Analysis of Stress in Speech Using Adaptive Empirical Mode Decomposition

James Z. Zhang, Nyaga Mbitiru, Peter C. Tay, Robert D. Adams
Department of Engineering and Technology
Western Carolina University, Cullowhee, North Carolina 28723

Abstract— **Stress in human speech can be detected by various methods know as Voice Stress Analysis (VSA). The detection is accomplished by measuring the frequency shift of a microtremor normally residing in the frequency range of 8 to 12 Hz when not stressed. Conventional detection methods include Fast Fourier Transform (FFT) or McQuiston-Ford algorithm. This paper presents a new method called Adaptive Empirical Mode Decomposition (AEMD) applied to voice stress detection. Because AEMD in essence is a time-frequency analysis method, it is possible to use this method for real-time voice stress detection.**

I. Introduction

Voice stress detection has found its applications in many areas including voice activated military equipment, psychological testing and deception detection. Voice stress analysis (VSA) is accomplished by measuring fluctuations in the physiological microtremor present in speech. A microtremor is a low amplitude oscillation of the reflex mechanism controlling the length and tension of a stretched muscle caused by the finite transmission delay between neurons to and from the target muscle. Microtremors are present in every muscle in the body including the vocal chords and have a frequency of around 8 – 12 Hz. During times of increased stress this microtremor shifts in frequency. This change in frequency transfers from the muscles in the vocal tract to the voice produced. Stress can thus be detected by analyzing the change in microtremor frequency of an individual's voice. Conventional algorithms include Fast Fourier Transform (FFT) as well as McQuiston-Ford algorithm. However, the accuracy of voice stress detection depends on the algorithms in use as well as the effectiveness of the examiners. An analysis method that is capable of consistently determining the stress level of an individual at low to medium stress levels regardless of examiner's effectiveness is needed. In this paper, we present a new method called Adaptive Empirical Mode Decomposition (AEMD) for voice stress detection. Because AMED can resolve frequency and amplitude contents of a signal at a given time, it may provide a means of accurately detecting voice stress in real-time.

II. Physiological Microtremor

Stress is defined as the disruption of (homeostasis) by physical or psychological stimuli. Physical factors such as noise, excessive heat/cold, and psychological factors such as emotion and sleep deprivation alter the internal equilibrium of the body causing a stress response. The General Adaptation Syndrome (GAS) is a model by Hans Seyle [1] that identifies the various stages of the stress response. The first stage is known as the alarm stage, here the body identifies the stressor or threat and goes into a state of alarm. Adrenaline is produced in order to prepare the body for fight or flight, this causes blood flow to be diverted to the large muscles of the body as the body prepares to run away or fight. In addition to adrenaline, another hormone known as cortisol is also produced. Cortisol is known as the 'stress hormone' and increases blood pressure and blood sugar in order to restore the body's homeostasis after stress. The second stage is known as the resistance stage, during this stage the body attempts to cope with the stress by adaptation. As the body tries to cope with the stress, it uses up its resources. The third stage is appropriately known as the exhaustion stage. This occurs when the body's resources are used up and it is unable to maintain normal function.

The first and second stages of stress are of particular interest as increased muscle tension occurs during these stages. This increased tension affects all muscles in the body including the vocal chords. An increase in tension may directly or indirectly affect the production of speech. In particular increased tension affects the physiological microtremor that is present in speech. A microtremor is a low amplitude oscillation of the reflex mechanism controlling the length and tension of a stretched muscle caused by the finite transmission delay between neurons to and from the target muscle. Microtremors are present in every muscle in the body including the vocal chords and have a frequency of around 8 – 12 Hz. During times of increased stress this microtremor decreases in amplitude [2]. This change in frequency transfers from the muscles in the vocal tract to the voice produced. Stress can thus be detected by analyzing the change in microtremor frequency of an individual's voice.

III. Voice Stress Analysis

Detection of stress by voice analysis has numerous applications most notably in military, law enforcement and emergency services. Military applications are of greatest interest as individuals are often under some sort of stress, be it physical or emotional. Physical stress can stem from sleep deprivation, environmental extremes and exhaustion. Emotional stress arises from fear or confusion from conflicting information. Elevated stress levels can adversely affect the performance speech recognition equipment which is a concern especially when the equipment is used in failsafe applications such as military equipment. Additional coding can be incorporated into the equipment to identify stressed

978-1-4244-5827-1/09/$26.00 ©2009 IEEE

speech and perform the appropriate corrections in order to maintain optimum performance.

