Dişken et al. EURASIP Journal on Audio, Speech, and Music Processing
(2017) 2017:23
DOI 10.1186/s13636-017-0120-6

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

# A robust polynomial regression-based voice activity detector for speaker verification

Gökay Dişken[1*], Zekeriya Tüfekci[2] and Ulus Çevik[3]

**Abstract**

Robustness against background noise is a major research area for speech-related applications such as speech recognition and speaker recognition. One of the many solutions for this problem is to detect speech-dominant regions by using a voice activity detector (VAD). In this paper, a second-order polynomial regression-based algorithm is proposed with a similar function as a VAD for text-independent speaker verification systems. The proposed method aims to separate steady noise/silence regions, steady speech regions, and speech onset/offset regions. The regression is applied independently to each filter band of a mel spectrum, which makes the algorithm fit seamlessly to the conventional extraction process of the mel-frequency cepstral coefficients (MFCCs). The k-means algorithm is also applied to estimate average noise energy in each band for spectral subtraction. A pseudo SNR-dependent linear thresholding for the final VAD output decision is introduced based on the k-means energy centers. This thresholding considers the speech presence in each band. Conventional VADs usually neglect the deteriorative effects of the additive noise in the speech regions. Contrary to this, the proposed method decides not only for the speech presence, but also if the frame is dominated by the speech, or the noise. Performance of the proposed algorithm is compared with a continuous noise tracking method, and another VAD method in speaker verification experiments, where five different noise types at five different SNR levels were considered. The proposed algorithm showed superior verification performance both with the conventional GMM-UBM method, and the state-of-the-art i-vector method.

**Keywords:** Polynomial regression, Robust speaker recognition, Voice activity detection

## 1 Introduction

Automatic speaker recognition systems' performances are greatly improved in the last two decades, especially with the introduction of the modeling methods such as universal background model (UBM) [1] and i-vectors [2]. For the front end, mel-frequency cepstral coefficients (MFCCs) [3] are extensively preferred by the researchers in speaker or speech recognition systems. Despite the high performance of the MFCCs in controlled environments, their performance degrades in adverse conditions, such as convolutive channel noise, additive background noise, and reverberation. Many different techniques have been developed to overcome this issue such as using a voice activity detector (VAD), extracting robust features, and speech enhancement. Some of them are mentioned in the following paragraphs.

Instead of using the MFCCs, other types of features are proposed by the researchers to increase the robustness of the recognizers [4–8]. Also, since the MFCCs are widely adopted, many researchers have made effort to improve its robustness under noise by modifying, or changing, some processes in the conventional scheme [9–13]. Interested readers may refer to [14] for the recent progress in the feature extraction techniques for robust speaker recognition.

Another way to improve the performance of the recognizers is to enhance the speech. These methods usually require an estimation of the level and frequency response of the noise. When the noise has been estimated, the enhancement can be achieved via methods such as spectral subtraction [15] or Wiener filtering as in [16]. Once the signal is enhanced, it can be treated as a clean signal and can be used in further operations (i.e., feature

* Correspondence: gdisken@adanabtu.edu.tr
[1]Department of Electrical-Electronics Engineering, Adana Science and Technology University, Adana, Turkey
Full list of author information is available at the end of the article

extraction). A popular method for the noise estimation is based on the minimum statistics as proposed in [17] or based on its improvement [18]. Statistical differences are taken into account in [19] for non-stationary noise power spectral density estimation. Relying on the speech/non-speech decisions of the frames (by considering the harmonics for the voiced speech and durations for the unvoiced speech), a noise tracking algorithm is proposed [20]. Speech presence probability is considered in [21], which depends on the relation of the adjacent spectral components.

The methods referred in the previous paragraph are aiming to track the noise spectrum even under the speech activity. They are fast adapting to the changes in the noise level but may degrade the information during the speech regions. Since both the noise and the speech are present at the same time, exactly separating them is not possible. On the other hand, separating the speech-dominant and the noise-dominant frames can also be beneficial. It is also possible to estimate an average of the noise from the noise-dominant frames, or the noise parameters can be updated between the speech regions. Various voice activity detection algorithms have been developed for this purpose. In [22], long-term signal probabilities are used as the discrimination criteria and the decision is assigned to the frame in the middle of a long window. Similarly, in [23], long-term signal variability is used but the decision is assigned to the long window. Energy-based detection and Gaussian mixture model (GMM)-based statistical model are combined for the VAD in [24]. Modeling the feature distribution with a bi-Gaussian model can be also used as a VAD, where the Gaussian with the lower mean corresponds to noisy frames [25]. Four different voicing measures and a spectral feature are concatenated and mapped to a one-dimensional space via principal component analysis in [26], and the resulting feature distribution is bi-Gaussian. In addition to the four voicing measures used in [26], MFCCs and two pitch trackers are used as features for the VAD in [27]. A comparison of some standard VAD methods for speaker recognition can be found in [28], where the bi-Gaussian modeling-based VAD is reported to be the best performing one. Similar to the VAD, vowel-like regions are used in [29] and improved in [30] by including the non-vowel-like regions. Missing data approach is also investigated in several studies [31–34], where a binary time-frequency mask is constructed for the noisy spectrum to indicate reliable and unreliable features. The unreliable features are then reconstructed, or marginalized (ignored in score computation).

Usually, VADs look for clues of speech presence (i.e., harmonicity, periodicity, energy, long-term variability) but do not give any idea about if the found speech region is useful for recognition, or how many bands include speech information. As mentioned before, exactly separating the noise and the speech is not possible. Therefore, an output speech of a robust VAD may not cover the same speech/speaker information as its clean equivalent does. Especially for the low SNR signals, directly using the VAD output is not suitable, because of the nonlinear transformation effect of the additive noise on the low energy frames [35]. Missing data approach tries to compensate this issue; however, its performance highly depends on the estimation of the spectral mask, where a poor estimation reduces the recognition performance [36].

In this paper, a polynomial regression-based algorithm is proposed as a solution to the aforementioned problem, that is, separating the useful speaker/speech information from the additive noise. A polynomial is fitted on the energy output over a range of several frames, and the output is not strictly frame-dependent. Also, suppression of sudden energy ripples is possible with this method, which is achieved by smoothing in some studies [27]. Besides the polynomial regression, another important novelty of the proposed algorithm is the pseudo SNR-dependent linear thresholding that takes account of the speech presence in each band. Since the algorithm is unsupervised, it is suitable for any SNR level and any stationary additive noise, unlike supervised methods such as neural networks, which may perform inadequately under unseen environments. To verify the performance of the proposed method, speaker verification experiments with both male and female data were made. Five different noise types at five different overall SNR levels were added to the test files. Both conventional UBM and state-of-the-art i-vector frameworks were investigated.

## 2 Polynomial regression on mel spectrum

The proposed method can be separated into two parts. As the first part, the polynomial regression, k-means classification, and speech enhancement applied on the filter bank outputs are explained in Sections 2.1, 2.2, 2.3, 2.4, and 2.5. For the second part, based on the results from the first part of the proposed algorithm, the final VAD output decision is given in Section 2.6.

### 2.1 Motivation

The VADs, in general, use parameters, such as periodicity, harmonicity, and long-term variability, or energy levels to detect speech regions, then make frame-based speech/non-speech decision. For the voiced phonemes, these parameters could be sufficient, because of their harmonic structure and relatively high energy. However, the unvoiced phonemes may be treated as noise since they do not possess harmonic content as the voiced phonemes do and have less energy. To overcome this problem, polynomial regression is applied on the output

of each filter band of the mel scale spectrum, and the frames are grouped together by finding the best fitting polynomial over the range of five to ten frames. For a 25-ms frame with a 10-ms overlap, this range is equal to 65–115 ms, which covers the average duration of a vowel-like (i.e., vowel, semivowel, and diphthong) region in a continuous speech [29]. Another important issue is that the temporal contour of speech energy of a filter bank output approximately resembles a bell-like shape in the spectrum, i.e., energy level rising from the speech onset point and vice versa. This phenomenon was also considered in [37], where an end point detection method was proposed with the aid of the aforementioned rising and falling energy levels (called beginning and ending edges, respectively). With a basic energy thresholding for VAD, low energy frames, especially unvoiced phonemes, are likely to be suppressed in low SNR levels. In the proposed algorithm, the grouped frames are represented by the average energy of the fitted polynomial, which can be thought as an energy boost for the low energy frames in a speech region. Also, this representation avoids the misclassifications of sudden energy ripples (i.e., noise frame with a higher energy than its neighbor noise frames, or speech frame with a lower energy than its neighbor speech frames). It should be noted that a single polynomial may not capture the entire bell-like shape of a speech segment, due to the frame range (5–10) limitations. Instead, the minimum expectation is to capture at least the rising and falling edges and the peak regions.

