ANALYSIS OF PITCH-SYNCHRONOUS MODULATION EFFECTS BY USING ANALYTIC FILTERS

Unto K. Laine
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O. Box 3000, FIN-02015 Espoo, Finland
e-mail: Unto.Laine@hut.fi

ABSTRACT

Gabor type of analytic filters are used together with Teager-Kaiser energy operator based DESA-1 and two other algorithms to study pitch-synchronous modulation effects in speech. The methods are tested with synthetic and natural speech. Common problems and limitations of these methods are discussed.

1. INTRODUCTION

During the last years an increasing number of speech analysis results have shown new and partially surprising outcomes related to nonlinear, time-varying aspects of speech. Interesting results are reached by applying so called Teager-Kaiser (TK) energy operator to analyze the pitch-synchronous AM-FM modulation effects of formants in voiced speech segments [1, 2]. The method is even applied to a speech analysis-synthesis system [3, 4] which according to the authors has shown that the pitch-synchronous modulation effects in formant parameters are perceptually important and will improve the naturalness of reconstructed speech.

Maragos et al. have published a review on the TK operator and on its application to instantaneous format amplitude (AM) and formant frequency (FM) estimation [2]. Their paper contains analysis examples of synthetic and natural speech and discusses some of the limitations of the method.

However, the application of the TK-operator to speech analysis has gained also such results which probably can not be explained on the basis of the present knowledge of speech acoustics, e.g., Foote et al. [3] have pointed out that the usage of the TK operator may lead to artifacts which are not properties of the analyzed speech. Some results even contradict those gained by widely used 'standard' methods, like adaptive inverse filtering, different variants of linear prediction (LP) or time-frequency methods, like auditory filterbanks.

An auditory filterbank with frequency resolution chosen especially for speech analysis (frequency resolution not higher than the formant bandwidths and time resolution the highest possible within that frequency resolution) and with analytic channel responses provides possibilities to compute instantaneous frequency and amplitude for each channel at each time instant and thus follow even the details of many pitch-synchronous effects [6, 7].

These methods have revealed interesting pitch-synchronous details in vowel sounds. Some of these modulation effects are probably related to the subglottal resonances. Also some nonlinear effects, like the secondary excitation by turbulent flow at the glottal opening, can be monitored.

The surprising results gained by the TK operator and also some new outcomes of the novel auditory filterbank designed for speech analysis motivated this study. The aim was to get a closer picture of the nature of the tools and also to try to explain and understand some of the results from the speech acoustics point of view.

This paper shortly reviews the TK operator and the related *discrete energy separation algorithm,* DESA-1. The algorithm is designed for analysis of the AM-FM properties of a signal generated by a second order resonator system (e.g., harmonic oscillator, formant resonator or, complex exponential).

The algorithm is tested by using synthetic as well natural speech samples. A simple *analytic resonator model* is discussed and compared with the DESA-1. The analytic model seems to work in a more stable and robust way than the DESA-1. Both methods will provide new interesting results which can (at least partially) be explained based on the present knowledge of speech production.. However, both methods may lead to serious artifacts when not applied with care. A very demonstrative case is reported below.

Many questions are still left open and need a lot of multidimensional, multimethodic and multidisciplinary work before a better and deeper understanding is reached.

2. THE ENERGY OPERATOR

The continuous time Teager-Kaiser energy operator is defined by

$$\Psi_c[x(t)] = (dx/dt)^2 - x(t)d^2x(t)/dt^2 \qquad (1)$$

The operator can track the total energy of a simple harmonic oscillator (e.g., a mass-spring system) which generates an undamped sinusoidal signal with an amplitude A and frequency w. In other words $Y[A\cos(wt + q)] = (Aw)^2$ = the total energy (per half-unit mass) of the oscillating system. It has also been shown that the operator can estimate the instantaneous amplitude $a(t)$ and the instantaneous frequency *w*(*t*) of a time-varying second order system [8].