Deception detection is a valuable application in both military and law enforcement. Voice stress analysis can be used to detect elevated stress levels caused by deception. Voice stress analysis is non-intrusive and does not require any physical connection to the subject in question so is therefore also suitable for clandestine functions.

Voice stress analysis can be used in emergency services to direct calls to priority operators based on the stress level of the individual. This allows for priority responses to priority calls therefore increasing the quality of service [3].

Most VSA products use the physiological microtremor in combination with other voice features such as pitch, tone, and fundamental frequency as a descriptor of an altered psychological state. Using signal processing as well as knowledge of microtremors, the products claim to be able to identify if an individual is in a stressed state [4].

IV. Empirical Mode Decomposition

Empirical Mode Decomposition method was first proposed by N. E. Huang in 1998 [5]. The basic concept of EMD is to identify proper time scales that reveal physical characteristics of the signals, and then decompose the signal into modes intrinsic to the functions, which are referred to as Intrinsic Mode Functions (IMF). IMFs are signals satisfying the following conditions:

1) in the whole dataset, the number of extrema and the number of zero crossings must either be equal or differ at most by one,
2) at any point, the mean value of the envelope defined by local maxima and the envelope defined by the local minima is zero.

Adaptive EMD Procedures: AEMD is an iterative or “sifting” process described as follows:

1) Upper and lower envelopes of the unstressed voice signal $h_x(t)$ are constructed with its maxima and minima using cubic spline function.
2) Mean of the envelopes m_i is subtracted from $h_x(t)$ to obtain a new signal $h_i(t)$.
3) Determine if $h_i(t)$ is an IMF using the criteria described above.
4) If $h_i(t)$ is an IMF, it is subtracted from the original signal $h_x(t)$, and the resulted new signal $h_x(t)$ goes through the above procedures until another IMF is obtained.
5) Each IMF is checked if it is in the microtremor frequency band (8 – 12 Hz). If not, the algorithm adaptively adjusts the stopping criteria until the in-band IMF representing a microtremor is detected. This IMF is used as the reference.
6) The stopping criteria obtained in 5) and procedures 1) – 5) are applied to the Speech Signals Under Test (SSUT), and the IMFs obtained from SSUT are compared with the reference to determine if voice stress exists.

The stopping criteria consist of several important parameters including the absolute amplitude of the remaining signal, the mean value of the envelope, the cross-correlation coefficient between the remaining signal and the original signal, and the Standard Deviation (SD) between two consecutive results in the sifting process. SD is the most sensitive criterion and can be expressed in (1).

$$SD = \sum_{t=0}^{T} \left[\frac{\left| h_{1(k-1)}(t) - h_{1k}(t) \right|^2}{h_{1(k-1)}^2(t)} \right] \tag{1}$$

The AEMD process is shown in figure 1.

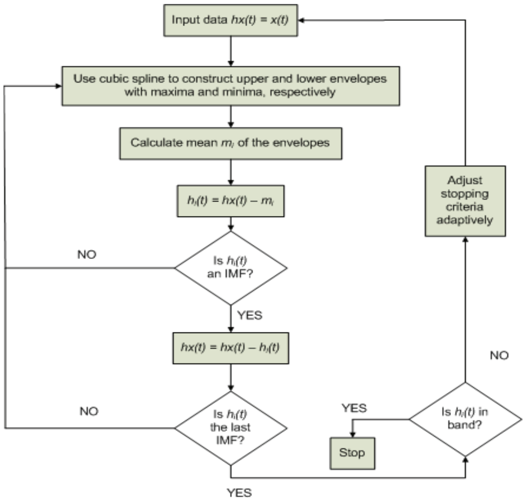

Input data $hx(t) = x(t)$
Use cubic spline to construct upper and lower envelopes with maxima and minima, respectively
Calculate mean m_i of the envelopes
$h_i(t) = hx(t) – m_i$
NO
Is $h_i(t)$ an IMF?
YES
$hx(t) = hx(t) – h_i(t)$
NO
Is $h_i(t)$ the last IMF?
YES
Stop
YES
Is $h_i(t)$ in band?
NO
Adjust stopping criteria adaptively
YES

Fig. 1. Adaptive Empirical Mode Decomposition

V. Experimental Design

Experimental data for this research came from two sources, one is the Speech Under Simulated and Actual Stress (SUSAS) database, and the other is “Deception Interviews of Volunteers” developed for this research. The former consists of simulated stress and actual stress, the later was the recordings of the volunteers’ verbal answers to the questions using the Modified Zone of Comparison (MZOC) protocol. Results of detecting “simulated stress,” “actual stress,” and “deception interviews” are presented in the next section.

