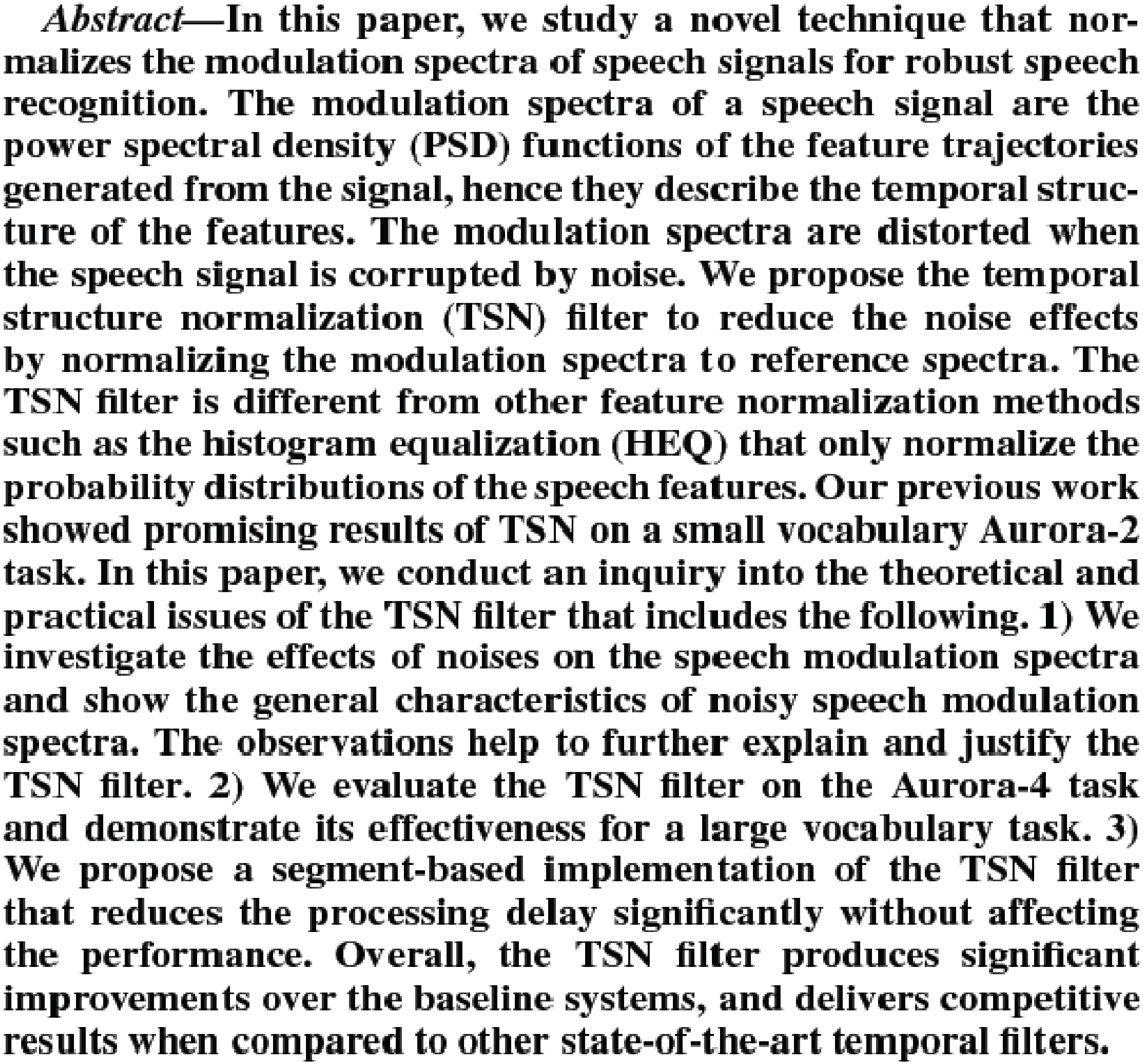
Normalization of the Speech Modulation Spectra for Robust Speech Recognition

