Normalization of the Speech Modulation Spectra for Robust Speech Recognition

Xiong Xiao, Eng Siong Chng, Haizhou Li

Reporter:邱聖權

Professor:陳嘉平

Introduction

- This paper introduces a temporal filter to normalize the modulation spectra of a speech utterance.

Modulation spectra

- Speech is a wide-band signal, it is more appropriate to analyze the amplitude modulation for individual frequency bands.
- The energy envelope of each band can be seen as the modulating signal of the band.
- The collection of the modulation spectra from all frequency bands forms the joint acoustic-modulation frequency representation of the speech signal.

Modulation spectra

- Short-time power spectral density (spectrogram) of a time domain signal can be obtained as follows

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

where $x(i)$ is the time domain signal,

i,t,f are the sample index, frame index, and frequency index.

Modulation spectra

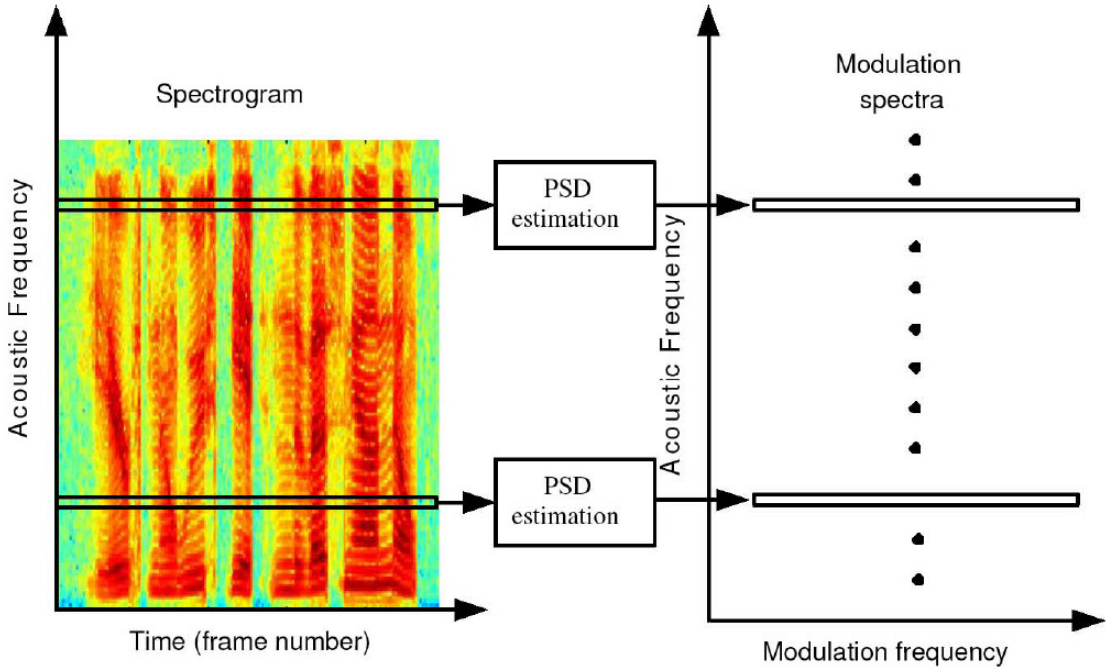
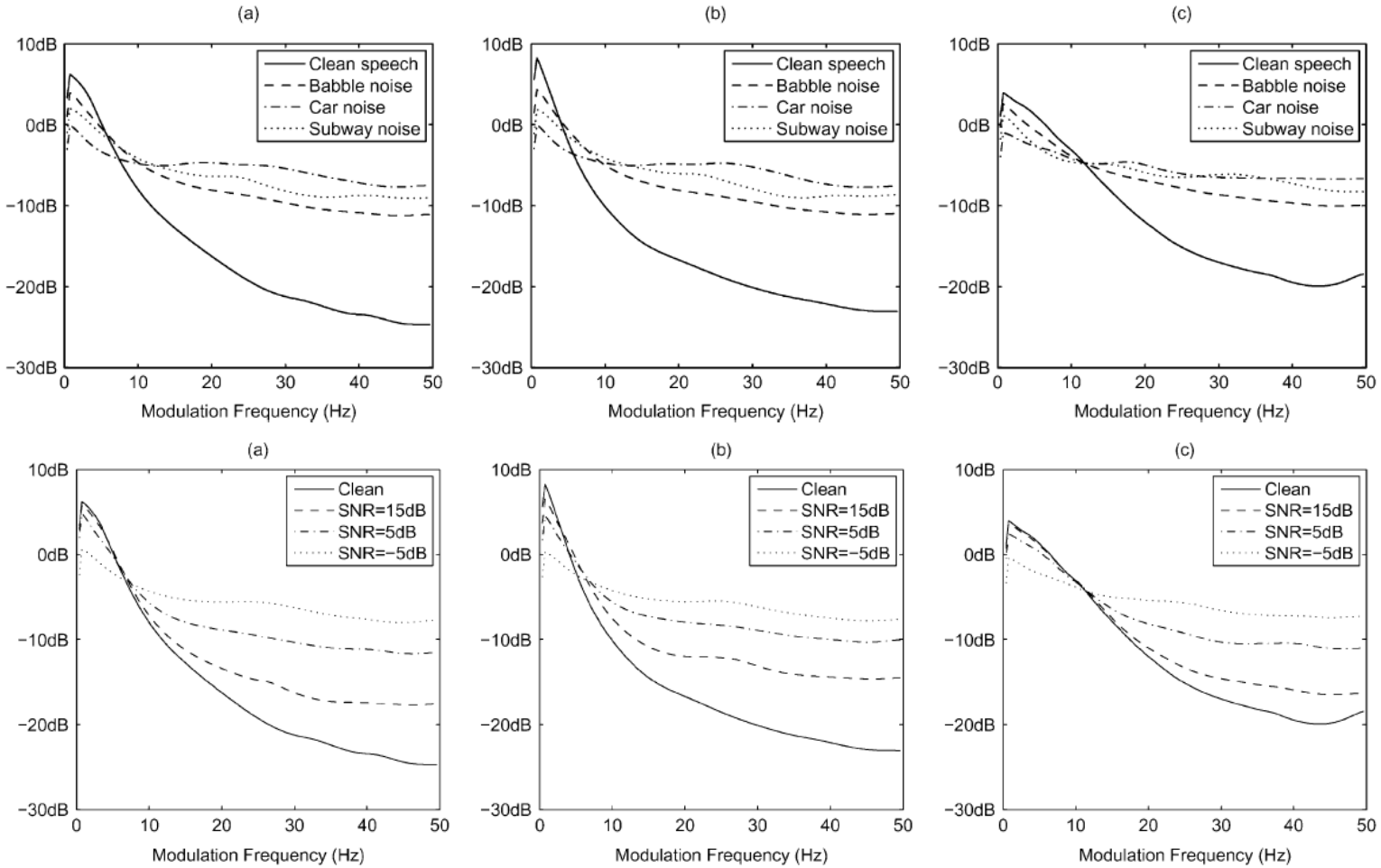


