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)

and multi-layer perceptrons (MLP) classifier (Ng et al., 2012). To build efficient models these methods require a large amount of labeled training data. On the contrary, unsupervised methods do not require training samples of the speech signal and/or noise (Ying et al., 2011; Ahmadi and Spanias, 1999). Here, speech detection is generally performed by first clustering the extracted feature from the noisy speech signal into speech and non-speech classes using a predefined criteria, such as the average background noise levels (estimated from silence regions of speech) (Shi et al., 2006; Sadjadi et al., 2013). Next an optimal decision threshold is decided for classification on the test data. In order to further improve the performance, some methods learn statistical models after initial clustering is done (Ying et al., 2011). In addition, few methods extract features after using noise suppression as a part of the process e.g., use of enhanced speech spectra derived from Wiener filtering based on estimated noise statistics (Soon et al., 1999). Due to advantages such as no requirement of labeled training data, less training time and computational complexity, unsupervised methods are preferred over supervised methods (Sadjadi et al., 2013).

Recent works in Teng and Jia (2013) and You et al. (2012) have proposed methods based on sparse coding for VAD rather than V/NV detection. These methods are discussed here mainly due to their similarity in using a sparse coding framework (but on speech signal directly). In these methods, the noise-robustness of VAD algorithm is achieved by features extracted from a noise-suppressed representation of noisy speech signal using sparse decomposition over an overcomplete dictionary. However, both the methods require several examples of clean speech to learn the dictionary. Moreover, the method in Teng and Jia (2013) based on non-negative matrix factorization (NMF), employs a separate dictionary for noise examples, which is not suitable in practical scenarios, as generally the type of noise is not known *a priori*. Method in Teng and Jia (2013) uses the maximum and mean value, while method in You et al. (2012) uses power spectrum energy of the estimated sparse representation as a VAD feature, which might not be robust in low SNR conditions, and in the presence of highly non-stationary noises having sparse or speech like characteristics.

The major contributions of the proposed work are: (1) We propose a novel sparse feature to estimate glottal activity from compressive speech signals. (2) Propose a metric to measure the unique characteristic of the sparse feature vector for voiced and nonvoiced regions of the speech signal. (3) Propose a signal adaptive dictionary motivated by the speech production mechanism for V/NV detection. (4) A dynamic adaptive threshold scheme, which exploits the temporal behavior of the speech signals, is proposed for V/NV classification. (5) Extensive experiments are performed to demonstrate the performance of the proposed V/NV method under different noisy conditions (stationary and non-stationary).

The rest of the paper is organized as follows. Section 2 briefly explains the modeling of speech signal using CS. The proposed feature for V/NV detection is explained in Section 3. Section 4 presents the signal adaptive dictionary used to extract sparse feature vector for voiced speech signals, and Section 5 presents the proposed dynamic thresholding scheme for V/NV classification. Section 6 shows the experimental results and the summary of paper is given in Section 7.

## 2. Modeling speech signals using CS

CS takes advantage of the sparsity of a signal $\mathbf{s} \in \mathbf{R}^n$ in an overcomplete dictionary $\mathbf{\Psi} \in \mathbf{R}^{n\times d}(d = n$ for complete dictionary), for various signal processing applications (Donoho, 2006). Here, the raw speech samples are transformed to a sparse representation using $\mathbf{\Psi}$. However, in CS framework, both the sparse representation and the signal can be efficiently recovered from compressed measurements $\mathbf{y}$, sampled below Nyquist rate using a measurement matrix $\mathbf{\Phi} \in \mathbf{R}^{m\times n}$ with $m \ll n$ (Donoho, 2006; Christensen et al., 2009; Sreenivas and Kleijn, 2009). This problem is formulated as:

$$\mathbf{y} = \mathbf{\Phi s} = \mathbf{\Phi\Psi\alpha} = \mathbf{A\alpha} \quad \text{subject to} \quad \mathbf{s} = \mathbf{\Psi\alpha}, \tag{1}$$

where $\boldsymbol{\alpha}$ is the sparse vector obtained after projecting the speech signal on the dictionary $\mathbf{\Psi}$, and $\mathbf{A}$ is the overall reconstruction matrix. The assumption of sparsity means that only $k$ coefficients, with $k \ll n$, of $\boldsymbol{\alpha}$ are sufficient to reconstruct $\mathbf{s}$ with a small error. According to CS theory, if the matrix $\mathbf{\Phi}$ satisfies restricted isometry property (RIP), and is incoherent with the dictionary $\mathbf{\Psi}$, the signal can be recovered by linear programming methods, and is formulated as (Candès et al., 2006; Elad, 2010):

$$\hat{\mathbf{s}} \approx \mathbf{\Psi}\hat{\boldsymbol{\alpha}} \quad \text{where } \hat{\boldsymbol{\alpha}} \text{ is computed as,} \tag{2}$$

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}}\ f(\boldsymbol{\alpha}) \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A\alpha}\|_2^2 < \epsilon$$

Here $\epsilon$ is the error tolerance constant, while $f()$ is a function (e.g., $l_1$-norm) that promotes sparsity in sparse vector. Further, it has been shown that CS also enables reconstruction of sparse signals from noisy measurements (Donoho, 2006). The reconstruction quality is still comparable to that of the signal's optimal sparse approximation (such as obtained by keeping only the largest coefficients in the sparse vector) (Elad, 2010).

## 3. Proposed feature for V/NV detection using CS framework

It has been conjectured in the literature that, there is a significant glottal activity (i.e., the vibration of vocal folds) during the production of voiced speech, while the same is not true in case of nonvoiced (including silence and unvoiced) regions of the speech signals (Ananthapadmanabha and Yegnanarayana, 1979; Yegnanarayana et al., 2001). Hence, one of the ways to

detect V/NV regions could be to extract features from the given speech signal, which characterize the vibration of vocal folds during speech production. This characteristic is associated as the source component in speech production mechanism (Ananthapadmanabha and Yegnanarayana, 1979). Various approaches (with their own advantages and disadvantages) to extract the source and system characteristics from the given speech signal can be found in Sri Rama Murty et al. (2009), Ananthapadmanabha and Yegnanarayana (1979), Christensen et al. (2009), Yegnanarayana et al. (2001), and Cabral et al. (2014).

In this work, we propose a method based on CS which associates Eq. (1) as a source-filter model of speech production for V/NV detection. Here, sparse vector $\boldsymbol{\alpha}$ can be associated with the source characteristics provided a suitable dictionary $\boldsymbol{\Psi}$ (which can be associated with system/filter characteristics) is chosen. Works reported in Christensen et al. (2009), Sreenivas and Kleijn (2009), and Giacobello et al. (2010), have demonstrated the ability of CS framework to efficiently characterize the source and system characteristics of the speech signals. However, none of them have analyzed the behavior of sparse vector for different regions of the speech signal. We have observed that for an appropriately chosen dictionary, the behavior of the sparse vector $\boldsymbol{\alpha}$ for a voiced segment differs from nonvoiced segments, and can be used as a robust indicator of glottal activity in speech segments. Analysis is performed using two different dictionaries namely: Discrete cosine transform (DCT) and WLP, which will be explained in subsequent sections.

Fig. 1 shows the block diagram representation of the proposed V/NV method. Here V/NV detection is performed using compressed measurement vector $\mathbf{y}$ sensed from speech signal using a random measurement matrix. Hence, given the compressed measurements, a source feature vector is extracted using a sparse coding algorithm. Apart from the estimation of sparse feature, the proposed method also efficiently estimates the sparsifying dictionary from the compressed measurements only. However, one can also use pre-learned or analytical dictionaries with the proposed method. Next, an information theoretic metric is computed to measure the unique behavior of the extracted feature. The resultant metric is then compared against a threshold for V/NV decisions. The value of threshold is modified dynamically by exploiting the temporal behavior of the speech signal.

### 3.1. Source characteristics using sparse vector

The proposed method is based on the fact that, the estimated sparse vector $\hat{\boldsymbol{\alpha}}$ using CS/sparse coding framework exhibits high energy in the voiced regions as compared to the unvoiced regions of speech. This could be due to the fact that vocal folds vibrate (which is associated with source feature) during production of voiced sounds. For simplicity, we will first examine this behavior of sparse vector using DCT as dictionary, where a speech signal is represented as a linear combination of various sinusoids. Figs. 2(c) and (d) shows sparse vectors derived for a voiced and unvoiced frame (of duration 25 ms) of a speech signal (taken from CMU-ARCTIC database sampled at 8 kHz) respectively. The instants of glottal closure (GCI), extracted using the Differentiated Electroglottograph (DEGG) signal, are also marked in the figure. It can be observed that number of coefficients with significant amplitude values is proportional to the strength of glottal activity. Also, the coefficients of the sparse vector corresponding to voiced frame have more variance. The same is not true in the case of unvoiced speech. This observation can be illustrated better by plotting the histogram (Fig. 2(e) and (f)) of the values of sparse vector, corresponding to the voiced and the unvoiced frames. Similarly, it was observed that for a transition frame (voiced to unvoiced, unvoiced to voiced, voiced to silence or silence to voiced) the above mentioned behavior i.e., variance of sparse vector coefficients lies between the variance obtained for voiced and unvoiced regions.