The discrete time TK energy operator is defined by

$$\Psi_d[x(n)] = x^2(n) - x(n-1)x(n+1) \qquad (2)$$

and finally the DESA-1 algorithm by:

$$y(n) = x(n) - x(n-1)$$

$$\Omega(n) \approx \arccos(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[y(n)]}) \qquad (3)$$

$$|a(n)| \approx \sqrt{\frac{\Psi[y(n)]}{1-\cos^2(\Omega)}}$$

Interestingly, the discrete TK operator comes out also when solving the instantaneous parameters (a_1, a_2) of a second order recursive system.

$$x(n) = a_1\, x(n-1) + a_2\, x(n-2)$$

$$x(n+1) = a_1\, x(n) + a_2\, x(n-1) \qquad (4)$$

we get

$$a_1 = \frac{\Sigma[x(n)]}{\Psi[x(n-1)]}, \quad a_2 = \frac{\Psi[x(n)]}{\Psi[x(n-1)]} \qquad (5)$$

$$\Sigma[x(n)] = x(n)x(n-1) - x(n+1)x(n-2)$$

where a new operator S[x(n)] called *the spread energy operator* is introduced. It tracts the energy by using four samples of *x*(*n*) whereas the TK operator needs three samples.

The instantaneous *r*(*n*) which controls the bandwidth is now given by r(n)=Sqrt[a_2] and the instantaneous frequency by *w*(*n*)= arccos[a_1/(2 *r*(*n*))]. In the case of harmonic oscillator *r*(*n*) equals one. Also this algorithm was tested and it works about as well as DESA-1.

3. TEST WITH SYNTHETIC SIGNALS

In order to better understand the results given by the TK operator we simulated the speech production with a simple glottal (flow) pulse model which was connected to the first formant resonator filter. The sampling frequency was 22.05 kHz. The glottal excitation pulse had the shape given by the polynomial $g(n) = 9.0*((n/c)^2 - (n/c)^3)$. This part of 95 samples was followed by 105 zeroes to model the closed phase of the glottis. The formant resonator was tuned to 750 Hz and had a constant *r*-value of 0.98.

Figure 1. illustrates the time waveform of the resonator output, its Fourier spectra, the instantaneous *r* and frequency given by a simple analytic model, and finally the result given by the DESA-1 algorithm. Both methods used the same signal, the first in the analytic (complex) form and the second its real component only. A Gaussian window was used to emphasize the frequency area of the first formant and to create an analytic Gabor type filter for the first algorithm.

