5-min knowledge sharing/discussion

Week 8	**12012842**	**房淑晗**
	12012001	**曹昕琰**
	12012341	**陈佳琪**
	12012103	**施永祺**
Week 9	11911616	杨振宇
	12010417	邵明磊
	12012146	杨佳怡

Lab VIII
Algorithms for Estimating Speech Parameters

DONG Yunyang

dongyy@mail.sustech.edu.cn

North Tower 316, College of Engineering

TencentMeeting: 614-588-8061

2023-04-04

Purpose of this lab...

1. Test and master the processing of median smoothing on speech parameter estimation
2. Test the commonly-used algorithms on pitch estimation (i.e., autocorrelation method, ceptrum method)

Problem 1

10.1. (MATLAB Exercise) The purpose of this MATLAB exercise is to compare linear, median, and combination smoothers on short-time zero-crossing estimates. Using the speech file `test_16k.wav`, compute the short-time zero-crossing rate (per 10 msec of speech) using a frame length of 10 msec and a frame shift of 5 msec. Plot the resulting short-time estimate of zero-crossing rate per 10 msec interval. (It should exhibit a lot of high frequency variation since the estimation interval is short.) Now design a lowpass filter to preserve the low frequency band until about $f = 0.1F_s$, or about 1.6 kHz for this 16 kHz sampled signal. Design a lowpass filter to remove the band from $f = 0.2 * F_s$ to $f = 0.5 * F_s$. (Hint: You should find that you get about 40 dB out-of-band rejection using a filter length of $L = 41$ samples and this should be adequate for this exercise.) Use this filter to smooth the short-time zero-crossings rate contour and plot the resulting smoothed curve as a sub-plot on the page with the original zero-crossings rate contour. Now median-smooth the original zero-crossings rate contour using a combination of a running median of seven samples followed by a running median of five samples. Plot the resulting curve as the third sub-plot. Finally use a combination smoother (of the type discussed in this chapter) and plot the resulting smoothed contour. What differences do you see in the four contours? Which contour best preserves the most salient characteristics of the original short-time zero-crossing rate contour?

Linear Smoother

Median Smoother

Combination Smoother

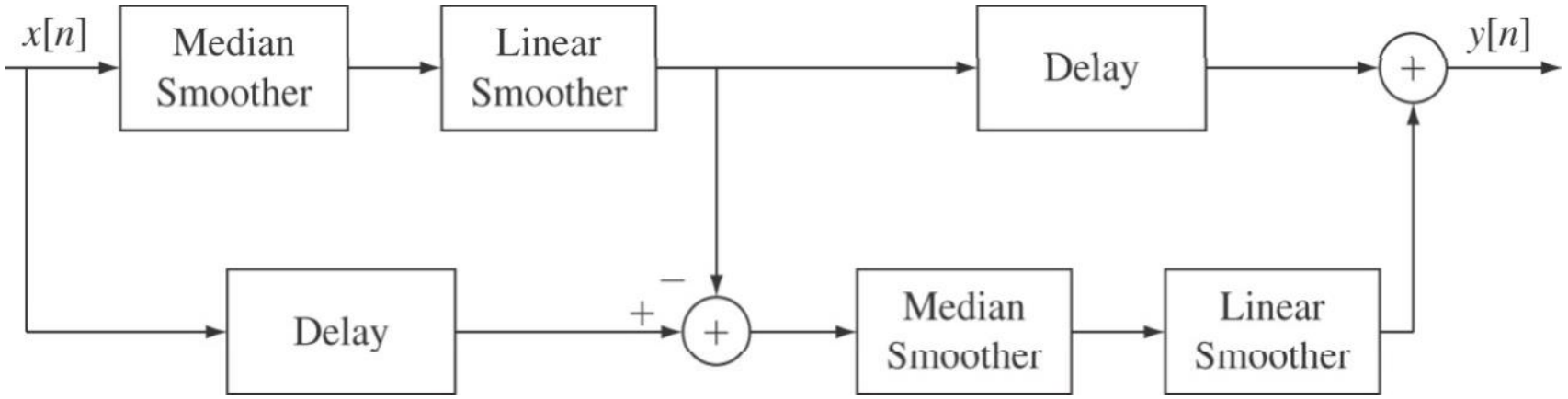
y = MedianSmoother (x, n)

