Impact of language on detecting agitation

Rik de Sain
University of Twente
The Netherlands
r.desain@student.utwente.nl

ABSTRACT

The paper is explores the possibilities of machine learning models for agitation detection in different languages. The aim is to identify the voice features that are constant across languages and use them to develop an algorithm that can detect agitated speech by considering different languages. The research is conducted in three phases: a preparatory stage, feature extraction phase and algorithm developing phase. In the preparatory stage data sets having emotions corresponding to agitation behaviour in German and English languages were obtained. Among all voice based activities of agitation, 'rapid speech' was selected. In feature extraction stage, voice features or properties relevant to rapid speech (pitch, loudness) will be extracted. In the algorithm development phase, several machine learning models will be trained and tested to predict a level of agitation in both the languages. In-lining with the hypothesis, results indicate the difference in accuracy as we change language i.e. by using pitch and loudness as voice features in the English language 79% accuracy is achieved whereas the same features result in a 72% accuracy score for the German language. The Support Vector Classifier was most accurate in both languages.

Keywords

Agitation detection, Machine learning, Voice processing, Language

1. INTRODUCTION

The literature defines agitation as "a state of anxiety or nervous excitement". Agitation is a major and common symptom of various neuropsychiatric disorders like dementia, Alzheimer's and cognitive impairment [9]. Approximately 12% of the world's population is suffering from a neuropsychiatric disorder [4]. When compared to other diseases neurophysiatric disorders are some of the most prominent when it comes to disability and burden. It not only impacts the life of patients [22] but also diminishes the quality of life for their caregivers [9, 8]. Detecting agitation in earlier stages can prolong a patient's stay in their own homes and helps caregiver in providing adequate patient care and restore their quality of life.

The current way of detecting agitation is by monitoring a patient's behavior through behavioral scales, like the "Scale for Observing Agitation in Persons with DAT" [13] and the Cohen-Mansfield scale [7]. In scales for measuring agitation and anxiety vocal-activities are very common. In SOAPD, seven broad categories of human behaviour were identified, in which three categories: high-pitched or loud noise, repetitive vocalization and negative words, are based on voice monitoring. The high-pitch category contains activities like shouting, yelling, crying; The repetitive vocalization category contains activities like repeated words, rapid speech and whining and the negative words category contains activities like using abusive words and threatening language [13]. Other than human voice, motor activities and vital signs are also commonly used for agitation monitoring [19, 13, 7]. In a research by Sakr et al., vital signs like heart rate, breathing rate and skin galvanic response were monitored by using a wearable device to predict agitation [19]. In an another research by Lisa et al., the idea of non-contact monitoring system for agitated patients in hospitals was explored. They developed and tested a vision based system attached to patient's bed that can track vital sign changes in combination with voice monitoring [14]. Furthermore, for early stage agitation detection, a behavioral and environmental Sensing intervention was developed by using wearable sensing technologies [12].

These existing systems are obtrusive in nature. A wearable system monitoring vital signs has very low acceptance rate by patients [10] whereas observing a patient through the behavioural scales requires the presence of caregivers. Due to the human-observation nature of these scales the results are prone to human error. The results obtained from these scales can be biased due to a caregiver's own perception of agitation. Other factors, such as patient's behaviors i.e. the patient is not agitated when observed by a caregiver, are also biasing.

To overcome these drawbacks, a technology-driven unobtrusive system that can automatically detect the level of agitation and alert the caregiver is required. This research will use 'human-voice' as an unobtrusive way of detecting agitation. Furthermore, for effective implementation of such systems it is important to validate it universally i.e. with different languages. So far, to the best of our knowledge, no research work is done in comparing the various languages and understanding its impact on agitation predication systems. To do that 'rapid speech' is used as a common activity found among agitated patients. We hypothesize that different languages might use different voice features to predict a level of agitation. This leads us to the research question:

RQ: What is the accuracy of machine learning algorithms in classifying agitated German and English speech by using pitch and loudness as voice features?

28^{th} *Twente Student Conference on IT* Febr. 2^{nd}, 2018, Enschede, The Netherlands.

2. BACKGROUND WORK

Speakers from different languages express themselves differently. According to research the expression of emotions is determined by your culture [5]. Therefore people from different countries express emotions differently. When looking specifically at voice features like pitch research suggest that native speakers of different language speak at a different average pitch [16] and they have pitch variation when imposing questions [11].

