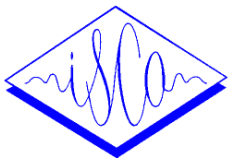


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Nonlinear Speech Analysis Using the Teager Energy Operator with Application to Speech Classification under Stress

Douglas A. Cairns and John H. L. Hansen

Robust Speech Processing Laboratory
Department of Electrical Engineering
Duke University, Box 90291
Durham, North Carolina 27708-0291

ABSTRACT

In this study, the problem of reliably classifying speech as normal or speech under stress is examined. It is hypothesized that speech production is composed of linear and nonlinear components, and that the nonlinear component changes measurably between normal and stressed speech. It is proposed that the Teager Energy operator could be utilized to quantify the change between normal and stressed speech. The stress styles considered are speech produced under loud, angry, Lombard effect and clear conditions. Results show that loud and angry speech can be reliably differentiated from neutral speech, while clear speech is difficult to differentiate from neutral speech. The results also show that Lombard effect speech can be reliably classified, but performance varies across speakers.

1. INTRODUCTION

In the course of a normal day, a person can experience stress in a number of ways. Emotional and physical stress are common influences in daily life. These stresses can be communicated to a listener through acoustic cues. Environmental stress due to background noise (called the Lombard effect [7]) can also be conveyed through acoustic cues. Researchers have studied acoustic phenomena such as fundamental frequency (F0), amplitude, concentration of spectral energy, formant location and bandwidth, and others in the hope of finding a reliable indicator of stress. The results of these studies have shown some promising acoustic indicators, although some of the results are contradictory.

Pitch has been the most common acoustic cue studied. Williams and Stevens [15, 16] performed several studies using simulated and actual speech under stress. Their work in simulated stress showed that each emotion had a distictive F0 contour. Their work using speakers under actual stress conditions indicated that F0 rose under stress. However, this finding was not duplicated by Hecker et al. [4] and Streeter et al. [12], who performed similar experiments on speakers under actual stress conditions.

*This work supported in part by a grant from The Whitaker Foundation.

Spectral properties such as energy content and formant bandwidth and location have also received attention. Both Pisoni et al. [10] and Stanton et al. [11] noted that spectral energy shifts to higher frequencies for consonants under the Lombard effect. Hansen showed that there are statistically significant changes in formant bandwidths and locations for both simulated and actual stressed speech [2]. While these spectral properties show definite trends for speech under stress, speakers tend to manifest stress in very individual ways, making it difficult to generalize how effective spectral properties are at predicting stress.

Since the previous acoustic studies of speech under stress have been inconclusive, it is suggested that there is another factor that conveys stress to a listener. It is suggested that speech production consists of a linear and a nonlinear component. It is hypothesized that the nonlinear component changes appreciably when a speaker is under stress. In this study, the nonlinear Teager Energy operator is utilized to measure the change between normal and stressed speech. The following sections will discuss the Teager Energy operator as well as an overall classification system for speech under stress.

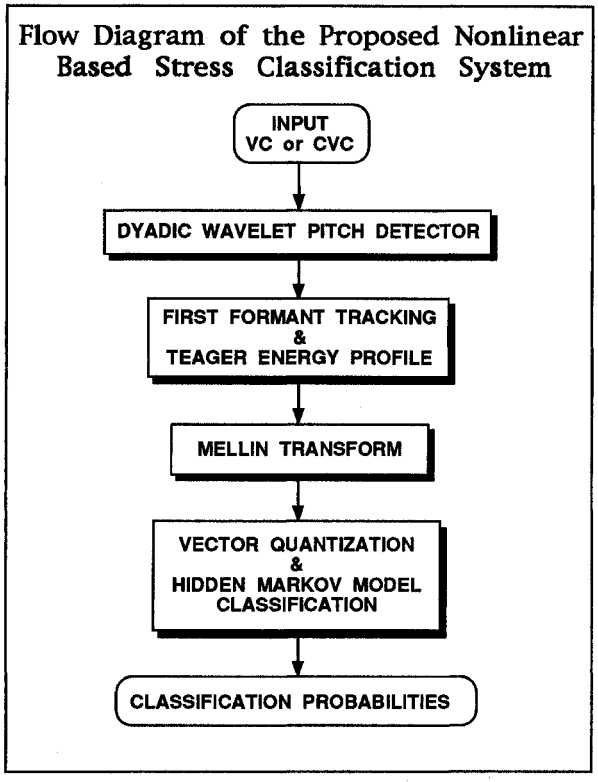
2. TEAGER ENERGY OPERATOR

The Teager Energy Operator, which provides an instantaneous estimate of the 'energy' of a signal, has the form [6]

