ISSN 2347 - 3983
Volume 8. No. 4, April 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter43842020.pdf
https://doi.org/10.30534/ijeter/2020/43842020

Speech Emotion Recognition using Support Vector Machines

Aaron Don M. Africa, Anna Rovia V. Tabalan, Mharela Angela A. Tan
De La Salle University, Manila
2401 Taft Ave., Malate, Manila 1004,
Philippines, aaron.africa@dlsu.edu.ph

ABSTRACT

The technology of recognition is one that has been developed continuously over the years and with its various applications in a wide variety of fields opens up massive opportunities to bridge the gap between humans and computers. Albeit common knowledge that computers are designed to make everyday life easier, there is still an indubitable lack of deep understanding due to the computer's lack of knowledge in complex emotions present with human beings and this often prohibits computers to offer specific help that is suitable for its user. Therefore, it's important to further develop today's technology and one promising way to accomplish this task is to utilize Speech Recognition to recognize and classify emotions as well. This way, the computer essentially understands the user enough to give valuable aid instead of just preset actions. Support Vector Machine is one of the leading classifying algorithms in today's time, boasting the highest accuracy rate which makes it the most viable option for this field of study.

Key words : Speech Recognition, Support Vector Machine, Emotions.

1. INTRODUCTION

Artificial Intelligence systems have developed quite extensively over the years and it has been one of the breakthroughs throughout humanity's technological advancement that were a couple of eras ago, only existed in the form of stone axes for cave dwellers. The rapid development of our technology has allowed humans to increase their dependency on computers to make everyday lives easier. This type of serving technology is used for countless reasons but one field that is not majorly focused on is serving the user's emotional health. Emotional health is characterized as the wellbeing of a person based on the state of his/her emotions. Although not majorly focused on, emotional health is a serious subject in need of attention.

Speech emotion recognition is one of the fields that AIs are currently being used on. Emotions include Anger, Happiness, Sadness, Calm, Anger and Disgust. Past research showed that fear and disgust have low recognition rates compared to other common emotions [1]. Other past studies have been conducted to automatically detect the emotion in one's speech. Most of the research focused on the classification's performance however a more efficient speech analysis technique has not yet been developed completely [2]. Many of this speech emotion recognition features depend on the pitch, frequency, time, loudness, etc. One study considers the intensity of emotions to quantify emotions [3].

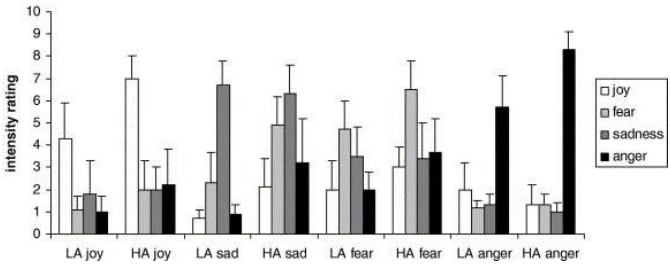

Figure 1: High and Low Intensity Emotions

This field of study is especially significant for people with special circumstances and in need of extended help that another human may not be able to give constantly. For example, a study conducted in 2014 focusing on "alexithymia index" in students showed that a significant number of students are prone to alexithymia, which is an indicator that one has some sort of deficit in their emotional health. [4]. Another assessment concerning this field done in 2016 shows that children exposed to trauma during their childhood develop weaker physical, mental, and emotional health [5]. In this current era, we're in, stress and trauma are ever so present at all points in human life, in some cases, this is inevitable which is why it's vital for effective external help, help that advanced technology will be able to provide.

An umbrella term encompasses this specific field of technology, E-Health or electronic health. This serves as a general term for electronic tools, systems, and processes used in the health department which helps health professionals with their work and provides improved productivity and efficiency at the same time. A research done in 2019 showed that health professionals do support the use of technology in helping them with providing care to their patients, however, it should not be a complete substitute for a face-to-face session or be the sole decision-maker instead of a professional's autonomy over which procedures or medications need to be administered to the patient. Professionals also pointed out that aside from developing this technology, e-health literacy must also be given importance to develop the necessary skillset to

use these tools [6]. One method of e-health is using music therapy to aid people with selective needs. A research conducted in 2019 focused on playing the music of different genres to participants and rating their emotional response [7].