In past research it shows that the human voice indicates various human emotions such as anger, happiness, sadness and fear [17, 21]. The moods of people can also be detecting solely through the use of voice, as proven by Rojas et al.[18]. Vocal activities with emotions also indicate various behaviours. For example, laughing while talking reflects healthy relations between individuals, shouting while talking reflects anger and talking fast reflects nervousness or anxiety [20]. In previous research [9] it is shown that people experiencing agitation also exhibit signs consistent with emotional distress. Therefore emotions like anger and fear can signal agitation. Another research [23] found a correlation between voice and emotion, more specifically they use the frequency of someone's voice to determine their emotion. Combining these correlations we see that the frequency of someone's voice could indicate agitation as well.

The major agitation determining voice factors as suggested by both the SOAPD [13] and the Cohen-mansfield [7] scales are the pitch and loudness of the patient's voice. Using that in this research, we will be determining agitation by using pitch and loudness in 'rapid speech' behaviour.

3. METHOD

The research will consist of three phases:

- *The collection of data*:
  In this phase a valid data set of English and German is gathered. The data set consists of audio fragments rated with an agitation score, either 1 when deemed agitated or 0 when deemed as not agitated. This data is separated into two sets: the training set and the testing set. The training set is used to train the machine learning algorithms. The testing set is used to test the accuracy of the model. In order to find out the optimal features for agitated speech detection, for each data set three types of testing is done by varying the voice feature: only pitch, only loudness, and both pitch and loudness together.

- *Feature extraction:* In this phase the voice data is pre-processed followed by features extraction. These extracted features will be fed into machine learning algorithms.

- *Algorithm for rapid speech detection:*
  In this phase different machine learning algorithms are trained using the data set. Further, these models are used to determine the level of agitation of testing set. The accuracy score is recorded.

3.1 The collection of data

The database used in this research will need to consist of a selection of audio fragments and a measurement for the level of agitation expressed in the fragment. The database is split in to two parts, the first part (80%) is designated for the training of the machine learning algorithms, the second part (20%) is used to test the generated models. Since there was no database readily available that indicates agitation, emotional databases consistent with agitation behaviour i.e. anger and fear emotions were taken. The English data set used for this is the RAVDESS [15] data set. This data set contains speech of 24 professional actors, 12 male and 12 female, talking in 8 different emotions: neutral, calm, happiness, sadness, anger, fear, disgust and surprise. The German data set used is the "EmoDB" [6] data set. This data set contains speech of 10 speakers, 5 male and 5 female aged 21 to 32, talking in 7 different emotions: neutral, anger, boredom, happiness, sadness, fear, and disgust. A copy of English data set is made using only 10 speakers, 5 male and 5 female with only anger and fear emotion to equalize the number of speakers in both data sets.

3.2 Feature extraction

The features were extracted from the audio files by using the libROSA library [1] to load the audio files as a floating point time series, this data contains both the loudness and the pitch. To isolate the pitch and the loudness we used librosa. piptrack which results in an array of pitches.

To train the classifiers we first needed to pre-process the data as all the entries need to be equally long, therefore we separated the data into subsets of two seconds to gain a floating point time series of 20000 frames that can be used to train the classifiers.
To determine whether a sample is considered agitated we looked at the emotions in the data sets. Anger and fear where determined to indicate rapid speech and thus agitation.

3.3 Algorithm for rapid speech detection

The training of the machine learning algorithm is done in the Python programming language and library is used. For the machine learning models the SciPy [2] library is used. A number if machine learning classifiers were trained and tested. They are:

- *Decision Tree classifier*: A decision tree classifier is a classifier that classifies data by putting the results in a tree where the trees make choices on the data. The choices that every node makes are trained.

- *Gaussian Naive Bayes classifier*: A Gaussian Naive Bayes classifier classifies the data with a normal distribution.

- *K Nearest Neighbors classifier*: A k-nearest neighbors classifier lets the k nearest neighbors of a node vote on the outcome of the node. In this research we set k as 5.

- *Linear Distribution classifier*: A linear distribution classifier tries to classify the result through a linear combination of parameters of the input. These parameters are trained.

- *Logistic Regression classifier*: A logistic regression classifier uses a logistic function to classify the data.

- *Support Vector classification*: A support vector classifier represents data as points in space and uses a gap to classify which binary answer a new data point belongs to. This gap is trained. In this research we used a support vector machine with degree 3.

- *Multilayer Perceptron classifier*: A multilayer perceptron classifier is a neural network consists of an input layer, one or more hidden layers and one output layer of nodes. These nodes have an activation

function. These functions are trained to create a classifier. In this research we set the multilayer perceptron to have 100 hidden layers.

Among these classifiers Support Vector classification is the only classifier that is strictly designed for binary decisions [3]. We compare the models against a so called Zero classifier, this is a classifier that simply picks the option that occurs most in the data.

