.1. Error in sparse reconstruction

Section 3details that modulation domain is sparser than the frequency domain and hence it provides a better sparse recovery. However, it remains to show that the sparse reconstruction error of the modulation domain is indeed upper bounded by the sparse

reconstruction error in frequency domain. Consider a modulation transformed signal $\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$ at frequency ω, modulation frequency ν and time index l. In practice, the exact sparsity $\mathcal{P}_{\mathbf{x}}$ of $\mathbf{x}_{\text{MOD}}(\omega, \nu, l)$ is unknown. If $\mathcal{P}_{\mathbf{x}}$ is known, then the measurement samples M could be set to $\mathcal{P}_{\mathbf{x}}$, for maximal compression. Typically, the sparse reconstruction assumes a $\mathcal{P}$-sparse reconstruction for the signal, where $\mathcal{P}_{\mathbf{x}} < \mathcal{P} < M \leq N$. As shown in Stankovic and Stankovic (2015), for the case where $\mathcal{P}_{\mathbf{x}} > M$, there will be an error term resulting from the non-sparse reconstruction. Readers interested in non-sparse reconstruction may refer to Sytankovic (2015). For ease of exposition, the indices (ω, ν, l) will be suppressed in the ensuing variables.

Assuming all the reconstruction conditions are fulfilled, then the error, $\sigma_{\mathcal{P}}$ between the $\mathcal{P}$-sparse reconstructed signal $\mathbf{y}_{\text{MOD}}$ and the $\mathcal{P}_{\mathbf{x}}$-sparse desired signal $\mathbf{x}_{\mathcal{P}_{\mathbf{x}}}$ can be derived as (Stankovic and Stankovic, 2015; Stankovic et al., 2016)

$$\sigma_{\mathcal{P}} = \left\| \mathbf{y}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 = \mathcal{P}\frac{N-M}{MN} \left\| \mathbf{x}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2. \tag{20}$$

Eq. (20) attests to the fact that sparse reconstruction error is directly proportional to the sparsity difference in the observed and the desired signals. This is because the $\mathcal{P}$-sparse reconstruction detects the $\mathcal{P}$ largest components and uses them to reconstruct the output signal, $\mathbf{y}_{\text{MOD}}$. Naturally, if the sparsity difference is high, then those large components become more difficult to be detected. In an ideal situation where the signal $\mathbf{x}_{\text{MOD}}$ is exactly $\mathcal{P}_{\mathbf{x}}$-sparse, i.e., $\mathbf{x}_{\text{MOD}} = \mathbf{x}_{\mathcal{P}_{\mathbf{x}}}$ then the following holds

$$\left\| \mathbf{y}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 = 0. \tag{21}$$

Eq. (21) indicates that if the sparsity of the incoming signal can be known in advance, then the measurement matrix can be designed to compress the signal exactly to its sparsity level or intrinsic information. Note that the error term in Eq. (20) can be generalized to signals with unknown sparsity but also non-sparse (Stankovic et al., 2016). Evidently, non-sparse signals will naturally yield a higher error term as the sparsity of a non-sparse signal is far greater than $\mathcal{P}$.

4.2. Upper error bound for modulation sparse reconstruction

This section derives the upper error bound for modulation sparse reconstruction. As demonstrated in the previous section, the sparse reconstruction error is directly linked with the sparsity level of the signal.

Theorem 1. *Denote* $\mathbf{y}_{\text{FRE}}$ *and* $\mathbf{y}_{\text{MOD}}$ *as the reconstructed sparse signal representation from the frequency and modulation transforms, respectively. Let the ideal* $\mathcal{P}_{\mathbf{x}}$*-sparse signal be* $\mathbf{x}_{\mathcal{P}_{\mathbf{x}}}$*. The error term for a* $\mathcal{P}$*-sparse reconstruction in the modulation domain is then upper bounded by the error term to that of in the frequency domain as*

$$\begin{aligned} \left\| \mathbf{y}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 &< \left\| \mathbf{y}_{\text{FRE}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 \\ \sigma_{\text{MOD},\mathcal{P}} &< \sigma_{\text{FRE},\mathcal{P}}. \end{aligned} \tag{22}$$

Proof of Theorem 1.. Following (20), the error term for the reconstructed frequency transformed signal, $\sigma_{\text{FRE},\mathcal{P}}$ can be expressed as

$$\left\| \mathbf{y}_{\text{FRE}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 = \mathcal{P}\frac{N-M}{MN} \left\| \mathbf{x}_{\text{FRE}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2. \tag{23}$$

Let the sparsities of $\mathbf{x}_{\text{FRE}}$ and $\mathbf{x}_{\text{MOD}}$ be $\mathcal{P}_2$ and $\mathcal{P}_1$, respectively, where $\{\mathcal{P}_1, \mathcal{P}_2\} < \mathcal{P}$. It follows from Section 3 and Properties 1–3, the sparsities of the frequency transformed signal, $\mathcal{P}_2$ and modulated signal, $\mathcal{P}_1$ satisfy the following

$$\mathcal{P}_1 < \mathcal{P}_2. \tag{24}$$

Since $\left\| \mathbf{x}_{\text{FRE}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 = \sum_{i=\mathcal{P}_2+1}^{N} |NA_i|^2$ and $\left\| \mathbf{x}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 = \sum_{i=\mathcal{P}_1+1}^{N} |NA_i|^2$, where A denotes the signal's coefficients, it is straightforward to see from Eq. (24) that $\left\| \mathbf{x}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 < \left\| \mathbf{x}_{\text{FRE}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2$. Thus, the following relationship follows fromEqs. (20), (23) and (24)

$$\begin{aligned} \mathcal{P}\frac{N-M}{MN} \left\| \mathbf{x}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 &< \mathcal{P}\frac{N-M}{MN} \left\| \mathbf{x}_{\text{FRE}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 \\ \left\| \mathbf{y}_{\text{MOD}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 &< \left\| \mathbf{y}_{\text{FRE}} - \mathbf{x}_{\mathcal{P}_{\mathbf{x}}} \right\|_2^2 \\ \sigma_{\text{MOD},\mathcal{P}} &< \sigma_{\text{FRE},\mathcal{P}}. \end{aligned} \tag{25}$$

□

Theorem 1states that the error of the $\mathcal{P}$-sparse reconstruction in the modulation domain is upper bounded by the error for the frequency domain sparse reconstructed signal. However, note that if both transformed signals have the sparsity to that of the ideal signal, then the error term reduces to zero. In practice, it is not straightforward to know the sparsity of the signal exactly. Typically, the sparsity level for a $\mathcal{P}$-reconstruction is assumed from empirical observations, such that $\mathcal{P} < M$. Intuitively, the notion of sparsity for practical signals rests on the compressibility of a signal. If the given signal is less compressible, then the same can be deduced for its sparsity. Naturally, a less compressible representation has more terms for its approximation and as shown in Theorem 1 the reconstruction error will be higher.

Table 2
A comparison of the eigenvalue spread of speech in the frequency and modulation domains.

	Average eigenvalue spread			
SNR	0dB	10dB	20dB	Clean
Frequency Domain	1.01×10^{15}	0.91×10^{15}	0.87×10^{15}	1.69×10^{15}
Modulation Domain	3.61×10^{15}	3.60×10^{15}	3.58×10^{15}	5.11×10^{15}

4.3. A note on the eigenvalue spread

A sparser modulation domain can also be viewed as a more compact feature space akin to a sparse low rank model, which in theory occupies the same dimension as the original observation. This compactness can be measured in terms of eigenvalue spread as a sparse presentation would have a low rank covariance matrix and hence a high eigenvalue spread.