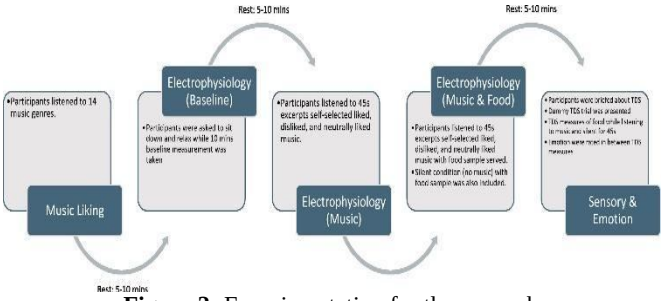

Figure 2: Experimentation for the research

This issue has taken the interests of many technological researchers and they have been looking into achieving a more human-like performance of these AI systems by eliminating the evident barrier between humans and computers. Performing this requires the computer to perceive human emotions and have an adaptive response that would further provide the user with a more suitable and interactive companion. To address this concern, many studies developed embedded systems with various emotion recognition algorithms. Most of this research focuses on comparing the accuracy rates of multiple algorithms to determine the best possible one to use. One of the best possible algorithms is Support Vector Machine (SVM), which operates by drawing a line to divide the classes it's trained to classify.

2. LITERATURE REVIEW

A novel feature selection method for speech emotion recognition is a study done by Turgut Özseven [8]. They proposed a different method of statistically selecting features from a speech that differs from the current one being used in such a way that the features selected will be decreased whilst still having significant improvement in the accuracy. The paper argued that speech emotion analysis requires several features and these features aren't always useful or important for the application it's being used for. Moreover, different emotions can affect different features which in turn will undermine the accuracy of the system. Their solution to this proved to be successful compared to the related research in this field. Figure 2 shows a significant reduction in the features needed for speech analysis for their proposed method.

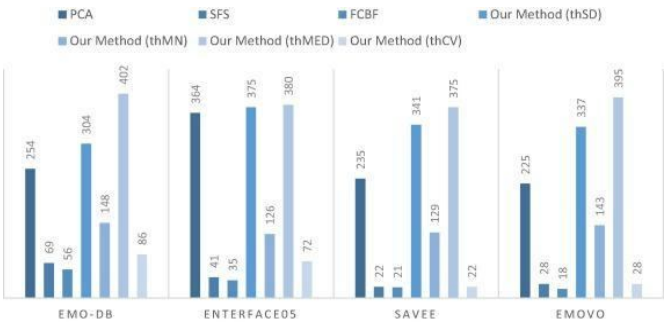

Figure 3: The decrease in Features Selected

Classification of silent speech using a support vector machine and relevance vector machine is a research done in 2014 by Mariko Matsumoto and Junichi Hori [9]. This is a study of feasibility for artificial speech using an imaginary voice. This is done with a derivative of the support vector machine which is the relevance vector machine with a gaussian kernel. Both the vector machines performed well but the RVM-G had reduced ratio of vectors to training data, this also provided a higher accuracy. However, classification accuracies with RVM-G proved to be quite weaker than SVM-G when a significantly fewer number of training data is available. From the results gathered and shown in Figure 3, the RVM-G's performance was better than SVM-G but in other cases, SVM-G's performance is slightly better and there was no significant difference between the two. This study is especially important because it measures the feasibility of the algorithm as well as the accuracy when multiple classes are present, which is usually the case for speech emotion recognition.

