Incremental Acoustic Subspace Learning for Voice Activity Detection using Harmonicity-based Features

Jiaxing Ye, Takumi Kobayashi, Masahiro Murakawa, Tetsuya Higuchi

National Institute of Advanced Industrial Science and Technology (AIST), Japan
(jiaxing.ye, takumi.kobayashi, m.murakawa, t-higuchi}@aist.go.jp

Abstract
This paper presents novel voice activity detection (VAD) approach based on incremental subspace learning using harmonicity-based features. Harmonic structure is well known as noise robust speech feature. We develop novel harmonicity-based feature based on temporal-spectral co-occurrence patterns. At statistical decision stage, many conventional statistical VAD methods rely on Gaussian model; however, owing to the non-Gaussian nature in speech, Gaussian model becomes faulty and produces incorrect VAD results. We reformulate the VAD by incremental subspace learning. The candid covariance-free incremental PCA (CCIPCA) subspace method is employed to adaptively model the input sound by a subspace. Subsequently, a speech activity measure can be established based on the distance from input sound to the adaptive subspace. Notably, the CCIPCA subspace update interval is set to 0.5 second in this work and the deviation distance is computed afterwards. In such short time scale, environmental sound present more Gaussian-like/stationary pattern and therefore can be well accommodated by adaptive subspace, conversely, speech always exhibit non-stationary characteristic which lead to distinct deviation to the adaptive acoustic subspace, and thus, can be effectively distinguished. We experimentally compared our scheme with various VAD methods over real-world data. The results validate the effectiveness of the proposed approach.

Index Terms: voice activity detection, Harmonic feature, higher-order local auto-correlations, CCIPCA

1. Introduction
Voice activity detection (VAD) is designated to distinguish speech from background noises. By producing effective indications for further process, VAD scheme reduces the bandwidth usage in data communication as well as the computation cost in various discontinuous transmission applications, such as in voice over IP (VOIP) and mobile telephony. Also, the performance of VAD contributes to the accuracy of speaker recognition and automatic speech recognition (ASR) by effectively removing non-speech segments. Conventional VAD approaches operate well for high signal-to-noise-ratio (SNR) situations (SNR>10dB), however, the performance degenerates sharply under the hostile acoustic conditions (SNR<10dB). Motivated by increasing demands from various applications, noise robust VAD received considerable attentions from speech processing research field in recent years.

Typical modern VAD systems consist of two main phases: the acoustic feature extraction phase and the speech/non-speech decision phase. Over the past two decades, various VAD features have been developed, such as energy-based features of low band to full band energy (LFER), spectral-domain features for noise power estimation and subtraction [1] and delta cepstrum features capturing the dynamic changes raised by speech activities [2]. In addition, given the fact that speech signals are always non-stationary with wide-time-frequency variations, spectral entropy is introduced to measure the degree of non-stationarity for noise robust VAD [3]. Speech harmonicity, well known as the most discriminant VAD feature characterizing the unique harmonic structures of speech signal, has been extensively investigated lately [4]. VAD features can also be extracted not only from individual frame (10-20ms), but from a set of contiguous frames (hundreds of milliseconds). These so-called long-term features accumulate the variability difference between speech and non-speech signals and exhibit more robust VAD performance. Several VAD features can be combined together to attain better results, e.g. a VAD method operates on sub-band spectral entropy covering long-term range of hundreds of milliseconds proposed in [5].

Plenty of decision rules have been developed to judge the absence/presence of speech under low SNR situations. Statistical model-based VAD methods are intensively studied ones, in which the VAD is cast into by hypothesis testing over statistical models of speech absence and presence based on likelihood ratio test (LRT) [6]. The latter work extended the single frame-based LRT to consecutive frames-based version (MO-LRT) [7]. The method reported in [4] combines harmonicity-based feature with MO-LRT scheme and achieves some improvements. Besides, many sophisticated machine learning techniques have also been employed for VAD task, such as support vector machine [8], nonetheless, the algorithms’ complexity become the obstacle for practical applications.

In this paper, we propose a noise robust VAD approach based on incremental acoustic subspace learning using novel harmonicity-based features. Harmonicity is widely regarded as discriminative speech feature, which is credible even under extremely noisy conditions. We develop temporal-spectral co-occurrence features to extract the distinguishable speech harmonicity based on higher-order local correlations (HLAC) feature mask patterns. Prototype of HLAC feature is well proposed in computer vision field for describing joint spatial features in local region [9]. In this work, we adapt HLAC features to characterize the variation patterns in local time and frequency domains, particularly the harmonic structures, in acoustic signals.

