Disclaimer

- The material provided in this document is not my original work and is a summary of some one else’s work(s).
- A simple Google search of the title of the document will direct you to the original source of the material.
- I do not guarantee the accuracy, completeness, timeliness, validity, non-omission, merchantability or fitness of the contents of this document for any particular purpose.
- Downloaded from najeebkhan.github.io

CEPSTRAL ANALYSIS SYNTHESIS ON THE MEL FREQUENCY SCALE

Presented by

Najeeb Khan

2014-5-20

Abstract

- The log spectrum on the Mel frequency scale is considered to be an effective representation of the spectral envelope of speech
- This analysis synthesis system uses the Mel log spectrum approximation (MLSA) filter which was devised for the cepstral synthesis on the Mel frequency scale
- The filter coefficients are easily obtained through a simple linear transform from the Mel cepstrum

Abstract (contd…)

- The MLSA filter has
  - Low coefficient sensitivity
  - Good coefficient quantization characteristics
  - Spectral distortion due to interpolation is small
  - Same quality speech is synthesized at 60-70 % of data rates in the conventional cepstral vocoder or the LPC vocoder

Introduction(1)

- The log spectrum is considered to be a reasonable representation of the spectral envelope of speech
- The cepstrum has good characteristics for parametric representation of speech, since it is defined as a Fourier transform of the log spectrum
- The log spectrum is efficiently approximated by the LMA filter from the cepstral parameter
- LMA filter is of pole-zero, and it is an accurate and efficient model for the log spectral envelope of speech

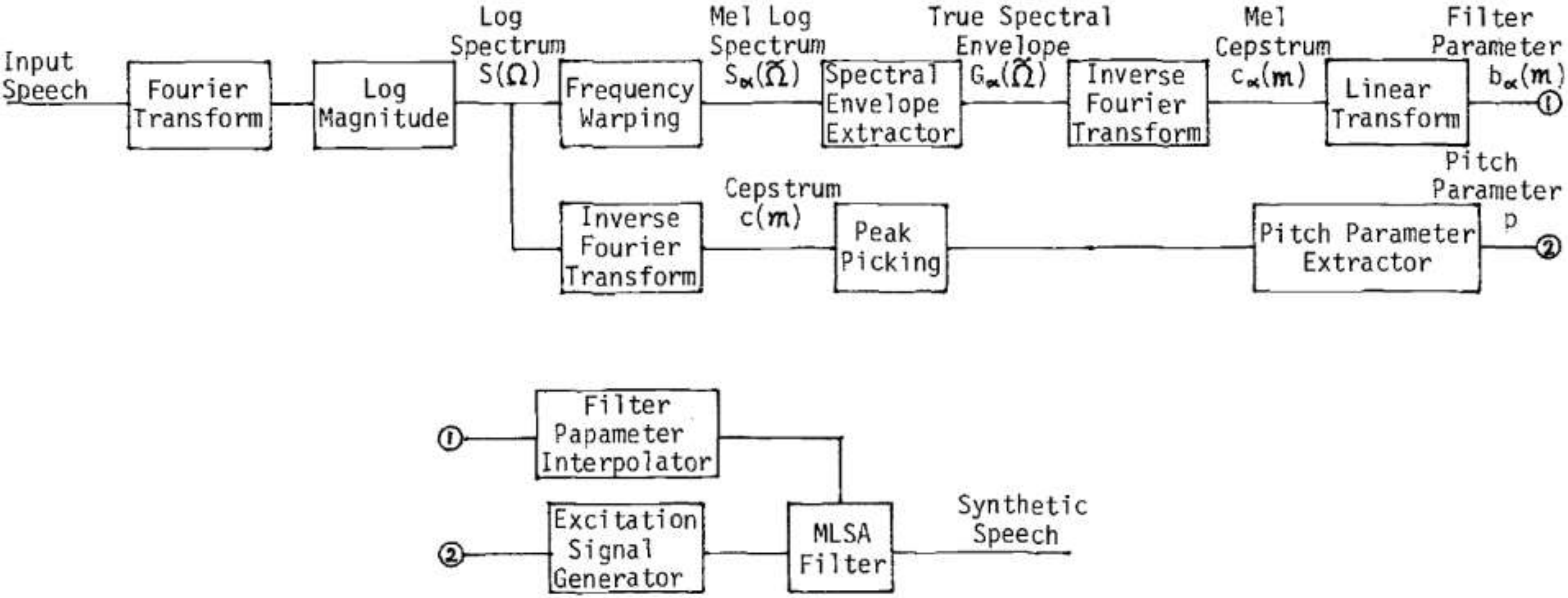
Introduction(2)