Author: Xiong Xiao , Eng Siong Chng , Haizho Li

Professor : 陳嘉平

Reporter : 許峰閤

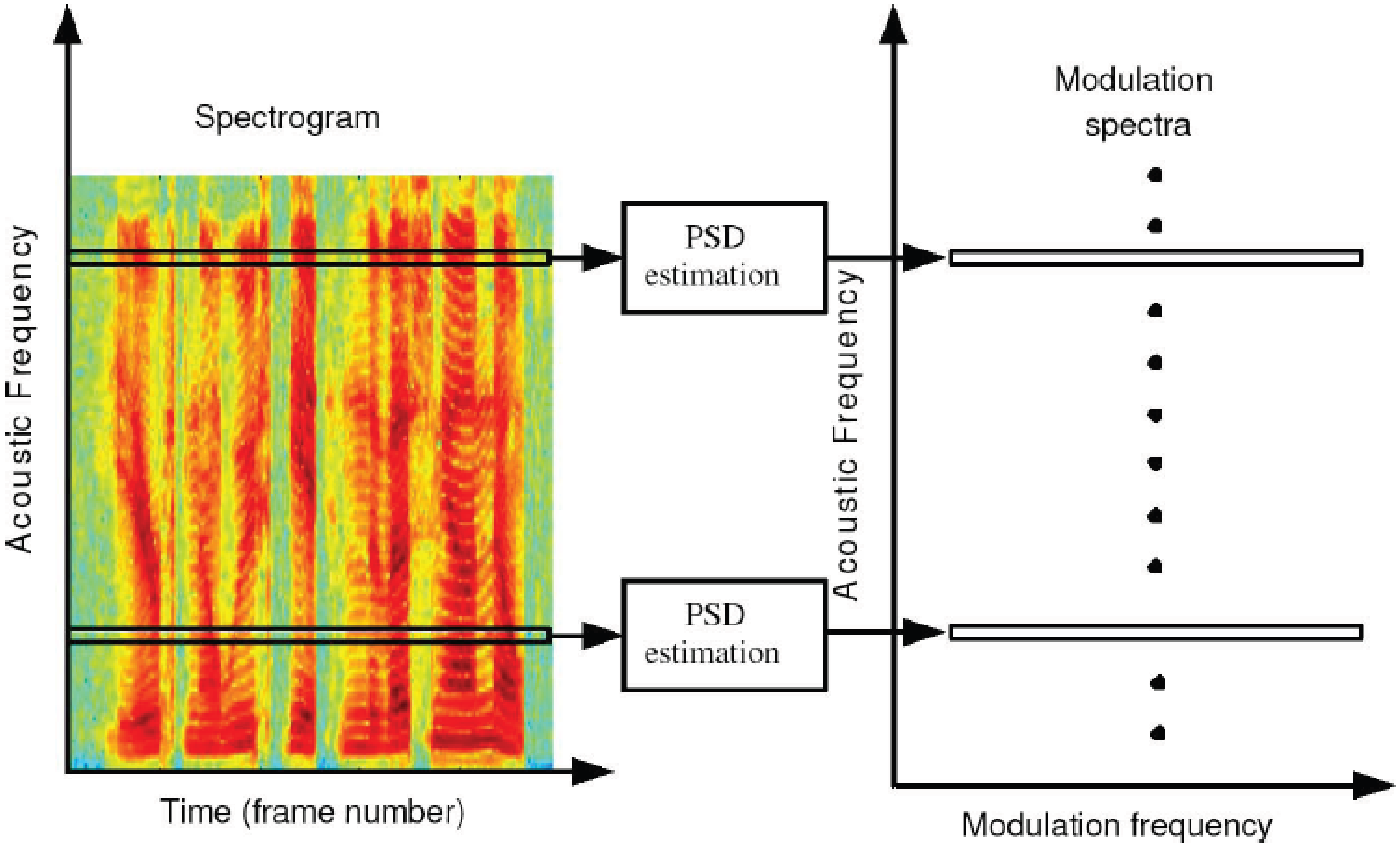


Abstract—In this paper, we study a novel technique that normalizes the modulation spectra of speech signals for robust speech recognition. The modulation spectra of a speech signal are the power spectral density (PSD) functions of the feature trajectories generated from the signal, hence they describe the temporal structure of the features. The modulation spectra are distorted when the speech signal is corrupted by noise. We propose the temporal structure normalization (TSN) filter to reduce the noise effects by normalizing the modulation spectra to reference spectra. The TSN filter is different from other feature normalization methods such as the histogram equalization (HEQ) that only normalize the probability distributions of the speech features. Our previous work showed promising results of TSN on a small vocabulary Aurora-2 task. In this paper, we conduct an inquiry into the theoretical and practical issues of the TSN filter that includes the following. 1) We investigate the effects of noises on the speech modulation spectra and show the general characteristics of noisy speech modulation spectra. The observations help to further explain and justify the TSN filter. 2) We evaluate the TSN filter on the Aurora-4 task and demonstrate its effectiveness for a large vocabulary task. 3) We propose a segment-based implementation of the TSN filter that reduces the processing delay significantly without affecting the performance. Overall, the TSN filter produces significant improvements over the baseline systems, and delivers competitive results when compared to other state-of-the-art temporal filters.

