Estimation of formant frequencies by means of a wavelet transform of the speech spectrum

M. Bensaid, J. Schoentgen and S. Ciocea
Université Libre de Bruxelles
Institute of Modern Languages and Phonetics, Laboratory of Experimental Phonetics
50, av. F.-D. Roosevelt, B-1050 Brussels, Belgium
tel. +32 2 650 2010 fax. +32 2 650 2007
jschoent@ulb.ac.be

Abstract—**The objective is to present a method that extracts the spectral envelope of a speech signal. The method is based on the wavelet transform, which is a family of multi-resolution analysis methods. The extraction is founded on the observation that spectral envelope and source-related spectral components vary on different frequency scales. The difference between cepstral and wavelet analysis is that the latter is not based on the hypothesis that a speech signal is the outcome of a convolution of the source signal with the vocal tract impulse response. The method of analysis was tested on a corpus of [a],[i],[u] vowels sustained by healthy and dysphonic speakers. Results show that the envelopes extracted via cepstral and wavelet analysis are very similar except in the case of [u], for which the first two formants extracted by means of wavelet analysis are shifted slightly towards higher values and the formant peaks are somewhat closer.**

Keywords— **Signal Analysis, Cepstrum, Wavelet transform, Formant.**

I. Introduction

The formants are the peaks in the speech spectrum that are the effects of the eigenmodes of the vocal tract. The formant frequencies are the corresponding eigenfrequencies. Vowels, semi-vowels, lateral approximants, nasals and trills are among the speech sounds that are characterized by formant frequencies and their transitions.

The measurement of formant frequencies is a signal processing problem which has been approached via several methods, both temporal and spectral [1]. Among the latter, the best-known are based either on linear predictive coding or cepstral analysis of speech. Both methods detect the local maxima of the spectral envelope. For linear predictive coding, the positions of the spectral maxima agree with the frequencies of the complex conjugate poles. For the cepstrum, the positions of the maxima must be determined via peak-picking.

Cepstral analysis is founded on the observation that the amplitudes of the spectral components vary on distinct frequency scales. One scale is related to the amplitudes of the partials and noise component of the source signal while another involves amplitude changes typical of the spectral envelope. Since these variations occur on qualitatively different frequency scales, it has been shown that it is possible to separate them within the framework of a mathematical representation in which the amplitude changes typical of both scales are additive.

Indeed, a speech signal is the outcome of a convolution of a source signal with an impulse response on the part of the vocal tract. The Fourier transform of the signal is therefore the product of the Fourier transforms of the source signal and impulse response. The multiplicative link is preserved inside the amplitude spectrum and is turned into an additive relation via the logarithm. The logarithm of the amplitude spectrum is therefore the sought after mathematical representation, in the framework of which the relation between the source spectrum and the envelope is additive. To position both constituents on an axis so that they can be separated by putting one or the other to zero, the logarithm of the amplitude spectrum is then (inverse) Fourier transformed. The unwanted (high quefrequency) cepstral coefficients are zeroed and the desired constituent of the spectrum - its envelope -

is obtained by (direct) Fourier transforming the nonzeroed cepstral coefficients [2].

To sum up, the foundation of cepstral analysis is the existence of separate frequency scales for the source and the envelope, and the additivity of both within a chosen mathematical representation; this additivity is the consequence of the postulate that source signal and vocal tract response are convolved. Interestingly, this postulate can be dispensed with within the framework of a multi-scale analysis of the spectrum. Indeed, the additivity then falls within the province of inter-scale relations, and is therefore not founded on the log-transform of the amplitude spectrum.

At present, the best-known multi-scale analysis method is wavelet analysis, which is a family of methods which differ according to the wavelet that is made use of and whether the representation is invertible or not [3], [4]. Here, we show how wavelet analysis can be applied to the extraction of the spectral envelope. The results show that the wavelet-based method obtains spectral envelopes that are similar to envelopes arrived at by means of cepstral analysis, and that the method is robust, i.e. a fixed set of scales exists which reconstructs the envelope, while the cepstral method requires the adjustment of the cut-off quefrequency from vowel to vowel so as to avoid spuriously inserting or omitting peaks.

