Optimum Classification of Stress Types in Speech Using Machine Learning and AI

***Mrs. N. P. Dhole , Mrs. S. N. Kale**

[*]Assistant Professor, Department of Electronics and Telecommunication Engineering, PRMIT&R Badnera, (444701), Maharashtra, India,
Email: npdhole@mitra.ac.in
Associate Professor, Department of Applied Electronics, Sant Gadge Baba Amravati University, Amravati, (444602), Maharashtra, India, Email:sujatakale@sgbau.ac.in

Abstract

Human speech is many times the reflection of stress through which the person is going through. Proper evaluation of these speech signals into stress types is necessary in order to ensure that the person is in a healthy state of mind. In this work we propose a novel highly accurate speech to stress classification algorithm, which uses machine learning (ML) and artificial intelligence (AI) combined with sophisticated feature extraction techniques. The machine learning and AI based approach introduced in this paper, uses an intelligent combination of feature selection and neural optimization algorithm which assists the system to learn the speech patterns in real time, and self-train itself in order to improve the classification accuracy of the overall system. We compared our approach with standard neural nets and fuzzy inference classifiers and obtained more than 25% improvement in classification performance. The proposed system is suitable for real time speech and is language and word independent.

Keywords: - Stress Classification, Machine Learning, Feature Selection, Neural Nets, And Fuzzy Inference.

1. Introduction

Classification of human stress levels is a critical component of research for many psychologists and related medical practitioners. This classification can be done from user's speech, physical behaviour, sleep patterns and various other human interactions. While speech and physical behaviour are considered to be the primary parameters for evaluation of stress, sleep patterns, heart rate, and other measurements are secondary parameters for the same. Speech and physical behaviour is considered to be primary because of the fact that these parameters give a near-to-instant analysis of the user stress, while the secondary parameters need to perform some level of pattern analysis over a series observation in order to detect the stress levels and stress types. Out of the speech and physical behaviour patterns, the later demands complex level of processing, right from segmentation, pre-filtering & post-filtering, feature extraction, restricted environmental conditions and other parameters which either affect image processing or biomedical signal processing. These effects reduce the signal capturing capability of the system, thereby reducing the overall classification performance. Thus, speech based systems are best suited to perform the task of stress detection and classification in real time.

This paper is solely based on stress detection and classification from human speech, due to the fact that stress causes Diminished Immunity, Headache, Fatigue, Weight gain, Dyslipidemia, Hypertension, Heart Disease, Psoriasis/Eczema, Digestive problems, and many more diseases. In medical terms, stress is a state of disharmony or a threat to homeostasis which causes physiological

changes increase alertness, focus, and energy and results in perceived demands which may exceed the perceived resources. Some stress types are fruitful, while others have damaging effects, for example, eustress is manageable stress and can lead to growth and enhanced competence, while distress is uncontrollable, prolonged, or overwhelming stress and is destructive. There is also acute stress which arises due to immediate response to a threat or challenge and chronic stress which is due to ongoing exposure to stress, and may seem unrelenting. Thus, it is necessary to detect and control stress in order to have a healthy lifestyle.

This paper proposes a novel algorithm to detect and classify the human speech into different stress classes, and thereby provide a preliminary analysis of the type of stress which the person might be undergoing. Doing this can help the person to analyze the stress and obtain remedies for the same. The next section describes various approaches which have been proposed for stress classification, the section next to that is dedicated to introduce our novel ML and AI based classification system, and finally the results of our technique are compared with the standard neural network and fuzzy inference classifiers to evaluate the performance of the proposed classifier. In the last section of this text we conclude with our observations, and with a few points of finer research which can be undertaken by the readers of this text.