- The cepstrum has the following good features as the spectral envelope parameter
  - It is considered to be a good parameter of a pole-zero model which represents the log spectral envelope of the speech accurately and efficiently
  - The LMA filter can be used for a high quality speech direct synthesis from the cepstral parameter
  - The cepstral parameter sensitivity of the log spectrum is very small and the cepstrum quantization effect is also small
  - Spectral distortion caused by interpolation of the cepstral parameters of two successive frames is small

Introduction(3)

- Although the cepstrum has many good features, it is not always an efficient parameter for speech analysis synthesis
- The order of the cepstrum is larger than that of the LPC parameter for a high quality speech analysis synthesis
- The log spectrum on a Mel frequency scale is considered to be a more effective representation of the spectral envelope of speech than that on the linear frequency scale
- The Mel cepstrum has a comparatively low order hence it is an efficient parameter

Mel Cepstral Analysis Synthesis System



Input Speech
Fourier Transform
Log Magnitude
Log Spectrum $S(\Omega)$
Frequency Warping
Mel Log Spectrum $S_\alpha(\tilde{\Omega})$
Spectral Envelope Extractor
True Spectral Envelope $G_\alpha(\tilde{\Omega})$
Inverse Fourier Transform
Mel Cepstrum $c_\alpha(m)$
Linear Transform
Filter Parameter $b_\alpha(m)$
①
Inverse Fourier Transform
Cepstrum $c(m)$
Peak Picking
Pitch Parameter Extractor
Pitch Parameter p
②
①
Filter Papameter Interpolator
②
Excitation Signal Generator
MLSA Filter
Synthetic Speech

Spectral Envelope Extraction by Improved Spectral Method

- The Mel scale can be approximated by the phase characteristics of a first order all-pass filter

$$H^{(\alpha)}(z) = (z^{-1} - \alpha)/(1 - \alpha z^{-1})$$

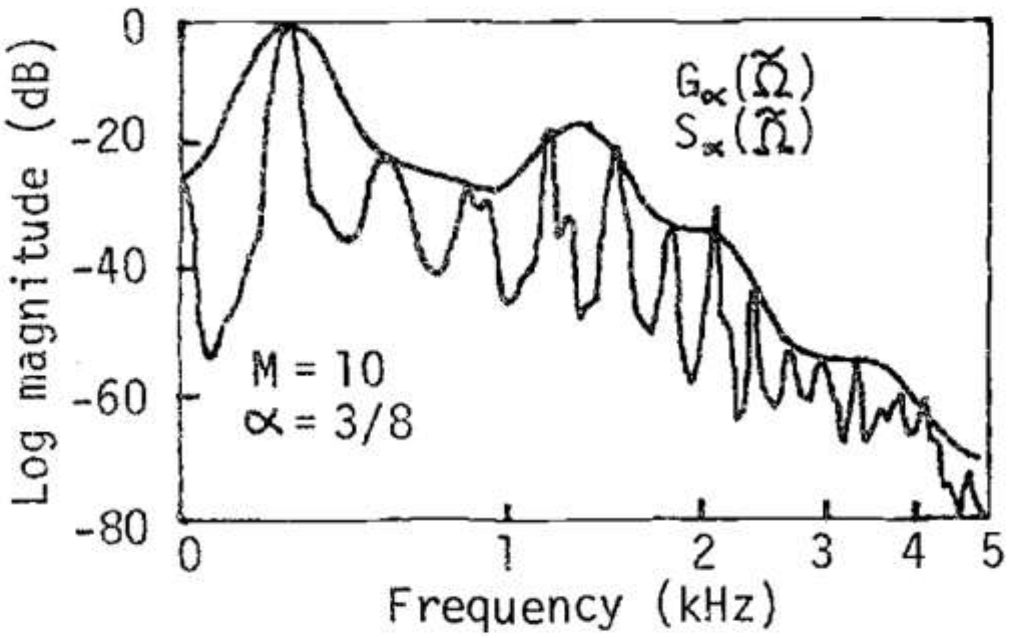
$$\beta_\alpha(\Omega) = -\arg H^{(\alpha)}(e^{j\Omega}) = \tan^{-1}\frac{(1-\alpha^2)\sin\Omega}{(1+\alpha^2)\cos\Omega - 2\alpha}.$$