4. RESULTS

The results showed that Gaussian naive Bayes and support vector classification provides more accurate results among all. The results for both of them canbe found in Table 1, Table 2 and Table 3. The full results can be found in the tables in the appendix. Table 1 contains the accuracy's of the classifiers when detecting agitation using features of the English language. 25% percent of the entries in the English data set are considered agitated. This means random guessing would result in an accuracy of 62.5% and a Zero classifier would have an accuracy of 75%. We can see that in the results of the 24 English speakers an accuracy of 83%, 8% better than the Zero classifier, is reached by the Support Vector classification for features loudness and pitch. Furthermore we see that the accuracy of all classifiers went below the Zero classifier while considering loudness and pitch individually. Considering the smaller set of 10 English speakers, the table 3 shows that the maximum accuracy has gone down to 79%, 5% better than the Zero classifier, this is still by the Support Vector Classifier. However with this smaller data set there are classifiers that have a better accuracy than the Zero classifier on only pitch or only loudness. Namely the Gaussian Naive Bayes and the Support Vector classifier when considering only loudness and the k-neighbors classifier when considering only pitch.

Table 2 contains the accuracy's of the classifiers when detecting agitation using features of the German language. This data set contains 33.3% agitated speech samples. This means that random guessing would result in an accuracy of 55.6% and a Zero classifier would have an accuracy of 66.7%. We can see that the best result is obtained by both the Support Vector classifier and the Gaussian Naive Bayes classifier with 72%, 5.3% above the Zero classifier. Interestingly all classifiers except for those two had very low results when considering both loudness and pitch. When we consider only loudness the average accuracy rises by 11% from 55% to 66% and this average accuracy raises to 75% when only considering pitch. When only considering pitch every classifier except for the Linear Discriminant classifier scores better at accuracy than the Zero classifier.

Table 1. Accuracy of two of the classifiers on different features of English speech.

Classifier	Gaussian Naive Bayes	Support Vector Classification
Loudness and Pitch	82%	83%
Loudness	74%	73%
Pitch	57%	72%

Table 2. Accuracy of two of the classifiers on different features of German speech.

Classifier	Gaussian Naive Bayes	Support Vector Classification
Loudness and Pitch	72%	72%
Loudness	66%	65%
Pitch	77%	75%

Table 3. Accuracy of two of the classifiers on different features of a smaller sample of English speech.

Classifier	Gaussian Naive Bayes	Support Vector Classification
Loudness and Pitch	78%	79%
Loudness	68%	78%
Pitch	63%	74%

5. CONCLUSION

From the results we can see a clear distinction in accuracy between both the using different features and the different languages. In German there is a significant distinction between the accuracy of the models when trained with the combination of loudness and pitch or trained with loudness only, this distinction is smaller in English. When looking at the accuracy of training with only the pitch in comparison to training with both features combined we see that in English the accuracy went down as opposed to German where the accuracy went up.

From this we can conclude that in English both loudness and pitch are important for the detection of agitation whereas in German only using the pitch is better than also considering the loudness. This means that different languages have different features that can be used to detect agitation.

6. LIMITATIONS

There are some big limitations to this research. First of all the lack of a database annotated with agitation means that the tested property, in this case rapid-speech, will only be one indication of agitation. This leaves a lot of determining factors out of the research. A great way to test or train the classifiers is also to have the same person say something in both tested languages, both in an agitated state and in a neutral state and compare the accuracy of the classifiers. But cohering to the current situation of global isolation this was not a possibility for this research.

7. FUTURE RESEARCH

As no direct database is available for agitation, it is highly desirable from future works to make such a database, this could be done by using for example SOAPD [13] or the Cohen-Mansfield scale [7]. Furthermore it is recommended that the languages taken are further apart in origin. German and English are both both originate in the class of west Germanic languages and therefore there are a lot of similarities between the two. The study could be repeated using languages from Asia or Africa. A sound study method by collecting a database properly will reflect more on impact of language on agitation detection. Lastly it would be recommended to add voice features, like for example tremor or repetition, as these are described in the literature as good indicators for agitation as well.

8. ACKNOWLEDGEMENTS

I am extremely grateful for the help of my supervisor N. Sharma for giving me all the necessary advice and support for the writing of this paper and for keeping me on track to my goals.

I would also like to extend my thanks to A. Chiumento and D. le Viet for assisting on questions and making sure that progress was made.

9. REFERENCES

[1] librosa.org.

[2] Scipy.org.

[3] Us5649068a - pattern recognition system using support vectors.

[4] Cross-national comparisons of the prevalences and correlates of mental disorders. who international consortium in psychiatric epidemiology, 2000.

[5] J. S. Boster. Emotion categories across languages, May 2007.

[6] F. Burkhardt, A. Paeschke, M. A. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech: Semantic scholar, Jan 1970.

[7] J. Cohen-Mansfield. Cohen-mansfield agitation inventory. *PsycTESTS Dataset*, 1986.