Definition of Modulation Spectra

- 調變頻譜是針對每個特徵中,在一個頻率範圍內做PSD(Power Spectral Density)形成一個調變的頻率

$$|X(t,f)|^2 = |\mathbf{STFT}[x(i)]|^2$$

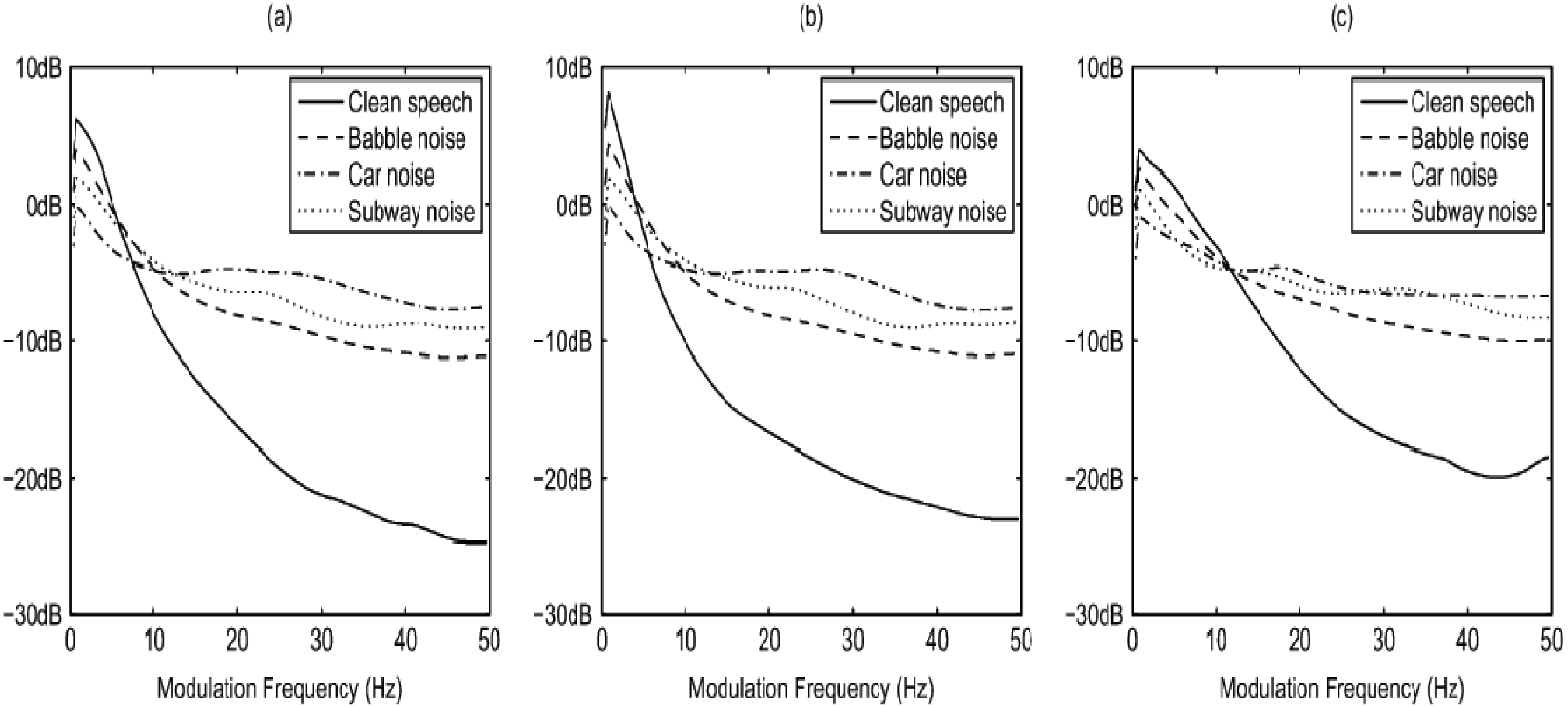

Spectrogram
Acoustic Frequency
Time (frame number)
PSD estimation
PSD estimation
Modulation spectra
Acoustic Frequency
Modulation frequency

Modulation Spectra of Clean Speech and Noises

- 接著要對乾淨的語音及噪音分別做調變,來觀察兩者的不同及變化

- 我們發現語音的能量大部分都集中在較低的調變頻率中(1~16Hz)

