International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 11 | Nov 2020 www.irjet.net p-ISSN: 2395-0072

A Review on Detecting Aggression in Voice

Aswathi Balakrishnan[1], Anoop K[2]

[1]M Tech Student, Dept. of Electronics and Communication, College of Engineering Thalassery, Kerala, India
[2]Assistant Professor, Dept. of Electronics and Communication, College of Engineering Thalassery, Kerala, India

---

Abstract - *Aggression is a harmful communication style that is caused because of the negative emotions, like fear, anger, pain, and frustration when accompanied by high arousal. The aggression of an individual can be viewed from his/her body gestures, facial expressions, voice etc. This study has been undertaken to investigate different methods of detecting aggression in voice. The major steps include the feature extraction process and the classification process. Different types of estimation techniques are used for the extraction process in which glottal inverse filtering (GIF) was found to be more effective. Classification process is carried out using suitable classifiers*

Key Words: Aggression detection, Classifiers, Estimation, Feature Extraction, GIF etc

1. INTRODUCTION

Aggression is an action or response by an individual that is unpleasant to another person. It as an intentional behavior aimed at causing either physical or psychological pain. It is broadly divided into two categories.

1) Controlled-instrumental aggression

2) Reactive-impulsive aggression

Controlled-instrumental aggressions are purposeful or goal oriented. It is used to achieve a goal or secure some reward. Reactive-impulsive aggressions are sudden or unpredictable that is inappropriate. Here we concentrate on the second type of aggression that is the impulsive aggression. Impulsive aggression mainly occurs in speech, specifically during an argument.

There are several methods for detecting aggression in voice like using of inverse filtered speech features, using sensors and semantic information, analysis of neural network, using microphones, analysis of acoustic information, using overlapping speech etc. However all these methods consists of two major steps i) Feature extraction process ii) Classification process. In the feature extraction process the features required for the detection are extracted using various techniques. Inverse estimation technique or inverse filtering algorithm, forward selection method, image acquisition and processing, functional localizer scanning, auto regressive models, analysis using window functions etc are some of the techniques used here. After the extraction process these features are classified using suitable classifiers. HMM, GMM, SVM, RF, BN, DBN, are some of the classifiers used here.

2. LITERATURE REVIEW

Björn Schuller, Gerhard Rigoll, and Manfred Lang developed a hidden Markov model-based speech emotion recognition [1]. Here two methods are propogated and compared. The features extracted in the first method are the raw pitch and energy contour of the speech signal. In order to calculate the contours frames the speech signal is analyzed every 10ms using a Hamming window function. Logarithmic mean energy within a frame is used to calculate the values of energy. Average magnitude difference function (AMDF) is used to achieve pitch contour. Within the first method 13 pitch related features and 7 energy related features of the raw contours are derived. Mean duration of voiced sounds, average pitch, standard deviation of duration, standard deviation of pitch, relative pitch maximum, relative pitch minimum, position of minimum pitch, position of maximum pitch, maximum of absolute pitch derivation, standard deviation of distance between reversal points , mean distance between reversal points, mean of absolute pitch derivation, rate of voiced sounds are the derived pitch features. Relative maximum of derivation of energy, average of derivation of energy, position of maximum of derivation of energy, standard deviation of derivation of energy, maximum of absolute second derivative of energy, mean distance between reversal points, standard deviation of distance between reversal points are the derived energy features. Compared to the energy related features the pitch related features showed more potential. Classification is done using Gaussian mixture models and gave 86% recognition rate. Continuous hidden Markov model is used in the second method. Several states are considered using low-level instantaneous features instead of global statistics and gave 77.8% overall recognition rate.

P.W.J. van Hengel and T.C. Andringa introduced a SI guard system for verbal aggression detection in complex social environments [2]. The system detects the presence of aggressive shouting in realistic and uncontrollable environments. It consists of 1) signal processing stage that simulates some form of auditory attention in the form of foreground separation, 2) methods to extract cues for verbal aggression and 3) a decision mechanism. For searching of verbal aggression cues foreground signal is used as a basis. The properties used are, salience and height of the pitch, the audibility, the level and three measures for the spectral shape and distortion of the harmonic pattern. The physical setup

consists of a microphone that is weather-proofed, far field and low cost, with a 50 dB dynamic range. It is connected to a specially designed and weatherproof analysis hardware. Detections are routed via IP to a central gateway they are logged and administrated there. To give an audio or video alarm on detection a user interface can be configured this provides an observer to add comments and access the logs. This system is considered to be the first successful detection system for a non-trivial target in an unconstrained environment.

