Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis

Yoshinori Shiga and Simon King

Centre for Speech Technology Research
University of Edinburgh, Edinburgh, U.K.
yoshi@cstr.ed.ac.uk

Abstract

This paper proposes a novel approach for estimating the spectral envelope of voiced speech independently of its harmonic structure. Because of the quasi-periodicity of voiced speech, its spectrum indicates harmonic structure and only has energy at frequencies corresponding to integral multiples of F_0. It is hence impossible to identify transfer characteristics between the adjacent harmonics. In order to resolve this problem, Multi-frame Analysis (MFA) is introduced. The MFA estimates a spectral envelope using many portions of speech which are vocalised using the same vocal-tract shape. Since each of the portions usually has a different F_0 and ensuing different harmonic structure, a number of harmonics can be obtained at various frequencies to form a spectral envelope. The method thereby gives a closer approximation to the vocal-tract transfer function.

1. Introduction

The accurate estimation of the vocal-tract transfer function (VTTF) is an important issue to explain the mechanism of speech production. If we assume that in the source-filter model[1] the source is a periodic impulse train, the estimation of the VTTF is equivalent to that of a spectral envelope. The envelope can be obtained easily by analysing the speech waveform using well-known methods and hence is often used instead of the VTTF.

It has been pointed out, however, that estimation of the spectral envelope is interfered with by the harmonic structure (e.g. in [2]). Voiced sound shows quasi-periodicity in the time domain and its spectrum consists of line-spectra, or harmonics, which only have energy at frequencies corresponding to integral multiples of F_0. It is therefore impossible to precisely identify the transfer characteristics between the adjacent harmonics. Since, as F_0 increases, the number of harmonics decreases and the gaps between adjacent harmonics widen, it becomes harder to estimate a detailed envelope. Even identifying the location of formants can be difficult in some cases[3]. For this reason, speech with high F_0 such as the female voice makes accurate estimation of the envelope very difficult.

In estimating the spectral envelope of a voiced sound, conventional methods merely interpolate between harmonic spectra in the frequency domain using parameters such as cepstral coefficients or linear predictive coefficients. Thus the envelope obtained by those methods may vary depending on the harmonic structure of the observed speech even if the vocal tract system maintains an identical transfer characteristic. Fig.1 shows examples of spectra obtained by analysing artificial speech produced with a periodic impulse train through an all-pole filter, in which poles are at frequencies corresponding to the first and second formants of the sound [ʌ]. Each graph shows the spectrum of the filter output generated at a different fundamental frequency. The thick line shows the LPC spectrum of the output, the thin line the FFT spectrum of the output and the broken line the transfer characteristic of the filter. Evidently, LPC spectra estimated by a conventional method vary greatly depending on whether harmonics appear at the frequencies around the formant peaks of the filter characteristic. The estimated peaks are dulled if no harmonic exists at a peak frequency, and moreover the frequency of formant peaks, being affected by a harmonic with local maximum amplitude, tend to be incorrectly estimated.

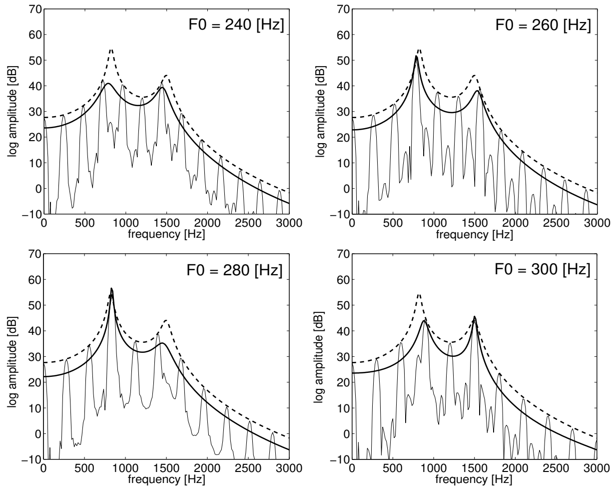

Figure 1: *Spectra of artificial voiced sound with different F_0s*

In this paper, we deal with the estimation of the spectral envelope of voiced speech. Our proposed method employs many portions of speech vocalised using the same vocal tract shape. Each of the portions having usually different F_0, we can obtain a sufficient number of harmonics at various frequencies to form a real spectral envelope which corresponds to the VTTF. The envelope is then estimated by interpolating and smoothing all the harmonic spectra of all the portions.