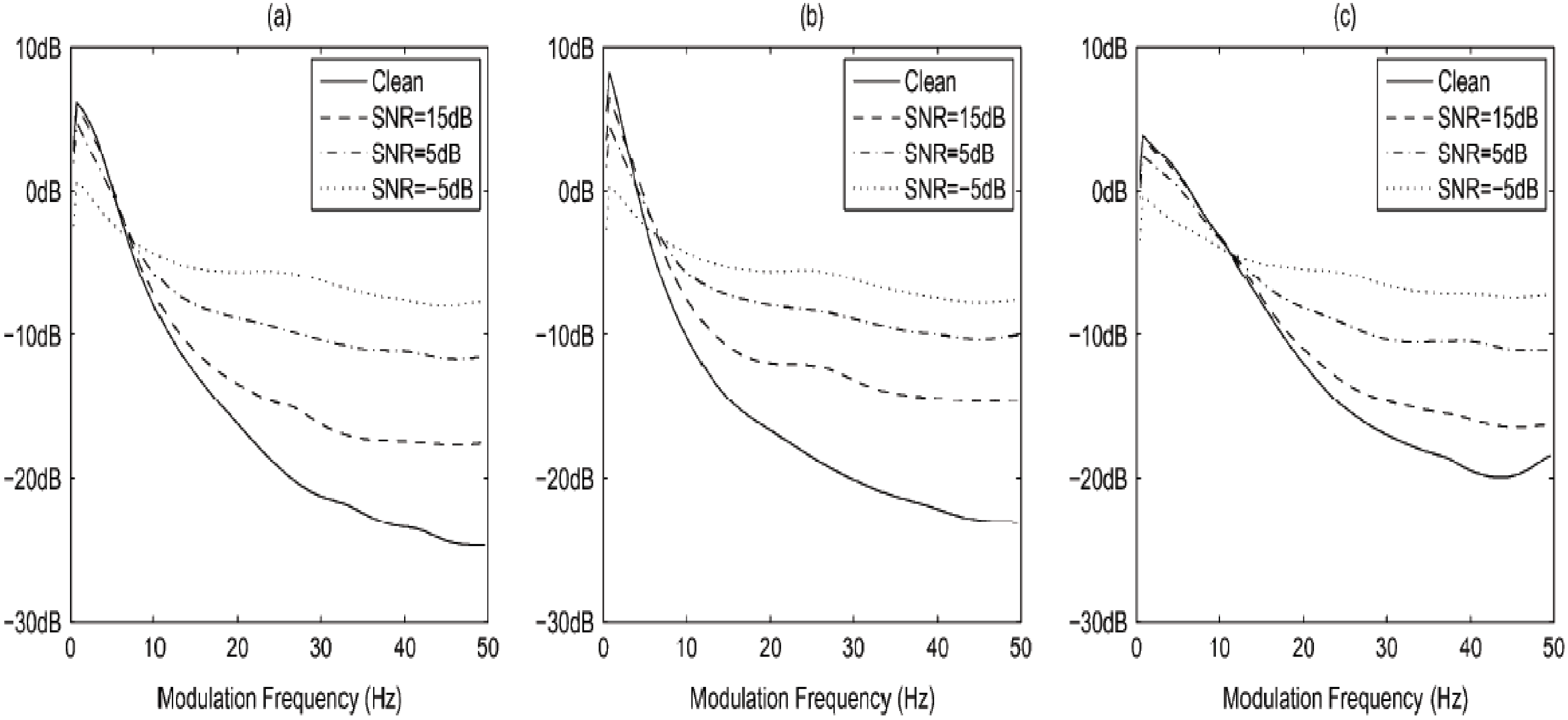
- 噪音的能量走勢在任合的調變頻率中都較平滑


(a)
(b)
(c)
10dB
0dB
-10dB
-20dB
-30dB
0
10
20
30
40
50
Modulation Frequency (Hz)
Clean speech
Babble noise
Car noise
Subway noise

