Voice feature extraction from agitated speech

Radu Seteanu
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands

r.seteanu@student.utwente.nl

ABSTRACT

This research paper aims to identify voice features by literature survey and validate them for human agitation detection. The prosodic features pitch and loudness were identified through literature survey and then voice snippets were analysed and tested in offline (pre-recorded) speech for validation. These voice features are correlated with quantifiable sound wave metrics such as frequency and amplitude, respectively. By identifying distinct variations in these metrics, agitation behaviour can be extrapolated. Using the results found, the agitation state of the person is to be analysed. Such an agitation detection system can benefit in improving the quality of life of caregivers and elderlies.

KEYWORDS

Voice feature, voice recognition, agitation detection

1. INTRODUCTION

Agitation is a common neuropsychiatric disorder that usually manifests in the elderly and which is associated with Alzheimer's disease and dementia. It leads to diminished quality of life for patients and their loved ones. In literature, agitation has been defined as: "(1) occurring in patients with a cognitive impairment or dementia syndrome; (2) exhibiting behaviour consistent with emotional distress; (3) manifesting excessive motor activity, verbal aggression, or physical aggression; and (4) evidencing behaviours that cause excess disability and are not solely attributable to another disorder (psychiatric, medical, or substance-related)."[1] Thus, agitation is demarcated by two broad categories of behaviours: motor and verbal. Among them, verbal behaviour or aggression can be monitored more unobtrusively and hence it is the focus of this paper. Whilst behaviour monitoring techniques include wearables that detect heart rate, video cameras and audio recording equipment, for the scope of this research, only audio recording technology will be analysed.

The human voice is an indicator of various emotions such as fear, sadness, anger, happiness, etc. These emotions reflect the mood of a person and hence can be used to determine their possible behaviour. For example: when a person is angry, he may be agitated. The prosodic voice features such as pitch, loudness and timber are used for voice analysis.[2] Depending on these features, a preliminary emotion or set of emotions can be attributed to the voice. Therefore, as the first step through the literature review, a clear link between emotions and agitation related voice behaviours such as shouting, screaming, crying is required and established. It has been found that anger and fear are directly related to these agitation behaviours**.**[3], [4]

The focus of this project is to identify voice features that can be extracted from speech in order to establish if the people talking exhibit signs of agitation. Further, an algorithm is built to provide the proof for whether the identified features are sufficient to detect agitation.

2. RESEARCH QUESTIONS

1. What are the most suitable voice features for agitation detection?
2. What is the accuracy of monitoring agitation with speech?
   2.1. Are there any differences in accuracy while detecting agitation in live speech compared to pre-recorded speech?

3. LITERATURE SURVEY

Fifteen papers have been surveyed on the topics of voice feature detection, emotion recognition and the relation between emotions and agitated behaviour. The goal of this was to establish a research trail that contains the relevant information to be able to use voice recognition to identify emotions that are linked to agitation.

Rojas et al.[5] created a system that analyses the voice of elderlies and detects sadness via their speech. After classifying a series of emotions and then processing the audio in terms of arousal and valence, the emotions were mapped to audio characteristics. Arousal and valence make up the two poles of the circumflex model of affect (*Figure 1*)[6], where arousal represents the intensity of an emotion and valence is the pleasure felt while expressing an emotion. This dual filter approach proved to have a high success rate. openSMILE[7] was used to extract the audio features from the files and a series of databases of audio samples pre classified per emotion were used to train and check the emotion classifier.

33rd Twente Student Conference on IT, July. 3rd, 2020, Enschede, The Netherlands. Copyright 2020, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science.

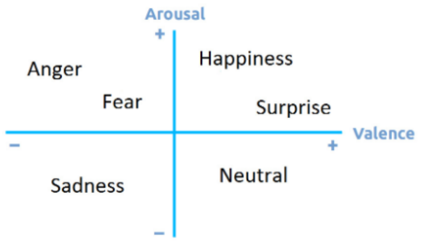


Figure 1: Circumflex model of affect

A different approach was taken by Omiya et al.[8] that analysed the vocalisations of three types of long vowels, specifically /Ah/, /Eh/, and /Uh/, in order to streamline the testing process by removing the variable of different words being spoken. openSMILE was used to extract acoustic features of the voices that were collected from a set of institutionalised patients suffering from clinical depression. The study reinforces the idea that voice analysis can be used to extrapolate psychiatric illness conditions such as depression or agitation.

