Reverberant Speech Enhancement by Temporal and Spectral Processing

Author : P. Krishnamoorthy, S. R. Mahadeva Prasanna

Professor: 陳嘉平

Reporter: 吳國豪

Outline

- Introduction
- Spectral Processing of Reverberant Speech
- Temporal Processing of Reverberant Speech
- Experiments

Introduction

- **Reverberation** is one of the primary factors that degrade the quality of speech when collected by a distant microphone.

- This paper presents an approach for the enhancement of reverberant speech by **temporal** and **spectral** processing.

Introduction

- The spectral subtraction-based spectral processing methods **reduce the late reverberation** by estimating and subtracting the late reverberant spectrum from the degraded speech spectrum.

- Temporal processing involves **enhancement** of high signal-to-reverberation ratio (SRR) regions.

Spectral Processing of Reverberant Speech

- The spectral subtraction-based methods **estimate the power spectrum of the late reverberation** and then **subtract it** from the power spectrum of reverberant speech.

- The impulse response of the room $h(n)$ can be split into two parts $h_e(n)$ and $h_l(n)$, so that

$$h(n) = \begin{cases} h_e(n), \ for \ 0 \le n \le N_1 \\ h_l(n), \ for \ N_1 \le n \le L \\ 0, \ otherwisw \end{cases}$$

Spectral Processing of Reverberant Speech

- The short-time power spectral density (PSD) of **the reverberant speech signal** can be expressed as

$$S_{yy}(l,k) = S_{ye}(l,k) + S_{yl}(l,k)$$

where $S_{ye}(l,k)$ and $S_{yl}(l,k)$ are the PSD of the early and late reverberant components, respectively. Indexes *l* and *k* refer to time frame and frequency bin, respectively.

- The PSD of late reverberant:

$$\hat{S}_{yl}(l,k) = \gamma\omega(l - N_1) * \left|Y(l,k)\right|^2$$

where $Y(l,k)$ is the short-time Fourier transform of $y(n)$.

Spectral Processing of Reverberant Speech

- The smoothing function $\omega(l)$:

$$\omega(l) = \begin{cases} \dfrac{l+a}{a^2} e^{\frac{-(l-a)^2}{2a^2}}, & l > -a \\ 0, & otherwise \end{cases}$$

- The spectral subtraction process can be described as a filtering operation in the frequency domain by

$$\hat{S}(l,k) = Y(l,k)G(l,k)$$

Spectral Processing of Reverberant Speech

$$G(l,k) = \sqrt{\frac{|Y(l,k)|^2 - \hat{S}_{yl}(l,k)}{|Y(l,k)|^2}} = 1 - \frac{1}{\sqrt{r(l,k)}}$$

$$r(l,k) = \frac{|Y(l,k)|^2}{\hat{S}_{yl}(l,k)}$$

- The enhanced spectra obtained using the above relation may contain some negative values. The simplest solution is to **half-wave rectify** these values to ensure a nonnegative magnitude spectrum.

Spectral Processing of Reverberant Speech

- This nonlinear processing of negative values, however, creates small, isolated peaks in the spectrum occurring at random frequency locations in each frame. This type of noise is commonly referred as **musical noise**.

- The modification consists of using a spectral floor to prevent the gain from descending below a lower bound, as proposed in

$$\left|\hat{S}(l,k)\right| = \begin{cases} \left|Y(l,k)\right| G(l,k), \ if \ \left|Y(l,k)\right| G(l,k) > \beta \left|Y(l,k)\right| \\ \beta \left|Y(l,k)\right|, \ otherwise \end{cases}$$

Temporal Processing of Reverberant Speech

- Sum of peaks in the DFT spectrum:
  - The sum of amplitudes of the major peak locations will be **higher** in high SRR regions than other SRR regions.
  - This property is exploited in the identification of **high SRR regions** of the reverberant speech.
  - The high SRR regions are identified by using **the sum of the ten largest peaks** in the DFT spectrum.

Temporal Processing of Reverberant Speech

- In the proposed method, the indicators of the high SRR regions are first further enhanced. This is achieved with the help of **the first-order difference** (FOD) of the indicators obtained.