In case of noisy speech signal, the derived sparse vector will also be corrupted by noise, because Eq. (2) does not address the presence of noise. However, the estimated sparse vector corresponding to the voiced region of the speech signal is less affected by noise. It can be observed from the sparse vectors (Figs. 3(c) and (d)) estimated for a voiced and an unvoiced frame corrupted by additive white noise (Figs. 3(a) and (b)). The effect of noise is more in an unvoiced frame for which one cannot hold a sparse representation in presence of noise as evident from Fig. 3(d). One can conclude that, the strength of the significant coefficients of sparse vector of noisy speech signal is similar to its clean counterpart, except for some spurious coefficients. This is because, due to vocal fold vibrations, compared to unvoiced frames, voiced speech segments exhibit higher signal to noise ratio (SNR)

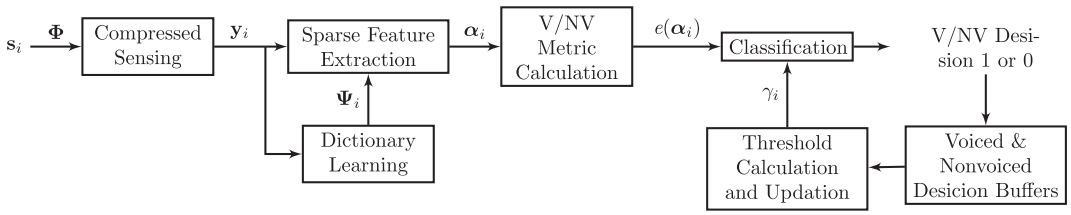


Fig. 1. Block diagram representation of the proposed V/NV detection method. Here $i$ denotes processing at $i^{th}$ frame of the speech signal.

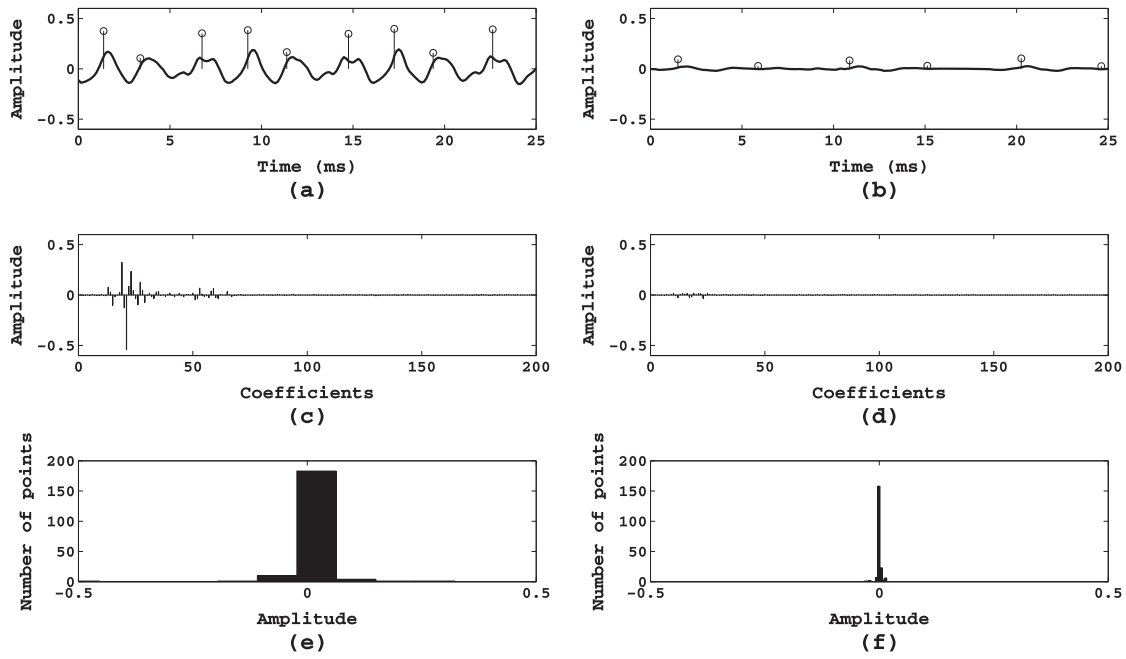


Fig. 2. (a) and (b) Show V and UV frame of a clean speech signal and GCIs with their normalized strength. (c) and (d) Show their sparse vectors derived using DCT dictionary with the corresponding histograms shown in (e) and (f), respectively.

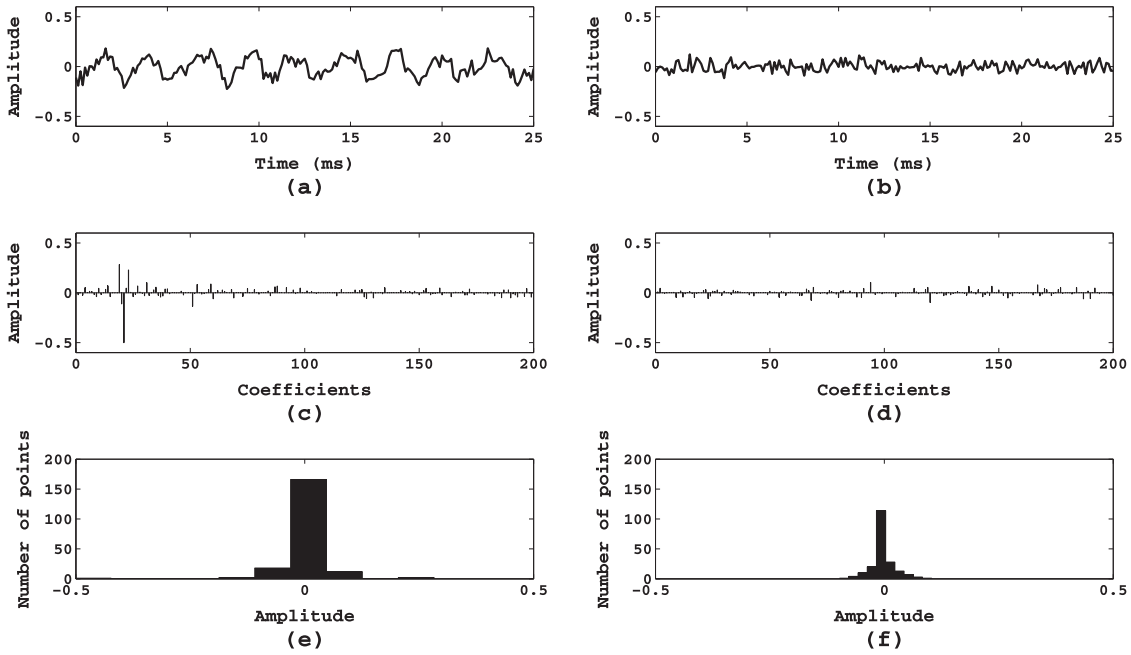


Fig. 3. (a) and (b) Show V and UV frame of a speech signal corrupted with white noise (SNR: 0 dB). (c) and (d) Show their sparse vectors derived using DCT dictionary with the corresponding histograms shown in (e) and (f), respectively.

(Ananthapadmanabha and Yegnanarayana, 1979). Hence, inherent behavior of sparse vector for a voiced frame is not much affected by presence of noise, provided the dictionary is chosen well. In a degraded environment, as the level of noise increases, nonvoiced regions (having low energy) partially or totally merge with the noise. Therefore, it is very difficult to detect unvoiced regions as compared to voiced regions in the presence of noise. This is also the reason that recently proposed sparse coding based VAD algorithms (such as in Teng and Jia (2013) and You et al. (2012)) have poor performance in detecting unvoiced regions.

### 3.2. Proposed V/NV metric

Following the analysis given in the previous section, for clean speech V/NV classification can be done by using a threshold on the number of significant coefficients of the estimated sparse vector for a given speech frame.

However, this procedure fails in the presence of noise as the sparse vector will also be corrupted. Hence, considering variance of the significant coefficients of sparse vector is a better metric as compared to considering a sparsity measure (number of significant or non-zero values) of the sparse vector directly. We propose a metric which quantifies the distribution of sparse vector under the uncertainty induced by noise. To this aim, consider the Renyi's entropy (Jizba and Arimitsu, 2004) of the sparse vector for a $i^{th}$ frame defined as:

$$e_i = \frac{1}{1-p}\log\|\hat{\boldsymbol{\alpha}}_i\|_p, \tag{3}$$

Selecting an appropriate value of $p$ (i.e., $p^{th}$-norm of sparse vector) is very crucial for V/NV metric. The scaling factor $\left(\frac{1}{1-p}\right)$ can be discarded due to the fact that it is constant for all the frames.

