Distress Recognition from Speech Analysis: A Pairwise Association Rules-Based Approach

Daniel Isozo Machanje, Joseph Onderi Orero, Christophe Marsala

Distress Recognition from Speech Analysis: A Pairwise Association Rules-Based Approach

Daniel Isoso Machanje
Faculty of Information Technology
Strathmore University
Nairobi, Kenya
Email: dmachanje@strathmore.edu

Joseph Onderi Orero
Faculty of Information Technology
Strathmore University
Nairobi, Kenya
Email: jorero@strathmore.edu

Christophe Marsala
Sorbonne Université, LIP6
CNRS
F-75005 Paris, France
Email: Christophe.Marsala@lip6.fr

***Abstract*—Monitoring of the elderly has previously been done using visual, physiological devices, and physical nurse rounds. These have breached privacy and caused further health scares. The use of speech to read distress emotions is an alternative that, if utilized well, would provide effective monitoring while preserving the subject's privacy. Classical speech features have been used to feed machine learning algorithms to facilitate emotion detection. However, given the difference that individuals exhibit with their speech feature baselines with regard to the manifestation of distress on their speech given, it would make it difficult to normalize the features. The same case would apply for speech of an individual under different emotional circumstances. This paper proposes a novel approach where association rules drawn from speech features are used to derive the correlation between features and feed these correlations to machine learning techniques for distress detection. In achieving this, extraction of periodic segments is done, where each of these segments' features is paired with adjacent segments features, and their correlation percentages compared with other pairs within the same sound file, in order to establish a correlation between emotion features using association rules, creating defined rules that indicate specific emotions.**

***Keywords*—Association Rules; Speech; Emotion; Distress; Elderly; Monitoring.**

I. Introduction

According to the World Bank, the elderly are people who are aged 65 and above. The population of the elderly has steadily increased from 4.974% in 1960 to 8.696% of the total global human population as of 2017 [1]. Given the average retiring age of 62 years for women and 64 years for men [2], most elderly people at these ages are retired from active economic activities. In Africa, the aged bracket represented 3.5% in 2015 representing 15 millions [3] while in Kenya, the same bracket represented 3% in 2017 representing a population of 1.42 millions [4].

Health and wellness of the elderly is a core focus globally. This has been clearly outlined in the third global goal for sustainable development among other 16 that were formulated by world leaders; good health and well-being [5]. As they get older, the elderly's functionality reduces in a phase they are referred to as the *transitioners* [6]. Eventually, in their sunset years, the elderly become fully dependent on younger citizens with regard to the four aspects as earlier mentioned; social, identity, routine, and daily activities. This phase, in which they are usually referred to as the *strugglers* [6], is the most delicate, and as such, require round the clock monitoring and care. A depiction of these phases is made in Figure 1. The care the elderly require arises from the distress they start exhibiting during their sunset years.

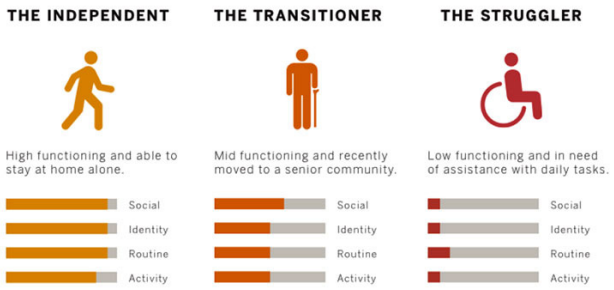

Fig. 1: Aging phases among the elderly [6]

Distress is an emotional state resulting to negative effect with mood ranges that include disgust, anger, scorn, guilt, fearfulness, and depression [7]. Distress can also be defined as a mental distress, suffering or anguish resulting to unpleasant mental reactions such as fright, nervousness, anger, grief, anxiety, worry, mortification, shock, humiliation, indignity and physical pain [8]. The detection of distress has gained traction in the last few years as a result of its ability to communicate a need or want for individuals to get immediate help. Many advancements have been made with an aim to try and detect distress at high accuracy, with the most focus being on visual signals [9], courtesy of an established infrastructure of closed-circuit televisions used for monitoring. Physiological signals have also been used, with the main challenge being the technicalities involved in installing body sensors that would easily read diverse human physiological signals [10]. Amidst these advancements, speech has been given a wide berth, with interest in this area just gaining ground in the past decade. This is despite the strong justification speech has such as few resources required for collection and processing, and the need for privacy that physiological and visual users sometimes have to forego given their intrusion. Use of speech in distress detection has the potential of a myriad of applications, key being in health monitoring systems and security monitoring.

Detection of compound emotions from human speech has been perceived as a complicated routine in the realm of emotion detection. Compound emotions are emotional states with a composition existence of more than one emotion; a case in context being *distress* [11]. Unlike the ease that comes with detection of primary emotions as has been described in many emotion experiments, where feature extraction is in one dimension, the detection of distress from speech requires the extraction of features in two dimensions given the diversity of emotions it might be harboring. In the same study, 2D dimension was proposed, where two perspectives: excitement level and polarity are proposed in the detection of distress. This new approach proves to result in higher accuracy results in comparison with other extraction techniques.

