CEPSTRAL ANALYSIS SYNTHESIS ON THE MEL FREQUENCY SCALE, AND AN ADAPTATIVE ALGORITHM FOR IT.

Summarized overview of the IEEE-publicated papers
"Cepstral analysis synthesis on the mel frequency scale" by Satochi IMAI (Japan, 1983),
and "An adaptative algorithm for mel-cepstral analysis of speech" by Toshiaki FUKADA,
Keiichi TOKUDA, Takao KOBAYASHI and Satoshi IMAI (Japan, 1992).

Cecilia CARUNCHO LLAGUNO, Graz (Austria), April 2008.

What is cepstral analysis?

Ceptral analysis is a modelation of speech based on the use of cepstrum, which is defined as the inverse Fourier transform of the logarithm of the Fourier transform module.

$$\text{cepstrum of signal} = \mathrm{F}\{ \log [\mathrm{F}^{-1}(\text{signal}) + j \cdot 2\pi \cdot m]\}$$

It has therefore good characteristics for parametric representation of speech. Some other interesting features are:

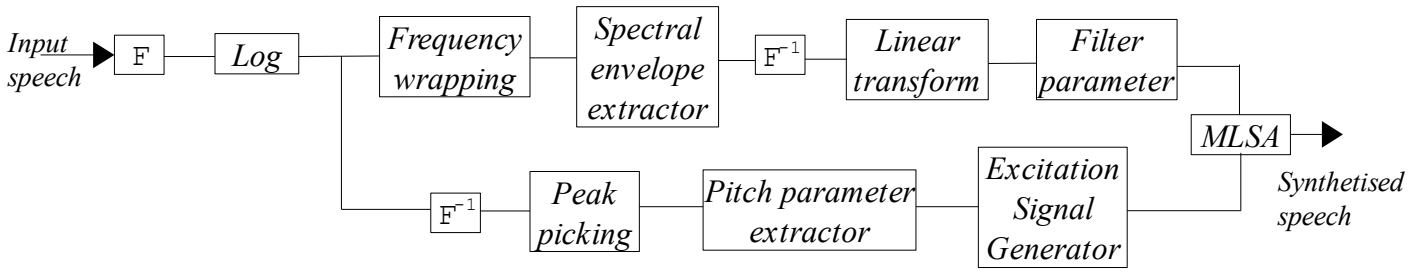
1. It represents the log spectral envelope of speech accurately and efficiently,
2. the LMA filter can be used for a high quality speech direct synthesis,
3. both the sensitivity and the cepstrum quantization noise are small, and
4. the spectral distortion is also small.

Note that it will be better to use Mel frequency scale rather than linear frequency scale, since the representation of the spectral envelope of speech is more effective.

What is the Mel scale?

Psychophysical studies have shown that human perception of the frequency content of sounds, either for pure tones or for speech signals, does not follow a linear scale. This research has led to the idea of defining subjective pitch of pure tones. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called "Mel" scale. As a reference point, the pitch of a 1kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Other subjective pitch values are obtained by adjusting the frequency of a tone such that it is half or twice the perceived pitch of a reference tone (with a known mel frequency).

Mel cepstral analysis synthesis system.



Note that the MLSA filter is used as the speech synthetizer. The filter coefficients are obtained through a simple linear transform from the mel cepstrum which is defined as the Fourier cosine coefficients of the "true" spectral envelope of the mel log spectrum.

Spectral envelope extraction

The mel scale can be approximated by the phase characteristics of a first order all-pass filter, with phase $\beta(\Omega)$:

$$\beta(\Omega)=\tan^{-1}\frac{(1-\alpha^2)\sin\Omega}{(1+\alpha^2)\cos\Omega-2\alpha}$$

And the mel frequency scale ω is represented by $\omega=\beta(\Omega)$. Thus the mel log spectrum can be defined as a function of ω. We will assume that this is a smooth function represented by a trigonometric polimomial of order M:

$$G(\omega)=\sum_{m=0}^{M} c(m)\cos(m\omega)$$

It is also very important to extract the spectral envelope properly. The cepstral method estimates a spectral envelope minimizing the mean square error, so that poles and zeros can be expressed with te same accuracy. The present cepstral method can automatically and stably extract the spectral envelope satisfactory without being affected by the fine structure in either voiced or invoiced sounds.

Mel log spectrum approximation

In order to optimize the quality of the synthetized speech, , the mean spare approximation to the desired log spectral envelope must be minimum. The best results are achieved with the MLSA filter, whose coefficients are obtained by a simple linear transform from the mel spectrum. This filter has two main advantages: i) very low coefficient sensitivities and ii) good coefficient quantization characteristics.