Let us consider the case of fixed cardinality, whereby using a hard thresholding sparse recovery algorithm, the solution of sparse vector retains only $k$ significant coefficients. Hence, for $p = 0$, Eq. (3) becomes the log of cardinality (i.e., $\log(k)$) of the sparse vector and this relation is known as Hartley entropy (Jizba and Arimitsu, 2004). On the contrary for $p = 1$, Eq. (3) becomes Shannon entropy (Jizba and Arimitsu, 2004). Hartley entropy counts or considers all the non-zero coefficients of the sparse vector. Shannon entropy (employing $l_1$-norm) is the average unpredictability considering the contribution of all the significant coefficients. In both cases, if some of these coefficients have appeared due to noise then they should be avoided during estimation of metric for better classification of voiced activity. Hence, the extreme values of $p$ are not good choices for an efficient V/NV metric. It has been shown that the change in entropy of a signal vector due to the presence of noise is approximately equal to the change in the logarithm of $l_p$-norm of that vector ($\hat{\boldsymbol{\alpha}}$ in our case) (Dolinar, 1991). Thus, the V/NV metric for sparse vector $\hat{\boldsymbol{\alpha}}$ of $i^{th}$ frame can be chosen as:

$$e_i = \log\|\hat{\boldsymbol{\alpha}}_i\|_p \quad 0 < p < 1 \tag{4}$$

Here, setting $0 < p < 1$ gives us a flexibility to consider variance of significant components of sparse vector $\hat{\boldsymbol{\alpha}}$ in presence of noise. It is illustrated using Fig. 4, which shows the V/NV decisions for a segment of noisy speech signal using proposed V/NV metric for different values of $p$. Here, V/NV decision is performed using a thresholding scheme, which will be explained in Section 5. Clearly choosing $p = 0.3$ performs better compared to $p = 0$ or 1 in capturing the variance of significant sparse vector coefficients under the influence of noise.

To investigate further, Table 1 shows the averaged mean and variance of the proposed V/NV metric for 50 voiced and nonvoiced (silence and unvoiced) frames of clean and noisy speech utterances, taken from CMU-ARCTIC database (CMU ARCTIC). It can be observed that for extreme values of $p$ (i.e., 0 and 1), the difference between the mean values of V/NV metric for voiced and nonvoiced frames is very less. Also, the variance of the V/NV metric firsts increase and then decrease as $p$ increases. However, variance of the metric changes much faster for voiced speech

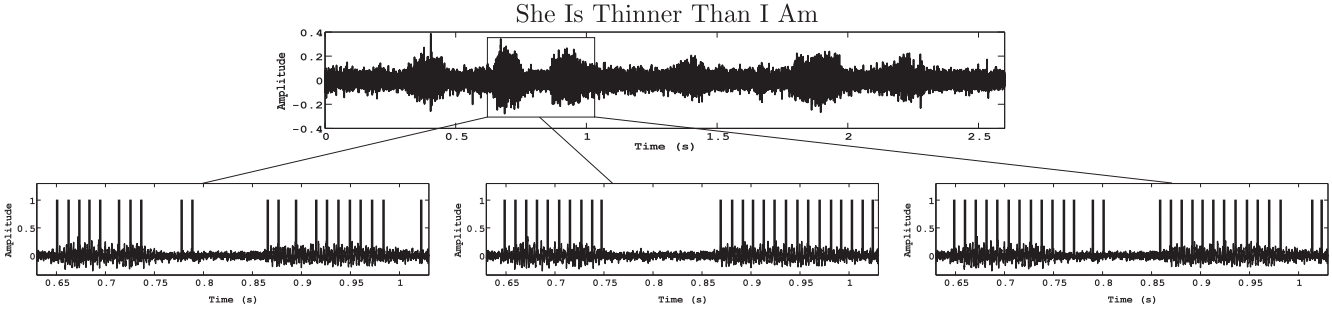


Fig. 4. Framewise V/NV markings {1 (V), 0 (NV)} for a segment of noisy speech in presence of white noise (SNR: 0 dB) using proposed V/NV metric: left bottom-($p = 0$), middle bottom-($p = 0.3$) and right bottom-($p = 1$).

Table 1
Averaged mean and variance of V/NV metric (for 50 V and NV speech frames) shown for different values of $p$.

| $p$ | Clean | | | | With white noise SNR: 0 dB | | | | With babble noise SNR: 0 dB | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NV | | V | | NV | | V | | NV | | V | |
| | Mean | Var | Mean | Var | Mean | Var | Mean | Var | Mean | Var | Mean | Var |
| 0.1 | 40.2 | 0.26 | 41.52 | 0.304 | 42.4 | 0.008 | 42.94 | 0.055 | 42.73 | 0.08 | 43.08 | 0.047 |
| 0.2 | 17.9 | 0.27 | 18.72 | 0.307 | 19.4 | 0.009 | 20.10 | 0.071 | 19.78 | 0.16 | 20.23 | 0.10 |
| 0.3 | 10.3 | 0.34 | 11.97 | 0.318 | 11.8 | 0.014 | 13.21 | 0.057 | 11.90 | 0.15 | 12.98 | 0.08 |
| 0.5 | 4.39 | 0.29 | 5.26 | 0.265 | 5.71 | 0.017 | 7.18 | 0.065 | 5.86 | 0.27 | 7.04 | 0.07 |
| 0.7 | 2.72 | 0.28 | 3.01 | 0.253 | 3.30 | 0.012 | 3.55 | 0.052 | 3.21 | 0.22 | 3.34 | 0.06 |
| 0.9 | 2.47 | 0.07 | 1.79 | 0.241 | 2.23 | 0.09 | 2.27 | 0.051 | 2.01 | 0.09 | 2.07 | 0.05 |

as compared to unvoiced speech. In order to have a good classification accuracy variance of the metric should be small and distance between means of voiced and nonvoiced frames should be large. Experimentally, we have observed that the value of $p$ between 0.3 and 0.5 gives the best classification accuracy. These observations are consistent for both stationary (e.g., white) and non-stationary (e.g., babble) noises.

## 4. Proposed dictionary

The important prerequisite for the proposed V/NV method is a suitable choice of dictionary, which can provide efficient sparse representation for speech signals, capturing the behavior of voiced sounds under the influence of noise. In literature, sparse representations of speech signals have been studied using dictionaries, which are either signal independent transforms (e.g., DCT) also called analytical dictionaries, or signal dependent transforms such as learned dictionaries (Tosic and Frossard, 2011). Analytical dictionaries have fast numerical implementation but cannot efficiently adapt to different signal variation, especially in the noisy environments (Tosic and Frossard, 2011). To address this issue, various approaches have been proposed to learn a data dependent dictionary (Sreenivas and Kleijn, 2009; Tosic and Frossard, 2011; Jafari and Plumbley, 2011). Although, this provides an efficient sparser representation but, if the characteristics of signals under consideration are exploited, the performance of a learned dictionary in different applications can be improved further (Sreenivas and Kleijn, 2009).

In the context of CS based speech signal processing, a dictionary should exploit the characteristics of speech production mechanism. One such potential approach is to model voiced speech as the output of a time-varying (signal dependent) linear system excited by impulse like signal (Christensen et al., 2009). It can be incorporated by considering Eq. (2) as source-filter model of speech production, where the vocal tract impulse response of the system/filter acts as the dictionary for sparse representation of speech signals (Giacobello et al., 2010). Impulse response of the vocal tract system can be computed using various speech modeling approaches, which tries to model the spectral envelope of speech signal in terms of very few coefficients/-model parameters (such as Linear predictive (LP) model, Liljencrants–Fant (LF) model or cepstrum model) (Cabral et al., 2014). However, our work is based on linear predictive analysis of the speech signal.

### 4.1. Learning LP dictionary using CS framework

In traditional LP analysis, a speech signal $s(n)$ is expressed as a convolution of the impulse response of vocal tract system $h(n)$ (estimated from coefficients $a_q$) of order $L$, and excitation signal $r(n)$, as: $s(n) = h(n) \circledast r(n)$ or in matrix form as $\mathbf{s} = \mathbf{Hr}$ (Sreenivas and Kleijn, 2009). Here $\mathbf{H} \in \mathbb{R}^{n \times n}$ is a convolution matrix which can be constructed from the truncated impulse response (which ensures stability) of the infinite impulse response (IIR) LP synthesis filter, expressed in $z$-domain as (Giacobello et al., 2010; Raj and Sreenivas, 2011):

$$G(z) = \frac{1}{1 - \sum_{q=1}^{L} a_q z^{-q}} \tag{5}$$

Considering Eq. (1), the dictionary $\mathbf{\Psi}$ can be associated with impulse response matrix $\mathbf{H}$ and the sparse feature vector $\boldsymbol{\alpha}$ can be associated with the residual or excitation signal $\mathbf{r}$ (Sreenivas and Kleijn, 2009). Recent work in Giacobello et al. (2010), has shown the application of CS in speech coding where the dictionary (impulse response matrix computed from clean speech signal) is known *a priori*. Similarly, approach in Sreenivas and Kleijn (2009), uses a pre-estimated line spectral frequency (LSF) code book of vocal tract impulse response derived from training data to construct the dictionary. However, in CS framework estimation of filter coefficients (or impulse response matrix) from compressed measurements is not straight forward as compared to its estimation from speech signal itself (Giacobello et al., 2010). The proposed method aims at performing V/NV detection using the compressed measurements only. Hence, we propose an iterative approach, whereby using a limited number of CS measurements (random projections), both the dictionary and corresponding sparse feature vector is estimated (on frame by frame basis).