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1). \quad (1)$$

The operator was first used by H. Teager [13] to show the modulation patterns of individual formants. Teager's experimental work in speech and hearing, along with the modulation patterns obtained by the Teager Energy operator, led Teager to propose a nonlinear model of speech production. Teager theorized that vortices are formed throughout the vocal tract, and these vortices modulate the airflow passing through the vocal tract, producing sound. Thomas [14] and McGowan [9] showed by simulation and theoretical calculation that vortices exist in the vocal tract. In addition, McGowan further demonstrated that the acoustic waveform is composed of sound generated by a linear source added to sound generated by a nonlinear source. The hypothesis suggested here builds on McGowan's model to conjecture that the sound generated

by the nonlinear source(s) changes measurably between normal and stressed speech. The Teager Energy operator has been used to measure the change in the modulation pattern of the first formant (Teager Energy profile) for normal and stressed speech. This measure has been experimentally shown to be successful in classifying normal and stressed speech [1].



3. ALGORITHM FORMULATION

Since the modulation pattern of the first formant has been observed to be quasi-periodic with a period corresponding to the pitch period, a system was developed to track the first formant for the duration of a vowel, and extract a measure of the modulation pattern for each pitch period. The key elements of this system are the wavelet-based pitch detector, the formant tracker, the Mellin transform parameterization, and the VQ/HMM classifier. Each of the elements is discussed in greater detail below.

Pitch Detector

For this system, it is necessary to calculate the location of a recognizable event during the pitch period. In this way, modulation patterns (Teager Energy profiles) can be compared across utterances with full confidence that the starting and ending points correspond to the same relative position in the pitch period. The pitch detector used was a derivative of the wavelet-based pitch detector of Kadambe and Boudreax-Bartels [5]. A two-pass version of this algorithm was found to be necessary to accurately mark time based pitch boundaries in stressed speech. This pitch detection algorithm estimates the glottal closure instant (GCI). Figure 2 shows a plot of a word and a corresponding plot of the GCI. Further implementation details of the pitch detector are discussed in [1].

Formant Tracker

The formant tracking algorithm is based on an AM-FM model for speech proposed by Maragos, Kaiser, and Quatieri [8]. In this model, each formant is modelled as an AM-FM signal. Maragos et al. also showed that the Teager Energy operator could be used to separate the AM and FM contributions (called the Energy Separation Algorithm (ESA)) by

$$f(n) \approx \frac{1}{2\pi T} \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right), \quad (2)$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{1 - (\frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]})^2}}, \quad (3)$$

where $y(n) = x(n) - x(n-1)$ represents the difference in adjacent speech samples, $\Psi[\bullet]$ is the discrete Teager energy operator, $f(n)$ is the FM contribution at sample n, and $a(n)$ is the AM contribution at sample n. To determine the AM and FM contribution of an individual formant, the speech signal must be bandpass filtered in the region of the formant since the Teager Energy operator is sensitive to multicomponent signals. Hanson, Maragos, and Potaminiaos [3] expanded on this concept and developed an iterative procedure based on equations 2 and 3 to track formants. The algorithm used a standard LPC formant tracker to provide an initial esitmate of a formant center frequency. The ESA equations were then applied and the new formant center frequency obtained from

$$f_c^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} f(n). \quad (4)$$

Here, N is the length of the speech segment, and $f_c^{(i+1)}$ is the estimated formant center frequency at iteration $i+1$. The new formant center frequency was used in conjunction with the ESA equations to refine the formant center frequency until it changed by less than 10 Hz. The appeal of this approach is that a good estimate of the formant center frequency is obtained and the Teager Energy profile is generated as a byproduct. Figure 3 shows typical Teager Energy profiles from voiced sections of VC utterances for normal and angry speech.

Mellin Transform

Since many researchers have shown that pitch changes between normal and stressed speech, it is necessary to normalize the effect of pitch to study the modulation pattern of the first formant. The Mellin Transform, having the form [17],

$$M\{f(e^x)\} = \int_{-\infty}^{\infty} f(e^x) e^{sx}\, dx. \quad (5)$$

was chosen because it is a scale invariant transform and was successful in a similar shape classification task involving ship radar signatures. Figure 4 shows the Mellin transform coefficients corresponding to the Teager Energy profiles shown in Figure 3.

VQ/HMM Classifier

A classifier was formulated that consists of a vector quantizer (VQ) with a 128 entry codebook (Euclidean distance measure), followed by a five state, left-to-right, discrete observation hidden Markov model (HMM). The output of the HMM was stored for processing off-line. The algorithm given in [1] was used to make the final stress/neutral decision. This approach was shown to achieve good performance, but it should be noted that other classifiers (neural networks, ML classifiers, etc.) are equally valid.

