ON THE USE OF WIDEBAND SIGNAL FOR NOISE ROBUST ASR

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

Outline

- Introduction
- Investigated approaches
- Recognition experiments

Introduction(1/2)

- The use of wider frequency range (wideband signal) in telecommunication applications is expected soon.
- We consider that the wideband (w-b) signal ranges from 0 to 8kHz.

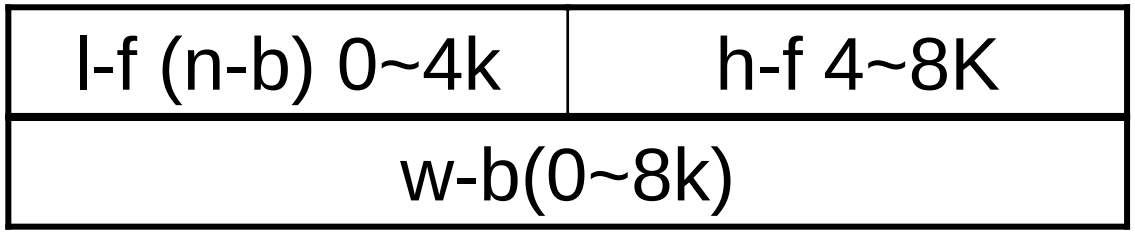
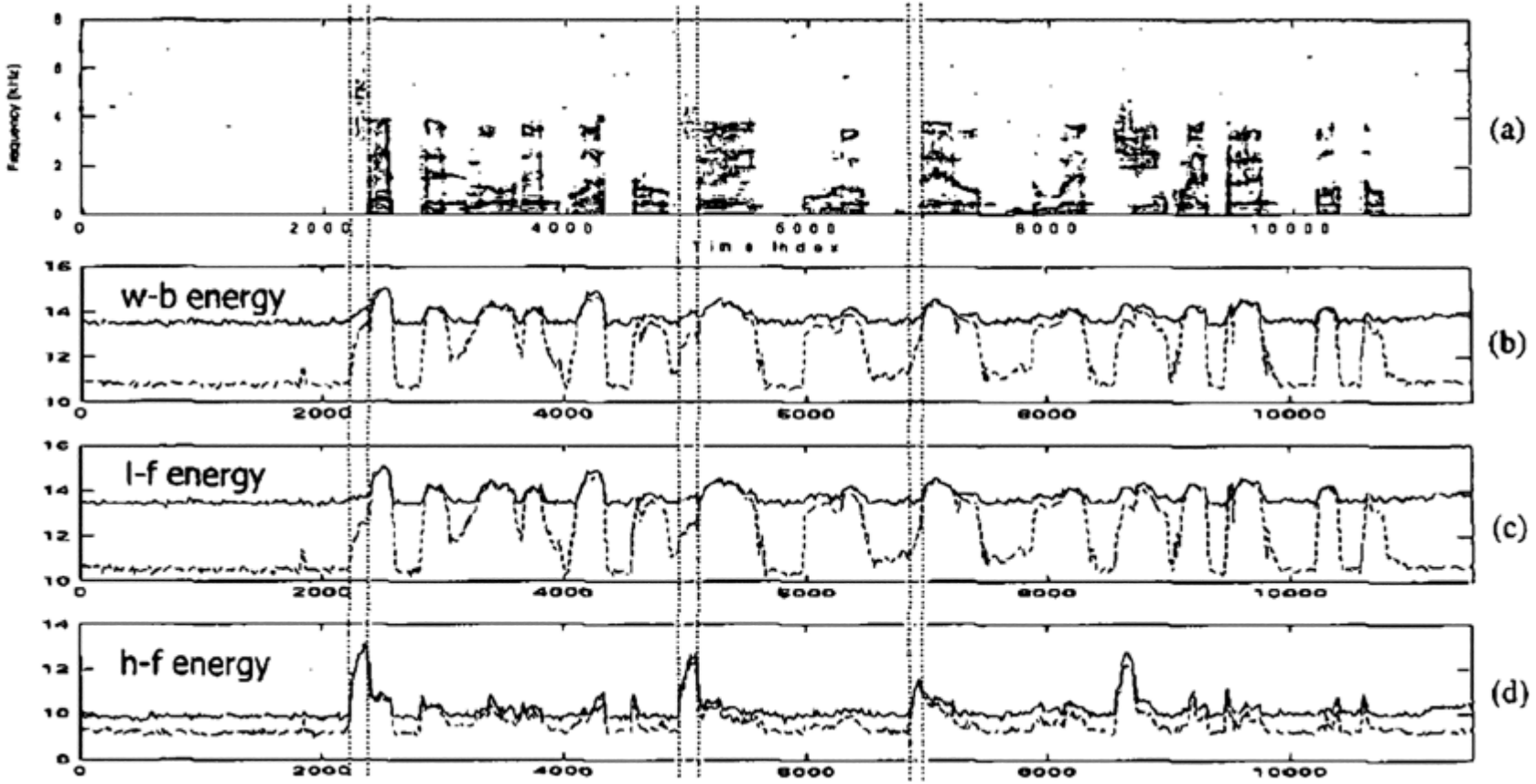


