Proceedings of 1990 IEEE Digital Signal Processing Workshop, New Paltz, NY, Sep. 1990

Detecting Nonlinearities in Speech using an Energy Operator[0]

Petros Maragos, Division of Applied Sciences, Harvard University, Cambridge, MA 02138
Thomas F. Quatieri, MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02173
James F. Kaiser, Bell Communications Research, 445 South St., Morristown, NJ 07960

In his work on nonlinear modeling of speech production, Teager [1] used the following energy operator E on speech-related signals $x(n)$:

$$E[x(n)] = x^2(n) - x(n-1)x(n+1) \tag{1}$$

Kaiser [2] has analysed E and shown that it yields the energy of simple oscillators that generated the signal $x(n)$; this energy measure is a function of the frequency as well as the amplitude composition of a signal. Kaiser also showed that E can track the instantaneous frequency in single sinusoids and chirp signals, possibly exponentially damped. Teager [1] applied E to signals resulting from bandpass filtering of speech vowels in the vicinity of their formants. The output from the energy operator frequently consisted of several *pulses* per pitch period, with decaying peak amplitude. Teager suggested that these energy pulses indicate modulation of formants caused by nonlinear phenomena such as rapidly varying separated air flow in the vocal tract.

In our work we interpret these energy pulses by using a frequency modulation (FM) model for the time-varying formants. Specifically, consider the following *exponentially-damped FM signal with sine modulation*

$$x(n) = Ae^{-an}\cos[\phi(n)] = Ae^{-an}\cos[\Omega_0 n + \beta\sin(\Omega_m n) + \theta] \tag{2}$$

where Ω_0 is the center (or carrier) frequency, $\beta = \Delta/\Omega_m$ is the modulation index, Δ is the frequency deviation, and Ω_m is the frequency of the modulating sinusoid. The instantaneous frequency is $\Omega(n) = d\phi(n)/dn = \Omega_0 + \Delta\cos(\Omega_m n)$. If Ω_m is sufficiently small so that $\cos(\Omega_m) \approx 1$ and $\sin(\Omega_m) \approx \Omega_m$, and we apply the energy operator E to $x(n)$, we obtain

$$E[x(n)] \approx A^2e^{-2an}\sin^2[\Omega_0 + \Delta\cos(\Omega_m n)] = A^2e^{-2an}\sin^2[\Omega(n)]. \tag{3}$$

Thus E can track the instantaneous frequency $\Omega(n)$ of FM-sine signals. Figs. 1a,b,c show the (square root of the) energy operator ouput $\sqrt{E[x(n)]}$ where $x(n)$ is the FM-sine signal in (2) with $A = 10, a = 0.002$, $\Omega_0 = 0.2\pi$, $\Omega_m = 0.02\pi$, $\theta = 0$, and (a) $\Delta = \Omega_0$, (b) $\Delta = 0.2\Omega_0$ and (c) $\Delta = 0.01\Omega_0$. The outputs of Fig. 1 due to input synthetic FM-sine signals correspond roughly to measurements made on actual speech using E. The instantaneous frequency $\Omega(n)$ plays the role of a time-varying formant. Thus, if we view Ω_0 as the center value of some "formant", then the operator E followed by an envelope detector, followed by dividing with the envelope and inverse square sine will yield the time-varying formant (within a single pitch period) as the instantaneous frequency $\Omega(n)$ of the FM signal. Fig. 2 shows (a) a segment of a speech vowel /e/ sampled at $F_s = 30$ kHz and (b) the output from $\sqrt{E}$ when applied to a band-pass filtered version of (a) extracted around a formant at $F_0 = 3400$ Hz using a Gabor filter with impulse response $\exp(-b^2n^2)\cos(\Omega_0 n)$, where $\Omega_0 = 2\pi F_0/F_s$ and $b = 1000/F_s$. Fig. 3 is similar to Fig. 2 but for a sustained vowel /a/ sampled at 30 kHz and for a 830 Hz frequency band extracted around 2660 Hz using the sine wave transform. For both Figs. 2 and 3 there are 2-4 pulses per pitch period, and the exponentially-damped sine squared model (3) can approximately explain the shape of these measured energy pulses. There have also been cases where we have observed only one major pulse

[0]This work was supported in part by the ARO under Grant DAALO3-86-K-0171 (Center for Intelligent Control Systems), in part by the NSF under Grant MIPS-86-58150 with matching funds from Bellcore, DEC, Sun, and Xerox, and in part by the Department of the Air Force.

per pitch period. These observations can be partially explained from the FM-sine model by comparing Figs. 1a,b,c. These figures show that as the frequency deviation Δ increases (compared with the center frequency Ω_0), the multiple sine pulses become more detectable. If Δ is very small, the exponentially-decaying envelope dominates and yields only one pulse per period.

References:

[1] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", *NATO Advanced Study Institute on Speech Production and Speech Modelling*, Bonas, France, July 1989; Kluwer Acad. Publ., Boston, MA, 1990.

[2] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", *Proc. ICASSP-90*, vol. 1, pp. 381-384, Albuquerque, NM, Apr. 1990.