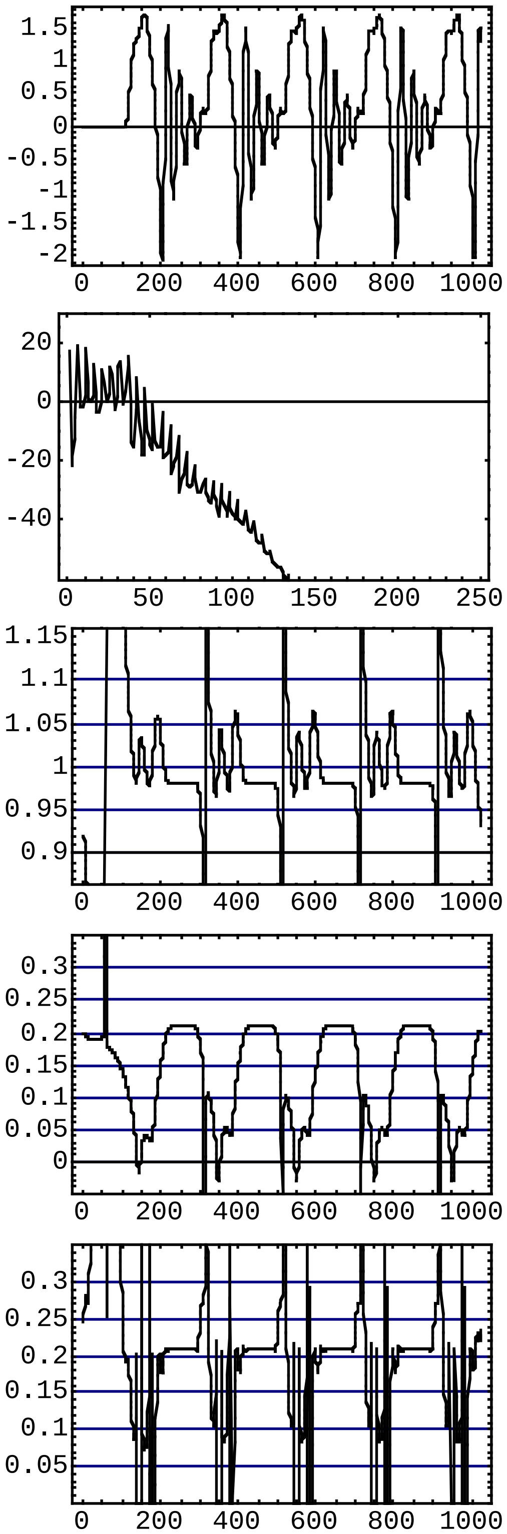

Figure 1. Analysis of synthetic speech (see text).

The analytic model assumes that the output is of the form *z*(*n*) = *r* exp[I w_0] *z*(*n*-1). Now the instantaneous *r* which is related to the instantaneous bandwidth and the instantaneous w_0 can be easily solved.

$$r(n) = abs[z(n)/z(n-1)]$$

$$\omega_0(n) = \arg[z(n)/z(n-1)] \qquad (6)$$

The instantaneous amplitude is simply $A(n) = \mathrm{abs}[z(n)]$. The analytic signal $z(n)$ is easily produced from the Fourier transformed signal by windowing it with a Gaussian window. The window must be placed one-sided (not symmetrically in the positive and negative frequencies) at the formant frequency to be analyzed. Finally, the $z(n)$ signal is created by inverse Fourier transform.

When the r-value is below one, the system is stable. This occurs during the closed glottal phase. However, the rapid glottal opening gives a sharp excitation which drives the (model) system instantaneously to an unstable state. The instabilities continue during the whole open period and the model shows parametric oscillations. The oscillation frequency equals approximately the formant frequency. During the closed phase the model gives good estimates for the actual r-value.

The instantaneous frequency of the model varies pitch-synchronously. During the closed period, when the resonator is autonomic (runs purely based on its internal state and energy) both methods give correct values for the instantaneous frequency. However, during the open period the models lose the tract and are not able to follow the formant frequency. The reason to this is that the formant is damped and the broad frequency domain window is not able to filter out the strong low-frequency components of the glottal excitation. Thus the model starts to follow the strongest signal components. This is a clear example of a possible artifact related to the methods. These artifacts should not be interpreted as a true FM in the formant frequency.

The lowest frame in Fig. 1 illustrates the unfiltered (raw) data given by the DESA-1. We noted that the noiseness of the DESA-1 result can be reduced by analyzing both the real and imaginary component of $z(n)$ and averaging the results.

4. ANALYSIS OF A VOWEL

The speech material (isolated vowels) produced by one male speaker was carefully recorded in an anechoic chamber by using a B&K condenser microphone and a preamplifier. The recordings were made with digital tape recorder. The signals are phase linear from about 20Hz upwards.

Figure 2. illustrates the results given by the analytic model. The instantaneous r has similar strong oscillations during the open period as what was seen in the simulation. The oscillation frequency practically equals the first formant frequency. The closed period is clearly shorter than in the simulation and it also includes oscillations. These are related to the second formant which is not damped enough because the frequency domain window is broad. The r-value drops clearly at the glottal opening before the strong first-formant oscillations start.

