A new perceptually motivated MVDR-based acoustic speech recognition

Speech communication 50(2008) 142-152

Author : Umit H. Yapanel, John H.L. Hansen*

Professor:陳嘉平

Reporter:葉佳璋

Outline

- Introduction
- Minimum variance distortionless response (MVDR)
- Previous MVDR-based acoustic front-ends
- Direct warping of the FFT spectrum
- Experiments

Introduction

- The most crucial information need for ASR is a representation of the vocal tract transfer function(VTTF).
- The VTTF is mainly encoded in the short-term spectral envelope and extracting the short-term spectral envelope accurately and robust is important.
- Moreover, incorporating perceptual considerations, such as Mel and Bark scales, into the acoustic front-end leads to improved accuracy.

Introduction

- This paper proposes a new acoustic front-end based on the Minimum variance distortionless response (MVDR) spectrum estimation method.

- The perceptual scales were integrated through the used of a non-linearly spaced filterbank, in the PMVDR front-end, on the other hand, this step is eliminated by directly warping the FFT power spectrum.

Minimum variance distortionless response (MVDR)

- The MVDR spectrum is a good way of performing all-pole modeling on the speech spectrum.
- Unlike FFT analysis where fixed bandpass filter are used regardless of the characteristics of the incoming signal.
- MVDR obtains the power spectrum estimates by using data-dependent bandpass filters.

Minimum variance distortionless response (MVDR)

- The signal power at a frequency ω_l is determined by filtering the signal by a specially FIR filter $h(n)$ and measuring the power at its output.
- $h(n)$ is designed to minimize its output power subject to the constraint that its response at the frequency of interest ω_l has unity gain :

$$H(e^{j\omega_l}) = \sum_{k=0}^{M} h(k) e^{-j\omega_l k} = 1$$

- This is the *distortionless constraint*

Minimum variance distortionless response (MVDR)

- The distortionless filter $h(n)$ is obtained by solving the following constrained optimization problem

$$\min_{h} h^H R_{M+1} h \text{ subject to } v^H(\omega) h = 1$$

$$\text{where } h = [h_0, h_1, \ldots h_M]^H \;, v(\omega) = [1, e^{j\omega}, e^{j2\omega}, \ldots, e^{jM\omega}],$$

$$R_{M+1} \text{ is the } (M+1) \times (M+1) \text{ Toeplitz autocorrelation matrix of the data}$$

- The solution is $$h_l = \frac{R_{M+1}^{-1} v(\omega_l)}{v^H(\omega_l) R_{M+1}^{-1} v(\omega_l)}$$

Minimum variance distortionless response (MVDR)

$$P_{MV}(\omega_l) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H_l(e^{j\omega})\right|^2 S_{xx}(e^{j\omega})d\omega$$

$$= \frac{1}{v^H(\omega_l)R_{M+1}^{-1}v(\omega_l)}$$

$$= \frac{1}{\sum_{k=-M}^{M}\mu(k)e^{-j\omega_l k}} = \frac{1}{\left|B(e^{j\omega})\right|^2}$$

$$\mu(k) = \begin{cases} \frac{1}{P_e}\sum_{i=0}^{M-k}(M+1-k-2i)\times a_i a_{i+k}^* & , \text{for } k = 0,\ldots,M \\ \mu^*(-k) & , \text{for } k = -M,\ldots,-1 \end{cases}$$

a_i is the LP coefficient, P_e is the prediction error variance

Previous MVDR-based acoustic front-ends

- Several studies have considered incorporating the merits of MVDR spectrum into the speech recognition framework.
- The first use of MVDR in speech parameterization was for power spectrum estimation.
- The FFT spectrum in the MFCC computation method was simply replaced by a high-order MVDR spectrum computation method.
- The remainder of the feature extraction algorithm was the same as the MFCC, therefore called MVDR-MFCCs.

Previous MVDR-based acoustic front-ends

- A second study employing MVDR methodology for feature extraction.
- The features developed is called Perceptual MVDR-based cepstral coefficients(PMCCs).
- In the PMCC front-end, the MVDR methodology is used for spectral envelope extraction rather than for spectrum estimation.
- The implementation of PMCCs is very similar to PLP in that they both represent the spectral envelope using an all-pole model.

