A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions

Rezwan Matin
Ingram School of Engineering
Texas State University
San Marcos, USA
r_m727@txstate.edu

Damian Valles
Ingram School of Engineering
Texas State University
San Marcos, USA
dvalles@txstate.edu

***Abstract*— Children who fall into the autism spectrum have difficulty communicating with others. In this work, a speech emotion recognition model has been developed to help children with Autism Spectrum Disorder (ASD) identify emotions in social interactions. The model is created using the Python programming language to develop a machine learning model based on the Support Vector Machine (SVM). SVM has proven to yield high accuracies when classifying inputs in speech processing. Individual audio databases are specifically designed to train models for the emotion recognition task. One such speech corpus is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which is used to train the model in this work. Acoustic feature extraction will be part of the pre-processing step utilizing Python libraries. The libROSA library is used in this work. The first 26 Mel-frequency Cepstral Coefficients (MFCCs) and the zero-crossing rate (ZCR) are extracted and used as the acoustic features to train the machine learning model. The final SVM model provided a test accuracy of 77%. This model also performed well when significant background noise was introduced to the RAVDESS audio recordings, for which it yielded a test accuracy of 64%.**

Keywords—emotion recognition, speech, autism spectrum disorder, SVM, MFCC, ZCR.

I. INTRODUCTION

The field of speech processing has already been through decades of research. Some of the popular tasks include automatic speaker recognition (ASR) [1], language recognition [2], mental stress detection [3], and others. Traditionally, the Hidden Markov Model (HMM) has been extensively used in speech processing [4][5]. However, after the advent of machine learning, HMM has been all but replaced. Machine learning models have been performing well in all speech detection tasks. The author in [6] used the Binary Tree-Structured Support Vector Machine (BTSS) to classify voice and music in data. In [7], the authors used the Support Vector Machine (SVM) classifier along with two datasets – the General Sounds dataset (containing audio of applause, birds, cars, etc.), and Audio Scenes dataset (containing audio of park, bar, station, etc.). An audio file can be converted into a spectrogram, which then allows researchers to apply image processing techniques to analyze the data. An example is using Convolutional Neural Networks (CNN) to classify audio, as presented in [8]. Authors in [9] stated that the Mel-frequency spectrogram showed better classification accuracy than the classical spectrogram.

Speech emotion recognition is a field of research that has been of enormous interest to the research community. Emotion recognizers find their application in many areas - call centers assessing customer satisfaction, e-learning systems, assistive robotics, security agencies, military organizations, and many more [10][11][12]. In this study, the emotion recognition system is intended to aid children with Autism Spectrum Disorder (ASD). Once the model is ready, it can be used to train these children so that they get better at identifying human emotions in face-to-face conversations.

This research work involves developing a speech emotion recognition model using the SVM learning algorithm to assist children with ASD in identifying various emotions. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) speech emotion corpus was used to train and test the speech emotion recognition model. It involves 1,440 speech recordings of 24 actors in eight different emotional states. The final SVM model was able to predict emotions with a classification accuracy of 77% using only a handful of acoustic features. The SVM model was later tested on the same speech recordings but with some added background noise. This was done to simulate real-world social interactions. The model showed satisfactory performance, considering the amount of noise introduced to the data. The accuracy dropped to 64% with the noisy data. In future work, the model will be expanded to include more speech corpora.

II. BACKGROUND

The human species thrive as a community, and for that reason, communication is of utmost importance. Through social interactions, information is exchanged and feelings are conveyed to others. The detection of emotional *valence*—the intrinsic attractiveness/'goodness' (positive valence) or averseness/'badness' (negative valence) of a person, object, or situation provides crucial information for decision making [13]. Emotions with the same valence (e.g., anger and fear) produce a similar influence on judgments and choices [14]. The *arousal*, on the other hand, is simply the intensity of each affective state.

Human interactions are of two types – verbal and non-verbal. Non-verbal communication can be further broken down into facial expressions and body language. Reference [15] listed six universally recognized facial expressions - happiness, sadness, surprise, anger, fear, and disgust. The work in [16] defined the 7-38-55 rule of personal communication (also known as the 7% rule). This rule states words account for 7% of communication, while the majority of the information is the vocal tone (55%) and body language (38%).

978-1-7281-4201-3/20/$31.00 ©2020 IEEE

The ability to detect emotions in social interaction is often second nature; however, children with ASD struggle to pick up on these emotional cues. According to the National Institute of Mental Health, "*Autism Spectrum Disorder (ASD) is a development disorder that affects communication and behavior*" [17]. ASD is a developmental disorder that is usually detected at an early age, within the first couple of years after birth. Children with ASD can have a range of learning, communicating, and problem-solving challenges. Until now, there is no medical test that can be used to diagnose ASD. Doctors rely on a child's development and behavior to make a proper diagnosis. The reason why people with ASD have a hard time recognizing emotions in speech is still unclear [18]. Fig. 1 shows some of the challenges faced by children with Autism Spectrum Disorder.