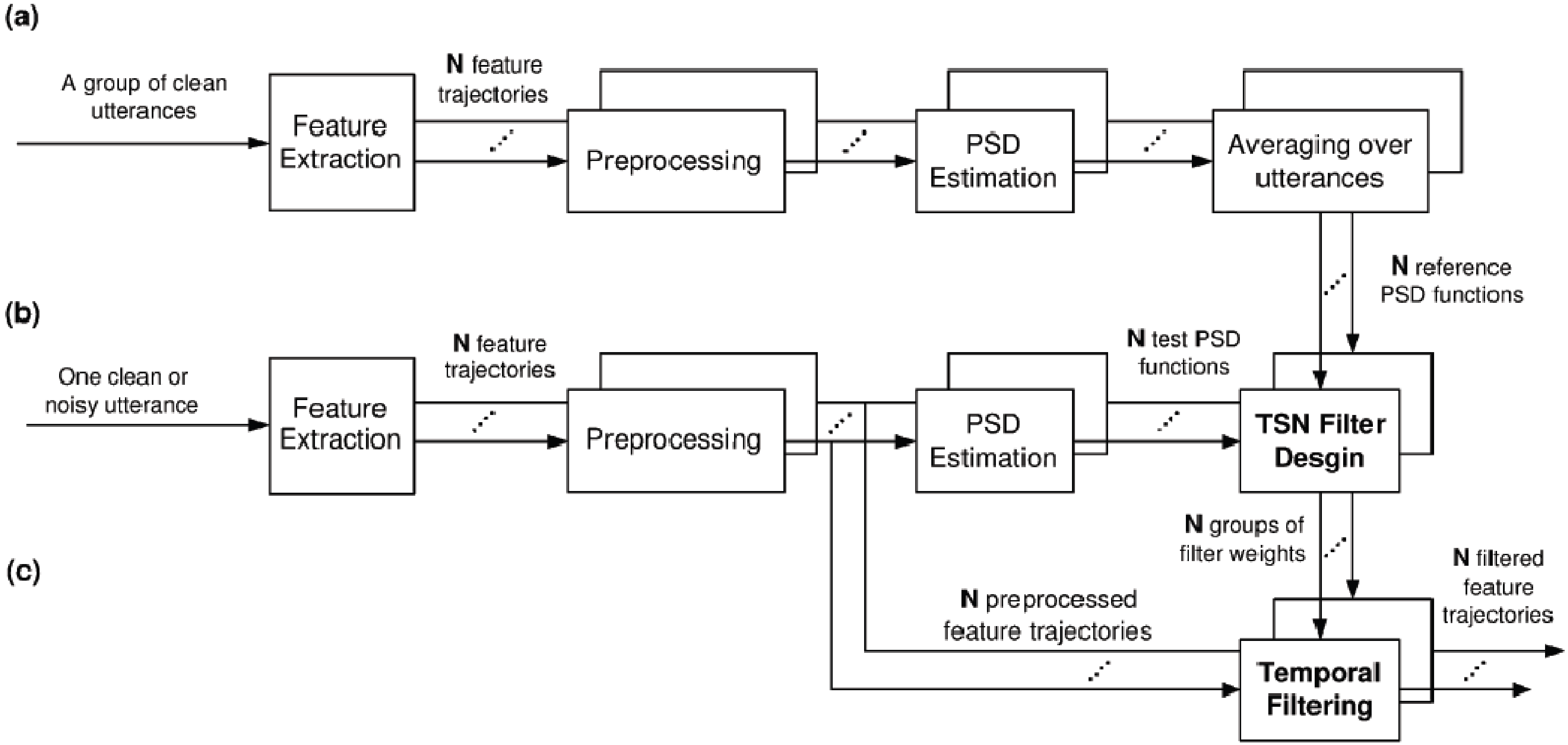
Modulation Spectra of Noisy Speech

- 針對乾淨的語音受到噪音的影響以後,再來對Modulation Spectra做分析
- 發現能量還是集中在低Modulation Frequency的地方,而且走勢的平緩程度與SNR成反比

$$P_y = P_x + P_n + Q$$


(a)
(b)
(c)
10dB
0dB
-10dB
-20dB
-30dB
0
10
20
30
40
50
Modulation Frequency (Hz)
Clean
SNR=15dB
SNR=5dB
SNR=−5dB

TSN Framework



TSN Filter Design

- 這邊設計一個時間結構過濾器(TSN Filter),來減少clean及noisy feature之間的差異
- 首先Filter中的振幅響應決定於Modulation Spectra