2. Multi-frame analysis

2.1. Outline of the proposed method

In order to resolve the aforementioned problem, we first locate portions of speech which are vocalised using almost the same vocal-tract shape. Although the similarity of the vocal tract shape can be roughly estimated from its phonetic contexts, we employ data that tell us the actual shape with greater reliability. The data used in this study include articulators' positions measured with an electromagnetic articulograph (EMA) system[4].

Since in the frequency domain each of the portions generally has a different F_0 and ensuing different spacing between adjacent harmonics, we can increase the number of harmon-

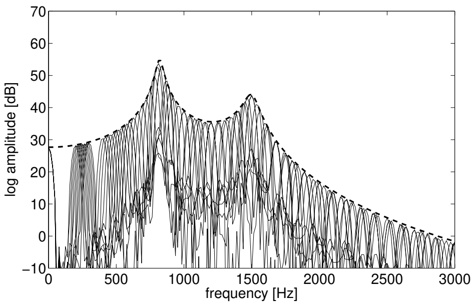

Figure 2: *Overlapped spectrum of speech in various F_0s*

ics to form a spectral envelope representing the VTTF. Let us consider again the above-mentioned artificial speech. Fig.2 shows overlapped 1024-point FFT results calculated from the filter outputs in different F_0s (200-300 Hz at 10 Hz intervals). As it is clear from this figure, it becomes possible to obtain more detailed spectral envelopes equivalent to the VTTFs from the (quasi-)periodic system outputs with different F_0, when the transfer function of the system remains constant.

Spectral envelopes are obtained by interpolating and smoothing all the harmonics included in all collected portions using cepstra based on a least-squares criterion. Because of the use of multiple frames in spectral envelope analysis, we call this approach *Multi-frame Analysis* (MFA).

2.2. Spectral envelope estimation

2.2.1. Cepstrum

We adopt *analysis frames* as the above-mentioned speech *portions*, and employ the *cepstrum*[5] as an expression of the spectral envelope for the purpose of smoothing and interpolating harmonic spectra of multiple frames. The cepstrum is a suitable parameter for representing zero and pole characteristics of a system and can easily be developed into a perceptual scale (Mel-scale)[6]. These merits allow the cepstrum to be widely applied in the field of the speech technology (e.g. [7]).

Let $X(e^{j\omega})$ denote the Fourier transform of the speech waveform. $\hat{X}(e^{j\omega})$, the natural logarithm of $X(e^{j\omega})$, is then defined as:

$$\hat{X}(e^{j\omega}) = \ln|X(e^{j\omega})| + j\arg[X(e^{j\omega})] \qquad (1)$$

whilst $\hat{X}(e^{j\omega})$ is given as the Fourier transform of the complex cepstrum $\hat{x}[n]$ by:

$$\hat{X}(e^{j\omega}) = \sum_{n=-\infty}^{\infty} \hat{x}[n]e^{-j\omega n} \qquad (2)$$

Taking into consideration the properties of the complex cepstrum that $\hat{x}[n]$ is a real number and the sum of an even function $\hat{x}_e(n)$ and an odd function $\hat{x}_o(n)$, the following equations are obtained on referring to Eq.(1) and (2):

$$\ln|X(e^{j\omega})| = \sum_{n=-\infty}^{\infty} \hat{x}_e[n]\cos(\omega n) \qquad (3)$$

$$\arg[X(e^{j\omega})] = -\sum_{n=-\infty}^{\infty} \hat{x}_o[n]\sin(\omega n) \qquad (4)$$

where,

$$\hat{x}_e[n] = \frac{\hat{x}[n]+\hat{x}[-n]}{2},\ \hat{x}_o[n] = \frac{\hat{x}[n]-\hat{x}[-n]}{2} \qquad (5)$$

Eq.(3) and (4) represent the log-amplitude spectrum and phase spectrum respectively.