To quantify the compactness between frequency and modulation domains, consider the covariance matrix of the frequency data, $\mathbf{x}_{\text{FRE}}$ with sparsity $\mathcal{P}_2$ as follows

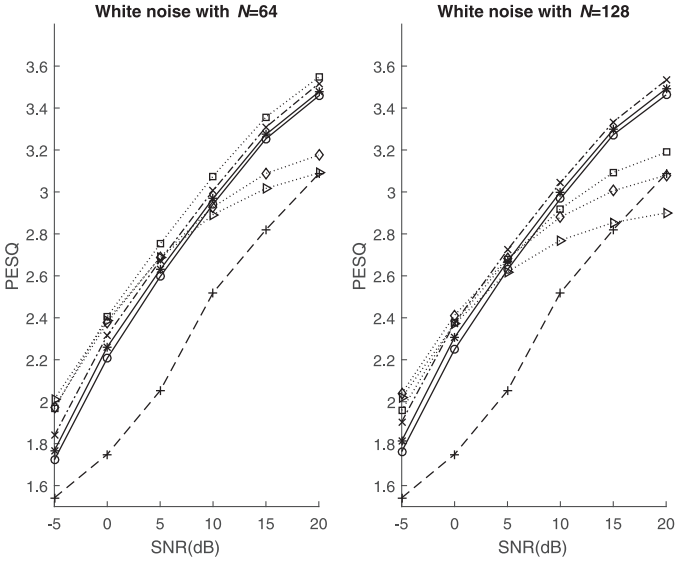

Fig. 1. The PESQ score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in white noise.

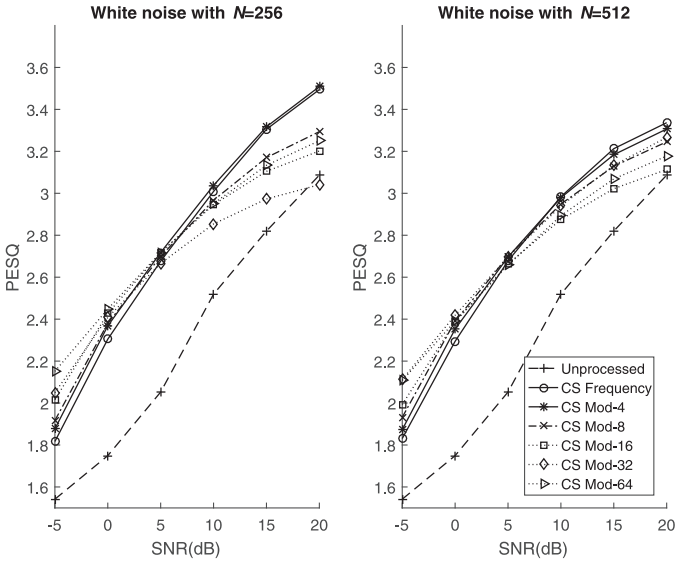

Fig. 2. The PESQ score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in white noise.

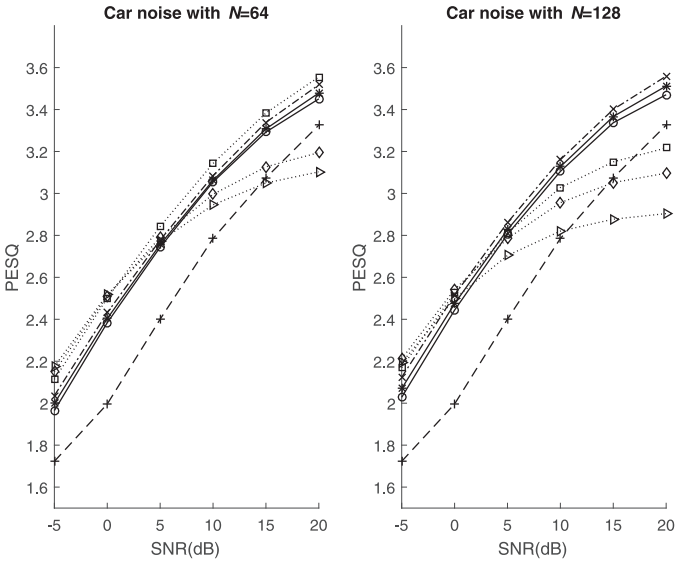

Fig. 3. The PESQ score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in car noise.

$$\mathbf{R}_{\mathbf{x}_{\text{FRE}}} = \mathbf{x}_{\text{FRE}}\mathbf{x}_{\text{FRE}}{}^{H} \tag{26}$$

where $(\cdot)^H$ denotes the Hermitian transposition. Since the modulation domain is sparser than the frequency domain, then the covariance matrix of the modulation domain can be viewed in the form of principal submatrix of $\mathbf{R}_{\mathbf{x}_{\text{FRE}}}$. Assume that the modulation representation is $\mathcal{P}_1$-sparse. Then $\mathbf{R}_{\mathbf{x}_{\text{MOD}}}$ is the $\mathcal{P}_1 \times \mathcal{P}_1$ principal submatrix of $\mathbf{R}_{\mathbf{x}_{\text{FRE}}}$ (Moghaddam et al., 2006). This is possible since $\mathcal{P}_1 < \mathcal{P}_2$, and $\mathbf{R}_{\mathbf{x}_{\text{MOD}}}$ is obtained by deleting the rows and columns corresponding to the zero indices.

From Thompson (1992), it has been proven that the eigenvalue spread of the principal submatrix is greater than the full matrix itself. Thus, it is straightforward to see that

$$\text{Sp}(\mathbf{R}_{\mathbf{x}_{\text{MOD}}}) \geq \text{Sp}(\mathbf{R}_{\mathbf{x}_{\text{FRE}}}) \tag{27}$$

where $\text{Sp}(\cdot)$ denotes the eigenvalue spread of the matrix, i.e., the difference between its largest and smallest eigenvalues. The greater eigenvalue spread in the modulation from Eq. (27) reveals that modulation domain is indeed more "compact" compared to the frequency domain. By using subspace as an analogy, a greater eigenvalue spread suggests a greater demarcation between speech and noise, which was first mooted by Zhang and Zhao (2013). From the redundancy viewpoint, a more compact representation translates to less redundant information, which in turn reduces the error term in its sparse reconstruction as shown in Theorem 1.

For numerical illustrations, an experiment was conducted on the values of eigenvalue spread in the frequency and modulation domains. A total of ten speech signals from the TIMIT database were used to get the average eigenvalue spread in both domains. Table 2 shows the eigenvalue spread of the covariance matrix for both domains with $\mathcal{N} = 256$ and $\mathcal{K} = 8$, which corresponds to the modulation domain being sampled at 4Hz. For a comparison purpose, the eigenvalue spread was calculated for every covariance matrix at each frequency and averaged to get the mean eigenvalue spread in both domains. Numerical results demonstrate the validity of Eq. (27), with the eigenvalue spread of the modulation domain consistently higher than the frequency domain by at least three and a half times across the different SNRs (with babble noise). The results also reaffirms the difference in the eigenvalue spread is more pronounced in the clean speech compared to the noisy speech, as that is the case for the greatest demarcation with no noise in the background.

