

An Algorithm for V/UV/S Segmentation of Speech

by

Meng Wang, Donald Bitzer,

David McAllister, Robert Rodman, Jamie Taylor

Voice I/O Lab, N.C. State University, Raleigh, NC 27606

TEL=9195157480 FAX=9195157896 mwang3@unity.ncsu.edu Categories 2 & 14

Let $f(n)$ be a sampled voice signal. Our goal is to identify the voiced (V) portions of f (as opposed to the silence (S) and unvoiced (UV) portions). In the following discussion, the sampling rate is 22050 Hz, quantized at 8 bits. A window of length of 880 is twice the maximum period of the minimum frequency of 50 Hz we will track in the time domain. We use the algorithm described below to provide a robust estimate of the fundamental period to start our glottal pulse (GP) or pitch tracker described in [1].

We compute $R(f)$, the discrete Fourier transform (DFT) for samples in the window. Let $T(f) = |R(f)|$ having eliminated the last half of R because of symmetry. If the window contains more than one occurrence of the fundamental period of a voiced utterance, there will be “aliasing” in T , where we define “aliasing” to be any inaccuracies in the frequency analysis of a periodic signal resulting from a poor choice of window size. A plot of $T(f)$ in Figure 1 is of voiced speech.

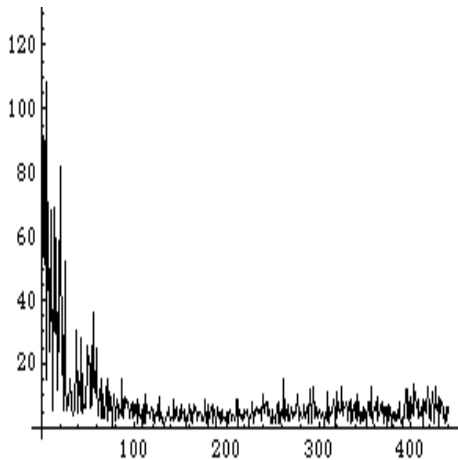


Figure 1(a) DFT with aliasing.

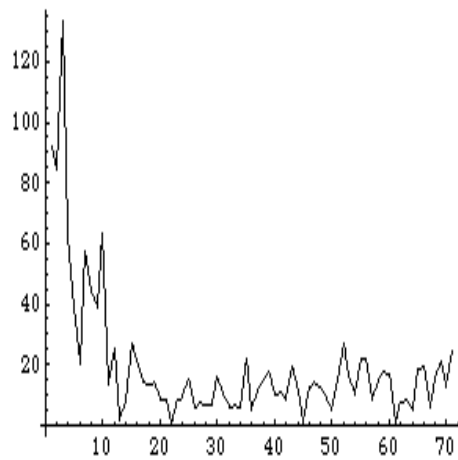


Figure 1(b) DFT w/o aliasing

Notice Figure 1(b) is the envelope of Figure 1(a). The aliasing is seen in Figure 2 by observing the large area under the DC term and under the first peak.

The real Cepstrum $c(f)$ is the logarithm of the power spectrum, $c(f) = 2 \ln |R(f)|$. From the properties of the Cepstrum, we know $c(f)$ consists of two components: a slowly varying component which corresponds to the spectral envelope and a rapidly varying component which corresponds to the pitch harmonic peaks [2].

Since the logarithm is monotonically increasing, $R(f)$ also has two components: one which corresponds to the spectral envelope and another which corresponds to the pitch harmonic peaks. These components can be separated by filtering, which is the traditional way to proceed, or by a second Fourier transform, which we have found to be more resistant to noise.

$$Y(f) = T(R(f)).$$

A graph of $Y(f)$ in the case of a voiced utterance is shown as follows :