2.2.2. Estimating the envelope of the amplitude spectrum

Let us determine a cepstrum which approximates the amplitude of all the harmonics included in M speech frames based on Eq.(3) using the least squares method. This can be considered the expansion of the cepstrum estimation in [8].

Let a_k^l and f_k^l denote, respectively, an observed log-amplitude and frequency of the l-th harmonic ($l = 1, 2, ..., N_k$) within the speech frame $k(= 1, 2, 3, ..., M)$, and T the sampling period. Then, the total error ε, the sum of the squares of the amplitude approximation error for all the harmonics of all the frames, is as follows:

$$\varepsilon = \sum_{k=1}^{M}\sum_{l=1}^{N_k} \frac{w(f_k^l)}{N_k}\left[a_k^l - d_k - \sum_{n=-p}^{p} \hat{x}_e[n]\cos(\Omega_k^l n)\right]^2 \qquad (6)$$

where

$$\Omega_k^l = 2\pi f_k^l T$$

In Eq.(6) we introduce weighting functions, $w(f)$ for attaching importance to the low frequency band, and $(1/N_k)$ for evaluating each frame equally regardless of the number of harmonics. The factor d_k is an offset that adjusts the total power of each frame so as to minimise the error ε. Eq.(6) is expressed in terms of vectors and matrices as:

$$\varepsilon = \sum_{k=1}^{M}(\boldsymbol{a}_k - d_k\boldsymbol{u}_k - \boldsymbol{B}_k\boldsymbol{c})^T\boldsymbol{W}_k(\boldsymbol{a}_k - d_k\boldsymbol{u}_k - \boldsymbol{B}_k\boldsymbol{c}) \qquad (7)$$

where the vector $\boldsymbol{c}$ indicates 0- to p-order cepstral coefficients,

$$\boldsymbol{c} = [\ \hat{x}_e[0]\ \hat{x}_e[1]\ \hat{x}_e[2]\ \cdots\ \hat{x}_e[p]\ ]^T$$

and both $\boldsymbol{a}_k$ and $\boldsymbol{u}_k$ are N_k dimension vectors,

$$\boldsymbol{a}_k = [\ a_k^1\ \ a_k^2\ \ a_k^3\ \cdots\ a_k^{N_k}\ ]^T,\ \boldsymbol{u}_k = [\ 1\ \ 1\ \ 1\ \cdots\ 1\ ]^T$$

and

$$\boldsymbol{B}_k = \begin{bmatrix} 1 & 2\cos(\Omega_k^1) & 2\cos(2\Omega_k^1) & \cdots & 2\cos(p\Omega_k^1) \\ 1 & 2\cos(\Omega_k^2) & 2\cos(2\Omega_k^2) & \cdots & 2\cos(p\Omega_k^2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos(\Omega_k^{N_k}) & 2\cos(2\Omega_k^{N_k}) & \cdots & 2\cos(p\Omega_k^{N_k}) \end{bmatrix}$$

$$\boldsymbol{W}_k = \frac{1}{N_k}\begin{bmatrix} w(f_k^1) & & & \\ & w(f_k^2) & & 0 \\ & 0 & \ddots & \\ & & & w(f_k^{N_k}) \end{bmatrix} \qquad (8)$$

The equation can be solved by reducing it to a problem of weighted least squares. The normal equation is thereby:

$$\left(\sum_{k=1}^{M}\boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{B}_k\right)\boldsymbol{c} = \sum_{k=1}^{M}\boldsymbol{B}_k^T\boldsymbol{W}_k(\boldsymbol{a}_k - d_k\boldsymbol{u}_k) \qquad (9)$$

By solving this, the cepstrum coefficients $\boldsymbol{c}$ can be found. Using $\boldsymbol{c}$ obtained, the factor d_k that minimises the entire error of each frame k is calculated as:

$$d_k = \frac{\boldsymbol{u}_k^T\boldsymbol{W}_k(\boldsymbol{a}_k - \boldsymbol{B}_k\boldsymbol{c})}{\boldsymbol{u}_k^T\boldsymbol{W}_k\boldsymbol{u}_k} \qquad (10)$$

From the above, we can obtain a cepstrum which best approximates all the harmonic amplitude spectra of all the M frames according to the following procedure: 1) Substitute $\mathbf{0}$ for $\boldsymbol{c}$ (initial value); 2) Obtain d_k using Eq.(10); 3) Calculate ε of Eq.(7), and terminate the procedure if ε converges; 4) Find $\boldsymbol{c}$ by solving Eq.(9); 5) Substitute 0 for $\hat{x}_e[0]$ (amplitude normalization); 6) Return to Step 2.