Iulia Lefter, Gertjan J. Burghouts, Leon J.M Rothkrantz proposed Automatic audio-visual fusion for aggression detection using meta-information [3]. Here a new method for audio visual sensor fusion is proposed and applied to automatic aggression detection. The fusion starts from low level sensor features and ends with the high level multimodal assessment. An intermediate step is proposed to discover the structure in the fusion process which is called meta features. There exist a set of 5 meta features which are Audio focus (AF), Video focus (VF), Context (C), History (H), Semantics (S).The results of the intermediate layer are fused by a RF classifier. The main advantage is that it showed a positive impact over the standard fusion techniques that is a 58% improvement over feature level fusion and a 40% improvement over decision level fusion.

As a modification to [3], Aggression detection in speech using sensor and semantic information [4] was developed by Iulia Lefter , Gertjan J. Burghouts , Leon J.M Rothkrantz . Here along with the audio and video semantic information are also considered. Semantic information include both acoustic and linguistic information that is both the non verbal and verbal information are considered. The acoustic feature set consists of features like speech duration, mean, standard deviation, slope, range of pitch (F0) and intensity, mean and bandwidth of the first four formants F1–F4, jitter, shimmer, high frequency energy (HF500), harmonics to noise ratio (HNR), Hammarberg index, center of gravity and skewness of the spectrum. These features are computed on segments of length equal to 2 second. Linguistic features have been clustered in 6 classes: Positive emotions, negative emotions, actions, context, cursing, nonverbal. The techniques and algorithms used are similar as that of [3]. Here also results are fused by the RF classifier. This technique shows better performance and accuracy because of the fusion of more features.

J. F. P. Kooija, M. C. Liema, J. D. Krijndersb, T. Andringab, D. M. Gavrilaa proposed a Multi-modal human aggression detection [5] system named CASSANDRA. It is a smart surveillance system that can be used to detect cases of aggressive human behavior in public environments. It consists of a audio, video and sensor fusion unit. To track persons in 3D and to extract features regarding the limb motion relative to the torso the system uses overlapping cameras. From the audio side, it classifies instances of speech, singing, screaming, and kicking-object. The video and audio cues are fused with contextual cues. Dynamic Bayesian Network (DBN) produces an estimate of the total aggression level. The overall processing rate is on average about 4 s per frame, using un-optimized C and MATLAB.

Aggression recognition using overlapping speech[6] was proposed by Iulia Lefter and Catholijn M. Jonker. In most of the cases segments containing overlapping speech are not considered because of the difficulties in the estimation of pitch related features but here it is used as the key feature. Here it is developed as a Virtual Reality therapy system that helps patients in forensic clinics to deal with aggression tendencies. Along with the acoustic features a feature vector consisting of overlap information is also added. 3 categories of overlapping speech is considered. Short feedback, Premature turn taking and Competing ovelapp. In the first case no turn change is present only one speaker talk continuously, in the second type the current speaker is not allowed to complete his/her turn and in the last case two speakers talk at a time or simultaneously to impose themselves. Classification for aggression level recognition and predicted overlap is performed using a Random Forest classifier. Classification results are frequently affected by data unbalance. Compared to other acoustic feature sets, overlap information had the highest information gain for aggression prediction and it was among the best feature subsets of 3-4 features obtained with automatic feature selection. It is proved that overlapping speech is of great importance in the identification and determing the emotional state.

Subhasmita Sahoo and Aurobinda Routray detected aggression in voice using inverse filtered speech features [7] .Here an automatic method for detection of aggression is proposed using features extracted from pressure distribution in vocal tract. These variations in air pressure distribution is computed by inverse estimation of the speech signal. Principal component analysis has been used to reduce the dimension of extracted features. The air pressure variations are classified into aggression or calm using a Hidden Markov Model. The system detected aggression with about 93% accuracy.

An aggression detector using microphone [8] was developed by Jeff Kao and Jack Gillum. It was implemented using a Louroe Digifact A microphone. The detector includes a microphone, a sound-processing component, a machine-learning algorithm and a thresholding component. The microphone receives the audio signal and then the sound processing component extracts the sound features and these features are used by the machine learning algorithm to predict verbal aggression, the settings for the algorithm is contained in the thresholding component. Each set of audio features is considered as a frame of sound and is used to predict whether that segment of the sound input is aggressive. Once trained, based on the audio features the classification algorithm

generates a score ranging from 0.0 to 1.0 for each frame. This score represents an overall percentage for identifying aggression from 0% to 100%. In operation, a percentage exceeding a set threshold for a long period of time results in the prediction of aggression by the device. During the operation the algorithm is tuned using the threshold setting. The main disadvantage of the system is that it frequently produced false positives for sounds such as laughing, and loud discussions.

3. COMPARISON

Extracting the raw pitch, energy contour of the speech signal and using a GMM classifier [1] for aggression detection the recognition rate was 86%. Whereas in the second method a HMM classifier is used considering several states using low-level instantaneous features and the recognition rate was 79.8%.