- Since FOD represents the **slope**, the positive to negative going zero transition in FOD locates the **peaks** in the sum of DFT spectrum values.

Temporal Processing of Reverberant Speech

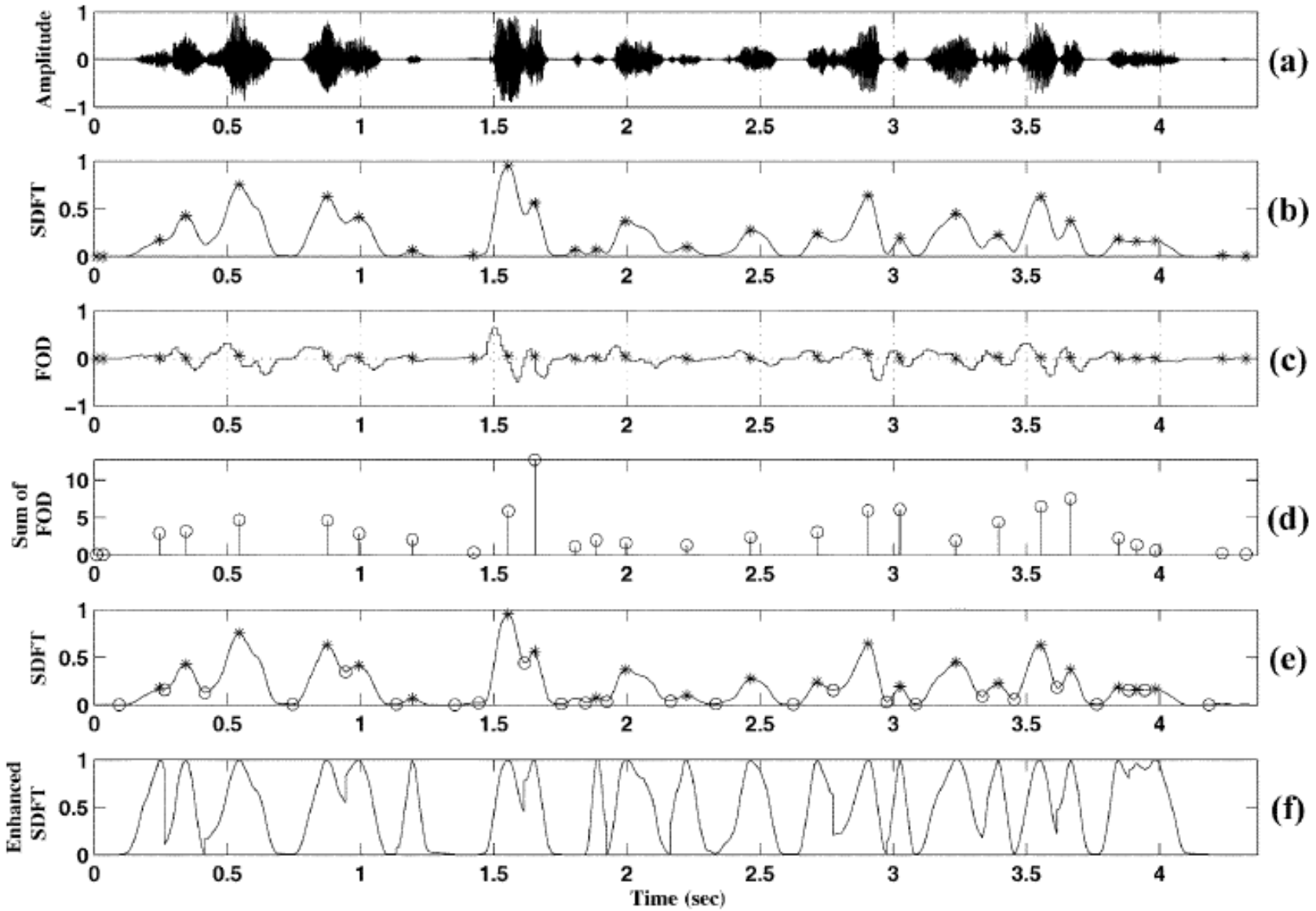
- The sum of absolute FOD values computed for a duration of 5 ms on either side with reference to each positive to negative going zero crossing point.