Description of PMVDR

- Different from the earlier approaches, PMVDR front-end completely removes the filterbank processing step and directly performs warping on the FFT power spectrum.

- The main aim of the filterbank is to average out the harmonic information(i.e. , the pitch) that exists in the FFT spectrum and to track the spectral envelope.

Direct warping of the FFT spectrum

- Since the filters are spaced closely at low frequencies , the effectiveness of filterbandk in smoothing the pitch information is significantly reduced for high-pitch speakers.

- PMVDR is an appropriate spectral envelope modeling approach for a broad range of speech phoneme classes, especially for high-pitched speech.

Direct warping of the FFT spectrum

- One way of incorporating perceptual consideration is to implement through a first order all-pass system.
- This approach is simple and feasible for our purpose. In fact, both Mel and Bark scale are determined by changing the single parameter α of the system.
- The form, H(z), and the phase response, $\hat{\omega}$, of the first order system are given as,

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}}$$

$$\omega = \tan^{-1}\left\{\frac{(1-\alpha)^2 \sin(\hat{\omega})}{(1+\alpha)^2 \cos(\hat{\omega}) - 2\alpha}\right\}$$

Direct warping of the FFT spectrum

- Warping via interpolation is a simple and fast method to implement directly warping.
- A step-by-step algorithm that describes how warping can be efficiently implemented via interpolation can be given as follows

➢ 1. Take the FFT of the input speech frame of length N to obtain the FFT power spectrum. N should be selected as the nearest possible power of 2, thus providing N spetral points(i.e. S[k], k=0, …, N-1) in linear power spectrum space.

Direct warping of the FFT spectrum

- 2. Calculate N linearly spaced spectral points over the warped frequency space by dividing the entire 2π warped frequency range into N equi-spaced points,

$$\hat{\omega}[i] = 2i\pi / N, \qquad i = 0, \ldots, N-1$$

- 3. Compute the linear frequencies and FFT indexes that corresponds to these warped frequencies using

$$\omega[i] = \tan^{-1}\left\{ \frac{(1-\alpha)^2 \sin(\hat{\omega}[i])}{(1+\alpha)^2 \cos(\hat{\omega}[i]) + 2\alpha} \right\}, \qquad i = 0, \ldots, N-1,$$

$$\hat{k}[i] = \omega[i]N / 2\pi, \qquad i = 0, \ldots, N-1$$

Direct warping of the FFT spectrum

- 4. For the final step, perform an interpolation of the nearest linear spectral values to obtain the warped spectral value

$$k_l[i] = \min(N-2, \hat{k}[i]), \qquad i = 0, \ldots, N-1$$

$$k_u[i] = \max(1, k_l[i]+1), \qquad i = 0, \ldots, N-1$$

$$\hat{S}[i] = (k_u[i] - \hat{k}[i])\,S[k_l[i]] + (\hat{k}[i] - k_l[i])\,S[k_u[i]]$$

where $k_l[i]$ is the lower nearest linear FFT bin, $k_u[i]$ is the nearest upper linear FFT bin and $\hat{S}[i]$ is the value of the warped power spectrum that corresponds to FFT bin i.

PMDR algorithm

- The proposed PMVDR algorithm can be summarized as follows:

1. Obtain the perceptually warped FFT power spectrum via interpolation.
2. Compute the perceptual autocorrelation lags by taking the IFFT of the perceptually warped power spectrum.
3. Perform an Mth order LP anlysis via Levinson-Durbin recursion using perceptual autocorrelation lags.
4. Calculate the Mth order MVDR spectrum.

PMDR algorithm

5. After obtaining the MVDR coefficients from the perceptually warped spectrum , we take the FFT of the parametrically expressible MVDR spectrum. After taking log, we apply IFFT to return back to the cepstral domain.
6. Take the first N, generally 12 excluding 0^{th} cepstrum, cepstral coefficients as the output of the PMVDR front-end. This is the cepstral truncation step.