A. Speech Under Simulated and Actual Stress

SUSAS database were used to isolate and identify features unique to stressed voice. The database contains voice samples of 44 speakers, both male and female speaking in five different domains; Talking Styles, Single Tracking Task, Dual Tracking Task, Actual Speech Under Stress, and Psychiatric analysis. The scope of the database allows for an analysis of stress under different types of stress and different speaking styles. Voice samples from the SUSAS database have a 16 bit sample depth and 8 kHz sample rate. The voice samples are stored in uncompressed Pulse Code Modulation (PCM) format to preserve as much information as possible.

B. Detection of Deception Interviews

MZOC was the question protocol used for interviews. The first stage consisted of a pre-interview, during which the volunteer was asked to fill out a questionnaire regarding age, health and normal activities. During this time the working theory behind the research was explained and a review of the questions to be asked in the main interview was conducted. Half of the volunteers were asked to be part of a scenario which involved attempting to deceive the interviewer about items concealed on their person. Those that were part of the scenario were told not to answer truthfully about taking or concealing any items on their person. The remaining volunteers were instructed to answer truthfully about all questions asked. Once the interview was completed, a post-interview was conducted and the results of the session presented to the volunteers.

VI. Results

In this section, results of this research are presented. Figure 2 and figure 3 shows the general concept of using AEMD for stress detection. Figure 2 shows the FFT of a speech under "Neutral," "Medium-Stressed," and "Stressed" conditions. This figure shows the fact that the microtremor ("Neutral" at ~ 12Hz) shifts in frequency to ~15Hz ("Medium") and ~17Hz ("Stressed").

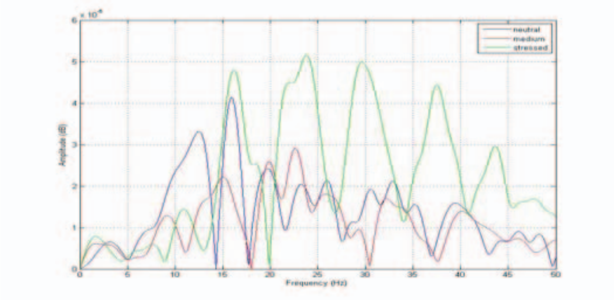
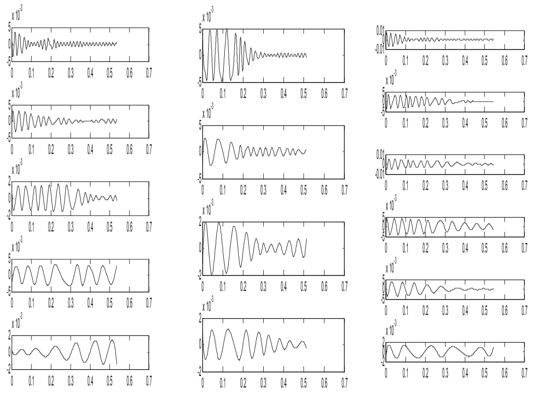


Fig. 2. FFT analysis of a speech under "Neutral," "Medium," and "Stressed" conditions



(a) Neutral (b) Medium Stress (c) High

Fig. 3. IMFs obtained from AEMD using Neutral, Medium, and High stress samples

In figure 3, the IMFs obtained from AEMD are shown. Frequency shifts can be observed in medium and high stress samples comparing with the neutral voice sample. The results are comparable to those obtained by FFT, however, unlike FFT where time resolution is lost due to averaging, AEMD method preserves the "time" resolution and show frequency and amplitude changes along time axis.

The samples used from the simulated portion of the database included 15 different aircraft communication words spoken in 7 different talking styles, and spoken under 3 levels of simulated stress (cond50, cond70 and Lombard). Neutral and training samples were also included. Figure 4 and 5 show the AEMD results of detection of voice stress of male speakers with general US accent and New York accent, respectively.

It can be seen that the angry, cond50, cond70, Lombard and training voice all have frequencies above the 8 to 12 Hz range which is expected for stress type speech. The slow utterance has a frequency below the 8 to 12 Hz range which is also considered a stress response. The stress in the training speech can be attributed to the individual actively making an effort to speak perfectly as the utterance is to be used in speech recognition training. The high frequencies can be attributed to a heightened stress state by the individual at the time the recordings were taken but are most likely due to the accents affecting the performance of the algorithm.