- If two successive peaks occur within 50 ms then the peak with the lower FOD value is eliminated based on the assumption that occurrence of two high SRR regions unlikely within a 50 ms interval.

Temporal Processing of Reverberant Speech

- The star $(*)$ symbols in Fig. 2(e) show the peak locations after eliminating the undesirable peaks. With respect to each of these local peaks the nearest negative to positive going zero transition points on either side are identified and are marked by circles in Fig. 2(e).

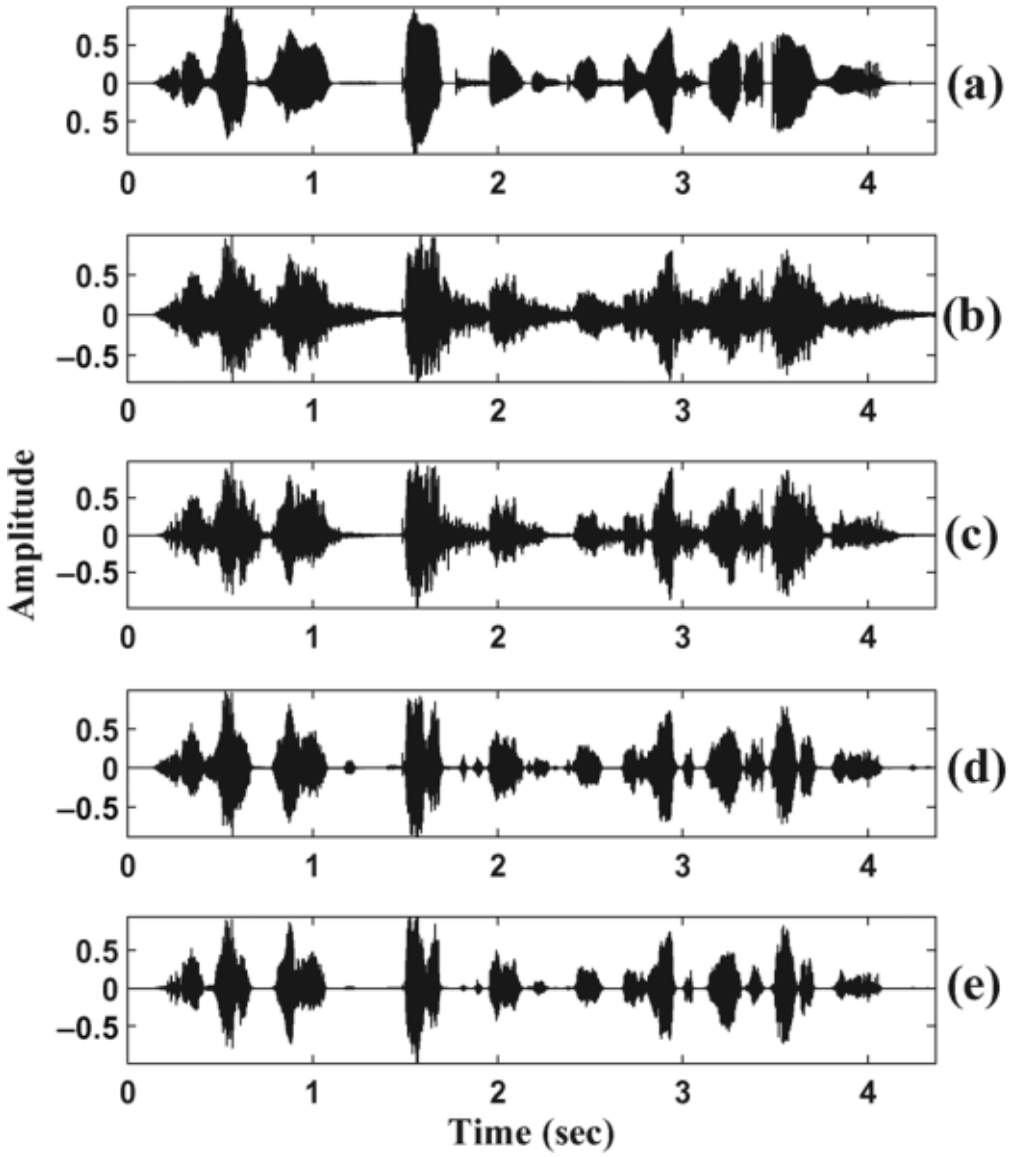
- The regions between the circles are enhanced by taking the normalized value of that particular region and is shown in Fig. 2(f).