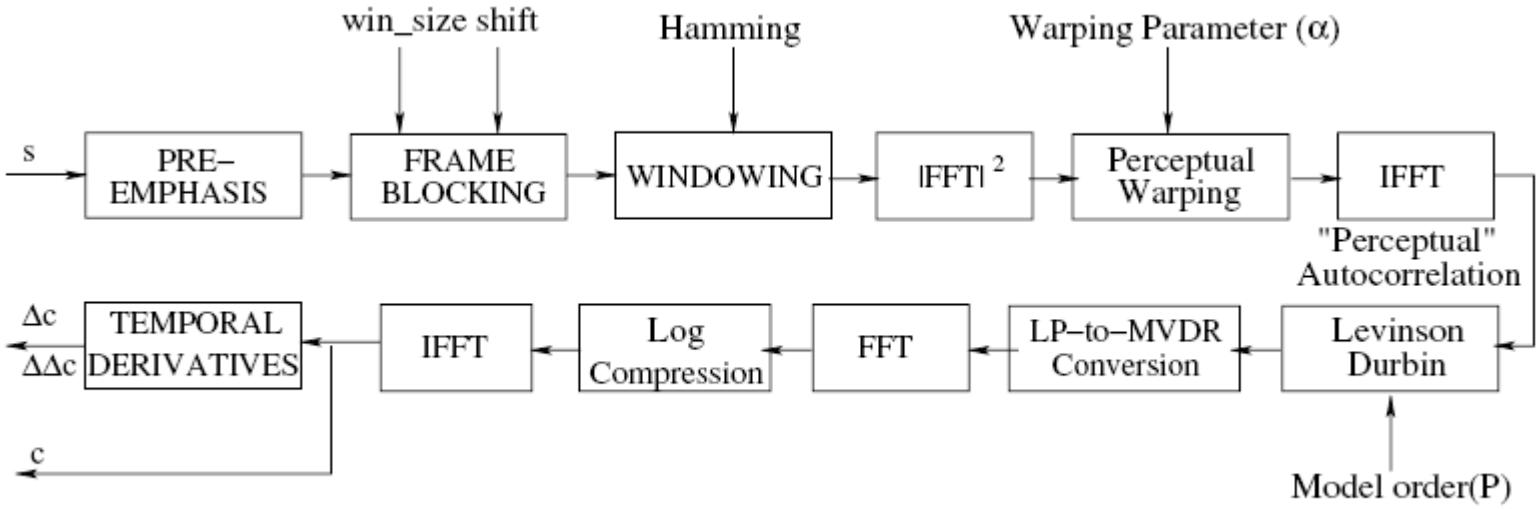

Fig. 3. Schematic diagram of PMVDR front-end.