Another approach by Pokorny et al.[9] was to utilise bags-of-audio-words to determine if there are negative emotions exhibited in speech. A bag-of-words is a simplified text representation that is used in natural language processing and analysis. The results of the research have shown a 65.6% accuracy in detecting negative emotions.

A study by Lexutan R. has tried to determine which audio features best represent the human voice. This was done by first collecting a dataset of audio recordings of human voices. Then a selection of feature extraction algorithms was used to compare which has the highest success rate in detecting emotions when combined with an emotion detection machine learning algorithm. The datasets used contained audio recordings made by actors and the data was pre labelled with the emotions that were exhibited. This was used to determine which particular features can optimally classify emotions such as happiness, sadness or a neutral state. The results of the research have shown that the most effective feature set out of the ones that were tested is: "pitch: mean value, standard deviation, maximum value, minimum value, range, median value; energy: mean value, standard deviation, maximum value, minimum value, range, median value; ratio of voiced/unvoiced segments – total of 13 features".[10]

A different study by Cheolwoo Jo and Jianglin Wang has used a different approach to try and detect emotions through audio analysis of voice recordings. They have used both the voice source and the vocal tract characteristics of audio recordings to determine emotions. The emotions that they have targeted are: neutral, happiness, anger, sadness, fear and boredom. They have collected a dataset of trained actors recording the six emotions mentioned above in various sentences. They have used Jitter, Shimmer, NHR (Noise-to-Harmonic ratio) pitch and pitch range as the audio features used by their algorithm to determine which emotions are detected. They have shown that anger is best observed by a high pitch range, whilst shimmer was best used to detect fear, as it had equal values for all the other emotions that were exhibited. Fear also showed the highest NHR values.[2]

Further, to establish a link between agitation behaviour and emotions, the study conducted by Volicer et al.[11] was used. They describe three main vocal tell-tale signs of agitation:

1. *High-pitched or loud noises,* which are characterised as louder or higher pitched than normal conversation levels. This includes calling out, shouting, yelling, crying out and screaming.
2. *Repetitive vocalizations*: comprised of repeated requests for information, repeated words, whining, muttering, grumbling, mumbling, rapid speech, crying and self-talk.
3. *Negative words*: words expressing negativity or using tones that are argumentative or demanding. This includes name calling (in a derogatory sense), swearing, cursing, profane language, hostile or threatening language, abusive or obscene language, argumentative and heated tones and being demanding.

This shows that emotional speech (when a person is angry or afraid) can be mimicked with agitated speech such as high-pitched voices. For example, when a person is angry, they tend to use a high pitch and when a person is agitated, they also tend to use high pitch. A way to detect agitation would be to detect whether certain emotions that are common in agitated patients are detected. Another method would be to analyse emotions exhibited by people and see if there are rapid switches in the emotions detected. This would indicate that someone is agitated as it constitutes rapid mood swings.

An emotion is a complex psychological state that has been associated with the nervous system. Emotions in general emerge from various neuropsychological changes attributed to thoughts, feelings, behavioural responses, and a degree of pleasure or displeasure. Whilst they occur on a biological level on the inside, each emotion is followed by tell-tale signs that are exhibited physically, be it through a smile in the case of happiness or crying in the case of sadness. In total there are 6 major emotions: happiness, sadness, surprise, fear, anger, and disgust. These emotions are also exhibited by people suffering from neuropsychiatric disorders such as dementia. It is important to note, however, that patients suffering from these disorders usually have trouble controlling their emotions or the way they express these emotions.[12] As such, it is possible to use emotion recognition to detect if such a patient is having an outburst of some kind, such as a period of agitation.