II. Wavelet analysis

Wavelet analysis is the multi-resolution analysis of a signal [5], [6]. Basically, the analysis consists of expanding the signal by means of oscillating waveforms (called wavelets) which are significantly different from zero on a finite interval only [7]. The experimenter controls the positions of the wavelets and the length of the period during which they are different from zero. Since the number of oscillations is fixed, the wavelet oscillates slowly on long, and rapidly on short, wavelet supports. As a consequence, the projection of a one-dimensional signal on a one-dimensional wavelet is expected to be small when the rate of change of the signal at a given position is either faster or slower than the typical rate of change of the wavelet. On the contrary, the projection is expected to assume values that are big when the rate of change of both are similar [8]. Since the positions of the wavelets along the signal and their scales (fine or coarse) are determined by the experimenter, he can build a wavelet representation redundant both in time (if the signal is temporal) and in scale. The outcome is an analysis that is both prolix (since the same stretch of signal has been analyzed by means of several wavelets at different positions and scales) and non-invertible because of the many-to-one relation between the wavelet coefficients and the signal samples.

One solution is to resort to a set of wavelets which form an orthogonal base. The Daubechies wavelets, for instance, are orthogonal, zero beyond a finite interval and almost smooth [7]. A (fast) invertible wavelet analysis thus consists of the following. Firstly, the selection of a number of signal samples which is a power of two. This number determines the number of scales since progressing from the coarsest to the finest consists of doubling the resolution at each step. Secondly, the performance of a fast wavelet transform which yields a number of coefficients equal to the number of samples. The number of coefficients from the coarsest to the finest scales increases as 1,2,4,8,etc. For any given scale, the positions of the wavelets are equidistant. Signal constituents can be re-synthesized scale by scale and, together, reproduce the original exactly. Separating constituents, e.g. de-noising, consists of reproducing constituents by means of subsets of scales, and processing or storing them separately.

The speech spectrum is a sequence of partials and noise components of the source signal whose amplitudes have been shaped by the vocal tract transfer function. Since the transfer function is characterized by few poles, the spectral envelope changes slowly with frequency compared to the frequency components of the source, the amplitudes of which may fluctuate periodically or noisily with the frequency. As a consequence, envelope and source components involve different frequency scales. It is therefore appropriate to carry out a multi-resolution analysis of the amplitude spectrum. This analysis consists of performing a fast wavelet transform of the spectrum. Here, the signal is not a sampled temporal quantity, but a quantity evolving with frequency. The same algorithms apply, however.

Cepstral analysis was carried out in parallel in order to compare the spectral envelopes obtained by both methods.

III. Methods

The speech signals were sustained vowels [a],[i],[u] produced by healthy and dysphonic speakers in view of an examination of the effects of laryngeal pathologies on the speech signal. Cepstral analysis was carried out as follows.

a) Sampling of the sustained vowel signals at 20 kHz.
b) Hamming or Bartlett windowing of the analysis interval of 8192 samples.
c) Performance of a fast Fourier transform of the windowed signal.
d) Calculation of the amplitude spectrum via the module of the complex spectrum.
e) Calculation of the logarithm of the amplitude spectrum.
f) Performance of the inverse Fourier transform on the log-amplitude spectrum.
g) Zeroing of the cepstral coefficients above a fixed quefrequency; the discrete cut-off quefrequency was typically equal to, or less than, 80; the cut-off was adjusted manually when required.
h) Performance of a fast Fourier transform of the liftered cepstrum; the outcome is the spectral envelope, which is symmetrical with reference to the zero frequency axis.

Wavelet analysis was carried out as follows.

a) Sampling of the sustained vowel signals at 20 kHz.
b) Hamming or Bartlett windowing of the analysis interval of 8192 samples.
c) Performance of a fast Fourier transform on the windowed signal.
d) Calculation of the amplitude spectrum via the module of the complex spectrum.
e) Optional calculation of the logarithm of the amplitude spectrum; multi-resolution analysis could be carried out on the amplitude spectrum or on the log-transformed one.
f) Performance of a fast wavelet transform; the number of input samples was 8192 and the number of scales 13; the wavelet was the 20 coefficient Daubechies wavelet.
g) Cut-off between scales 7 and 8; reconstruction of the spectral envelope by means of the 7 lowest scales comprising a total of 128 wavelets.