As a preliminary experiment, the proposed algorithm is going to be tested on a sample speech signal to illustrate its effectiveness. The speech file used in the tests is an utterance spoken by a male speaker from the Noizeus corpus [38], with an 8 kHz sampling frequency. The mel spectrum of the signal is shown in Fig. 1a, which is obtained by using 25 ms long frames with 10 ms overlap, windowing with a Hamming window, taking 1024 points Fourier transform, and filtering with 26 filters that are equally spaced on the mel scale between 300 Hz and 4 kHz. Note that these parameters are the same throughout this section and also for the speaker verification experiments in Section 3. The lynx noise from the NOISEX-92 database [39] is mixed with the clean speech to obtain a noisy speech signal with a desired overall SNR level (5 dB in this case). Note that the NOISEX-92 database has a 16 kHz sampling frequency. Noise files were down-sampled to 8 kHz to match the sampling rates of speech and noise data. To achieve this, first, the NOISEX-92 data were filtered by a low pass filter (almost ideal low pass) with a cut-off frequency of 4 kHz. Then, the data were down-sampled by a factor of two. The mel spectrum of the degraded signal is illustrated in Fig. 1b.

The effect of the noise can be easily observed by comparing the spectra given in Fig. 1. The speech signals with low energies vanish into the noise. However, the regions with higher energies are still visible in the spectrum. The ultimate goal of the proposed algorithm is to detect and enhance these frames by using the polynomial regression, followed by the spectral subtraction.

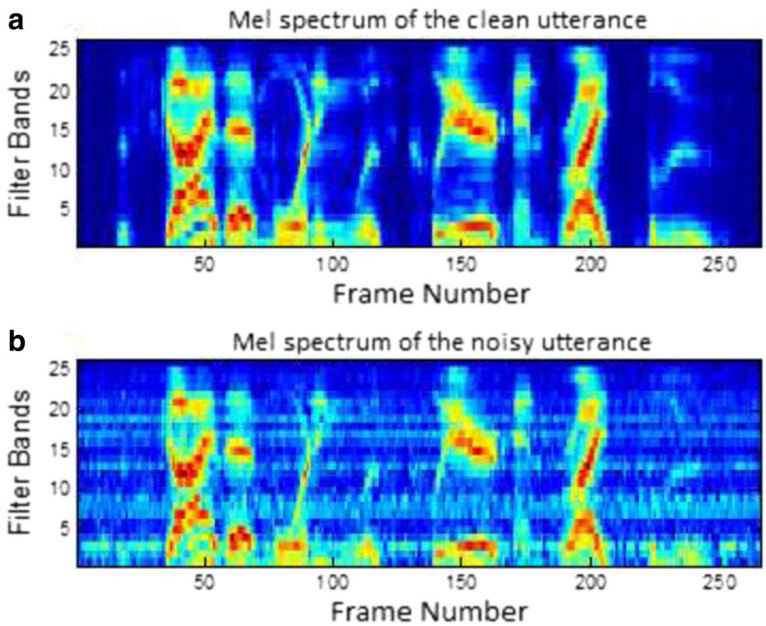


**Fig. 1** Mel spectra of the **a** clean signal and **b** the noisy signal with a 5 dB overall SNR

### 2.2 A brief review of polynomial regression

Before detailing the proposed algorithm, reminding the general polynomial regression expressions with least squares sense could be beneficial. A $k$th order polynomial is defined as $a_0 + a_1x + ... + a_kx^k$, where $x$ is the intermediate variable, and $a$'s are the coefficients of the polynomial. The least squares method minimizes the summed difference (error) of the observed value and the estimated value. This error can be defined as in Eq. 1.

$$E = \sum_{i=1}^{n} \left[y_i - \left(a_0 + a_1x_i + ... + a_kx_i^k\right)\right]^2 \tag{1}$$

where $\boldsymbol{y}$ is the observation vector (filter bank output vector for a given band), $n$ is the length of the vector (number of frames considered for regression, 5 to 10), and $i$ is the index (frame index). In the proposed method, $x_i = i$. The optimum value of a coefficient is the value that makes the partial derivative of the error, with respect to the coefficient, equal to zero (Eq. 2).

$$\begin{aligned} \frac{\partial E}{\partial a_0} &= -2\sum_{i=1}^{n}\left[y_i - \left(a_0 + a_1x_i + ... + a_kx_i^k\right)\right] = 0 \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \frac{\partial E}{\partial a_k} &= -2\sum_{i=1}^{n}\left[y_i - \left(a_0 + a_1x_i + ... + a_kx_i^k\right)\right]x_i^k = 0 \end{aligned} \tag{2}$$

Taking the terms with $y$ to one side, these equations can be written in a matrix form as given in Eq. 3. Equation 3 can be expanded as $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{a}$, where $\boldsymbol{X}$ is defined as in Eq. 4. Then, the coefficients of the polynomial can be found as $\boldsymbol{a} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \cdots & \sum_{i=1}^{n} x_i^k \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_i^k & \sum_{i=1}^{n} x_i^{k+1} & \cdots & \sum_{i=1}^{n} x_i^{2k} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^k y_i \end{bmatrix} \tag{3}$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix} \tag{4}$$

### 2.3 Application of polynomial regression on mel spectrum

The order of the polynomial, "$k$," is chosen as two in the proposed method. The reason behind this choice is that a first-order polynomial is a straight line, which is not suitable to follow the variations of a highly nonstationary signal. On the other hand, as the order of the polynomial increases, the fitted signal takes values closer to the real signal but the computational load increases. For the proposed algorithm, only an approximation of the general trend over the frames is sufficient; hence, the polynomial order was chosen as two. Also, it should be noted that no substantial advantage was found by using a third-degree polynomial.

The regression is applied in each band of the mel spectrum, directly to the smoothed filter bank energies. It can be expanded to the conventional spectrum; however, increased number of the frequency bins will also increase the computation time. Also, the mel spectrum is a part of the MFCC extraction process; hence, the proposed algorithm fits seamlessly in the conventional framework.

Let $S(t,m)$ denote the filter bank output of the noisy speech signal for frame $t$ and filter $m$. The filter bank outputs are first smoothed to reduce the ripples (Eq. 5).

$$S^s(t,m) = \sum_{n=-2}^{2} p_n S(t+n,m) \tag{5}$$

where, $S^s(t,m)$ is the smoothed filter bank output, $p_n$ is the smoothing coefficient with $p_{-2} = p_2 = 0.1$, $p_{-1} = p_1 = 0.2$, and $p_0 = 0.4$. The smoothed filter bank outputs are then subjected to the polynomial regression in each filter band independently. The regression error is defined as the normalized distance between the smoothed filter bank outputs and the fitted polynomial (Eq. 6).

$$e_{N,m} = \frac{\sqrt{\sum_{i=0}^{N-1}\left(S^s(t+i,m) - F_N(t+i,m)\right)^2}}{N}, \qquad \mathrm{N} = 5, 6, ..., 10 \tag{6}$$

where $F_N(t,m)$ is the value of the fitted polynomial at frame $t$ and filter bank $m$. $N$ is the number of the frames used for regression, and $e_{N,m}$ is the error observed for

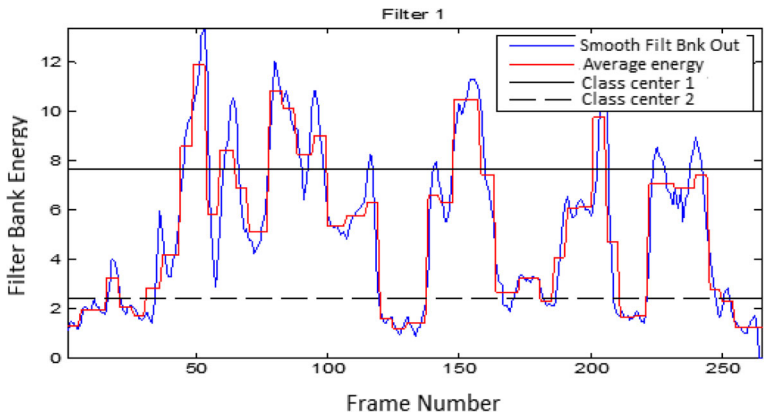


**Fig. 2** Results of the regression and the k-means in filter 1

$N$-length fitting. The consecutive frames, which give the minimum $e_{N,m}$, are grouped and represented by the average of the fitted polynomial (i.e., average energy of the group). Starting from the next frame, this process continues until all frames have been represented by their respective average polynomial energy. As an example, consider first 25 frames. $F$ is calculated from 1st frame to 5th, then 1st to 6th,…1st to 10th. Assume that minimum error is found in the range of 6 frames. So, frames 1, 2,…6 represents a group. Next, $F$ is calculated from 7th frame to 11th, then 7th to 12th, and so on. In the end, the number of the segments is at most $D/5$, where $D$ is the total number of the frames.

