SNR-DEPENDENT WAVEFORM PROCESSING FOR IMPROVING THE ROBUSTNESS OF ASR FRONT-END

Author : Dusan Macho & Yan Ming Cheng
Professor : 陳嘉平
Reporter : 楊治鏞

Outline

- Introduction
- Basic idea
- Algorithm description
- Experiment

Introduction(1/2)

- Noise reduction
  - Spectral Subtraction
  - Wiener Filtering
- Thus, the assumption of good speech/noise detector is the fundamental weakness of these techniques.

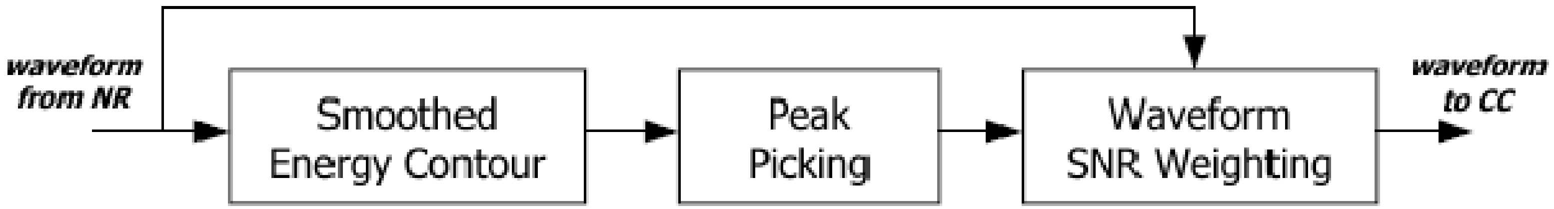
Introduction(2/2)

- In this paper, we explore a time domain based method as a complementary approach to the spectrum based speech enhancement techniques

Basic idea

- Waveform
- SNR
- SNR-dependent Waveform Processing (SWP)

Algorithm description



Smoothed energy contour

- Teager energy operator

$$E_{Teag}(n) = |s^2_{nr\_of}(n) - s_{nr\_of}(n-1) \times s_{nr\_of}(n+1)|, \quad 1 \le n < N_{in} - 1$$

$$E_{Teag}(0) = |s^2_{nr\_of}(0) - s_{nr\_of}(0) \times s_{nr\_of}(1)|$$

$$E_{Teag}(N_{in}-1) = |s^2_{nr\_of}(N_{in}-1) - s_{nr\_of}(N_{in}-2) \times s_{nr\_of}(N_{in}-1)|$$

