Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 757–768

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Compressive speech enhancement

Siow Yong Low [a,*], Duc Son Pham [b], Svetha Venkatesh [c]

[a] Curtin University, Sarawak Campus, Miri, Malaysia
[b] Curtin University, Department of Computing, WA, Australia
[c] Deakin University, Centre for Pattern Recognition and Data Analytics, Victoria, Australia

Received 17 August 2012; received in revised form 18 February 2013; accepted 3 March 2013
Available online 26 March 2013

Abstract

This paper presents an alternative approach to speech enhancement by using compressed sensing (CS). CS is a new sampling theory, which states that sparse signals can be reconstructed from far fewer measurements than the Nyquist sampling. As such, CS can be exploited to reconstruct only the sparse components (e.g., speech) from the mixture of sparse and non-sparse components (e.g., noise). This is possible because in a time-frequency representation, speech signal is sparse whilst most noise is non-sparse. Derivation shows that on average the signal to noise ratio (SNR) in the compressed domain is greater or equal than the uncompressed domain. Experimental results concur with the derivation and the proposed CS scheme achieves better or similar perceptual evaluation of speech quality (PESQ) scores and segmental SNR compared to other conventional methods in a wide range of input SNR.

Keywords: Compressed sensing; Speech enhancement; Sparsity

1. Introduction

In many practical applications, speech signal is degraded due to unwanted interference and hence speech enhancement is important to enhance intelligibility and the overall perceptual quality of degraded speech. Speech enhancement can be broadly categorized into multiple channel and single channel approaches (Benesty et al., 2005; Brandstein and Ward, 2001). As the name multi-channel implies, multi-channel solutions require more than one microphone for signal observations. In this case, multi-channel techniques primarily exploit spatial information to separate the signal of interest from other interfering signals. A common method to perform spatial filtering (or beamforming) is to make use of the array geometry to form a beam towards the target signal. This technique has been widely studied and considerable noise suppression is reported (Veen and Buckley, 1988; Davis et al., 2005; Dam et al., 2004). However, beamforming based methods rely on apriori information about the acoustical environments and the target signal localization. This means that beamforming methods may be susceptible to potential modeling errors (e.g., model mismatch error) particularly when the transfer function from source to sensor is difficult to model (Hoshuyama et al., 1999; Low et al., 2002). Some multi-channel approaches, which do not require source localization have been proposed in Lotter et al. (2003) and Low and Nordholm (2005).

Single channel techniques, on the other hand, offer a much more computationally appealing solution since only one microphone is needed (Paliwal et al., 2010; Cohen, 2003; O'Shaughnessy, 2000). Recent advances in single-channel techniques have seen attractive speech enhancement applications such as cochlear implants (Kokkinakis and Loizou, 2008). One popular single channel speech enhancement technique is the spectral subtraction (Boll, 1979). It was originally suggested by Boll and has since

* Corresponding author. Tel.: +60 128708345.
E-mail addresses: siowyong@curtin.edu.my (S.Y. Low), DucSon.Pham@curtin.edu.au (D.S. Pham), svetha.venkatesh@deakin.edu.au (S. Venkatesh).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.03.003

gained huge acceptance due to its simplicity (Paliwal et al., 2010; Yang, 1993; Lu, 2011). Basically, spectral subtraction relies on the assumption that the target signal and noise signal are uncorrelated. Therefore, if the noise spectral component is estimated correctly, the target signal can be enhanced by subtracting the estimated spectral noise from the noisy spectral observations. Typically, a voice activity detector (VAD) is used detect the presence of speech and non-speech periods for the estimation of noise statistics. However, any mis-detection by the VAD will result in erroneous updates, which in turn causes spectral subtraction to become defective. A comprehensive review on single channel techniques can be found in Principi et al. (2010).

