Robust and High-resolution Voiced/Unvoiced Classification in Noisy Speech Using A Signal Smoothness Criterion

A. Sreenivasa Murthy, S. Chandra Sekhar and T. V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore, India
{asmuvce, tvsree}@ece.iisc.ernet.in, chandrasekhar.seelamantula@epfl.ch

Abstract

We propose a novel technique for robust voiced/unvoiced segment detection in noisy speech, based on local polynomial regression. The local polynomial model is well-suited for voiced segments in speech. The unvoiced segments are noise-like and do not exhibit any smooth structure. This property of smoothness is used for devising a new metric called the variance ratio metric, which, after thresholding, indicates the voiced/unvoiced boundaries with 75% accuracy for 0dB global signal-to-noise ratio (SNR). A novelty of our algorithm is that it processes the signal continuously, sample-by-sample rather than frame-by-frame. Simulation results on TIMIT speech database (downsampled to 8kHz) for various SNRs are presented to illustrate the performance of the new algorithm. Results indicate that the algorithm is robust even in high noise levels.

Index terms: voiced, unvoiced, local polynomial model, regression, signal-to-noise ratio.

1. Introduction

In many real-world technological applications involving speech-activated automatic systems, speech recognition engines, automatic transcription systems, efficient low bitrate speech coding for mobile telephone applications etc., classification of speech into voiced and unvoiced regions is necessary. Voiced and unvoiced segments possess different spectral and temporal properties. They carry intelligibility and linguistic information by different amounts and hence require different processing techniques. Since many of these applications run in a real-world environment where ambient noise is always present, there is a huge demand for noise-robust techniques.

Numerous techniques have been developed for classification of speech signals into voice/unvoiced segments. These are techniques based on likelihood-ratio test [1, 2], multiple features like cepstral peak, zero crossing rate and energy [3], instantaneous frequency amplitude spectrum (IFAS) [4], dominance spectrum [5], long-term spectral divergence measure [6] etc. In [7], the speech/ pause discrimination is posed as an unsupervised learning problem. Each frame is divided into a number of subbands and log energy is computed for every subband. An observation vector for each frame is obtained using these subband log-energies as features. Hard C-means clustering algorithm is applied to a set of initial pause frames’ observation vectors to obtain the noise-subspace model. The noise model is updated during speech pause frames detected later—thus it also works in nonstationary environments. The energy envelope of the frame to be classified is computed as the maximum value of the observation vectors of $2m + 1$ frames centered around the frame to be discriminated. The Euclidean distance between the energy envelope and the average noise cluster center is compared with a threshold to decide voicedness of the frame. The detection threshold depends not only on the data to be processed, but also on the noisiest and cleanest conditions under which the technique has to work.

One of the practical difficulties of the above method is that the decision requires an $m$ frame delay and more computational effort to make a voiced-unvoiced decision. Also the detection threshold requires a priori knowledge about the environment in which the algorithm works.

In this paper, we formulate the problem of voiced/unvoiced segment detection in noisy speech in the framework of local-polynomial regression [8] which exploits the smoothness property of the voiced segments. The local-polynomial approximation model has been successfully used in many recent research works [9, 10]. The difference signal between the modeled and the measured signals is used to compute a variance-ratio metric which is defined as the ratio of the variance of the difference signal and the variance of the noisy signal. Statistical analysis of this metric is performed to obtain a threshold for classification.

2. Problem Formulation

Let $s[n]$ denote the samples of a speech signal and $w[n]$ that of ambient noise. The noisy speech samples are denoted by $x[n]$ and they are given by

$$x[n] = s[n] + w[n], 0 \leq n \leq N - 1.$$  \hspace{1cm} (1)

where we assumed an additive noise model. The noise is assumed to be white Gaussian with zero mean and variance $\sigma_w^2$. $N$ is the length of the speech signal and not the frame size.

