Nonlinear analysis and classification of speech under stressed conditions

Douglas A. Cairns and John H. L. Hansen
Robust Speech Processing Laboratory, Department of Electrical Engineering, Box 90291, Duke University, Durham, North Carolina 27708-0291

(Received 22 October 1993; revised 2 January 1994; accepted 15 July 1994)

The speech production system is capable of conveying an abundance of information with regards to sentence text, speaker identity, prosodics, as well as emotion and speaker stress. In an effort to better understand the mechanism of human voice communication, researchers have attempted to determine reliable acoustic indicators of stress using such speech production features as fundamental frequency ($F0$), intensity, spectral tilt, the distribution of spectral energy, and others. Their findings indicate that more work is necessary to propose a general solution. In this study, we hypothesize that speech consists of a linear and nonlinear component, and that the nonlinear component changes markedly between normal and stressed speech. To quantify the changes between normal and stressed speech, a classification procedure was developed based on the nonlinear Teager Energy operator. The Teager Energy operator provides an indirect means of evaluating the nonlinear component of speech. The system was tested using VC and CVC utterances from native speakers of English across the following speaking styles; neutral, loud, angry, Lombard effect, and clear. Results of the system evaluation show that loud and angry speech can be differentiated from neutral speech, while clear speech is more difficult to differentiate. Results also show that reliable classification of Lombard effect speech is possible, but system performance varies across speakers.

PACS numbers: 43.72.Ar, 43.72.Kb

INTRODUCTION

Stress and its manifestation in the acoustic speech signal has been the subject of many studies. Researchers have attempted to determine reliable indicators of stress by analyzing acoustic variables such as fundamental frequency (Lieberman and Michaels, 1962; Hecker *et al.*, 1968; Williams and Stevens, 1969; Williams and Stevens, 1972; Streeter *et al.*, 1983; Pisoni *et al.*, 1985; Hansen and Clements, 1987; Hansen, 1988; Stanton *et al.*, 1988; Hansen, 1989), amplitude (Lieberman and Michaels, 1962; Pisoni *et al.*, 1985; Hansen and Clements, 1987; Hansen, 1988), concentration of spectral energy (Scherer, 1981; Hansen and Clements, 1987; Hansen, 1988; Stanton *et al.*, 1988), and others (Kuroda *et al.*, 1976; Hansen and Clements, 1987; Hansen, 1988; Stanton *et al.*, 1988; Hansen, 1989). In these studies, stress refers to speech spoken under one (or more) of the following conditions; emotional (anger, fear, sorrow), task-induced (completion of a task with a time constraint), or environmental [high level of background noise as in the Lombard effect (Lombard, 1911)].

Fundamental frequency ($F0$) has been the most common acoustic variable studied. Williams and Stevens (1969) performed an experiment on data collected from radio transmissions of pilots experiencing flight problems (some of which were fatal). They found that $F0$ rose in stressful situations. They also noted that $F0$ changes were smooth in normal speech, while $F0$ changes could be erratic in stressed speech. In another study, Williams and Stevens (1972) analyzed the speech of professional actors simulating emotions. This study reinforced the finding that $F0$ is different between normal and stressed speech. However, they also found that each emotion had distinctive characteristics (i.e., sorrow—low $F0$, flat $F0$ contour; anger—high $F0$, large range of $F0$ values). Streeter *et al.* (1983) performed a study similar to the flight recording experiment of Williams and Stevens. They analyzed the recorded speech of system operators responsible for electric power distribution before the 1977 blackout of New York City. In contrast to Williams and Stevens, Streeter *et al.* could not find a consistent rise of $F0$ under stress. Hecker *et al.* (1968) reported similar results in a task-induced stress experiment.

In conjunction with $F0$, variables such as amplitude and the distribution of spectral energy have been studied. Analysis of speech produced in high levels of background noise has shown significant changes in several acoustic variables. The change in speech characteristics needed to communicate effectively in the presence of background noise is called the Lombard effect (Lombard, 1911). Pisoni *et al.* (1985) found that amplitude, duration, and pitch changed in Lombard speech. In addition, spectral energy shifted to higher frequency for consonants. In a similar study of loud and Lombard speech, Stanton *et al.* (1988) noted the same shift in spectral energy. Scherer (1981) suggested that this shift in spectral energy, along with pitch, were the two most promising indicators of stress.

Previous research directed at the problem of analysis of speech under stress has generally been limited in scope, often suffering from one to five problems. These include: (i) limited speaker populations, (ii) sparse vocabularies, (iii) reporting only qualitative results with little statistical confirmation, (iv) limited numbers and types of speech parameters considered, and (v) analysis based on simulated or actual conditions with little confirmation between the two. To address

0001-4966/94/96(6)/3392/9/$6.00

TABLE I. A selected set of speech production features previously analyzed using stressed speech from speakers in the SUSAS database.