2. Literature review

While most vendors of VSA (Voice stress analysis) technology omit specific details on how their systems work, by studying the basic literature, key information can be extracted on the theory behind the technology. Voice stress analysis originated from the concept that when a person is under stress, micro-muscle tremors (MMT) occur in the muscles that make up the vocal tract which are transmitted through the speech. VSA literature [2] points to a descriptor as the physiological basis for the MMT. This paper describes "a slight oscillation at approximately 10 cycles per second" (i.e. physiological tremors) during the normal contraction of the voluntary muscle. All muscles in the body, including the vocal chords, vibrate in the 8 to 12 Hz range. It is these MMT that the VSA vendors claim to be the sole source of detecting if an individual is lying. In moments of stress, especially if a person is exposed to jeopardy, the body prepares for fight or flight by increasing the readiness of its muscles to spring into action. This in turn causes the muscle vibrations to increase. According to the Merck Manual [3], "enhanced physiologic tremors may be produced by anxiety, stress, fatigue, or metabolic derangements or by certain drugs. VSA systems claim to measure these tremors transmitted through the speech. VSA systems can be broken into two separate categories, energy-based systems and frequency-based systems. Here we will discuss both technologies. The majority of the systems evaluated, as reported by the vendors, are based on the detection of the MMT. From the tests and research conducted during Phase I of our study, it was discovered that it was not the measurement of the MMTs that the signal processing was detecting, but a change in the energy of the spectrum envelope between 20 Hz and 40 Hz [4].

When a voice response is processed through a series of filters, a waveform is displayed that represents the level of stress in the voice. In the event of a non-stress response, the waveform takes on the shape of a Christmas tree (see Figure 1). As stress increases in the subject, the processed waveform takes on a flatter shape. This waveform will flatten out until extreme (hard) stress occurs. See Figure 1(c). Medium stress levels can take on forms that can be a combination of the two extremes. See Figure 1(b). A waveform that shows signs of significant stress is labelled as deceptive when the response corresponds to a relevant question as compared to responses from control questions. This type of technology is known as energy-based VSA. Frequency based VSA systems derive their results from multiple features, primarily identifying changes within frequency bands and the distribution of the frequencies within those bands. Amir Lieberman of Israel developed one such system evaluated during this study. Lieberman identifies the underlying technology as Layered Voice

Analysis (LVA). Dr. J. H. L. Hansen, of the Robust Speech Processing Laboratory, Centre for Spoken Language Understanding, University of Colorado at Boulder, also identified a multi-faceted approach to VSA in his research during the initial phase of this study that included the examination of frequencies and their distribution [5]. This system provides textual responses ranging from truth, through confusion to false statement, with various intervening responses. When examining the data a continuum of stress responses was identified and a comparison of the position of the relevant responses on this continuum was made to the position of the control response to determine if the answers were truthful or deceptive.

The following VSA systems are available,

- VSA-2 000 – is a hardware-based system that uses a digital readout to evaluate the stress levels of the speaker. Tests are recorded and the audio input is processed.
- Computerized Voice Stress Analyzer (CVSA) - is an energy-based detection system that produces a filtered waveform for evaluation and is computer based. Testing is live and multiple waveforms can be displayed on a single page. The audio is passed through the sound card and is automatically directed to the CVSA system.
- Digital Voice Stress Analyzer (DVSA) - is an energy based system that uses waveforms and is capable of displaying multiple patterns per page. Testing with this system is accomplished from data that is recorded, digitized, and then stored in the computer.
- TiPi 6.40 - is a frequency-based system and is the most recent iteration of the TrusterPro. This system automatically segments narrative responses and analyzes each phrase for stress. The system processes live audio in real-time mode and can digitize audio in off-line mode.
- Digoenes Lantern - is an energy-based system that utilizes either live audio feeds or digitally stored audio. The waveforms are displayed individually.

It was not the objective of this review to recommend one Psychophysiological Stress Detection technology over another. Rather, it was our intent to provide users with an unbiased evaluation of VSA technology along with enough information to assist them in making decisions on what type of system to employ. Our results indicated that, given the VSA systems tested, results indicate that this technology, with a trained and experienced examiner is capable of detecting deception or truthfulness in a subject at a rate better than chance. The experience an examiner has with VSA technology plays a key role in their ability to detect deception. These instruments alone are not “lie detectors”. The decision as to whether a subject is being truthful or lying should only be made by a trained examiner. This decision should be based upon reviewing the data presented by the instrument, the demeanour of the subject, and other evidence from the case. VSA systems are capable of providing an examiner with a waveform or other response that may be a reasonable reflection of the stress level being experienced by the subject, in a majority of the cases. The correct interpretation of this indicator is the responsibility of the examiner. The goal in using a VSA system or polygraph should be to convince the subject that they cannot deceive the operator, and that the instrument will detect their deception and their best avenue is to confess to the crime. This study has shown that VSA systems will produce results that trained operators can employ with confidence to obtain confessions. The results of these examinations should not be considered “proof positive” of innocence or guilt. While some consider these tools as an aid in focusing an investigation on proving someone guilty, officers should not lose sight of other suspects or evidence that may indicate otherwise.