Fig. 1. Computation of the modulation spectra of a speech signal. The left panel is the spectrogram of a speech signal. The modulation spectra of the speech signal shown on the right panel are the PSD functions of the spectrogram trajectories.

Modulation spectra

- The trajectory of spectrogram coefficients in a single acoustic frequency bin represents the energy envelope of the speech signal in the acoustic frequency band centered at the bin.
- The PSD function of each trajectory is the modulation spectrum for the acoustic frequency bin.
- The PSD functions for all the bins are the modulation spectra of the speech signal

Modulation spectra of clean speech, noises, and noisy speech



Mel filterbank log Mel filterbank Cepstral coefficient

Normalization of PSD

- The temporal structure normalization (TSN) filter is introduced to normalize the PSD.
- The magnitude response of the filter is designed from the PSD functions of the feature trajectories.

Training of reference PSD

- The PSD of training data is estimated to form the reference PSD, then the PSD of testing data is normalized to the reference PSD by TSN filter.
- The reference PSD is obtained from averaging the PSD of training data.
- The averaging process reduces other factors affecting the PSD functions, such as spoken content and speaker's characteristics.

TSN framework

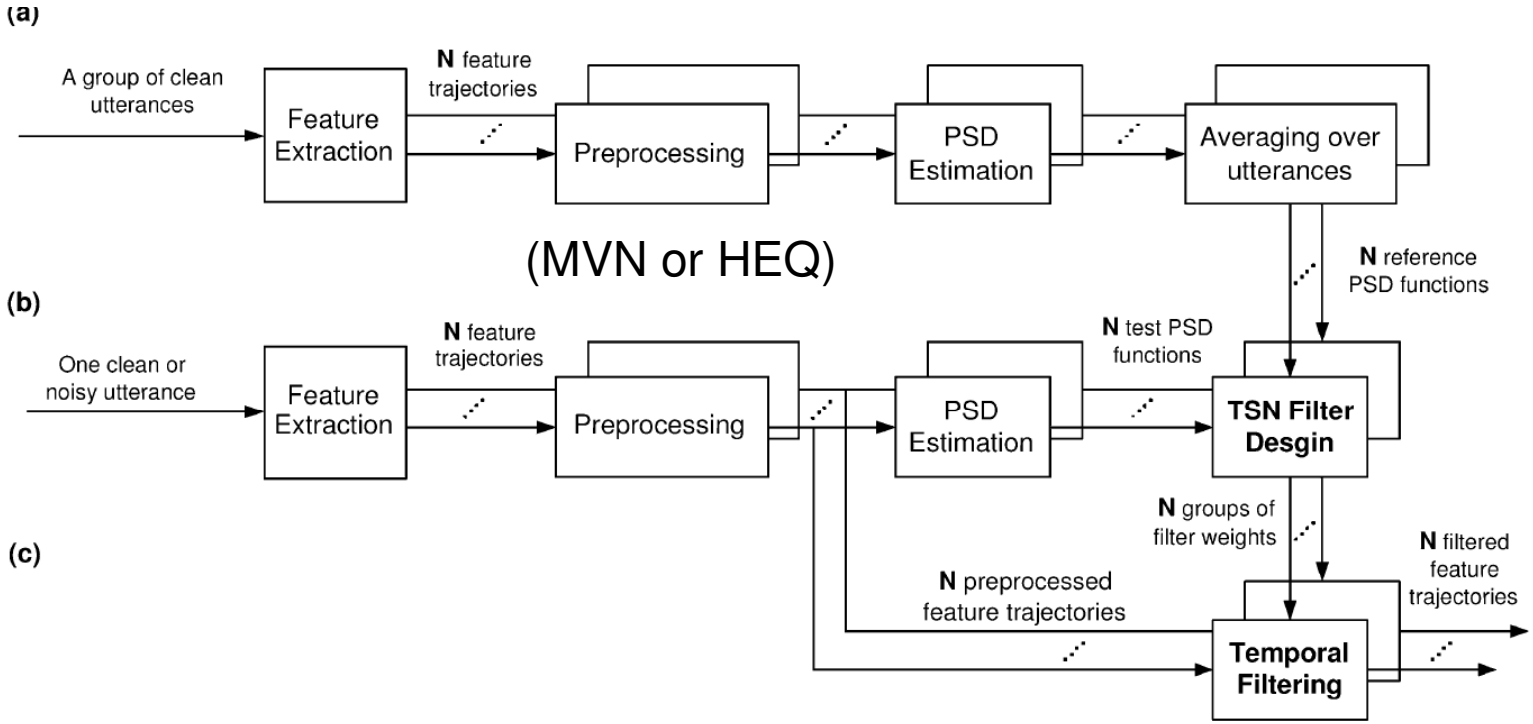


Fig. 4. Illustration of the TSN framework. The framework is divided into three steps. (a) The offline training of the reference PSD functions. (b) The designing of the TSN filters for current utterance, one filter for each feature trajectory. (c) The filtering of the feature trajectories.

TSN filter design

- Magnitude response of TSN filter:

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

- The FIR filter coefficients can be found from IDFT of the magnitude response.
- Combination with other temporal filter (optional)

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

- The normalized PSD:

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

TABLE I

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). $w(\tau, j) = \text{IDFT}(|H(k, j)|)$.

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.

- Using a shorter filter can avoid extreme normalization (step 5 and 6 in table 1)
- The weights of FIR is symmetric and the most significant weights are concentrated around the center.
- A Hanning window is applied to the truncated weights to reduce truncation effect.
- Finally, the sum of the windowed weights are normalized to one to ensure that features are properly scaled.

Segment based implementation

- The TSN filter is implemented utterance-by-utterance, which leads to a large processing delay.
- Dividing the utterances into multiple overlapping and equal-sized segments can reduce the processing delay.