The main challenge that faces the use of classical speech feature extraction for distress emotion detection is normalization. There are lots of variation of speech from one individual to another due to difference manifestations of distress to the human speech as a result of different upbringing and training, cultures, diverse lingual backgrounds, and origins. Such manifestation of distress on the human speech, specifically on the laryngeal musculature, glottis, articulation rate, and pitch have been explained with great detail in [12] [13] [14]. The challenge arises by the difference in baselines of the possible features when extracting speech features from the human voice, as a given measurement value in an individual during distressful moment might differ with the measured value of another individual during a similar scenario. There is also the possibility of the same individual exhibiting different baselines in similar distressful scenarios depending on immediate prior or current circumstances the individual might be experiencing. Additionally, double extraction and multi-dimension processing in classifying compound emotions might prove resource-intensive. This also requires that a given speech set has met threshold of extraction before the subsequent processes can take place; which in some case does not happen given the tense application scenarios where such technologies are applies which include health and security.

The emergence of association rules was a welcome idea in data mining given the vast application areas, especially in large database mining. With such rules, the discovery of associations between items in a database becomes easier as compared to the extraction of features from these items, which usually require a lot of resources and time.

This paper aspires to propose the use of association rules to derive the correlation between speech features; rules that will eventually be used to feed machine learning techniques for distress emotion detection, and eventually, the detection of various other compound emotions. The following is the paper's plan: in Section 2, we present related works that are able to highlight various aspects with regard to distress definition, speech features, various applications of association rules, and diverse machine learning approaches that have been used in detecting distress as an emotion. In Section 3, a detail of our proposal on the use of association rules as an alternative of classical speech features is discussed, before a conclusion is drawn in Section 4.

II. Related Works

This section has three main parts: distress, speech features, and association rules. In the distress part, a detailed manifestation of the distress emotion on human voice is explained. Speech features entail the classical speech features that have been used so far in detecting distress. The final subsection; association rules, entail the description of the rules and give a brief overview of its application areas.

A. Effects of Emotional Distress on Human Speech

It has been proven that emotions in humans affect their voice, and distress is not an exception to this. There are several ways that distress in humans affect their voice [13]. Muscle tension is one of the key effects imparted by distress on humans. For instance, the laryngeal musculature is adversely affected during distress situations, thus limiting adjustment of vocal cords [12], an effect that affects the voice quality in a certain way. Apart from muscular tensions, unsettledness suffered by victims in distress causes jitters [13]. Distress also generally affects the glottis, a significant body part in the area of the throat. Effect on the glottis causes voice modulation [12].

Fluctuating respiration by the subject under distress affects the subglottal pressure. This, by extension, affects the pitch, speech duration, and articulation rate [14]. Pitch can simply be defined as the highs and lows of sound judgment with association of musical melodies [15] or scientifically as sound property that can be ordered in a frequency related scale (fundamental frequency) [14]. Articulation is a sequence of signals with regard to mouth and mouth-related movements that recurs to make sounds of speech. It is also referred to as temporal pattern [14].

In their research investigating the effects of stress on speech, [16] hypothesize a stressor/stress relation in which they propose four orders: *zero-order*, *first-order*, *second-order*, and *third order* [16]. These stressor orders are indicated in Figure 2, with brief explanations on their effects to speech. In the zero-order, physical change to an individual's speech production apparatus occurs as a result of the stressor, resulting in, for instance, vibration that would cause variation of the elements that produce sounds, thus changing the speech. In the first-order, physiological changes to the apparatus for speech production occur as a result of the stressor, thus causing changes such as the vocal cords distending. In this order, diversity starts appearing in different individuals speeches as a result of the differences in their physical characteristics.

In the second-order, [16] suggest psychological changes occur as a result of the stressor. This occurs through the individual's consciousness, thus varying in great diversity as a result of the individual's current situation. This order is the one mostly connected to the emotions of the human individual. The third-order includes additional effects that may be caused by additional interpretation of the context, leading to different levels of voice raising or lowering.

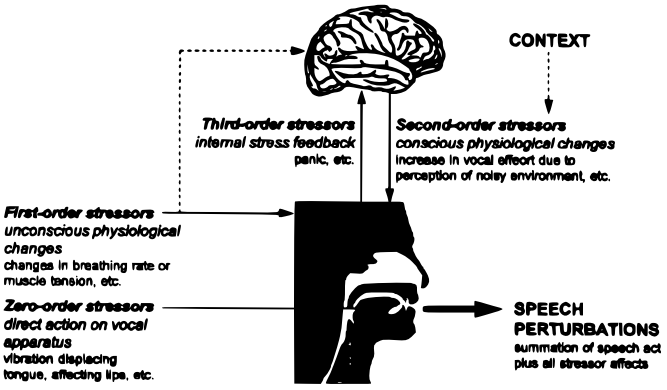

Fig. 2: Stressor orders alongside their effects on speech [16]

Hansen and Patil (2007) further support and emphasize Murray et al.'s (1996) hypothesis by rephrasing the categories of stress orders [14]: physical stressors, unconscious physiological stressors, conscious physiological stressors, and internal stress feedback stressors. These are highlighted in Figure 3.