This study has also shown that when training and experience of an examiner are taken into consideration, test results indicate that over time there is a marked improvement in an examiners ability to correctly identify deceptive subjects. Likewise, the comparison between the two analysts of different skill levels also indicates that experience may be a factor in improving accuracy. Observations made during the study of other analysts seems to indicate that the more opportunities

one is given to run tests, examine charts and receive feedback (ground truth), the better the examiner becomes.

3. Machine learning and AI based stress classifier

The proposed classifier is based on real time learning; it uses a combination of feature selection and neuron optimization. It has the following components,

- Feature selection unit (FSU)
- Neuron optimization unit (NOU)
- Machine learning look up tables (MLUTs)
- AI selection unit (ASU)

The feature selection unit or FSU extracts the mel frequency cepstral components (MFCCs) from the input speech. The MFCCs consists of the frequency responses, the reconstructed frequency bands and the pitch of the input speech. In total, for a 1 second speech signal sampled at 44100 Hz, we obtain more than 22 million feature values. These feature values describe the speech independent of the spoken languages and words; thereby can be used for language and word independent classification of speech. But, due to its large feature length, these feature vectors are unusable for neural network training and thus there is a need of feature reduction. Wavelet transforms not only preserve the patterns of the signals but also assist in reducing the feature length in order to obtain well defined features which are widely usable for classification purposes. From our study and observations, the daubichies 8 (DB8) variant of the wavelet transform is the most suitable for feature reduction and pattern preservation. Thus, our FSU uses the DB8 transform in order to produce variable length feature patterns for training the neural network classifier.

The neuron optimization unit or NOU handles all the operations related to neuron selection and processing. Our proposed classifier uses a series of trained neural networks, each of which have intelligently selected neurons for the best selected features and most optimum accuracy. The NOU performs neuron selection by observing the resulting accuracy, and then refining the number of neurons in each layer of the network until a sufficient level of accuracy is achieved. These observations are then stored in the machine learning look up tables (MLUTs) which have the following structure,

- Selected features
- Number of neurons in input layer
- Learning function
- Number of training samples
- Obtained accuracy

The MLUTs are filled using the following algorithm,

1. Let the number of iterations be Ni
2. Let the number of solutions per iteration be Ns
3. Let the maximum number of samples in training set be Nmax, and the minimum number of samples for training be Nmin
4. For each solution 's', in Ns, perform the following steps,
    a) Select a random number between Nmin and Nmax = Nr
    b) Apply DB8 wavelet transform to each of the training samples, until the features of each sample are reduced to Nr
    c) Select random number of neurons, such that selected neurons are more than Nr
    d) Train the neural net with Nr features, and evaluate its accuracy Ai
5. Evaluate the mean accuracy, and discard all solutions where accuracy is less than the mean accuracy

6. Repeat steps 4 and 5 for Ni iterations, and tabulate the results in the MLUT

Once the MLUT is filled with sufficient entries, the AI selection performs classification of the evaluation speech signal. This performs neural net evaluation for all the neural nets stored in the MLUT, and obtains the classification results from each of the entries. From all these results, the most frequently occurring class is selected, and the solutions which classified the evaluation sequence to that class are marked with +1, rest others are marked with -1. After several evaluations, all the negatively marked solutions are discarded, and the MLUT filling algorithm is applied to fill in those discarded slots. But, during this step, the previously evaluated signals are also taken into the training set, so that the system can self-learn and be adaptive to changing inputs.

We tested our proposed algorithm on different databases, and the results & observations are described in the next section. We also compared the proposed approach with existing neural network based approaches, and obtained a 25% improvement in overall system performance in terms of classification accuracy of the system.