Anxiety is a type of fear, and it been associated with agitation in a study on elderlies with dementia by Twelftree et al.[4] As their research paper points out, anxiety is common in people suffering from neurological conditions, with a rate of between 38% in Alzheimer patients and 72% in vascular dementia patients. Anxiety has been associated with a reduced capacity to perform the tasks of day to day life of patients and it has also been associated with increased healthcare costs. Agitation has also been described as an expression of underlying anxiety. The research done uses statistical analysis on data collected from 40 participants with mild-to-moderate cognitive impairment. The main finding of the research was that agitation and anxiety are correlated in dementia patients.

Anger has also been linked with agitation in various studies aggregated in a paper on the fundamentals of anger by A. M. Shahsavarani and S. Noohi. Anger is one of the fundamental emotions expressed by humans. It can range from mild irritation to an outburst of rage. It has been shown there are physical side effects such as hypertension that result from anger which can "make thinking and decision-making procedures difficult and may harm physical and mental health". Anger is common in clinical cases of psychiatric patients.[3]

It can be observed that a clear link between voice-based emotion recognition and agitation appears to be missing in the literature. With the help of the above papers, the information necessary to make the link has been analysed and presented. The following section presents a proof of concept that will attempt to fill in the gap in the literature. The general flow of an approach to this, as extrapolated from the literature review, is presented in the diagram in Figure 2.

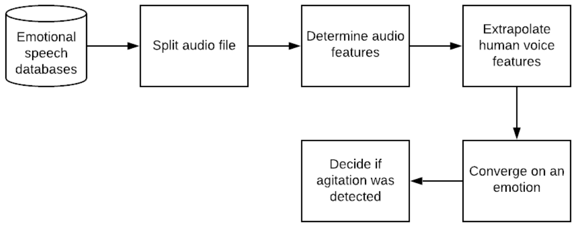

Emotional speech databases
Split audio file
Determine audio features
Extrapolate human voice features
Converge on an emotion
Decide if agitation was detected

Figure 2: Diagram of a theoretical algorithm

4. PROOF OF CONCEPT

The goal of this part of the research paper is to analyse if the literature findings are indeed valid for the purpose of identifying agitation in humans. In order to do this, an algorithm is developed that will be used to determine the validity of the findings and how accurately agitation can be detected through speech. This algorithm represents a proof of concept that will take the findings of the literature review and test their validity on the application of detecting agitation through voice-based emotion recognition. The algorithm will first take an audio file as input, it will then split the audio files into a number of audio samples. It will run audio analysis on these samples to determine values of certain audio features: frequency, amplitude and waveform. These are then extrapolated into the features of the human voice: loudness, changes in pitch and timbre respectively. Using these values, the algorithm will attempt to determine if any of the three emotions (anger, fear or sadness) is expressed in the audio recording. The expected output of the algorithm will be whether agitation has been detected in the audio file. After the algorithm has been tested on a subset of the audio files that have been left out of the main training set, tests will ideally also be done on live audio. The research team will have live conversations with around five people in order to determine the usability of the developed algorithm in live situations. Due to the current geopolitical circumstances surrounding government recommended self-isolation, the research team will attempt to carry out these conversations online via internet audio calls if possible.

4.1 Dataset collection

In order to achieve this, the algorithm will need to be developed on a training set and then tested on a different data set. For this, two databases SAVEE and RAVDESS that have audio-based emotion recordings were selected. SAVEE (Surrey Audio-Visual Expressed Emotion) is an English language database of audio and video samples. The accent of the actors is British. These are recordings of people expressing: anger, disgust, fear, happiness, neutral emotion, sadness and surprise. There is a total of 480 total audio files recorded by 4 male actors in this dataset. Of those, there are a total of 60 anger emotion recordings, 60 fear emotion recordings, 60 sadness emotion recordings and 60 neutral emotion recordings. The audio files are all in .wav (Waveform audio file) format, recorded in 16-bit at 44kHz. RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is an English language database of audio and video recordings of North American actors. For the purpose of this research, only the audio recordings are analysed. There are 1440 audio files from 24 actors (12 females and 12 males, all adults), each recorded twice for a grand total of 2880 files. The speech emotions in the dataset are: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. Each emotion is also expressed at two intensity levels: normal and strong, with the exception being the neutral emotion. There are 192 neutral emotion recordings and 672 recordings of anger, sadness and fear. The audio files are all in .wav (Waveform audio file) format, recorded in 16-bit at 48kHz.[13] In total, out of the 1104 total audio files that represent either a neutral emotion, anger, sadness or fear, 900 are used to train the proof of concept algorithm and the remaining 204 will be used to test it and to analyse the results of the findings.

