Psychological Stress Speech Analysis: A Review

Bhagyalaxmi Jena[1]
[1]Silicon Institute of Technology,
Bhubaneswar

Sudhansu Sekhar Singh[2]
[2]School of Electronics Engineering,
Kiit University, Bhubaneswar

***Abstract*: This paper deals with psychological stress speech signal in a stressful activity. Stress speech signal is different from normal speech signal . The stress can be cognitive or noise induced. Here speaker's stress is based on certain changes in short-time spectrum of vowel phonemes. Two different methods were used to compute the spectrum of each selected signal: Fourier transformation and chirp transformation. Comparation between two spectrum is used to detect the stress of a signal. Speech under stress, gives the higher frequencies observed in the envelope of the chirp spectrum due to enhanced pitch modulation . In this a new database of speech known as "Exam Stress"is created consisting of data collected at our bput exam. Spectrum of speech signal changes gives the indication of emotional condition of a person. The speaker's stress can be detected from each segments of vowels by comparing in the two different transformation .Our long-term goal is to automatically detect and quantify the actual stress influencing a person.**

I. INTRODUCTION

Since the signal is not periodic and non-stationary in nature, it is desired to analyze. Speech is the general way of communication of Human being. Speech Signal is 1-D signal. Stress is the non-specific response of the body to any demand for change.It is complex because the characteristic of speech signal is not periodic. It is the new area of research. Stress is the "non-specific response of the body to any demand for change". Stress is not merely a reaction to something bad, but merely a reaction to a change in situation.

Stress is not only a change in a body response but more specifically a "physical, mental, or emotional strain or tension vowels, are produced. Speech spectrum is the product of the excitation spectrum and the vocal tract frequency response. The purpose of speech is communication. There are several ways of characterizing the communication potential of speech.

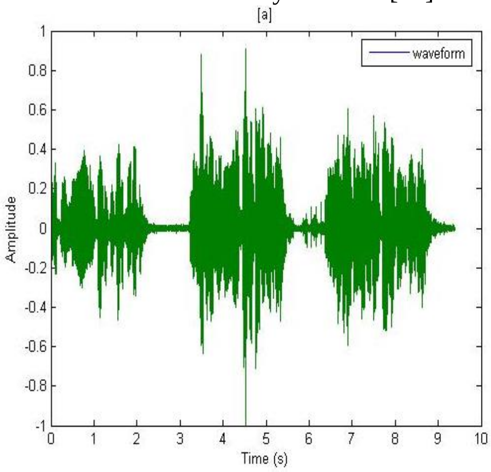
The symbols from which every sound can be classified are called "Phoneme" [11].

Each language has its own distinctive set of Phonemes, typically numbering between 30 and 50. For example, English can be represented by a set of around 42 Phonemes.

In speech communication system, the speech signal is transmitted, stored, and processed in many ways. Technical concerns lead to a wide verity of representation of speech signal. In general, there are 2 major concerns in many systems [12]:

1) Maintaining the content of the message signal.
2) Convenient or suitable representation of speech signal for flexible and recoverable transmission or storage, without introducing serious degradation while processing.

The Representation of the speech signal must be such that the information content can easily be extracted by human listeners or automation by machine [13].



(Fig 1. Time domain Representation of speech)

II. STRESS

Stress may be defined as a condition that forces a speaker to change speech production from neutral conditions. When a speaker is in a 'quiet room' without any task obligations, then the speech which is produced is regarded neutral. Two stress effect areas emerge when we apply this definition namely Psychological and Physiological [19].

2.1. Psychological Stress

Perceptually induced stress occurs when a speaker feels that his environment is different from 'normal environment' in such a way that his intentions to produce speech differs from neutral conditions. The reasons for perceptually induced stress include are emotion, actual task workload (e.g., a pilot in an aircraft cockpit),environmental noise (i.e., the Lombard effect).

2.2. Physiological Stress

Physiologically stress happens because of the physical impact of human body .This leads to deviations from neutral speech production. The different cause of physical stress may be vibration.