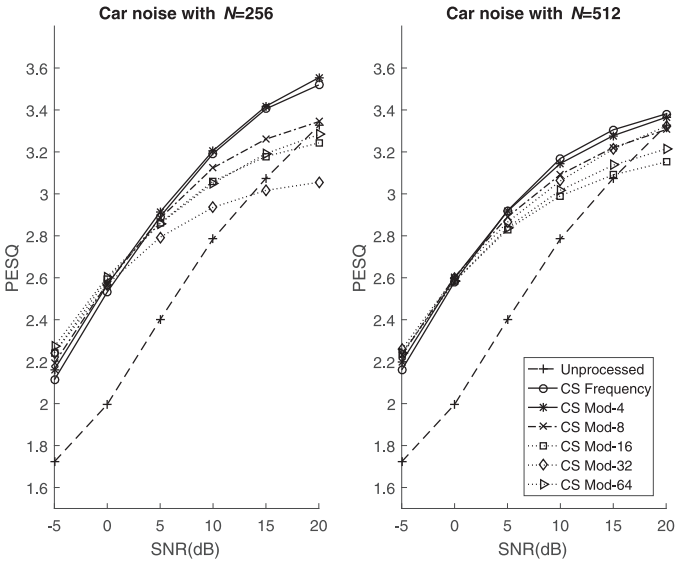

Fig. 4. The PESQ score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in car noise.

5. Experiments and discussions

5.1. Experiment settings

A total of forty speech signals, twenty males and twenty females from the TIMIT database were used as stimuli. Each speech sequence was sampled at 8 kHz with a length of 6.25 s. The short-time Fourier transform with varying number of frequency points and four times over-sampling was used to realize both the frequency and modulation envelopes. Four different kinds of noise from the NOISEX database namely babble noise, white noise, car noise and train noise were chosen to represent the varying stationarity of noise encountered in the real world. Babble noise would be the most non-stationary followed by train noise and car noise. White noise in this case is the most stationary but it is omnipresent across the signal's spectrum. In the following experiments, noise was artificially added to the clean speech to a range of SNR levels for the evaluation of the proposed method.

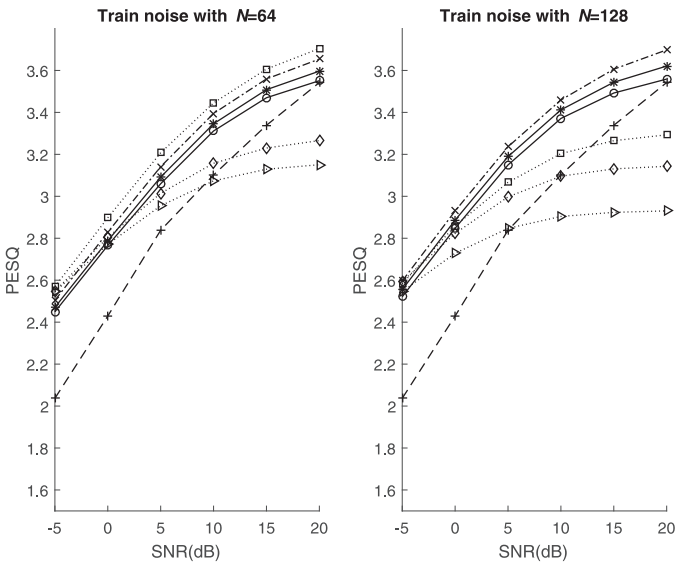

Fig. 5. The PESQ score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in train noise.

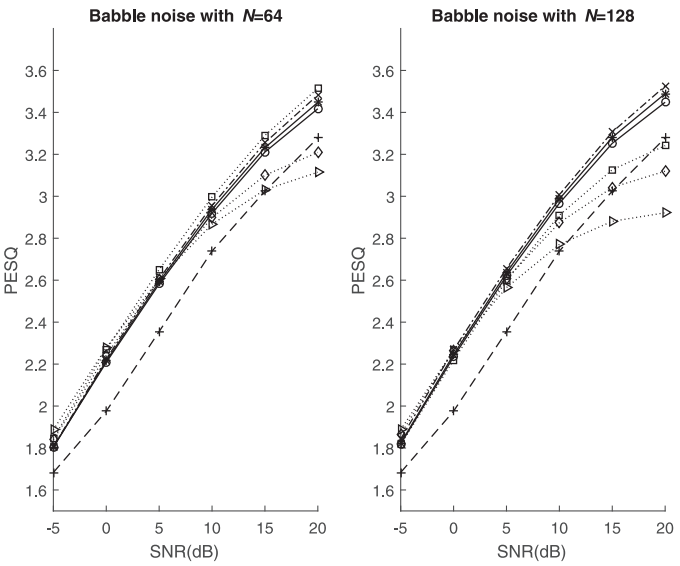

Fig. 7. The PESQ score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in babble noise.

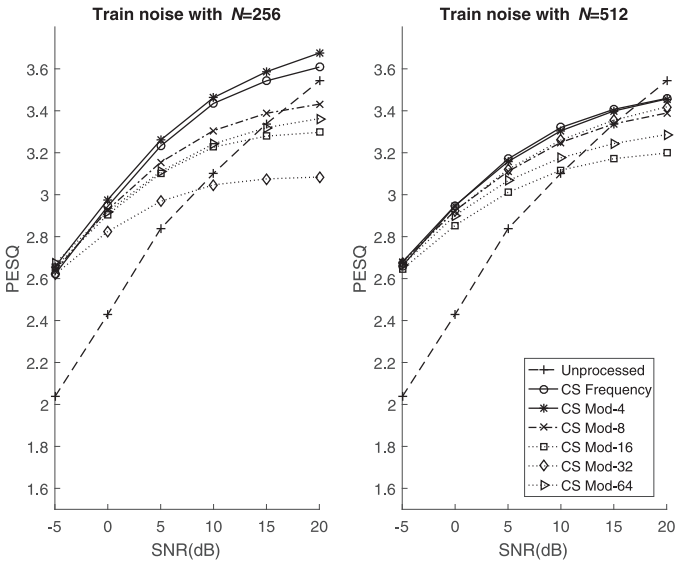

Fig. 6. The PESQ score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in train noise.

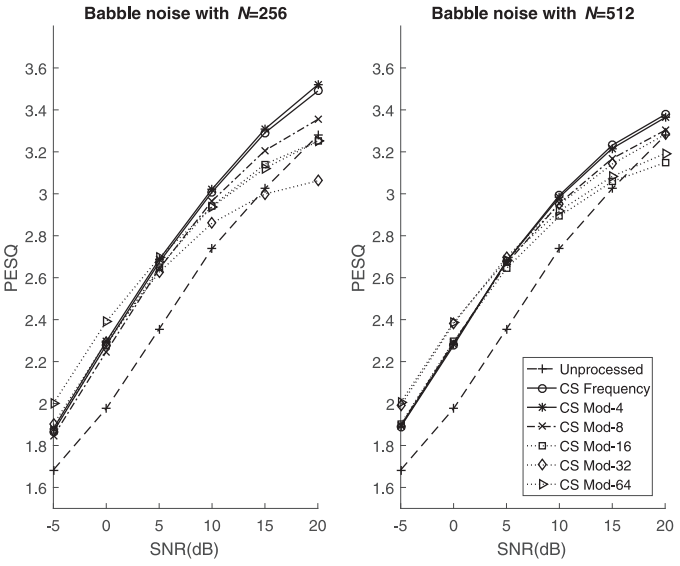

Fig. 8. The PESQ score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in babble noise.

5.2. Performance measures

The performance of the proposed system was evaluated in terms of its overall quality, the amount of noise suppression and the overall intelligibility improvement by using the PESQ (ITU, 2000), the STOI (Taal et al., 2011b) and the segmental SNR (Loizou, 2007), respectively. PESQ is an automated computation algorithm developed to replace human subjects in the evaluation of the mean opinion score (MOS). PESQ measures the perceived speech quality, i.e., aspects of pleasantness or naturalness of speech (Benesty et al., 2005). STOI, on the other hand, has been developed to measure the intelligibility of speech, i.e., how comprehensible or understood is the speech in question. The conventional segmental SNR is the averages of the measurements of SNR over short frames. Essentially, segmental SNR represents how well the noise has been suppressed, with speech distortion factored in. Each of these measures represents different performance aspects of a speech enhancement system. For instance, for face to face communication such as hearing aids, STOI or PESQ would be a more suitable measure compared to segmental SNR. In the case of speech recognition systems, the segmental SNR would be the preferred measure.