4.2 Algorithm development

The algorithm has as a main goal to determine if an audio file given as input contains speech that can be categorised as agitated by using pitch and loudness as parameters. Given the format of the dataset that was compiled for this proof of concept, the audio contains emotional human speech that has the emotion expressed tagged onto each file. The algorithm is developed in a Node.js programming environment.

The first step is taking a raw .wav audio file. Depending on which database is used, the algorithm needs to be set up to accept either 44kHz audio or 48kHz audio, based on the format that each database has used for their recordings. The human voice has a frequency range from around 100Hz to around 17KHz, depending on age and gender. The algorithm applies a band pass filter to limit the audio range that it analyses from 300Hz to 4000Hz. This is done because the human voice features are mainly concentrated in this smaller range. This filter also eliminates a lot of potential background noise and it also makes the computation much faster. There still exists the possibility of background noise in theory, however, due to the controlled way of colleting the dataset, the algorithm assumes that there is an insignificant amount of noise in the background.

The next step is taking the audio file and splitting it up into equal sized “windows”. This is done via a windowing algorithm, which is necessary to facilitate the next step. The windowing algorithm separates each audio file in windows of 16384 (2 to the power of 14) samples each. This value was chosen through experimentation during the development of the algorithm and it was found that this represents the best accuracy, when compared to smaller window sizes. The window size and the sample rate (44kHz or 48kHz) are used to compute variables such as the number of windows per second and the window duration. This duration is of approximately 20ms per window.

It is important to note that an issue known as spectral bleeding occurs when an audio file is broken up into windows. This could be solved by applying a hamming algorithm to the windows in order to even out the starting and ending frequencies that appear when the cut is made. It was determined, however, that since the same bleeding occurs both in the training set and in the testing set, the differences would cancel each other out. Thus, for the sake of reducing complexity, hamming has been left out. The purpose of this step is to enable the algorithm to transform the standard audio file format which is a function of amplitude and time (a discrete signal) (Figure 4) into a function of amplitude and frequency (a discrete spectrum) (Figure 5). This is done by applying a Discrete Fourier Transform on each of the windows created in the previous step. The following is the mathematical formula of the Discrete Fourier Transform (Figure

3), where N is the size of the window, X(n) represents the nth bin of frequencies and x(k) is kth sample of the audio signal.

$$X(n)=\sum_{k=0}^{N-1}x[k]e^{-j(2\pi kn/N)}$$

Figure 3: Fourier Transform

For efficiency purposes, the algorithm utilises a Fast Fourier Transform (FFT) which has a time complexity of O(N log N) compared to the O(N^2) complexity of simply applying the standard mathematical formula given above. N here represents the number of data sets to be computed. After this, we have a set of both frequency and amplitude for every window. Since the frequency represents the pitch and the amplitude represents the loudness, it is possible to analyse the data. What the algorithm does next is that it computes a weighted average of the frequency of each window. The main variable is the frequency while the weight is the amplitude of the signal. This is done in accordance with the literature review as it has been shown that frequency is the key factor in determining anger in audio recordings.

After this step, there will be a value assigned to each window which represents the average frequencies that are most represented in intensity of speech. The values of all the windows are now averaged out to compute a value that represents the audio signature of the recording. This is the value that the algorithm uses to determine agitation (Figure 7).