It is well known that human speech has its fundamental frequency ($F_0$) and first few formants in the lower frequency band, which plays an important role in discriminating voiced sounds (Ananthapadmanabha and Yegna narayana, 1979; Paliwal et al., 2009). Thus, one should use a non-linear frequency scale (e.g. bark scale) in LP analysis for efficient modeling of voiced sounds, by giving more emphasis to the lower frequency band as compared to the higher band (Härmä et al., 2000). This can be done by using frequency warping (transforming uniform frequency scale to bark scale), which is performed by replacing the unit delays ($z^{-1}$) in conventional LP modeling with first-order all-pass filters $D(z)$ (known as the Laguerre filter) (Härmä et al., 2000). The synthesis filter is now expressed as:

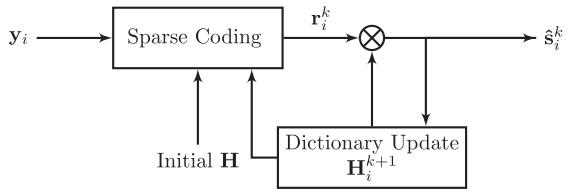


Fig. 5. Block diagram representation of the dictionary learning procedure. Here $i$ denotes processing at $i^{th}$ frame of the speech signal and $k$ denotes the iteration number.

$$G(z) = \frac{1}{1 - \sum_{q=1}^{L} a_q D(z)^q}, \quad \text{where} \quad D(z) = \frac{-\beta + z^{-1}}{1 - \beta z^{-1}} \tag{6}$$

Positive values of warping coefficient $\beta$ results in longer group delay for low frequencies and shorter group delay for high frequencies (Härmä et al., 2000). Thus, a better resolution is obtained at lower frequencies at the expense of poor resolution at high frequencies. Hence, the proposed approach employ warped-LP (WLP) analysis to compute the vocal tract impulse response. Also, compared to LP analysis, the residual signal obtained in case of WLP analysis is sparser (Härmä et al., 2000) and thus more robust for V/NV detection in noisy environments.

**Algorithm 1.** Iterative estimation of WLP dictionary and sparse source feature from CS measurements.

**Inputs:** CS measurement vector $\mathbf{y}$ for a frame of speech signal along with measurement matrix $\mathbf{\Phi}$
**Outputs:** Dictionary $\mathbf{H}^k$, source feature $\mathbf{r}^k$ after $k^{th}$ iteration

1: Initialization: $k = 0$, $\epsilon = 0.001, \mathbf{h}^0 = []$, $\mathbf{H}^0 = \text{DCT}$ and $\mathbf{A}^0 = \mathbf{\Phi}\mathbf{H}^0$.
**Perform iterations**
2: $\mathbf{r}^k \leftarrow \underset{\mathbf{r}}{\text{argmin}} \ \|\mathbf{r}\|_1 \quad \text{subject to} \quad \mathbf{y} = \mathbf{A}^k\mathbf{r}^k$
3: $\hat{\mathbf{s}}^k = \mathbf{H}^k\mathbf{r}^k$
4: Compute WLP coefficients $\hat{\mathbf{a}}^k$ from current estimate of $\hat{\mathbf{s}}^k$
5: Estimate vocal tract impulse response $\mathbf{h}^{k+1}$ from predictor $\mathbf{a}^k$
6: Update dictionary $\mathbf{H}^{k+1} = conv(\mathbf{h}^{k+1})$
7: $k = k + 1$
**Untill** $\|\mathbf{y}\|_2 / \|\mathbf{y} - \mathbf{A}^k\mathbf{r}^k\|_2 < \epsilon$

Note: $conv()$ stands for convolution matrix

Fig. 5 shows the dictionary learning procedure and details are given in Algorithm 1. The dictionary is initialized to DCT matrix, and a sparse recovery algorithm is used to solve for the residual signal on compressed speech measurements by minimizing $l_1$-norm of the residual signal. Now, given the sparse vector speech is reconstructed using the current dictionary and used for estimation of the filter coefficients using the traditional method of the least squares[1] (Raj and Sreenivas, 2011). The updated filter coefficients are then used to estimate the impulse response in order to build the dictionary. This procedure continues iteratively until the convergence, which can be achieved by either fixing number of iterations or when SNR in the measurement space (i.e., $\|\mathbf{y}\|_2 / \|\mathbf{y} - \mathbf{A}\hat{\mathbf{r}}\|_2$) reaches to a predefined value. In case of noisy speech, due to inherit denoising advantage of CS/sparse coding framework (Low et al., 2013), the proposed procedure also repeatedly reduces the effect of noise in each iteration, allowing a robust estimation of both the dictionary and the corresponding sparse feature vector.

[1] WLP implementation can be found in WarpTB toolkit: http://www.acoustics.hut.fi/software/warp.

## 5. Proposed method for V/NV classification

Classification of a given speech region as voiced or unvoiced using the proposed V/NV metric can be done by selecting a suitable threshold value. However, a fixed value of threshold $\gamma$ might not perform well for entire speech utterance under different noisy conditions. This issue is addressed by considering the observation that, speech in general is processed on a very short duration and compared to voiced regions, nonvoiced regions may extend up to several frames (Kristjansson et al., 2005). Also, in most cases characteristics of sparse vector obtained from nonvoiced frames (including unvoiced and silence frames) remains stationary. Thus, one can exploit this temporal information to update the threshold. Hence, we propose to update the threshold $\gamma$ using an adaptive threshold selection scheme. The threshold $\gamma$ at $i^{th}$ frame is updated based on the V/NV metric computed for previous 20 frames as:

$$\begin{aligned} &\gamma_i = (1 - \lambda)e_{v_i} + \lambda e_{nv_i} \\ &e_{v_i} = \min_{j \in V}(e_j) \quad \text{and} \quad e_{nv_i} = \max_{j \in NV}(e_j), \\ &\forall_j \quad j = i-1, i-2, \ldots i-20 \end{aligned} \tag{7}$$

where $\lambda$ is the parameter for combination. For initialization of buffers, few initial frames of the observed signal are assumed as silence frames. In case all the previous 20 frames are nonvoiced, max value of $e_j$, where as if they are voiced, min value of $e_j$ (multiplied by appropriate factor) is selected as $\gamma$. However, in cases where previous few frames are both voiced and nonvoiced (e.g., for transition regions), $\gamma$ is computed using Eq. (7). We experimentally found that $\lambda = 0.7$ results in maximum accuracy for V/NV decisions. This choice of $\lambda$ selects the threshold between mean values of the $e_j$ of nonvoiced and voiced speech. This is because, on average the $e_j$ of noisy voiced regions is more than that of noisy nonvoiced regions, thus $\gamma$ should be more than the mean $e_j$ of noise. Further, we assume that the degree of non-stationarity of the nonvoiced regions does not change drastically over few frames.

## 6. Experimental results

The performance of the proposed method is evaluated using speech utterances (from 7 different speakers) taken from CMU-ARCTIC (CMU ARCTIC) database resampled at 8 kHz, and corrupted by eight type of additive noises (white, F16, pink, factory, volvo, babble, machine-gun, leopard) taken from the NOISEX-92 database at different SNR levels. The percentage of voiced speech samples in each of the utterances is maintained at 40% by

appending requisite duration of silence before the addition of noise samples. The DEGG signal corresponding to each speech signal is used for deriving the ground truth to minimize human error in manual labeling. $12^{th}$ order WLP filter with a warping coefficient of $\beta = 0.35$ is considered for dictionary learning. Speech is processed on a short time frame basis, where framing is achieved by applying a 25 ms long Hanning window. In all the experiments, frame overlap is set to 70% unless otherwise specified. The sensing matrix $\mathbf{\Phi}$ is chosen to be a random Gaussian matrix with a compression ratio of $m/n = 0.5$ and the V/NV metric is computed with $p = 0.3$. In this paper our experiments are divided into three parts: Section 6.1 highlights the advantages of the proposed WLP dictionary for V/NV detection. Section 6.2 evaluates the proposed method for V/NV detection under different noisy environments. Finally, Section 6.3 shows a comparative analysis of the proposed method along with existing methods for V/NV detection.