Figure 2

Wideband spectrogram of Italian digit sequence "sette nove quattro sei uno zero due cinque Ire otto" - close-talking microphone(a).



Introduction(2/2)

- In the presence of noise, the combination of the l-f speech information, which is high-SNR, with the h-f speech information, which is low-SNR, would cause that the w-b speech features become more affected by noise than the n-b speech features.

- Therefore, the benefit of adding the h-f speech information into the ASR speech representation is questionable when Considering noisy speech recognition.

Investigated approaches

- When comparing Figures 2(b) and 2(d)
- we can see that the identification of the h-f speech sounds in the case of low-SNR signal (solid lines) can be done much better from the h-f energy contour than from the w-b energy contour

Investigated approaches

- a) the noise robust front-end designed originally for the n-b signal can be reused for processing of the l-f part of w-h signal

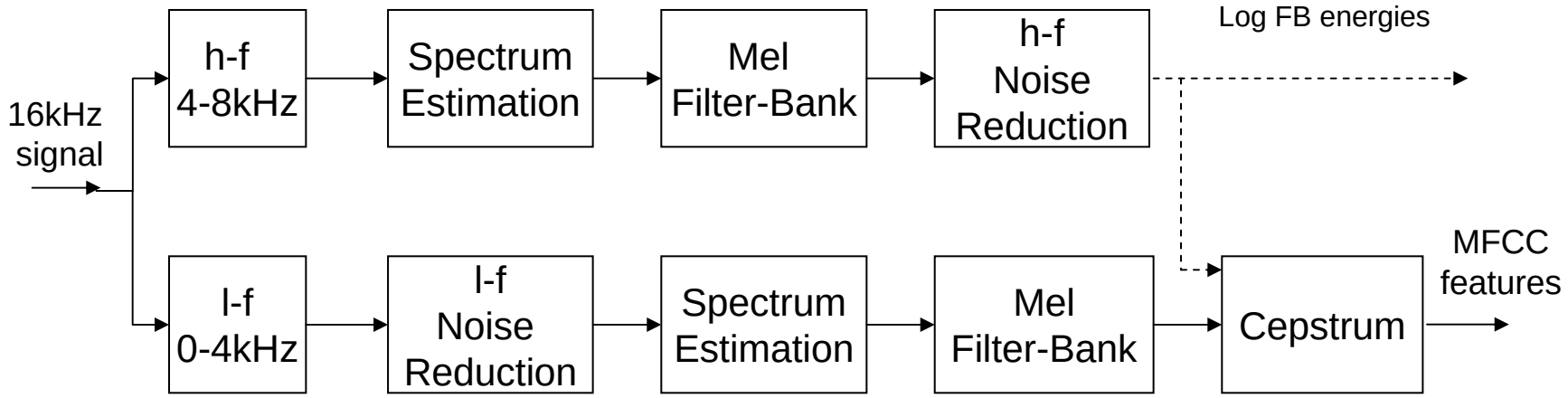
- b) due to a relatively lower information content, the processing of the h-f part of w-h signal may be less complex than that of the l-f part

Description of approaches(1/3)

- Both l-f and h-f parts of w-b signal are obtained by filtering the input w-h signal by a couple of quadrature mirror filters.