5.3. Results and discussions

5.3.1. Modulation band selection in terms of PESQ

Figs. 1–8 show the PESQ scores for the case of noisy speech in white noise, car noise, train noise and babble noise for different modulation frequency points, respectively. The PESQ scores were plotted against varying number of subbands for the frequency points and modulation frequency points. The results show varying PESQ performance across the different subbands combination. Clearly, the performance of the

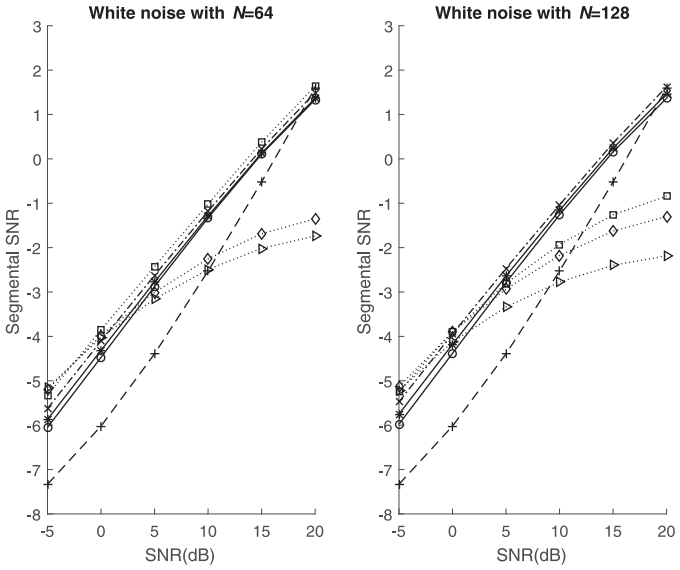

Fig. 9. The segmental SNR for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in white noise.

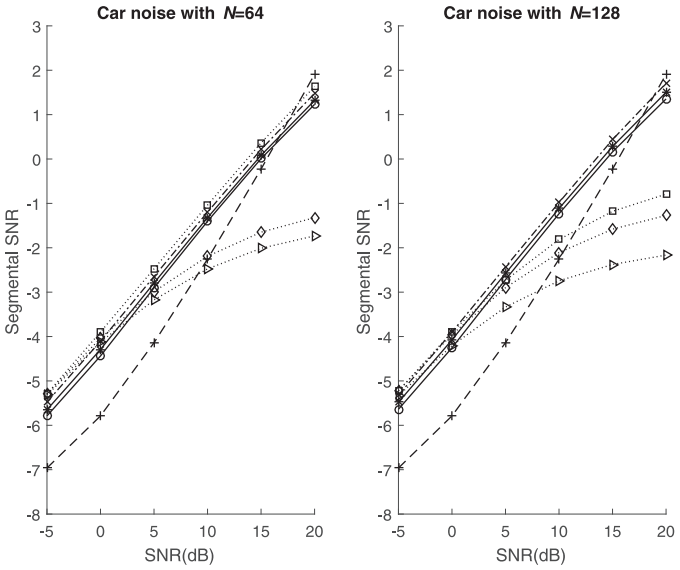

Fig. 11. The segmental SNR for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in car noise.

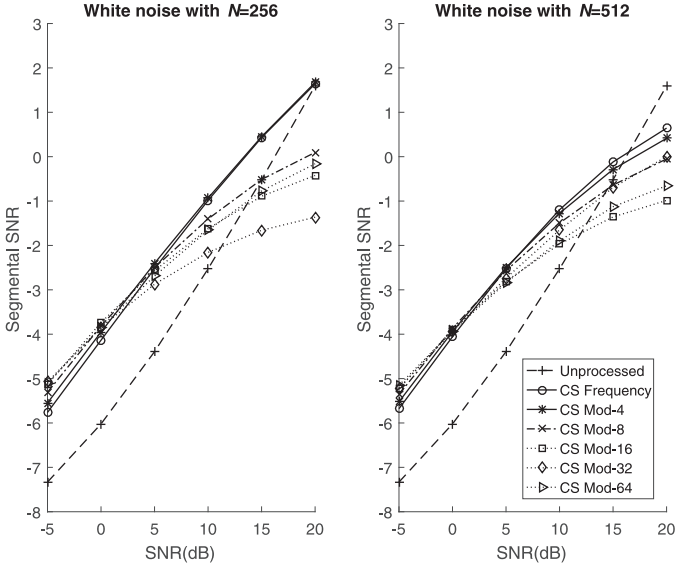

Fig. 10. The segmental SNR for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in white noise.

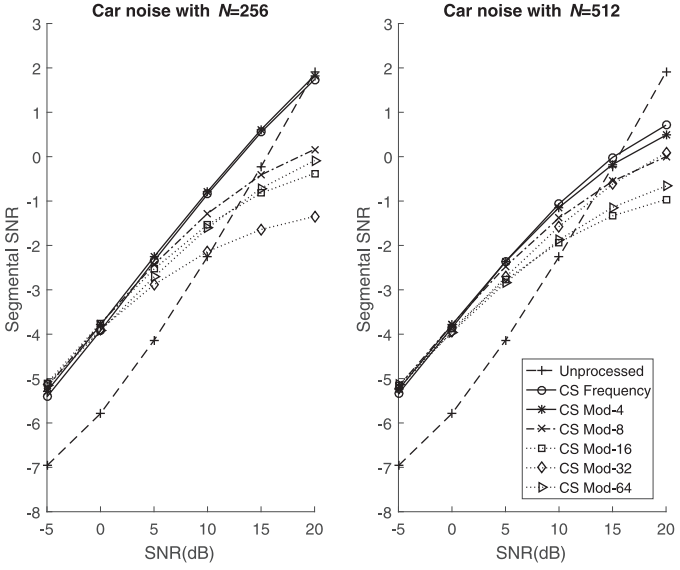

Fig. 12. The segmental SNR for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in car noise.

modulation domain compressed sensing outperforms to that of the frequency domain. For example, consider the case of car noise with SNR= −5dB, the PESQ improves by approximately 30% compared to only 17% improvement in the frequency domain. In other words, the performance in the modulation domain almost doubles to that of the frequency domain. Across the different noise types, the PESQ improvement for a specified value of $\mathcal{N}$, the best performance is achieved for $\mathcal{K}$, which corresponds to processing the modulation envelopes at 3.9Hz, 7.8Hz, and 15.6Hz. The results corroborate with previous studies, which suggest that modulation frequency between $4 - 16$Hz to be the most relevant to intelligibility as it reflects the humans' syllables rate (Wojcicki and Loizou, 2012).

5.3.2. *Modulation band selection in terms of segmental SNR*

The segmental SNR performances are shown in Figs. 9–16. The performance in terms of segmental SNR is very similar to the PESQ where best improvement is registered for when the modulation is sampled between 7.8Hz and 15.6Hz. The results show that the performance is very dependent on the selection of the modulation bands, which in this case must match to that of on the modulation sampling range of $4 - 16$Hz. Broadly, as a general rule of thumb, modulation frequencies should be chosen to reflect the syllabic temporal information of speech. This opens up future work for optimizing the appropriate modulation frequency points for different environmental settings, which may have different syllabic temporal rate.