At VAD decision phase, Gaussian model played a fundamental role in conventional statistical model-based VAD methods. However, Gaussian model becomes unreliable for characterizing speech signal, which is typically non-Gaussian distributed, and thus deteriorate the VAD accuracy. To cope with such drawback, we reformulate VAD task based on online subspace learning, by which the input sound is incrementally characterized by an acoustic subspace with short-term updating intervals (<1 sec). Based on the short-term time division, environmental sounds are anticipated to present...
more stationary property and can be characterized well by the subspace. Conversely, due to the significant variability, the speech would always exhibit distinct deviation distance to the subspace and can therefore be correctly discerned. To realize this scheme, the candidate covariance-free incremental PCA (CCIPCA) [10] is adopted to statistically model the input sound by a subspace, which offers three main merits: first, through performing CCIPCA, a low rank data representation preserved with global structure is obtained, which is favorable for describing the intrinsic time-frequency distributions of acoustic signal and therefore establishes the basis for speech activity (outlier) detection; second, as an unsupervised incremental learning algorithm, CCIPCA endows the proposed VAD scheme with self-adaption property for working under various noise conditions; third, since CCIPCA operates fast based on simple matrix multiplications and additions, it is convenient for practical applications. An effective speech activity measure can be established based on the deviation distance from the input sound to adaptive acoustic subspace generated by CCIPCA. We incorporate hang-over scheme to retrieve the unvoiced segments, in which none harmonicity is presented. An adaptive threshold is applied in this work to retrieve the unvoiced segments, in which none harmoincity is presented. Hence, the deviation distance, that is residual of x to the adaptive acoustic subspace V, that indicating the membership measure of x to be speech, can be defined as:

\[
d(x) = \| x - P(x) \| = \begin{cases} x - P_{H_0}(x) & x \in H_0 \\ x - P_{H_1}(x) & x \in H_1 \end{cases}
\]  

(3)

We explain the decision formula f in (3). The non-speech sound (H0), with predominant patterns well accommodated by the adaptive acoustic subspace, would produce quite small (near zero) deviation distance. In contrast, the wide variability/non-stationary nature in speech activity (H1) would always exhibit significant deviation value and can therefore be successfully discerned.

In the following paragraphs, we introduce the acoustic feature (x) and the VAD decision rule in (3) developed in this study.

2.2. Chart flow of the proposed VAD approach

Fig.1 depicts the flow chart of the proposed VAD framework. At acoustic feature extraction stage, the acoustic signal is transformed to spectrogram by using short-time Fourier transform (STFT). Among three neighbor frames, we extract the proposed harmonicity-based features by HLAC. Then, Mel filter bank is applied to reduce the feature dimension. Subsequently, we utilize CCIPCA subspace method to incrementally model the acoustic signal by a subspace, based on which the speech activities can be effectively discerned through examining the deviation values. With a proper post processing over deviation values, we can obtain the final VAD result.

2.3. Harmonicity-based feature

In this section, we propose a harmonicity-sensitive VAD feature for characterizing the co-occurrence patterns in time and frequency domains. Most acoustic features have been extracted at each frame, such as in MFCC, in consequence, the temporal-spectral dynamic features are deteriorated. We adapt higher-order local correlations (HLAC), which is well developed for describing the geometrical joint pattern in computer vision [9], to exploit rich co-occurrence characteristics of sound signal in adjacent regions cross the time and frequency domains. Particularly, the proposed feature is effective for exploiting the discriminative harmonic structures, which are favorable for VAD.
Let time and frequency be denoted by $t$ and $v$, respectively, and $f(t,v)$ indicates the power at $(t,v)$ point over spectrogram. We employ higher-order local auto-correlations to characterize the joint co-occurrence patterns in time and frequency domains:

$$x(a_1, a_2) = f(r)f(r+a_1)f(r+a_2),$$

where $a_1$ and $a_2$ are displacement vectors and thus $f(r+a_1)$ and $f(r+a_2)$ indicate adjacent neighbors of the reference point $f(r)$. We limit $a$ within a $3x3$ local neighborhood of $r$ on the time-frequency plane, which is anticipated to be highly correlated. We introduce 9 feature patterns, as shown in Fig. 3, which are extracted at respective frequencies ($v$). Mel filter banks are then applied to the features generated by each mask pattern for reducing the dimension. Then, we concatenate all Mel-scale features at every frame $(t)$ to obtain the proposed feature vector.