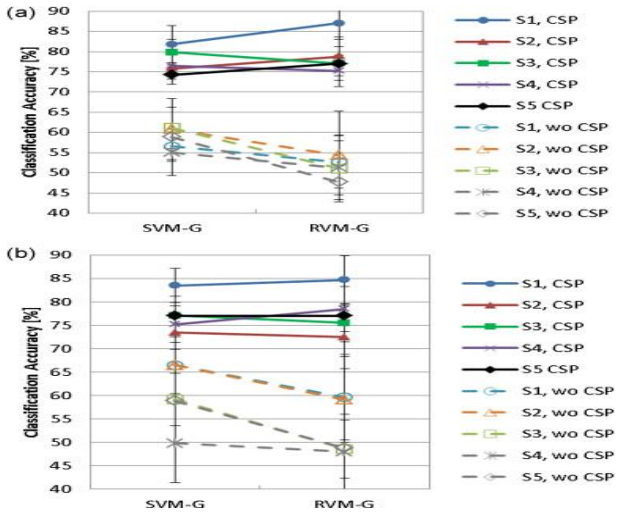

Figure 4: Comparison between RVM-G and SVM-G

Bagged support vector machines for emotion recognition from the speech is a research done by Anjali Bhavan, Pankaj Chauhan, Hitkul, Rajiv Ratn Shah in 2019 [10]. The study uses three databases for the research and features extracted from these databases are reduced and processed even further. With this, they used a bagged ensemble as the basis for Ensemble learning, which proved to be better than single estimators, and support vector machine with a gaussian kernel as the main algorithm. Ensemble learning works by combining useful learning methods from multiple models to create a newer and more efficient model. From the system overview of the research, as shown in Figure 4, one of the most important parts of speech emotion recognition are the features that are valuable to the analysis. One of the most popular ones is the Mel-Frequency Cepstral Coefficients (MFCCs) which gave a 91.3% and 95.1% accuracy two out of three databases used. All the accuracies are plotted and shown in Figure 5.

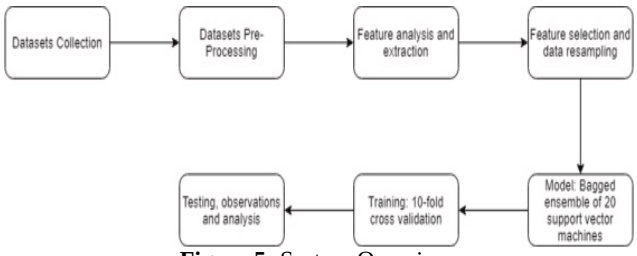

Figure 5: System Overview

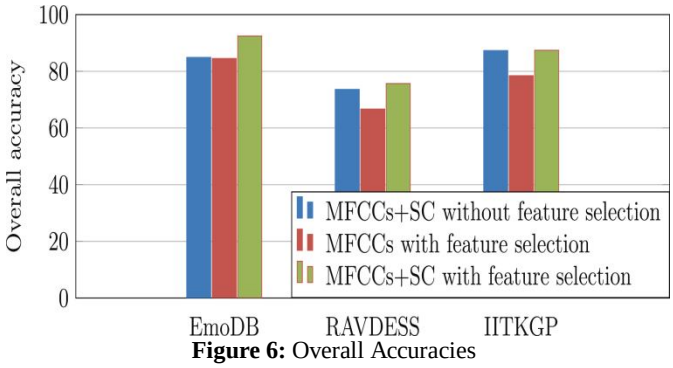

Figure 6: Overall Accuracies

Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition is another research done by Turgut Özseven in 2018 [11]. The study focuses on determining the effects of different methods of analysis on speech emotion recognition, specifically texture analysis. The accuracy of said methods is then determined by support vector machines. Figure 6 shows the process of feature extraction of the system. In their study, much like the previous one, they used MFCCs with other formant frequencies. The researchers used SVM which is derived from statistical learning theory. SVM is commonly used with multiple classes which makes it the best option for this kind of study. The kernel used for this study is a linear kernel due to the number of features needed.