### 6.1. Robustness of warped-LP dictionary for V/NV detection

Compared to conventional LP analysis, the feature vector estimated using WLP dictionary is more sparse. This is because residual signal in case of WLP analysis is more close to the true source of excitation signal (a quasi-periodic impulse train in case of voiced speech) (Härmä et al., 2000). Figs. 6(c) and (d) show the sparse vector obtained for a noisy voiced and unvoiced frame using the proposed WLP dictionary. It can be observed that even in presence of noise the extracted feature vector for a voiced frame efficiently captures the source characteristics, and is less affected by noise as compared to the unvoiced frame. In Section 3.1, it was explained that for speech modeling using CS framework, the dictionary $\mathbf{\Psi}$ could be of sine/cosine functions constructed with or without the knowledge of sinusoidal component phases, frequencies and actual pitch $F_0$ of the speaker (e.g., Gabor, DCT or DFT dictionary) (Tosic and Frossard, 2011). However, comparison of histogram plot in Figs. 6(e) and (f) along with Fig. 3, illustrates that WLP dictionary is more robust against noise as compared to DCT dictionary.

To investigate further, Fig. 7 shows the obtained V/NV metric for a sample speech waveform using both the dictionaries. One can observe that V/NV metric in case of WLP dictionary is more discriminating as compared to the DCT dictionary. This is because in the proposed dictionary, only voiced speech frames can be sparsely represented and thus the deviation between the V/NV metric

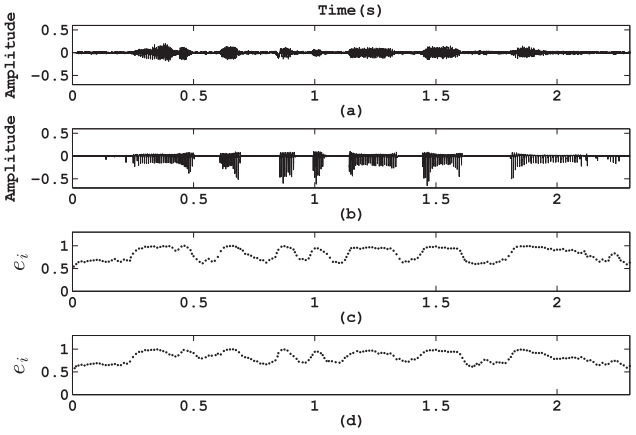


Fig. 7. (a) Speech signal corrupted by babble noise (SNR: 0 dB). (b) DEGG signal. Normalized V/NV metric obtained using (c) WLP dictionary and (d) DCT dictionary.

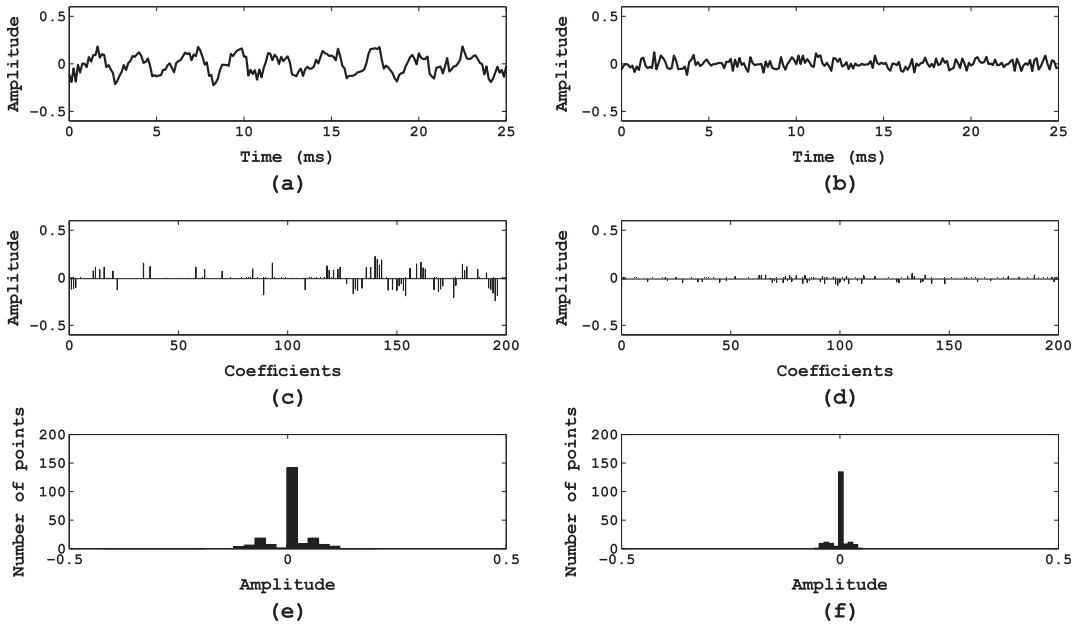


Fig. 6. (a) and (b) Show V and UV frame of a speech signal corrupted with white noise (SNR: 0 dB). (c) and (d) Show their sparse vectors derived using WLP dictionary with the corresponding histograms shown in (e) and (f), respectively.

computed for voiced and nonvoiced frames is generally large. In order to have fast implementation, DCT can be preferred, but in this case the information of the speech source (sparse vector) and speech production system (dictionary) is coupled (Girin et al., 2007). This coupling produces ambiguous results for characterizing voiced sounds in certain cases, which results in a lower V/NV accuracy. This is one possible explanation for analytical dictionaries being more popular in speech synthesis instead of applications such as VAD or V/NV detection.

### 6.2. Performance under different types of noises

In order to evaluate the robustness of the proposed method against different types of additive noises, three different parameters namely number of correct classified voiced samples (CORRECT), front end clipping or number of samples clipped at the front of a voiced speech burst (FEC) and number of nonvoiced samples detected as voiced (NDV) are employed (Mousazadeh and Cohen, 2013). In practice, high values of CORRECT and low values for NDV and FEC are desired. Table 2 shows these scores averaged over SNR values of −5 dB, 0 dB, 5 dB. These scores are calculated by considering a tolerance interval of 5 ms around the starting and ending boundaries (as per ground truth) for a voiced speech region.

While expected results are achieved for stationary noises (e.g., white noise), results for non-stationary noises (e.g., factory and volvo noises) are also close to the case of stationary noise counterpart. However, value of CORRECT is less in case of highly non-stationary noises such as babble, machinegun and leopard noise. Because, it is assumed that only speech can have a sparse representation while the same is not true for noise (Abrol et al., 2013). However, babble noise has characteristics like speech and machinegun noise has impulse like characteristics. Hence, good accuracy is achieved only in case of stationary noisy environments. This observation can be made from Fig. 8, which shows the V/NV decisions for a speech waveform under different noisy conditions. In comparison to clean speech, the V/NV markings obtained using the proposed method for most of the voiced segments under different noisy conditions are correct. However, some wrong V/NV decisions are obtained in nonvoiced regions, especially for noises like babble, machinegun, pink and factory. This is because such noises confuses with properties of sparse vector of voiced region, which leads to poor V/NV detection performance.

### 6.3. Comparison with the existing methods

In this section, we examine the performance of the proposed method compared to the few existing methods. The performance is measured in terms of missed detection ($P_m$) and false alarm rate ($P_f$) as $P = 1 - (0.4P_m + 0.6P_f)$ (Dhananjaya and Yegnanarayana, 2010). $P_m$ denotes the ratio of the samples that belong to voiced regions but are incorrectly detected as nonvoiced to the total number of samples in the speech signal. $P_f$ denotes the ratio of the samples that belong to nonvoiced regions but are incorrectly detected as voiced to the total number of samples in the speech signal. Further, as in the previous Section 6.2, classification accuracy is calculated by considering a tolerance interval of 5 ms around the starting and ending boundaries of a voiced speech region across all methods.