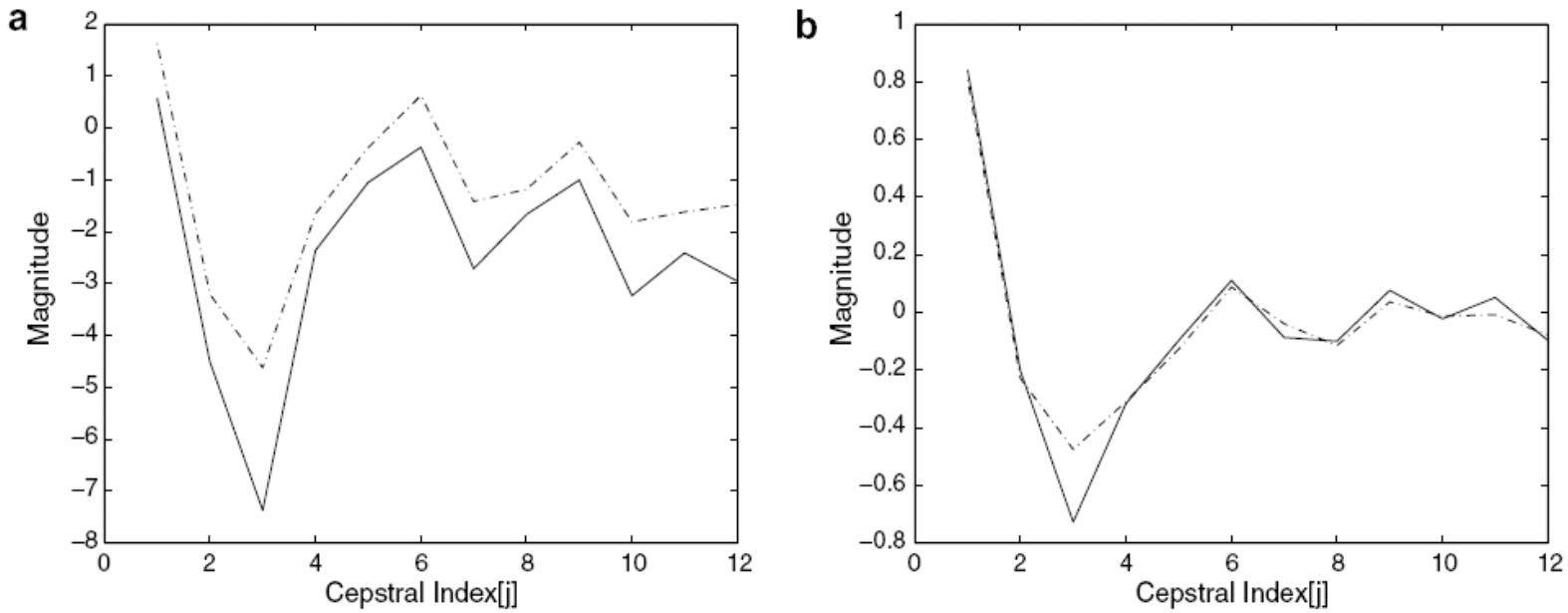

Fig. 4. Cepstrum values for a clean (solid) and 5 dB car-noise degraded (dash-dotted) voiced sound frame from /AA/ (a) variation of MFCCs, (b) variation of PMVDRs.

Experiments

- For noisy speech experiments, we use the CU-Move Extended Digits Corpus Database(Cu-Move 2004)
- There are 5 parts in the database

  command and control words.

  digit strings being mostly phone number.

  street addresses with mostly spelling .

  phonetically balanced sentence.

  wizard of Oz interactive navigation.
- A total of 500 speakers produced over 600GB of data

Table 1
WERs (%) for CU-Move task with different front-ends

Gender/Systems	MFCC	PMCC	PMVDR	Rel. Imp.
Female	9.16	7.85	5.47	40.3
Male	13.22	12.03	10.16	23.1
Overall	11.12	9.87	7.74	30.4

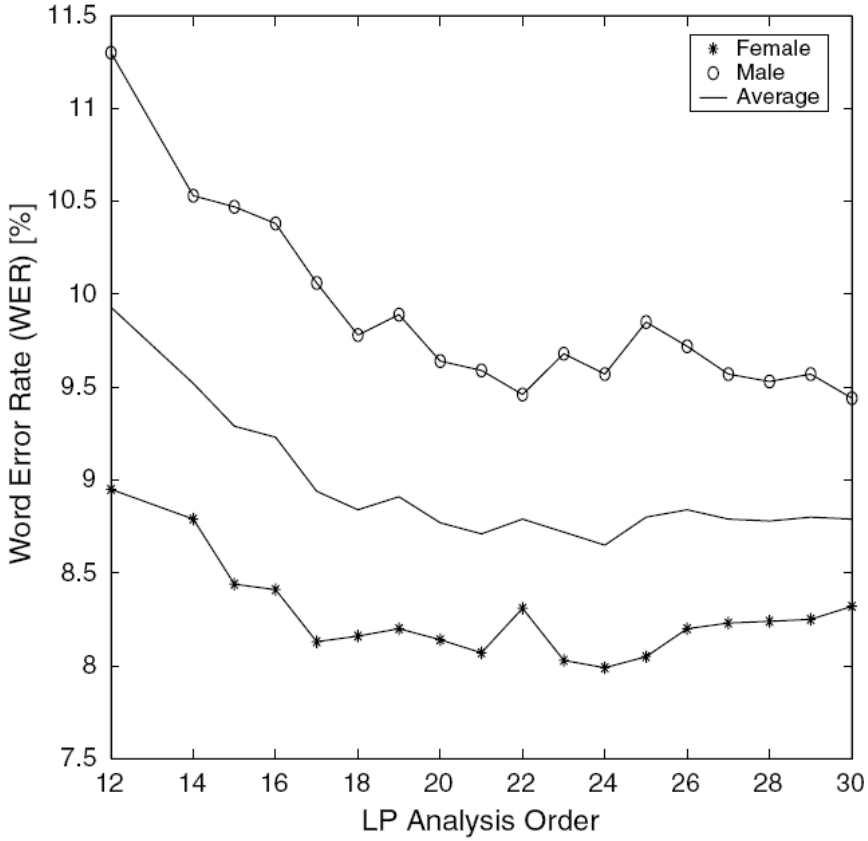

Fig. 5. Variation of WER (%) with LP analysis order, females (∗), males (○) and overall (solid).