Speech production analysis domain		Speaking condition				
		Neutral	Clear	Lombard	Loud	Angry
Fundamental frequency (Hz)	F_0(vowel)	142	150^a	163^a	209^a	283^a
	σ_{F_0}(vowel)	15	22^a	24^a	44^a	56^a
Glottal source-spectrum roll-off (dB/oct.)		−12.1	−11.9	−9.2^a	−9.5^a	−9.4^a
Duration (ms)	d(word)	478	666^a	572^a	650^a	662^a
	σ^2(word)	18.	40.^{a}	24.^{a}	28.^{a}	41.^{a}
	d(vowel)	160	202	198	253^a	271^a
	σ(vowel)	7.9	17.^{a}	13.^{a}	19.^{a}	23.^{a}
	d(diphthong)	192	199	249^a	294^a	315^a
	σ^2(diphthong)	3.3	3.3	3.5	5.6	7.0
	d(consonant)	71	128^a	73	73	62
	σ^2(consonant)	1.8	10.^{a}	2.6	3.7^a	3.3^a
Intensity (dB)	I(word)	77.7	77.0	78.4	80.5^a	81.1^a
	I(vowel)	79.7	79.8	79.7	81.6^a	82.1^a
	I(diphthong)	80.1	80.3	80.8	83.4^a	83.4^a
	I(semivowel)	80.0	78.4	78.4	79.5	81.3
	I(consonant)	62.9	62.2	62.9	61.3	63.9
Vocal tract spectrum (Hz)						
	$\bar{F}1$(/IY/) location	411	387^a	412	431^a	586^a
	$\bar{F}2$(/IY/) location	1970	2086^a	2006^a	2071^a	2078^a
	$\bar{F}3$(/IY/) location	2607	2667^a	2644^a	2686^a	2661^a
	$\bar{F}4$(/IY/) location	3368	3379	3376	3414^a	3357
	$\bar{B}1$(/IY/) bandwidth	52	105^a	73^a	86^a	102^a
	$\bar{B}2$(/IY/) bandwidth	222	356^a	139^a	174^a	166^a
	$\bar{B}3$(/IY/) bandwidth	496	613^a	250^a	355^a	464
	$\bar{B}4$(/IY/) bandwidth	366	531^a	185^a	219^a	392

[a]Indicates a statistically significant shift from neutral.

these issues, Hansen and Clements (Hansen and Clements, 1987; Hansen, 1988; Hansen, 1989) considered an analysis of acoustic and perceptual correlates of speech under various emotion and stress conditions. These studies included analysis of a database of simulated and actual stressed speech recordings using a predefined vocabulary set. [This database is called SUSAS, for Speech Under Simulated and Actual Stress; approximately half of which consists of style data from Lincoln Laboratories (Lippmann *et al.*, 1987; Hansen, 1988, 1994).] The areas of speech under stress included various talking styles (slow, fast, soft, loud, angry, clear, question), tracking workload stress inducing tasks, speech spoken in noise, and subject motion-fear tasks. These studies were performed on five factors of speech production that include pitch, glottal source, intensity, duration, and spectral features. Well over 200 features were considered across simulated and actual stress conditions. Results showed that such features as glottal source spectral slope to be significantly different under loud, angry, and Lombard effect speaking conditions. Individual phoneme duration varies significantly under clear, angry, and loud conditions. Formant structure was shown to vary significantly for a number of stressed speaking conditions. Table I summarizes selected speech production features from previous studies (Hansen, 1988) for the five speech conditions (neutral and four stressed speaking styles) used in this study. Features that are different from neutral in a statistically significant manner are appropriately marked. The results show that when a speaker produces speech under stressed conditions, a variety of production domains are used to indicate the presence of stress.

These studies showed that though certain production domains are traditionally modified when a speaker is under stress, not all speakers exhibit the same level of production variation for a given stressed speaking condition. In addition to the five reasons listed above, there is another potential explanation for the inconclusiveness of past acoustic studies. In the speech production process, there is a net airflow through the glottis. The linear acoustic model of speech production says that this flow only causes sound when forced through a constriction (i.e., fricative production). However, if the propagation of the glottal flow through the vocal tract created vortices of air in the region of the false vocal folds, sound could be actively produced from a source other than the glottis. This phenomena of sound creation by vortex action is nonlinear and cannot be measured by any of the techniques employed to date. Teager (Teager and Teager, 1983a), who suggested that these vortices modulated airflow in the vocal tract causing sound, developed the Teager Energy operator. The operator was used to show modulation patterns in the energy of individual formants. In this study, we propose to utilize the Teager Energy operator to measure the energy of the first formant. Experimental evidence has shown that the energy of the first formant could be a useful basis for classifying speech as normal or stressed. The next section

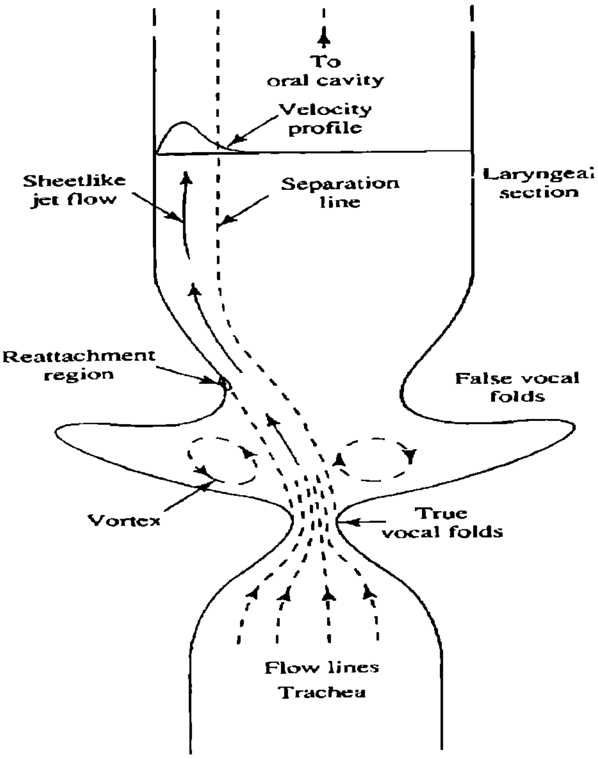

FIG. 1. Nonlinear model of sound propagation along the vocal tract (Kaiser, 1983).

gives the formulation and motivation of the Teager Energy operator.