Recently, there is a new sampling theory called compressed sensing (CS). CS states that super-resolved signals and images can be reconstructed from far fewer measurements than the Nyquist sampling (Donoho, 2006). Whilst compressed sensing/sparsity learning has been a celebrated theory recently and its applications to images are popular, its applications to speech and audio signals are difficult and limited (Sreenivas and Kleijn, 2009; Jancovic et al., 2012). Most of the recent applications are mainly on linear predictive coding of speech in the residual domain (Giacobello et al., 2012; Griffin et al., 2011) or dictionary design (Christensen et al., 2009) and to the best of the authors' knowledge, none of the applications address speech enhancement.

At the very heart of CS sampling is its sparsity assumption and CS theory shows that sparse signals with a small set of linear measurements can be reconstructed with an overwhelming probability (Candès et al., 2006; Candès and Tao, 2006). Suffice to say, under the presence of both sparse and non-sparse components, only sparse signals will be reconstructed. Interestingly, since speech signals are generally sparse in the time-frequency representation (Pham et al., 2009), while many types of noise is non-sparse, CS may hold the potential as a speech discriminator. This means that CS can be designed to reconstruct only the sparse components (speech) from the mixture of sparse and non-sparse components (noise). Potentially, the speech enhancement process can be made to rely upon the strength of CS to maintain only the sparse components and its weakness in preserving the non-sparse components.

In the spirit of speech enhancement, this paper investigates the feasibility of using CS to perform speech enhancement. The most related work to CS based speech enhancement technique is the wavelet based method by Wu et al. (2011). However, the work in Wu et al. (2011) is mainly empirical and is not supported by a theoretical framework. Most importantly, we have found that the global sparsity model (batch processing) adopted in Wu et al. (2011) is inferior compared to the local sparsity modeling in each subband. In the subsequent section, we also outline the differences between wavelet and STFT transformations. This paper shows that on average, the signal to noise ratio (SNR) of the output of time-frequency CS is greater than the SNR of the original noisy signal. The derivation details that the SNR improvement is directly proportional to the sensing dimension, M and the largest eigenvalue of the observation matrix, $\lambda_{\max}$. Further, the theoretical finding is extended and validated by proposing a CS based algorithmic solution that performs speech enhancement. It is worth mentioning that the CS based method relies only on the sparsity of speech of interest and bypasses the need for noise estimation or a VAD. Whilst the performance of the proposed CS scheme is not exceedingly superior in comparison with other speech enhancement algorithms, it is interesting to note that CS could be readily used for improving the SNR. More importantly, both the perceptual evaluation of speech quality (PESQ) scores and the segmental SNR improvement concur with the theoretical finding. Also, an inherent positive byproduct from this scheme is the reduction of sample measurements as the compression/enhancement is performed on the signals.

The paper is organized as follows. Section 2 provides a general background on CS and the proposed CS speech enhancement scheme is detailed in Section 2. Section 3 presents its performance evaluation and lastly, Section 5 summarizes the findings.

2. Background

2.1. Overview of compressed sensing

CS theory states that if a signal has a sparse representation in one basis then it can be recovered from a small number of projections onto a second basis, which is incoherent with the first (Donoho, 2006; Candès, 2006; Rachlin and Baron, 2008). Therefore, it is possible to reconstruct a signal from far fewer samples or measurements than conventional methods use. In other words, the number of measurements can be much lower than the number of samples needed if the signal is sampled at the Nyquist rate. Such capability brings about the benefits of reduced storage space and transmission bandwidth due to the compression achieved.

At the very heart of this compression capability lie two major assumptions, i.e., sparsity, which pertains to the signals of interest and incoherence, which is related to the sensing modality (Candès and Wakin, 2008). Sparsity refers to the idea that the information rate of a continuous time signal may be much smaller than suggested by its bandwidth. Hence, this assumption can be extended to many natural signals that are sparse or compressible in the sense that they have concise representations when expressed in the proper basis. Incoherence on the other hand expresses the idea that objects having a sparse representation must be spread out in the domain in which they are acquired (Candès and Wakin, 2008; Candès and Tao, 2006).

To define sparsity, let us consider a $N \times N$ matrix $\mathbf{\Psi}$ whose columns form an orthonormal basis. Thus a K-sparse signal, $\mathbf{x}(n) \in \mathbb{R}^N$ can be expressed as