P.W.J. van Hengel and T.C. Andringa [2] made a SI gard system where the foreground signal was used for the detection. The physical set up included a microphone and it is connected to hardware. The system was tested in a pilot project for eighteen days and it produced ninety six detection in which two was essential, twenty three was useful, forty four justified alarm and twenty seven false alarm. Also, no alarm was missed.

In the detection of aggression using meta information [3] a large audio – video data base was considered. Here the aggression level was marginalized on a three point scale as low, medium and high based on audio only, video only and both. This approach of making use of meta information in fusion methodology showed a great improvement over other standard fusion techniques. 88.5% accuracy was obtained.

Along with the meta information fusion methodology [3] here [4] sensor and semantic information are also considered. The dataset used was same. Here to predict the level of aggression and the context and the semantic meta features the audio models was used in terms of prosody and words. RF classifier was used. By the addition of extra features the performance improved from 86% to 92%

CASSANDRA [5] combines both audio and video features with the help of a Dynamic Bayesian Network. The system was tested at a platform of the Amsterdam – Amstel train station and evaluated with different configurations like audio features only, kinetic energy feature only, kinetic energy and audio features, video features only , all audio and video features. Considering the audio features only the mean, standard deviation of the error and the root mean squared error (RMSE) was 0.162, 0.154 and 0.225 respectively.

In the detection of aggression using overlapping speech [6] the dataset consist of dyadic interactions between professional aggression training and some other local participants. The annotation was performed on a 3 point scale ranging from no aggression, medium aggression and high aggression .The results showed 21% samples with no aggression, 54% medium aggression and 25% high aggression.

Subhasmita Sahoo and Aurobinda Routray [7] used one hundered and twenty audio audio clips with 30 minute duration and detected aggression in these clips.this was extracted from intervies of political personalities. Here air pressure variation was the key feature for determining aggression. HMM classifier was used for the classification and gave 93% accuracy .

The detection system proposed by Jeff Kao and Jack Gillum [8] was particularly for the school environment. The dataset contained recorded clips of fourty students between the age of fifteen and eighteen from two different schools. The system recorded scream or shout, laughter, loud discussion , cheering, singing, coughing etc. There was a total of 65 aggressive shout or scream in which the system triggered for 35 instances. The main disadvantage of the system was that it frequently produced false positives for s laughing, coughing, cheering and loud discussions also high-pitched shrieks did not trigger the algorithm. These can be avoided by improving the algorithm.

4. CONCLUSIONS

Different technique of detecting aggression in voice is discussed here with each technique being an improved version of the previous one. There are various methods for the feature extraction process that ranges from local feature extraction to neural network based methods. It can be viewed that the accuracy and performance of the system is improved by combining more acoustic features. The early detection of aggression in voice is important because it could prevent physical aggression or assault. Also a system with enough intelligence to understand and detect aggression could help an individual to avoid a social embarrassing situation. These technologies can be utilized in health care institutes, city surveillance, schools, prisons etc.

REFERENCES

[1] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model based speech emotion recognition,”. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings.(ICASSP’03), vol. 2. pp. II–1.

[2] P.W.J. van Hengel and T.C. Andringa, “Verbal Aggression Detection In Complex Social Environments”, 2007 IEEE Conference on Advanced Video and Signal

Based Surveillance, Volume: 1, Pages: 15-20, DOI:10.1109/AVSS.2007.4425279

[3] I. Lefter, L.J.M. Rothkrantz, and G.J. Burghouts, "Automatic Audio-Visual Fusion for Aggression Detection Using Meta-Information" 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, DOI : 10.1109/AVSS.2012.13

[4] I. Lefter, L.J.M. Rothkrantz, and G.J. Burghouts "Aggression detection in speech using sensor and semantic information," in Text, Speech and Dialogue. Springer, 2012, pp. 665–672.

[5]J.F.P. Kooij a , M.C. Liema, J.D. Krijnders b, T.C. Andringa b and D.M. Gavrilaa, "Multi – Modal Human Aggression Detection", Elsevier, Computer Vision and Image Understanding, Volume 144, March 2016, Pages 106-120, DOI: 10.1016/j.cviu.2015.06.009

[6] Iulia Lefter and Catholijn M. Jonker "Aggression Recognition Using Overlapping Speech", IEEE, 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), DOI: 10.1109/ACII.2017.8273616

[7] Subhasmita Sahoo, Aurobinda Routray "Detecting Aggression in Voice Using Inverse Filtered Speech Features" IEEE Transactions on Affective Computing (Volume: 9, Issue: 2, April-June 1 2018),

DOI: 10.1109/TAFFC.2016.2615607

[8] Jeff Kao, Jack Gillum "Reverse-Engineering an Audio Aggression Detection Algorithm", ProPublica, Computational journalism Symposium 2020, March 2020, Boston, Mass. USA.