



Tracking of Involuntary Formant Frequency Variations and Application to Parkinsonian Speech

Laurence Cnockaert^{1*}, Jean Schoentgen^{1**}, Pascal Auzou^{2,3}, Canan Ozsancak^{3,4}, Francis Grenez¹

¹Université Libre de Bruxelles, Department Waves and Signals, Brussels, Belgium

²Service d'exploration fonctionnelles neurologiques, Fondation HOPALE, Berck sur Mer, France

³EA 2683, CHRU Lille, France

⁴Service de Neurologie, Centre Hospitalier de Boulogne sur Mer, France

lcnockae@ulb.ac.be

Abstract

The objective of this paper is to present a formant frequency estimation method, developed with a view to track small variations due to involuntary vocal tract movement. The formant frequency estimation is based on the instantaneous frequencies obtained by means of a complex wavelet transform and is synchronised with the glottal cycle. Results for synthetic speech signals show the precision of the formant frequency estimation method. Results are presented for speech signals from both normophonic speakers and speakers with Parkinson's disease.

Index Terms: formant tracking, continuous wavelet transform, disordered speech.

1. Introduction

The estimation of the characteristics of the vocal tract from the speech signal is an important research topic, because of its utility for the comprehension and the modelling of the speech production mechanism. In the clinical domain, precise descriptions of the variations of the vocal tract can also be useful to characterize neurological diseases presenting tremor, like Parkinson's disease. The vocal tract is usually described by means of the formants, which are the peaks in the speech signal spectrum and correspond to the resonances of the vocal tract. The aim of this paper is to track the involuntary movements of the vocal tract during the production of sustained vowels, by means of the formant frequencies. A reliable method of formant frequency estimation is presented.

The vocal tract properties vary with time for two main reasons. First, the shape of the vocal tract changes during speech production because of the movements of the articulators. Secondly, variations occur at the glottal cycle rhythm, because of the oscillation of the vocal cords between a closed and an open phase [1]. Indeed, when the vocal tract is closed at the glottis, the speech signal results from the free resonances in the vocal tract. On the other hand, during the open phase, the vocal tract is acoustically coupled with the glottis and the trachea and the resonances are modified. Therefore, in order to study properly the variations of the vocal tract and its free resonances, the formant frequencies should be tracked during the closed phase of the glottis.

The formant tracking method should be able to track variations with frequencies up to about 15Hz. In order to obtain these high performances in the temporal tracking of the formant parameters, the effective length of the analysis frames must be shorter than the glottal cycle, and must be synchronised with the glottal cycle [2]. Moreover, as the aim is to track formant frequency variations, the ability of the method to track variations of the formant frequencies is more important than the accuracy of the estimated formant frequencies. This is the reason why the performance criterium for the formant frequency estimation is based on the quality of the tracking of the formant frequency variations, and not on the vicinity of the estimated frequency with the reference.

In this paper, a non-stationary method of formant frequency estimation is presented. The method is based on the instantaneous frequencies obtained by means of a complex wavelet transform, and is synchronised with the closed phase of the glottal cycle. The performance of the method is shown on synthetic speech signals. Results are presented for real speech signals from normophonic speakers and speakers with Parkinson's disease.

2. Formant frequency estimation

2.1. Continuous wavelet transform

The instantaneous frequency $IF(t)$ of a band-pass signal $s(t)$ is usually defined by means of its Hilbert transform $H[s(t)]$ [3].

$$\Phi(t) = \arg[s(t) + jH[s(t)]] \quad (1)$$

$$IF(t) = \frac{1}{2\pi} \frac{d\Phi(t)}{dt} \quad (2)$$

The IF can also be defined using a continuous wavelet transform $CWT(\lambda, t)$ using an analytical wavelet $\psi_a(t)$ [4]. The continuous wavelet transform of a signal $x(t)$ is defined as

$$CWT(\lambda, t) = \int_{-\infty}^{+\infty} x(u) \frac{1}{\sqrt{\lambda}} \psi^* \left(\frac{u-t}{\lambda} \right) du, \quad (3)$$

where $\psi(t)$ is the mother wavelet, and where $CWT(\lambda, t)$ is the wavelet transform coefficient for a scale factor λ , at time t . The amplitude and phase of the complex CWT coefficients obtained using an analytical mother wavelet are the envelope and instantaneous phase of the spectral components of the signal in the frequency-band centred on the central frequency f_c of the wavelet [5]. The time-derivative of the phase of the complex CWT coefficients is therefore an estimate of the instantaneous frequency of

*The first author is a fellow with the *Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture* (Belgium).