- Basic concept is depicted in Figure ***3***

Description of approaches(2/3)



Description of approaches(3/3)

- use linear spectral subtraction(SS)
- To minimize the mismatch, we apply an adjustment to the h-f filter-bank energies, which is based on an encoding/decoding scheme.

Spectral subtraction in h-f

Linear spectral subtraction is applied on the h-f filter-bank energies $E_h(\mathrm{k})$ like

$$E_{ss\_h}(k) = \max\{E_h(k) - \alpha \cdot \hat{N}_h(k), \beta \cdot E_h(k)\},$$

$$1 \le k \le K_h$$

K_h is the number of bands, $\alpha = 1.5$, $\beta = 0.1$

The noise estimation $\hat{N}_{\mathrm{h}}(\mathrm{k})$ is updated by using only the frames labeled by voice activity detector as noise.

Encoding scheme in h-f

$$S_h(k) = \ln(E_h(k)), \quad 1 \le k \le K_h$$

$$Code(k, j) = S_{l\_aux}(k) - S_h(j), \; 1 \le k, j \le K_h$$

where $S_{l\_aux}(k)$ are $K_l = K_h = 3$ auxiliary log filter-bank energies from the 2-4kHz frequency range of the l-f singal before applying l-f noise reduction.

Decoding scheme in h-f

$$S_{code\_h}(h) = \sum_{j=1}^{K_h} w_{code}(j)(S_{den\_l\_aux}(j) - Code(j,k))$$

$$1 \le k \le K_h$$

where $S_{den\_l\_aux}(k)$ are filter-bank energies of the de-noised l-f spectrum, and they are aligned in frequency with $S_{l\_aux}(k)$, $w_{code}(j)$ are frequency-dependent weights and their sum equals 1.0

Integration of SS and E/D scheme in h-f

The mismatch was introduced by the different noise reductions used in the l-f and h-f parts of w-b signal. In practice, the h-f energies from spectral subtraction $S_{ss\_h}(k)$ are adjusted as follows:

$$S_h(k) = \lambda \cdot S_{code\_h}(k) + (1-\lambda) \cdot S_{SS\_h}$$

$$1 \le k \le K_h$$

where $\lambda = 0.7$ was determined experimentally

Recognition experiments

- SpeechDat Car (SDC) digit databases
- Spanish,Finnish,Danish,Italian
- well-match (WM)

  medium mismatched (MM)

  highly mismatched (HM)
- VAD

Experiment and results

- 8kHz and 16kHz data
- window size change: 200 to 400
- Window shift change:80 to 160
- FFT order change: 256 to 512
- Filter-bank bands:23 and 30
- Number of static cepstral coefficients:12+

Table 1

Sampling Frequency	SP			FI			IT			DA			Average of Abs			Aver of Rel
	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	
8k, baseline	92.94	83.31	51.55	92.74	80.51	40.53	92.26	73.39	51.76	87.28	67.32	39.37	**91.31**	**76.13**	**45.80**	--
16k, fb 23	93.76	84.00	50.23	93.69	84.54	35.02	93.59	76.71	41.29	89.19	70.18	30.81	**92.56**	**78.86**	**39.34**	**4.59**
16k, fb 30	93.95	83.29	53.68	92.78	82.63	34.52	93.13	76.67	40.24	88.63	71.35	32.74	**92.12**	**78.49**	**40.30**	**2.63**

Experiment and results 2

- First approach, we tested the addition of the de-noised h-f log filter-bank energies to cepstral features. The best results were obtained by adding two h-f log energies.
- In the second approach, three de-noised h-f log filter-bank energies were added to the 23 de-noised l-f log filter-bank energies.

Table 2

Sampling Frequency	SP			FI			IT			DA			Average of Abs			Aver of Rel
	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	
8k, baseline	96.03	92.65	88.33	95.53	85.98	87.07	96.45	89.85	87.72	92.74	82.03	80.71	**95.19**	**87.63**	**85.96**	--
16k, 1st approach	95.68	91.63	82.70	97.19	88.10	89.54	96.67	91.73	90.89	92.35	83.33	75.73	**95.47**	**88.70**	**84.72**	**2.25**
16k, 2nd approach	95.90	93.31	87.00	97.17	90.97	88.66	97.32	92.85	89.24	93.87	85.68	78.07	**96.07**	**90.70**	**85.74**	**13.96**