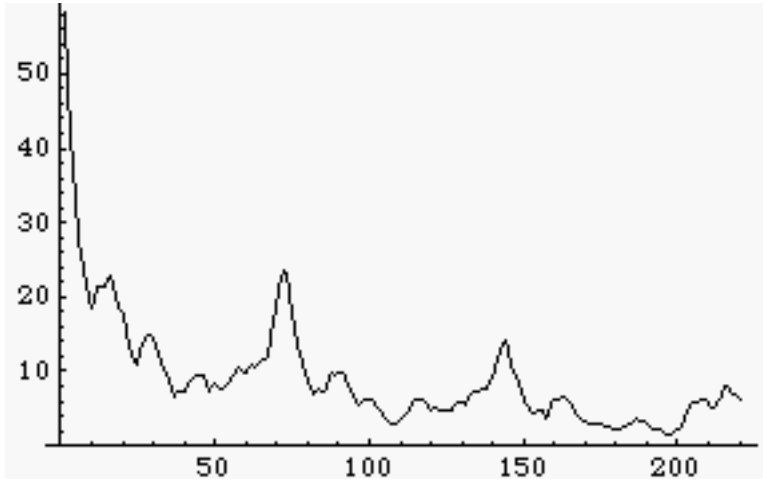


Figure 2
Plot of 2nd DFT of voiced speech, low noise

The cluster of harmonics near the origin (the DC term) is the transform of the spectral envelope. Hence, the value of this DC term represents the total energy of the spectral envelope in Figure 1(b). The narrow peak at t_0 (about 70 in this case.) is the transform of the harmonic peaks, so its maximum value represents the total energy of the harmonic peaks in Figure 1(a). Here the separation between the pitch peak and the envelope transform is always great enough so that the former can easily be distinguished.

Since unvoiced speech and silence won't produce harmonics in its DFT, the graphs of the second DFT will not have a salient peak. (See Figure 3)

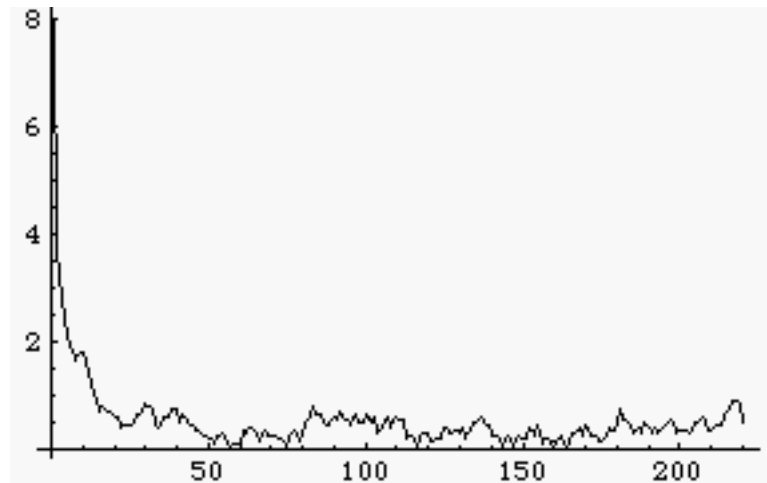


Figure 3
Plot of $Y(f)$ for unvoiced speech, low noise

Assume the amplitude of the above DC term is D , and that of the above peak at t_0 is P . We compute the ratio: P/D . If this ratio is large, then the energy generated by harmonics is an important part of the total energy, indicating the presence of a voiced utterance. On the contrary, if the energy is small relative to energy of the envelope, it is not possible to decide if the utterance is voiced or not.

The algorithm is as follows:

L is a list of the fundamental periods derived from $Y(f)$ computed from each group of 110 samples. If the samples represent UV or S, the value is zero.

1. Let the window W contain 880 samples. After computing $Y(f)$ for a window W , we shift it right 110 samples and begin again. The number of window computations becomes

$$N = \text{Round}[(\text{Length}[\text{signal}] - 880)/(880 - 770)]$$

where $\text{Length}[\text{signal}]$ is the total number of samples in the input signal.

2. For each window, compute the absolute value of the DFT of $f(t)$ and keep the first half. Call this $R(f)$. This yields 440 nonnegative real valued samples.

3. Take the absolute value of the DFT of $R(f)$, and keep the first half, giving 220 nonnegative real valued samples. Call this $Y(f)$.