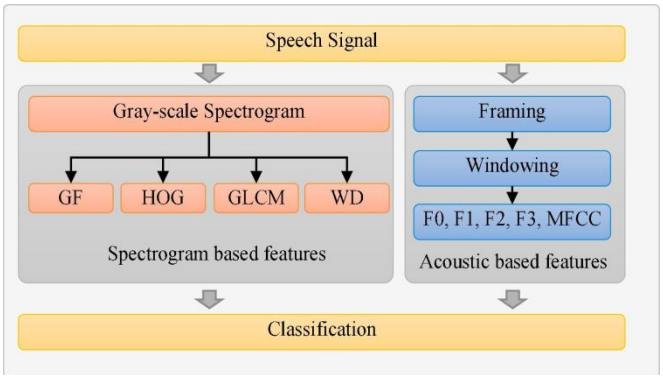

Figure 7: Feature Extraction Process

Boosting the selection of speech-related features to improve the performance of multi-class SVMs in emotion detection is another study done in 2009 by HalisAtlun and GökhanPola [12]. The study focuses on the importance of improving the feature selection process rather than the classifier itself. The researchers proposed a new method of feature selection by combining four existing methods and three classifiers. The result of the study showed that prosodic and subband energy features are most likely to be extracted by the algorithms. Figure 7 shows the performance of Framework 1 which is a "one vs set" framework and Framework 2 which is a "one vs one" framework. As seen from the result, a one-vs-one approach has a much better performance than other multi-task classifiers.

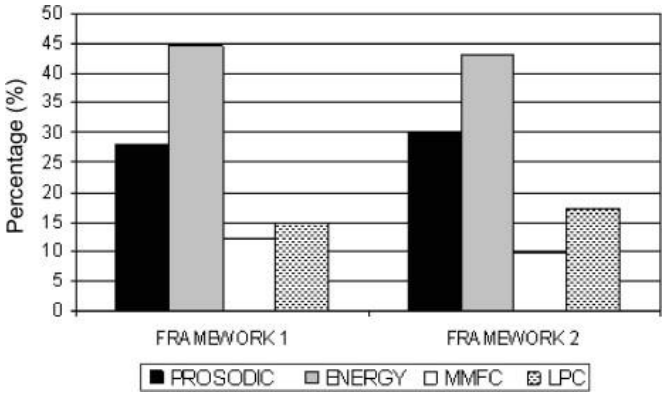

Figure 8: Framework 1 and Framework 2 Comparison

3. THEORETICAL CONSIDERATION

The main feature is the Support Vector Machine. SVM is a machine learning algorithm that uses structural risk minimization. SVM works but mapping an N- dimension input into a higher feature space using various kernel functions. Afterward, the algorithm will try to find the best possible generalization to separate the classes into its respective hyperplanes. To train a support vector machine algorithm, several methods can be used. One promising way is to use a k-nearest neighbor with a gaussian kernel. A study done in 2008 compares various methods such as LS-SVM, FLS-SVM, and LS+k-NN-SVM with that of clustered KSVM and concluded that the latter had the best accuracy out of all the standard methods [13]. Factors such as the size of any dataset that may be used or the complexity of the hyperplane or hypersurface can alter and increase the number of support vectors that will be required. [14]. This is illustrated in Figure 7 which shows how the required number of support vectors changed and the performance concerning that change.

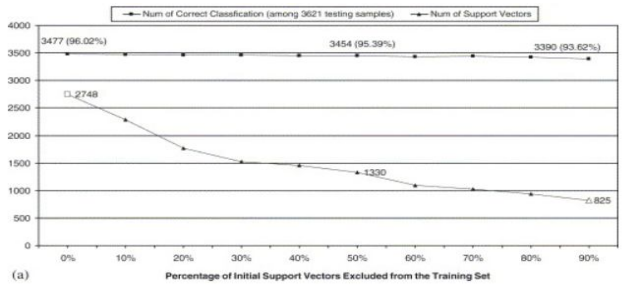

Figure 9: Performance of SVM concerning the number of support vectors

SVMs are especially useful with applications that require a generalization of training sets, because of this, the likelihood

of overfitting is greatly decreased. Another figure that illustrates the process of SVM is shown in Figure 8 which shows how it maps data into another high-dimensional space [15,16,17].