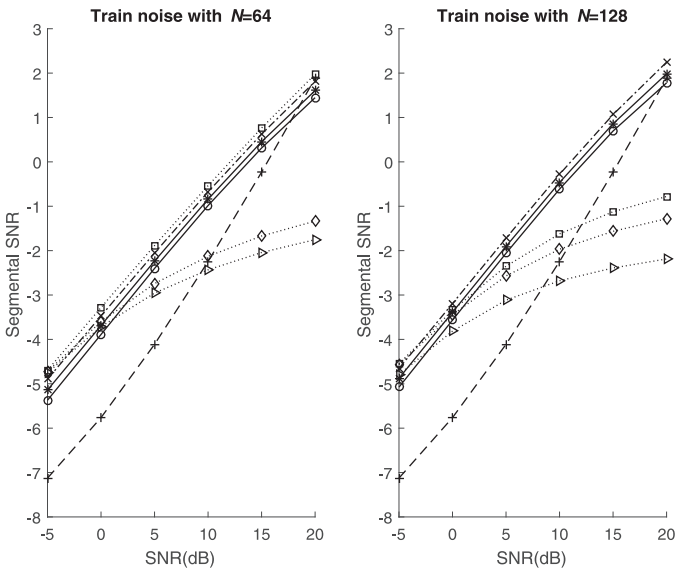

Fig. 13. The segmental SNR for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in train noise.

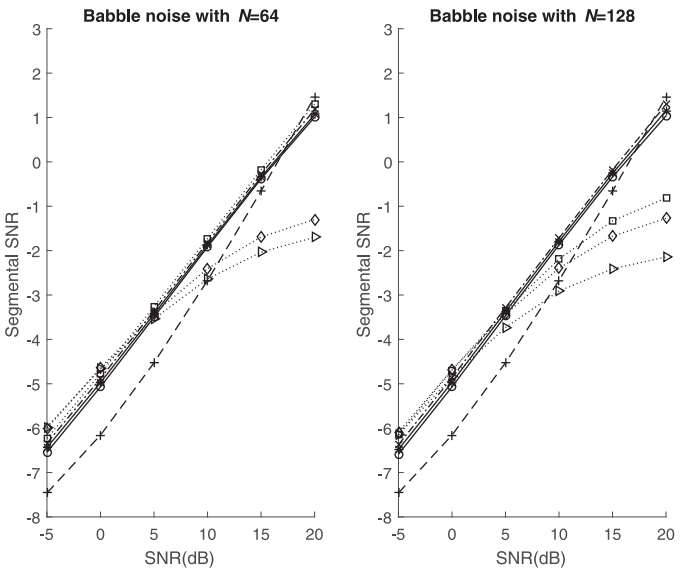

Fig. 15. The segmental SNR for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in babble noise.

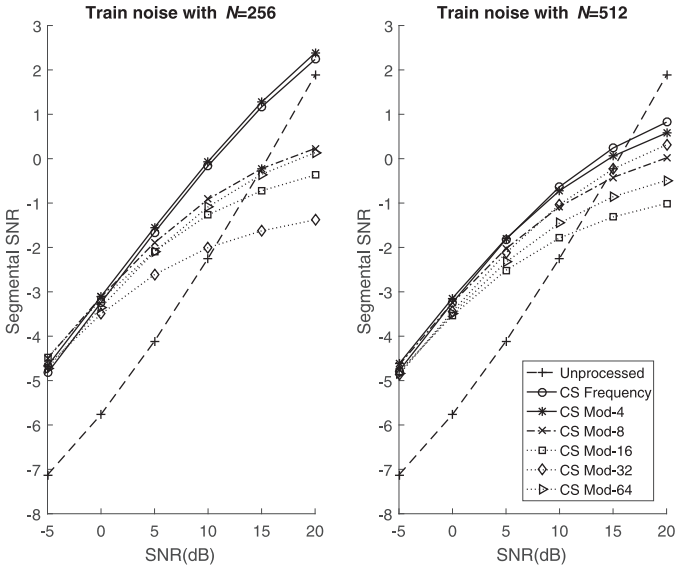

Fig. 14. The segmental SNR for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in train noise.

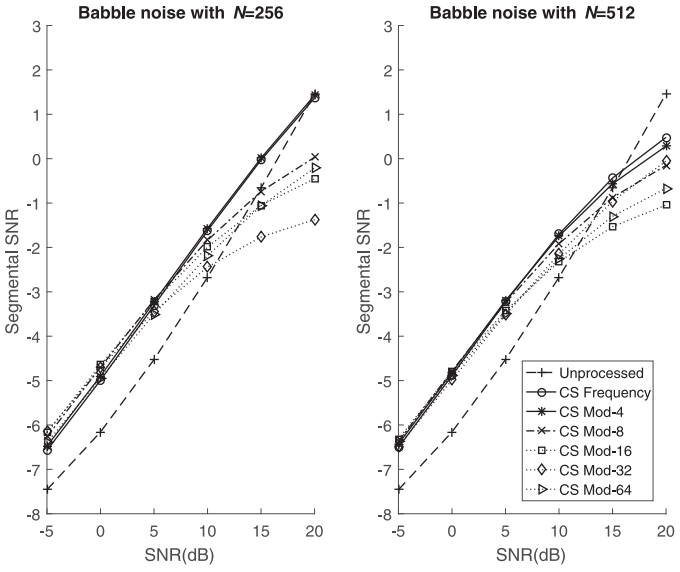

Fig. 16. The segmental SNR for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in babble noise.

5.3.3. Modulation band selection in terms of STOI

Figs. 17–24 show the performance of the compressive speech enhancement for both frequency and modulation domains in terms of STOI. The investigation reveals that improvement follows closely as in the case of PESQ and segmental SNR where the best performance is registered for parameters which, translate to processing the modulation envelopes at $7.8 - 15.6$Hz. The results also show that babble noise achieves the least STOI improvement. This can be understood from the fact that babble noise is highly non-stationary. In a separate study (Low and Yiu, 2017), it has been found that highly non-stationary noise such as babble tend to be more sparse than stationary noise such as car noise. Consequently, babble could be mistaken as a sparse component just like speech. This is especially the case under poor SNR conditions, where both sparse speech and babble noise components may be jointly reconstructed. In general, the performance in terms of STOI is marginal and at times degrading for high SNRs 15dB and above. The performance curves suggest that, very little or none is achieved when the SNR is high.

5.3.4. A note on the modulation band selection

The experimental results suggest an average best performance which corresponds to sampling the modulation envelope at $3.9 - 15.6$ Hz. This is consistent with the fact that most of the speech intelligibility information reside in this particular modulation frequency band (Dudley, 1940). Outside this range, there is a decline in the intelligibility performance. For instance, consider the PESQ, the segmental SNR and the STOI performance for $N = 64$ for the four types of noise. The poorest score is obtained for when the modulation

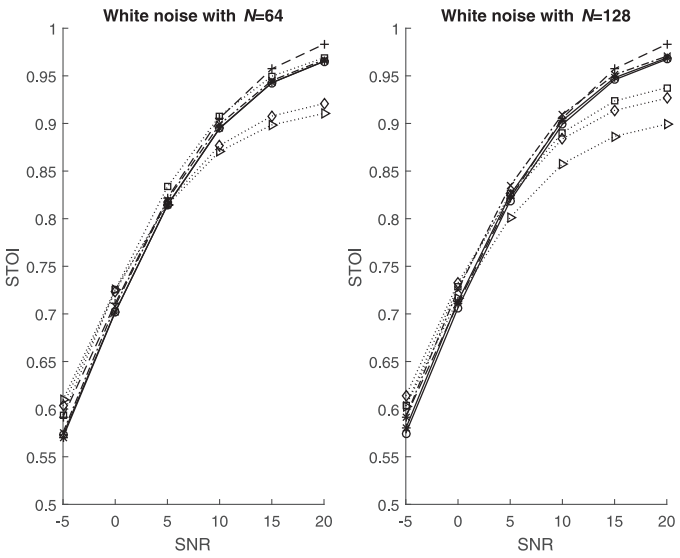

Fig. 17. The STOI score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in white noise.