#### 6.3.1. Experiment 1

In this experiment, we compare the performance of the proposed method along with existing VAD methods for V/NV detection. Table 3 shows these scores, averaged over eight types of noises at different SNR levels. Since VAD methods are optimized to detect speech regions (containing both voiced and unvoiced speech), in order to have a fair comparison, all the three methods (i.e., NMF (Franois et al., 2013), SVM (Saeedi et al., 2013) and SGMM (sequential Gaussian mixture model) (Ying et al., 2011)) are trained or optimized to obtain maximum V/NV classification. These are natural candidates for comparison as they fall under the categories of sparse coding, supervised and unsupervised methods, respectively. For experiments we considered a dataset consists of 300 utterances, 60% of which were used for training in case of supervised NMF and SVM approaches. NMF method extract a sparse feature from a noise-suppressed representation of noisy speech signal via speech spectrum decomposition over an overcomplete dictionary, which is concatenation of two dictionaries learned from clean speech and noise samples. For V/NV detection voiced and nonvoiced regions of speech are considered for learning the dictionaries. NMF based methods (such as in Teng and Jia (2013) and Franois et al. (2013)) are batch algorithms which perform well, but relies on pre-estimation of dictionaries for both noise and clean speech signal for efficient sparse representation, which might not be a suitable choice in the practical scenario. The supervised method proposed in Saeedi et al. (2013) employ SVM models trained in different background noises for speech/non-speech classification. For

Table 2
Objective V/NV parameters in %, averaged over all SNRs for eight types of noises using WLP dictionary.

| Metric | White | F16 | Pink | Volvo | Factory | Babble | Machinegun | Leopard |
|---|---|---|---|---|---|---|---|---|
| CORRECT | 97.2 | 95.4 | 93.5 | 92.1 | 90.4 | 88.6 | 85.2 | 84.7 |
| NDV | 7.3 | 11.7 | 13.1 | 15.6 | 17.2 | 17.7 | 22.1 | 24.3 |
| FEC | 0.42 | 0.53 | 0.61 | 0.73 | 0.82 | 0.97 | 0.96 | 0.98 |

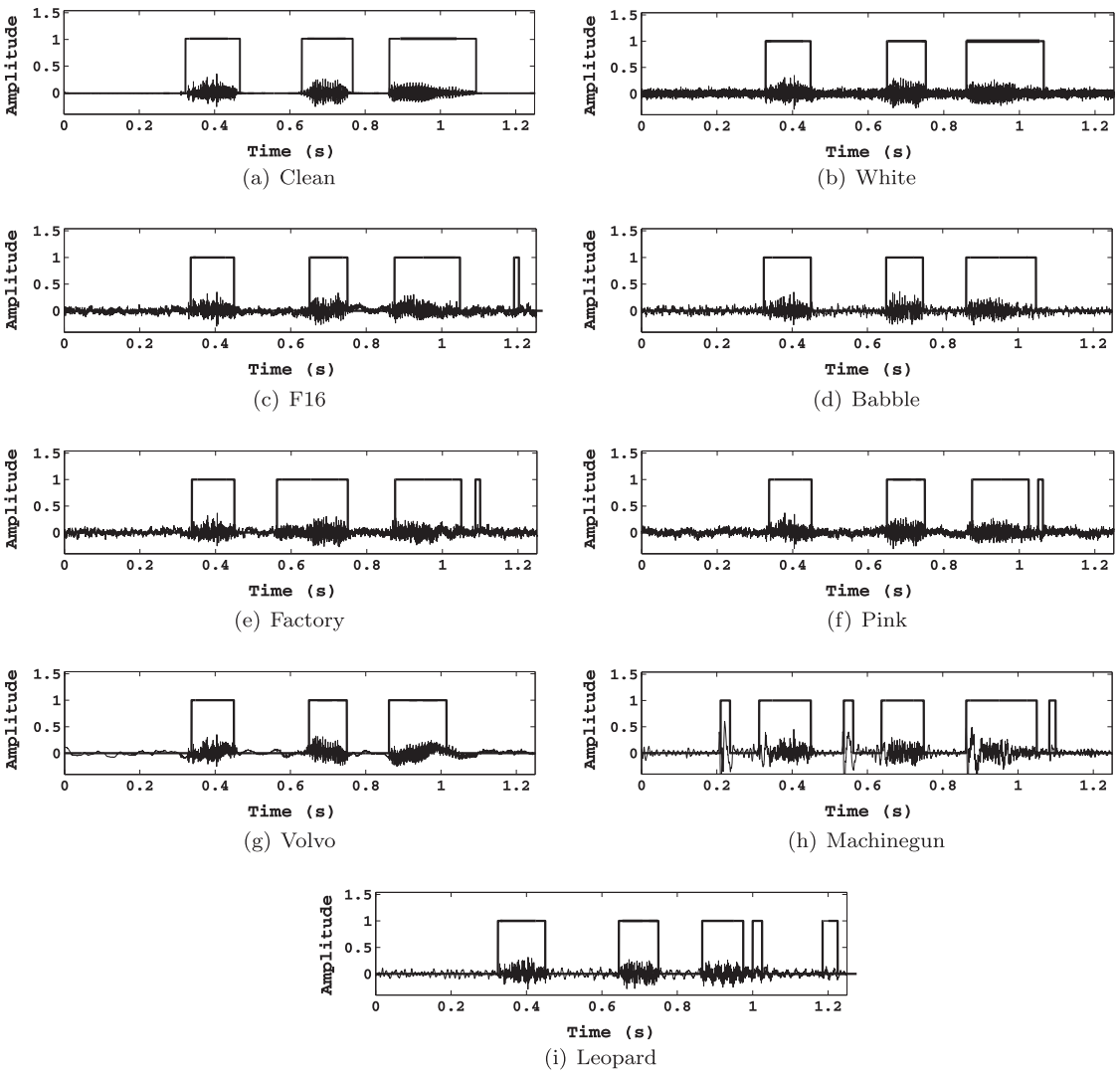


Fig. 8. V/NV decisions for a segment of speech signal using the proposed method: (a) clean. (b)–(i) corrupted by various types of noises (SNR: 0 dB).

V/NV detection the SVM models were trained from voiced and nonvoiced regions of speech under different noisy conditions. Method in Ying et al. (2011) employ a sequential expectation–maximization approach to update the parameters of a two component GMM for VAD. Hence, for V/NV detection the initialized GMM's are adapted to classify voiced and nonvoiced regions instead of speech and non-speech regions of a given utterance.

It should be noted that while the proposed method uses compressed speech samples, the existing methods are applied on raw speech samples directly. One can argue that existing methods can still be applied after reconstructing the speech signal back from compressed measurements. This itself requires a suitable dictionary (preferably learned from clean training data) to efficiently reconstruct the signal. In addition, this procedure increases the processing burden of a V/NV detection algorithm. In contrast, we have assumed the situation where only compressed speech samples are available. Hence, in order to have a fair comparison, V/NV detection was also performed on the recovered speech signal from CS measurements using a DCT dictionary. These results are shown in Table 3 as Proposed-I, NMF-I, SVM-I and SGMM-I.

It can be observed from Table 3 that the proposed method employing WLP dictionary, outperforms existing methods. Superior performance of the proposed method using WLP dictionary over DCT dictionary supports the claim that the estimation of sparse vector from compressively sensed speech is very much influenced by the choice of dictionary. Moreover as argued earlier, V/NV detection results obtained on recovered speech using proposed or any existing method are poor. However, the results for the case of method Proposed-I are still better than methods NMF-I, SVM-I and SGMM-I. This is mainly because we assume that no suitable dictionary is available *a priori* and recovered speech signal from CS samples using DCT dictionary does not preserve the speech characteristics especially in noisy environment. Moreover, learning dictionary iteratively using CS samples instead of directly from the recovered speech, is better in suppressing the effect of noise. This is evident from the superior performance of the proposed method using WLP dictionary over the method

Table 3
Averaged V/NV classification accuracy in % (for all noises) of the proposed and the baseline methods at different SNR values over 10 trials.

| Noise SNR (dB) | Proposed (with WLP) | Proposed (with DCT) | NMF (Franois et al., 2013) | SVM (Saeedi et al., 2013) | SGMM (Ying et al., 2011) | Proposed-I (with WLP) | NMF-I | SVM-I | SGMM-I |
|---|---|---|---|---|---|---|---|---|---|
| 5 | 92.3 | 89.7 | 91.5 | 90.7 | 90.2 | 90.9 | 89.3 | 88.2 | 87.8 |
| 0 | 90.2 | 87.6 | 89.1 | 88.4 | 87.7 | 88.5 | 87.6 | 87.1 | 86.4 |
| −5 | 86.4 | 83.3 | 85.4 | 83.5 | 83.6 | 85.2 | 84.5 | 81.2 | 79.6 |

Table 4
V/NV classification accuracy in % of the proposed and the baseline methods based on ZFR at different SNR values over 10 trials.

| Method | SNR (dB) | Classification accuracy under different noisy conditions | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | White | F16 | Pink | Volvo | Factory | Babble | Machinegun | Leopard |
| Proposed-WLP 70% frame overlap | 5 | 94.3 | 94.1 | 93.7 | 93.4 | 92.7 | 92.2 | 87.2 | 85.3 |
| | 0 | 92.4 | 92.2 | 91.6 | 90.2 | 89.6 | 89.2 | 85.3 | 83.4 |
| | −5 | 90.1 | 89.7 | 89.4 | 87.4 | 86.5 | 84.2 | 80.4 | 78.6 |
| Proposed-WLP 50% frame overlap | 5 | 93.2 | 92.8 | 92.4 | 91.4 | 90.6 | 89.9 | 86.5 | 84.1 |
| | 0 | 91.4 | 91.1 | 90.7 | 90.2 | 89.4 | 88.7 | 84.1 | 80.9 |
| | −5 | 89.1 | 88.7 | 88.3 | 87.9 | 87.3 | 83.4 | 78.7 | 77.1 |
| ZFR (Energy) (Sri Rama Murty et al., 2009) 70% frame overlap | 5 | 92.8 | 91.3 | 91.1 | 90.5 | 89.4 | 88.7 | 85.2 | 83.8 |
| | 0 | 90.4 | 90.1 | 89.4 | 89.1 | 88.7 | 86.5 | 82.5 | 79.4 |
| | −5 | 88.7 | 87.4 | 87.2 | 86.3 | 85.2 | 81.6 | 77.3 | 75.4 |
| ZFR (GCI) (Dhananjaya and Yegnanarayana, 2010) | 5 | 94.5 | 94.3 | 93.8 | 93.6 | 92.9 | 92.4 | 87.2 | 84.7 |
| | 0 | 92.3 | 92.2 | 91.6 | 90.3 | 89.7 | 89.1 | 84.2 | 81.2 |
| | −5 | 90.3 | 89.7 | 89.3 | 87.3 | 86.5 | 83.6 | 79.3 | 77.4 |

Proposed-I. This confirms that, it is better to perform V/NV detection on CS samples than on reconstructed speech.