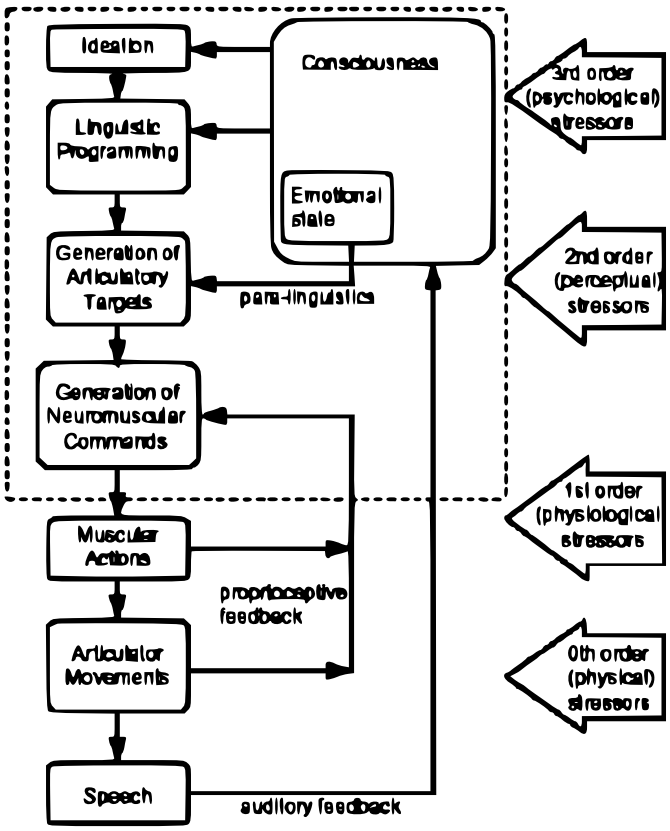

Fig. 3: Speech production process with the effect of stressors [14]

B. Existing State of Affairs for Distress Detection

This part explains past machine learning techniques that have been used in distress detection. The text further goes ahead and describes the process of classical speech feature extraction, and finally gives detailed information on various classical speech features that have been used in the past for distress detection.

1) Feature Extraction and Selection: Feature extraction is the process of obtaining speech features from sound signals. This process can be done by a multitude of tools, with the most popular one being Praat [17]. Such environments have defined baselines for extracting most popular speech features, with the provision to write and run independent scripts for more customized speech feature extractions.

After the extraction is done, and before training and testing proceeds, feature selection is a common process that involves choosing the best performing features and reducing dimensionality. There exists a number of feature selection techniques that include *forward selection (FS)* and *promising first selection (PFS)* [18]. Choice of feature selection techniques vary from one experiment to the other depending on preference and the accuracy that the tools offer in given contexts.

2) Speech Features: Classical speech features used in speech emotion detection are categorized into two main categories: phonetic features and prosodic features [19]. Phonetic features describe the types of different sound such as vowels, consonants, and the general pronunciation, while prosodic features describe the musical aspect of the speech. By musical aspects, the following features are implied: pitch, power/energy of the voice, speaking rates, spectral tilt, voiced trajectory duration and others.

There are various categories of speech segmentations that research in distress has focused on: whole signal versus utterances/sequences and global versus local configurations. Whole signals represent entire audio files in a database, while utterances/sequences are split segments with regard to different emotions in a given audio file [20]. Global configuration is segmentation of the audio file into utterances, while local configuration is the segmentation of speech into small fixed-size time windows. Global configurations/utterances have been known to give the best accuracy in detection [21].

The identification of emotions in an emotional speech databases is done through labeling and annotation. Usually, human labelers, based on a given criteria, are used. These could be professional or amateur labelers who could be used to eliminate biasness. Sometimes, databases are presented to the research community without labels and the researchers themselves have to label and annotate the data independently. Criteria of choosing *qualified* audio files ranges from number of agreed labelers to the percentage average of labeling accuracy.

Data extracted from speech is presented in diverse ranges that usually requires preprocessing in order to acquire optimal results from various machine learning techniques. The three main data preprocessing techniques include rescaling, normalization, and standardization.

Given the richness of features in speech audio files, sometimes too many features are extracted. Most of the features end up not being important for the final detection of the distress emotion. Therefore, to optimize the model, various

experiments utilize various feature selection tools. These include brute-force iterative feature selection algorithm [22], promising first selection (PFS) [18], forward selection [18], Fisher selection [20], ReliefF algorithm [21], and principal component analysis (PCA) [21]. The most efficient feature selection is forward selection as proposed by [18] through the research that they got to compare between promising first selection (PFS) and forward selection (FS). This is the same feature selection used by Machanje et al. (2018).