$$|H(k)| = \sqrt{P_{\mathrm{ref}}(k)/P_{\mathrm{test}}(k)}.$$

- 也可與其他的Filter合併,本篇論文中採取與ARMA Filter作合併

$$|H'(k)| = |H(k)||G(k)|$$

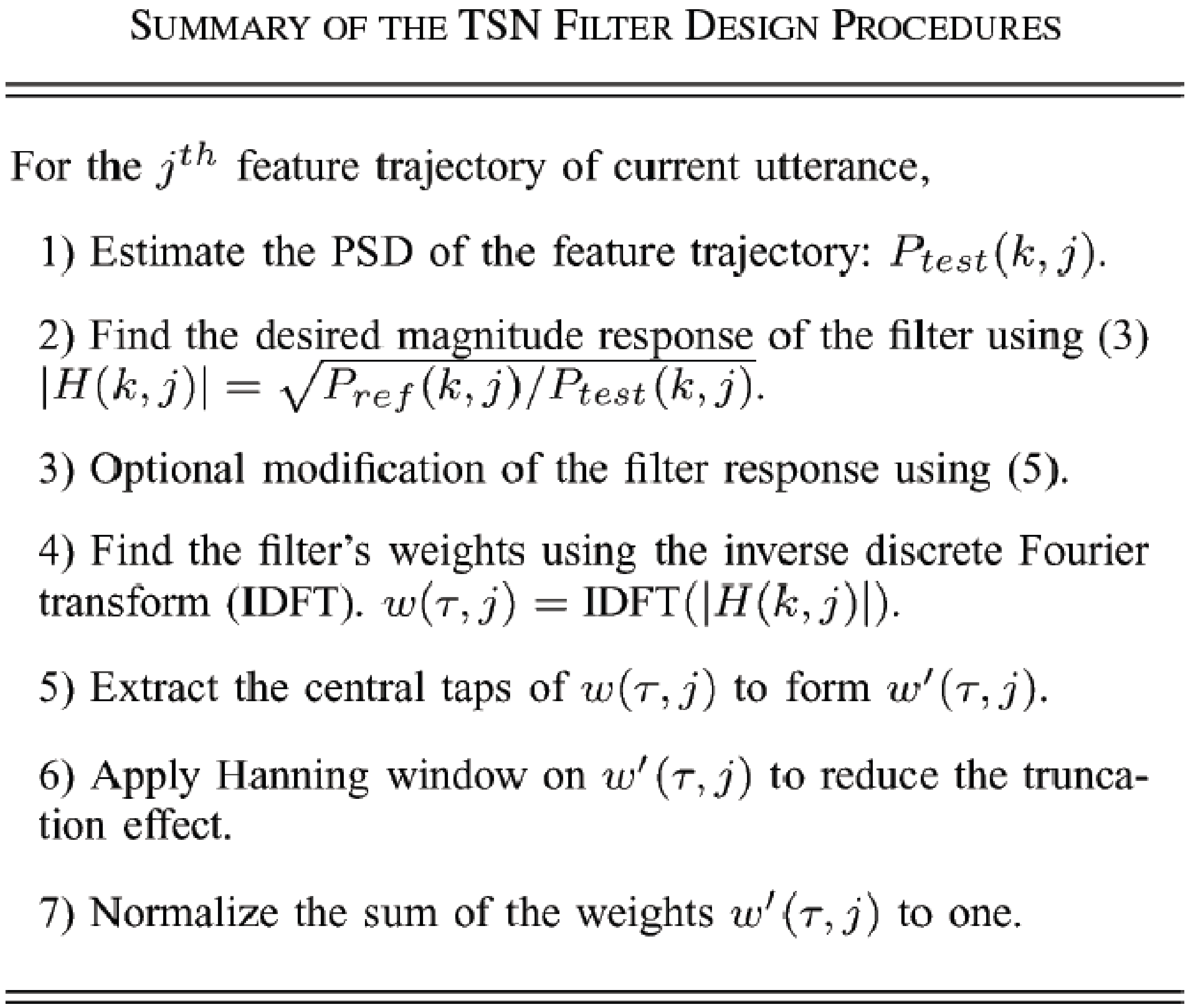
TSN Filter Design

- 而正規化後的振幅響應可表示為下式

$$P_{\mathrm{norm}}(k) = |H(k)|^2 P_{\mathrm{test}}(k) = P_{\mathrm{ref}}(k)$$

- FIR Filter Coefficients可由振幅響應做IDFT所得到

- 接著使用Hamming Window來消除Truncation Effect



SUMMARY OF THE TSN FILTER DESIGN PROCEDURES

For the j^{th} feature trajectory of current utterance,

1) Estimate the PSD of the feature trajectory: $P_{test}(k, j)$.

2) Find the desired magnitude response of the filter using (3) $|H(k, j)| = \sqrt{P_{ref}(k, j)/P_{test}(k, j)}$.

3) Optional modification of the filter response using (5).

4) Find the filter's weights using the inverse discrete Fourier transform (IDFT). $w(\tau, j) = \mathrm{IDFT}(|H(k, j)|)$.

5) Extract the central taps of $w(\tau, j)$ to form $w'(\tau, j)$.

6) Apply Hanning window on $w'(\tau, j)$ to reduce the truncation effect.

7) Normalize the sum of the weights $w'(\tau, j)$ to one.