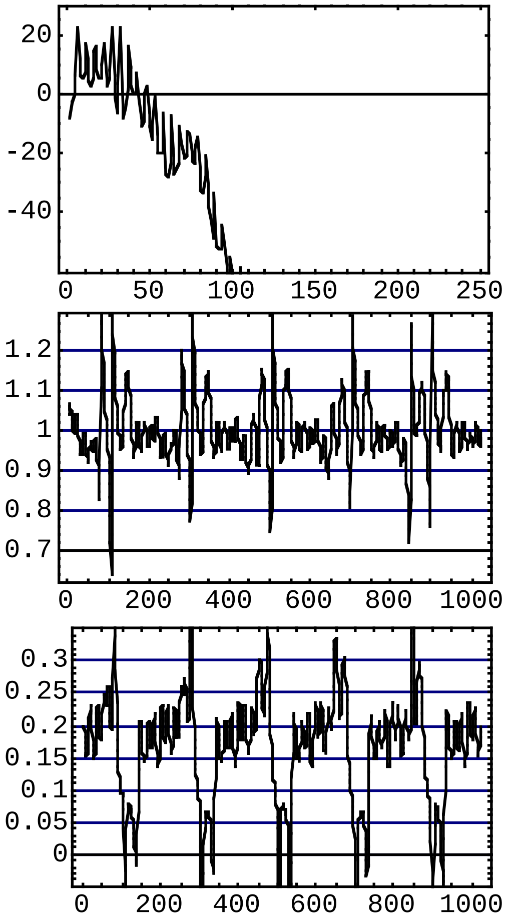

Figure 2. The analysis of Finnish /ae/ (see text).

The strongest oscillation in the open period occurs about in the middle of it which tells that the opening is much softer than in the case of simulation.

The instantaneous frequency (lowest frame) includes quite strong oscillations as well. Both formant frequencies are present and modulated pitch-synchronously. During the strongest glottal excitation the model follows the glottal waveform, not the first formant. The open glottal period is clearly shorter than in the simulation.

Finally, the model proposes a linear, pitch-synchronous frequency modulation for the first formant during the closed period. Is this a fact or an artifact? Speech acousticians have discussed quite much about the pitch-synchronous vertical movements of the tissues around the glottal orifice during the phonation. This could explain at least a part of the FM. The detected sweep happens to run in the right direction: the vocal folds are closed at their lower parts first and opened at their highest parts. When the simultaneous vertical movements of the tissues are included, the vocal tract is longer at the glottal closure than at the opening. Thus there should be some pitch-synchronous FM.

The results of the analysis of real data can be well understood when compared to the simple simulation made. The results given by the analytic model can be further processed in order to filter out the irrelevant oscillations.

5. SOME PROBLEMS AND ARTIFACTS

Most of the problems found during this study where related to the shape and position of the Gaussian window. When the /ae/ vowel was analyzed with a relatively narrow (500 Hz) window positioned around 350 Hz a strong pitch-synchronous AM was found (see Fig. 3.). The sinusoids of 350 Hz were present only during the open glottal period. This awaked a question, could there be such a strong subglottal resonance which radiates energy only during the open periods (amplitude modulated by the time waveform of the glottal opening)?

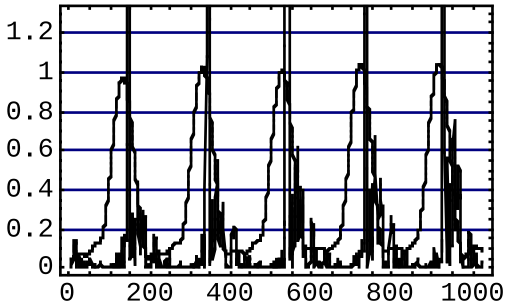

Figure 3. An artifact, 'subglottal resonance' seen just before the glottal closure analyzed by the TK energy operator.

In order to answer this question two synthetic speech samples one including the 'subglottal resonance' were generated and analyzed with an auditory filterbank. The produced auditory time-frequency distributions were compared to the distribution of the natural vowel. The comparison gave a clear result: there is no subglottal resonance around that frequency (see Fig. 4.).

