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
0167-6393/

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

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

where $\mathcal{D}_{\text{ENV}}\{\cdot\}$ 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

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

The time-limited window $w(n - kR_1)$ is with a hop size of R_1, length $\mathcal{N}$, $\omega \in \omega_0, \ldots, \omega_{\mathcal{N}-1}$ 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

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

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 $\nu \in \nu_0, \ldots, \nu_{\mathcal{K}-1}$.

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

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

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 $1 - 16$Hz carry the most speech information as they reflect the