In Fig. 2, the results of the polynomial regression are visualized for a better understanding of the proposed algorithm, using the same degraded signal shown in Section 2.1. The blue-colored line shows the smoothed filter bank outputs of the first filter. The red line shows the frame groups represented by their respective polynomial's average energy (note the horizontal parts). Same illustration is made for the filter 24, given in Fig. 3.

### 2.4 Two class k-means algorithm

After representing all of the frames within a band as energy levels, a threshold is necessary to determine the noise-dominant or the speech-dominant groups. In order to define a threshold level, the well-known k-means algorithm is used, influenced from the success of the bi-Gaussian modeling approach [28]. The k-means algorithm can be thought as a simplified version of Gaussian mixture models, where the weights and the variances of the classes are assumed to be equal. Also, since only two classes are needed and the number of the data points (frames) is low (at most $D/5$ points as described above), a simple classification should be sufficient. The k-means algorithm gives two energy levels as class centroids. The frames that belong to the higher energy class are assumed to be speech-dominant frames. On the other hand, low energy speech regions

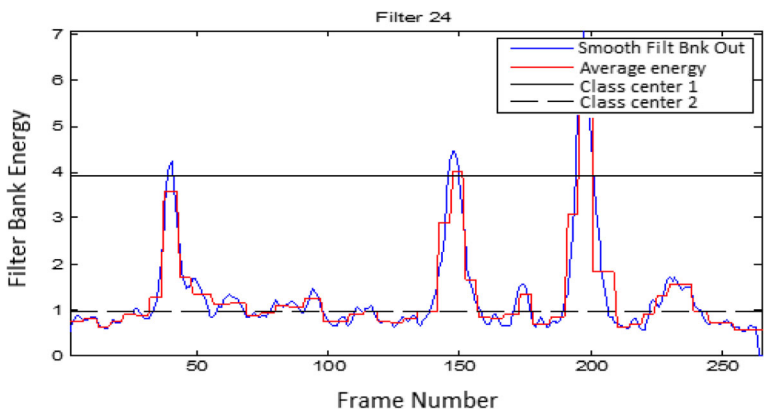


**Fig. 3** Results of the regression and the k-means in filter 24

contribute to the lower energy class. Therefore, the lower energy center is chosen as the threshold and only the frames below this threshold are treated as noise-dominant frames. Then, the average noise energy in the band can be calculated by using these frames.

The horizontal lines in Figs. 2 and 3 show the class centroids found by the k-means algorithm. The lower line (dashed), which is the centroid of the lower energy class, is used as the threshold level for the filter band. The frames below the threshold are used to estimate the average noise energy in each band.

It can be observed from the figures that as the speech energy decreases, the class centroids come closer to each other. This situation is clearly seen by comparing the class centers of the two filter bands. Filter 24 covers the higher frequencies, where less speech energy is found usually. Therefore, the difference between the class centroids is reduced. As the overall SNR is decreased, it is expected that the class centroids in the other filters also come closer. This information will be used in Section 2.6 to estimate how noisy the signal is.

### 2.5 Speech enhancement

Once the noise energy for a band has been estimated, the spectral subtraction method can be used to enhance the speech information. To avoid the spectrum becoming too small, an energy floor is also included as expressed in Eq. 7.

$$S^e(t,m) = \max(S^s(t,m) - N(m), 0.001S^s(t,m)) \quad (7)$$

where $S^e(t,m)$ is the enhanced signal and $N(m)$ is the estimated noise energy for $m$th filter.

At the end of the regression and k-means processes, each of the filter bank outputs is divided into one of the two classes, indicating the reliable and unreliable components similar to the missing data method as mentioned in Section 1. However, the binary matrix of the proposed algorithm is used in a way that it aids the final decisions on the frames. Let the ones in the matrix denote the reliable components and the zeros denote the unreliable components. If the ones dominated a frame, it is highly probable that the frame is a speech frame. However, as the noise level increases, the number of ones also tends to increase, due to misclassifying the noise frames as speech in filter bands. Therefore, a constant threshold value for all SNR levels is not suitable for the final decision. A pseudo SNR-dependent threshold called "clarity level" is defined to tackle the aforementioned problem, which is detailed in the following subsection. The procedures described above to obtain binary representation of the frames can be summarized as a block diagram given in Fig. 4.

### 2.6 Clarity level and VAD output

The clarity level is a parameter that is related to the k-means class centroids in a band. Mathematically, it is defined as in Eq. 8.

$$L = \frac{\sum_{m=1}^{M} \log_{10}\left(C^{hi}(m)/C^{low}(m)\right)}{M} \quad (8)$$

where $L$ is the clarity level, $C^{hi}(m)$ is the centroid of the speech-dominant class, and $C^{low}(m)$ is the centroid of the noise-dominant class, of $m$th filter, and $M$ is the total number of filters in the filter bank. It should be noted that $C^{low}(m)$ is also used as the threshold as described in Section 2.4.

As the noise level in the signal increases, the value of the $L$ decreases. This is mostly due to the fact that the class centroids get closer to each other since the speech vanishes into the noise. Therefore, if the clarity level is high, the signal is treated as a high SNR signal and a small number of ones found in a frame of the binary matrix is sufficient to determine that the frame is dominated by the speech. If the clarity level is low, more evidence is needed to depict a frame as the speech, so the total number of ones in a frame should be more.

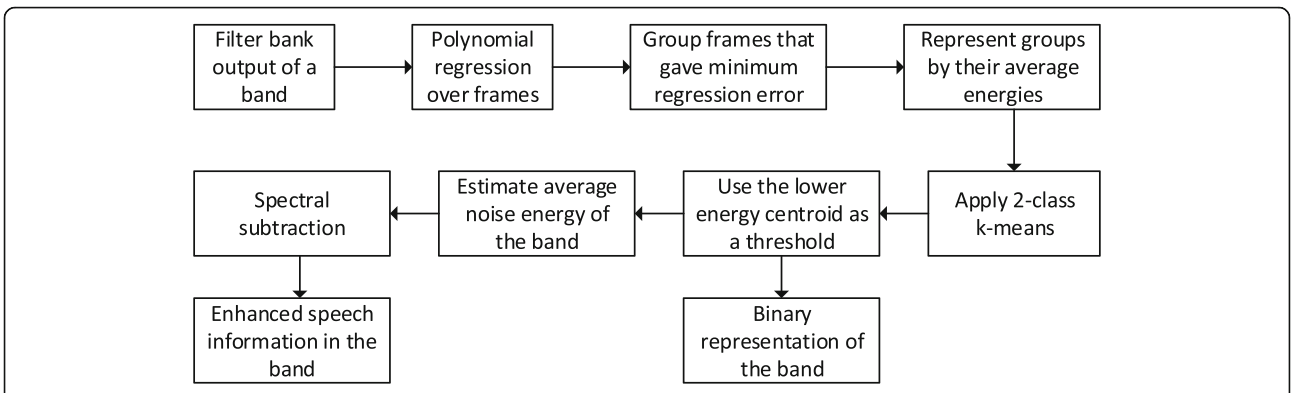


**Fig. 4** Block diagram of the processes to observe enhanced speech information and binary representation of the frames in a given band

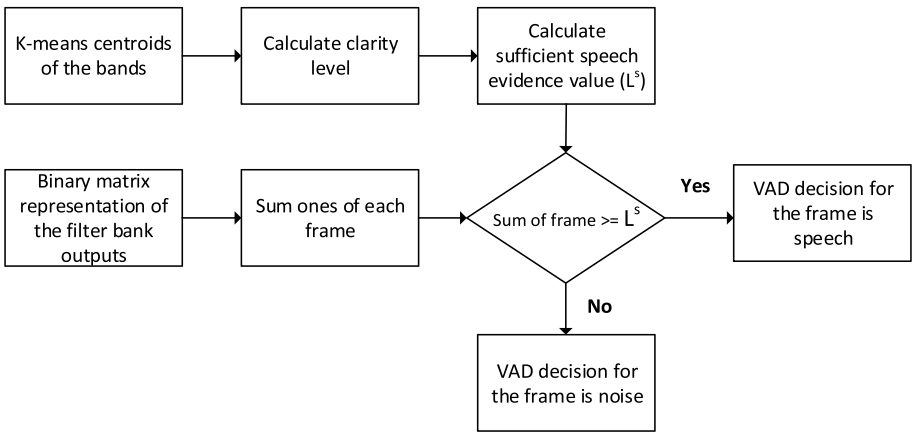


**Fig. 5** VAD final decision processes as a block diagram

The relation between the clarity level and the sufficient speech evidence value for a signal can be expressed as in Eq. 9, which is found by setting the limits for the best and worst SNR cases as 7 and 23, respectively, then simply calculating the equation of the line between them.

$$L^s = \begin{cases} 7, L > 0.8 \\ \text{round}\,(28.36 - 25.45^* L), 0.8 \geq L \geq 0.25 \\ 23, L < 0.25 \end{cases} \quad (9)$$

where $L^s$ is the minimum number of ones required for a frame to be detected as speech (sufficient speech evidence number) and is calculated per utterance. After comparing the ones in the frames of the binary matrix with the sufficient speech evidence value, frames detected as the noise are eliminated and the other frames observed as the final VAD output are subjected to discrete cosine transform to obtain the MFCCs.