4. Let P be the maximum value of $Y(f)$ between sample 51 and sample 120. If the set of samples in W is voiced speech, then the peak occurs at a half of the fundamental period. Hence, at 22 KHz, the fundamental period must lie between sample 102 and sample 240.

5. Compute the ratio of P to D, the DC term, and compare it to a threshold t . If the ratio is greater than t then the utterance segment is voiced, and we append the t_0 value at peak of DFTDFT plot to L. That t_0 value is the estimate of half of the fundamental period of the voiced segment. Otherwise, append zero to L. The value of t is determined by experiment, which indicates if using a smaller value, unvoiced will be considered as voiced, and if using a larger value, we will lose information.

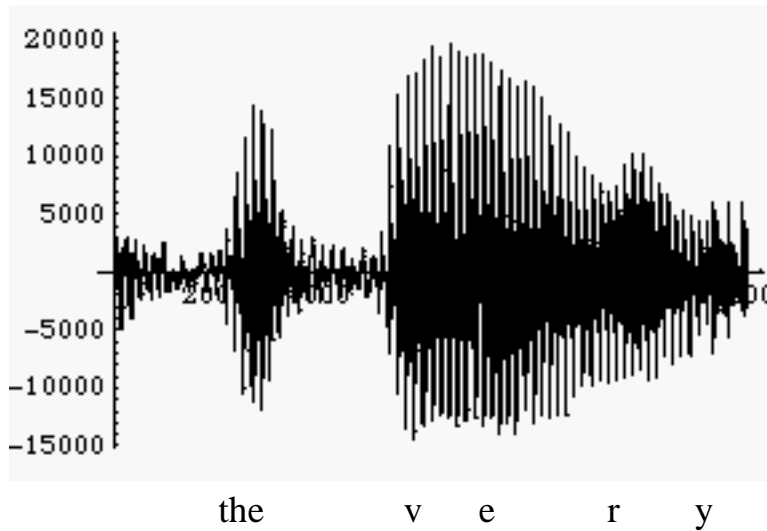


Figure 4

Signal containing voiced speech and silence, low noise
 "...the very ..."

We used Mathematica 4.0 on a G4 MAC to compute the following results. The threshold t to determine if a window is voiced is 0.25.

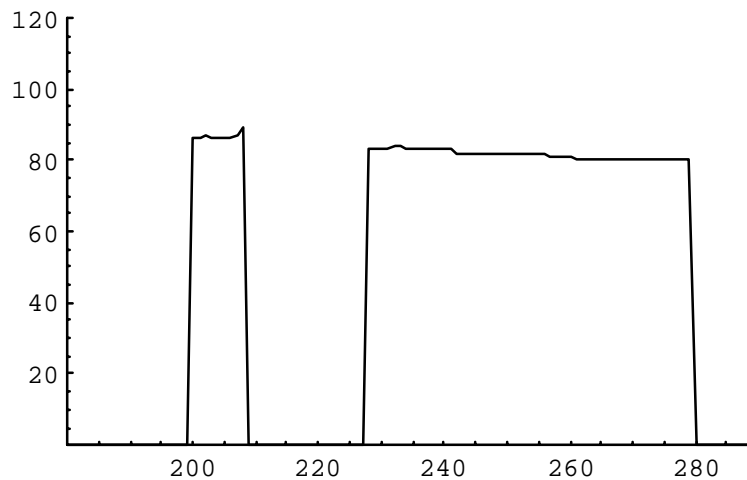


Figure 5

A plot of L for the signal in Figure 4.
Twice the vertical ordinate is the period of the fundamental

Conclusion

This algorithm only uses traditional DSP method to detect voiced utterances in speech signals. The accuracy is also very good. The method is superior to the Cepstrum method, which is applicable only when no noise is present. It provides a robust estimate of the period of the glottal pulse if the signal is voiced.

References:

[1] Rodman, R., McAllister, D., Bitzer, D., and Chappell, D. A High-Resolution Glottal Pulse Tracker. *Proceedings of the International Conference on Spoken Language Processing (ICSLP2000)*, October, 2000.

[2] Thomas Parsons, "Voice and Speech Processing", 1986, pp.203 - 204.