[8] J. Cohen-Mansfield and N. Billig. Agitated behaviors in the elderly: I. a conceptual review. *Journal of the American Geriatrics Society*, 34(10):711–721, 1986.

[9] J. Cummings, J. Mintzer, H. Brodaty, M. Sano, S. Banerjee, D. Devanand, S. Gauthier, R. Howard, K. Lanctôt, C. G. Lyketsos, and et al. Agitation in cognitive disorders: International psychogeriatric association provisional consensus clinical and research definition. *International Psychogeriatrics*, 27(1):7–17, 2014.

[10] J. George, S. Long, and C. Vincent. How can we keep patients with dementia safe in our acute hospitals? a review of challenges and solutions. *Journal of the Royal Society of Medicine*, 106(9):355–361, 2013.

[11] V. J. V. Heuven and E. V. Zanten. Speech rate as a secondary prosodic characteristic of polarity questions in three languages. *Speech Communication*, 47(1-2):87–99, 2005.

[12] N. Homdee, R. Alam, J. A. Hayes, T. Hamid, J. Park, S. Wolfe, H. Goins, N. Fyffe, T. Newbold, T. Smith-Jackson, A. Bankole, M. S. Anderson, and J. Lach. Agitation monitoring and prevention system for dementia caregiver empowerment. *Computer*, 52(11):30–39, 2019.

[13] A. Hurley, L. Volicer, L. Camberg, J. Ashley, P. Woods, G. Odenheimer, W. Ooi, K. McIntyre, and E. Mahoney. Measurement of observed agitation in patients with dementia of the alzheimer type. 5:117–132, 01 1999.

[14] L. Kroll, N. Böhning, H. Müßigbrodt, M. Stahl, P. Halkin, B. Liehr, C. Grunow, B. Kujumdshieva-Böhning, C. Freise, W. Hopfenmüller, and et al. Non-contact monitoring of agitation and use of a sheltering device in patients with dementia in emergency departments: a feasibility study. *BMC Psychiatry*, 20(1), 2020.

[15] S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess), Apr 2018.

[16] H. Noh and D.-H. Lee. Cross-language identification of long-term average speech spectra in korean and english. *Ear and Hearing*, 33(3):441–443, 2012.

[17] V. A. Petrushin. Us7222075b2 - detecting emotions using voice signal analysis, 2007.

[18] V. Rojas, S. F. Ochoa, and R. Hervás. Monitoring moods in elderly people through voice processing. *Ambient Assisted Living and Daily Activities Lecture Notes in Computer Science*, page 139–146, 2014.

[19] G. E. Sakr, I. H. Elhajj, and U. C. Wejinya. Multi level svm for subject independent agitation detection. *2009 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, 2009.

[20] M. Schaller, D. Keltner, J. Haidt, and M. N. Shiota. Evolution and social psychology. *Social Functionalism and the Evolution of Emotions.*, page 115–142, 2013.

[21] E. R. Simon-Thomas, D. J. Keltner, D. Sauter, L. Sinicropi-Yao, and A. Abramson. The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, 9(6):838–846, 2009.

[22] R. M. Suinn. The terrible twos—anger and anxiety: Hazardous to your health. *American Psychologist*, 56(1):27–36, 2001.

[23] E. L. van den Broek. Emotional prosody measurement (epm): A voice-based evaluation method for psychological therapy effectiveness.

10. APPENDIX

Table 4. Accuracy of the classifiers on different features of English speech.

Classifier	Decision tree	Gaussian Naive Bayes	k-Neighbors	Linear Discriminant	Logistic Regression	Support Vector Classification	Multilayer Perceptron
Loudness and Pitch	67%	82%	72%	74%	75%	83%	69%
Loudness	71%	74%	74%	75%	72%	73%	72%
Pitch	74%	57%	73%	74%	72%	72%	74%

Table 5. Accuracy of the classifiers on different features of German speech.

Classifier	Decision tree	Gaussian Naive Bayes	k-Neighbors	Linear Discriminant	Logistic Regression	Support Vector Classification	Multilayer Perceptron
Loudness and Pitch	58%	72%	35%	50%	50%	72%	48%
Loudness	74%	66%	65%	64%	65%	65%	65%
Pitch	79%	77%	70%	64%	81%	75%	76%

Table 6. Accuracy of the classifiers on different features of a smaller sample of English speech.

Classifier	Decision tree	Gaussian Naive Bayes	k-Neighbors	Linear Discriminant	Logistic Regression	Support Vector Classification	Multilayer Perceptron
Loudness and Pitch	78%	78%	74%	73%	77%	79%	70%
Loudness	70%	78%	76%	76%	74%	78%	74%
Pitch	67%	63%	78%	66%	71%	74%	65%