A different approach was taken in determining fear and sadness. As per the literature review, instead of analysing the frequency, an analysis is made on the peak to peak variation of the amplitude of the discrete spectrum.[2] The rest of the steps are the same, however, the weighted average of frequency is replaced by the peak to peak variability of the amplitude. Another method that was used was to determine the average difference in peaks of weighted frequencies. All of these values combined should offer a clear view of whether a person is agitated.

All of the above-mentioned computations are achieved for a 3 second audio file in an average of under 100 milliseconds. This means that the algorithm is in theory capable of running and providing output in real time.

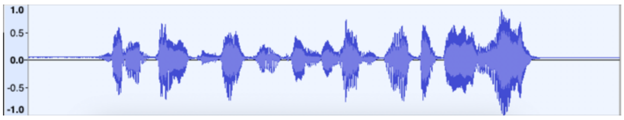


Figure 4: Discrete Signal (before FFT)

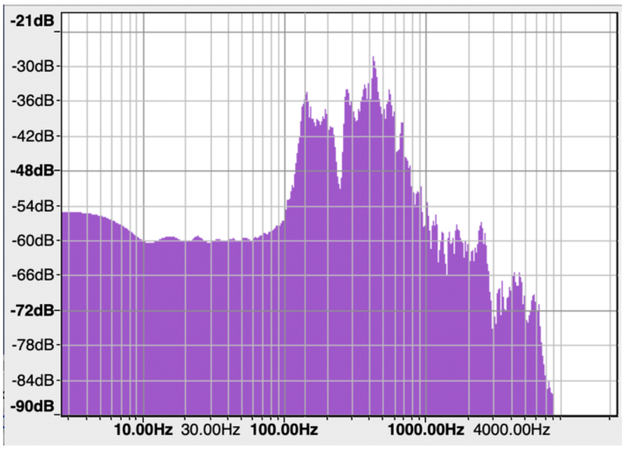


Figure 5: Discrete Spectrum (after FFT)

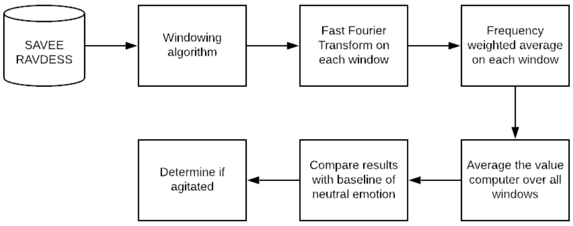
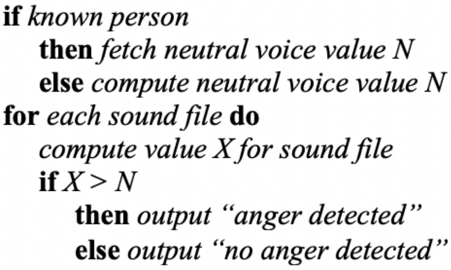

SAVEE RAVDESS
Windowing algorithm
Fast Fourier Transform on each window
Frequency weighted average on each window
Average the value computer over all windows
Compare results with baseline of neutral emotion
Determine if agitated

Figure 6: Diagram of proof of concept



```
if known person
    then fetch neutral voice value N
    else compute neutral voice value N
for each sound file do
    compute value X for sound file
    if $X > N$
        then output "anger detected"
        else output "no anger detected"
```

Figure 7: Anger detection logic

5. RESULTS

In order for the algorithm to determine if someone is agitated, a baseline needs to be established of the person speaking with a neutral tone first. This is used as the basis on top of which the other emotions can be detected. With the given frequency weighted average technique anger can be detected in someone, though the accuracy is not very high. With fear and sadness, none of the techniques have proved to be reliable enough. One of the possible issues can be a lack of relevant data. On closer inspection, the algorithm showed false positives by identifying some happiness recordings as being sad. Fear was also misidentified with surprise. The techniques used for sadness and fear proved to unreliable. A consolidated table for the relevant results is presented below (Table 1).