The best-case limit of $L^s$, 7, is determined by assuming the speech signal should cover at least several bands for a clean signal, and it is found that this limit is not critical as the worst-case limit (choosing 6, 7, or 8 gave similar verification performances). On the other hand, the worst-case limit ($L < 0.25$) indicates a severely degraded signal; therefore, much more evidence is needed. Eighteen and 23 were examined as the worst-case limits, and $L^s = 23$ improved speaker verification performance in terms of equal error rate (EER) for the low SNR levels (− 5 and − 10 dB) as high as absolute 10% compared to $L^s = 18$. Hence, $L^s = 23$ is chosen as the worst-case limit. This result also supports the assumption made above; more speech evidence in a frame is needed for the low SNR levels. Note that total number of the filters in the bank is 26, and the worst-case limits were not reached for the SNR levels used in the experiments, they rather adjust the slope of the threshold line. The values of $L$ were determined on a sample signal described in Section 2.1, which is not included in any part of the speaker verification process. Values higher than 0.8 indicate almost clean speech (SNR > 15 dB), and values lower than 0.25 correspond to noisy speech (SNR < − 10 dB).

The VAD decision process is summarized as a block diagram in Fig. 5.

As a preliminary experiment, noisy speech spectrum, enhanced speech spectrum, and final VAD output decisions are shown together in Fig. 6 for the utterance degraded with the lynx noise at 0 dB overall SNR value. It is clear that the proposed algorithm can detect relatively high energy regions, but more proofs are needed to understand if the proposed method adds robustness to recognition systems or deteriorates the speech/speaker information. Therefore, speaker verification

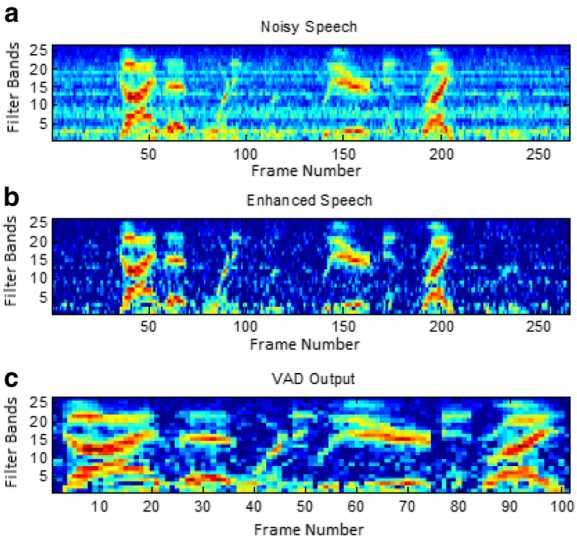


**Fig. 6** **a** Noisy speech signal with a 0 dB overall SNR, **b** enhanced speech signal, and **c** final VAD output (only the frames detected as speech-dominant) of the proposed algorithm

experiments are conducted with five different noises and five different SNR levels for each noise in the next section.

## 3 Speaker verification experiments

### 3.1 Description of the experimental setup

In the experiments, 250 male and 250 female speakers from the NIST SRE 1998 database were used, where there are approximately 5 min training data per speaker. For the tests, speech segments with 30 s duration were used. There are 1308 test speech files for male speakers and 1379 test files for female speakers. For each test file, there is one trial for the target speaker and nine trials for the non-target speakers. A more detailed analysis of the database can be found in [40]. A simple energy-based VAD as given in [41] was applied to clean training data to eliminate silence. Since the training data is clean, the type of VAD will not make much difference on the verification performance. For each speaker, only one model is trained by using the speaker's training data. The proposed and baseline methods are applied only to the test utterances. In the test phase, the features were extracted only from the speech-dominant frames for all methods. The feature vectors used in the experiments were 26-dimensional (13 MFCCs, excluding the zeroth coefficient and their deltas).

For the back end, GMM-UBM [1] and the i-vector [2] methods were considered. For the GMM-UBM, two gender-dependent UBM models with 1024 mixtures were trained by pooling all available gender-dependent training data. The speaker models were then adapted from their respective gender's UBM with a relevance factor of 16. For the i-vector framework, same UBMs and the pooled training data were used to train the total variability matrix in 20 iterations, and then, 100 dimensional i-vectors were

**Table 1** Male speaker verification results of GMM-UBM method in terms of percent EER (minDCF) for the proposed algorithm, Drugman's VAD method [27], and Rangachari's noise tracking method [21]. The last columns show the relative percent EER reduction rates compared to Drugman's VAD and Rangachari's method, respectively

| Noise type | SNR level (dB) | Proposed algorithm | Drugman's VAD | Rangachari's noise tracking | EER reduction compared to Drugman's | EER reduction compared to Rangachari's |
|---|---|---|---|---|---|---|
| Lynx | − 10 | 34.25 (0.64) | 46.10 (0.85) | 47.4 (0.87) | 25.70 | 27.74 |
| | − 5 | 25.30 (0.47) | 32.18 (0.60) | 39.22 (0.72) | 21.38 | 35.49 |
| | 0 | 15.29 (0.28) | 14.60 (0.27) | 22.47 (0.42) | − 4.72 | 31.95 |
| | 5 | 8.41 (0.15) | 8.41 (0.15) | 13.45 (0.24) | 0 | 37.47 |
| | 10 | 5.42 (0.10) | 6.50 (0.12) | 9.93 (0.18) | 16.61 | 45.41 |
| F16 | − 10 | 41.28 (0.78) | 48.31 (0.88) | 48.16 (0.89) | 14.55 | 14.28 |
| | − 5 | 31.88 (0.60) | 41.82 (0.80) | 45.18 (0.84) | 23.77 | 29.43 |
| | 0 | 20.87 (0.39) | 24.38 (0.46) | 33.4 (0.60) | 14.40 | 37.51 |
| | 5 | 11.85 (0.22) | 11.54 (0.21) | 18.19 (0.34) | − 2.68 | 34.85 |
| | 10 | 6.95 (0.13) | 7.8 (0.14) | 12.46 (0.23) | 10.89 | 44.22 |
| Car | − 10 | 5.96 (0.10) | 6.27 (0.11) | 8.94 (0.16) | 4.94 | 33.33 |
| | − 5 | 4.74 (0.08) | 5.88 (0.10) | 8.35 (0.15) | 19.38 | 43.23 |
| | 0 | 4.35 (0.08) | 5.50 (0.10) | 8.18 (0.15) | 20.91 | 46.82 |
| | 5 | 4.05 (0.07) | 5.27 (0.09) | 7.95 (0.14) | 23.15 | 49.05 |
| | 10 | 4.05 (0.07) | 5.12 (0.09) | 7.95 (0.14) | 20.90 | 49.05 |
| Babble | − 10 | 36.85 (0.69) | 48.08 (0.87) | 47.85 (0.88) | 23.35 | 22.98 |
| | − 5 | 26.83 (0.50) | 38.45 (0.72) | 43.94 (0.87) | 30.22 | 42.84 |
| | 0 | 17.50 (0.33) | 19.49 (0.36) | 28.28 (0.51) | 10.21 | 38.11 |
| | 5 | 10.01 (0.18) | 10.16 (0.18) | 14.52 (0.27) | 1.47 | 31.06 |
| | 10 | 6.72 (0.12) | 7.26 (0.13) | 10.93 (0.20) | 7.44 | 38.51 |
| Stitel | − 10 | 42.66 (0.79) | 47.17 (0.86) | 45.18 (0.84) | 9.56 | 5.57 |
| | − 5 | 33.71 (0.62) | 37.23 (0.69) | 37.15 (0.69) | 9.45 | 9.26 |
| | 0 | 19.95 (0.37) | 19.26 (0.36) | 20.41 (0.38) | − 3.58 | 2.25 |
| | 5 | 9.40 (0.17) | 9.71 (0.18) | 11.62 (0.21) | 3.19 | 19.10 |
| | 10 | 5.96 (0.11) | 6.95 (0.12) | 9.32 (0.17) | 14.24 | 36.05 |