$s[n]$ consists of a sequence of voiced and unvoiced phonemes. The objective is to find the boundaries between voiced and unvoiced segments in $\hat{s}[n], 0 \leq n \leq N - 1$ given $x \triangleq \{x[n], 0 \leq n \leq N - 1\}$. We would like to state here that our problem formulation is different from the usual formulation in the sense that we do not perform a frame-by-frame analysis. We process the signal continuously and obtain voiced/unvoiced decision boundaries without regard to any frame boundaries. The signal $s[n]$ is a concatenation of several voiced and unvoiced segments. In addition, each voiced/unvoiced segment has a different duration. Our objective is to obtain high-resolution noise-robust decision boundaries.

To enable V/UV (voiced/unvoiced) decision-making, we need a test statistic $T(x)$ that quantifies the voiced/unvoicedness
of $x$. The decision rule then takes the form:
$$ T(x) > \gamma, \text{ for voiced signals and,} $$
$$ T(x) < \gamma, \text{ for unvoiced signals,} $$
(2)
where $\gamma$ is a signal-dependent threshold.

### 3. Local polynomial model

Let us compare the waveforms of voiced and unvoiced segments. Few randomly-picked samples of voiced and unvoiced segments are shown in Fig. 1(a) and (b) respectively. We observe that the voiced segment waveforms have a smooth structure whereas the unvoiced segments are mostly noisy. We propose to model the smooth structure in voiced segments by polynomials. Any other smooth function model can also be used but we use polynomials owing to their simplicity. Since the signal is nonstationary and it has time-varying properties, one must perform local polynomial regression [8]. The polynomial model fails to be effective for unvoiced segments - it is precisely this feature that we use to distinguish voiced and unvoiced segments. Local polynomial modeling is more efficient compared to global polynomial modeling because it is computationally faster and stabler to use low-order polynomial models over short durations than use large-order polynomials over a long duration. The local polynomial model approach is similar in spirit to the short-time Fourier transform (STFT) except that the choice of the basis functions is different i.e., we use polynomials instead of sinusoids. The signal can be modeled over a short duration by using low-order polynomials. Unlike the STFT, in a local polynomial model (LPM), the number of coefficients (=model order+1) is not necessarily equal to the number of samples. For example, in STFT analysis using a frame-size of 160 samples, the number of STFT coefficients is necessarily 160. In the LPM technique, if we use a frame size of 30 samples, the number of polynomial coefficients can be as small as four (for a third-order polynomial) with a small error. Thus, the LPM provides a more compact representation. The LPM approach allows the model coefficients to change with time which makes it suitable for nonstationary signal processing. With this motivation, we write the following model for the signal:

$$ s[n] = \sum_{k=0}^{p} a_k[n] n^k + e[n], \quad 0 \leq n \leq N - 1, $$
(3)
where $e[n]$ denotes the modeling error. $e[n]$ can also be used to absorb the noise in voiced consonants since it cannot be modeled efficiently by polynomials. The coefficients \{a_0[n], a_1[n], \ldots, a_p[n]\} are functions of time so as to adapt to the time-varying properties of $s[n]$. We can rewrite (1) using (3) as:

$$ x[n] = \sum_{k=0}^{p} a_k[n] n^k + e[n] + w[n], \quad 0 \leq n \leq N - 1. $$
(4)

Given \{x[n], 0 \leq n \leq N - 1\}, we perform polynomial regression over \( n - \frac{L}{2} \leq n \leq \frac{L}{2} \) to estimate the coefficients which are denoted by \{\hat{a}_k[n], k = 0, 1, 2, \ldots, p\}. These coefficients are then used to estimate $s[n]$ as:

$$ \hat{s}[n] = \sum_{k=0}^{p} \hat{a}_k[n] n^k, \quad 0 \leq n \leq N - 1. $$
(5)

The LPM can be performed at every instant $n \in [0, N - 1]$ or on a frame-by-frame basis with an overlap. In this paper, we perform LPM at every instant since we are interested in computing high-resolution VUV boundaries. The estimate $\hat{s}[n]$ is a close approximation to $s[n]$ for voiced segments. In unvoiced segments, the estimate $\hat{s}[n]$ is close to zero since the signal is noise-like and lacks a smooth structure. Therefore, the voiced/unvoiced hypothesis testing problem can be rewritten as:

- $\mathcal{H}_0$: Unvoiced segment
  \[ \hat{s}[n] \approx 0; d[n] \triangleq x[n] - \hat{s}[n] \approx x[n], 0 \leq n \leq N - 1. \]  
  (6)
- $\mathcal{H}_1$: Voiced segment
  \[ d[n] \triangleq x[n] - \hat{s}[n] \approx w[n] + e[n], 0 \leq n \leq N - 1. \]  
  (7)
where $d[n]$ is the difference between the noisy signal and $\hat{s}[n]$. The waveforms of the noisy signal $x[n]$ (at 2.4 dB SNR) in voiced and unvoiced segments are shown in Fig. 1(c) and (d). The signals computed by local polynomial regression are shown in Fig. 1(e) and (f). In our experiments, $L = 6$ and $p = 3$.

### 4. Test statistic

Let $\sigma^2_x, \sigma^2_e$ and $\sigma^2_w$ denote the variances of $x[n], e[n]$ and $d[n]$ respectively. We have:

- $\mathcal{H}_0$: Unvoiced segment
  \[ d[n] = x[n] - \hat{s}[n] \approx x[n] \Rightarrow \sigma^2_d \approx \sigma^2_x. \]  
  (8)
- $\mathcal{H}_1$: Voiced segment
  \[ d[n] \approx w[n] + e[n], \Rightarrow \sigma^2_d = \sigma^2_w + \sigma^2_e \ll \sigma^2_x. \]  
  (9)

Now, consider the variance ratio:

$$ \xi = \frac{\sigma^2_d}{\sigma^2_x} $$
(10)
We propose the use of $\xi$ as a test statistic for distinguishing voiced segments from the unvoiced segments. Given $x[n]$, we compute smooth running-sample variances to estimate $\sigma_d^2$, $\sigma_s^2$, and $\xi$ as:

$$
\xi[n] = \frac{\sum_{m=n-L+1}^{n+L} h[m](d[m] - \mu_d)^2}{\sum_{m=n-L+1}^{n+L} h[m](x[m] - \mu_s)^2},
$$

where $\mu_d$ and $\mu_s$ are the sample means of $d[m]$ and $x[m]$ respectively. $h[n]$ is a window function used to smooth the variance estimates—we chose a $(L+1)$ point Hamming window.

5. Threshold computation

The test statistic $\xi[n]$, is plotted in Figs. 2(c) and 3(c) for a noisy speech segment. Clearly, $\xi[n]$ shows a discriminating behaviour depending on whether the segment is voiced or unvoiced. Next, we need to obtain a threshold for V/U detection, for which, we compute the distribution of $\xi[n]$ by a histogram approximation. The histogram is shown in Fig. 4, normalized to possess unity sum. The histogram shows two distinct modes, ignoring the minor variations. The mode at the lower end corresponds to the voiced segments whereas the mode corresponding to the unvoiced segment is at the upper end. We have devised the following experimental procedure for computing the threshold for a given sentence.

Let $\mu$ be the position of the highest peak in the histogram $f(\xi)$ and let $\sigma$ be the standard deviation of $\xi$ about $\mu$ (computed using the histogram). The highest-valued peak may occur at the lower-end or at the higher-end of the histogram depending on the signal-to-noise ratio. Taking this fact into account, we have developed an experimental choice of the threshold $\gamma$ as follows:

$$
\gamma = \begin{cases} 
\min (\mu + 3\sigma, 0.5) & \text{if } \mu < 0.5, \\
\max (\mu - 3\sigma, 0.5) & \text{if } \mu > 0.5.
\end{cases}
$$

The hypothesis-testing problem now has the following solution:

If $\xi[n] > \gamma$, then decide $\mathcal{H}_0$, else decide $\mathcal{H}_1$.