3) Machine Learning Techniques for Distress Detection: [23] has done a detailed comparison of various machine learning techniques that have so far been used in detecting emotions in general. In their comparison, they identified the following techniques as having already been used in detecting distress: Hidden Markov Model (HMM), Gaussian Mixture Models (GMM), Artificial Neural Networks (ANN), and Support Vector Machine (SVM). Under each of these techniques, [23] were able to come up with respective constraints that surrounded experimentations with regard to emotion detection [23]. For instance, GMM has in the past depicted limitations to modeling independent vectors and also gives a challenge in determining the optimum number of Gaussian components. On the other hand, HMM has been found to have topological restrictions that restrict analysis from only left to right and also a challenge in determining the optimal number of states within a data set. ANN has been cited to be only effective to nonlinear mappings, while SVM has had discrepancies in detection accuracies when presented with data that exhibits speaker dependence and independence.

In the detection of distress, various researches have used a diversity of machine learning techniques. Ang et al. (2002) used Decision tree Model to detect annoyance and frustration, gaining an accuracy of 76%. In this experiment, 49,553 utterances were used, with a 75-25% split for training and testing model being used. The dataset used was generated through acting, comprising of air travel arrangements through phone calls. Lee et al. (2001) used Linear Discriminant Classifier (LDC) and K-nearest neighbor (KNN) to detect negative and non-negative emotions, attaining an accuracy detection of 77%. The dataset utilized in this case was natural speech comprised of 1187 real telephone calls brought together by Speech Works. The features used comprised pitch and energy.

Subsequent researchers in Clavel et al. (2008), Lefter et al. (2011) and Alkaher et al. (2016) have used GMM and SVM to detect fear, neutral vs. Emotional, and Distress (fear/anxiety) respectively, with detection accuracies of 71% and 97.9%. Lefter et al. (2011) utilized a dataset with 3000 utterances from a South African call center. Each utterance comprised of averagely 4.30 seconds, with a total of 215.02 minutes. Clavel et al. (2008) usde a fictional corpus with 400 audio-visual files. The prosodic features used comprised of pitch, intensity, voiced trajectory, and duration. Alkaher et al. (2016) used the Berlin Database of Emotional Studies, containing 535 utterances with 7 emotions; anger, joy, boredom, sadness, disgust, fear, and neutral. Teager Energy Operator (TEO) based features were used, with a 75-25% model used for training and testing. All these emotions treated by these researchers were assumed to represent distress in the given contexts that their data was collected. Machanje et al. (2018) used their proposed 2D approach, with *fuzzy* K-NN as the classifier to attain a detection accuracy of 86.64% using the Berlin Database of Emotional Studies [24]. This is a simulated database containing 535 utterances with the emotions anger, fear, joy, boredom, disgust, sadness, and neutral.

C. Association Rules

The main idea behind association rules is the ability of discovering relevant associations between diverse attribute pairs [25]. This is implied by the expression $X \Rightarrow Y$ in the context where X and Y are item sets, implying that, any transactions that contains X, also tend to contain Y [26].

At the onset of association rules by the real proponents [26], the main application that association rules were devoted to were to solely reveal patterns in transactional databases especially in the business, more specifically, retail industry. The classical explanation to demonstrate the association as drawn in the first paragraph of this subsection was an analysis of diverse goods within a retail business; to find if the customers that bought a certain category of goods a also bought a certain kind of goods b. By getting such relationships, it would be easier for a retail chain to decide on the complementary goods to stock when restocking, so as to ensure more sales to the customers and also ensure reliability by ensuring that the customer is able to precisely get what he/she wanted, informed by the probability calculated through an association rule.

With regard to application areas, association rules now cover a larger spectrum of applications apart from the initial intention within the retail industry and business in general. They have been used in medical diagnosis, big data mining, such as census data, market analysis, and nutrition. This paper endeavors to add on a novel application area.

III. Association Rules-Based Approach

The association rules-based approach we are proposing targets to eliminate the key challenge of normalization that is experienced when aggregating statistics for classical speech features. This challenge arises from the difference in baselines that different individuals portray during times of distress. To expound on this, an individual A from a different background can have a digital measurement value of e when distressed, while an individual B, during a very similar scenario that A was in, could exhibit a different baseline of f in his/her production of distressed speech. In a different scenario, it can go as far as a different individual, D, producing distressed speech at a given scenario g that would have a completely different baseline the same person would produce in a different distressful environment.

In our proposal, association rules drawn from speech features are used to derive the correlation between features and feed these correlations to machine learning techniques for distress detection. To achieve this, during model training and testing, each sound file is split into segments of defined

duration. Speech features are then extracted from each of these segments. Each of these segments' features are then paired with the adjacent segments' features, and their correlation percentages compared with other pairs within the same sound file in order to establish a trend within a given emotion using association rules. This is iterated till all the segments' features on a single sound file are compared to other segments, each of the same time length *t*. During the derivation of association rules among the pairs, the key considerations are the presence and absence of certain feature measurement value changes within each segment. Presence is detected by the increase of the measured value by a certain speech feature, for instance, pitch, while absence is portrayed by the decrease in measured value of a certain feature, say jitter. These pairs are then compared to achieve diverse itemsets that will eventually give rise to percentage correlations from one segment to the other.