Problem 1

Median Smoother (L points):

$$y[n] = M_L\{x[n]\} = \text{med}_{m=0}^{L-1}\, x[n-m],$$

Combination Smoother:



Problem 2

10.7. (MATLAB Exercise: Autocorrelation-Based Pitch Detector) Program an autocorrelation-based pitch detector using the modified autocorrelation function. Compare the results using both the original speech file and those obtained from a bandpass filtered version of the speech file.

The steps to follow for implementing the pitch detector are the following:

1. specify whether the talker is a male or female (the program uses the talker gender to set ranges for the pitch period contour);
2. read in the speech file (including determining the speech sampling rate, `fs`);
3. convert the sampling rate to a standard value of `fsout=10000` Hz for this exercise;
4. design and implement a bandpass filter to eliminate DC offset, 60 Hz hum, and high frequency (above 1000 Hz) signal, using the design parameters:
   - stopband from 0 to 80 Hz
   - transition band from 80 to 150 Hz
   - passband from 150 to 900 Hz
   - transition band from 900 to 970 Hz
   - stopband from 970 to 5000 Hz
   - filter length of `n=301` samples

PicthDetector_Autocorrelation (s, fs, gender)

Resample -> Filter -> (Window) -> Autocorrelation searching
-> Confident -> Set zero for below confident -> Median smooth

5. save both the full band speech and the bandpass filtered speech files for processing and comparison;
6. play both the original and bandpass filtered speech files to be sure that the filtering worked properly;
7. block the signal into frames of length `L=400` samples (corresponding to 40 msec in duration), with frame shift of `R=100` samples (corresponding to 10 msec shift duration);
8. compute the frame-by-frame modified correlation between the frames of signal specified as:

$$s_1[n] = [s[n], s[n+1], \ldots, s[n+L-1]],$$
$$s_2[n] = [s[n], s[n+1], \ldots, s[n+L+\text{pdhigh}-1],$$

where `n` is the starting sample of the current frame, and `pdhigh` is the longest anticipated pitch period (based on the gender of the speaker) and is specified as:

$$\texttt{pdhigh} = \begin{cases} \texttt{fsout/75} & \text{for males} \\ \texttt{fsout/150} & \text{for females} \end{cases}$$

(Hint: Use the MATLAB function `xcorr` to compute the modified correlation between $s_1[n]$ and $s_2[n]$ as it is considerably faster than any other implementation);

9. search from `pdlow` to `pdhigh` to find the maximum of the modified autocorrelation function (the putative pitch period estimate for the current frame), along with the value of the modified autocorrelation at the maximum (the confidence score), using the range estimate of:

$$\texttt{pdlow} = \begin{cases} \texttt{fsout/200} & \text{for males} \\ \texttt{fsout/300} & \text{for females,} \end{cases}$$

do all operations on both the original file and the bandpass filtered speech file;

10. convert the confidence score (the value of the modified autocorrelation at the maximum) to a log confidence, set a threshold at 0.75 of the maximum value of the log confidence score, and set the pitch period to zero for all frames whose log confidence scores fell below the threshold;
11. plot the resulting pitch period contour along with the log confidence scores for both the original speech file and the bandpass filtered speech file; how do these contours compare?
12. use a 5-point median smoother to smooth the pitch period scores as well as the confidence scores;
13. plot the median smoothed pitch period scores along with the median smoothed confidence scores;
14. save the computed pitch period and confidence scores in the file `out_autoc.mat`.

Which processing works best; i.e., using the full band original speech file or using the bandpass filtered speech file? How much difference do you observe in the resulting pitch period contours?