4. EXPERIMENT

Nine male speakers of English were selected from the *SUSAS* (Speech Under Simulated and Actual Stress) database [2]. Six VC (vowel-consonant) or CVC (consonant-vowel-consonant) words were chosen for evaluation purposes. For each word, there were twelve normal tokens, and two tokens of the following stress styles; loud, angry, Lombard effect, and clear. The Lombard effect was simulated by having speakers listen to 85 dBA SPL of pink noise played through headphones while speaking. Of the twelve normal tokens for a word, six were used to train the VQ/HMM classifier. The remaining normal data and all the stress data were used to evaluate the performance of the system.

5. RESULTS AND DISCUSSION

The results of the system evaluation are shown in Figure 1. Several conclusions can be drawn from this data. First, loud and angry speech can be reliably differentiated from normal speech. Second, clear speech can not be consistently differentiated from normal speech. Finally, Lombard effect speech can be distinguished from neutral speech, but performance varies across speakers. It may be necessary to utilize additional features such as spectral shape to reliably classify Lombard effect speech.

In this study it has been hypothesized that speech is composed of linear and nonlinear components, and that the nonlinear component changes measurably under stress. While the experimental framework cannot conclusively resolve this issue, a promising application of the Teager Energy operator has been shown. Beyond the current application, it is conceivable that this algorithm could be used as a front end for a speech recognition system. The classification algorithm could control the recognition model selected, depending on the result of the stress/normal decision. Other possible areas for applying some, or all, of this approach include improved speech coding/synthesis and the detection of vocal pathology.

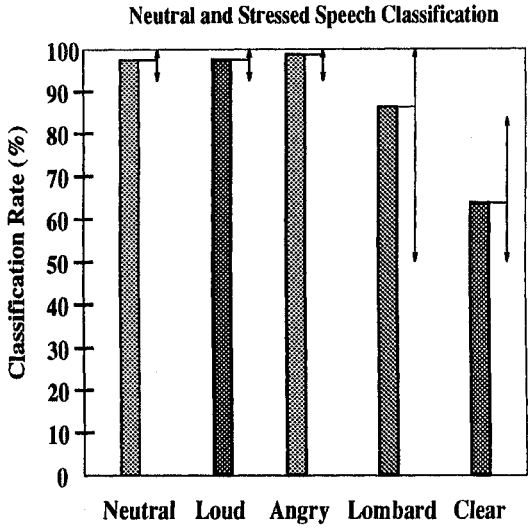


Figure 1: Mean classification rates across speaking styles.

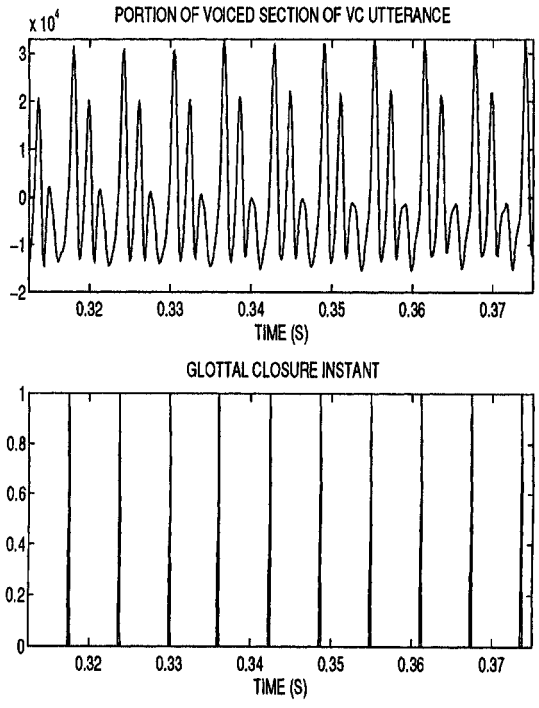


Figure 2: GCI for a portion of a VC utterance (impulses indicate location of GCI).

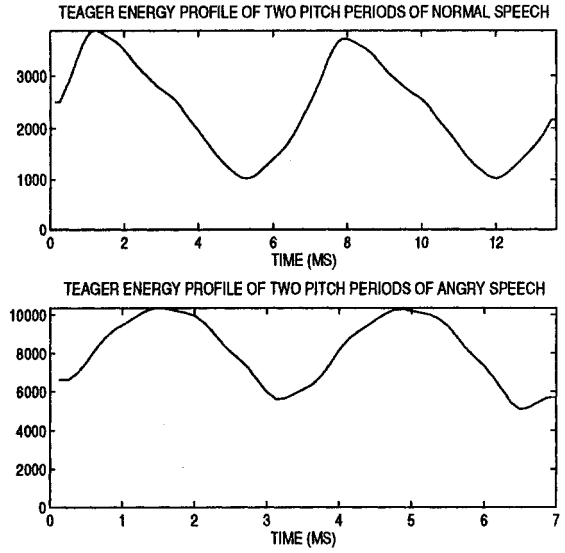


Figure 3: Teager Energy profile of two pitch periods of normal and angry speech.