### 6.3.2. Experiment 2

In this experiment, we compare the proposed method with two recent V/NV detection methods proposed in Dhananjaya and Yegnanarayana (2010) and Sri Rama Murty et al. (2009). Similar to proposed method, both these methods also rely on glottal activity detection for V/NV classification. Both the methods in Dhananjaya and Yegnanarayana (2010) and Sri Rama Murty et al. (2009) use zero frequency resonator (ZFR) filtered signal to extract instants of glottal closure (epochs), which are known to be robust against different types of degradations especially in voiced regions of speech. It was argued in Dhananjaya and Yegnanarayana (2010) that for a speech signal corrupted by two different noise functions, the GCIs detected during the voiced regions show less drift as compared to the large drift in case of nonvoiced regions. Voiced regions are thus detected as regions with low values for GCI drift and jitter. In contrast, V/NV detection by method in Sri Rama Murty et al. (2009) is done by computing energy of the ZFR filtered signal (over a short duration) and comparing it against a fixed threshold. However, zero frequency filtering introduces a dominant low frequency trend. This trend is removed by a mean subtraction process, which require the global average pitch period to be known *a priori*. The pitch parameter is very critical in the sense that an inappropriate value degrades the performance (Dhananjaya and Yegnanarayana, 2010). Classification accuracy of ZFR method in Dhananjaya and Yegnanarayana (2010) is calculated based on number of epochs (GCIs) detected as voiced or non-voiced. The lowest time resolution achieved by ZFR method in Dhananjaya and Yegnanarayana (2010) is the minimum duration between two successive epochs because the V/NV decision is taken at each epoch location. In contrast, the lowest time resolution achieved by the proposed method and the method (Sri Rama Murty et al., 2009) depend on percentage of frame overlap as decisions are taken on frame-by-frame basis. Hence, in order to have a fair comparison the performance measures (i.e., $P_m$ and $P_f$) are calculated in terms of number of samples as explained earlier.

Table 4 shows the classification accuracy (i.e., $P$ in %), for eight types of additive noises at different SNR levels. Classification accuracy for the proposed method was evaluated at various percentage of frame overlap. It was observed that at 50% frame overlap the proposed method results in poor V/NV classification accuracy as compared to the method in Dhananjaya and Yegnanarayana (2010). However, better performance was achieved as compared to the method in Sri Rama Murty et al. (2009). We found that compared to high, at low percentages of frame overlap, V/NV markings in case of proposed method drifted beyond the tolerance interval of 5 ms in certain cases resulting in poor accuracy. In case of 70% frame overlap the proposed method achieves comparable performance as in case of Dhananjaya and Yegnanarayana

(2010) under different noisy environments. In general, frame overlap beyond 70% is not preferable due to increase in processing delay of the algorithm. Further, it is important to note that the proposed method performs best at low SNR levels especially for highly non-stationary (e.g., babble, machinegun and leopard) noises.

## 7. Conclusion

We have presented an unsupervised CS/sparse coding based method for V/NV detection that requires no training data, and is robust against noisy environments. The proposed method exploits the distinct behavior of feature vector extracted from the voiced and the nonvoiced regions of compressively sensed speech signals. It has been shown that if a suitable dictionary is chosen, the sparse vector contains the source characteristics of the speech signals which is very decisive in V/NV detection. The behavior of sparse vector for different regions of speech is quantified using an information theoretic based criterion. The final V/NV classification is done using an adaptive threshold selection scheme, which exploits the temporal information of speech signals. Both, feature extraction and dictionary learning (which is motivated by speech production mechanism) are performed during frame based processing of the compressively sensed speech signals. However, the proposed V/NV detection method can also be extended to speech signals directly. It should be noted that as the level of noise increases one can not completely reduce the effect of noise. However, due to inherent properties of CS/sparse coding framework, the proposed iterative dictionary learning procedure is able to reduce the effect of noise to some extent. V/NV detection in presence of highly non-stationary noises (like babble and machinegun) is still a challenging task as such noises confuses with the properties of sparse vector of the voiced region. 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, which in turn can be used for applications such as speaker verification.

## References


Abrol, V., Sharma, P., Sao, A.K., 2013. Speech enhancement using compressed sensing. In: Proceedings, 14th INTERSPEECH, ISCA, Lyon, France, pp. 3274–3278.

Ahmadi, S., Spanias, A.S., 1999. Cepstrum-based pitch detection using a new statistical V/UV classification algorithm. IEEE Trans. Speech Audio Process. 7 (3), 333–338. http://dx.doi.org/10.1109/89.759042.

Ananthapadmanabha, T., Yegnanarayana, B., 1979. Epoch extraction from linear prediction residual for identification of closed glottis interval. IEEE Trans. Acoust. Speech Signal Process. 27 (4), 309–319. http://dx.doi.org/10.1109/TASSP.1979.1163267.

Arifianto, D., 2007. Dual parameters for voiced–unvoiced speech signal determination. In: Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. IV-749–IV-752. http://dx.doi.org/10.1109/ICASSP.2007.367021.

Asaei, A., Bourlard, H., Cevher, V., 2011. Model-based compressive sensing for multi-party distant speech recognition. In: Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 4600–4603. http://dx.doi.org/10.1109/ICASSP.2011.5947379.

Atal, B., Rabiner, L., 1976. A pattern recognition approach to voiced–unvoiced–silence classification with applications to speech recognition. IEEE Trans. Acoust. Speech Signal Process. 24 (3), 201–212. http://dx.doi.org/10.1109/TASSP.1976.1162800.

Benyassine, A., Shlomot, E., Su, H.-Y., Massaloux, D., Lamblin, C., Petit, J.P., 1997. ITU-T recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. IEEE Commun. Mag. 35 (9), 64–73. http://dx.doi.org/10.1109/35.620527.

Cabral, J.P., Richmond, K., Yamagishi, J., Renals, S., 2014. Glottal spectral separation for speech synthesis. IEEE J. Sel. Top. Signal Process. 8 (2), 195–208. http://dx.doi.org/10.1109/JSTSP.2014.2307274.

Candès, E.J., Wakin, M.B., 2008. An introduction to compressive sampling. IEEE Signal Process. Mag. 25 (2), 21–30.

Candès, E., Romberg, J., Tao, T., 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory 52 (2), 489–509. http://dx.doi.org/10.1109/TIT.2005.862083.

Christensen, M.G., Stergaard, J., Jensen, S.H., 2009. On compressed sensing and its application to speech and audio signals. In: Proceedings, Forty-Third Asilomar Conference on Signals, Systems and Computers, California, USA, November 2009, pp. 356–360. http://dx.doi.org/10.1109/ACSSC.2009.5469828.

CMU ARCTIC Speech Synthesis Databases. <http://festvox.org/>.

Dhananjaya, N., Yegnanarayana, B., 2010. Voiced/nonvoiced detection based on robustness of voiced epochs. IEEE Signal Process. Lett. 17 (3), 273–276. http://dx.doi.org/10.1109/LSP.2009.2038507.

Dolinar, S., 1991. Maximum-entropy probability distributions under $l_p$-norm constraints. In: TDA Progress Report, NASA, Computational System Research Section, pp. 74–87.

Donoho, D.L., 2006. Compressed sensing. IEEE Trans. Inform. Theory 52 (4), 1289–1306. http://dx.doi.org/10.1109/TIT.2006.871582.

Elad, M., 2010. Sparse and Redundant Representations – From Theory to Applications in Signal and Image Processing. Springer.

Franois, G.G., Dennis, L.S., Gautham, J.M., 2013. Speaker and noise independent voice activity detection. In: Proceedings, 14th INTERSPEECH, ISCA, Lyon, France, September 2013.

Giacobello, D., Christensen, M.G., Murthi, M.N., Jensen, S.H., Moonen, M., 2010. Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction. IEEE Signal Process. Lett. 17 (1), 103–106. http://dx.doi.org/10.1109/LSP.2009.2034560.

Girin, L., Firouzmand, M., Marchand, S., 2007. Perceptual long-term variable-rate sinusoidal modeling of speech. IEEE Trans. Audio Speech Lang. Process. 15 (3), 851–861. http://dx.doi.org/10.1109/TASL.2006.885928.

Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U.K., Huopaniemi, J., 2000. Frequency-warped signal processing for audio applications. J. Audio Eng. Soc. 48 (11), 1011–1031.

Jafari, M.G., Plumbley, M.D., 2011. Fast dictionary learning for sparse representations of speech signals. IEEE J. Sel. Top. Signal Process. 5 (5), 1025–1031. http://dx.doi.org/10.1109/JSTSP.2011.2157892.

Jancovic, P., Kokuer, M., 2006. Voicing-character estimation of speech spectra: application to noise robust speech recognition. In: Proceedings, IEEE International Conference on Acoustics, Speech and Signal (ICASSP), vol. 1, pp. I–I. http://dx.doi.org/10.1109/ICASSP.2006.1660006.

Jizba, P., Arimitsu, T., 2004. Observability of rényis entropy. Phys. Rev. E 69 (2), 026128.

Kinnunen, T., Chernenko, E., Tuononen, M., Fränti, P., Li, H., 2007. Voice activity detection using MFCC features and support vector

machine. In: Proceedings, International Conference on Speech and Computer (SPECOM), vol. 2, Moscow, Russia, pp. 556–561.
Koolagudi, S.G., Reddy, R., Rao, K.S., 2010. Emotion recognition from speech signal using epoch parameters. In: Proceedings, International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. http://dx.doi.org/10.1109/SPCOM.2010.5560541.
Kristjansson, T., Deligne, S., Olsen, P., 2005. Voicing features for robust speech detection. In: Proceedings, 9th INTERSPEECH, Lisbon, Portugal, pp. 369–372.
Ke, L., Swamy, M.N.S., Ahmad, M.O., 2005. An improved voice activity detection using higher order statistics. IEEE Trans. Speech Audio Process. 13 (5), 965–974. http://dx.doi.org/10.1109/TSA.2005.851955.
Low, S.Y., Pham, D.S., Venkatesh, S., 2013. Compressive speech enhancement. Speech Commun. 55 (6), 757–768.
Mandal, S., Bhavsar, A., Sao, A.K., 2014. Hierarchical example-based range-image super-resolution with edge-preservation. In: IEEE International Conference on Image Processing (ICIP), pp. 3867–3871. http://dx.doi.org/10.1109/ICIP.2014.7025785.
Mousazadeh, S., Cohen, I., 2013. Voice activity detection in presence of transient noise using spectral clustering. IEEE Trans. Audio Speech Lang. Process. 21 (6), 1261–1271. http://dx.doi.org/10.1109/TASL.2013.2248717.
Ng, T., Zhang, B., Nguyen, L., Matsoukas, S., Zhou, X., Mesgarani, N., Veselý K., Matejka, P., 2012. Developing a speech activity detection system for the DARPA RATS program. In: Proceedings, 13th INTERSPEECH, Portland, Oregon.
Paliwal, K., Shannon, B., Lyons, J., Wojcicki, K., 2009. Speech-signal-based frequency warping. IEEE Signal Process. Lett. 16 (4), 319–322. http://dx.doi.org/10.1109/LSP.2009.2014096.
Pradhan, G., Prasanna, S.R.M., 2013. Speaker verification by vowel and nonvowel like segmentation. IEEE Trans. Audio Speech Lang. Process. 21 (4), 854–867. http://dx.doi.org/10.1109/TASL.2013.2238529.
Prasanna, S., Reddy, B.V.S., Krishnamoorthy, P., 2009. Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Trans. Audio Speech Lang. Process. 17 (4), 556–565. http://dx.doi.org/10.1109/TASL.2008.2010884.
Raj, C.S., Sreenivas, T.V., 2011. Time-varying signal adaptive transform and IHT recovery of compressive sensed speech. In: Proceedings, 12th INTERSPEECH, Florence, Italy, pp. 73–76.
Ramírez, J., Gorriz, J.M., Segura, J.C., 2007. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In: Robust Speech Recognition and Understanding. InTech.
Sadjadi, S., Hansen, J., 2013. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process. Lett. 20 (3), 197–200. http://dx.doi.org/10.1109/LSP.2013.2237903.
Saeedi, J., Ahadi, S., Faez, K., 2013. Robust voice activity detection directed by noise classification. J. Signal Image Video Process., 1–12 http://dx.doi.org/10.1007/s11760-013-0479-5.
Shahnaz, C., Zhu, W.P., Ahmad, M.O., 2006. A multifeature voiced/unvoiced decision algorithm for noisy speech. In: Proceedings, IEEE International Symposium on Circuits and Systems (ISCAS), pp. 2525–2528. http://dx.doi.org/10.1109/ISCAS.2006.1693137.
Sharma, P., Abrol, V., Sao, A.K., 2015a. Supervised speech enhancement using compressed sensing. In: Proceedings, IEEE Twenty First National Conference on Communications (NCC), pp. 1–5. http://dx.doi.org/10.1109/NCC.2015.7084919.
Sharma, P., Abrol, V., Sao, A.K., 2015b. Compressed sensing for unit selection based speech synthesis system. In: Proceedings, European Signal Processing Conference (EUSIPCO).
Sharma, P., Abrol, V., Dileep, A.D., Sao, A.K., 2015c. Sparse coding based features for speech units classification. In: Proceedings, 16th INTERSPEECH, ISCA.
Shi, Y., Soong, F.K., lai Zhou, J., 2006. Auto-segmentation based partitioning and clustering approach to robust endpointing. In: Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, Toulouse, France, pp. I–I. http://dx.doi.org/10.1109/ICASSP.2006.1660140.
Soon, I.Y., Koh, S.N., Yeo, C.K., 1999. Improved noise suppression filter using self-adaptive estimator of probability of speech absence. Signal Process. 75 (2), 151–159. http://dx.doi.org/10.1016/S0165-1684(98)00230-8.
Sreenivas, T.V., Kleijn, W.B., 2009. Compressive sensing for sparsely excited speech signals. In: Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Taipei, Taiwan, April 2009, pp. 4125–4128. http://dx.doi.org/10.1109/ICASSP.2009.4960536.
Sri Rama Murty, K., Yegnanarayana, B., Joseph, M.A., 2009. Characterization of glottal activity from speech signals. IEEE Signal Process. Lett. 16 (6), 469–472. http://dx.doi.org/10.1109/LSP.2009.2016829.
Suk, S.-Y., Chung, H.-Y., Kojima, H., 2007. Voice/non-voice classification using reliable fundamental frequency estimator for voice activated powered wheelchair control. Proceedings, 3rd International Conference on Embedded Software and Systems (ICESS). Springer, Verlag, Berlin, Heidelberg, pp. 347–357.
Teng, P., Jia, Y., 2013. Voice activity detection via noise reducing using non-negative sparse coding. IEEE Signal Process. Lett. 20 (5), 475–478. http://dx.doi.org/10.1109/LSP.2013.2252615.
Tosic, I., Frossard, P., 2011. Dictionary learning. IEEE Signal Process. Mag. 28 (2), 27–38. http://dx.doi.org/10.1109/MSP.2010.939537.
Yang, G., Leich, H., Boite, R., 1995. Voiced speech coding at very low bit rates based on forward–backward waveform prediction. IEEE Trans. Speech Audio Process. 3 (1), 40–47. http://dx.doi.org/10.1109/89.365382.
Yang, J., Liao, X., Yuan, X., Llull, P., Brady, D.J., Sapiro, G., Carin, L., 2015. Compressive sensing by learning a gaussian mixture model from measurements. IEEE Trans. Image Process. 24 (1), 106–119. http://dx.doi.org/10.1109/TIP.2014.2365720.
Yegnanarayana, B., Sharat Reddy, K., Kishore, S.P., 2001. Source and system features for speaker recognition using AANN models. In: Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, Utah, USA, pp. 409–412. http://dx.doi.org/10.1109/ICASSP.2001.940854.
Ying, D., Yan, Y., Dang, J., Soong, F.K., 2011. Voice activity detection based on an unsupervised learning framework. IEEE Trans. Audio Speech Lang. Process. 19 (8), 2624–2633. http://dx.doi.org/10.1109/TASL.2011.2125953.
Ykhlef, F., Bendaouia, L., 2012. Evaluation of time domain features for voiced/non-voiced classification of speech. In: Proceedings, International Conference on Signals and Electronic Systems (ICSES), pp. 1–4. http://dx.doi.org/10.1109/ICSES.2012.6382213.
You, D., Han, J., Zheng, G., Zheng, T., 2012. Sparse power spectrum based robust voice activity detector. In: Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Koyoto, Japan, pp. 289–292. http://dx.doi.org/10.1109/ICASSP.2012.6287874.
Zeng, L., Zhang, X., Chen, L., Fan, Z., Wang, Y., 2012. Scrambling-based speech encryption via compressed sensing. EURASIP J. Adv. Signal Process. 2012 (1), 1–12.