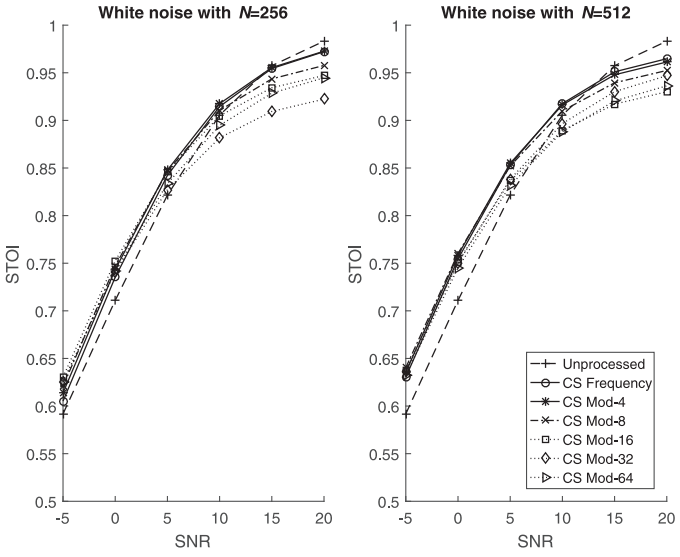

Fig. 18. The STOI score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in white noise.

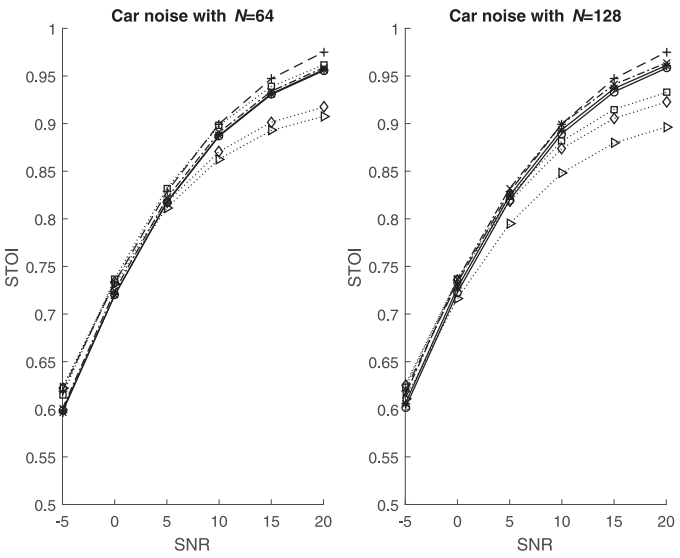

Fig. 19. The STOI score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in car noise.

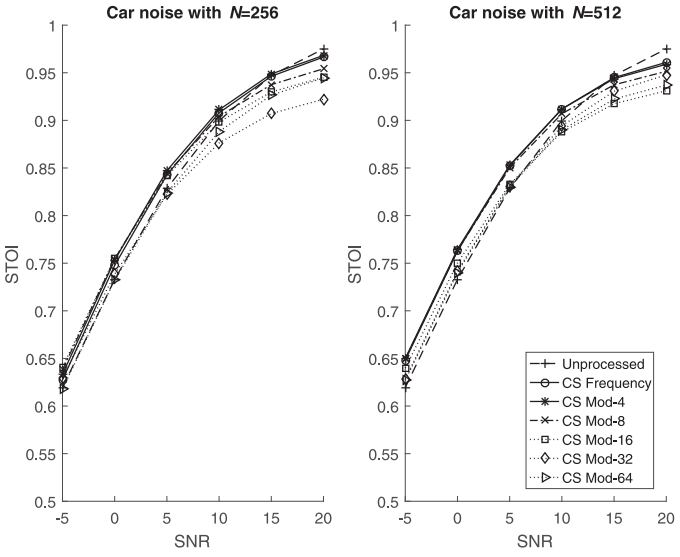

Fig. 20. The STOI score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in car noise.

envelope is sampled at 1.95Hz, i.e., $\mathcal{K} = 64$. Importantly, for speech intelligibility measure, the modulation sampling must be chosen in the aforementioned range. However, for low SNRs, the performance is very marginal. This observation concurs with the previous findings (Low et al., 2013) where for very low SNR such as − 5dB, very little improvement can be achieved as the demarcation between sparse and non-sparse components reduces. The same can be argued for very high SNR i.e., 20dB, where there is hardly any non-sparse components for CSE to suppress.

The performance of the proposed modulation CSE is compared with other reference methods namely, the frequency domain CSE approach (Low et al., 2013), the minimum statistics based Wiener solution (Martin, 2001) and the log minimum mean square error (MMSE) approach (Ephraim and Malah, 1985). For this experiment, the most challenging non-stationary noise, babble was added to the speech corpus across a range of SNRs. Tables 3, 4 and 5 tabulate the PESQ, segmental SNR and the STOI scores for the various methods. It it evident from the results that the proposed modulation CS consistently yield the highest scores across the performance measures and SNRs. Interestingly, the proposed method performed less than MMSE for very poor SNR condition in terms of STOI score. As explained in Low et al. (2013), this is attributed to the fact that under poor SNRs, the sparse speech components may be heavily masked by noise, which makes the discrimination difficult. On the other hand, when the SNR is 20dB, the non-sparse noise content may look very similar to speech, again resulting in very little distinction between the two. Note however, as explained in Low et al. (2013), the proposed modulated CSE will not work for transient noise as the noise shares the same sparsity profile as speech. The proposed method just like its predecessor (Low et al., 2013) will only enhance sparse components. All in all, the experimental

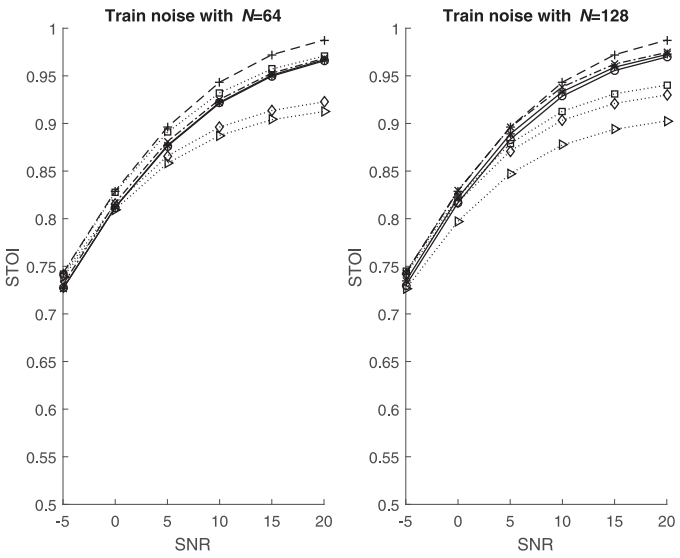

Fig. 21. The STOI score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in train noise.