Using these association rules correlations, the classical classifiers can then be fed for training and testing, serving as a non-discriminative way of detecting distress as compared to the classical methods where aggregate feature measurement values such as mean, median, maximum, and minimum were used to feed classifiers for training and testing of a model. This supports the diversity in speaker independence and ensures accuracy despite the origin of given individuals, as detection is considered based on a certain rise (presence) or fall (absence) of certain speech feature pairs. This process flow is demonstrated in Figure 4

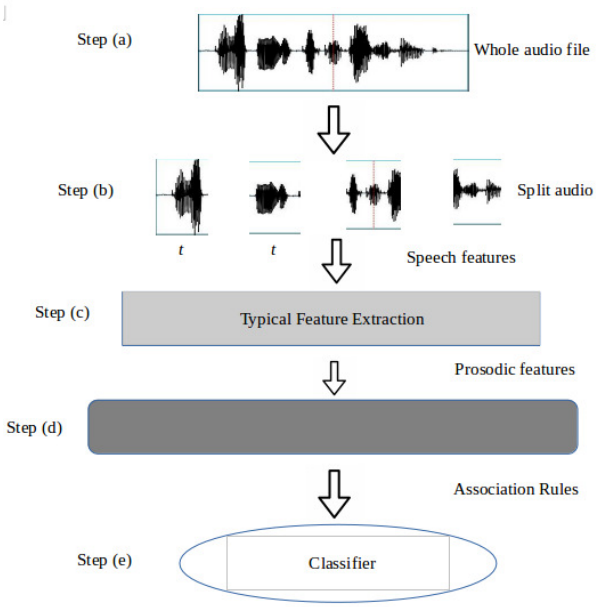

Fig. 4: Proposed approach

IV. Experiment

A. The Berlin Database of Emotional Studies

The Berlin Database of Emotional Studies [24] is the database of focus. It is a simulated database, containing 535 utterances of the emotions anger, fear, joy, sadness, disgust, boredom, and neutral. The actors involved included 5 males and 5 females under controlled environment, pronouncing German phrases under instructions to mimic the emotions as mentioned [24].

B. Sound Splitting

To facilitate sound splitting, and anticipate feature extraction (next step after sound splitting), each of the 535 files has their textgrids generated. A textgrid [17] object facilitates the marking of various periods within a sound file to enhance accurate feature extraction or better definition or description of a sound file. This process is done using Praat scripts on the Praat tool [17].

The 535 sound files have different durations, with the least lasting for 1.2255 seconds, while the longest sound file lasts for 7.6794375 seconds. The majority number of sound files, however, last between 1.5 seconds to 3.5 seconds. In deciding the optimal split duration, a consideration was made based on the time step; a measurement value interval for a frame that determines the pitch concentration per given second, which is usually also affected by pitch floor. Pitch floor is a measurement value that determines the length of an analysis window, usually represented by the lowest frequency that is usually targeted by each sample. Given the pitch floor in this case being set at 75 Hz, the time step is equivalent to 0.1 seconds, prompting Praat to calculate 100 pitch values in every second. A sound file with duration of 0.04 seconds should, therefore, be long enough to contain at least 3 periods (3/75). This duration is sufficient to calculate at least the speech feature of pitch, which is by far the most relevant prosodic feature in speech analysis [27]. A study of the influence of this value in the whole process will be conducted in future works.

Given the lengths of the sound files contained in BDES, and given the metrics to obtain optimal sound file length, the sound splits from the larger sound files are 0.4 (10 times longer than required threshold) seconds long, ensuring that the threshold to obtain the optimal length of a sound file to be analysed for speech analysis is met, and also ensuring that there is no potential of having completely voiceless segments of the sound. In previous experiments, voiceless segments have been eliminated, but given the novelty that this research aims to introduce (association rules), it is important to have all segments inclusive so that the comparison from segment to segment is more conclusive enough to come up with relevant association rules.

The segments split in this experiment will be a minimum of 3 on the 1.2255 seconds file and a maximum of 19 on the longest file of 7.6794375 seconds. The remaining section of the file not meeting the threshold of 0.4 seconds is truncated and dropped from any further experimental phases. The praat script used to split the sound files alongside their textgrids names the new files as parts of the previous name, and generates a csv file for each initial sound file containing the starting and ending point of each part.

A total of 3,453 sound segments with lengths of 0.4 seconds each were split from the original 535 sound files of the BDES. Alongside these are corresponding 3,453 textgrid segments. Figures 5 and 6.