2.2.3. *Estimating the envelope of the phase spectrum*

Next, we determine a cepstrum which approximates the phase of all the harmonics included in M speech frames. Whilst the left side of Eq.(4), $\arg[X(e^{j\omega})]$, represents *unwrapped* phase spectrum, observed phases are *wrapped* phase, which is expressed using $\mathrm{ARG}[X(e^{j\omega})]$ in [5]. Unwrapping harmonic phases for each individual frame leads to lack of accuracy because of the influence of spacing between harmonics. Hence we cannot estimate the envelope of phase spectrum in the same way as we did that of amplitude spectrum. We therefore adopt the following method instead.

Let θ_k^l and f_k^l denote, respectively, an observed wrapped phase and frequency of the l-th harmonic ($l = 1, 2, 3, ..., N_k$) within the speech frame $k(= 1, 2, 3, ..., M)$. First of all, we take a moving average of every observed phase, which is expressed as a point on the unit circle of the complex plane, under a weighting factor $(1/N_k)$ using the following equation.

$$\phi(f) = \frac{\displaystyle\sum_{k=1}^{M}\sum_{l=1}^{N_k} \frac{1}{N_k} G(f_k^l - f) e^{j(\theta_k^l - 2\pi f_k^l \tau_k)}}{\displaystyle\sum_{k=1}^{M}\sum_{l=1}^{N_k} \frac{1}{N_k} G(f_k^l - f)} \tag{11}$$

Here, the function $G(f)$ indicates a window function for the moving average, and τ_k a delay factor that adjusts the global tilt of the phase spectrum so as to minimise the error ϵ below. Let its amplitude be normalised,

$$\bar{\varphi}(f) = \frac{\phi(f)}{|\phi(f)|} \tag{12}$$

We consider $\bar{\varphi}(f)$ the mean phase-spectrum of all the harmonics in the M frames. In addition, we define the error between the mean phase-spectrum $\bar{\varphi}(f_k^l)$ and all the observed harmonic phase θ_k^l as follows:

$$\epsilon = \sum_{k=1}^{M}\sum_{l=1}^{N_k} \frac{w(f_k^l)}{N_k} \left\{ \frac{1}{2\pi f_k^l} \mathrm{ARG}\left[\frac{e^{j(\theta_k^l - 2\pi f_k^l \tau_k)}}{\bar{\varphi}(f_k^l)} \right] \right\}^2 \tag{13}$$

The error ϵ decreases and converges by iterating the procedure with two steps, a step to adjust the observed phase θ_k^l with the delay τ_k so that ϵ in Eq.(13) is minimised for each frame k, and a step to calculate the mean spectrum $\bar{\varphi}(f)$ applying τ_k to Eq.(11) and (12). If the phase θ_k^l is near the mean spectrum $\bar{\varphi}(f_k^l)$ and their difference is within the range of $-\pi$ to π, we can obtain the correcting value t_k, which minimises ϵ, for the delay τ_k as:

$$t_k = \frac{\displaystyle\sum_{l=1}^{N_k} \frac{w(f_k^l)}{2\pi f_k^l} \mathrm{ARG}\left[\frac{e^{j(\theta_k^l - 2\pi f_k^l \tau_k)}}{\bar{\varphi}(f_k^l)} \right]}{\displaystyle\sum_{l=1}^{N_k} w(f_k^l)} \tag{14}$$

As an initial value of τ_k, we use the following:

$$\tau_k = \frac{\theta_k^1 + \pi}{2\pi f_k^1} \tag{15}$$

Here, θ_k^1 and f_k^1 correspond, respectively, to the observed phase and frequency of the fundamental harmonic.