I. TEAGER ENERGY OPERATOR

The Teager Energy operator, which provides a measure of the energy of a speech signal, was motivated by experiments in speech and hearing by Teager and Teager (1980, 1981, 1983b, 1990). In these experiments, H. Teager demonstrated that the airflow in the vocal tract is separated and adheres to the walls of the vocal tract. Given these observations, the geometry of the vocal tract, and the results of some experiments with whistle cavities, Teager proposed the model of speech production shown in Fig. 1. In this model, air exits the glottis as a jet and attaches to the nearest wall of the vocal tract. As the air passes over the cavity between the true vocal folds and the false vocal folds, vortices of air are created. The bulk of the air continues propagating towards the lips while adhering to the walls of the vocal tract.

The key element in this model is the vortex action. A traditional model of speech production allows sound to be actively produced in an unconstricted vocal tract only at the glottis. Teager asserted that vortices in the region of the false vocal folds are also actively producing sound that causes modulations in the speech signal. Teager's view of speech production is supported by the work of Thomas (1986) and McGowan (1988). Thomas numerically simulated fluid flow in the vocal tract and found that vortices are created. McGowan, using principles of fluid mechanics, showed that vortices exist. He also showed that the vortices are capable of actively producing sound.

Teager developed an energy measurement motivated by his speech research and some experiments in hearing to find evidence of speech modulation patterns. J. Kaiser (1990) first documented the form of the energy operator as follows,

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

where $\Psi\{\bullet\}$ is the Teager Energy operator, and $x(n)$ is the sampled speech signal.

The Teager Energy operator has been shown to contain significant cross-terms when applied to multicomponent signals (Kaiser, 1990). Therefore, to determine the Teager Energy profile for a single component of a multicomponent signal, the other components of the signal must be filtered out. For speech, this means that a bandpass filter must be applied to a formant to remove the influence of the other formants. When applied to a single formant, the output of the operator has shown multiple excitation pulses within a pitch period (Maragos *et al.*, 1991; Teager and Teager, 1990). Teager has suggested that the multiple excitation pulses represent evidence of a nonlinear speech production phenomena (Teager and Teager, 1990).

In this study, it is suggested that speech production consists of both linear and nonlinear components. In other words, speech is a combination of linear acoustic production, and sound generated by vortex action [McGowan (1988) makes a similar argument]. It is further hypothesized that the nonlinear component changes appreciably between normal and stressed speech. To quantify the changes between neutral and stressed speech, the Teager Energy operator was used. As a result of speech production experiments, Teager suggested that the flow in the vocal tract switches walls at the first formant frequency. Since the nonlinear component of speech arises from this flow, the proposed system extracts the first formant over an entire voiced speech segment. In an effort to eliminate the effect of variable pitch on the results, an analysis frame that is pitch synchronous is used, and frame duration is normalized using the Mellin Transform. The following section describes the proposed nonlinear normal/stressed classification system.

II. CLASSIFICATION SYSTEM

The system developed for the normal/stressed speech classification task is shown in Fig. 2. As can be seen, there are four main processing steps. When an utterance is presented to the system, two assumptions are made: (1) the system knows the text of the word spoken, and (2) the word is a vowel-consonant (VC) or a consonant-vowel-consonant (CVC) utterance. The first assumption eliminates the uncertainty of which word is presented to the system. If this assumption was not made, speech recognition would have to be performed prior to speech classification. Since the performance of speech recognition has been shown to degrade on stressed speech (Hansen, 1988; Hansen and Clements, 1989; Cairns and Hansen, 1992), this would introduce an additional source of error to the system. The second assumption is required in order to eliminate voiced production variability due to varying levels of lexical stress in multisyllable words.

The processing proceeds in this manner: pitch information is extracted from the word. An analysis window of two

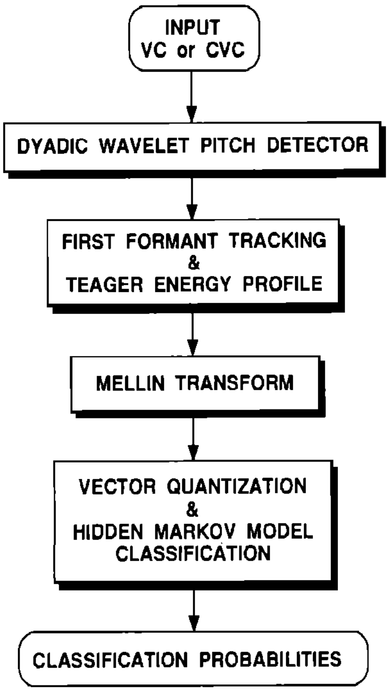

FIG. 2. Speech classification system flow diagram.

pitch periods is selected and the first formant located. The Teager Energy profile is extracted and parameterized by the Mellin Transform. A vector quantizer is used to map the feature data to a codebook entry. The analysis window is shifted by one pitch period and the procedure repeated. When pitch information is exhausted, the sequence of vector quantized observations is submitted to a hidden Markov model classifier. Each of the main processing steps will now be discussed.