![Figure 2: The proposed co-occurrence features over spectrogram.](image)

It is noteworthy that the proposed feature is favorable for characterizing the harmonic structures in speech signal, by which the distinct (large) feature value can be produced when speech harmonicity is presenting as illustrated in Fig. 3.

![Figure 3: Harmonicity feature extraction by proposed feature.](image)

### 2.4. Incremental acoustic subspace learning

In Sec. 2.1, we outlined the incremental subspace learning formulation for VAD. In this part, we explicitly explain the detail of incremental learning process in proposed VAD approach.

With random speech inserted, environmental sound is always presenting diverse acoustic signal characteristics with time-variant compositions, and thus, an unsupervised incremental learning scheme is demanded for modeling such variable signal. The candid covariance-free incremental PCA (CCIPCA) [10] subspace method renders such specifications with low computation cost.

Let $x(i = 1, \ldots, J \ldots N) \in \mathbb{R}^m$ denote $N$ frames of $M$-dimensional VAD feature vectors. We choose first $J$ frames (initiation time) for starting up the incremental eigenvector and eigenvalue estimation. The eigenvalues $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_J)$ and eigenvectors $U = [u_1, \ldots, u_M], u \in \mathbb{R}^M$ can be calculated by:

$$RU = U\Lambda, \quad R \equiv \left\{ \sum_{i=1}^{J} x_i x_i^T \right\},$$

where $x_i^T, i \in (1, \ldots, J)$ is the transpose of $x_i$. The contribution rate $\eta_i$ is defined as:

$$\eta_i = \frac{\sum_{i=1}^{J} \lambda_i / \sum_{i=1}^{J} \lambda_i}{K}, \quad (6)$$

We sort eigenvectors by eigenvalues in decreasing order, and then, keep first $K$ principle eigenvectors $U_k = [u_1, \ldots, u_k]$ with contribution rate of $\eta_k > 0.99$ to express predominant acoustic patterns. Subsequently, the eigenvectors/eigenvalues can be incrementally updated in a recursive form based on new input feature vectors as:

$$\hat{u}(t + \Delta t) = \frac{\Delta t - 1}{\Delta t} u(t) + \frac{1}{\Delta t} X_{t+\Delta t} X_{t+\Delta t}^T u(t) / \| u(t) \|, \quad (7)$$

where $u(t)$ and $\hat{u}(t + \Delta t)$ denote the eigenvector at time $t$ and the updated eigenvector at time $t+\Delta t$, respectively, and $X_{t+\Delta t} = [x_{t+1}, \ldots, x_{t+\Delta t}]$ denotes the newly input feature vectors from time $t+1$ to $t+\Delta t$. To accelerate the estimation convergence, an efficient way is to use an amnesic average and, therefore, (7) is changed into:

$$\hat{u}(t + \Delta t) = \frac{\Delta t - 1 - l}{\Delta t} u(t) + \frac{1 + l}{\Delta t} X_{t+\Delta t} X_{t+\Delta t}^T u(t) / \| u(t) \|, \quad (8)$$

where $l$ is the positive amnesic parameter usually set to 2 [10]. The eigenvalues can be updated by $\hat{\Lambda}(t + \Delta t) = \| u(t + \Delta t) \|$. Then, the projection operator to the adaptive acoustic subspace is derived as $P_{t+\Delta t} = U_{t+\Delta t} U_{t+\Delta t}^T$, in which $U_{t+\Delta t} = [u(1+\Delta t), \ldots, u(L+\Delta t)]$ denotes the estimated $K$ principle eigenvectors at time $t+\Delta t$, therefore, we obtain the analytical expression of speech measure function defined in (3), which computes the deviation distances between the adaptive subspace and input sound from $t+1$ to $t+\Delta t$ as:

$$d_i^2 = \text{diag} \left( \| X_i^T - P_{t+\Delta t} \cdot X_i^T \| \right),$$

$$= \text{diag} (X_{t+\Delta t}^T X_i^T - X_i^T \hat{U}_{t+\Delta t} X_{t+\Delta t} X_i^T). \quad (9)$$

We assume that the first 1 second sound to be non-speech, which is used to initialize the eigenvectors and eigenvalues by setting $J$ in (5). The adaptive subspace updating interval is set to 0.5 sec which can be realized by $\Delta t$ in (7); then, for the 0.5 sec latest input sound, the speech activity measure can be computed by (9). Together with short-term timescale updating scheme, the proposed subspace learning approach effectively differentiates the variability properties between speech and environmental sound. On text may be above or below the table.