Segment-Based Implementation

- 在前面提到的TSN Filter中,都是Utterance by Utterance的,但是這樣會造成系統的延遲

- 在這邊我們將一個Utterance切成許多一樣大小並且有重疊的Segment來減少延遲的時間

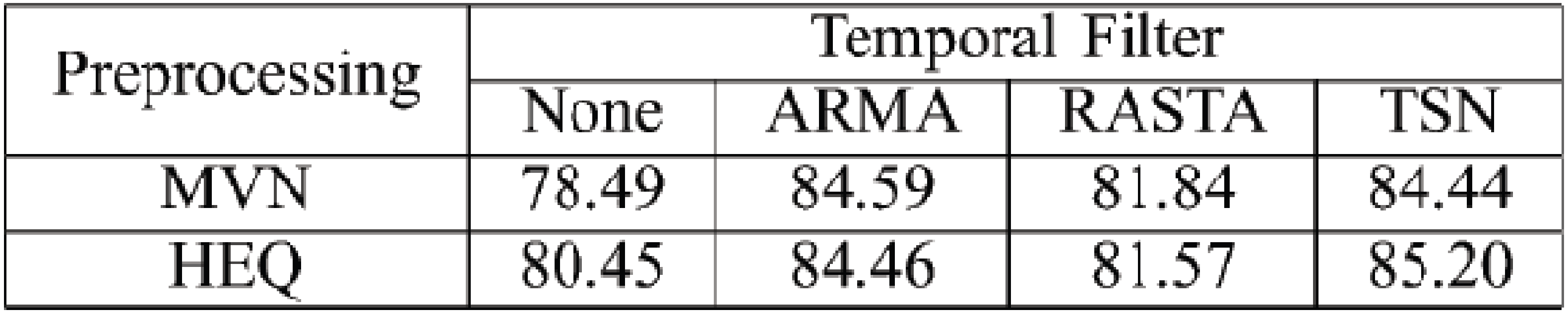
Experiments

- 語料庫使用AURORA 2以及AURORA 4

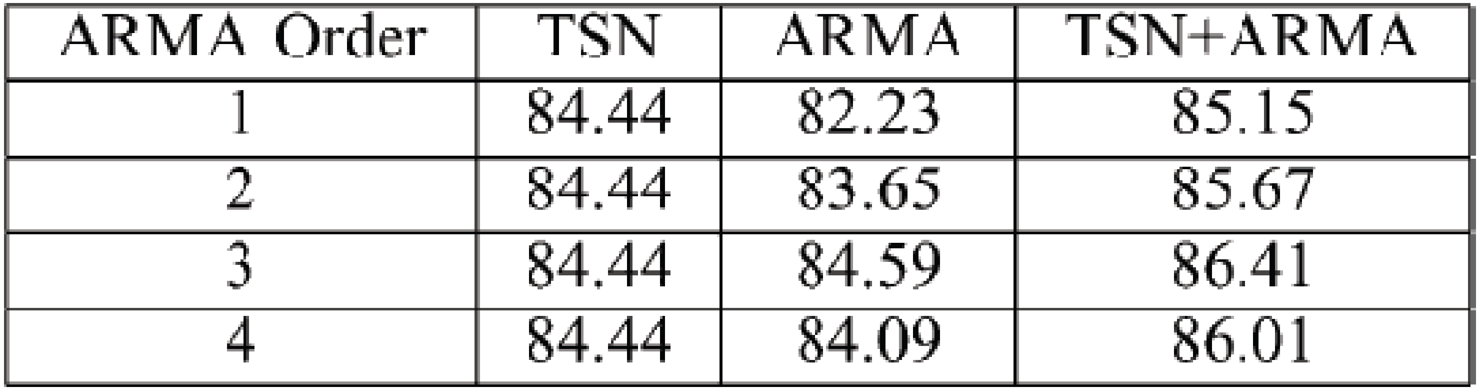
- AURORA 2是一串連續的英文數字,每個Utterance長度約為1.8秒

- AURORA 4是大詞彙的連續英文語音,每個Utterance長度約為7.6秒

Experiments



Preprocessing	Temporal Filter			
	None	ARMA	RASTA	TSN
MVN	78.49	84.59	81.84	84.44
HEQ	80.45	84.46	81.57	85.20



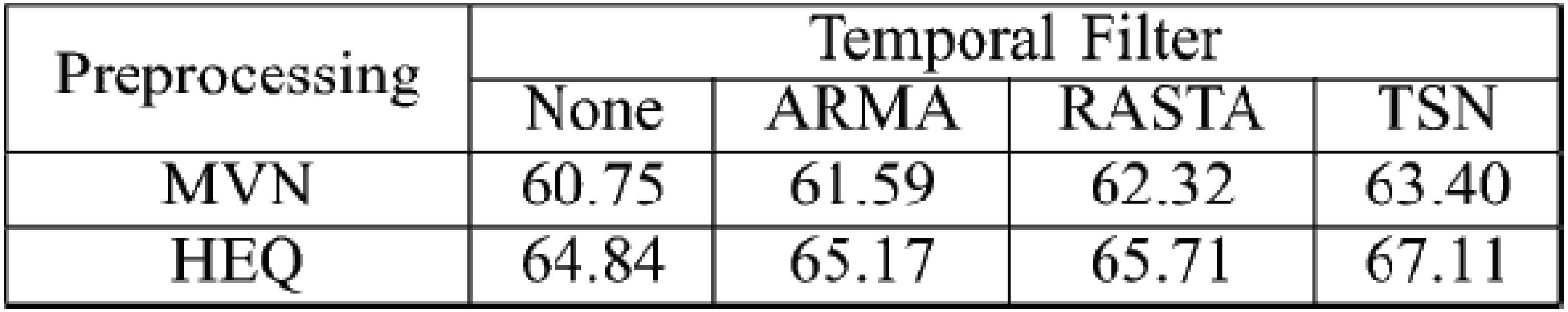
ARMA Order	TSN	ARMA	TSN+ARMA
1	84.44	82.23	85.15
2	84.44	83.65	85.67
3	84.44	84.59	86.41
4	84.44	84.09	86.01

Experiments

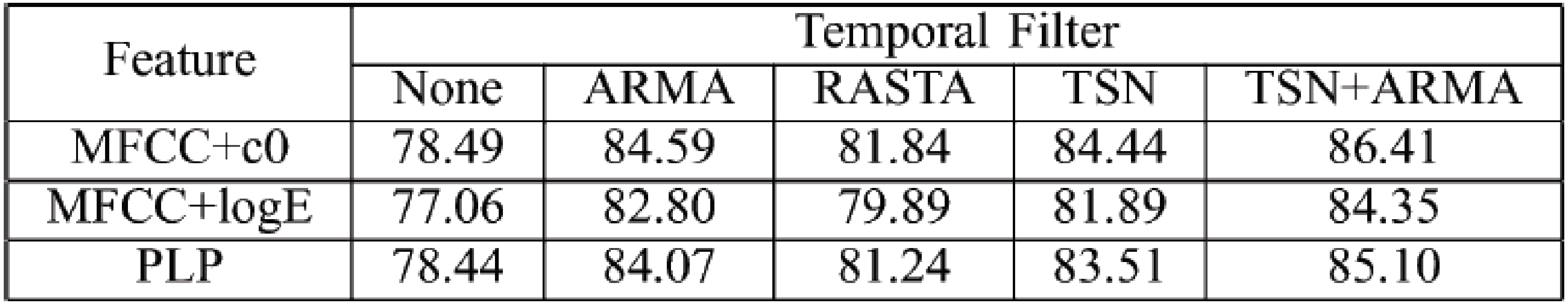
- 在AURORA 4中因為Utterance較長,所以在前置處理中可以先使用HEQ做處理