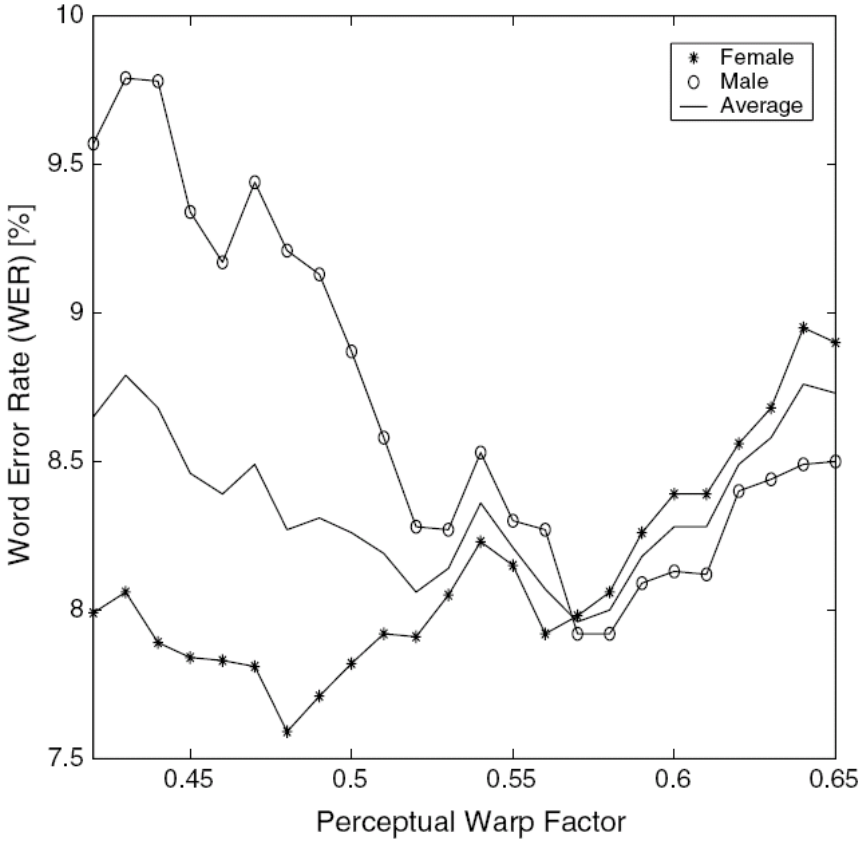

Fig. 6. Variation of WER (%) with the perceptual warp parameter (α), females (dashed), males (dash-dotted) and overall (solid).

Table 2
WERs (%) for CU-Move task with PMVDR Optimized settings

Gender/Systems	MFCC	PMCC	PMVDR	Rel. Imp.
Female	9.16	7.85	5.57	39.2
Male	13.22	12.03	8.76	33.7
Overall	11.12	9.87	7.11	36.1