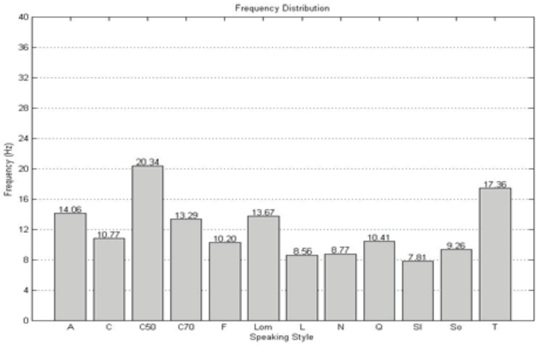


Fig. 4. Simulated Stress: Word "ten" in 12 speaking styles – male with general US accent

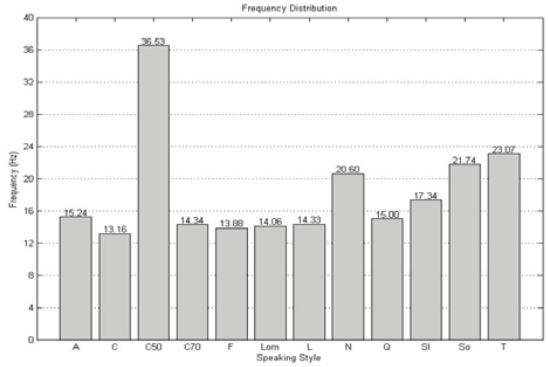


Fig. 5. Simulated Stress: Word "ten" in 12 speaking styles – male with New York US accent

Stressed Speech Styles Key for figure 4 and 5:
N – *Neutral* So – *Soft* C – *Clear* C50 – *Moderate stress condition*
Sl – *Slow* L – *Loud* Q – *Question* C70 – *High Stress Condition*
F – *Fast* A - *Angry* T – *Training voice* Lom – *Lombard Noise effect*

The actual stress portion of the SUSAS database involves both male and female individuals uttering the 35 word vocabulary whilst performing a dual tracking task (medium and high stress), and while on roller coasters. Figure 6 and 7 show a female and male under high and medium dual tracking task stress, neutral and while on the 'scream machine' uttering the word "change," respectively.

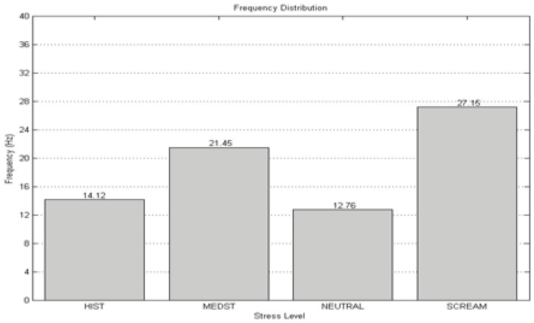


Fig. 6. Actual Stress: Female's speech samples acquired under four conditions

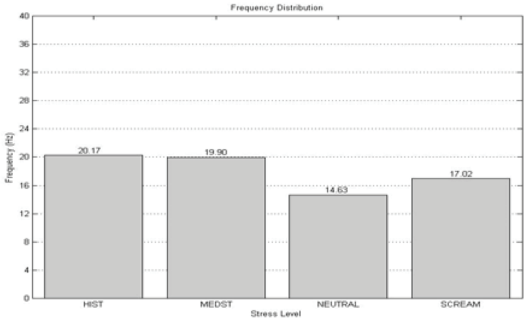


Fig. 7. Actual Stress: Male's speech samples acquired under four conditions

Stress Level Key for figure 6 and 7:
HIST – *Acquisition and Compensatory Task*
MEDST – *Acquisition and Compensatory Task*
NEUTRAL – *Soft* SCREAM – *Scream Machine*

As can be observed from figure 6, the female's microtremor frequency is slightly above the 8 to 12 Hz range in the neutral stress level. However, when stressed, the microtremor frequency moves above the frequency range as expected.

The male's microtremor frequency in figure 7 is seen to be above the 8 to 12 Hz range in all stress levels. In the neutral stress level, this can be attributed to recording of the voice prior to a high stress task which would cause anxiety and hence detectable stress. Both plots show that frquency shifts are consistent with the expected fluctuations, higher frequencies when in a state of stress and lower while calm. Activities conducted prior to or after the neutral samples were recorded affects the quality of the sample by introducing residual stress or anxiety respectively.