III. SHORT TIME ANALYSIS

Speech signal is dynamic with voiced segments and unvoiced segments. The variation in the speech signal is due to vocal cord vibration and vocal tract shape. Non-periodic variations are not under the control of speaker, where as voiced segment speech signals are directly under speaker‟s control. Speech analysis is used to extract related parameters of periodic speech. Speech analysis usually assumes the speech signal properties change slowly with time, hence allowing the examination of short time window of speech to extract parameters presumed to remain fixed for the duration of the window. Most of the techniques yield parameters averaged over the course of the time window. To model dynamic parameters we must divide the signal into successive windows or frames so as to calculate the parameters for the relevant change in the signal.

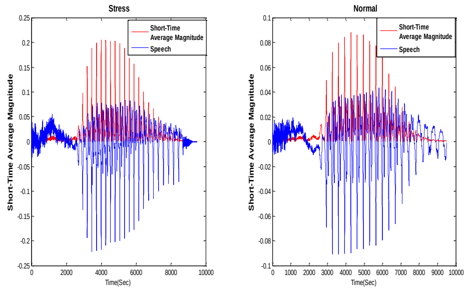
3.1 Time Domain Analysis

The time domain analysis transforms a speech signal into set of parameter signals, which varies very slowly in time than the original signal. This allows more efficient storage or manipulation of the relevant speech parameters than the original signal. To capture the relevant aspects of speech we require several parameters which can be obtained by sampling the signal at lower rate.

3.1.1 Short Time Average Magnitude:

Short Time Average Magnitude (STAM) is used for detecting the starting point and ending point of the speech signal [2].

This measurement is used to classify voiced and unvoiced segments of the speech, therefore unvoiced speech has smaller short- time energy. For the length of the window a practical choice is 10-20 msec for sampling frequency 16kHz.[3]



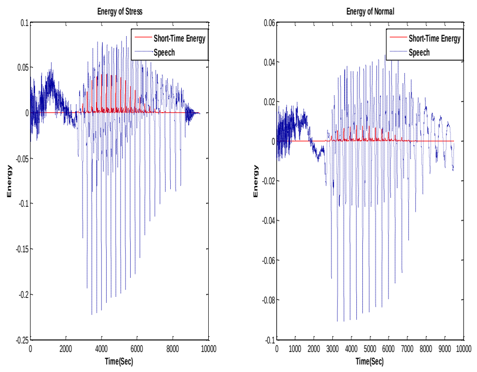
(Fig 3.1.1 Short Time Average Magnitude)

3.1.2 Short Time Energy of Speech Signals:

The short time energy measurement of a speech signal can be used to determine voiced vs. unvoiced speech. Short time energy can also be used to detect the transition from unvoiced to voiced speech and vice versa[2]. The energy of voiced speech is much greater than the energy of unvoiced speech.

The window must be long enough to encompass several pitch periods to produce a smooth representation of the amplitude of the signal. At the same time the window must be short enough to reflect rapid changes in amplitude that occur at the voiced/unvoiced boundaries. The selection of the window size is a compromise since a high pitched female or child's voice may have a pitch period as small as 16 samples at an 8 kHz. sampling rate up to 200 samples for a low pitched male voice. A window size of 160 samples or about 20 msec. is a good compromise. One of the advantages of using a tool like Octave to prototype algorithms is that is that it makes it easier to experiment with parameters like the window size[2].

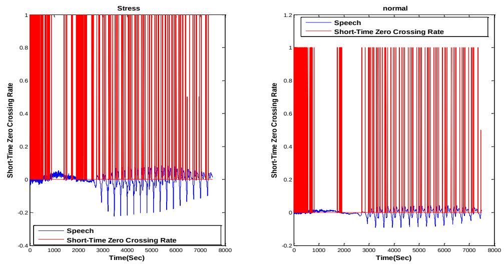
One problem with the short time energy function is that it is very sensitive to large signal levels since the sample values are squared[3]. This isn't a problem in Octave since Octave scales audio samples to +/- 1. In addition a multiply operation is required for each sample.