IV. Results and discussion

a) Figures 1(a) to 3(a) show the log-amplitude spectrum of the vowels [u],[a],[i] respectively. Overlaid is the spectral envelope of the log-amplitude spectrum arrived at by means of wavelet analysis. The horizontal axis is the frequency axis in Hz and the vertical one the log-amplitude in arbitrary units. Figures 1(b) to 3(b) show the same spectra with the overlaid envelope obtained via cepstral analysis. As far as the first three formants are concerned, the peak positions arrived at by means of cepstral and wavelet analysis agree to within a few Hz. The only systematic disagreement occurs for the vowel [u]. Here, the peaks of the two first formants are always shifted to higher frequencies and closer together when the envelope is arrived at via wavelet analysis than when it is obtained by means of cepstral analysis. Indeed, cepstral analysis is known to exaggerate the inter-peak distance in the case of the vowel [u].

b) Figure 4(a) shows the amplitude spectrum of the vowel [a] and, overlaid, the spectral envelope arrived at by means of wavelet analysis. The horizontal axis is the frequency axis in Hz and the vertical one the spectral amplitude in arbitrary units.

Figure 4(b) shows (crossed line) the spectral envelope obtained via cepstral analysis of the vowel [a] the spectrum which is displayed in Figure 4(a). The dashed line shows the logarithm of the spectral envelope arrived at via the wavelet analysis of the amplitude (not the log-amplitude) spectrum. It will be seen that the positions of the main spectral peaks match. This is remarkable considering that the logarithm is a non-linear operator which, generally speaking, does not commutate with others. Indeed, in this case wavelet analysis was performed before the log-transform. Finally, the solid line is the envelope obtained via the wavelet analysis of the log-amplitude spectrum. The agreement between the envelopes arrived at by means of the wavelet and cepstral analyses of the log-spectrum is better than a few Hz almost everywhere.

c) The difference between both methods is that according to the hypotheses underlying cepstral analysis, the signal is the output of a convolution operator (i.e. linear operator). This means that, firstly, if this hypothesis is not correct, cepstral analysis can be expected to fail and, secondly, it is necessary to compute the log-spectrum since it is only within the log-spectrum that the convolution turns into a sum. Wavelet analysis therefore is a more general method which makes fewer assumptions than the cepstral method.

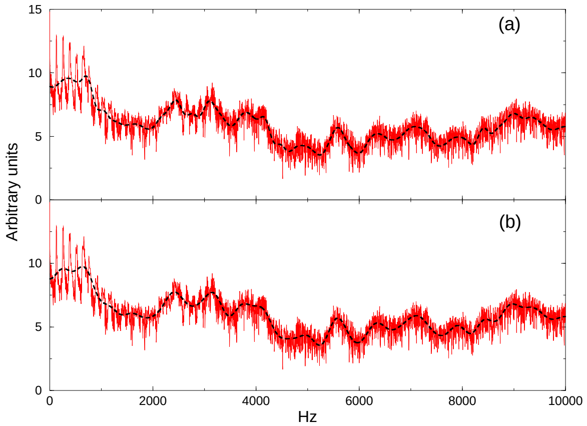

Fig. 1. (a) Log-amplitude spectrum (dotted line) of the vowel [u] and the spectral envelope (dashed line) arrived at by means of wavelet analysis; (b) Log-amplitude spectrum (dotted line) of vowel [u] and the spectral envelope (dashed line) arrived at by means of cepstral analysis.

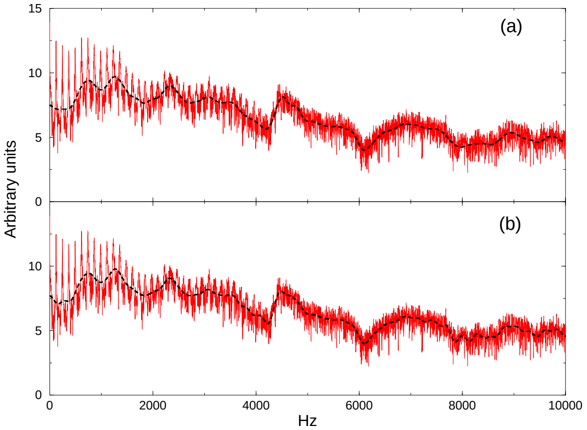

Fig. 2. (a) Log-amplitude spectrum (dotted line) of the vowel [a] and the spectral envelope (dashed line) arrived at by means of wavelet analysis; (b) Log-amplitude spectrum (dotted line) of vowel [a] and the spectral envelope (dashed line) arrived at by means of cepstral analysis.