**Table 2** Female speaker verification results of GMM-UBM method in terms of percent EER (minDCF) for the proposed algorithm, Drugman's VAD method [27], and Rangachari's noise tracking method [21]. The last columns show the relative percent EER reduction rates compared to Drugman's VAD and Rangachari's method, respectively

| Noise type | SNR level (dB) | Proposed algorithm | Drugman's VAD | Rangachari's noise tracking | EER reduction compared to Drugman's | EER reduction compared to Rangachari's |
|---|---|---|---|---|---|---|
| Lynx | − 10 | 36.91 (0.69) | 43.80 (0.82) | 47.71 (0.88) | 15.73 | 22.63 |
| | − 5 | 27.55 (0.52) | 34.51 (0.64) | 41.40 (0.78) | 20.16 | 33.45 |
| | 0 | 16.67 (0.31) | 18.63 (0.34) | 28.86 (0.53) | 10.52 | 42.23 |
| | 5 | 9.86 (0.18) | 10.80 (0.20) | 17.76 (0.32) | 8.70 | 44.48 |
| | 10 | 6.60 (0.12) | 7.61 (0.14) | 11.16 (0.20) | 13.27 | 40.86 |
| F16 | − 10 | 42.13 (0.79) | 47.50 (0.88) | 48.73 (0.89) | 11.30 | 13.54 |
| | − 5 | 33.57 (0.63) | 42.78 (0.79) | 46.62 (0.86) | 21.52 | 27.99 |
| | 0 | 23.71 (0.45) | 29.51 (0.55) | 37.63 (0.69) | 19.65 | 36.99 |
| | 5 | 13.63 (0.25) | 15.15 (0.28) | 23.93 (0.44) | 10.03 | 43.04 |
| | 10 | 8.41 (0.15) | 8.77 (0.16) | 13.92 (0.26) | 4.10 | 39.58 |
| Car | − 10 | 6.89 (0.12) | 5.94 (0.11) | 8.70 (0.16) | − 15.99 | 20.80 |
| | − 5 | 5.14 (0.09) | 5.57 (0.10) | 8.33 (0.15) | 7.72 | 38.29 |
| | 0 | 4.78 (0.08) | 5.51 (0.10) | 8.12 (0.15) | 13.24 | 41.13 |
| | 5 | 4.49 (0.08) | 5.58 (0.10) | 8.04 (0.15) | 19.53 | 44.15 |
| | 10 | 4.56 (0.08) | 5.58 (0.10) | 8.12 (0.14) | 18.28 | 43.82 |
| Babble | − 10 | 37.05 (0.69) | 46.33 (0.86) | 48.22 (0.89) | 20.03 | 23.16 |
| | − 5 | 27.70 (0.52) | 38.50 (0.70) | 44.01 (0.81) | 28.05 | 37.06 |
| | 0 | 18.05 (0.33) | 22.84 (0.42) | 33.43 (0.62) | 20.97 | 46 |
| | 5 | 11.38 (0.21) | 12.25 (0.23) | 20.66 (0.38) | 7.10 | 44.91 |
| | 10 | 7.25 (0.13) | 8.19 (0.15) | 12.18 (0.23) | 11.47 | 40.47 |
| Stitel | − 10 | 41.55 (0.78) | 45.68 (0.86) | 46.84 (0.87) | 9.04 | 11.29 |
| | − 5 | 30.96 (0.58) | 36.26 (0.68) | 40.24 (0.76) | 14.61 | 23.06 |
| | 0 | 19.50 (0.36) | 21.32 (0.40) | 27 (0.5) | 8.53 | 27.77 |
| | 5 | 10.95 (0.20) | 11.89 (0.22) | 16.24 (0.30) | 7.90 | 32.57 |
| | 10 | 6.67 (0.12) | 8.48 (0.16) | 10.51 (0.20) | 21.34 | 36.53 |

extracted from each utterance. Linear discriminant analysis (LDA) was used to reduce the channel mismatch effects, and probabilistic LDA was employed for scoring the i-vectors. MSR Identity Toolbox [42] was used in all of the classification and scoring phases.

Two methods were selected to compare the performance of the proposed VAD algorithm. One of these methods was the noise tracking algorithm proposed in [21] (called Rangachari's method from here on). This algorithm works on the frequency bins of the conventional

**Table 3** SNR-based relative percent EER reduction rates for GMM-UBM method compared to Drugman's VAD (second and fourth column) and Rangachari's method (third and fifth column). Computed by averaging over all noise types

| SNR level (dB) | Relative EER reduction compared to Dugman's (male) | Relative EER reduction compared to Rangachari's (male) | Relative EER reduction compared to Drugman's (female) | Relative EER reduction compared to Rangachari's (female) |
|---|---|---|---|---|
| − 10 | 15.62 | 20.78 | 8.02 | 18.28 |
| − 5 | 20.84 | 32.05 | 18.41 | 31.97 |
| 0 | 7.44 | 31.33 | 14.58 | 38.82 |
| 5 | 5.02 | 34.31 | 10.65 | 41.83 |
| 10 | 14.02 | 42.65 | 13.69 | 40.25 |
| Average | 12.59 | 32.22 | 13.07 | 34.23 |

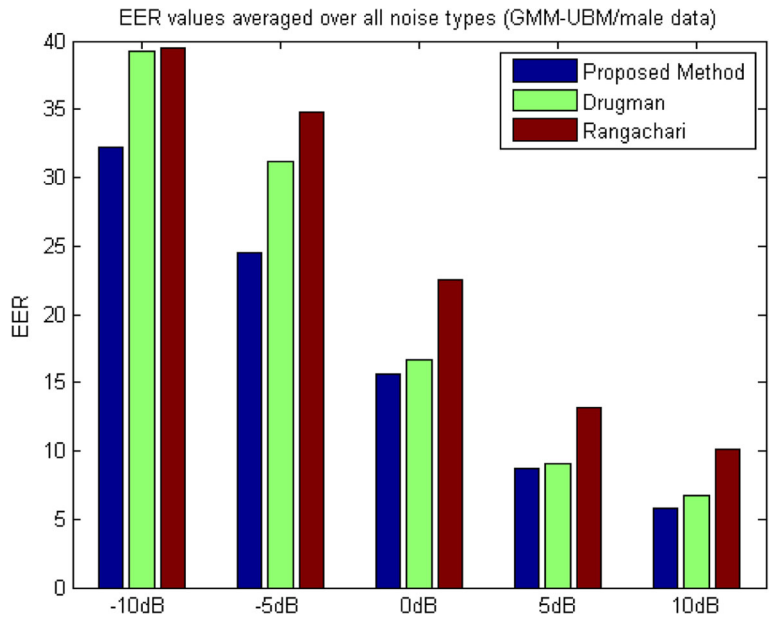


**Fig. 7** EER values averaged over all noise types for the male data with GMM/UBM backend

spectrum, but, to make a fair comparison, it was modified to work on the mel spectrum, similar to the algorithm proposed in this paper. Also, the noise tracking algorithm does not explicitly indicate the speech-active regions (although it gives a speech presence probability, it is not suitable to use it directly as a VAD output). Therefore, after the noise spectrum was estimated, and the speech was enhanced with the spectral subtraction method, the frames that had an energy higher than the average energy of all frames were accepted as speech regions. All of the other parameters were the same as given in [21].

The other method was another recently proposed VAD [27] (called Drugman's method from here on). Four voicing measures of [26], MFCCs, and two pitch trackers were used as the features, and a neural network with a single hidden layer of 32 neurons was used to obtain posterior speech probabilities of frames. The reason for choosing this method for comparison was that it had achieved superior results against the four state-of-the-art

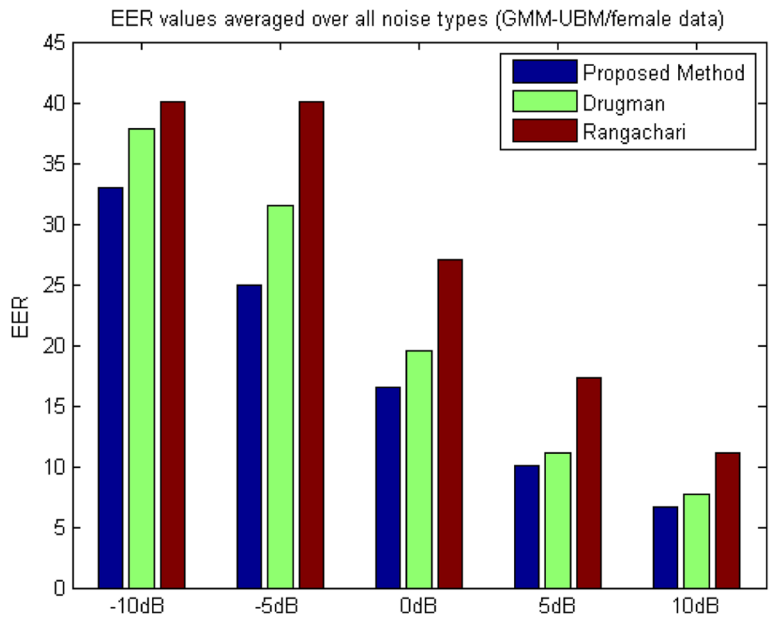


**Fig. 8** EER values averaged over all noise types for the female data with GMM/UBM backend

**Table 4** Results of the proposed algorithm with and without the polynomial regression using GMM-UBM method. Female data degraded with the lynx noise are used

| SNR level (dB) | With polynomial regression | Without polynomial regression |
|---|---|---|
| − 10 | 36.91 (0.69) | 43.36 (0.81) |
| − 5 | 27.55 (0.52) | 34.08 (0.64) |
| 0 | 16.67 (0.31) | 20.30 (0.37) |
| 5 | 9.86 (0.18) | 11.38 (0.21) |
| 10 | 6.60 (0.12) | 7.61 (0.14) |

VAD methods, as reported in [27]. Also, its memory and computational demands are low, and it includes smoothing processes before and after the neural network phase to remove spurious values and isolated misclassifications, respectively. The method is implemented by using the code obtained from the author's website [43]. For the spectral subtraction, frames decided as non-speech were used to estimate average noise energy.