**The second author is a Senior Research Associate with the *Fonds National de la Recherche Scientifique* (Belgium).



the signal in that frequency-band. The evolution of the IF in different frequency-bands of the signal can thus be studied by means of the CWT coefficients.

Here, the complex Morlet wavelet is used [6]:

$$\psi_{\omega_c}(t) = C e^{-i\omega_c t} \left[e^{-\frac{t^2}{2\sigma_t^2}} - \sqrt{2} e^{-\frac{\omega_c^2 \sigma_t^2}{4}} e^{-\frac{t^2}{\sigma_t^2}} \right]. \quad (4)$$

The scale λ of the wavelet is determined by the central frequency $f_c = \frac{\omega_c}{2\pi}$, which is the frequency of oscillation of the wavelet. The product $\omega_c \sigma_t$ fixes the link between the width of the envelope of the wavelet and its oscillation frequency f_c and must be constant for a wavelet family. C normalizes the energy. The effective duration of the wavelet can be defined as $2\sigma_t$.

2.2. Application to formant frequency estimation

Using continuous wavelet transforms, the amplitude and instantaneous frequency can be calculated for different frequency bands of the speech signal. In the neighbourhood of the wavelet frequencies that fit best the cyclicity of a signal, the amplitude of the CWT coefficients presents a maximum and the instantaneous frequencies obtained by means of the CWT phase coefficients are very close to the cyclicity of the signal. The vocal frequency of a speech signal can thus be obtained [7]. For smaller scales, if the frequency of a formant lies in the frequency band of the wavelet, the resulting instantaneous frequency will be very close to the formant frequency. Instantaneous values of the formant frequencies can thus be obtained. Using the instantaneous frequency gives a better frequency resolution than the frequency step of the CWT evaluation [7].

The formant frequency estimation is performed as follows:

First, characteristic values of the first three formant frequencies are estimated by means of a traditional method.

Then, instantaneous formant frequencies are estimated. Three distinct CWT are calculated around each characteristic formant frequency, with different values of the parameter $\omega_c \sigma_t$. For the first formant, $\omega_c \sigma_t$ is chosen such that the effective durations of the wavelets are shorter than one glottal cycle. For the second formant, the frequency resolution must be high enough to dissociate the second formant from the first. For the third formant, the wavelet bandwidth is chosen equal to 800Hz . The instantaneous formant frequency traces are given by the instantaneous frequency based on the phase of the CWT coefficients whose amplitudes are at a maximum in the interval around each formant.

Finally, the instantaneous formant frequency traces are sampled, in order to isolate the values of the formant frequencies corresponding to the closed phase of the glottis for each cycle. The sampling is based on the detection of the instant of glottal closure. This is achieved by the detection of the instants of maximum energy of the first formant. As the wavelet corresponding to the sampling instant should be situated in the closed phase, the instantaneous formant frequency trace is sampled 1.5ms after each F_1 energy maximum.

3. Synthetic speech

In this section, results illustrating the performance of the formant estimation method for synthetic speech are presented. The aim is to show and understand the limits of the method, by decomposing the difficulties encountered in real speech signals.

The synthetic signals are sustained vowels [a]. They are obtained by a source - vocal tract model. The source signal is given

by the time-derivative of the glottal flow model of Liljencrants and Fant. The vocal tract is obtained by a cascade of time-varying second-order IIR filters, each filter modelling one formant. To model the source-vocal tract interaction, the bandwidth of the formants is modulated in synchrony with the source. Two different values of the formant bandwidth characterize thus the open phase and the closed phase of the glottis.

The influence of the parameters of the synthetic signal on the estimated formant frequencies is investigated by means of the following cases:

- constant vocal frequency F_0 , constant formant frequencies,
- constant F_0 , variable formant frequencies,
- variable F_0 , constant formant frequencies.

The performance is evaluated by the quality of the tracking of the formant frequency variations, and not by the vicinity between the estimated formant frequency and the reference frequency.