$$\tilde{\Omega} = \beta_\alpha(\Omega)$$

Spectral Envelope Extraction by Improved Spectral Method

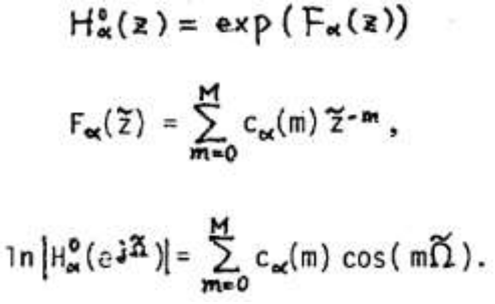
- The true envelope of the Mel log spectrum is

$$G_\alpha(\tilde{\Omega}) = \sum_{m=0}^{M} c_\alpha(m)\cos(m\tilde{\Omega})$$



Mel Log Spectrum Approximation (MLSA) Filter

- The transfer function of the MLSA filter is given by a rational function of the first order all-pass transfer function, it is necessary not to contain any delay free path in the feedback loop of the filter
- The MLSA filter is given by a transfer function approximating the exponential of a transfer function of a basic filter



$$H^{0}_{\alpha}(z) = \exp(F_{\alpha}(z))$$

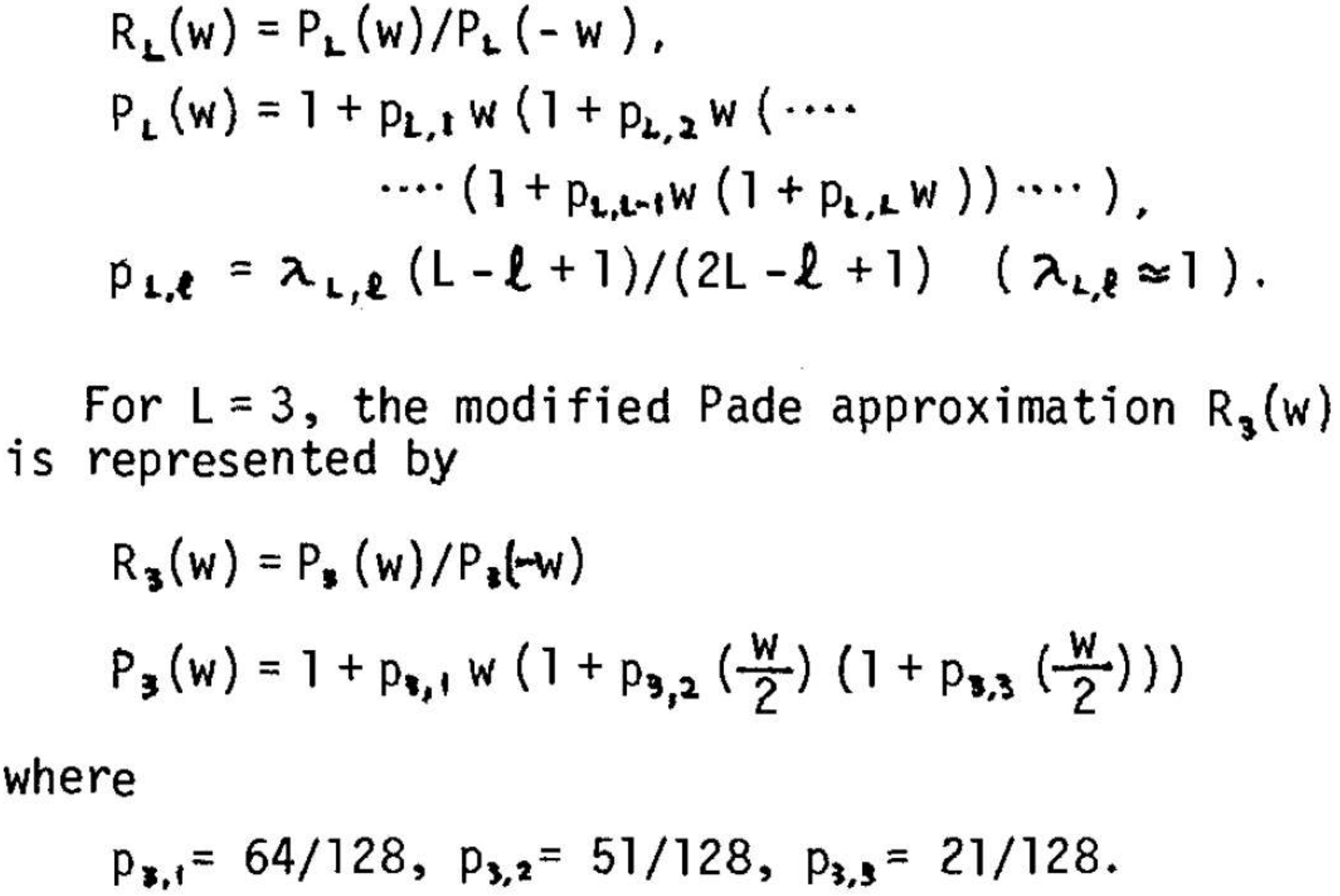
$$F_{\alpha}(\tilde{z}) = \sum_{m=0}^{M} c_{\alpha}(m)\,\tilde{z}^{-m},$$