Segment based implementation

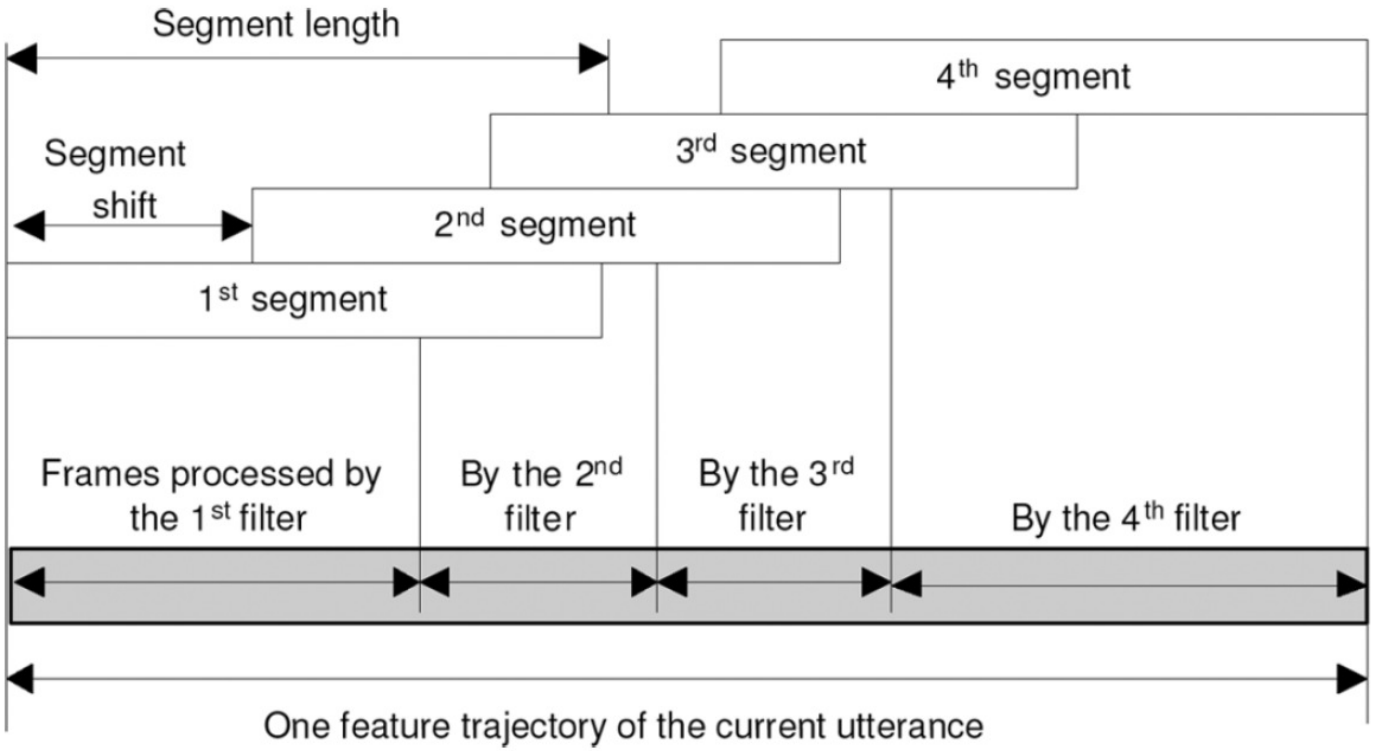


Fig. 5. Processing scheme of the segment-based implementation of the TSN framework.

Experiments

- Aurora2

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 filter order

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

- Aurora4

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

- Different features on Aurora2

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

Discussion

- Aurora-4 has longer utterances so that the histograms of the features are better estimated for HEQ.
- While feature smoothing can reduce the feature mismatches, it also leads to removal of useful speech information.
- Small-vocabulary Aurora-2 task can be aggressively smoothed (ARMA order = 3) while the large-vocabulary Aurora-4 task can only be mildly smoothed (ARMA order = 1).