3.1. Constant F_0 , constant formant frequencies

Fig. 1 to Fig. 3 illustrate the estimation of the formant frequencies for a synthetic signal, with F_0 equal to 120Hz and the formant frequencies equal to 700Hz , 1200Hz and 2500Hz . The bandwidth of all formants are equal to 100Hz and 150Hz , respectively for the closed and open phase of the glottis.

Fig. 1 shows the synthetic speech signal, together with the amplitude of its CWT, for $\omega_c \sigma_t = 10$. The stars mark the estimated formant frequencies. Peaks can be seen at the formant frequencies, and one can also clearly see the excitation instant, where more energy is present at every frequency.

Fig. 2 shows the amplitude and the instantaneous frequency of the CWT as functions of the wavelet central frequency, for a given moment. The dotted lines mark the formant frequency references. The vertical lines show the central frequencies of the wavelets for which the CWT amplitude presents a maximum. One can see the plateaus of the instantaneous frequency at the formant frequencies, by means of which the precision of the estimates is obtained.

The choice of the sampling instants is illustrated on Fig. 3. The energy and instantaneous frequency traces of the three formants are shown. The energy trace of the first formant presents one maximum per glottal cycle. The diamonds mark the estimated formant frequencies. The variation amplitude of the estimates is lower than 0.2Hz . The variations obtained for other value of the vocal frequency and of the formant frequencies have the same order of magnitude.

3.2. Variable formant frequencies

To test the effect of formant frequency variations, synthetic signals have been generated with one linearly varying formant frequency.

Table 1 shows results for synthetic signals, with the frequency of the first formant F_1 varying from 700Hz to 725Hz . Two values of F_0 and two values of the frequency of the second formant F_2 are considered: F_0 is equal to 100Hz or 125Hz , and F_2 is equal to 1100Hz or 1200Hz . F_3 is equal to 2500Hz . The first part of Table 1 shows the distance between the extreme deviations of the F_1 estimates from their references. The second part of Table 1 shows the distance between the extreme estimates of F_2 and F_3 .

Table 1 shows that the F_1 estimates track well the reference. The error is lower than 1Hz , which is slightly higher than the error for a synthetic signal with constant formants. Table 1 also shows that the F_2 estimates are influenced by the variations of F_1 . This

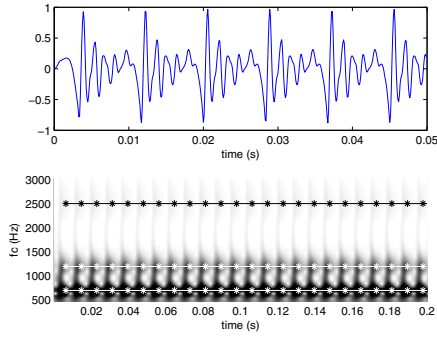
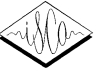


Figure 1: Synthetic speech signal and CWT amplitude for $\omega_c \sigma_t = 10$. The stars mark the estimated formant frequencies. High amplitudes are represented in black, low amplitudes in white.

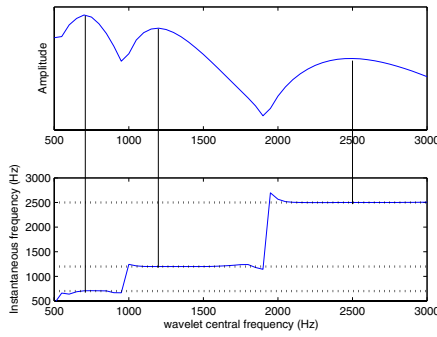


Figure 2: CWT amplitude and instantaneous frequency as functions of the wavelet central frequency, for a given moment. The dotted lines mark the formant frequency references. The vertical lines show the central frequencies of the wavelets for which the CWT amplitude presents a maximum.

effect increases when F_1 and F_2 get closer, or when F_0 is higher. The F_3 estimates are not influenced by the variations of F_1 .

The results for similar simulations with variations of F_2 or F_3 lead to the same conclusions: First, when F_0 is higher, there are more variations in the distance between the varying formant and its reference. Second, when the varying formant is closer to the estimated formant, the latter varies more.