Note that $\xi[n]$ is an instantaneous test statistic and hence we might get spurious classification in some regions, resulting in very short duration voiced/unvoiced segments. Spurious classification is corrected for, by imposing a speech signal-specific duration constraint. Any segment of duration less than 20ms is merged with the preceding segment. The duration constraint is based on the speech production mechanism according to which speech signal changes relatively slowly with time over durations of the order of a few milliseconds. The voiced/unvoiced classification boundaries of speech at average segmental SNR of 2.4dB and 9dB are shown in Figs. 2(d) and 3(d) respectively, for a single realization of the noise sequence.

To obtain the statistical performance of the algorithm, we simulate noisy speech at 0, 5, 10, 15 and 20 dB global SNR and generate 100 Monte-Carlo trials for each SNR. The percentage accuracy of classification compared with manual segmentation is studied in terms of percentage matches, miss, and insertions (false alarm). In the process of matching the boundaries, we allow for a tolerance of the order of a few milliseconds. We report results for three different tolerance durations: 20ms, 40ms and 50ms. The tolerance limits are used only for the purpose of comparing the results with manual segmentation. Otherwise, we can still retain the high-resolution voiced/unvoiced decision and need not quantize the decision boundaries in terms of these tolerances. The results (in percentages) corresponding to each value of the tolerance and SNR are shown in Tables 1, 2, and 3. We deduce that the proposed algorithm is quite robust to noise and its performance degradation with respect to a decrease in SNR is at an acceptable level.

6. Conclusions

We developed a new technique for obtaining the boundaries between voiced and unvoiced segments in the presence of noise. The technique employs a local polynomial model which adapts to the nonstationary nature of the signal. Localized processing also enables a fine-level voiced/unvoiced classification which is useful in segment- vocoder applications. The robustness of the
Figure 3: (a) Clean speech signal, (b) noisy speech signal (average segmental SNR = 9 dB), (c) test statistic $\xi$ and (d) voiced-unvoiced decision boundaries superimposed on the speech signal (1: unvoiced; 0: voiced). $L = 6$ and $p = 3$.

Table 2: Performance of the voiced/unvoiced classification algorithm for 40ms tolerance.

<table>
<thead>
<tr>
<th>SNR dB</th>
<th>%Matches</th>
<th>%Insertions</th>
<th>%Deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>75.08</td>
<td>0.00</td>
<td>35.66</td>
</tr>
<tr>
<td>5</td>
<td>87.34</td>
<td>0.30</td>
<td>12.36</td>
</tr>
<tr>
<td>10</td>
<td>81.30</td>
<td>0.00</td>
<td>18.60</td>
</tr>
<tr>
<td>15</td>
<td>89.22</td>
<td>0.20</td>
<td>10.58</td>
</tr>
<tr>
<td>20</td>
<td>97.80</td>
<td>0.10</td>
<td>02.10</td>
</tr>
</tbody>
</table>

Table 3: Performance of the voiced/unvoiced classification algorithm for 50ms tolerance.

<table>
<thead>
<tr>
<th>SNR dB</th>
<th>%Matches</th>
<th>%Insertions</th>
<th>%Deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>84.27</td>
<td>0.00</td>
<td>26.47</td>
</tr>
<tr>
<td>5</td>
<td>92.02</td>
<td>0.30</td>
<td>07.38</td>
</tr>
<tr>
<td>10</td>
<td>89.80</td>
<td>0.00</td>
<td>10.20</td>
</tr>
<tr>
<td>15</td>
<td>93.01</td>
<td>0.20</td>
<td>06.39</td>
</tr>
<tr>
<td>20</td>
<td>98.30</td>
<td>0.10</td>
<td>01.60</td>
</tr>
</tbody>
</table>

technique has been demonstrated by a Monte-Carlo performance analysis in the presence of white Gaussian noise, for various values of SNR. The performance degradation with respect to decrease in SNR is at an acceptable level. The performance for other types of noise will be reported separately. There is scope for improvement in the algorithm - we computed the thresholds experimentally using the noisy signal. Alternatively, one can devise efficient techniques for computing the threshold based on the optimization of a suitable criterion.

7. References