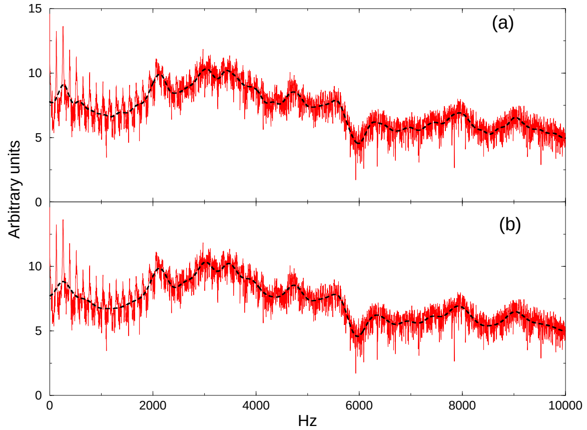

Fig. 3. (a) Log-amplitude spectrum (dotted line) of the vowel [i] and the spectral envelope (dashed line) arrived at by means of wavelet analysis; (b) Log-amplitude spectrum (dotted line) of vowel [i] and the spectral envelope (dashed line) arrived at by means of cepstral analysis.

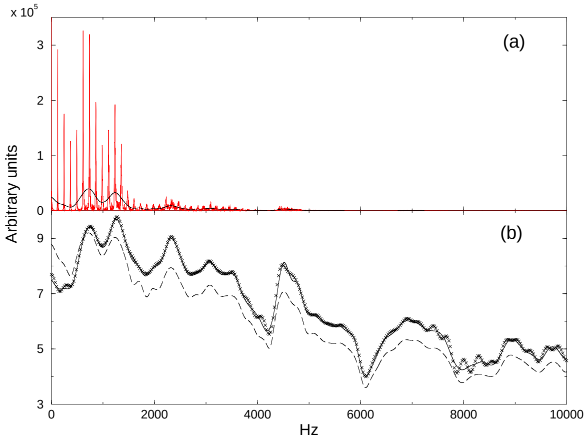

Fig. 4. (a) Amplitude spectrum of the vowel [a] (dotted line) and its envelope (solid line) obtained by means of wavelet analysis. (b) Spectral envelope of the log-amplitude spectrum of the vowel [a] obtained via cepstral (crossed line) and wavelet analysis (solid line). The dashed line is the logarithm of the spectral envelope shown in Figure 4(a).

d) We analyzed sustained vowels since they are adequate items considering our concern with the analysis of dysphonic voices. When the aim is to analyze connected speech, a 8192 sample window is too large. We experimented with sliding windows of down to 1024 samples and obtained satisfactory results. The standard window length of 512 samples (i.e. 25 ms) still yielded acceptable results.

e) The number of wavelet coefficients of the spectral envelope was 128, and this number must be compared to the number of cepstral coefficients, which was maximally 80. As far as storage requirements are concerned, cepstral coefficients appear to be more economical than wavelet coefficients, but algorithmically speaking the latter are obtained more easily.

References

[1] J. Deller, J. Proakis, J. Hansen, Discrete-time Processing of speech signals, Macmillan Publishing Company 1993.

[2] L.R. Rabines, R.W. Schafer, Digital Processing of Speech signals, Prentice Hall, 1978.

[3] I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, PA, 1996.

[4] G. Kaiser, A friendly Guide to Wavelets, Birkhauser, 1994.

[5] I. Daubechies, The wavelet transform, time frequency localization and signal analysis, IEEE Trans. Inform. Theory **36** (1990) 961-1005.

[6] S. Mallat, Multiresolution approximation and wavelets, Trans. Amer. Math. Soc. **315** (1989) 69-88

[7] I. Daubechies, Orthonormal bases of compactly supported wavelets, Comm. Pure Appl. Math (1988) (909-996).

[8] J.P. Antoine, Wavelet Analysis: A new tool in signal processing, Physicalia Mag. **16** (1994) 17-42.