Anger detection algorithm	Anger	Not anger (other emotions)
Correct result	73%	68%
Wrong result	27%	32%

Table 1: Results of anger analysis algorithm

6. LIMITATIONS

The current system is only able to detect anger reliably. Fear and sadness might get confused with other emotions. There are two major reasons for this: first is a lack of accurate data and second is that a more adequate or advanced algorithm is required. The paper by J. Cheolwoo and W. Jianglin that links audio features such as frequency and amplitude with specific emotions had its database in the South Korean language. Considering that each language has its own signature sound, the same features and techniques that work for one language might not be reliable when testing for other languages. It also needs to be taken into account that spectral bleeding was not countered. It can also have an impact on the accuracy of the overall system.[14] It is also important to note that the datasets that were collected were of actors acting out the emotions. There might be a contrast to real emotions exhibited genuinely in real life human speech.

Hence, using only frequency and amplitude is not enough to determine agitated speech reliably. The Fourier transform algorithm only provided information on those two voice features. It is possible to extrapolate other voice features such as timbre or tone [15], as this might provide different viewpoints on the audio data. The approach used represents only a subset of the possibilities of analysing the data. It could be that a different method to determine if someone is angry, sad or fearful could be implemented using the data available via the Fourier Transform, which might be more effective at the task.

Due to the current situation revolving around the COVID-19 pandemic and the resulting stay at home recommendation from the Dutch government, it proved impossible to make a reliable live audio test. It was impossible to visit any healthcare institution in order to record agitation data from patient. An attempt was made to analyse speech online by asking people to act agitated over the call. But it was discovered that without trained actors, the quality of acted emotions was subpar. It proved unreliable to use live audio of average healthy untrained people and thus the question of validity of the findings with regards to comparing live audio with pre-recorded audio was not able to be answered.

7. CONCLUSION

The literature review on agitation showed that patients expressing agitation usually express three key emotions: anger, sadness and fear. It is, thus, possible to extrapolate that if someone is expressing any of those emotions in a clinical environment, they might be agitated. A set of audio features, specifically frequency and amplitude, which represent pitch and loudness, respectively, has been identified as a good method to identify various human emotions. A proof of concept was made to determine if these emotions can be identified in order to determine if someone is agitated. The link from agitation to emotion recognition was made in the literature review and then tested in the algorithm development section. The results have shown that the chosen audio features do not create a reliable algorithm to detect agitation. More audio features are necessary to make this connection.

8. DISCUSSION & FUTURE WORK

Future work could encompass a greater variety of audio features. By analysing multiple dimensions of the audio files, an algorithm that more reliably detects agitation could be developed. Features that could be considered are waveform, timbre, various other signal properties or the mel-spectrum. A different approach can be taken by compiling a dataset of audio recordings that represent genuine human emotions, instead of those recorded by actors. There could be a major difference in the results when real emotional speech is used. Ideally, these recordings would be of patients who are agitated. Another approach that a research team can take is to utilise machine learning algorithms on audio files that are marked as agitated. When it comes to the accuracy of the readings, attempting to solve the spectral bleeding issue could improve the results. Another way that this could be solved, apart from applying a hamming algorithm, as discussed in the methodology section, would be by overlapping the windows. This would even out the differences, which might improve accuracy.

A possible application of an emotion detection system is outlined. It has been shown that social isolation and loneliness are linked to high impact neurological conditions such as depression.[16], [17] Therefore a possible application is for such a system to detect agitation in elderlies. This system can facilitate the connection between elderlies and their caregivers by providing live information with regards to the elderlies' state of agitation. It has been shown that an unobtrusive system is important when working with elderlies. People are discomforted by wearing microphones all day, so a system which can record only prosodic features can be successful.[18] Therefore, there is value in developing a technical voice based unobtrusive detection system that can improve the work conditions of caregivers by alerting them when an elderly is agitated.

9. ACKNOWLEDGEMENT

I would like to thank my supervisor Nikita Sharma for the help provided with the literature review and for the guidance given for writing this paper. I would also like to thank the track chairs, Dr. Le Viet Duc and Dr. Ir. Alex Chiumento, for their help in writing the paper and in reaching the milestones set by the university in a timely manner.