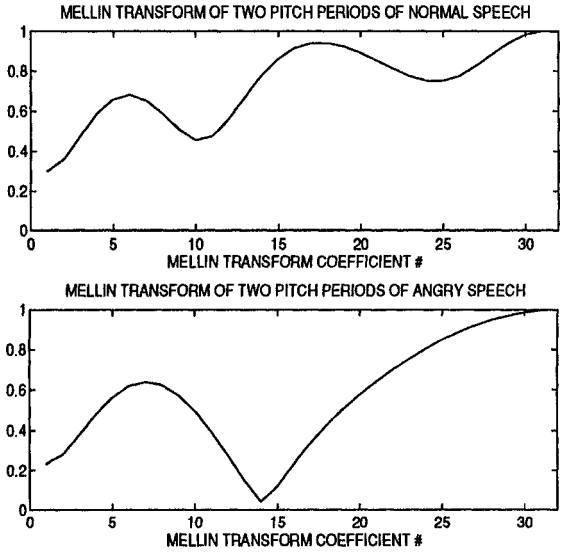


Figure 4: Mellin transform of Teager Energy profiles shown in Figure 3.

References

[1] D. A. Cairns and J. H. L. Hansen, "Nonlinear Analysis and Detection of Speech Under Stressed Conditions", submitted to *J. Acoust. Soc. Am.* , September 1993.

[2] J. H. L. Hansen, "Analysis and Compensation of Stressed and Noisy Speech With Application to Automatic Speech Recognition," Ph. D. Dissertation, Georgia Institute of Technology, 1988.

[3] H. Hanson, P. Maragos, A. Potamianos, "Finding Speech Formants and Modulations via Energy Separation: With Application to a Vocoder", *Proc. IEEE ICASSP-93*, vol. 2, pp. 716-719, 1993.

[4] M. H. L. Hecker, K. N. Stevens, G. von Bismark, and C. E. Williams, "Manifestations of Task-Induced Stress in the Acoustic Speech Signal", *J. Acoust. Soc. Am.* , vol. 44, no. 4, pp. 993-1001, 1968.

[5] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals", *IEEE Trans. on Information Theory*, vol. 38, no. 2, pp. 917-924, March 1992.

[6] J. F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", *Proc. ICASSP-90*, pp. 381-384, 1990.

[7] E. Lombard, "Le Signe de l'Elevation de la Voix", *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101-119, 1911.

[8] P. Maragos, J. Kaiser, and T. Quatieri, "On Separating Amplitude from Frequency Modulations Using Energy Operators", *Proc. IEEE ICASSP-92*, vol. 2, pp. 1-4, 1992.

[9] R. S. McGowan, "An Aeroacoustic Approach to Phonation", *J. Acoust. Soc. Am.* , vol. 83, no. 2, pp. 696-704, 1988.

[10] D. B. Pisoni, R. H. Bernacki, H. C. Nusbaum, M. Yuchtman, "Some Acoustic-Phonetic Correlates of Speech Produced in Noise", *Proc. ICASSP-85*, pp. 1581-1585, 1985.

[11] B. Stanton, L. H. Jamieson, G. D. Allen, "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions", *Proc. ICASSP-88*, pp. 331-334, 1988.

[12] L. A. Streeter, N. H. Macdonald, W. Apple, R. M. Krauss, K. M. Galotti, "Acoustic and Perceptual Indicators of Emotional Stress", *J. Acoust. Soc. Am.* , vol. 73, no. 4, pp. 1354-1360, 1983.

[13] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Production Mechanisms in the Vocal Tract", in *Speech Production and Speech Modeling*, pp. 241-261, 1990.

[14] T. J. Thomas, "A Finite Element Model of Fluid Flow in the Vocal Tract", *Computer Speech and Language*, vol. 1, pp. 131-151, 1986.

[15] C. E. Williams and K. N. Stevens, "On Determining the Emotional State of Pilots During Flight: An Exploratory Study", *Aerospace Medicine*, vol. 40, pp. 1369-1372, 1969.

[16] C. E. Williams and K. N. Stevens, "Emotions and Speech: Some Acoustic Correlates", *J. Acoust. Soc. Am.* , vol. 52, no. 4, pp. 1238-1250, 1972.

[17] P. E. Zwicke and I. Kiss Jr. , "A New Implementation of the Mellin Transform and its Application to Radar Classification of Ships", *IEEE Trans. on PAMI*, pp. 191-199, March 1983.