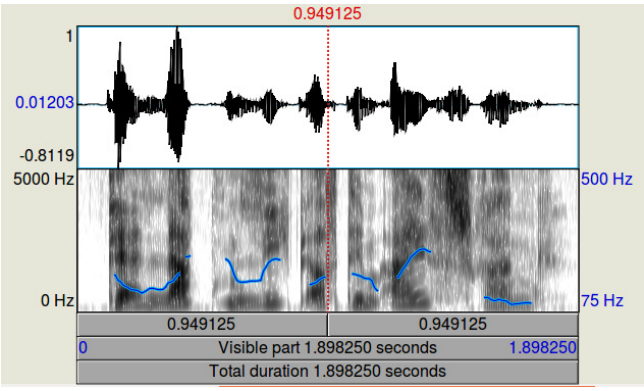

Fig. 5: A whole sound file wave: 03a01Fa.wav with length 1.898250 seconds

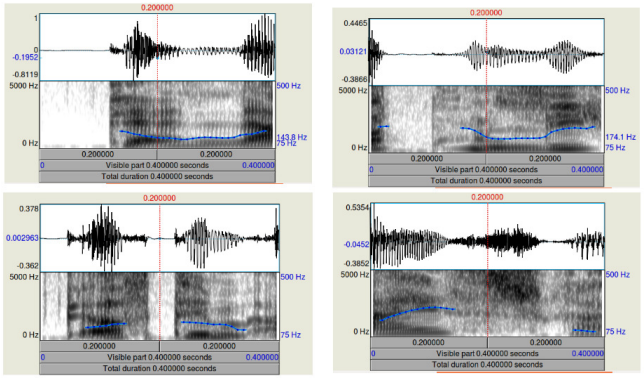

Fig. 6: 4 segments measuring 0.4 seconds each split from the file 03a01Fa.wav, starting from top-left, moving clockwise

C. Feature Extraction

The successive phase in the experiment is feature extraction. Feature extraction was done using Praat, and given the availability of textgrids alongside the *wav* sound files, this process required a simple praat script to extract the prosodic features. The extraction was done at an analysis width of a window range that was 0.005 seconds. The maximum frequency was set at 5000Hz, with the frequency step set at 20Hz. The dynamic range was set at 70dB. These, according to [17] provide an optimal environment to extract speech features.

A total of 27 prosodic features are extracted. These are *duration*, *pitch (F0)* (medium, mean, standard deviation, maximum, minimum), *jitter* (local, local.abs, rap, ppq5, ddp), *shimmer* (local, local.dB, apq3, apq5, apq11, dda), *voicing* (fraction of unvoiced frames, number of breaks, degree of breaks), *intensity* (dB.mean, dB.minimum, dB.maximum), and *pulses* (number, periods, mean period, sd of period) (ppq - Period Perturbation Quotient, apq - Amplitude Perturbation Quotient, abs - absolute, dB - decibels, rap - Relative Average Perturbation).

Pitch defines the quality of sound, usually stipulated by the vibration rates producing the sound. It can simply be defined as the tone degree of lowness or highness. *Jitter* is a loss or lack of a given sample within an audio stream, reducing sound quality. *Shimmer* is the distortiona of sound from its tone, providing sudden highness or lowness in the duration of a sound. Sound *intensity* is the amount of energy flowing through a given audio stream at a given time. A sound *pulse* describes the temporary variation, mostly increase, of a given sound stream beyond its normal level [28].

The data extracted is in continuous form. As the segments contain the same duration due to equal splitting in the segmentation step, this feature are ignored in the succeeding phases.

D. Association Rules Mining

This is the next step after typical feature extraction. This step is meant to provide association rules to imply the relationship the attributes have, resulting to each emotion. Prior to this step, the continuous data resulting from the feature extraction step was first converted in a format to ensure flawless association mining is performed.

In this prior step, for each attribute, continuous data that fluctuated evenly (up-down or down-up), the codeword *confluct* for consistent fluctuation was used. For fluctuations that were inconsistent (up-down, down-up, up-down), *inconfluct* was used to mark these. For data that increased in each segment of the same whole file, *conincrease* was used to mark these, for data that reduced in each segment, *condecrease* was used, for data that increased then disappeared, *incdisappear* was used, while for data that decreased then disappeared, *decdisappear* was used to mark them.

After the study of the patterns in the data, the number of rows for the BDES that was initially split to 3,453 was again reduced to 535 rows, which was the original number before segmentation. Each row consisted of 26 attributes, each marked with either of *confluct, inconfluct, conincrease, condecrease, incdisappear, and decdisappear*.

Apriori algorithm [26] was used in the association rules mining. In mining the rules, the best confidence of 67.4% was attained with the support of 0.2 (107/535). This translates to an appearance pattern of any of the 26 attributes of at least a fifth of the frequency. With the repetitive nature of the pattern, and with a possibility of having at least 0.166 (89/535) support for each of the markers for each attribute if divided equally, (107/535) provided the best confidence levels.