The above mean phase-spectrum $\bar{\varphi}(f)$ usually includes a certain degree of delay factor (the slope of a phase spectrum). In order to remove this, we calculate an N-point ($N = 512$) discrete spectrum by sampling the continuous spectrum $\bar{\varphi}(f)$ in Eq.(12) between 0 and the Nyquist frequency, following which phase unwrapping[5] is performed. The delay α is then calculated as the slope of a linear-regression line that best fits the phase spectrum. Finally, adding the delay α to τ_k, we cancel the delay component.

From the above, we can obtain a cepstrum which approximates all the harmonic phase spectra of M frames according to the following procedure: 1) Initialise τ_k using Eq.(15); 2) Calculate ϵ using Eq.(13) and proceed to Step 6 if ϵ converges; 3) Find the correcting value t_k using Eq.(14) and substitute $(\tau_k + t_k)$ for τ_k for all k; 4) Find the delay α and substitute $(\tau_k + \alpha)$ for τ_k for all k; 5) Return to Step 2; 6) Calculate cepstrum coefficients $\hat{x}_o[n](n = -p, -p+1, -p+2, ..., p)$ by performing inverse discrete Fourier transform to an N-point ($N = 512$) discrete spectrum obtained by sampling and unwrapping the continuous spectrum $\bar{\varphi}(f)$ in Eq.(12).

3. Experiment

We conducted an experiment applying the MFA to actual speech data that include articulatory information.

3.1. Data

The data used in the experiment is a corpus composed of 460 sentences uttered by a female speaker (fsew0). The corpus includes parallel acoustic-articulatory information, which was recorded using a Carstens EMA system at Queen Margaret University College, Edinburgh (see [4] for details). The sampling frequencies of the acoustic waveform and articulatory information are 16 kHz and 0.5 kHz respectively.

3.2. Method

Voiced sections were first extracted from the corpus and used to build a set of pairs of harmonic spectra and articulators' positions. The harmonic spectra were estimated from the waveform using the weighted least squares method in [9]. The width and spacing of the time window (Hanning) for this analysis were 20 ms and 8 ms respectively. The articulatory information was downsampled to the same period of 8 ms. Thereby 87208 voiced frames with parallel acoustic-articulatory information were obtained in total.

In order to identify frames with similar articulator settings, all the voiced frames were divided into 2048 clusters by applying LBG clustering method[10] to the articulatory data.

Finally, cepstrum coefficients (order of 40) were calculated by applying the MFA to all the frames in each cluster. In this calculation, we employed a Gaussian distribution with 0 Hz mean and 2 kHz standard deviation for the weighting function $w(f)$ in Eqs.(8)(13)(14), and a Gaussian window with 100 Hz standard deviation for the moving-average window $G(f)$ in Eq.(11).