3.3. Variable vocal frequency

To test the effect of the variation of the vocal frequency F_0 , synthetic signals have been generated with F_0 varying linearly.

Table 2 shows the span of the estimated formant frequencies for a synthetic signal with F_0 varying linearly between 95Hz and 105Hz. The reference formant frequencies are 700Hz, 1200Hz and 2500Hz. The estimated formant frequencies are not perfectly stable and vary with the vicinity between the formant frequencies and the harmonics of F_0 . This effect is more marked for lower formant frequencies, but stays inferior to 2Hz.

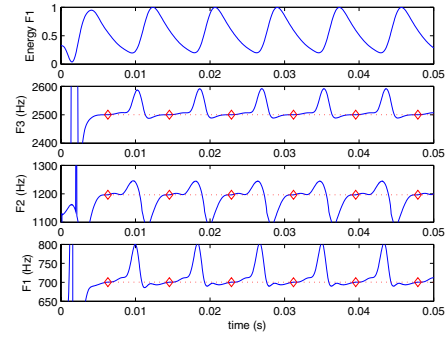


Figure 3: Instantaneous energy and frequency of the three first formants. The diamonds mark the estimated formant frequencies.

$\Delta(F_1 - F_{1,ref}) = (F_1 - F_{1,ref})_{max} - (F_1 - F_{1,ref})_{min}$		
	F0=100Hz	F0=125Hz
F2=1100Hz	0.25Hz	0.12Hz
F2=1200Hz	0.45Hz	0.97Hz
$\Delta F_2 = F_{2,max} - F_{2,min}$		
	F0=100Hz	F0=125Hz
F2=1100Hz	1.40Hz	5.41Hz
F2=1200Hz	0.18Hz	0.65Hz
$\Delta F_3 = F_{3,max} - F_{3,min}$		
	F0=100Hz	F0=125Hz
F2=1100Hz	0.008Hz	0.01Hz
F2=1200Hz	0.005Hz	0.03Hz

Table 1: Precision of the formant frequency estimates for synthetic speech, with F_1 varying linearly, for two values of F_0 and F_2 .

4. Application to real speech

The formant frequency estimation method has been applied to real speech signals. The performance of the method is illustrated for a speech signal from a normophonic speaker. Then results obtained for a speaker with Parkinson's disease and a normophonic speaker are presented and compared.

Fig.4 and 5 illustrate the formant frequency estimation method for a real speech signal. Fig. 4 shows a sustained vowel [a], together with the amplitude of its CWT, for $\omega_c \sigma_t = 10$. The stars mark the estimated formant frequencies. Fig. 5 shows F_0 and the estimated formant frequencies for the same sustained vowel, for a longer time interval. One can see that the formant frequency variations are much larger than the precision of the estimation method presented in section 3. The formant estimation method is thus able to track properly the involuntary variations of the formant frequencies for real speech.

$\Delta F = F_{max} - F_{min}$	
ΔF_1	1.4Hz
ΔF_2	1.7Hz
ΔF_3	0.03Hz

Table 2: Distance between the extreme formant frequency estimates for synthetic speech, with F_0 varying linearly from 95Hz to 105Hz.

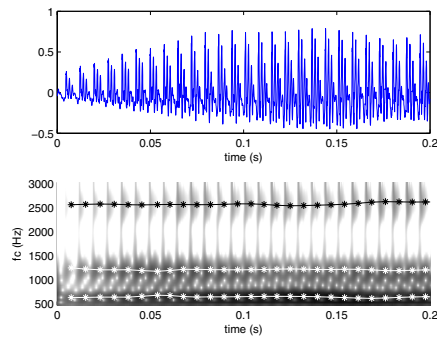


Figure 4: Real speech signal and CWT amplitude for $\omega_c \sigma_t = 10$. The stars mark the estimated formant frequencies. High amplitudes are represented in black, low amplitudes in white.

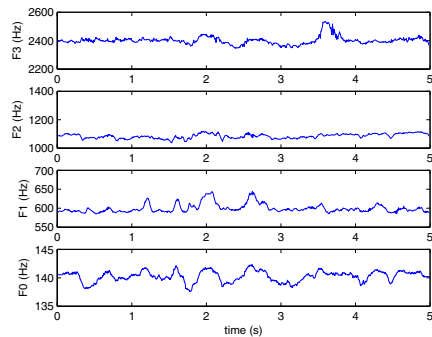


Figure 5: F_0 and first three formants for a real speech signal.