A. Pitch detection

A pitch detector is necessary for one important reason. The Teager Energy profile has been observed to exhibit a modulation pattern for each pitch period. It is hypothesized that this modulation pattern, and the evolution of the modulation pattern across an utterance, characterize a speaking style. It is therefore necessary to track the modulation pattern of the Teager Energy profile at the scale of a pitch period. The pitch detector used was a derivative of the algorithm by Kadambe and Boudreaux-Bartels (1990, 1992). Though other pitch estimation schemes exist (Hess, 1983), the rapid and varied movement of pitch under stress required a robust method with limited user supervision. The pitch detector developed was based on the Dyadic Wavelet Transform (D_yWT), which has the form

$$D_yWT(b,2^j)=\frac{1}{2^j}\int_{-\infty}^{\infty} x(t)\,\psi^*\left(\frac{t-b}{2^j}\right)dt. \tag{2}$$

Here, b is the time index, $x(t)$ is the signal, $\psi^*(t)$ is the complex conjugate of the wavelet, and 2^j is the scale parameter. In realizing the above equation, the D_yWT for a given scale was computed by convolving the time-reversed wavelet with the speech. This procedure was followed for the three scales that correspond to the range of fundamental frequency of the human voice. The D_yWT for the three scales were then windowed by a 128 point (16 ms) rectangular window. Maxima were compared across scales, and matching maxima indicated pitch epochs. The window was shifted by 64 points (8 ms) and the procedure was repeated. This process continued until the D_yWT information was exhausted. This approach was tested and found to mark pitch epochs consistently in neutral speech. However, for erratic pitch under some stressed conditions, it would sometimes miss pitch periods. After several experimental trials, a two-pass version of the algorithm was implemented. In the two-pass approach, the first pass is the original algorithm. The second pass utilizes the original algorithm with a 64 point (8 ms) rectangular window and a 32 point (4 ms) skip rate. This pass is applied to portions of an utterance that fall within previously marked pitch epochs, but have a pitch greater than 150% of the median pitch of the first pass. This version of the pitch detector was found to satisfactorily mark pitch epochs across neutral and stressed speech.

After successfully determining the pitch profile, each pitch epoch was migrated to the location of the previous zero crossing in the speech waveform. This was implemented because pitch epochs were not marked in the same location in a pitch period from utterance to utterance. This approach was found to give a consistent pitch period and hence a consistent Teager Energy profile across words.

B. Formant tracking/Teager Energy operator

Once pitch boundaries have been determined, the modulation pattern of the first formant could be considered. To track and isolate the first formant, a Teager Energy based formant-tracker developed by Hanson, Maragos, and Potamianos (1993a,b) was used. This work was based on an AM-FM model for speech. Maragos, Quatieri, and Kaiser (1992) found that the Teager Energy operator could be used to separate the AM and FM contributions via

$$f(n)\approx\frac{1}{2\pi T}\arccos\left(1-\frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]}\right), \tag{3}$$

$$|a(n)|\approx\sqrt{\Psi[x(n)]\Bigg/\left[1-\left(\frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]}\right)^2\right]}, \tag{4}$$

where $y(n)=x(n)-x(n-1)$, $\Psi[\bullet]$ 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. Hanson *et al.* found that, given an approximate estimate of a formant location, the FM contribution could be used to iteratively refine the formant center frequency via

$$f_c^{i+1}=\frac{1}{N}\sum_{n=1}^{N} f(n). \tag{5}$$

Here, N is the length of the speech segment, and f_c^{i+1} is the formant center frequency on iteration $i+1$. When the average instantaneous frequency (f_c^{i+1}) changed by less than 10 Hz, the formant center frequency was located (see Appendix A for exact details of the formant tracking procedure). Using this approach, the first formant was tracked for the vowel

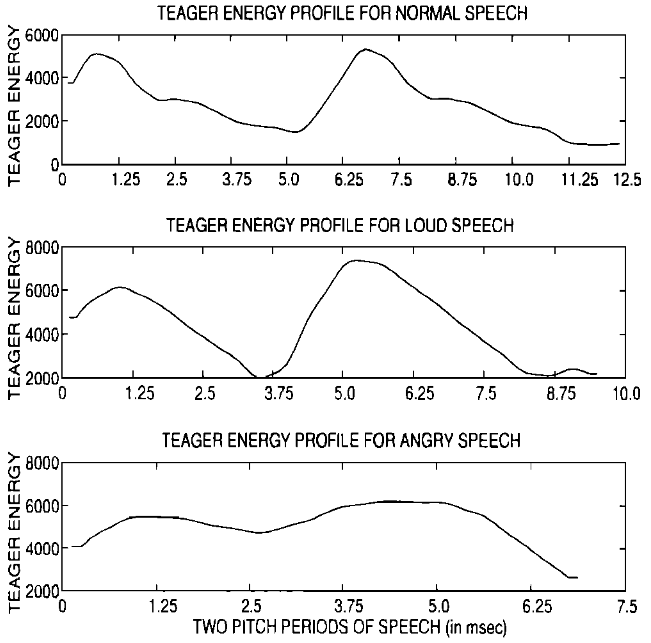

FIG. 3. Teager Energy profile for neutral, loud, and angry speaking styles.

section of a VC or CVC utterance, and the Teager Energy profile extracted for each analysis window. It should be noted that the Teager Energy operator gives the instantaneous "energy" of a signal, so the Teager Energy profile contains an energy value corresponding to each sample in an analysis window. Figure 3 shows examples of Teager Energy profiles output by the formant tracker.

C. Mellin Transform

As documented in the Introduction, several researchers have shown a link between fundamental frequency ($F0$) and stress. However, no consistent relationship has been shown across speakers between stress and $F0$ for any given stress condition. For this reason, we chose to effectively neutralize the influence of $F0$ on the analysis so that the nonlinear component of speech could be studied. A pitch synchronous analysis with a scale invariant transform removes duration variability caused by a changing pitch contour. The scale invariant transform utilized here is the Mellin Transform. The Mellin Transform was chosen because it showed considerable discriminatory power in a similar shape classification task (Zwicke and Kiss, 1983). The Mellin Transform has the form (Zwicke and Kiss, 1983)

$$M\{f(e^x)\}=\int_{-\infty}^{\infty} f(e^x)e^{sx}dx, \qquad (6)$$

where $f(e^x)$ is the signal to be transformed (i.e., the Teager Energy profile). Equation (6) can be shown to be equivalent to (Zwicke and Kiss, 1983)

$$G(\omega)=\frac{j}{\omega}\sum_{n=1}^{N}[\cos(\omega \ln n)-j\sin(\omega \ln n)]$$
$$\times[f(n)-f(n+1)], \qquad (7)$$

where $f(n)$ is the Teager Energy profile for a given analysis window at sample n, N is the length of the analysis window, and $G(\omega)$ is the Mellin Transform coefficient at frequency ω. This form of the Mellin Transform is called the Direct Mellin Transform (DMT). Taking the magnitude of both sides of Eq. (7), and dropping the $1/\omega$ term yields

$$|G(\omega)|$$
$$=\sqrt{\left\{\sum_{n=1}^{N}\cos(\omega\ln n)\Delta_n\right\}^2+\left\{\sum_{n=1}^{N}\sin(\omega\ln n)\Delta_n\right\}^2}. \qquad (8)$$