Problem 3

10.9. (MATLAB Exercise: Cepstrum-Based Pitch Detector). Program a pitch detector based on the (real) cepstrum of a speech signal. The real cepstrum is defined as the inverse FFT of the log magnitude spectrum of the signal, and the pitch period (for voiced speech sections) is found as the location of the peak of the cepstrum over the range of pitch periods appropriate for the gender of the speaker. A variation on standard pitch detectors is proposed in this exercise where the primary cepstral peak as well as a secondary cepstral peak (along with their peak amplitudes) are used as the basis for the pitch period decision.

The steps to follow in implementing the cepstrum-based pitch detector are similar to the ones used in the previous exercise (Problem 10.8), with the following exceptions:

- The signal processing parameters define the size of the FFT used to measure the spectrum and the cepstrum as `nfft=4000` to minimize aliasing.
- A threshold on the ratio of the primary cepstral peak (within the designated region of pitch periods) to the secondary cepstral peak is specified as `pthr1=4` and is used to define a region of "certainty (high confidence)" about pitch period estimates.
- The pitch period range for searching the cepstrum for the pitch peak is specified as the range $\texttt{nlow} \leq \texttt{n} \leq \texttt{nhigh}$, where the low and high of the pitch period range (as well as the cepstral search range) (for both males and females) is specified as:

PicthDetector_Cepstrum (s, fs, gender)

Resample -> (Equalize) -> Window -> Cepstrum searching
-> Reliable region -> Set zero for unreliable region -> Median smooth

$$\texttt{nlow} = \begin{cases} 40 & \text{for males} \\ 28 & \text{for females,} \end{cases}$$

$$\texttt{nhigh} = \begin{cases} 167 & \text{for males} \\ 67 & \text{for females.} \end{cases}$$

The process for finding primary and secondary cepstral peaks (within the specified range) is as follows:

1. Locate the maximum of the cepstrum over the specified range (`p1`), and record the quefrency at which the maximum occurs (`pd1`).
2. Zero the cepstrum over a range of ± 4 quefrencies around the maximum location found in the previous step, thereby eliminating the possibility that the secondary cepstral maximum will essentially be the same as the primary maximum.
3. Locate the secondary maximum of the processed cepstrum; record its quefrency (`pd2`), and its value (`p2`).

Figure P10.9(a) illustrates the results of the cepstral peak detection process. This figure shows a sequence of cepstral frames, displayed as a waterfall plot, with the primary cepstral peak indicated by a darkly shaded circle, and the secondary cepstral peak indicated by a lightly shaded circle. The continuity of pitch period, over frames, is clearly seen in parts of this figure.

The next step in the process is to define "reliable" regions of voiced speech. These regions are identified as those frames whose ratio of primary cepstral maximum to secondary cepstral maximum values exceeds the pre-set threshold of `pthr1=4.0`. Each of these reliable regions of voiced speech are extended by searching the neighboring regions (i.e., the frames prior to and following the reliable regions) and detecting adjacent frames whose primary or secondary pitch periods are within $\pm 10\%$ of the pitch period at the boundary frames. Whenever both putative pitch period estimates (i.e., primary and secondary) exceed the 10% difference threshold, the local search for additional adjacent frames is terminated. The next region of reliable voiced speech frames is searched and extended in a similar manner as above. This process is continued until all reliable regions have been extended to include neighboring pitch estimates that fall within the search criteria.

The final step in the signal processing is median smoothing of the pitch period contour using a 5-point median smoother.

Figure P10.9(b) shows plots of the resulting pitch period contour, along with the confidence scores (cepstral peak levels), for the waveform in the file `s6.wav`. It can be seen that, for the most part, regions of high cepstral value provide reliable pitch period estimates.