(Fig 3.1.2 Short Time Energy)

3.1.3 Short Time Zero Crossing Rate of Speech:

The short time average zero crossing rate of a speech signal can be used in conjunction with the short time average energy (or magnitude) to discriminate between voiced speech, unvoiced speech and silence.[3]



(Fig 3.1.3 Short Time Zero Crossing Rate)

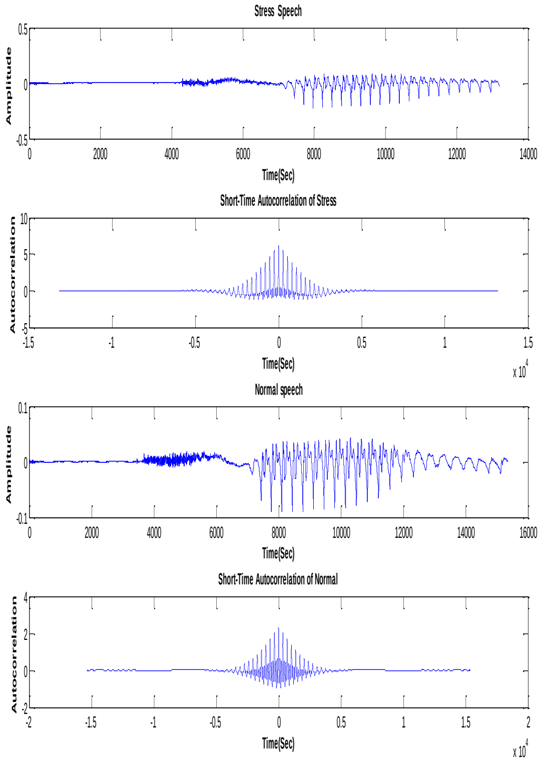
3.1.4 Short time auto correlation :

The mathematical tool used for correlation in signal processing, to analyze the functions or series of values in time domain signals. The mutual relationship between two or more random variables is known as Correlation. The correlation of a signal with itself is known as Auto-correlation[2] .

To find repeated patterns in a signal, autocorrelation is used to determine the signal buried under noise, or identifying the fundamental frequency of a signal which doesn't actually contain that frequency component, but implies it with many harmonic frequencies [5].

Multi-dimensional autocorrelations are defined similarly.

In autocorrelation when the window length becomes shorter, then the attenuation occurs. This happens, when the number of the samples used in the calculation decreases [3].



(Fig 3.1.4 Short time auto correlation)

3.2 Frequency Domain Analysis

3.2.1 Discrete Fourier Transform:

This section is concerned with the frequency domain sampling of an aperiodic finite energy sequence x(n). Fourier analysis is extremely useful for data analysis, as it breaks down a signal into constituent sinusoids of different frequencies.

3.2.2 Fast Fourier Transform (FFT):

The fast Fourier transform (FFT) is an efficient algorithm for computing the DFT of a sequence. Typically the essence of all FFT algorithms is the periodicity and symmetry of the exponential term and the possibility of breaking down a transform into a sum of smaller transforms for subsets of data.

FFT algorithms are based on the principle of

- Decimation-in-time
- Decimation-in-frequency

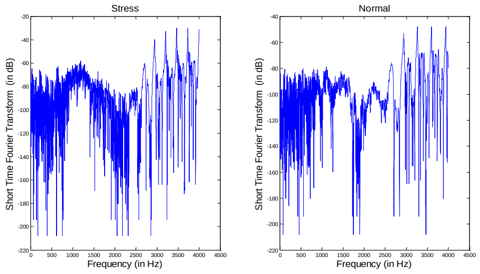
The total number of complex multiplication and addition in , DFT is N^2 and N(N-1) respectively ,where as in FFT total number of complex multiplication and addition is reduced to $(N/2)\log_2 N$ and $N\log_2 N$ respectivily. The FFT algorithms find application in a verity of areas, including linear filtering, correlation and spectrum analysis. Basically, the FFT algorithm is used as an efficient means to compute DFT and IDFT.