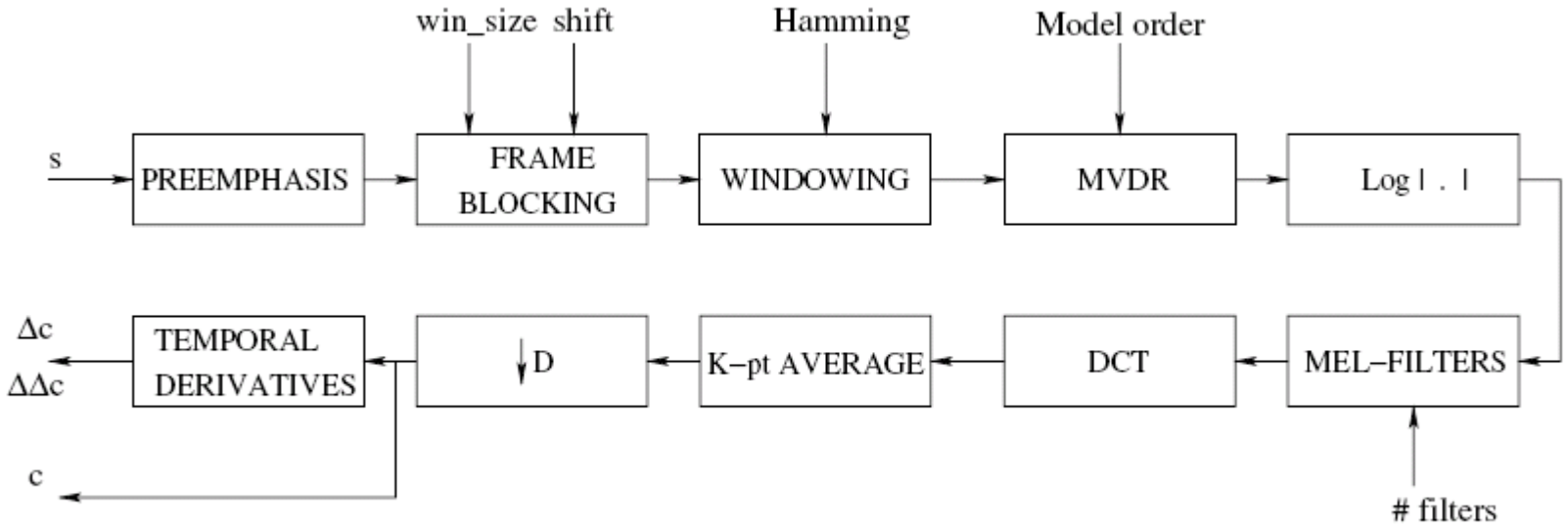

Fig. 1. Flow diagram of the MVDR–MFCC front-end.

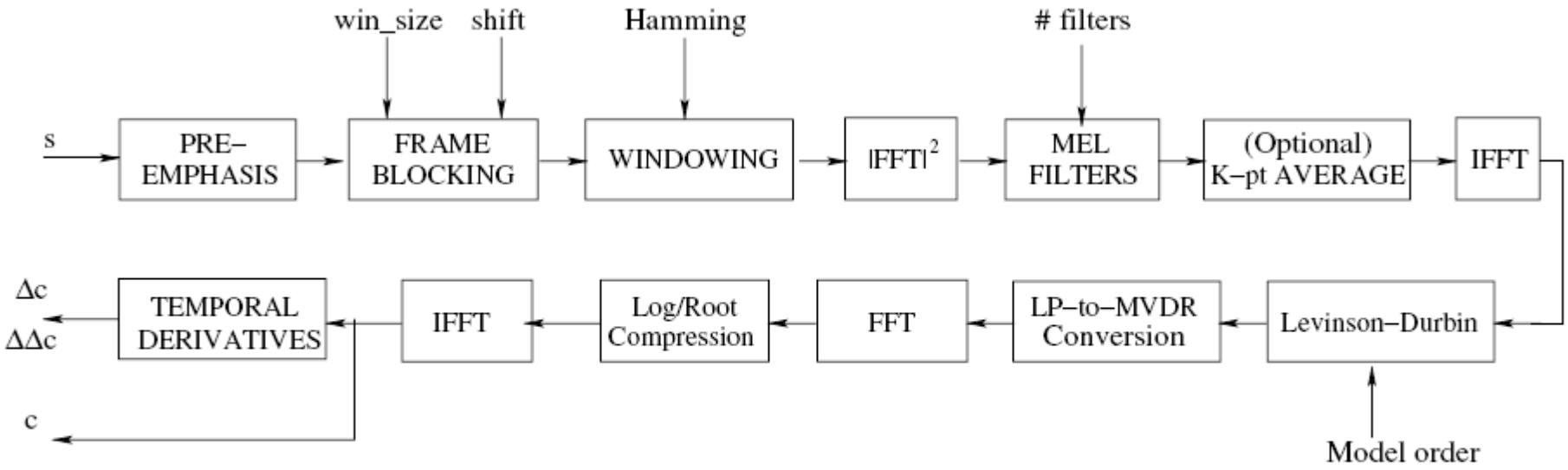

Fig. 2. Flow diagram of the PMCC front-end.