- 對Feature作平滑化可以降低mismatch的情況但是也會移除掉一些必要的資訊

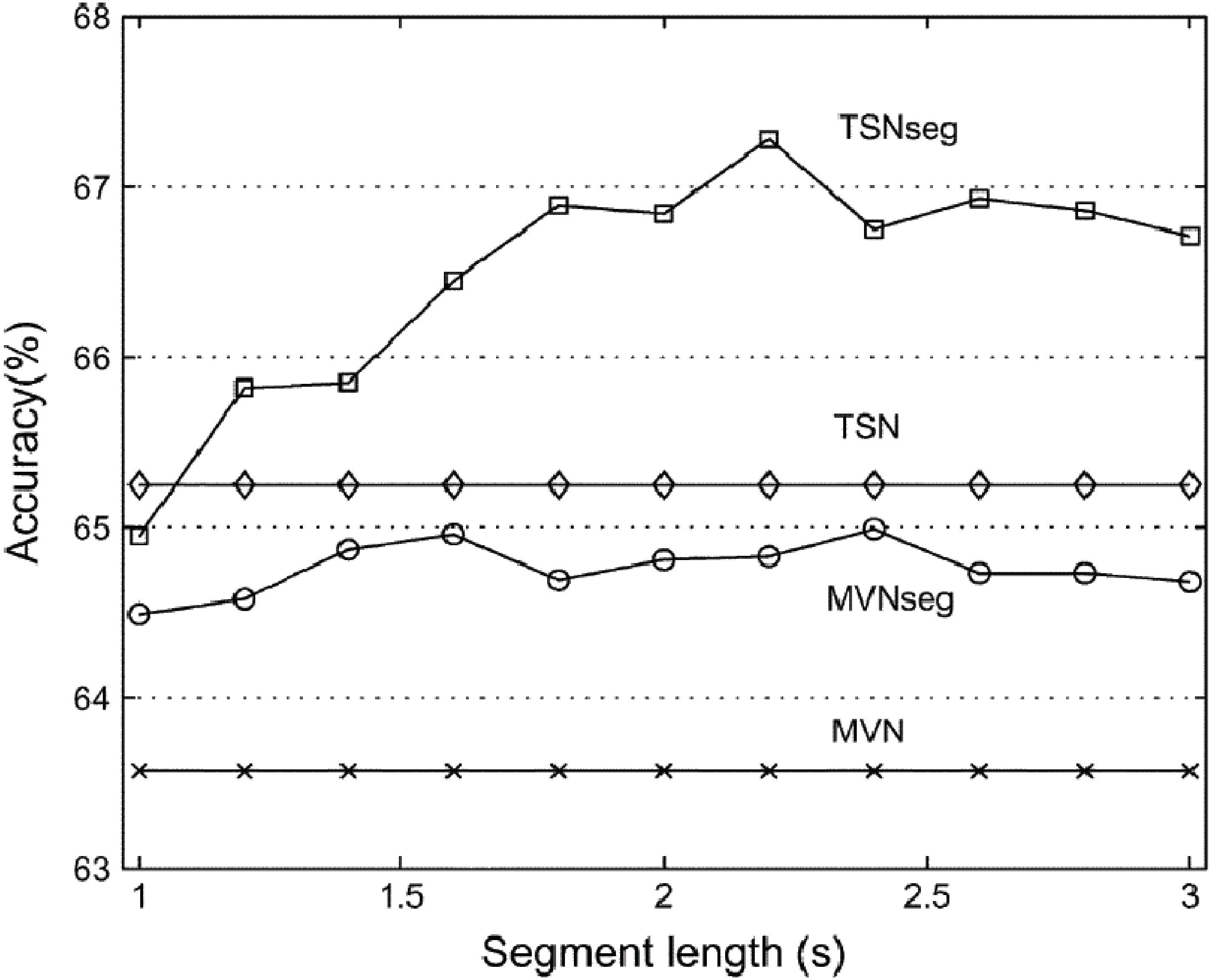
- 所以對AURORA 2的短語句我們可以加強特徵的平滑化(ARMA order =3),而AURORA 4則取order=1



Preprocessing	Temporal Filter			
	None	ARMA	RASTA	TSN
MVN	60.75	61.59	62.32	63.40
HEQ	64.84	65.17	65.71	67.11



Feature	Temporal Filter				
	None	ARMA	RASTA	TSN	TSN+ARMA
MFCC+c0	78.49	84.59	81.84	84.44	86.41
MFCC+logE	77.06	82.80	79.89	81.89	84.35
PLP	78.44	84.07	81.24	83.51	85.10


68
67
66
65
64
63
Accuracy(%)
1
1.5
2
2.5
3
Segment length (s)
TSNseg
TSN
MVNseg
MVN