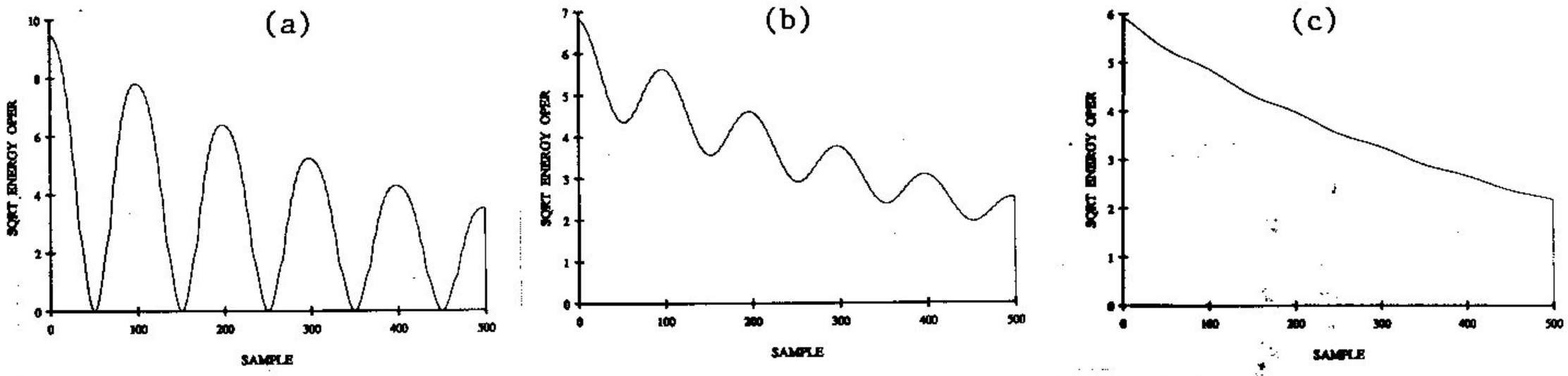

Fig. 1. Result of energy operator on synthetic FM signal. (a) 100% modulation; (b) 20% modulation; (c) 1% modulation.

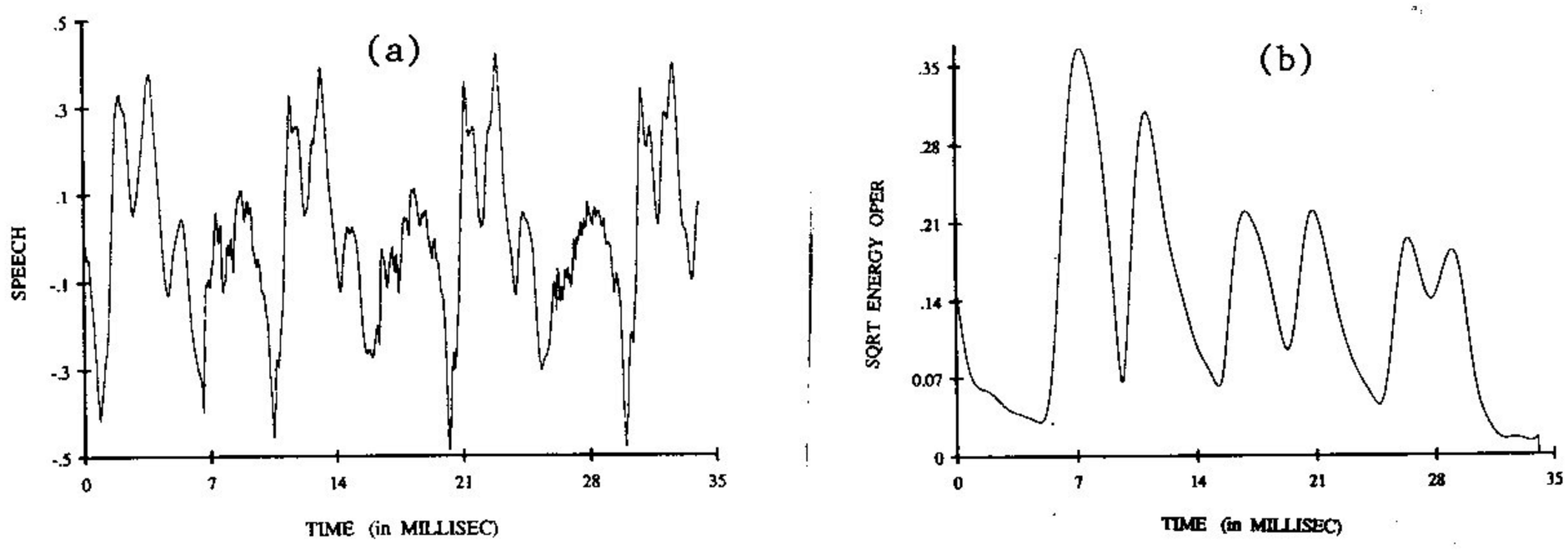

Fig. 2. Result of energy operator on speech vowel /e/. (a) original speech; (b) output of energy operator.

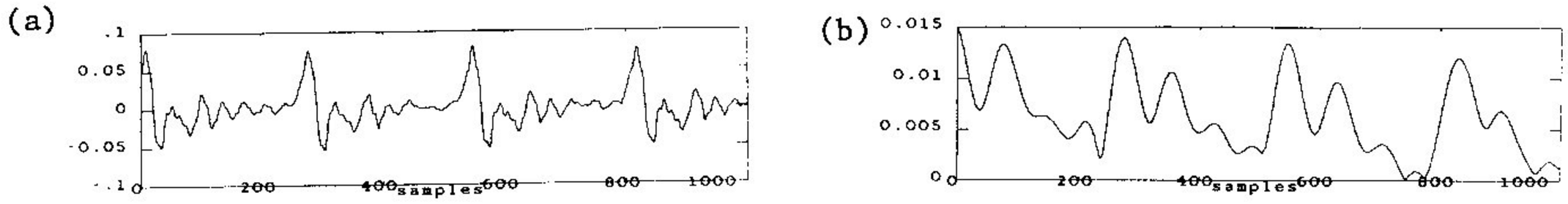

Fig. 3. Result of energy operator on sustained vowel /a/. (a) original speech; (b) output of energy operator.