Experiments of segment-based implementation

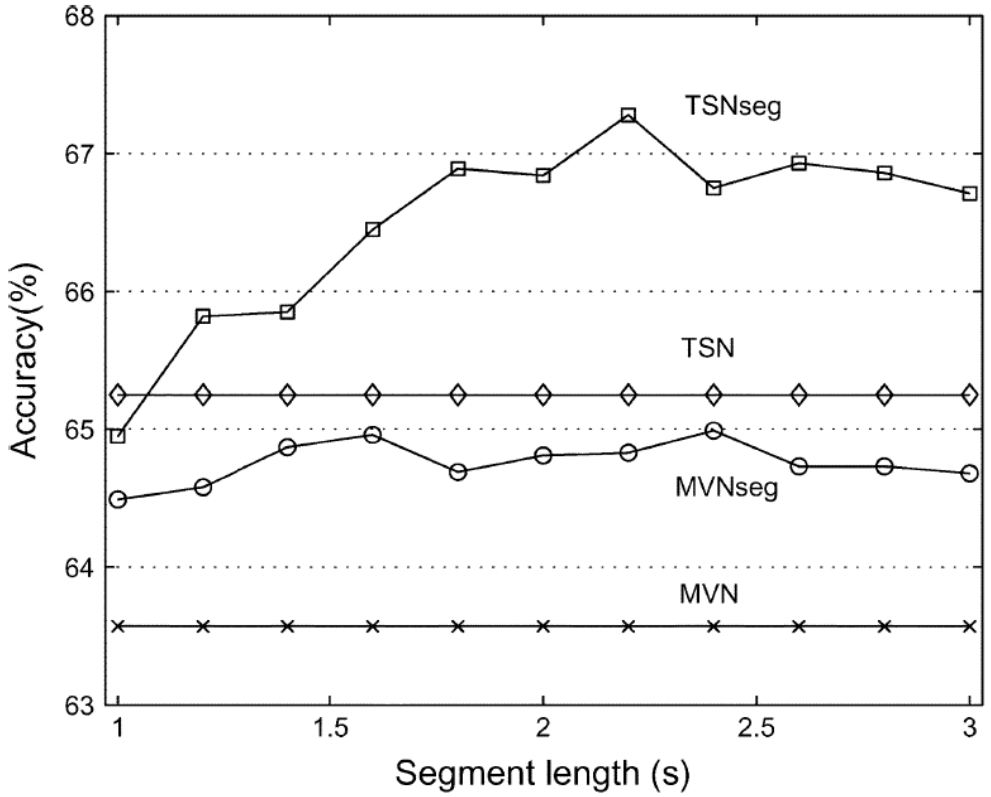


Fig. 6. Performance of the segment-based TSN with different segment length on the Aurora-4 task. The features are the MFCC with c0 energy plus their delta and acceleration features.

Filter length

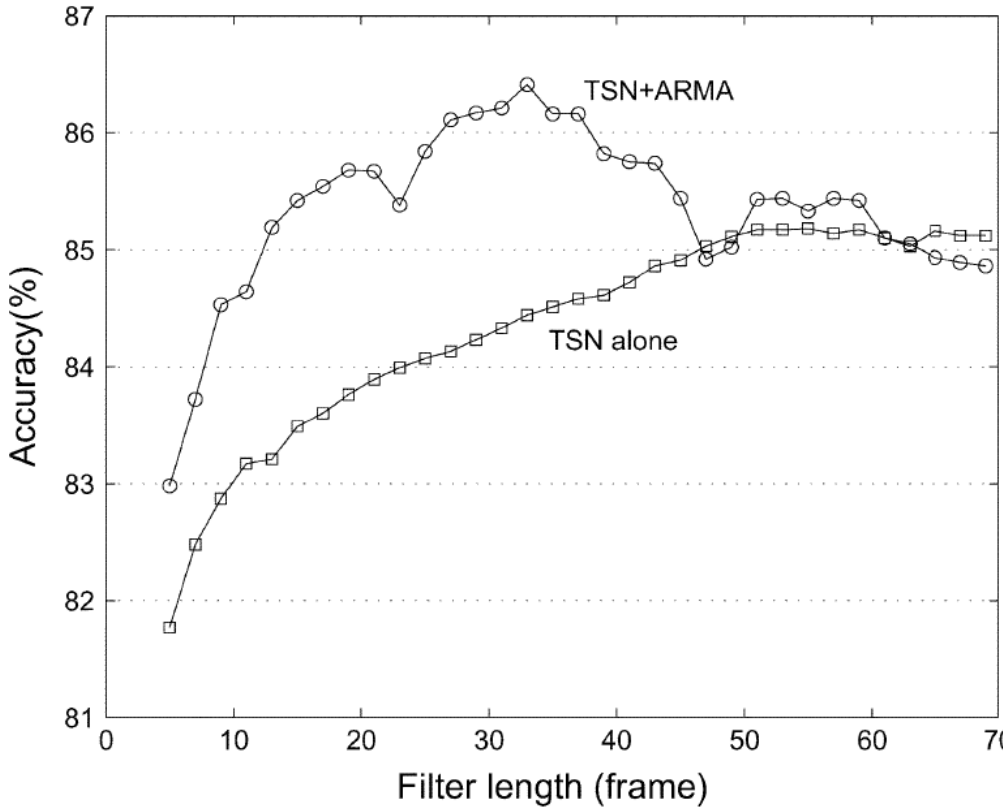


Fig. 7. Performance of two TSN configurations with different filter length. The features are the MFCC with c0 energy plus their delta and acceleration features the preprocessing is MVN, and the task is Aurora-2.

Conclusion

- The major advantage of the TSN filter is its ability to adapt to changing environments.
- The magnitude response is designed to reflect the temporal characteristics of the feature trajectories that carry information about the signal distortion.