The most frequent attributes appearing with the same marker among the itemsets included pitch mean, pitch median, jitter.local.abs, shimmer.local, voicing.fractunvoicedframes, intensity.dB.mean, and pulses.sdperiod. These six were able to achieve a support of over (120/535). The most predominant rules as drawn from the mining are:

- *pitch (mean) = confluct and jitter (rap) = confluct* $\Rightarrow$ *Anger*

- *pitch (max) = conincrease and intensity (mean) = conincrease and pulse (periods) = increase ⇒ Fear*

The entire experimental protocol is as indicated in Figure 7.

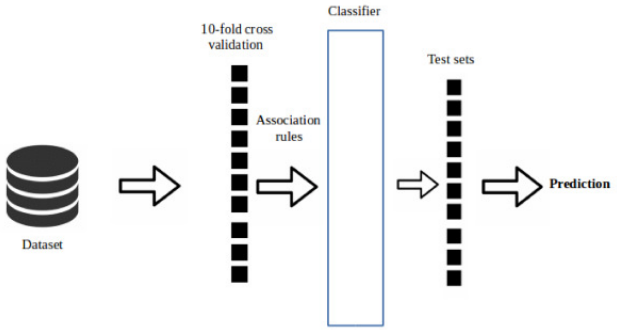

Fig. 7: Experimental protocol of the proposed approach

E. Classification Results and Discussion

Prior to the mining, the dataset was subjected to a 10-folds cross validation as indicated in the protocol.

The final step in this process is the classification of the emotions from the trained models using the data. For comparison of the proposed model with classical procedures, we shall have two parallel classifications, using three classification techniques: Fuzzy K-Nearest Neighbor, Support Vector Machine, and K-Nearest Neighbor.

The first classification is the one being proposed in this paper. The association rules, having been created from the continuous data from the feature-extracted dataset, are fed into the classifiers, and using the initially proposed 2D-approach [11], we make a prediction accuracy of the distress emotion. Rather than having continuous data represent the features as in classical classification process, the association rules in this case show the correlation of adjacent split segments of each sound file, showing the relationships spanning from one end of an audio file to the other end. For instance, after splitting an audio file from the database into 4 segments each 0.4 seconds long, the features extracted from each segment are presented as continuous data such as 156.101 for pitch (medium), 0.02626 for jitter (local), and 0.0543 for shimmer (local) for the first segment. Each of the other segments split from the same long audio file will also have its extracted data in similar continuous form. However, prior to generating association rules, each pair of the split segments is analyzed given the behaviour of the continuous data, and a description of the pattern given under each attribute, to replace the continuous data. As explained in part (IV) D, these patterns describe the various fluctuation patterns.

In the other classification, we use the classical features extracted, and after feature selection using *forward selection* technique, feed the optimal features into the classifiers to measure the prediction accuracy.

As indicated in Table I, the results of using the association rules in K-NN and SVM were higher than when the classical procedures were used in the prediction. However, the result using Fuzzy-KNN is a bit lower when using the proposed approach, but still, appear much more promising than the initial results using the classical classification techniques at all levels. This could be attributed to the fact that fuzzy-KNN in itself endeavors to discourage the assignation of crisp membership of the features, but rather, makes it more flexible for features to fit in a continuum of characteristics, some of which could be very closely related. This, therefore, justifies the possibility of the proposed association rules approach scoring a little lower than the classical approach when fuzzy-KNN is used as the machine learning technique for classification.

TABLE I: Experimental Results: Accuracy and Standard Deviation

Technique	**Classical**	**Association Rules**
***K*-NN**	74.28% (4.35)	75.44% (3.22)
SVM	73.21% (5.17)	76.18% (4.24)
Fuzzy *K*-NN	86.64% (3.31)	83.92% (2.84)

These results as shown in Table I are attributed to the improvement that the association rules bring in to enhance prediction of accuracy by introducing association of pairwise audio segments, other than using traditional continuous data as extracted during feature extraction. The key advantages that the use of association rules brings is the aspect of speaker independence. In the results as portrayed, this is manifested in the model accuracy stemming from the model trained to consider the rules fed into the system that would ensure no context whatsoever would be able to affect the outcome due to the consistency that such a model provides in detecting distress.

Also, due to the difficulty that is experienced in normalizing speech features, the application of association rules ensures an easier approach of comparing correlated characteristics in a specific speech to determine distress. The proposed approach also eliminates the need of feature selection as this might usher in biasness by the exclusion of some features. Association rule mining ensures that all prosodic features mined from a given speech are utilized, giving the emphasis of the importance each speech prosodic feature holds with regard to emotion detection, in this particular context, distress to be specific.

An aspect of time complexity also comes into play in the justification of this proposed approach. In the classical approach protocol, due to the step involving feature selection for optimal model training for accuracy prediction improvement, the procedure required more time to execute as this involved an extra algorithm for feature selection. For this approach, however, as the protocol depicts in Figure 4, the step involving feature selection, usually performed after feature extraction in the classical approach is skipped, therefore reducing the time and resources required to implement this approach. This makes it more efficient and effective as opposed to the classical approach.