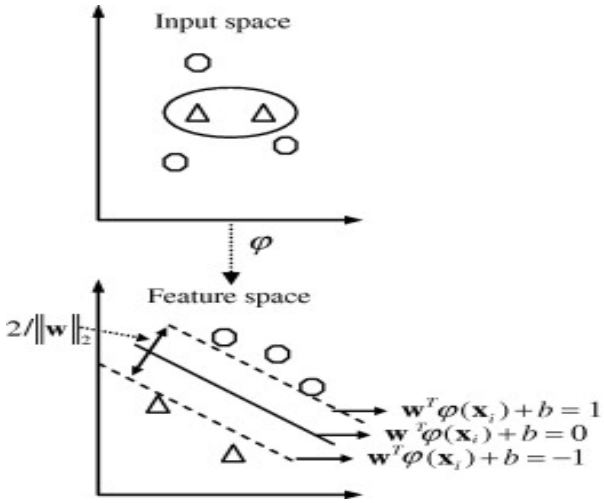

Figure 10: Training data mapped into a higher dimension space

Another main topic is the feature selection process for multi-class classification. The usual researches aim to increase the feature selection algorithms as well as its effectiveness. Another goal is to increase the accuracy of the classifiers when put through different methods and strategies One example study on this is in quadrotor based system detection [18]. However, many past studies have been conducted, there is still the problem of specifically determining the most useful feature in a speech analysis. This is severely important as features are much more consequential than the classifier itself. Because of this, irrelevant or trivial features are more than capable of significantly reducing the accuracy of any classifiers. In line with this, reducing the size of a feature set, using a fast and efficient classifier and understanding the process are the three main objectives in feature selection.

4. DESIGN CONSIDERATION

Feature selection is the most significant part of any speech or acoustic analysis. Figure 10 shows a proposed process for feature selection that takes into account two possible frameworks, one-vs-rest, and one-vs-one. One-vs-rest compares one class with the rest of the class while the one-vs-one compares each class with each other. The latter will need more computational power of course but it has been proven to provide better performance than the other approaches. After this, the actual feature selection is done, and a final set of features will be constructed, and the output will be fed through a multi-class classifier. This research can be done in a software simulation. It can follow the configurations of [19,20] and can use a Control System Feedback mechanism.

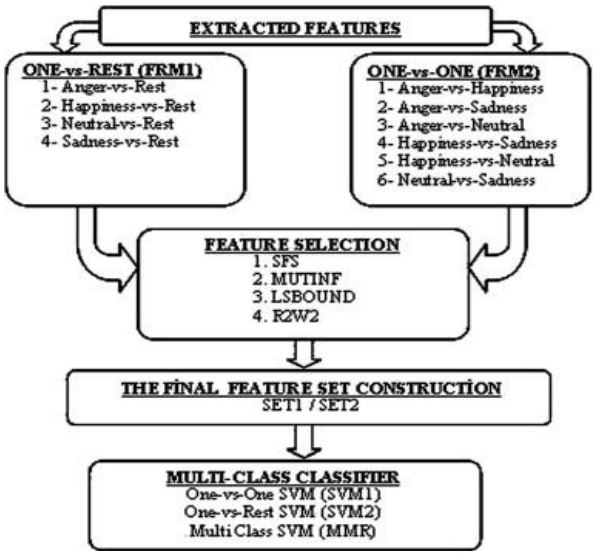

Figure 11: Feature selection process flow

5. CONCLUSION

Speech emotion recognition is a relatively new technology that is derived from the existing advancement in emotion recognition. With the multiple emotions and classes present, as well as many factors to be considered, the support vector machine is one of the leading algorithms and classifiers to be used in this application. With the complexity of human emotions and the features that can be extracted from speech, it is important to develop a new speech analysis that is much more efficient than the current ones. Additionally, it's also vital to understand that human emotions are extremely complex and even though this technology lets us quantify these emotions, they are ultimately much more in-depth for just one profession, hence the application of technology to this study must be done with the input of multiple professionals, from engineering to linguists, to psychology professionals.

REFERENCES

[1] L. Chen, X. Mao, Y. Xue, and L. Cheng, "Speech emotion recognition: Features and classification models," Digital Signal Processing. Vol. 22, No. 6, pp. 1154-1160, 2012. https://doi.org/10.1016/j.dsp.2012.05.007

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech Communication. Vol. 48, No. 9, pp. 1162-1181, 2006.

[3] T. Bänziger and K. Scherer, "The role of intonation in emotional expressions,"Speech Communication. Vol 46, Nos. 3–4, pp. 252-267, 2005. https://doi.org/10.1016/j.specom.2005.02.016

[4] E. Galván, "Alexithymia: Indicator of Communicative Deficit in Emotional Health," Procedia - Social and Behavioral Sciences. Vol. 132, pp. 603-607, 2014.