$$\ln\left|H^{0}_{\alpha}(e^{j\tilde{\Omega}})\right| = \sum_{m=0}^{M} c_{\alpha}(m)\cos(m\tilde{\Omega}).$$

Mel Log Spectrum Approximation (MLSA) Filter

- If the filter parameter C(m) is chosen as the mel cepstrum for the spectral envelope, the log magnitude on the mel frequency scale is identical to the mel log spectral envelope
- The exponentional is approximated by the pade approximant

Mel Log Spectrum Approximation (MLSA) Filter



$$R_L(w) = P_L(w)/P_L(-w),$$

$$P_L(w) = 1 + p_{L,1} w (1 + p_{L,2} w (\cdots$$

$$\cdots (1 + p_{L,L-1} w (1 + p_{L,L} w)) \cdots),$$

$$p_{L,\ell} = \lambda_{L,\ell} (L - \ell + 1)/(2L - \ell + 1) \quad (\lambda_{L,\ell} \simeq 1).$$

For L = 3, the modified Pade approximation $R_3(w)$ is represented by

$$R_3(w) = P_3(w)/P_3(-w)$$

$$P_3(w) = 1 + p_{3,1} w (1 + p_{3,2} (\frac{w}{2}) (1 + p_{3,3} (\frac{w}{2})))$$

where

$$p_{3,1} = 64/128, \; p_{3,2} = 51/128, \; p_{3,3} = 21/128.$$

Mel Log Spectrum Approximation (MLSA) Filter

$$F_{\alpha}(\tilde{z}) = F(z) = b_{\alpha}(0) + z^{-1}\sum_{m=1}^{M+1} b_{\alpha}(m)\,\tilde{z}^{-(m-1)}$$