The ideal form of the MLSA filter's transfer function is $H_\alpha^o(z) = e^{F_\alpha(z)}$, where $F_\alpha(z)$ is the so-called basic filter.

If the basic flilter is stable, then the MLSA will be stable and of minimum phase.

$$F_\alpha(z) = \sum_{m=0}^{M} c(m) z^{-m} \rightarrow \ln\left|H_\alpha^o(e^{j\omega})\right| = \sum_{m=0}^{M} c(m)\cos(m\omega) \qquad \textit{note that } \omega \textit{ is the frequency on mel scale}$$

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.

However, the ideal MLSA filter's transfer function is not realizable, so a Padé[1] approximation must be used. Thus, the transfer function is rewritten as:

$$F_\alpha(z) = b_\alpha(0) + z^{-1} \cdot \sum_{m01}^{M+1} b_\alpha(m) \cdot z^{-(m-1)} \qquad b_\alpha \ldots \textit{recursive filter parameter}$$

The filter coefficients $b_\alpha(m)$ decay almos in the same orther as the mel ceppstrum parameters $c_\alpha(m)$ do, which is actually rather fast. This means that the filter coefficients can be roughly quantized. Moreover, since the $b_\alpha(m)$ have almost the same statistical properties as those of the mel cepstrum, they can be used as a parametric representation of speech.

Data rate.

The filter coefficients are used as the spectral envelope parameter. It is shown from experimental results that the difference between the maximal and minimal values of each filter coefficient $b_\alpha(m)$ is bounded.

The filter coefficient sensitivities of the mel log spectrum are uniform for every order m. Consequently, the filter coefficients can be digitized by using a quantizer having the same quantization width q (< 1).

After experimental observations, the data amount is given by:

$$b_s = \begin{cases} (M+2)(2-[\log_2 q]+3) & \textit{if } M \le 9 \\ (M+2)(2-[\log_2 q]+14) & \textit{if } M > 9 \end{cases}$$

If the spectral envelope and pitch parameter are represented by b_s and b_p bit per frame transmitted every T seconds, the averall bit rate B of this system is given by:

---

1 A Padé approximant approximates a function in one variable. Given a function f and two integers m ≥ 0 and n ≥ 0, the Padé approximant of order (m, n) is the rational function:

$$R(x) = \frac{p_0 + p_1 x + p_2 x^2 + \cdots + p_m x^m}{1 + q_1 x + q_2 x^2 + \cdots q_n x^n}$$

which agrees with f(x) to the highest possible order. Equivalently, if R(x) is expanded in a Taylor series at 0, its first m + n + 1 terms would cancel the first m + n + 1 terms of f(x), and as such:

$$f(x) - R(x) = c_{m+n+1} x^{m+n+1} + c_{m+n+2} x^{m+n+2} + \cdots$$

$$B = \frac{b_s + b_p}{T}$$

The effect of varying T, q , M or b parameters has been studied, with the results shown in the following table.

T (ms)	M	q	Bp (bit)	B (kbits/s)	Speech quality
15	11	0.25	7	4	Very high
20	8	0.5	7	2	Fairly good
25	5	0.5	6	1.2	Still good

Spectral distortion

The spectral distortions are caused by the interpolation of the spectral envelope parameters of two succesive frames and by the quzntization of the spectral envelope parameter.

The approximate r.m.s. value of the spectral distortion caused by the quantization is referred to as D_Q (dB) and that caused by the interpolation is referred to as D_T (dB). They are given by the following expressions:

$$D_Q = \frac{q\sqrt{M+1}}{5} \quad ; \qquad D_T = 65\,T$$

Now, a method in which they apply the criterion used in the unbiased estimation of log spectrum to the spectral model represented by the mel-cepstral coefficients is proposed. To solve the nonlinear minimization problem involved in the method, an iterative algorithm whose convergence is guaranteed is given. Furthermore, the authors derive an adaptive algorithm for the mel-cepstral analysis by introducing an instantaneous estimate for gradient of the criterion. The adaptive mel-cepstral analysis system is implemented with an IIR adaptive filter which has an exponential transfer function, and whose stability is guaranteed.