Here, $\Delta_n=f(n)-f(n+1)$, represents the rate of change of the Teager Energy profile. The form shown in Eq. (8) has been called the Modified Direct Mellin Transform (MDMT). This form is very ameniable to discrete-time signals. Using the MDMT, the value of the transform was computed at intervals of $2\pi/32$ from 0 to 2π. Through the use of the MDMT, similar Teager Energy profiles will not be misclassified simply because the pitch based analysis window size is different.

D. Vector quantizer/hidden Markov model classifier

The final processing step involves data reduction and classification. Data reduction was accomplished by a 128-state vector quantizer. Vector quantization is a widely used technique in the areas of speech and image processing. For further details, the reader is referred to the study by Gray (1984). Classification was performed by an algorithm based on hidden Markov models. For these experiments, a five-state, discrete observation, left-to-right hidden Markov model was used. This type of algorithm has been used extensively in the area of speech recognition (Rabiner *et al.*, 1983). Since the proposed classification framework resembles a speech recognition task, it was deemed appropriate to use this approach. However, other classification schemes based on neural networks or other statistical pattern recognition techniques are equally valid. For more details on hidden Markov models for speech applications, see (Jelinek *et al.*, 1975; Jelinek, 1976; Levinson *et al.*, 1983; Rabiner and Juang, 1986).

III. EXPERIMENT

In the hypothesis from Sec. I, it is suggested that the nonlinear component of speech undergoes a fundamental change between neutral and stressed speech. The data used to test this hypothesis comes from the SUSAS (Speech Under Simulated and Actual Stress) database. The SUSAS database was established for the purposes of stress research (Hansen and Clements, 1987; Hansen, 1988; Hansen, 1989; Cairns, 1991; Cairns and Hansen, 1992; Hansen and Bria, 1992). The database is partitioned into five domains, encompassing a wide variety of stresses that include: various talking styles (slow, fast, soft, loud, angry, clear, question, in noise), single and dual tracking workload stress inducing tasks, emotional speech from psychiatric analysis sessions, and subject motion-fear tasks. A total of 32 speakers were employed to generate in excess of 16,000 utterances. From the database, nine native speakers of English were chosen. For each speaker, there were twelve neutral utterances of each word, and two utterances of the following stress speaking styles; loud, angry, Lombard effect, and clear. The Lombard speech data was obtained by having speakers wear headphones with 85 dB SPL of pink noise played while speaking (i.e., all recordings are noisefree). A total of six VC or CVC words were chosen for each speaker resulting in a total of 120 utterances per speaker.

Evaluations were conducted in the following manner. Three examples of neutral speech along with one example of each of the stress styles for each word were chosen to train the vector quantizer. The three neutral examples used to train the vector quantizer, along with three additional neutral examples for each word were used to train the hidden Markov models for each speaker. This approach yields a hidden Markov model that is speaker dependent. The training tokens were excluded from the final evaluation (i.e., an open recognition evaluation), which consisted of testing the remaining neutral and stressed speech. The final evaluation involved submitting the remaining neutral and stressed speech to the system. The hidden Markov model classifier outputs a probability for each utterance that was processed according to the following heuristic algorithm:

```
(1) Determine decision boundary for given VC or CVC
    (a) $thresh=min[LP(neutral)_1,...,LP(neutral)_n]$
        (i) $MAXCR(1)=0$,
        (ii) compute classification rate (CR) for each talking style,
        (iii) $TOTCR=CR(neutral)+CR(loud)+CR(angry)+CR(Lombard)+CR(clear)$,
        (iv) If $TOTCR>MAXCR(1)$, $MAXCR(1)=TOTCR$ and $THRESH(1)=thresh$,
        (v) $thresh=thresh-\epsilon$.
        (vi) If $thresh>-7.0$, return to 1(a)ii,
    (b) $thresh=min[LP(neutral)_1,...,LP(neutral)_n]+\epsilon$,
        (i) $MAXCR(2)=0$,
        (ii) compute classification rate (CR) for each talking style,
        (iii) $TOTCR=CR(neutral)+CR(loud)+CR(angry)+CR(Lombard)+CR(clear)$,
        (iv) If $TOTCR>MAXCR(2)$, $MAXCR(2)=TOTCR$ and $THRESH(2)=thresh$,
        (v) $thresh=thresh+\epsilon$,
        (vi) If $thresh<0.0$, return to 1(b)ii.
    (c) If $MAXCR(1)>MAXCR(2)$, $THRESHF=THRESH(1)$. Otherwise, $THRESHF=THRESH(2)$.
(2) Compute final classification rates using THRESHF as the decision boundary.
```

Here, LP(i) is the log-probability for the *i*th neutral utterance of a given word.

Application of the Teager Energy operator across the selected speaking styles revealed a difference in the Teager Energy profile of the first formant for neutral versus stressed speech. Figures 3 and 4 show examples of the Teager Energy profile for neutral, loud, Lombard, angry, and clear utterances of "on." These figures illustrate the different forms of the Teager Energy profile across speaking styles. The goal in this experiment is to determine whether the change in the Teager Energy profile can be used to classify speech as neutral or stressed for a given speaker.