3.3 Short-time Fourier transform

This transform gives the concept of a time varying frequency spectrum and the spectrogram. It gives the clarity about the effect of different windows on the spectrogram. It also gives the effectiveness of the window lengthn on frequency and time resolutions.

The plot of the magnitude of the STFT is called the Spectrogram.

$$\mathrm{spectrogram}\{x(t)\} = [X(\tau,\omega)]^2$$



(Fig 3.3Short-time Fourier transform)

3.3.1. Continuous-time STFT

Simply, in the continuous-time case, the function to be transformed is multiplied by a window function which is nonzero for only a short period of time [1]. The Fourier transform (a one-dimensional function) of the resulting signal is taken as the window is slid along the time axis, resulting in a two-dimensional representation of the signal.

3.3.2 .Discrete-time STFT

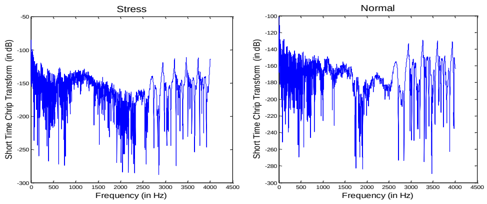
In the discrete time case, the data to be transformed could be broken up into chunks or frames (which usually overlap each other, to reduce artifacts at the boundary) [1]. Each chunk is Fourier transformed, and the complex result is added to a matrix, which records magnitude and phase for each point in time and frequency.

STFTs as well as standard Fourier transforms and other tools are frequently used to analyze music[7]. The spectrogram can, for example, show frequency on the horizontal axis, with the lowest frequencies at left, and the highest at the right. The height of each bar (augmented by color) represents the amplitude of the frequencies within that band [8]. The depth dimension represents time, where each new bar was a separate distinct transform. Audio engineers use this kind of visual to gain information about an audio sample[10], for example, to locate the frequencies of specific noises (especially when used with greater frequency resolution) or to find frequencies which may be more or less resonant in the space where the signal was recorded.

3.3.3 Short-time Chirp transform

The spectrogram can, for example, show frequency on the horizontal axis, with the lowest frequencies at left, and the highest at the right. The height of each bar (augmented by color) represents the amplitude of the frequencies within that band[8]. The depth dimension represents time, where each new bar was a separate distinct transform.

The chirp transformation is a generalization of the Fourier transformation, which corresponds to α=0.. This method is based on combining time-warping with the Fourier transform [1].



(Fig 3.3 Short-time Chirp transforms)

RESULTS

Table.1 (Time Domain Parameters)

	Normal speech	Stress speech
Average Magnitude	0.08	0.2
Energy	0.01	0.05
Zero Crossing Rate	Lesser	Higher
Autocorrelation	2	10

Table.2 (Frequency Domain Parameters)

	Normal speech	Stress speech
Fast Fourier Transform	75dB	90dB
Short time Fourier Transform	-30dB	-50dB
Short time Chirp Transform	-80dB	-100dB

Here in Short time Fourier Transform there is rapid change in both the speech during (1000-2500) Hz. where as in Short time Chirp Transform it is (1000- 2800 approx.) Hz.

CONCLUSION

In this paper we have analysed the vowel part of the recorded speech both in Time domain (i.e Time domain analysis) and Frequency domain (i.e Frequency domain analysis) . Within the time domain as well as in Frequency domain each case the result of stress speech is higher in comparision to normal speech.

APPLICATION

STFTs as well as standard FFT and others tools are used to analyze the speech frequencies [7]. The spectrogram shows the frequency on the horizontal axis, with the lowest frequencies at the left, and the highest frequencies at the right. The height of each bar (augmented by color) represents the amplitude of the frequencies within that band [8]. This kind of visual is to gain information about an audio sample is used by the audio engineers[10], for example, to locate the frequencies of specific noises (especially when used with greater frequency resolution)

FUTURE WORK

In future work we will implement some methodology to show the voiced , unvoiced and silence part of a given speech signal

REFERENCES

[1] Milan Sigmun:"Spectral analysis of speech under stress",IJCSNS International Journal of computer science and Network Security, VOL.7 No.4, April 2007. Institute of Radio Electronics.Brono University of technology 118, CZ-61200 Brono, Crezch Republic.