4. Results and Analysis

We tested our stress detection systems under 5 different categories, namely,

- Stress Type 1
- Stress Type 2
- Stress Type 3
- Stress Type 4
- No Stress

We took 100 voice samples per category, and tested them under different languages and different spoken words. The following accuracy results were obtained when we compared our system with Neural Networks and ANFIS,

Table 1. Comparison of NN and ANFIS with Proposed Algorithm

No. of entries in dB	No. of entries tested	Accuracy NN (%)	Accuracy ANFIS (%)	Accuracy Proposed (%)
10	12	60	50	85
20	25	65	52	91
30	39	68	54	93
40	55	70	56	94
50	65	72.5	57	95
75	85	75.8	60	96.5
90	100	77.2	62.4	97.5
100	120	78.1	65.2	97.8

We observe that the accuracy for ANFIS is lowest, while our algorithm outperforms NN by at least 20%. This is due to the fact that our machine learning algorithm uses deep nets for training, thus reducing the error in accuracy, and making the system more efficient in terms of classification accuracy. The accuracy saturates around 98%, and doesn't change much even after increasing the size of the database. Figure 1 shows the graph visualizes the accuracy variation,

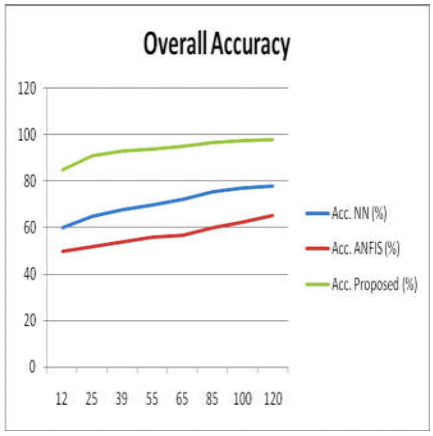

Figure 1. Overall Accuracy Variation of NN, ANFIS and proposed system

We tested our system for different number of solutions and iterations, and found the following relation of number of training sounds, iterations and solutions with accuracy,

Table 2. Results for different number of solutions and iterations of Proposed System

Sounds	Iterations	Solutions	Acc. (%)
100	10	10	68
100	20	10	69
100	50	10	70
100	100	10	70
100	200	10	70
100	50	20	70
100	100	20	75
100	200	20	75
100	300	20	75
100	50	30	78
100	100	30	80
100	200	30	84
100	100	40	86
100	200	40	89
100	300	40	91
100	300	60	95
100	450	85	97.8
100	500	90	97.8
100	600	100	97.8
100	1000	500	97.8

As the numbers of solutions are increased, the accuracy increases, but it mildly dependent on the number of iterations as well. Thus, in our case for 100 sound samples, the accuracy saturates at 97.8% for 450 iterations and 85 solutions. This number can change as the number of input samples are changed, and therefore it is for the network designer to test and train the machine learning and AI layer in order to get the satisfactory results based on varying number of iterations and varying number of solutions.

5. Conclusion

From our results it is evident that the accuracy of our proposed classifier is better than existing standard neural network and ANFIS classifiers. Our system outperforms the neural network classifier by more than 20% and outperforms ANFIS classifier by more than 30%. Our system also performs with good accuracy even under different languages and different spoken texts, and is able to distinguish between different stress types. In future, we plan to compare our performance with other bio-inspired classifiers like PSO, SVM and BFO in order to further optimize the accuracy of the system.

References

[1] National Institute for Truth Verification, http://www.cvsa1.com/cost.php.

[2] O. Lippold, "Physiological Tremor", Scientific American, Vol. 224, No. 3, March (1971).

[3] M. H. Beers and Robert Berkow, "The Merck Manual of Diagnosis and Therapy", John Wiley & Sons, 17th Edition, (1999).

[4] D. Haddad, et.al, "Investigation and Evaluation of Voice Stress Analysis Technology". Final Report for National Institute of Justice, Interagency Agreement 98- LB-R-013 Washington, DC, 2002, NCJRS, NCJ 193832.

[5] J. H. I. Hansen, et.al. "Methods for Voice Stress Analysis and Classification, as appendix to Investigation and Evaluation of Voice Stress Analysis Technology", Final Report for National Institute of Justice, Interagency Agreement 98-LB-R-013. Washington, DC, 2002. NCJRS, NCJ 193832.

[6] D. Lykken, "A Tremor in the Blood, Uses and Abuses of the Lie Detectors", New York, McGraw-Hill, (1981).

[7] "National Research Council of the National Academies", The Polygraph and Lie Detection, Washington DC, National Academies Press.