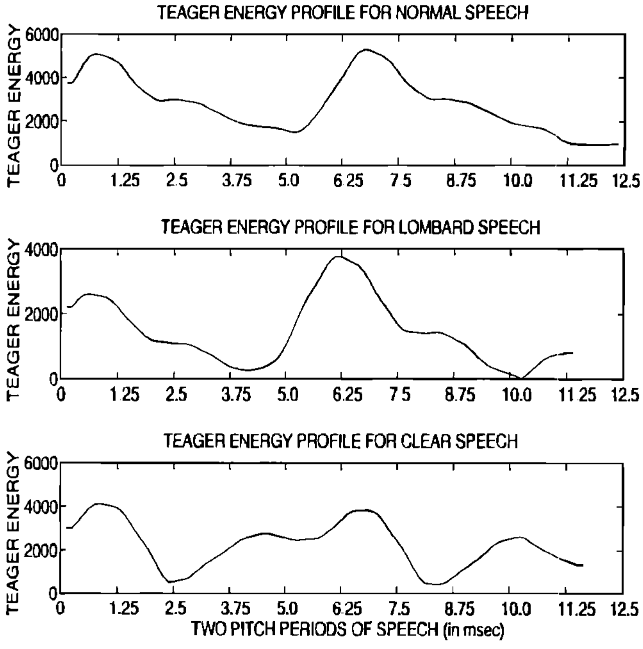

FIG. 4. Teager Energy profile for neutral, Lombard, and clear speaking styles.

IV. DISCUSSION

The results of the classification evaluations are shown in Table II and Fig. 5. From the data, several conclusions can be drawn. First, loud and angry stress styles are differentiated from neutral speech as evidenced by the overall classification rates of 98.1% and 99.1% respectively. This result shows that there is a clear difference in the Teager Energy profile between neutral and loud or angry speech. It is our contention that the change in the nonlinear component of speech causes the change in the Teager Energy profile. However, our experimental framework is not capable of proving this contention.

The second conclusion supported by the data is that Lombard effect speech is not as reliably differentiated from neutral speech as loud and angry speech. As shown in Fig. 5, there is a wide range in classification results for Lombard effect speech. However, if the results from speakers *S*3 and *S*4 are removed, the mean classification rate increases from 86.1% to 94.0%, and the range of classification rates shrinks considerably. The preceeding observation suggests several possibilities. First, that production variability due to the Lombard effect was not as pronounced in speakers *S*3 and *S*4 as it was in the other speakers. If this is the case, the relatively low classification rates are understandable. Another possibility is that the Teager Energy profile alone is not sufficient to reliably separate Lombard effect speech from neutral speech. It may be necessary to incorporate other features relating to spectral shape to reliably differentiate Lombard effect speech. There may be still other possibilities.

TABLE II. Classification results for nine speakers.

Speaker	Classification rate (%)				
	Neutral speech	Loud	Angry	Lombard	Clear
S1	100.0	100.0	100.0	83.3	50.0
S2	100.0	100.0	100.0	100.0	58.3
S3	100.0	100.0	100.0	50.0	58.3
S4	91.6	91.6	91.6	66.7	66.7
S5	100.0	100.0	100.0	83.3	83.3
S6	97.2	100.0	100.0	100.0	75.0
S7	97.2	100.0	100.0	100.0	50.0
S8	94.4	91.6	100.0	100.0	66.7
S9	97.2	100.0	100.0	91.6	75.0
Mean	97.5	98.1	99.1	86.1	64.8

More research is required to resolve which of the above explanations is the correct one.

The last conclusion supported by the data is that clear speech is not easily differentiated from neutral speech. It is suggested that the nonlinear component of speech is more pronounced in loud, angry, and Lombard effect speech as compared to clear speech. If the nonlinear component of speech is the dominant factor in the change in the shape of the Teager Energy profile, our contention explains the classification results for clear speech. However, it is noted that other factors could explain the change in the Teager Energy profile.

It has been hypothesized that speech production consists of a linear and a nonlinear component, and that the nonlinear component changes markedly between neutral and stressed speech. While the results presented here show a promising application of the Teager Energy operator, we cannot conclusively state that our hypothesis has been validated. It is possible that factors such as vocal fold vibration or vocalic register could have contributed to changes in the shape of the Teager Energy profile. Further research in this area may be required before a conclusive statement can be made about the nature of speech production.

In this study, it has been shown that the Teager Energy profile of the first formant is useful for differentiating neutral from loud, angry, and Lombard effect speech. The system developed to perform this task is independent of traditional acoustic features such as pitch, intensity, spectral energy, etc. Incorporation of traditional features could lead to further improvements in classification performance. Looking beyond the current classification task, this research has implications for the fields of speech synthesis and speech recognition. For speech synthesis, incorporation of Teager Energy profile information could aid in producing more natural sounding speech in text-to-speech systems. Speech recognition could benefit by using this classification scheme as a front end processor to determine the state of the speaker. The correct stress-dependent recognition model could then be selected based on the speaker's state. This approach could improve the accuracy of speech recognition under stress which has traditionally deteriorated under task-induced stressed speaking conditions (Hansen, 1988; Hansen and Clements, 1989; Cairns, 1991; Cairns and Hansen, 1992).

