Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 194–207

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Voiced/nonvoiced detection in compressively sensed speech signals

Vinayak Abrol *, Pulkit Sharma, Anil Kumar Sao

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi, India

Received 2 February 2015; received in revised form 16 April 2015; accepted 2 June 2015
Available online 8 June 2015

Abstract

We leverage the recent algorithmic advances in compressive sensing (CS), and propose a novel unsupervised voiced/nonvoiced (V/NV) detection method for compressively sensed speech signals. It attempts to exploit the fact that there is significant glottal activity during production of voiced speech while the same is not true for nonvoiced speech. This characteristic of the speech production mechanism is captured in the sparse feature vector derived using CS framework. Further, we propose an information theoretic metric, for V/NV classification, exploiting the sparsity of the extracted feature using a signal adaptive dictionary motivated by speech production mechanism. The final classification is done using an adaptive threshold selection scheme, which uses the temporal information of speech signals. While existing methods of feature extraction use speech samples directly, proposed method performs V/NV detection in compressively sensed speech signals (requiring very less memory), where existing time or frequency domain detection methods are not directly applicable. Hence, this method can be effective for various speech applications. Performance of the proposed method is studied on CMU-ARCTIC database, for eight types of additive noises, taken from the NOISEX database, at different signal-to-noise ratios (SNRs). The proposed method performs similar or better compared to the existing methods, especially at lower SNRs and this provide compelling evidence of the effectiveness of sparse feature vector for V/NV detection.

Keywords: Voiced/nonvoiced detection; Compressed sensing; Linear prediction; Sparse coding; Dictionary learning

1. Introduction

Compressed sensing (CS) is a radical way of sampling signals at less than the Nyquist rate (Candès and Wakin, 2008). In particular, CS enables us to reconstruct a signal via recovery of its sparse representation from very few measurements using an appropriate dictionary (Tosic and Frossard, 2011). Thus one does not require much memory to transmit or store CS measurements. Moreover, these measurements are robust to degradations such as random perturbations or noise (Donoho, 2006). CS or sparse signal representations have recently drawn much interest in the field of speech processing e.g., speech enhancement (Sharma et al., 2015a), speech synthesis (Sharma et al., 2015b), speech encryption (Zeng et al., 2012), speech recognition (Asaei et al., 2011; Sharma et al., 2015c) and image processing e.g., image super resolution (Mandal et al., 2014), hyperspectral imagery using Gaussian mixture modeling (Yang et al., 2015), etc.

In this paper, we propose a novel unsupervised voiced/nonvoiced (V/NV) method for compressive speech (both clean and noisy) signals. To the best of our knowledge, none of the previous papers have proposed such methods for compressive speech signals. Hence, the proposed method has promising applications, where only compressed speech samples (which require less memory) are available. For instance, it will help in extracting features only from the selected (voiced) region of speech signals,

* Corresponding author.
E-mail addresses: vinayak_abrol@students.iitmandi.ac.in (V. Abrol), pulkit_s@students.iitmandi.ac.in (P. Sharma), anil@iitmandi.ac.in (A.K. Sao).

http://dx.doi.org/10.1016/j.specom.2015.06.001
0167-6393/

which in turn can be used for applications such as speaker verification (Pradhan and Prasanna, 2013).

The proposed approach exploits the fact that there is significant glottal activity (i.e., the vibration of vocal folds) during production of voiced speech. On the other hand for nonvoiced segments, including both silence and unvoiced speech (UV) regions (such as voiceless fricatives and stops), vocal folds do not vibrate (Dhananjaya and Yegnan arayana, 2010; Ananthapadmanabha and Yegnanara yana, 1979). In various speech applications, such classification is preferred mainly because these regions correspond to different production mechanisms (Dhananjaya and Yegnanarayana, 2010). For instance, this distinction is employed in tasks such as telephony (reducing acoustic echo) (Benyassine et al., 1997), speech coding (reducing bandwidth by coding nonvoiced speech with fewer bits) (Yang et al., 1995), speech recognition (Jancovic and Kokuer, 2006; Atal and Rabiner, 1976), robotic aid for persons with disability (Suk et al., 2007), emotion recognition (Koolagudi et al., 2010), speaker verification (Pradhan and Prasanna, 2013) and pitch detection (Ykhlef and Bendaouia, 2012). It is interesting to note that even with very less and random compressed speech samples, one can efficiently capture the specific characteristic (significant glottal activity) of the speech production mechanism via derived sparse vector using the CS framework. In the proposed method, the sparse vector is shown to contain the source characteristics and, hence it shows a distinct behavior for voiced and nonvoiced regions of the speech signal. In order to measure this behavior, we propose an information theoretic measure, which efficiently captures the distribution of the sparse vector components. The final classification is done using an adaptive threshold selection scheme, which exploits the temporal information of the speech signals.