### 3.2 Experimental results

In the verification phase, the test data was degraded with Lynx, F16, car, babble, and Stitel noises from the NOISEX-92 database. The overall SNR levels were changed from − 10 dB to 10 dB with 5 dB steps. EER was selected as the main performance metric, which is a widely accepted measurement in the speaker verification literature. Also, a detection cost function (DCF) was defined as given in Eq. 10,

$$DCF = C_{FP}P_{FP|N}P_N + C_{FN}P_{FN|T}P_T \quad (10)$$

where $P_{FP|N}$ is the false positive rate (FPR), $P_{FN|T}$ is the false negative rate (FNR), the cost of the false acceptance is $C_{FP}$ = 10, the cost of the false rejection is $C_{FN}$ = 1, the

**Table 5** Male speaker verification results of i-vector method in terms of percent EER (minDCF) for the proposed algorithm, Drugman's VAD method [27], and Rangachari's noise tracking method [21]. The last columns show the relative percent EER reduction rates compared to Drugman's VAD and Rangachari's method, respectively

| Noise type | SNR level (dB) | Proposed algorithm | Drugman's VAD | Rangachari's noise tracking | EER reduction compared to Drugman's | EER reduction compared to Rangachari's |
|---|---|---|---|---|---|---|
| Lynx | − 10 | 30.04 (0.55) | 43.57 (0.81) | 43.57 (0.81) | 31.05 | 31.05 |
| | − 5 | 17.58 (0.33) | 28.36 (0.53) | 29.51 (0.55) | 38.01 | 40.42 |
| | 0 | 9.48 (0.17) | 14.37 (0.27) | 17.43 (0.32) | 34.03 | 45.61 |
| | 5 | 5.58 (0.09) | 8.25 (0.14) | 11.39 (0.21) | 32.36 | 51 |
| | 10 | 3.90 (0.06) | 5.35 (0.09) | 7.72 (0.14) | 27.10 | 49.48 |
| F16 | − 10 | 38.60 (0.73) | 48.16 (0.89) | 48.93(0.89) | 19.85 | 21.11 |
| | − 5 | 27.44 (0.52) | 38.07 (0.71) | 37.08 (0.70) | 27.92 | 26 |
| | 0 | 15.82 (0.30) | 21.71 (0.40) | 23.39 (0.43) | 27.13 | 32.36 |
| | 5 | 8.48 (0.15) | 11.31 (0.21) | 14.98 (0.27) | 25.02 | 43.39 |
| | 10 | 5.35 (0.09) | 7.41 (0.13) | 10.16 (0.18) | 27.8 | 47.34 |
| Car | − 10 | 3.74 (0.06) | 4.66 (0.08) | 6.04 (0.11) | 19.74 | 38.08 |
| | − 5 | 3.28 (0.05) | 4.43 (0.07) | 3.66 (0.06) | 25.95 | 10.38 |
| | 0 | 2.98 (0.05) | 4.20 (0.06) | 5.58 (0.09) | 29.04 | 46.59 |
| | 5 | 3.13 (0.05) | 4.05 (0.06) | 5.65 (0.09) | 22.71 | 44.60 |
| | 10 | 3.13 (0.04) | 3.97 (0.06) | 5.58 (0.09) | 21.15 | 43.90 |
| Babble | − 10 | 31.88 (0.60) | 47.24 (0.87) | 47.09 (0.88) | 32.51 | 32.3 |
| | − 5 | 19.95 (0.37) | 32.95 (0.61) | 42.66 (0.80) | 39.45 | 53.23 |
| | 0 | 10.85 (0.19) | 18.19 (0.34) | 20.18 (0.38) | 40.35 | 46.23 |
| | 5 | 5.65 (0.10) | 9.25 (0.17) | 12.00 (0.22) | 38.92 | 52.91 |
| | 10 | 4.35 (0.07) | 6.11 (0.11) | 8.56 (0.15) | 28.80 | 49.18 |
| Stitel | − 10 | 37.53 (0.71) | 46.56 (0.87) | 45.87 (0.86) | 19.39 | 18.18 |
| | − 5 | 22.24 (0.42) | 32.11 (0.60) | 31.72 (0.60) | 30.73 | 29.88 |
| | 0 | 11.23 (0.20) | 17.35 (0.32) | 15.75 (0.29) | 35.27 | 28.7 |
| | 5 | 5.81 (0.11) | 9.17 (0.17) | 10.24 (0.19) | 36.64 | 43.26 |
| | 10 | 4.05 (0.07) | 6.34 (0.11) | 7.41 (0.13) | 36.12 | 45.34 |

**Table 6** Female speaker verification results of i-vector method in terms of percent EER (minDCF) for the proposed algorithm, Drugman's VAD method [27], and Rangachari's noise tracking method [21]. The last columns show the relative percent EER reduction rates compared to Drugman's VAD and Rangachari's method, respectively

| Noise type | SNR level (dB) | Proposed algorithm | Drugman's VAD | Rangachari's noise tracking | EER reduction compared to Drugman's | EER reduction compared to Rangachari's |
|---|---|---|---|---|---|---|
| Lynx | − 10 | 31.32 (0.58) | 41.84 (0.78) | 44.52 (0.84) | 25.14 | 29.65 |
| | − 5 | 20.16 (0.38) | 29.44 (0.55) | 36.76 (0.86) | 31.52 | 45.15 |
| | 0 | 11.82 (0.22) | 15.80 (0.30) | 23.93 (0.44) | 25.19 | 50.60 |
| | 5 | 6.81 (0.12) | 8.34 (0.15) | 14.35 (0.26) | 18.34 | 52.54 |
| | 10 | 4.13 (0.07) | 4.85 (0.08) | 9.64 (0.18) | 14.84 | 57.15 |
| F16 | − 10 | 38.79 (0.71) | 46.26 (0.85) | 47.71 (0.88) | 16.14 | 18.69 |
| | − 5 | 27.99 (0.52) | 37.63 (0.70) | 42.20 (0.78) | 25.61 | 33.67 |
| | 0 | 17.4 (0.33) | 24.43 (0.46) | 31.54 (0.59) | 28.77 | 44.83 |
| | 5 | 9.93 (0.18) | 11.89 (0.22) | 19.29 (0.36) | 16.48 | 48.52 |
| | 10 | 5.87 (0.10) | 6.16 (0.11) | 12.54 (0.23) | 4.70 | 53.19 |
| Car | − 10 | 3.62 (0.06) | 3.77 (0.06) | 6.74 (0.12) | 3.97 | 47.29 |
| | − 5 | 2.82 (0.05) | 3.19 (0.05) | 6.23 (0.11) | 11.59 | 54.73 |
| | 0 | 2.75 (0.04) | 3.12 (0.05) | 6.09 (0.11) | 11.86 | 54.84 |
| | 5 | 2.75 (0.04) | 3.04 (0.05) | 6.02 (0.11) | 9.54 | 54.32 |
| | 10 | 2.75 (0.04) | 3.04 (0.05) | 6.09 (0.11) | 9.54 | 54.82 |
| Babble | − 10 | 33.21 (0.63) | 44.81 (0.84) | 46.12 (0.84) | 25.88 | 27.99 |
| | − 5 | 21.68 (0.40) | 34.15 (0.63) | 40.32 (0.75) | 36.51 | 46.23 |
| | 0 | 12.98 (0.24) | 20.08 (0.37) | 27.19 (0.51) | 35.35 | 52.26 |
| | 5 | 6.89 (0.12) | 9.42 (0.17) | 16.75 (0.31) | 26.85 | 58.86 |
| | 10 | 4.06 (0.07) | 5.07 (0.09) | 10.73 (0.19) | 19.92 | 62.16 |
| Stitel | − 10 | 33.21 (0.63) | 46.04 (0.85) | 45.17 (0.83) | 27.86 | 26.47 |
| | − 5 | 26.83 (0.50) | 34.37 (0.64) | 35.53 (0.65) | 21.93 | 24.48 |
| | 0 | 15.08 (0.28) | 18.92 (0.35) | 22.33 (0.42) | 20.29 | 32.46 |
| | 5 | 8.12 (0.14) | 10.37 (0.19) | 13.77 (0.26) | 21.69 | 41.03 |
| | 10 | 4.20 (0.07) | 6.23 (0.11) | 8.99 (0.17) | 32.58 | 53.28 |

a priori probability of target tests is $P_T$ = 0.1, and the a priori probability of nontarget tests is $P_N$ = 0.9. The other performance metric was the minimum of the DCF. The results of the experiments with the male speakers using the GMM-UBM method are given in Table 1 for all noise types and all SNR levels. The minDCF values are given in parenthesis. Also, the last two columns of the table show the relative percent EER reduction between the proposed algorithm and the others. Similarly, the experimental results for the female speakers using the GMM-UBM method are given in Table 2. Table 3 shows the relative percent EER