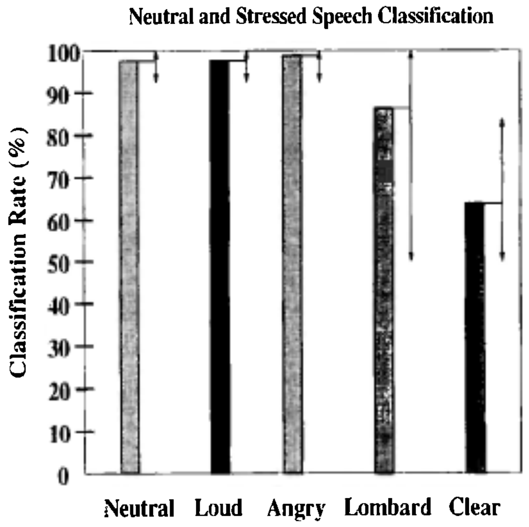

FIG. 5. Mean classification rates across speaking styles.

V. SUMMARY

This study has focused on evaluating the hypothesis that a nonlinear component of speech changes noticeably between speech spoken under neutral and stressed conditions. To evaluate this, a speech processing approach was developed to classify speech as being spoken in either neutral or stressed styles. The system employs the Teager Energy operator to quantify the nonlinearity (i.e., modulation pattern) of the first formant within vocalic sections of VC and CVC words. Results show that loud and angry speech can be differentiated from neutral speech, while clear speech is more difficult to differentiate. Results also indicate that Lombard effect speech can be reliably classified, although system performance varied across speakers. It is therefore suggested that the nonlinear component during speech production is more pronounced for loud and angry stressed speaking styles, hence a resulting improvement in classification performance. Though the experimental framework cannot conclusively prove the existence of a nonlinear component, the results do suggest a promising application of the Teager Energy operator and its ability to represent nonlinear speech dynamics. However, more research is required in the area of speech production dynamics to determine the contribution of the linear and nonlinear speech production components.

APPENDIX

The formant tracking procedure is as follows:

(1) Extract two pitch periods of speech with rectangular window.

(2) Obtain initial estimate of formant center frequency for analysis window (an LPC-based formant tracker was used to give initial estimates of $F1$ (McCandless, 1974).

(3) Filter frame of speech data with Gabor bandpass filter

$$g(n)=\exp[-(\alpha nT)^2]\cos(2\pi f_c Tn),\quad |n|\leq N, \qquad \text{(A1)}$$

where N is the length of the analysis window, T is the sampling interval, and f_c is the center frequency found on the previous iteration, or the estimate of $F1$ obtained from the LPC formant tracker for the initial pass through the algorithm. Choose α based on the following criteria (a) If $|f_c^{i+1}-F2|>500$ Hz and $f_c^{i+1}>1000$ Hz, $\alpha=1100$. (b) Otherwise, $\alpha=800$.

(4) Using Eq. (3), compute $f(n)$.

(5) Compute new estimate of first formant center frequency, f_c^{i+1}, using Eq. (5).

(6) If $|f_c^{i+1}-f_c^{i}|<10$ Hz, terminate procedure. Otherwise, return to (3).

(7) Move analysis window forward one pitch period and return to (1).

Cairns, D. A. (**1991**). "Real Time Speech Recognition Under Lombard Effect and in Noise," M. S. thesis, Department of Electrical Engineering, Duke University, Durham, N.C.

Cairns, D. A., and Hansen, J. H. L. (**1992**). "ICARUS: An Mwave Based Real-time Speech Recognition System in Noise and Lombard Effect," ICSLP-92, Inter. Conf. on Spoken Language Processing **II**, 703–706.

Gray, R. M. (**1984**). "Vector Quantization," IEEE ASSP Magazine, 4–29.

Hansen, J. H. L., and Clements, M. A. (**1987**). "Evaluation of Speech under Stress and Emotional Conditions," Proc. of the Acoust. Soc. Am., 114th Meeting, H15.

Hansen, J. H. L. (**1988**). "Analysis and Compensation of Stressed and Noisy Speech With Application to Automatic Speech Recognition," Ph.D. dissertation, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA.

Hansen, J. H. L. (**1989**). "Evaluation of Acoustic Correlates of Speech Under Stress for Robust Speech Recognition," IEEE Proc. of the Fifteenth Annual Northeast Bioengineering Conference, 31–32.

Hansen, J. H. L. (**1994**). "Morphological Constrained Feature Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect," IEEE Transactions on Speech and Audio Processing, **SA-2**(4), October 1994.

Hansen, J. H. L., and Clements, M. A. (**1989**). "Stress Compensation and Noise Reduction Algorithms for Robust Speech Recognition," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, 266–269.

Hansen, J. H. L., and Bria, O. (**1992**). "Improved Automatic Speech Recognition in Noise and Lombard Effect," EURASIP-92, The Sixth European Signal Processing Conference, 403–406.

Hanson, H., Maragos, P., and Potamianos, A. (**1993a**). "A System for Finding Speech Formants and Modulations via Energy Separation," IEEE Transactions on Speech and Audio Processing **2**(3), 436–442.

Hanson, H., Maragos, P., and Potamianos, A. (**1993b**). "Finding Speech Formants and Modulations via Energy Separation: With Application to a Vocoder," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing **2**, 716–719.

Hecker, M. H. L., Stevens, K. N., von Bismark, G., and Williams, C. E. (**1968**). "Manifestations of Task-Induced Stress in the Acoustic Speech Signal," J. Acoust. Soc. Am. **44**, 993–1001.

Hess, W. (**1983**). *Pitch Determination of Speech Signals: Algorithms and Devices* (Springer-Verlag, Berlin).

Jelinek, F., Bahl, L. R., and Mercer, R. L. (**1975**). "Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech," IEEE Transactions on Information Theory **21**, 250–256.

Jelinek, F. (**1976**). "Continuous Speech Recognition by Statistical Methods," Proc. IEEE **64**, 532–556.