[5] C. Campbell, Y. Roberts, F. Synder, J. Papp, M. Strambler, and C. Crusto, "The assessment of early trauma exposure on social-emotional health of young children," Children and Youth Services Review. Vol. 71, pp. 308-314, 2016. https://doi.org/10.1016/j.childyouth.2016.11.004

[6] L. Lolich, I. Riccò, B. Deusdad, and V. Timonen, "Embracing technology? Health and Social Care professionals' attitudes to the deployment of e-Health initiatives in elder care services in Catalonia and Ireland," Technological Forecasting and Social Change. Vol. 147, pp. 63-71, 2019.

[7] R. Wason, "Deep learning: Evolution and expansion," Cognitive Systems Research. Vol. 52, pp. 701-708, 2018. https://doi.org/10.1016/j.cogsys.2018.08.023

[8] T. Özseven, "A novel feature selection method for speech emotion recognition," Applied Acoustics. Vol. 146, pp. 320-326, 2019.

[9] M. Matsumoto and J. Hori, "Classification of silent speech using support vector machine and relevance vector machine," Applied Soft Computing. Vol. 20, pp. 95-102, 2014. https://doi.org/10.1016/j.asoc.2013.10.023

[10] A. Bhavan, P. Chauhan,Hitkul, and R. Ratn Shah, "Bagged support vector machines for emotion recognition from speech," Knowledge-Based Systems. 2019.

[11] T. Özseven, "Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition," Applied Acoustics. Vol. 142, pp. 70-77, 2018.

[12] H. Altun and G. Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection," Expert Systems with Applications. Vol. 36, No. 4, pp. 8197-8203, 2009. https://doi.org/10.1016/j.eswa.2008.10.005

[13] A. Africa, A. Alcantara, M. Lagula, A. Latina and C. Te, "Mobile phone graphical user interface (GUI) for appliance remote control: An SMS-based electronic appliance monitoring and control system," International Journal of Advanced Trends in Computer Science and Engineering. Vol. 8, No. 3, pp. 487-494, 2019. https://doi.org/10.30534/ijatcse/2019/23832019

[14] A. Africa, L. Bulda, M. Marasigan, and I. Navarro, "A study on number gesture recognition using neural network," International Journal of Advanced Trends in Computer Science and Engineering. Vol. 8, No. 4, pp. 1076-1082, 2019. https://doi.org/10.30534/ijatcse/2019/14842019

[15] A. Africa, P. Arevalo, A. Publico, and M. Tan, "A fuzzy neural control system," International Journal of Emerging Trends in Engineering Research. Vol. 7, No. 9, pp. 323-327, 2019. https://doi.org/10.30534/ijeter/2019/15792019

[16] L. Torrizo and A. Africa, "Next-hour electrical load forecasting using an artificial neural network: Applicability in the Philippines," International Journal of Advanced Trends in Computer Science and Engineering. Vol. 8, No. 3, pp. 831-835, 2019. https://doi.org/10.30534/ijatcse/2019/77832019

[17] A. Africa, G. Ching, K. Go, R. Evidente, and J. Uy, "A comprehensive study on application development software systems," International Journal of Emerging Trends in Engineering Research. Vol. 7, No. 8, pp. 99-103, 2019. https://doi.org/10.30534/ijeter/2019/03782019

[18] G. Guevarra, A. Koizumi, J. Moreno, J. Reccion, C. Sy, and J. Del Rosario, "Development of a quadrotor with vision-based target detection for autonomous landing," Journal of Telecommunication, Electronic and Computer Engineering. Vol. 10, No. 1-6, pp 41-45, 2018.

[19] A. Africa, P. Arevalo, A. Publico, and M. Tan, "A comprehensive study of the functions and operations of control systems," International Journal of Advanced Trends in Computer Science and Engineering. Vol. 8, No. 3, pp. 922-926, 2019. https://doi.org/10.30534/ijatcse/2019/89832019

[20] A. Africa, F. Espiritu, C. Lontoc, and R. Mendez, "The integration of computer systems into the expansive field of video games," International Journal of Advanced Trends in Computer Science and Engineering. Vol. 8, No. 4, pp. 1139-1145, 2019. https://doi.org/10.30534/ijatcse/2019/22842019