Amplitude
(a)
SDFT
(b)
FOD
(c)
Sum of FOD
(d)
SDFT
(e)
Enhanced SDFT
(f)
Time (sec)

Experiments

- An experiment is carried out to illustrate the objective of the proposed method using simulated room impulse response. Speech example is taken from the **TIMIT database**.

- Fig. 11(a) and (b) shows the clean speech and the corresponding reverberant speech obtained by convolving the obtained impulse response with a reverberation time of about 1 s.


Amplitude
0.5
0
0. 5
(a)
0
1
2
3
4
0.5
0
−0.5
(b)
0
1
2
3
4
0.5
0
−0.5
(c)
0
1
2
3
4
0.5
0
−0.5
(d)
0
1
2
3
4
0.5
0
−0.5
(e)
0
1
2
3
4
Time (sec)

T_{60}(sec)	**SDFT**
	Microphone Distance=1 m
0.2	86.00
0.4	85.45
0.6	85.31
0.8	85.03
1.0	85.58
	Microphone Distance=1.5 m
0.2	86.56
0.4	86.42
0.6	85.65
0.8	85.45
1.0	85.24
	Microphone Distance=2 m
0.2	86.70
0.4	85.65
0.6	84.47
0.8	81.27
1.0	80.92

Experiments

- The segmental SRR (SegSRR) of the frame is defined as

$$SegSRR(l) = 10\log_{10}\left[\frac{\sum_{n=lR}^{lR+N-1} S_d^2(n)}{\sum_{n=lR}^{lR+N-1} (S_d(n) - \hat{S}(n))^2}\right]$$

$$S_d(n) = S(n) * h_d(n)$$

N is the number of samples per frame and *R* is the frame rate in samples.

T_{60}(sec)	REV	SP	TP	TPSP
	Microphone Distance=1 m			
0.2	1.65	3.02	3.54	3.00
0.4	-1.66	0.11	0.24	0.14
0.6	-3.02	-0.72	-1.12	-0.64
0.8	-3.77	-1.12	-1.91	-0.99
1.0	-4.27	-1.44	-2.45	-1.24
	Microphone Distance=1.5 m			
0.2	0.33	1.54	2.02	1.55
0.4	-2.64	-0.40	-0.67	-0.33
0.6	-3.87	-1.07	-1.94	-0.99
0.8	-4.52	-1.39	-2.60	-1.19
1.0	-4.98	-1.64	-3.12	-1.40
	Microphone Distance=2 m			
0.2	-0.89	0.59	0.80	0.68
0.4	-3.99	-1.39	-2.20	-1.21
0.6	-5.05	-2.17	-3.35	-2.00
0.8	-5.59	-2.70	-3.95	-2.51
1.0	-5.94	-3.05	-4.33	-2.96

$$\omega(l) = \begin{cases} \dfrac{l+a}{a^2} e^{\frac{-(l-a)^2}{2a^2}}, & l > -a \\ 0, & otherwise \end{cases}$$

a	3	4	5	6	7	8	9	10
T_{60}(sec)	Segmental SRR (SegSRR)							
0.2	0.49	0.56	**0.59**	**0.59**	0.58	0.56	0.54	0.54
0.4	-1.43	**-1.39**	**-1.39**	-1.43	-1.47	-1.52	-1.56	-1.59
0.6	-2.23	-2.18	**-2.17**	-2.18	-2.23	-2.28	-2.33	-2.38
0.8	-2.84	-2.75	-2.70	**-2.68**	-2.69	-2.72	-2.76	-2.81
1.0	-3.24	-3.12	-3.05	**-3.00**	-3.02	-3.05	-3.07	-3.19