Junqua, J.-C., (**1993**). "The Lombard Effect and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Am. **93**(1), 510–524.

Kadambe, S., and Boudreaux-Bartels, G. F. (**1990**). "A Comparison of a Wavelet Transform event detection pitch detector with classical pitch detectors," Proc. 24th Asilomar Conf. on Signals, Systems, and Computers, 1073–1077.

Kadambe, S., and Boudreaux-Bartels, G. F. (**1992**). "Application of the Wavelet Transform for Pitch Detection of Speech Signals," IEEE Transactions on Information Theory **38**(2), 917–924.

Kaiser, J. F. (**1983**). "Some Observations on Vocal Tract Operation from a Fluid Flow Point of View," in *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, edited by I. R. Titze and R. C. Scherer (Denver Center for Performing Arts, Denver).

Kaiser, J. F. (**1990**). "On a Simple Algorithm to Calculate the 'Energy' of a Signal," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, 381–384.

Kaiser, J. F. (**1990**). "On Teager's Energy Algorithm and its Generalization to Continuous Signals," Proc. 4th IEEE Digital Signal Processing Workshop.

Kuroda, I., Fujiwara, O., Okamura, N., and Utsuki N. (**1976**). "Method for Determining Pilot Stress Through Analysis of Voice Communication," Aviation, Space, and Environmental Medicine **47**(5), 528–533.

Levinson, S. E., Rabiner, L. R., Sondhi, M. M. (**1983**). "An Introduction to the Application of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," The Bell System Technical Journal **62**(4), 1035–1074.

Lieberman, P., and Michaels, S. B. (**1962**). "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech," J. Acoust. Soc. Am. **34**(7), 922–927.

Lippmann, R. P., Martin, E. A., and Paul, D. B. (**1987**). "Multi-Style Training for Robust Isolated-Word Speech Recognition," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, 705–708.

Lombard, E. (**1911**). "Le Signe de l'Elevation de la Voix," Ann. Maladies Oreille, Larynx, Nez, Pharynx, **37**, 101–119.

Maragos, P., Quatieri, T., and Kaiser, J. (**1991**). "Speech Nonlinearities, Modulations, and Energy Operators," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, 421–424.

Maragos, P., Kaiser, J., and Quatieri, T. (**1992**). "On Separating Amplitude from Frequency Modulations Using Energy Operators," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, **2**, 1–4.

McCandless, S. S. (**1974**). "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Transactions on Acoustics, Speech and Signal Processing **ASSP-22**(2), 135–141.

McGowan, R. S. (**1988**). "An Aeroacoustic Approach to Phonation," J. Acoust. Soc. Am. **83**(2), 696–704.

Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., and Yuchtman, M. (**1985**). "Some Acoustic-Phonetic Correlates of Speech Produced in Noise," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, 1581–1585.

Rabiner, L. R., Levinson, S. E., and Sondhi, M. M. (**1983**). "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition," The Bell System Technical Journal **62**(4), 1075–1105.

Rabiner, L. R., and Juang, B. H. (**1986**). "An Introduction to Hidden Markov Models," IEEE ASSP Magazine, 4–16.

Scherer, K. (**1981**). "Vocal Indicators of Stress," in *Speech Evaluation in Psychiatry*, edited by J. K. Darby (Grune & Stratton, New York).

Stanton, B., Jamieson, L. H., and Allen, G. D. (**1988**). "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," IEEE Proc. International Conference on Acoustics, Speech, and Signal Processing, 331–334.

Streeter, L. A., Macdonald, N. H., Apple, W., Krauss, R. M., and Galotti, K. M. (**1983**). "Acoustic and Perceptual Indicators of Emotional Stress," J. Acoust. Soc. Am. **73**(4), 1354–1360.

Teager, H. M., and Teager, S. M. (**1980**). "Some Observations on Oral Air Flow During Phonation," IEEE Transactions on Acoustics, Speech, and Signal Processing, **ASSP-28**(5), 599–601.

Teager, H. M., and Teager, S. M. (**1981**). "The Effects of Separated Air Flow on Vocalization," in *Vocal Fold Physiology: Contemporary Research and Clinical Issues*, edited by D. M. Bless and J. H. Abbs (College Hill, San Diego).

Teager, H. M., and Teager, S. M. (**1983a**). "Active Fluid Dynamic Voice Production Models, or There is a Unicorn in the Garden," in *Vocal Fold Physiology*, edited by I. Titze and R. Scherer (Denver Center for the Performing Arts Press, Denver).

Teager, H. M., and Teager, S. M. (**1983b**). "A Phenomenological Model for Vowel Production in the Vocal Tract," in *Speech Sciences: Recent Advances*, edited by R. G. Daniloff (College-Hill, San Diego), pp. 73–109.

Teager, H. M., and Teager, S. M. (**1990**). "Evidence for Nonlinear Production Mechanisms in the Vocal Tract," in *Speech Production and Speech Modeling* (Kluwer, Boston), NATO Advanced Study Institute Series D, Vol. 55, pp. 241–261.

Thomas, T. J. (**1986**). "A Finite Element Model of Fluid Flow in the Vocal Tract," Comput. Speech Lang. **1**, 131–151.

Williams, C. E., and Stevens, K. N. (**1969**). "On Determining the Emotional State of Pilots During Flight: An Exploratory Study," Aerospace Medicine **40**, 1369–1372.

Williams, C. E., and Stevens, K. N. (**1972**). "Emotions and Speech: Some Acoustic Correlates," J. Acoust. Soc. Am. **52**(4), 1238–1250.

Zwicke, P. E., and Kiss, I. (**1983**). "A New Implementation of the Mellin Transform and its Application to Radar Classification of Ships," IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-5**(2), 191–199.