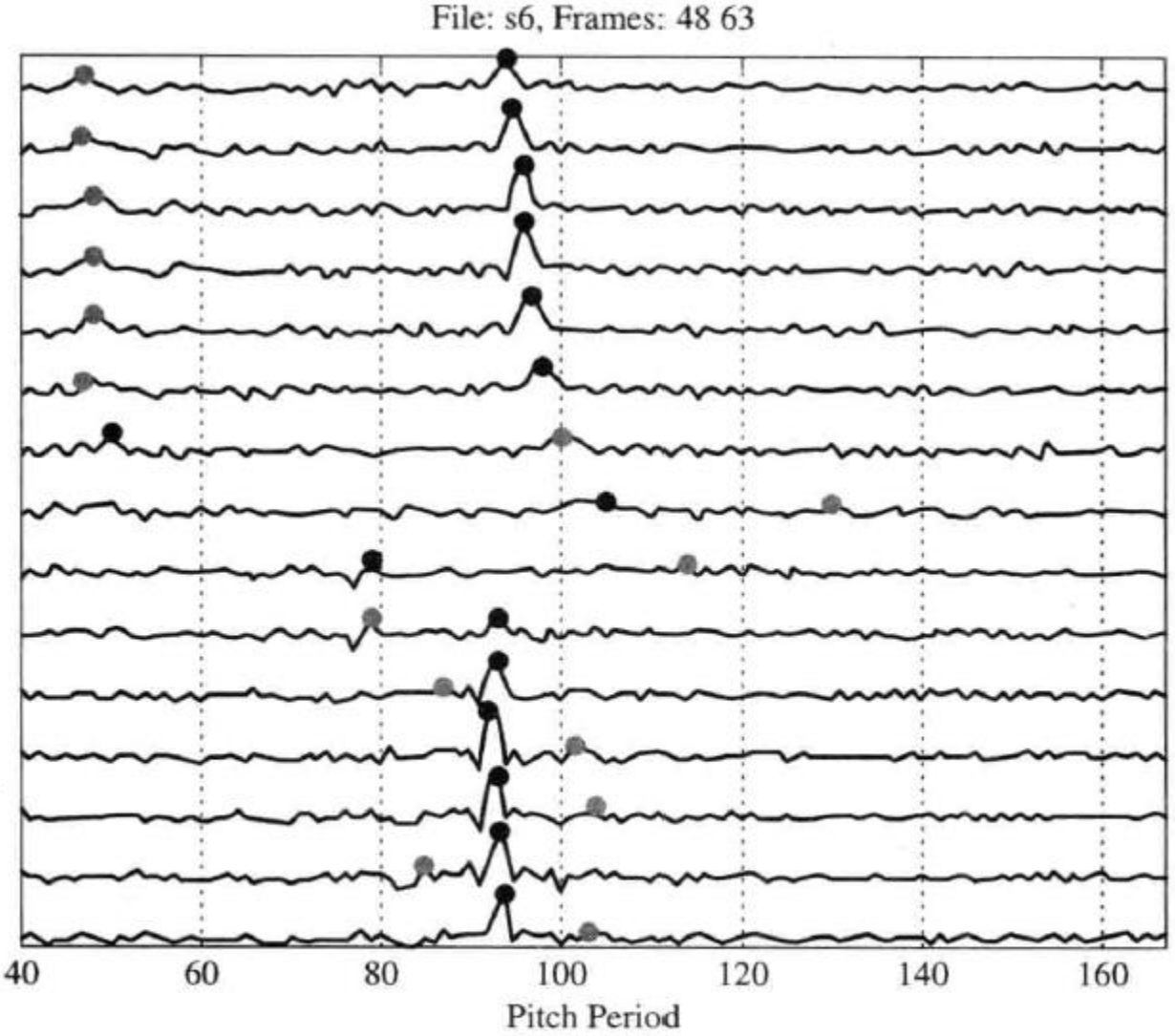


FIGURE P10.9(a)
Plots of sequence of cepstra in a waterfall display showing frame-by-frame cepstra with primary peak indicated by a darkly shaded circle, and secondary peak indicated by a lightly shaded circle.