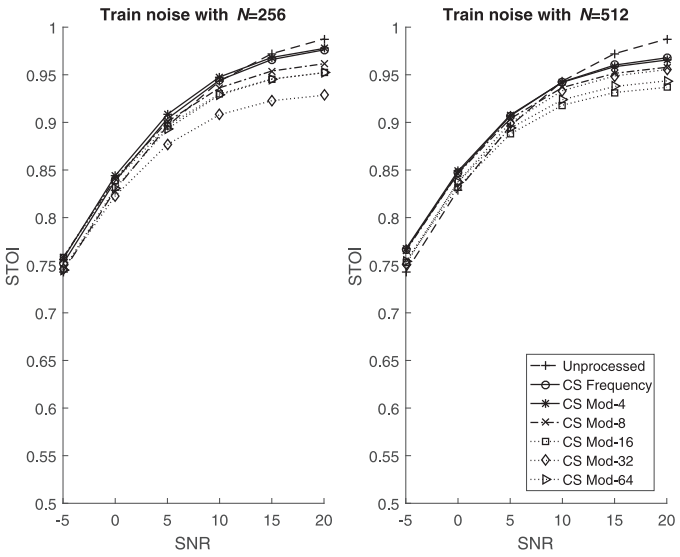

Fig. 22. The STOI score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in train noise.

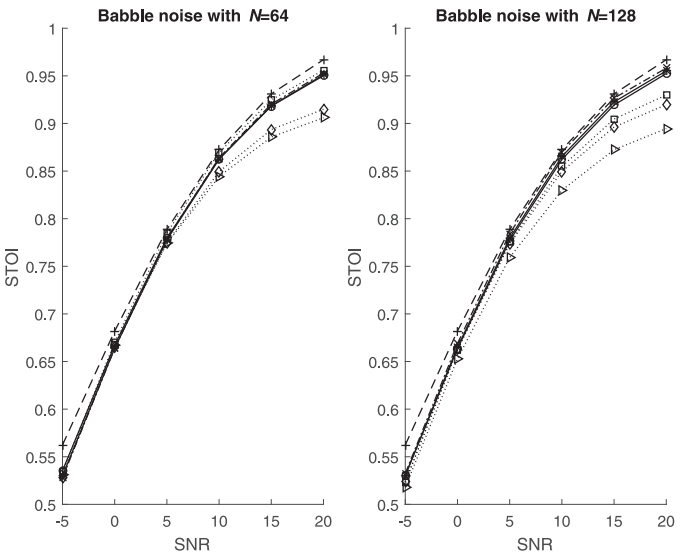

Fig. 23. The STOI score for $N = 64$ and $N = 128$ for different modulation frequency as a function of SNRs in babble noise.

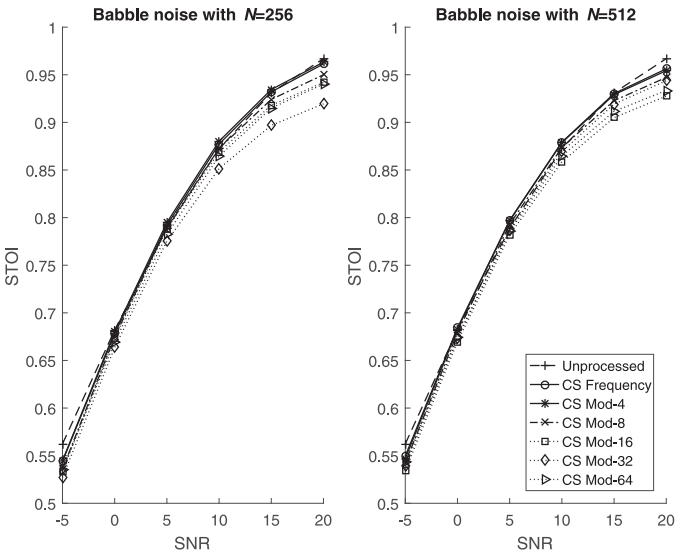

Fig. 24. The STOI score for $N = 256$ and $N = 512$ for different modulation frequency as a function of SNRs in white noise.

results substantiate the performance of the sparser modulation domain for use in the compressive speech enhancement setting.

6. Conclusions

This paper presents an investigation for sparse reconstruction in the modulation domain. First the sparsity of modulation domain is defined via its compressibility. The results reveal that the modulation domain is indeed sparser compared to the frequency domain as it is more compressible. Importantly, the investigation proves that modulation domain offers a stronger RIP condition necessary for sparse recovery with high probability. Further derivation show that the modulation based sparse reconstruction error is upper bounded to that of the frequency domain. The applicability of modulation domain sparse reconstruction

Table 3
The PESQ score of the proposed method versus other reference methods in babble noise for varying SNRs.

Noise type	Babble noise					
SNR	− 5dB	0dB	5dB	10dB	15dB	20dB
Unprocessed	1.8031	2.0611	2.4009	2.8288	3.0834	3.2691
Proposed	**2.0068**	**2.4117**	**2.8037**	**3.0775**	**3.2541**	3.3233
CS Frequency	1.9930	2.3909	2.7908	3.0666	3.2387	3.2827
Wiener	1.8178	2.0995	2.4532	2.8525	3.0783	3.2521
LogMMSE	1.9582	2.3020	2.7368	3.0077	3.2038	**3.3571**

Table 4
The segmental SNR of the proposed method versus other reference methods in babble noise for varying SNRs.

Noise type	Babble noise					
SNR	− 5dB	0dB	5dB	10dB	15dB	20dB
Unprocessed	−8.1200	−7.2171	−6.1007	−4.8309	−3.4542	−1.9948
Proposed	**−7.3974**	**−6.3774**	**−5.2504**	**−4.1067**	**−3.0021**	**−1.9921**
CS Frequency	−7.4519	−6.4321	−5.2989	−4.1447	−3.0406	−2.0513
Wiener	−7.8171	−7.0311	−5.9818	−4.7500	−3.4219	−2.0674
LogMMSE	−7.5869	−6.5948	−5.5718	−4.5684	−3.5961	−2.6461

Table 5
The STOI score of the proposed method versus other reference methods in babble noise for varying SNRs.

Noise type	Babble noise					
SNR	− 5dB	0dB	5dB	10dB	15dB	20dB
Unprocessed	0.6085	0.7305	0.8324	0.9050	0.9505	0.9762
Proposed	0.6000	**0.7365**	**0.8463**	**0.9180**	**0.9587**	**0.9796**
CS Frequency	0.5992	0.7351	0.8447	0.9166	0.9576	0.9787
Wiener	0.6092	0.7334	0.8346	0.9056	0.9501	0.9751
LogMMSE	**0.6108**	0.7349	0.8360	0.9065	0.9501	0.9738

is tested in a compressive speech enhancement scheme and results show improvement across the PESQ, segmental SNR and STOI.

Acknowledgment

S. Y. Low would like to acknowledge the Fundamental Research Grant Scheme (FRGS) under the Ministry of Higher Education (MOHE) Malaysia for the support of the research. The author would also like to thank Chia Cheng Han for the collation of some of the data.