Spectral estimation based on mel-cepstral representation

The model spectrum $H(z) = \exp \sum_{m=0}^{M} \tilde{c}(m) z^{\sim -m}$ can be expressed as follows:

$$H(z) = \exp \sum_{m=0}^{M} b(m)\cdot \phi_m(z) = K \cdot D(z)$$

$$\text{where} \quad \phi_m(z) = 1 \quad \text{if } m=0; \quad \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\left(\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right)^{-(m-1)} \quad \text{if } m \geq 1$$

$$\text{and} \quad K = \exp b(0) \quad ; \quad D(z) = \exp \sum_{m=1}^{M} b(m)\cdot \phi_m(z)$$

and where the cepstral (*c(m)*) and filter (*b(m)*) parameters are related through:

$$c(m) = b(m) \quad if\ m=M \quad and \quad b(m)+\alpha b(m+1) \quad if\ 0 \le m < M$$

In order to obtain an unbiased estimate, ε must be minimized, being:

$$\varepsilon = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{\left|D(e^{j\omega})\right|^2} d\omega$$

This minimization issue can be solved by means of the Newton-Raphson Method[2]. For the i-th result $b^{(i)}$, solving a set of linear equations

$$H \cdot \Delta b^{(i)} = -\nabla \varepsilon_{(b=b^{(i)})}$$

$$where \quad H \ldots Hessian\ matrix \quad H = \partial^2 \varepsilon / \partial b \partial b^T,$$

$$and\ the\ gradient\ \nabla \varepsilon \ \ is\ given\ by \quad \nabla \varepsilon = -2\vec{r},$$

$$where \quad \vec{r}(m) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_N(\omega)}{\left|D(e^{j\omega})\right|^2} \Phi_m(e^{j\omega})\, d\omega$$

we have the values: $\Delta b^{(i)} = [\Delta b^{(i)}(1), \Delta b^{(i)}(2), \ldots, \Delta b^{(i)}(M)]^T$, so $b^{(i+1)} = b^{(i)} + \Delta b^{(i)}$

Typically, few iterations are needed to obtain the solution, which makes this method quick and computationally efficient.

Adaptative mel-cepstral analysis algorithm

Replacing H with the unit matrix, the next result $b^{(i+1)}$ is given from the i-th result $b^{(i)}$:

$$b^{(i+1)} = b^{(i)} - \mu \nabla \varepsilon_{(b=b^{(i)})}$$

$$where \quad \mu \ldots adaptation\ step\ size$$

If we call *e(n)* to the output of the inverse filter *1/D(z)* driven by *x(n)*, we can interpret

2 The Newton–Raphson method is the best known method for finding successively better approximations to the zeros (roots) of a real-valued function *f(x)*:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$\varepsilon = E[e^2(n)] \text{ , so that } \quad \begin{array}{c} \nabla\varepsilon = -2\cdot E[e(n)e_\phi(n)], \\ \text{where } e_\phi(n) = [e_1(n), e_2(n), \ldots, e_M(n)]^T, \\ \wedge\ e_m(n) \ \ldots \text{output of the filter } \phi_m(z) \end{array} .$$

To derive an adaptative algorithm, an instantaneous estimate is introduced:

$$\tilde{\nabla}\varepsilon^{(n)} = -2e(n)e_\phi(n)$$

and, to suppress fluctuation of $b, \nabla\varepsilon$ is estimated using an exponential window:

$$\bar{\nabla}\varepsilon^{(n)} = -2(1-\tau)\sum_{i=-\infty}^{n}\tau^{n-i}e(i)e_\phi^{(i)} = \tau\bar{\nabla}\varepsilon^{(n-1)} - 2(1-\tau)e(n)e_\phi^{(n)}, \quad 0\le\tau<1 \ .$$

When the gain of the signal *x(n)* is time-varying, μ is normalized as:

$$\mu^{(n)} = \frac{a}{M\varepsilon^{(n)}}, \quad 0<a<1 \quad \text{where } \varepsilon^{(n)} \ldots \text{estimate of } \varepsilon \text{ at time } n: \ \varepsilon^{(n)} = \lambda\varepsilon^{(n-1)} + (1-\lambda)e^2(n), \quad 0\le\lambda<1$$

So the coefficient vector b(n) at time n can be updated:

$$\boxed{b^{(n+1)} = b^{(n)} - \mu^{(n)}\bar{\nabla}\varepsilon^{(n)}} \ .$$

And, since $K = \sqrt{\varepsilon_{min}}$, the mel cepstral coefficients $[\tilde{c}(m)]_{m=0}^{M}$ can be obtained.

Conclusion

The MLSA is a simple, great idea to be applied on low bit rate cepstral analysis synthesis systems since its parameters can be rougly quantized, and it has good stathistical features. Moreover, this system has fairly small spectral distortions.

In addition to this, the use of the above described algorithm will make the calculation of the cepstral coefficients so much easier and efficient, which has been awarded by experimental results.