**Table 7** SNR-based relative percent EER reduction rates for i-vector method compared to Drugman's VAD (second and fourth column) and Rangachari's method (third and fifth column). Computed by averaging over all noise types

| SNR level (dB) | Relative EER reduction compared to Drugman's (male) | Relative EER reduction compared to Rangachari's (male) | Relative EER reduction compared to Drugman's (female) | Relative EER reduction compared to Rangachari's (female) |
|---|---|---|---|---|
| − 10 | 24.50 | 28.14 | 19.79 | 30.01 |
| − 5 | 32.41 | 31.98 | 25.43 | 40.85 |
| 0 | 33.16 | 40.13 | 24.29 | 46.99 |
| 5 | 31.13 | 47.03 | 18.58 | 51.05 |
| 10 | 28.19 | 47.04 | 16.31 | 56.12 |
| Average | 29.87 | 38.86 | 20.88 | 45 |

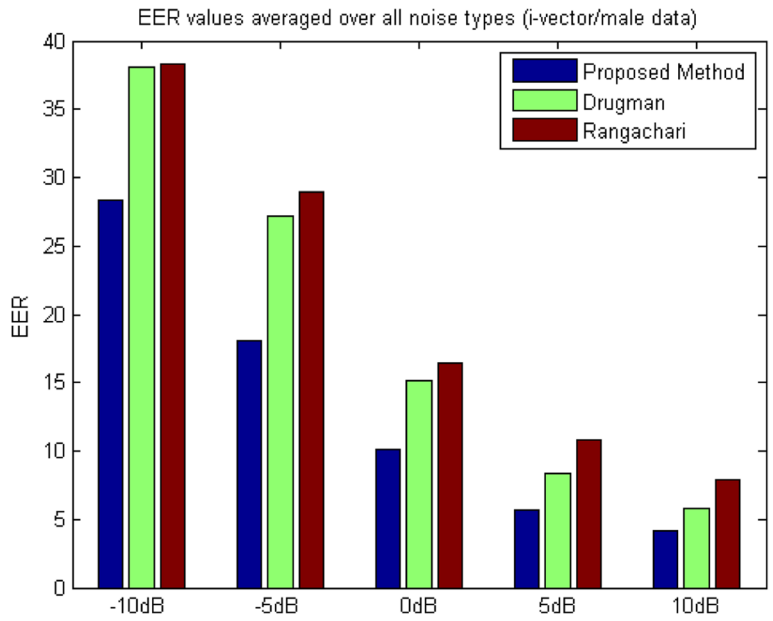


**Fig. 9** EER values averaged over all noise types for the female data with i-vector backend

reductions for each noise level (averaged over all noise types). In addition to the tables, EER values averaged over all noise types are given as bar graphs in Figs. 7 and 8 for the male speakers and the female speakers, respectively.

Before proceeding to the results for the i-vector method, it may be beneficial to verify that the polynomial regression is the main reason for the increased verification performance. For this purpose, the k-means classification is applied directly to the frames in the bands, without using the polynomial regression. All other parts of the algorithm (noise estimation, sufficient speech evidence thresholding, etc.) are the same. Table 4 shows the results of this case for the female data under the lynx noise. It is clear that directly classifying the frames as speech and noise is not as effective as the polynomial representation. The performance of the system without the polynomial regression is worse than the Drugman's VAD. Therefore, it is clear that the performance of the proposed algorithm can be attributed to

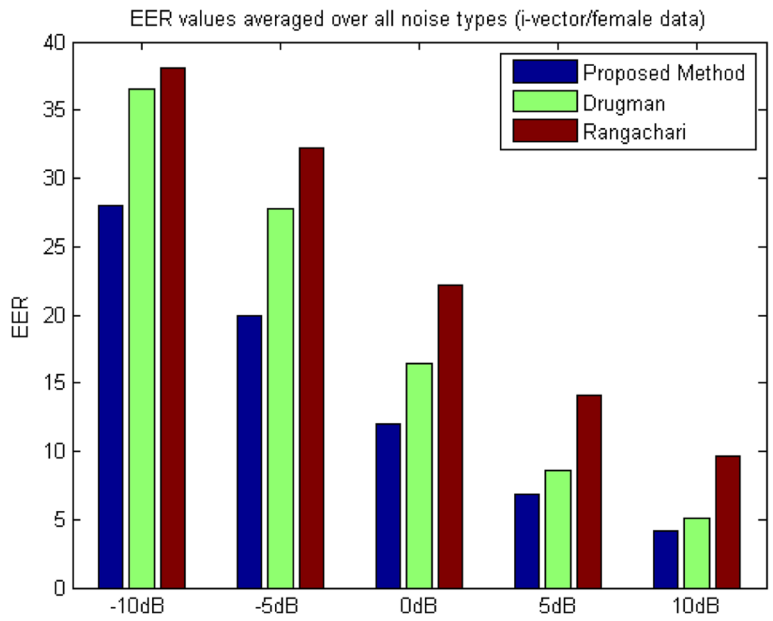


**Fig. 10** EER values averaged over all noise types for the female data with i-vector backend

both the polynomial regression and the sufficient speech evidence threshold.

The experimental results for the male and female speakers using the i-vector method are given in Tables 5 and 6, respectively. Table 7 shows the relative percent EER reductions for each noise level (averaged over all noise types). The relative EER reduction rates are also shown in the last two columns of Tables 5 and 6. EER values averaged over all noise types are given as bar graphs in Figs. 9 and 10 for the male speakers and the female speakers, respectively.

### 3.3 Discussion

As seen in the tables given above, the proposed algorithm showed better verification performances than the other methods except a few cases for the GMM-UBM method. It is also shown that all methods' performances increased by using the i-vector technique, but the proposed method benefits from the modeling capacities of the i-vectors more than the others. The superior performance of the proposed method may be due to the selection of useful speaker/speech information from the noisy speech signal. These regions were effectively extracted with the aid of the polynomial regression, and the sufficient speech evidence thresholding technique. Also, some of the low energy speech regions may be recovered by the polynomial regression if related frames also exceed the threshold. Contrary to this, frames with high energy speech only in a few bands may be discarded, due to the presence of the noise in most of the bands, which are expected to degrade the system's performance.

It is verified that the linear function chosen for determining the sufficient speech evidence is a good approximation to extract speech/speaker information under noise. Instead of a linear function between the end points, other types of functions, such as an exponential, may increase the verification performance. However, the authors prefer to analyze the performance differences of these functions in a future work to avoid an excessively long paper.

Drugman's VAD works with a conventional VAD's principle: detects speech regions, but does not produce any information about their usefulness, which probably deteriorates the verification performance. Hence, for the text-independent verification, a conventional VAD does not offer any advantages over the proposed algorithm.

Rangachari's noise tracking algorithm gave the worst overall verification performance. As discussed in Section 3.1, since the enhanced speech is treated as a clean speech, only the frames with an energy that is higher than the average energy were used for verification in this method. Instead of this kind of thresholding, a more suitable way to extract the speech-dominant regions may positively affect its results. Also, the speaker information may be damaged while the algorithm was trying to track the noise.

## 4 Conclusions

A novel algorithm is proposed in this work to extract speaker information in a robust manner. The core of the algorithm is the polynomial regression, applied in each filter band. Also, speech presence in each band is considered for a VAD-like final decision. The thresholding method, called sufficient speech evidence, increases the verification performance, especially for the low SNR levels. The algorithm also fits seamlessly to the conventional MFCC extraction scheme. The algorithm does not strictly search speech on/off points; instead, the focus is to extract most informative frames. Therefore, it is much more suitable for text-independent systems. In the experiments, the proposed algorithm was compared to a recently introduced neural network-based VAD and a speech enhancement algorithm that successfully tracks the additive noise signal. The proposed algorithm produced superior results than the others with both the conventional GMM-UBM system and the state-of-the-art i-vector system. It was verified that the frames selected by the proposed method captured more speaker information than a VAD.

**Acknowledgements**
Not applicable.

**Availability of data and materials**
Not applicable

**Authors' contributions**
Not applicable

**Funding**
Not applicable

**Authors' information**
Not applicable

**Ethics approval and consent to participate**
Not applicable

**Consent for publication**
Not applicable

**Competing interests**
The authors declare that they have no competing interests.

## Publisher's Note



**Author details**
[1]Department of Electrical-Electronics Engineering, Adana Science and Technology University, Adana, Turkey. [2]Department of Computer Engineering, Çukurova University, Adana, Turkey. [3]Department of Electrical-Electronics Engineering, Çukurova University, Adana, Turkey.