The formant frequency estimation has been applied to speakers with Parkinson's disease and control speakers. Differences are expected between both groups: A former study on the vocal tremor of F_0 has shown that the F_0 tremor features differ for parkinsonian and control speakers [8]. More tremor energy was observed at higher frequencies (7 – 15 Hz) for parkinsonian speakers, the cause being probably the presence of tremor of the laryngeal muscles. Likewise, a tremor of the articulators is expected for parkinsonian speakers and should be perceived in the formant frequencies.

Fig. 6 shows F_0 and the first three formants for a parkinsonian speaker. For this speaker, the standard deviations of the formant frequencies are 13 Hz, 53 Hz and 25 Hz. For the normophonic speaker on Fig. 5, the standard deviations of the formant frequencies are 14 Hz, 32 Hz and 18 Hz. Except for the second formant, these values don't differ significantly. The same result was obtained for the F_0 tremor amplitude in [8].

To characterize the frequency of the formant frequency variations by a unique value, the centre of gravity of the variation spectrum has been calculated in the frequency-band 3 Hz to 15 Hz. The results are [7.1 Hz – 7.5 Hz – 9.1 Hz] and [6.5 Hz – 5.8 Hz – 7.3 Hz], for the three formants of the parkinsonian and the normophonic speaker, respectively. The tremor frequencies are higher for the parkinsonian speaker than for the normophonic speaker. This result supports the hypothesis that the formant frequency traces of parkinsonian speakers present more high frequency variations. It

corroborates the previous results observed in the F_0 tremor in [8].

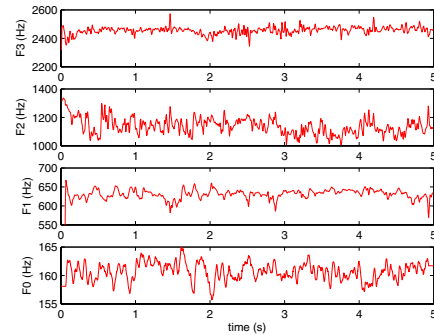


Figure 6: F_0 and first three formants for a parkinsonian speaker.

5. Conclusions

A formant frequency estimation method was proposed, with a view to track formant frequency variations due to vocal tract tremor. It is based on the instantaneous frequency obtained by means of a continuous wavelet transform and is synchronised on the glottal cycle. Results on synthetic and on real speech signal showed that the precision of the method is sufficient to track the formant frequency variations. Preliminary results for parkinsonian and normophonic speakers show differences between the two groups. Further investigations are being made on a larger corpus.

6. References

- [1] Pr. Rao and A. D. Barman, "Speech formant frequency estimation : evaluating a nonstationary method," *Signal Processing*, vol. 80, no. 8, pp. 1655–1667, 2000.
- [2] B. Yegnanarayana and R.N.J. Veldhuis, "Extraction of vocal tract system characteristics from speech signals," *IEEE trans. on speech and audio processing*, vol. 6, no. 4, pp. 313–327, july 1998.
- [3] B. Boashash, "Estimation and interpreting the instantaneous frequency of a signal - part i : Fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520 – 539, 1992.
- [4] T. Le-Tien, "Some issues of wavelet functions for instantaneous frequency extraction in speech signals," *Proc. IEEE Tencon 1997*, pp. 31–34, 1997.
- [5] St. Mallat, *A Wavelet Tour of Signal Processing*, San Diego: Academic Press, 2nd edition, 1999.
- [6] D.B. Percival and A.T. Walden, *Wavelet methods for time series analysis*, Cambridge University Press, 2000.
- [7] L. Cnockaert, F. Grenez, and J. Schoentgen, "Fundamental frequency estimation and vocal tremor analysis by means of morlet wavelet transforms," *Proc. ICASSP, Philadelphia (USA)*, pp. 393–396, 2005.
- [8] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, and F. Grenez, "Effect of parkinson's disease on vocal tremor," *Proc. MAVEBA 2005, Florence (Italy)*, pp. 177–180, 2005.