It should be noted that, the proposed method makes an assumption that only compressed measurements of the actual speech signal are available. Hence, the estimation of sparse vector using the framework of CS/sparse coding is very much influenced by the choice of dictionary (Donoho, 2006). We propose a signal-adaptive dictionary based on warped linear predictive (WLP) analysis. The proposed method uses an iterative method based on l_2-norm minimization for estimating both the dictionary and the corresponding sparse representation of the speech signal from its compressed measurements. In addition, the proposed method makes no assumption on the type of noise degrading the speech signal, neither requires clean speech samples to learn the dictionary. Further, note that the proposed V/NV detection method is also directly applicable to speech signals, for which it is easy to choose or learn the dictionary. However, we are interested in V/NV detection especially in compressive speech signals. Also, while dictionary learning, due to inherit denoising advantage of CS/sparse coding framework (Low et al., 2013), the proposed procedure repeatedly reduces the effect of noise in each iteration, allowing a robust estimation of sparse vector, and hence capturing the voiced characteristics efficiently.

1.1. Background and prior work

In general, existing V/NV algorithms (which process raw speech samples) consists of two subtasks: feature extraction and classification (Ramírez et al., 2007). The former task attempts to compute discriminating features for voiced and nonvoiced segments from the given speech signal, while the latter stage employs thresholding or statistical pattern recognition based approaches to give V/NV decisions (Ramírez et al., 2007). A good V/NV detector should be easy to implement, accurate and robust against noise (Dhananjaya and Yegnanarayana, 2010). Among all these, robustness against non-stationary noisy environments is the most difficult objective to accomplish. The existing methods available in the literature demonstrate good performance in the presence of stationary noise, but cannot deal with non-stationary noises (Dhananjaya and Yegnanarayana, 2010).

Researchers in the past have proposed various features, exploiting acoustic properties of voiced speech for V/NV classification. Elementary methods were based on autocorrelation function (ACF), average magnitude difference function (AMDF), line spectral frequencies, zero-crossing rate, full or low-band energy features extracted from the speech signal (Dhananjaya and Yegnanarayana, 2010; Ramírez et al., 2007; Kristjansson et al., 2005; Ykhlef and Bendaouia, 2012). Here V/NV decisions are generally taken based on an empirically chosen threshold or a fixed decision boundary in the space defined by the extracted features (Dhananjaya and Yegnanarayana, 2010). A major problem with these methods is in selecting a suitable value of threshold, which dictates the performance of V/NV detection. Moreover, the performance of these methods severely degrades in low SNR conditions (Ramírez et al., 2007). These issues have been addressed by various supervised/unsupervised classification methods, where the decision boundary is learned via pattern recognition/machine learning approaches (Atal and Rabiner, 1976; Li et al., 2005; Shahnaz et al., 2006; Arifianto, 2007). In addition, these methods employ acoustic features, which are more robust in noisy environments such as spectrum features (Kristjansson et al., 2005; Prasanna et al., 2009), mel-frequency cepstral coefficients and delta line spectral frequencies (Kinnunen et al., 2007). These methods are computationally expensive, and are more popular for voice activity detection (VAD) or speech activity detection (SAD), which distinguishes between voiced, unvoiced and silence (V–UV–S) regions. V/NV classification requires much less complexity as compared to V–UV–S classification, hence although not preferable, all the VAD methods can also be optimized for V/NV detection.

In supervised methods, classification is generally performed using Bayesian (Mousazadeh and Cohen, 2013), support vector machines (SVM) (Kinnunen et al., 2007)