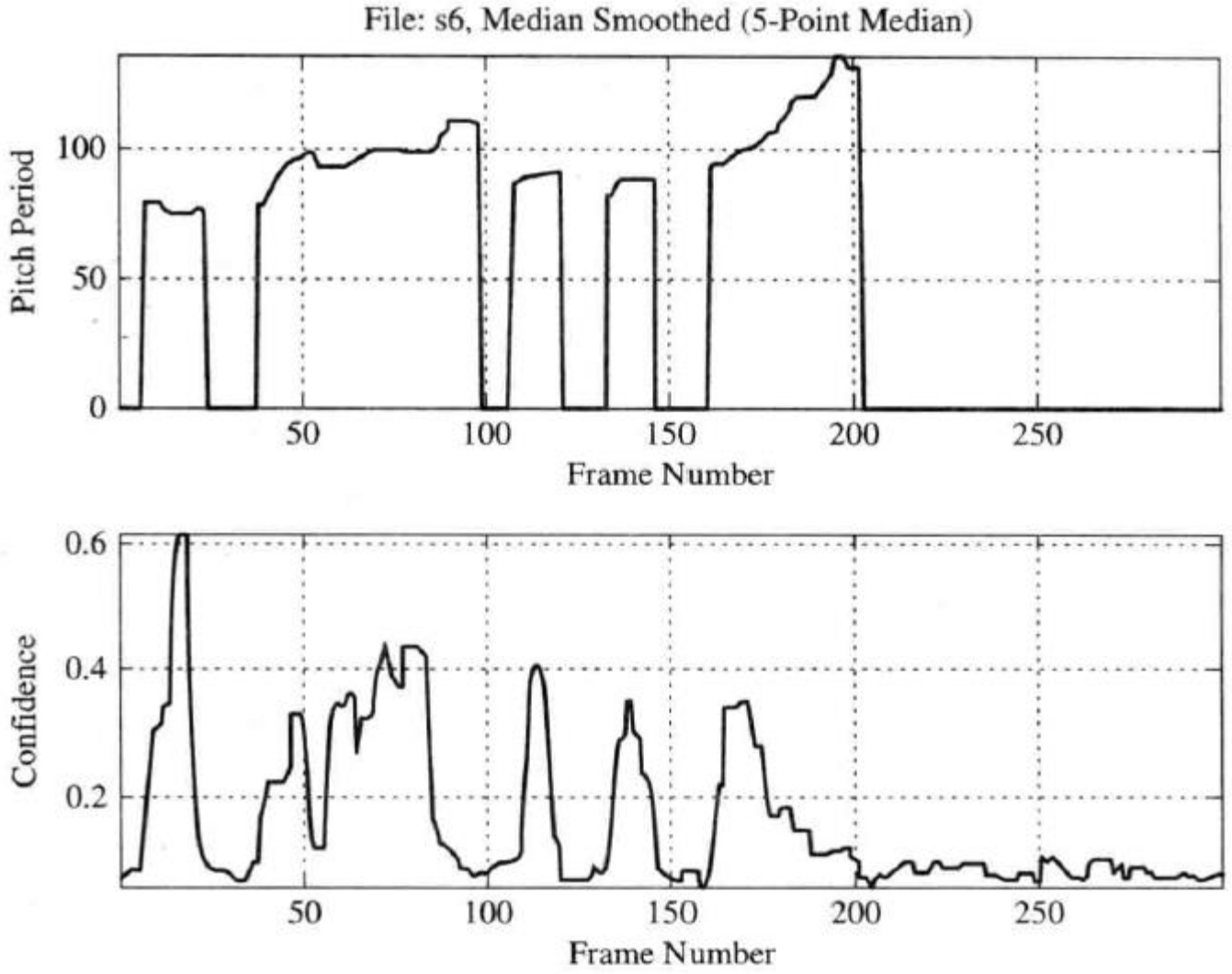


FIGURE P10.9(b)
Plots of pitch period estimates and cepstral magnitudes at the pitch peaks for the utterance in file `s6.wav`.