The detection of deception interviews were designed to simulate a law enforcement interview where the interviewee attempted to deceive the primary investigator. The design was very similar to that used by Janniro and Cestaro [3] with minor modifications made to increase jeopardy and hence the stress response. Figure 8 and 9 show the results of interviews of a deceptive subject (DS1) and a less deceptive subject (DS2), respectively.

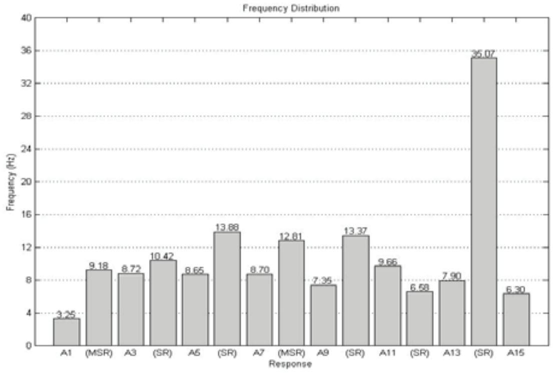


Fig. 8. Detection of deception results summary (DS1)

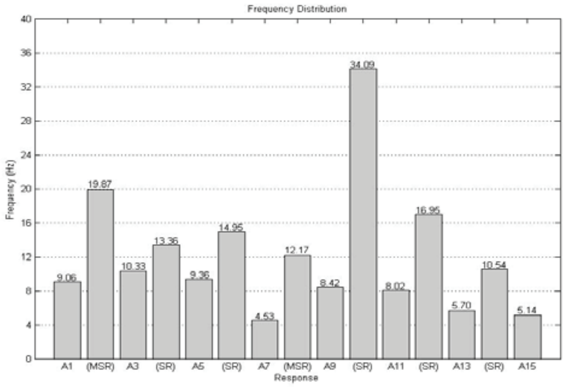


Fig. 9. Detection of deception results summary (DS2)

Stress Level Key for figure 8 and 9:
MSR – *Acquisition and Compensatory Task* *Ax* – Response to question "x"
SR – *Acquisition and Compensatory Task*

Figure 8 shows the frequency distribution of microtremor frequencies in the responses to interview questions for a deceptive subject (DS1). Responses labeled MSR are responses to questions that had some relevance and were designed to evoke a moderate stress response. Responses labeled "SR" are responses to direct questions designed to evoke a stress response. It can be observed that there is initial stress at the beginning of the interview. The microtremor frequency is below the 8 to 12 Hz range which is considered as a stress response. The subject clams down until the sixth response from which each response to each stress question has a frequency out of the 8 to 12 Hz range. The stress becomes very obvious in the last stress response and residual stress can be seen in the following sample (A15).

Subject DS2 was less capable in deceiving the primary investigator and the results can be observed in Figure 9. Each response to the stress questions has a frequency out of the 8 to 12 Hz range and is considered as a stress response. The one exception is the final stress response which has a normal unstressed frequency but is between two stress responses to irrelevant questions meaning there was residual stress from previous responses or the subject subconsciously realized the question pattern and adapted accordingly which is unlikely within such a short time frame.

VII. CONCLUSIONS

Successful detection of voice stress by finding microtremor frequency shift proved the effectiveness of AEMD method. The presence of the microtremor and the effect of stress on its frequency was demonstrated by the results of using samples from the SUSAS database as well as real-world application of the method through analyzing deceptive interview results.

There is a great potential for analyzing stress in speech using EMD. The improved adaptive stop criteria presented in this paper shows the intelligence of this method that greatly enhances the accuracy of the algorithm. The versatility of AEMD allows this algorithm to be modified and adapted to suit the microtremor related applications.

REFERENCES

[1] H. Seyle, “The general adaptation syndrome,” Annual Review of Medicine, vol. 2, pp. 327–342, Feb. 1951.

[2] O. Lippold, “Physiological microtremor,” Scientific American, vol. 224, no. 3,pp. 65–73, Mar. 1971.

[3] M. J. Janniro and V. L. Cestaro, “Effectiveness of detection of deception examinations using the computer voice stress analyzer,” Technical Report, Department of Defense Polygraph Institute, Nov. 1996

[4] J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” *IEEE J. Quantum Electron.*, submitted for publication.

[5] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis,” *Proc. Roy. Soc. Lond. A*, 1998, pp. 903-1005.