Received: 11 August 2017 Accepted: 3 October 2017
Published online: 11 October 2017

## References


1. Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Process., 10*(1–3), 19–41. https://doi.org/10.1006/dspr.1999.0361.
2. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process., 19*(4), 788–798. https://doi.org/10.1109/TASL.2010.2064307.
3. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process., 28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420.
4. Ganapathy, S., Mallidi, S. H., & Hermansky, H. (2014). Robust feature extraction using modulation filtering of autoregressive models. *IEEE/ACM Trans. Audio Speech Lang.Process., 22*(8), 1285–1295. https://doi.org/10.1109/TASLP.2014.2329190.
5. Tufekci, Z., Gowdy, J. N., Gurbuz, S., & Patterson, E. (2006). Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition. *Speech Comm., 48*(10), 1294–1307. https://doi.org/10.1016/j.specom.2006.06.006.
6. Alam, M. J., Kenny, P., & O'Shaughnessy, D. (2014). Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique. *Digital Signal Process., 29*, 147–157. https://doi.org/10.1016/j.dsp.2014.03.001.
7. Fazel, A., & Chakrabartty, S. (2012). Sparse auditory reproducing kernel (SPARK) features for noise-robust speech recognition. *IEEE Trans. Audio Speech Lang. Process., 20*(4), 1362–1371. https://doi.org/10.1109/TASL.2011.2179294.
8. Sadjadi, S. O., & Hansen, J. H. L. (2015). Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification. *Speech Comm., 72*, 138–148. https://doi.org/10.1016/j.specom.2015.04.005.
9. Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M., & Li, H. (2012). Low-variance multitaper MFCC features: a case study in robust speaker verification. *IEEE Trans. Audio Speech Lang. Process., 20*(7), 1990–2001. https://doi.org/10.1109/TASL.2012.2191960.
10. Hanilci, C., Kinnunen, T., Ertas, F., Saeidi, R., Pohjalainen, J., & Alku, P. (2012). Regularized all-pole models for speaker verification under noisy environments. *IEEE Signal Process. Lett., 19*(3), 163–166. https://doi.org/10.1109/LSP.2012.2184284.
11. Montalvão, J., & Rodrigues Araujo, M. R. (2012). Is masking a relevant aspect lacking in MFCC? A speaker verification perspective. *Pattern Recogn. Lett., 33*(16), 2156–2165. https://doi.org/10.1016/j.patrec.2012.07.023.
12. Ajmera, P. K., & Holambe, R. S. (2013). Fractional Fourier transform based features for speaker recognition using support vector machine. *Comput. Electr. Eng., 39*(2), 550–557. https://doi.org/10.1016/j.compeleceng.2012.05.011.
13. Sahidullah, M., & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Comm., 54*(4), 543–565. https://doi.org/10.1016/j.specom.2011.11.004.
14. Dişken, G., Tüfekçi, Z., Saribulut, L., & Çevik, U. (2016). A review on feature extraction for speaker recognition under degraded conditions. *IETE Tech. Rev.*, 1–12. https://doi.org/10.1080/02564602.2016.1185976.
15. Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process., 27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209.
16. Abd El-Fattah, M. A., Dessouky, M. I., Abbas, A. M., Diab, S. M., El-Rabaie, E.-S. M., Al-Nuaimy, W., et al. (2014). Speech enhancement with an adaptive wiener filter. *Int. J. Speech Technol., 17*(1), 53–64. https://doi.org/10.1007/s10772-013-9205-5.
17. Martin, R. (1994). Spectral subtraction based on minimum statistics. In *European Signal Processing Conference* (pp. 1182–1185). Edinburgh; EURASIP.
18. Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process., 9*(5), 504–512. https://doi.org/10.1109/89.928915.
19. Li, X., Girin, L., Gannot, S., & Horaud, R. (2016). Non-stationary noise power spectral density estimation based on regional statistics. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings* (pp. 181–185). Shanghai;IEEE.
20. Lin, Z., Goubran, R. A., & Dansereau, R. M. (2007). Noise estimation using speech/non-speech frame decision and subband spectral tracking. *Speech Comm., 49*(7–8), 542–557. https://doi.org/10.1016/j.specom.2006.10.002.
21. Rangachari, S., & Loizou, P. C. (2006). A noise-estimation algorithm for highly non-stationary environments. *Speech Comm., 48*(2), 220–231. https://doi.org/10.1016/j.specom.2005.08.005.
22. Ramírez, J., Segura, J. C., Benítez, C., de la Torre, Á., & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Comm., 42*(3–4), 271–287. https://doi.org/10.1016/j.specom.2003.10.002.
23. Ghosh, P. K., Tsiartas, A., & Narayanan, S. (2011). Robust voice activity detection using long-term signal variability. *IEEE Trans. Audio Speech Lang. Process., 19*(3), 600–613. https://doi.org/10.1109/TASL.2010.2052803.
24. Wu, J., & Zhang, X.-L. (2011). An efficient voice activity detection algorithm by combining statistical model and energy detection. *EURASIP J. Adv. Signal Process., 2011*(1), 18. https://doi.org/10.1186/1687-6180-2011-18.
25. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., et al. (2004). A tutorial on text-independent speaker verification. *EURASIP J. Adv. Signal Process., 2004*(4), 430–451. https://doi.org/10.1155/S1110865704310024.
26. Sadjadi, S. O., & Hansen, J. H. L. (2013). Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Process. Lett., 20*(3), 197–200. https://doi.org/10.1109/LSP.2013.2237903.
27. Drugman, T., Stylianou, Y., Kida, Y., & Akamine, M. (2016). Voice activity detection: merging source and filter-based information. *IEEE Signal Process. Lett., 23*(2), 252–256. https://doi.org/10.1109/LSP.2015.2495219.
28. M. Sahidullah, G. Saha, Comparison of speech activity detection techniques for speaker recognition (2012), Retrieved from http://arxiv.org/abs/1210.0297. Accessed 1 Aug 2017.
29. Prasanna, S. R. M., & Pradhan, G. (2011). Significance of vowel-like regions for speaker verification under degraded conditions. *IEEE Trans. Audio Speech Lang. Process., 19*(8), 2552–2565. https://doi.org/10.1109/TASL.2011.2155061.
30. 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. https://doi.org/10.1109/TASL.2013.2238529.
31. Ribas González, D., & Calvo de Lara, J. R. (2014). Feature classification criterion for missing features mask estimation in robust speaker recognition. *SIViP, 8*(2), 365–375. https://doi.org/10.1007/s11760-012-0299-z.
32. May, T., van de Par, S., & Kohlrausch, A. (2012). Noise-robust speaker recognition combining missing data techniques and universal background modeling. *IEEE Trans. Audio Speech Lang. Process., 20*(1), 108–121. https://doi.org/10.1109/TASL.2011.2158309.
33. Zhao, X., Shao, Y., & Wang, D. (2012). CASA-based robust speaker identification. *IEEE Trans. Audio Speech Lang. Process., 20*(5), 1608–1616. https://doi.org/10.1109/TASL.2012.2186803.
34. Yan, F., Zhang, Y., & Yan, J. (2014). A sub-band-based feature reconstruction approach for robust speaker recognition. *EURASIP J. Audio Speech Music Process., 2014*(1), 1–13. https://doi.org/10.1186/s13636-014-0040-7.
35. de la Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Benitez, M. C., & Rubio, A. J. (2005). Histogram equalization of speech representation for robust speech recognition. *IEEE Trans. Speech Audio Process., 13*(3), 355–366. https://doi.org/10.1109/TSA.2005.845805.
36. Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process., 22*(4), 745–777. https://doi.org/10.1109/TASLP.2014.2304637.
37. Li, Q., Zheng, J., Tsai, A., & Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Trans. Speech Audio Process., 10*(3), 146–157. https://doi.org/10.1109/TSA.2002.1001979.
38. Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Comm., 49*(7–8), 588–601. https://doi.org/10.1016/j.specom.2006.12.006.
39. Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Comm., 12*(3), 247–251. https://doi.org/10.1016/0167-6393(93)90095-3.
40. Doddington, G. R., Przybocki, M. A., Martin, A. F., & Reynolds, D. A. (2000). The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective. *Speech Comm., 31*(2–3), 225–254. https://doi.org/10.1016/S0167-6393(99)00080-1.

41. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Comm., 52*(1), 12–40. https://doi.org/10.1016/j.specom.2009.08.009.
42. Sadjadi, S. O., Slaney, M., & Heck, L. (2013). MSR identity toolbox v1. 0: A MATLAB toolbox for speaker recognition research. Proc. IEEE Signal Process. Soc. Speech Lang. Tech. Committee Newsl.
43. T. Drugman. http://tcts.fpms.ac.be/~drugman/Toolbox/. Accessed 8 Aug 2017.