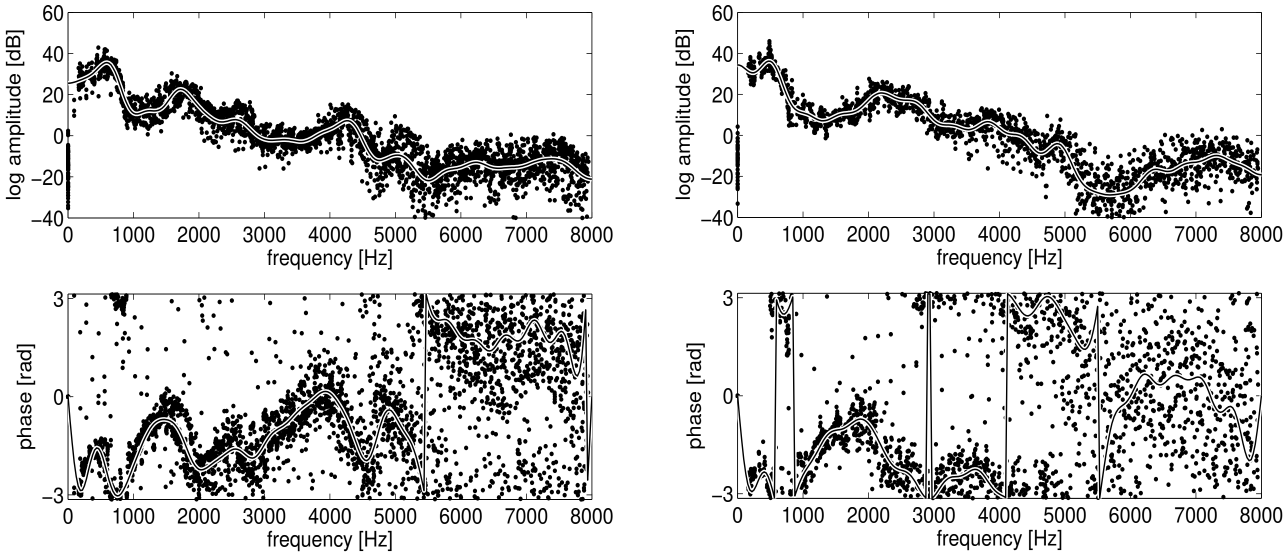

Figure 3: *Spectral envelopes of two articulatory clusters estimated using the MFA*

3.3. Results

Fig.3 shows two examples of spectral envelopes calculated from the cepstra obtained. In this figure, the dots represent observed harmonic amplitudes (biased by the factor d_k) in the upper graphs and phases (shifted by the delay τ_k) in the lower graphs. The solid line indicates the envelope of the amplitude spectrum (upper) and phase spectrum (lower) calculated by the MFA.

4. Discussion

The experimental results show smooth envelopes which approximate all the harmonics in each cluster. There appears, however, to be a tendency that some clusters have fairly large variances of observed harmonic spectra. A main cause of this is thought to be that harmonic spectra are distorted by the SN ratio of speech signal or some noises mixed in. However, it can also be considered that the variance of harmonic data in the *acoustic* space may become large in those clusters, because we adopt clustering in the articulatory space but do not pay any attention to the variance of data in the acoustic space. This will be the focus of future work.

We assumed at the beginning that the voice source of the source-filter model was a periodic impulse train and considered spectral envelopes as the VTTFs. However, it has been reported (e.g. in [11]) that the glottal waveform changes its shape depending mainly on the pitch and power of the waveform. In respect of this source problem, we are currently examining an approach[12] that can take into account the change of glottal characteristic using the MFA.

5. Conclusions

In this paper, we proposed a method to estimate the spectral envelope of voiced speech free from the effects of its harmonic structure. We discussed its theoretical aspects and conducted an experiment applying it to actual speech data.

Another aspect of the MFA is that it produces a reliable codebook which relates articulation with acoustic feature of speech. We plan to make use of this correspondence to elucidate the mechanism of speech production.

Acknowledgements

In carrying out this research, Y. Shiga is supported financially in part by the ORS Awards Scheme.

References

[1] Fant, G., *Acoustic Theory of Speech Production*, The Hague, Mouton, 1960.

[2] Makhoul, J., “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[3] Kent, R. D. and Read, C., *The Acoustic Analysis of Speech*, Singular Publishing Group, 1992.

[4] Wrench, A. A., “A new resource for production modelling in speech technology,” in *Proc. Workshop on Innovations in Speech Processing*, Stratford-upon-Avon, 2001.

[5] Oppenheim, A. V. and Schafer, R. W., *Discrete-Time Signal Processing*, Prentice Hall, 1989.

[6] Koishida, K., Tokuda, K., Kobayashi, T., and Imai, S., “CELP coding based on Mel-cepstral analysis,” in *Proc. ICASSP95*, May 1995, vol. 1, pp. 33–36.

[7] Shiga, Y., Hara, Y., and Nitta, T., “A novel segment-concatenation algorithm for a cepstrum-based synthesizer,” in *Proc. ICSLP94*, 1994, vol. 4, pp. 1783–1786.

[8] Galas, T. and Rodet, X., “An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sounds,” in *Proc. Int. Computer Music Conf.*, 1990, pp. 82–84.

[9] Stylianou, Y., “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[10] Linde, Y., Buzo, A., and Gray, R. M., “An algorithm for vector quantizer design,” *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.

[11] Miller, R. L., “Nature of the vocal cord wave,” *J. Acoust. Soc. Am.*, vol. 31, no. 6, pp. 667, 1959.

[12] Shiga, Y. and King, S., “Estimation of voice source and vocal tract characteristics based on multi-frame analysis,” in *Proc. Eurospeech2003*, Geneva, Switzerland, Sept. 2003, vol. 3, pp. 1749–1752.