References

Atlas, L., Shamma, S.A., 2003. Joint acoustics and modulation frequency. EURASIP J. Appl. Signal Process. 7, 668–675.
Baraniuk, R.G., Cevher, V., Duarte, M.F., Hegde, C., 2010. Model-based compressive sensing. IEEE Trans. Inf. Theory 56 (4), 1982–2001.
Baraniuk, R.G., Davenport, M.A., Duarte, M.F., Hegde, C., 2011. An Introduction to Compressive Sensing. OpenStax CNX e-textbook.
Benesty, J., Makino, S., Chen, J., 2005. Speech enhancement. Signals and Communication Technology. Springer-Verlag, Berlin. https://www.springer.com/gp/book/9783540240396.
Bentsen, T., May, T., Kressner, A., Dau, T., 2016. Comparing the influence of spectro-temporal integration in computational speech segregation. Interspeech 3324–3328.
Candes, E.J., 2008. The restricted isometry property and its implications for compressed sensing. C.R. Math. 346 (9–10), 589–592.
Candes, E.J., Romberg, J.K., Tao, T., 2006. Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. 59 (8), 1207–1223.
Candés, E., Romberg, J., Tan, T., 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory 52 (2), 489–509.
Candés, E.J., 2006. Compressive sampling. Proceedings of the International Congress of Mathematicians. Madrid, Spain.
Candés, E.J., Tao, T., 2006. Near-optimal signal recovery from randon projections: universal encoding strategies? IEEE Trans. Inf. Theory 52 (12), 5406–5425.
Candés, E.J., 2008. An introduction to compressive sampling. IEEE Signal Process. Mag. 21–30.
Chen, S., Donoho, D., Suanders, M., 2001. Atomic decomposition by basis pursuit. SIAM Rev. 43 (1), 129–159.
Dai, W., Milenkovic, O., 2009. Subspace pursuit for compressive sensing signal reconstruction. IEEE Trans. Inf. Theory 55 (5), 2230–2249.
Davis, A., Nordholm, S., Low, S.Y., Togneri, R., 2006. A multi-decision sub-band voice activity detector. Eur. Signal Process. Conf. (EUSIPCO) 1–5.
Donoho, D.L., 2006. Compressed sensing. IEEE Trans. Inf. Theory 52 (4), 1289–1306.
Duarte, M. F., Hegde, C., Cevher, V., Baraniuk, R. G., 2009. Recovery of compressible signals in unions of subspaces. In: Proceedings of the Conference on Information Sciences and Systems, 175–180.
Dudley, H., 1939. Remaking speech. J. Acoust. Soc. Am. 11 (2), 169–177.
Dudley, H., 1940. The carrier nature of speech. Bell Syst. Tech. J. 19 (4), 495–515.
Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-33, 443–445.
Gallun, F., Souza, P., 2008. Exploring the role of the modulation spectrum in phoneme recognition. Ear Hear. 29 (5), 800–813.
Gardner, T.J., Magnasco, M.O., 2006. Sparse time-frequency representations. Proc. Natl. Acad. Sci. 103, 6094–6099.
Giacobello, D., Christensen, M.G., Murthi, M.N., Jensen, S.H., Moonen, M., 2012. Sparse linear prediction and its applications to speech processing. IEEE Trans. Audio Speech Lang. Process. 20 (5), 1644–1657.
Gill, P.R., Wang, A., Molnar, A., 2011. The in-crowd algorithm for fast basis pursuit denoising. IEEE Trans. Signal Process. 59 (10), 4595–4605.
Greenberg, S., Kingsbury, B., 1997. The modulation spectrogram: in pursuit of an invariant representation of speech. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1647–1650.
Hermansky, H., 2011. Speech recognition from spectral dynamics. Sadhana 36 (5), 729–744.
Hurley, N., Rickard, S., 2009. Comparing measures of sparsity. IEEE Trans. Inf. Theory 55 (10), 4723–4741.
ITU, 2000. Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU Recommendation P. 862.
Karvanen, J., Cichocki, A., 2003. Measuring sparseness of noisy signals. In: Proceedings of the Symposium of Independent Component Analysis and Blind Signal Separation, 125–128.
Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D., 2007. An interior-point method for large-scale l1-regularized least squares. IEEE J. Sel. Top. Signal Process. 1 (4), 606–617.
Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press, Taylor and Francis, Boca Raton, Florida, USA.
Low, S.Y., Pham, D.S., Venkatesh, S., 2013. Compressive speech enhancement. Speech Commun. 55 (6), 757–768.
Low, S. Y., Yiu, K. F. C., 2017. A study on the compressibility of speech for compressive speech enhancement. In: Proceedings of the InterNoise, 1–7.
Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512.
Moghaddam, B., Weiss, Y., Avidan, S., 2006. Spectral bounds for sparse PCA: exact and greedy algorithms. Adv. Neural Inf. Process. Syst. 18.
Moritz, N., Anemüller, J., Kollmeier, B., 2011. Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 5492–5495.
Nilsson, M., Resch, B., Kim, M. Y., Kleijn, W. B., 2007. A canonical representation of speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 4, 849–852.
P.862, I.-T. R., 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. https://www.itu.int/rec/T-REC-P.862.
Paliwal, K., Wojcicki, K., Schwerin, B., 2010. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun. 52 (5), 450–475.
Paliwal, K.K., Schwerin, B., Wojcicki, K., 2012. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. Speech Commun. 54 (2), 282–305.
Panagakis, Y., Kotropoulos, C., Arce, G. R., 2009. Music genre classification via sparse representations of auditory temporal modulations In: Proceedings of the European Signal Processing Conference, 1–5.
Pham, D.T., El-Chami, Z., Guérin, A., Serviére, C., 2009. Modeling the short time fourier transform ratio and application to underdetermined audio source separation. Lecture Notes in Computer Science 5441/2009 Springer, Berlin.
Rix, A., Beerends, J., Hollier, M., Hekstra, A., 2001. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs In: Proceedings of theIEEE International Conference on Acoustics, Speech, and Signal Processing 2, 749–752.
Schimmel, S., Atlas, L., 2005. Coherent envelope detection for modulation filtering of speech. In: Proceedings of theIEEE International Conference on Acoustics, Speech, and Signal Processing, 221–224.
Schimmel, S.M., 2007. Theory of modulation frequency analysis and modulation filtering, with applications to hearing devices. Technical report. UNIVERSITY OF WASHINGTON.
Schwerin, B., Paliwal, K.K., 2014. Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement. Speech Commun. 58, 49–68.
Sephus, N. H., Lanterman, A. D., Anderson, D. V., 2013. Exploring frequency modulation features and resolution in the modulation spectrum In: Proceedings of the IEEE Digital Signal Processing and Signal Processing Education Meeting, 169–174.
Singh, M.K., Low, S.Y., Nordholm, S., Zang, Z., 2018. Bayesian noise estimation in the modulation domain. Speech Commun. 96, 81–92.

Sreenivas, T. V., Kleijn, W. B., 2009. Compressive sensing for sparsely excited speech signals In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 4125–4128.
Stankovic, L., Stankovic, I., 2015. Reconstruction of sparse and nonsparse signals from a reduced set of samples. ETF J. Electr. Eng. 21 (1), 147–169.
Stankovic, L., Stankovic, I., Dakovic, M., 2016. Nonsparsity influence on the ISAR recovery from reduced data. IEEE Trans. Aerosp. Electron. Syst. 52 (6), 3056–3070.
Sun, L.C., Lee, L.S., 2012. Modulation spectrum equalization for improved robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 20 (3), 828–843.
Sytankovic, L., 2015. Digital Signal Processing. CreateSpace, Amazon, South Carolina USA.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.
Thompson, R.C., 1992. The eigenvalue spreads of a hermitian matrix and its principal submatrices. Linear and Multilinear Algebra 32, 327–333.
Vinton, M., Atlas, L., 2001. A scalable and progressive audio codec In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 3277–3280.
Wang, Y., Brookes, M., 2018. Model-based speech enhancement in the modulation domain. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (3), 580–594.
Wojcicki, K., Loizou, P.C., 2012. Channel selection in the modulation domain for improved speech intelligibility in noise. J. Acoust. Soc. Am. 131 (4), 2904–2913.
Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2), 210–227.
Wu, D., Zhu, W. P., Swamy, M. N. S., 2011a. A compressive sensing method for noise reduction of speech and audio signals In: Proceedings of the IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS), 1–4.
Wu, S., Falk, T.H., Chan, W.Y., 2011. Automatic speech emotion recognition using modulation spectral features. Speech Commun. 53 (5), 768–785.
You, H., Alwan, A., 2009. Temporal modulation processing of speech signals for noise robust ASR. Interspeech 36–39.
Zhang, Y., Zhao, Y., 2013. Modulation domain blind speech separation in noisy environments. Speech Commun. 55 (10), 1081–1099.