V. Conclusion

This paper proposes a novel approach in distress emotion detection where association rules drawn from speech features are used to derive the correlation between features, which are then fed to machine learning techniques for distress detection. Classically, emotions have been predicted through the feeding of extracted features. This proposal aims to add an extra steps to ensure better prediction through the addition of steps to further segment the audio files and generate association rules that are then fed to the classifiers to make the predictions.

The use of association rules ensures that all distress speech, irregardless of time expanse, producer's origin, background and cultural variables are able to receive fare detection procedure. This is expected to be a trend-setter in ensuring speaker independence when it comes to the application of distress detection tools.

Further to this approach, healthcare monitoring of the elderly is bound to improve as the distress detected from the elderly usually isolated within their own residences or rooms in home cares will be much improved, enabling appropriate responses and action whenever incidences of attention come up.

References

[1] T. W. B. Group. (2018) The world bank group. [Online]. Available: https://data.worldbank.org/

[2] A. H. Munnell, "The average retirement age–an update," *Notes*, vol. 1920, pp. 1960–1980, 2015.

[3] U. N. E. C. for Africa, "The demographic profile of african countries," Economic Commission for Africa, 2016. [Online]. Available: https://www.uneca.org

[4] C. I. Agency. (2017) The world factbook. [Online]. Available: https://www.cia.gov/library/publications/the-world-factbook/fields/2010.html

[5] G. Yamey, R. Shretta, and F. N. Binka, "The 2030 sustainable development goal for health," 2014.

[6] Z. W. Alliance. (2018) Connected aging. [Online]. Available: https://z-wavealliance.org/connected-aging/, visited 2019-07-16.

[7] D. Watson and J. W. Pennebaker, "Health complaints, stress, and distress: exploring the central role of negative affectivity." *Psychological review*, vol. 96, no. 2, p. 234, 1989.

[8] E. Pearson. (1971) Torts-emotional distress. heinonline. [Online]. Available: https://heinonline.org, visited 2019-07-10.

[9] J. Aigrain, "Détection de stress dans la gestuelle à partir de vidéos," PhD dissertation, Sorbonne University, 2016.

[10] R. B. Knapp, J. Kim, and E. André, "Physiological signals and their use in augmenting emotion recognition for human–machine interaction," in *Emotion-oriented systems*. Springer, 2011, pp. 133–159.

[11] D. Machanje, J. Orero, and C. Marsala, "A 2d-approach towards the detection of distress using fuzzy k-nearest neighbor," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 2018, pp. 762–773.

[12] I. Lefter, L. J. Rothkrantz, D. A. Van Leeuwen, and P. Wiggers, "Automatic stress detection in emergency (telephone) calls," *International Journal of Intelligent Defence Support Systems*, vol. 4, no. 2, pp. 148–168, 2011.

[13] T. Johnstone and K. R. Scherer, "The effects of emotions on voice quality," in *Proceedings of the XIVth International Congress of Phonetic Sciences*. University of California, Berkeley San Francisco, 1999, pp. 2029–2032.

[14] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I*. Springer, 2007, pp. 108–137.

[15] B. Kollmeier, T. Brand, and B. Meyer, "Perception of speech and sound," in *Springer handbook of speech processing*. Springer, 2008, pp. 61–82.

[16] I. R. Murray, C. Baber, and A. South, "Towards a definition and working model of stress and its effects on speech," *Speech Communication*, vol. 20, no. 1-2, pp. 3–12, 1996.

[17] P. Boersma and D. Weenik, "Praat: a system for doing phonetics by computer. report of the institute of phonetic sciences of the university of amsterdam," 1996.

[18] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 240–243.

[19] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, no. 4, pp. 290–296, 2000.

[20] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[21] Y. Alkaher, O. Dahan, and Y. Moshe, "Detection of distress in speech," in *Science of Electrical Engineering (ICSEE), IEEE International Conference*. IEEE, 2016, pp. 1–5.

[22] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog." in *INTERSPEECH*. Citeseer, 2002.

[23] E. Trentin, S. Scherer, and F. Schwenker, "Emotion recognition from speech signals via a probabilistic echo-state network," *Pattern Recognition Letters*, vol. 66, pp. 4–12, 2015.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[25] M. Delgado, N. Manín, M. Martin-Bautista, D. Sanchez, and M.-A. Vila, "Mining fuzzy association rules: an overview," in *Soft Computing for Information Processing and Analysis*. Springer, 2005, pp. 351–373.

[26] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo *et al.*, "Fast discovery of association rules." *Advances in knowledge discovery and data mining*, vol. 12, no. 1, pp. 307–328, 1996.

[27] A. P. Vogel, P. Maruff, P. J. Snyder, and J. C. Mundt, "Standardization of pitch-range settings in voice acoustic analysis," *Behavior research methods*, vol. 41, no. 2, pp. 318–324, 2009.

[28] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.