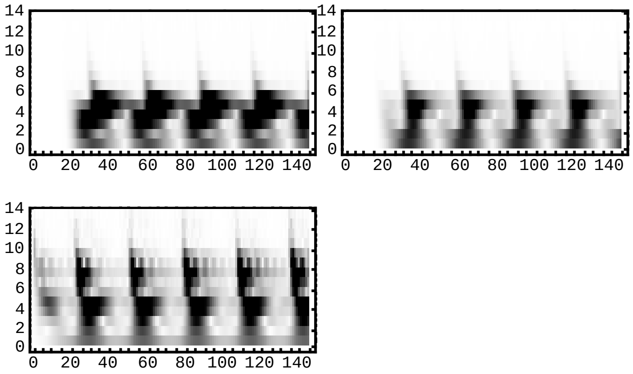

Figure 4. Auditory filterbank comparision of synthetic vowel with 'subglottal resonance' (upper left frame), without it (upper right frame) and the natural /ae/ vowel. (14 ERB-rate channels 130 Hz - 8.6 kHz, time step 0.63 ms)

When the analysis window (filter) in the frequency domain is too narrow it can easily create strong FM effects even though the analyzed synthetic vowel do not have any. The relatively simple reason to this is that the tested models are based on the second order system only. A new filter positioned around just those frequencies to be analyzed will always corrupt the analysis result more or less. After the filtering the energy operator is tracking a system of higher order than two. Thus the energy of the included filter will finally totally confuse the operator when the bandwidth of the filter approaches that of the formant. This is the reason why all the analysis results have a FM jump around the glottal closure. The sharp closure will excite the added filter as well and its impulse response will corrupt the instantaneous frequency estimate of the formant.

6. ACKNOWLEDGEMENT

The topic treated in this paper was pointed out by Dr. Paavo Alku. Many coworkers at the HUT Acoustics laboratory have helped this study. Mr. Aki Härmä and Mr. Tero Tolonen made many good comments on the text. The author is grateful to all those involved. This study has been supported by the Academy of Finland.

7. REFERENCES

[1] J. F. Kaiser, "*On Teager's energy algorithm and its generalization to continuous signals*," Proc. 4th IEEE DSP Workshop, Mohonk, New Paltz, NY, Sept., 1990.

[2] Maragos P., Kaiser J. F., Quatieri T. F., '*Energy separation in signal modulations with applications to speech analysis*", IEEE Trans. on Signal Processing, vol. 41, pp. 3024-3051, Oct. 1993.

[3] Foote T. J., Mashuo D. J., Silverman H. F, *"Stop classification using DESA-1 high resolution formant tracking",* Proc. of ICASSP-93, II, pp. 720-723.

[4] Hanson H. M., Maragos P. and, Potamianos A., "*Finding speech formants and modulations via energy separation: with application to a vocoder*," Proc. of IEEE ICASSP-93, pp. II-716-719, Minneapolis, Minnesota, 1993.

[5] Potamianos A., Maragos P., '*Speech analysis and synthesis using an AM-FM modulation model*", Proc. of Eurospeech'97, pp. 1355-1358, Rhodes, Greece, 1997.

[6] Laine U. K., *"Speech analysis using complex orthogonal auditory transform (COAT)."* Proc. of 1992 Int. Conf. on Spoken Language Processing,, Banff, Alberta, Canada, vol. 1, pp. 69-72.

[7] Laine U. K., '*Critically sampled PR filterbanks of nonuniform resolution based on block recursive FAMlet transform*", Proc. of Eurospeech'97, pp. 697-700, Rhodes, Greece, 1997.

[8] Maragos P., Kaiser J.F., and Quatieri T. F., "On amplitude and frequency demodulation using energy operators," IEEE Tr. Signal Processing, **41**, pp. 1532-1550, Apr. 1993.