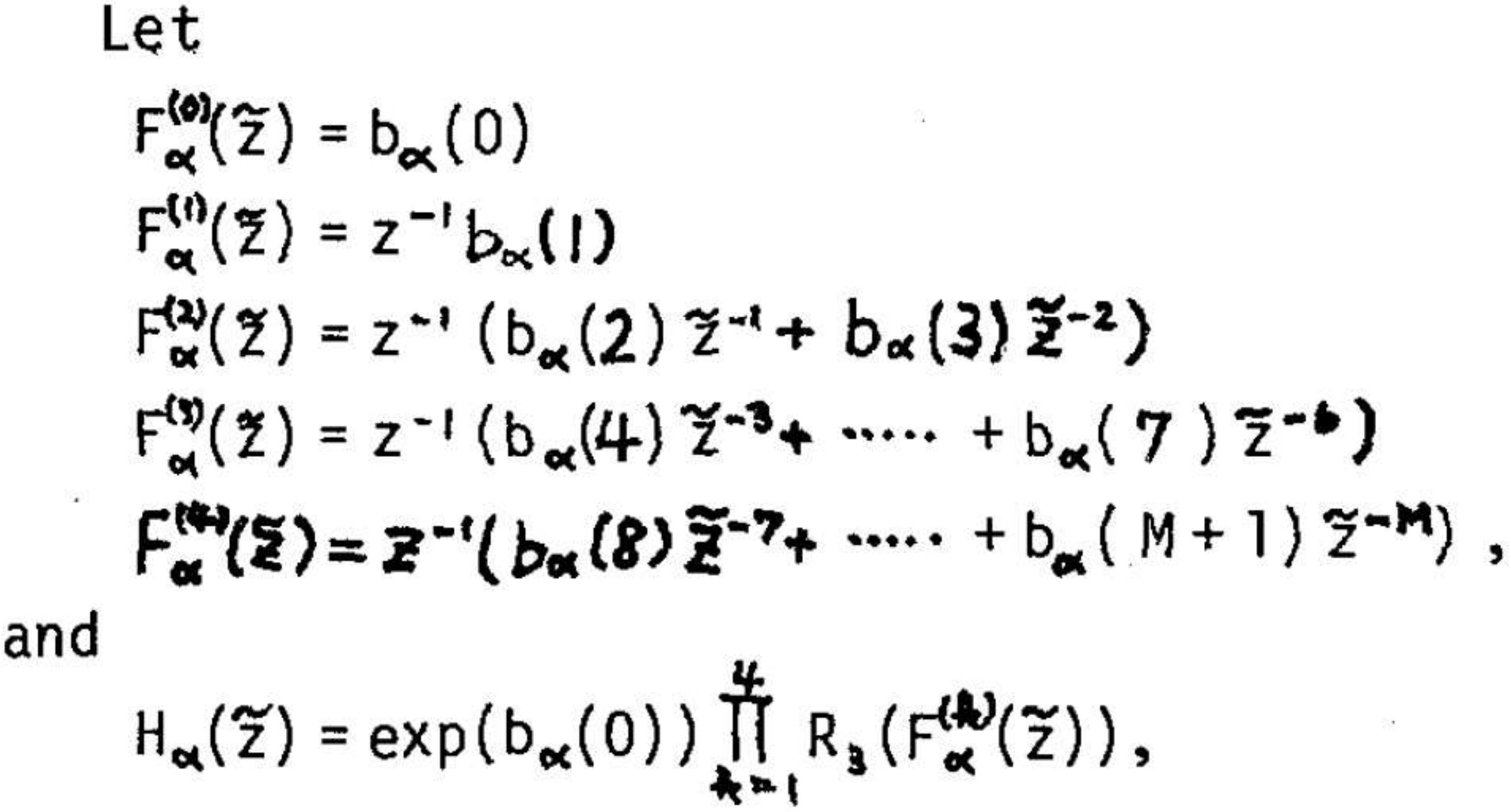
$$b_{\alpha}(M+1) = \alpha c_{\alpha}(M)$$

$$b_{\alpha}(m) = c_{\alpha}(m) + \alpha(c_{\alpha}(m-1) - b_{\alpha}(m+1))$$
$$(m = M,\ M-1,\ \cdots\cdots,\ 3,\ 2)$$

$$b_{\alpha}(1) = (c_{\alpha}(1) - \alpha b_{\alpha}(2))/(1-\alpha^{2})$$

$$b_{\alpha}(0) = c_{\alpha}(0) - \alpha b_{\alpha}(1).$$

Mel Log Spectrum Approximation (MLSA) Filter



Let

$$F_{\alpha}^{(0)}(\tilde{z}) = b_{\alpha}(0)$$

$$F_{\alpha}^{(1)}(\tilde{z}) = z^{-1} b_{\alpha}(1)$$

$$F_{\alpha}^{(2)}(\tilde{z}) = z^{-1} (b_{\alpha}(2)\,\tilde{z}^{-1} + b_{\alpha}(3)\,\tilde{z}^{-2})$$

$$F_{\alpha}^{(3)}(\tilde{z}) = z^{-1} (b_{\alpha}(4)\,\tilde{z}^{-3} + \cdots\cdots + b_{\alpha}(7)\,\tilde{z}^{-6})$$

$$F_{\alpha}^{(4)}(\tilde{z}) = z^{-1} (b_{\alpha}(8)\,\tilde{z}^{-7} + \cdots\cdots + b_{\alpha}(M+1)\,\tilde{z}^{-M}) ,$$

and

$$H_{\alpha}(\tilde{z}) = \exp(b_{\alpha}(0)) \prod_{k=1}^{4} R_3(F_{\alpha}^{(k)}(\tilde{z})) ,$$

Voiced-Unvoiced Decision

- When the averaged value of the spectral envelope in a fundamental frequency region (50—350 Hz) exceeds a threshold, the sound is voiced
- The voiced-to-unvoiced and unvoiced-to-voiced error rates are 1-2 % and 2-4 %, respectively

Speech Quality

- The frequency warping factor α is fixed at 0.375
- For T= 15 ms, M= 11, q = 0.25 and b = 7 bit, the speech quality is very high
- The synthesized speech is indistinguishable from the sound synthesized by a linear frequency cepstral vocoder using the 25th order cepstral parameter