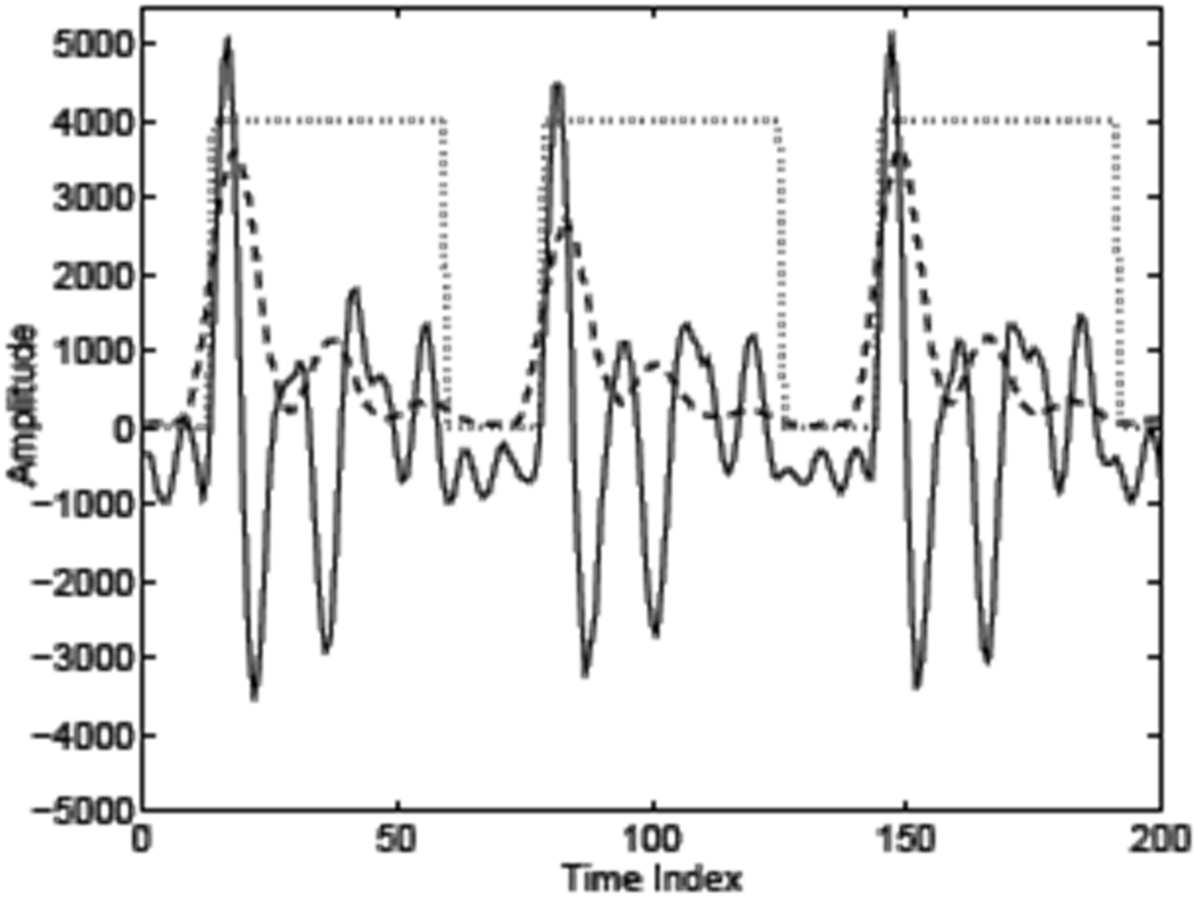
Smoothed energy contour

- The energy contour is smoothed by using a simple FIR filter of 9 like

$$E_{Teag\_Smooth}(n) = \frac{1}{9}\sum_{i=-4}^{4} E_{Teag}(n+i)$$

Peak picking

- A windowing function $w(n)$ is constructed for each frame in such a way that a rectangular unit window of width W is placed between each two adjacent maxima found within the frame.
- Each maximum is expected to be between 25 and 80 samples away from its neighbor.


5000
4000
3000
2000
1000
0
−1000
−2000
−3000
−4000
−5000
Amplitude
0
50
100
150
200
Time Index

Waveform SNR weighting

- $w(n)$
- W

$$\langle [pos_{MAX}(n_{MAX}) - 4],$$

$$[pos_{MAX}(n_{MAX}) - 4] + 0.8 \times [pos_{MAX}(n_{MAX} + 1) - pos_{MAX}(n_{MAX})] \rangle$$

$$0 \leq n_{MAX} < N_{MAX}$$

Waveform SNR weighting

$$s_{swp}(n) = f(\varepsilon)\cdot s_{highSNR}(n) + \varepsilon \cdot s_{lowSNR}(n)$$

$$= f(\varepsilon)\cdot w(n)s(n) + \varepsilon \cdot (1 - w(n))s(n)$$

$$f(\varepsilon) = \sqrt{\frac{\sum_n |s(n)|^2 - \varepsilon^2 \cdot \sum_n |(1-w(n))s(n)|^2}{\sum_n |w(n)s(n)|^2}}$$

$$0 < \varepsilon \le 1 \qquad f(\varepsilon) \ge 1$$

Waveform SNR weighting

- An important advantage of the SWP is that it does not need a speech/non-speech detector.
- SWP is applied after 2MWF, which would have already enhanced the SNR to the adequate level.

Experiment

- AURORA 2 database
- Multi-condition training (MCT)
- Clean speech training (CST)
- Testing: A, B, C
- Error reduction percentages

Table 1

Technique and parameter set	Multi-Condition Training			Clean Speech Training		
	A	B	C	A	B	C
2MWF (baseline)	26.37	21.54	33.81	47.03	53.76	37.04
2MWF+SWP, W=0.8, ε=0.9	27.71	24.57	35.09	50.86	55.43	43.86
2MWF+SWP, W=0.8, ε=0.8	***29.18***	***25.15***	***35.38***	***52.16***	***55.11***	***45.89***
2MWF+SWP, W=0.8, ε=0.7	26.62	23.47	33.89	52.92	54.94	47.17
2MWF+SWP, W=0.5, ε=0.9	27.78	25.20	34.52	50.72	55.59	44.16
2MWF+SWP, W=0.5, ε=0.8	27.81	24.95	33.64	51.81	55.52	46.13

Experiment

- SWP not only improves noise robustness but also increases the contrast between voiced and unvoiced speech that may help in clean speech recognition.

Table 2

Technique and parameter set	Multi-Condition Training	Clean Speech Training
	Clean Speech	**Clean Speech**
2MWF (baseline)	17.23	6.38
2MWF+SWP, W=0.8, ε=0.8	***32.43***	***13.01***

Experiment

- High SNR : W=0.8
- Low SNR : W=0.5
- Spectral subtraction

Table 3

Technique and parameter set	Multi-Condition Training			Clean Speech Training		
	A	B	C	A	B	C
2MWF, baseline	26.37	21.54	33.81	47.03	53.76	37.04
2MWF+SWP, ε=0.8, W=0.8	29.18	25.15	35.38	52.16	55.11	45.89
2MWF+SWP, ε=0.8, W_{SNR}=0.5-0.8	28.86	24.93	34.37	54.34	57.18	46.77
2MWF+SWP+SS, ε=0.8, W_{SNR}=0.5-0.8	27.26	23.78	33.25	***55.98***	***58.82***	***47.80***