[2] DOUGLAS O'SHAUHNESSY :" SPEECH COMMUNICATIONS (HUMAN AND MACHINE) " ,Martin Hagm¨uller, Erhard Rank, Gernot Kubin,Signal Processing and Speech Communication Laboratory, Graz University of Technology, 2nd edition,2004

[3] J. H. Hansen and S. E. Ghazale, "Getting started with SUSAS," *Proceedings of Eurospeech'97*. Rhodes, pp. 1743- 1746, 1997

[4] Thomas F. Quatiere: " Discrete –Time Speech Signal Processing, Principles, And Practice", 3rd Edition,2007

[5] Lawrence rabiner and Biing-Hwang Juang:"Fundamental Of Speech Recognition", 2nd edition,2005

[6] L.R. Rabiner and R.W. Scafer : "Digital Processing of speech signal",3rd edition, 2006

[7] John G.Proakis, Dimitris G. Manolakis,D.Sharma:" Digital signal Processing, priciples, Algorithms and Applications".

[8] T L N we,S W Foo, L C De Silva:" Detection of stress and emotion in speech using traditional and FFT Based Log Energy Features", ICICS-PCM 2003.15-18 December 2003,Singapore.

[9] Herman J.M. Steeneken,John H.L .Hansen:" Speech Under Stress Conditions:Overview Of The Effect On Speech Production And On System Performance"

[10] Milan Bostik:" VOICE STRESS RECOGNITION METHODS" Brno University of technology. Faculty of Electrical Engineering and communication. Department of Radio Electronics, Purkynova 118, 61200 Brno, Czech Republic

[11] Doug cairns and John H.L. Hansen:" Nonlinear Analysis And Detection Of Speech Under Stressed Conditions", Journal Of The Acoustical Society Of America, vol.96, no.6, pp,3392-3400,December 1994

[12] Sanjay A.Patil and John H.L.Hasan:" DETECTION OF SPEECH UNDER PHYSICAL STRESS : MODEL DEVELOPMENT , SENSOR SELECTION, AND FEATURE FUSION", Center for Robust Speech system (crss), Erik Jonsson school of electrical Engineering and computer science, university of Texas at Dallas. Richardson, TX75080

[13] J.W.A van Wees, L.J.M Rothrantz, P. Wiggers and R.J van Vark:" Voice Stress Analysis", Data and knowledge systems group , Delft university of technology, mekelweg 4, 2628 CD delft, the Netherlands

[14] John H.L. Hansen, Sahar E. Bou-Ghazale, Ruhi Sarikaya, and Bryan Pellom. Getting started with the SUSAS: " Speech under simulated and actual stress database". Technical Report RSPL-98-10, Robust Speech Processing Laboratory, Duke University, April 1998

[15] Speech Perception .Internet

[16] M. Pantic and L.J.M. Rothkrantz." Facial gesture recognition in face image sequences": A study on facial gestures typical for speech articulation. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 6, page 6 pp., 2002.

[17] J. A. Veltman and A.W. K Gaillard." Physiological indices of workload in a simulated flight task". *Biological Psychology*, 42(3):323–342, 1996.

[18] B.D. Womack and J.H.L Hansen." N-channel hidden markov models for combined stressed speech". *IEEE Trans Speech Audio Proc*, 7(6):668–677, 1999.

[19] Brian D. Womack and John H. L Hansen. "Classification of speech under stress using target driven features". *Speech Comm*, 20(1-2):131–150, 1996.

[20] Guojun Zhou, J.H.L. Hansen, and J.F Kaiser." Nonlinear feature based classification of speech under stress". *IEEE Trans Speech Audio Proc*, 9(3):201–216, 2001.