### 2.5. Threshold setting and post processing

An adaptive thresholding scheme is adopted to produce more robust VAD result. The threshold is initialized by the maximum deviation for the first 1 second non-speech sound which can be computed by $\text{Thr}_0 = \max(d_i)$, where $d_i$ denotes the 1s deviation distance sequence. We adopt the threshold updating scheme in [5] to adapt to the acoustic environment variations, which is based on the previous VAD outputs. Let the deviation values of the latest detected 4th speech and 4th non-speech sound be denoted by $S(k)$ and $N(k)$ and the threshold can be updated by:

$$\text{Threshold} = \alpha \times \min(S(k)) + (1 - \alpha) \times \max(N(k)) \quad (10)$$

where $\alpha$ is a stream weighting factor which is set to 0.5 and $k$ is set to 50 in our experiment.
In addition, a hangover scheme is adopted to retrieve the unvoiced speech, which delivers none harmonic structure. Based on the study in [11], the voiced speech is usually preceded by 300ms and followed by 500ms of unvoiced speech. We experimentally set the pre-hangover and post-hangover length to 150ms and 300ms.

3. Experiments

To validate the proposed method, we conduct VAD experiments under various acoustic environments with low SNRs. First, we randomly chose some speech clips from TIMIT dataset [12] and added silent intervals to build the 218s’ test corpora containing 58.9% speech. We manually labeled the speech activity for every 10ms and the pause regions between words, which are smaller than 100ms, are set to speech. The factory and babble noises are extracted from NOISEX-92 database [13]. In addition, to evaluate our method under real-world acoustic environments, we extracted public foyer and shopping center sound scenes from BBC Sound Effect Database [14]. All noises are added to clean speech with 5dB and 0dB SNRs.

The window length of STFT is set to 10ms. 15 filters are employed to build Mel-filter bank and therefore the feature dimension is 135 (9 feature mask patterns x 15 Mel-filters). The acoustic subspace updating interval is set to 0.5 second which equals the deviation distances computation period. We adopt true positive rate (TPR) and false positive rate (FPR) to produce receiver operating characteristic (ROC) curve for evaluating the proposed VAD approach, in which, TPR denoted the percentage of correctly detected reference speech frames, while the FPR is the percentage of reference non-speech frames mistakenly identified as speech. We drew extensive comparison with other VAD methods [4,5,6,15]. To plot the ROC curve, we disabled adaptive thresholding schemes in our method and in [5]. Comparison experiments had been performed under all four types of acoustic environments with 5dB and 0dB SNRs. Moreover, we validate the effectiveness of the proposed harmonicity-based feature through comparing to the acoustic spectrum.

Based on the comparison results presented in Fig.4, comparing with the conventional VAD methods, the proposed VAD scheme achieved superior performance and exhibited more robust characteristic for the increased noise intensity under all circumstances. Meanwhile, the proposed harmonicity-based features have also been validated, which greatly improved in VAD accuracy comparing to the case of applying spectrum feature.

4. Conclusions

This paper presented a novel noise robust voice activity detection approach. A harmonicity-based acoustic feature is proposed to characterize the discriminative harmonic structures in speech signal. At decision making stage, we established a speech activity measure based on the deviation distance from input sound to the adaptive acoustic subspace. The environmental sounds, which are deemed to be Gaussian/stationary for short-term timescale (<1 second), can be well accommodated by a subspace. In contrast, speech signal will always exhibit distinct deviation values due to the significant variability. Based on a proper post processing scheme, the proposed VAD approach outperformed conventional VAD schemes in the comparison experiments. The experimental results also validated the effectiveness of the proposed harmonicity-based features.

5. Acknowledgements

This work was partly supported by the Japan Science and Technology Agency (JST) A-STEP project.
6. References


[14]. The BBC sound effects library original series. Online: http://www.sound-ideas.com