10. REFERENCES

[1] J. , M. J. , B. H. , S. M. , B. S. , D. D. P. , . . . & P. E. Cummings, “Agitation in cognitive disorders: International Psychogeriatric Association provisional consensus clinical and research definition.,” *International Psychogeriatrics, 27(1), 7-17.*, 2015.

[2] J. Cheolwoo and W. Jianglin, “Measuring variations of voice source and vocal tract characteristics from Korean emotional voice,” in *Proceedings - ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications*, 2006, vol. 2, pp. 800–805, doi: 10.1109/ISDA.2006.253715.

[3] A. M. Shahsavarani and S. Noohi, “Explaining the Bases and Fundamentals of Anger: A literature Review,” 2014. Accessed: Jun. 10, 2020. [Online]. Available: http://www.ijmedrev.com/&url=http://www.ijmedrev.com/article_68914.html.

[4] H. Twelftree and A. Qazi, “Relationship between anxiety and agitation in dementia,” *Aging and Mental Health*, vol. 10, no. 4, pp. 362–367, Jul. 2006, doi: 10.1080/13607860600638511.

[5] V. Rojas, S. F. Ochoa, and R. Hervás, “Monitoring moods in elderly people through voice processing,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8868, pp. 139–146, 2014, doi: 10.1007/978-3-319-13105-4_22.

[6] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, doi: 10.1037/h0077714.

[7] F. Eyben, F. Weninger, M. Wöllmer, and B. Schuller, “openSMILE the Munich open Speech and Music Interpretation by Large Space Extraction toolkit TU München, MMK,” 2013. Accessed: Apr. 28, 2020. [Online]. Available: http://www.mmk.ei.tum.deTheofficialopenSMILEhomepagecanbefoundat:http://opensmile.sourceforge.net/.

[8] Y. Omiya *et al.*, “Estimating depressive status from voice,” in *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, Jan. 2019, pp. 2795–2796, doi: 10.1109/BIBM.2018.8621326.

[9] F. B. Pokorny, F. Graf, F. Pernkopf, and B. W. Schuller, “Detection of negative emotions in speech signals using bags-of-audio-words,” in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, Dec. 2015, pp. 879–884, doi: 10.1109/ACII.2015.7344678.

[10] R. M. Lexutan, “Comparative study regarding characteristic features of the human voice,” in *Proceedings of the 2015 7th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2015*, Oct. 2015, pp. WSD1–WSD4, doi: 10.1109/ECAI.2015.7301206.

[11] L. Volicer and G. L. Odenheimer, “Measurement of observed agitation in patients with dementia of the Alzheimer type.” [Online]. Available: https://www.researchgate.net/publication/291106096.

[12] E.-H. Kong, “Agitation in dementia: concept clarification,” *Journal of Advanced Nursing*, vol. 52, no. 5, pp. 526–536, Dec. 2005, doi: 10.1111/j.1365-2648.2005.03613.x.

[13] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english,” *PLoS ONE*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[14] F. J. Harris, “On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978, doi: 10.1109/PROC.1978.10837.

[15] T. R. Agus, C. Suied, S. J. Thorpe, and D. Pressnitzer, “Characteristics of human voice processing,” in *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, 2010, pp. 509–512, doi: 10.1109/ISCAS.2010.5537589.

[16] R. S. Tilvis, P. Routasalo, H. Karppinen, T. E. Strandberg, H. Kautiainen, and K. H. Pitkala, “Social isolation, social activity and loneliness as survival indicators in old age; A nationwide survey with a 7-year follow-up,” *European Geriatric Medicine*, vol. 3, no. 1, pp. 18–22, Feb. 2012, doi: 10.1016/j.eurger.2011.08.004.

[17] E. Y. Cornwell and L. J. Waite, “Social disconnectedness, perceived isolation, and health among older adults,” *Journal of Health and Social Behavior*, vol. 50, no. 1, pp. 31–48, Mar. 2009, doi: 10.1177/002214650905000103.

[18] Enamul Hoque, Robert F. Dickerson, and John A. Stankovic, “Vocal-Diary : A Voice Command based Ground Truth Collection System for Activity Recognition.”