In this research work, a speech emotion recognition solution is proposed, which would help children with ASD identify emotion in social interactions. It might serve as a tool to train children with special needs so that they can learn to recognize emotions when communicating with their family, friends, teachers, and caregivers.

Most of the utterances, speech recordings in audio databases contain very low noise since they are usually recorded in a soundproof studio environment. Therefore, it is not necessary to apply noise removal to these data in the model training phase. However, some researchers only work with the speech part of the recording and remove the non-speech part [20]. When implementing this work's model in a real-world scenario, the record will likely contain some Gaussian noise that could be removed for better data processing. Framing is essential for any audio processing. Any analysis done on audio frames is known as short-term analysis. A speech signal is a non-stationary signal - it is continually changing over time. Splitting the audio signal into small frames allows the signal to become "statistically stationary." Each audio frame is typically 20-30ms long, and adjacent frames are overlapping to ensure no loss of information. In this work, framing will be done by a library used in Python called libROSA, since this library was specifically designed to work with audio data and has most of the necessary tools. In practical applications, signal segments are finite and not periodic, and by applying FFT to a segment of a signal essentially wraps the end of the segment around to the start, generating a jump discontinuity. Such jumps result in undesirable background artifacts in the FFT amplitudes. To make the wrap-around smooth, one can fade the signal to zero at both segment ends by multiplying it with a window function [21]. Thus, windowing is performed after framing the signal by using the libROSA library.

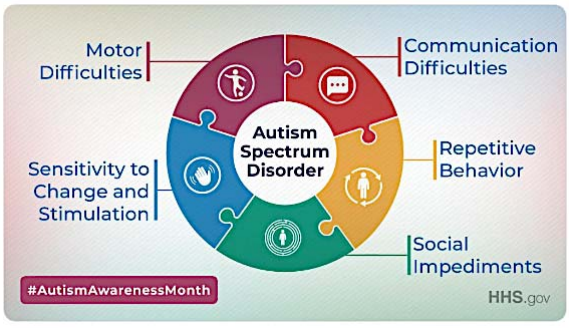

Fig. 1. Challenges faced by children with ASD [19].

The first step to training any machine learning model is to acquire features from data. Parameters or features of speech signals makes it possible for the machine to learn the patterns of different types of speech. This, in turn, allows the machine to predict the input speech class with a certain degree of accuracy. Thousands of acoustic features are used in speech research - energy, pitch, Mel-frequency Cepstral Coefficients (MFCC), formants, speech rate, jitters, shimmers, and many more [22]. Even though a large set of features guarantees capturing more acoustic characteristics from the speech data, it has some drawbacks as well. Firstly, with a vast amount of features, the interpretation of the results will become increasingly difficult. Furthermore, large feature sets cause the classifiers to over-adapt to the training data, decreasing the ability of the machine learning model to generalize, resulting in overfitting. For this work, only the MFCCs and the zero-crossing rate (ZCR) are used.

Cepstrum is a sequence of numbers that characterize a frame of speech. The MFCCs represent the short-term power spectrum of a sound, based on a discrete cosine transform (DCT) of a log power spectrum on the Mel-frequency scale. The expression in (1) shows how to calculate the Mel-spectrum of the m^{th} frame for $r = 1, 2, \ldots, R$.

$$MF_m[r] = \frac{1}{A_r}\sum_{k=L_r}^{U_r} |V_r[k]X_m[k]|^2 \quad (1)$$

Here, $V_r[k]$ is the weighing function for the r^{th} filter ranging from DFT index L_r to U_r, and A_r is a normalizing factor for the r^{th} Mel-filter. MFCCs compress the amount of information in a Fourier transform of a frame of speech to a small set of values. The presence of noise affects the performance of MFCCs as audio features.

When it comes to speech-emotion recognition, the support vector machine (SVM) has always proven to yield great results. It is therefore chosen as the learning algorithm for this research work. The SVM is considered as an extension of the perceptron. The perceptron minimizes the classification errors, whereas the SVM maximizes the margin. Fig. 2 shows how the SVM gives the optimal hyperplane with maximum margin. The close training samples are called support vectors. Decision boundaries with large margins tend to have a lower generalization error. Models with small margins are more prone to overfitting. For linearly inseparable data the SVM can be kernelized. The SVM algorithm has a few hyperparameters that can be tweaked to get the optimal performance for a given dataset. One such parameter is the C-parameter. It is the inverse of λ (the regularization parameter). If the test accuracy is much lower than the training accuracy (i.e. high variance), decreasing the value of the C-parameter would increase the regularization strength and most likely solve the overfitting issue. Another hyperparameter that is associated with some kernel SVM models is the γ (gamma). It is a cut-off parameter for the Gaussian sphere. The higher the value of γ, the tighter and bumpier the decision boundaries are [24].

Even though the SVM hyperparameters can be tuned and tested with the model manually, it can be a rather long and tedious process. Grid search is a technique by which the user can define a range of hyperparameter values and perform a parametric sweep in order to find the optimal hyperparameter values. One drawback of using grid search is that it requires more computational time to find all the possible combinations of hyperparameter values being tested by the machine. The test set is fed to the model and the classifier tries to predict the class label emotion for each input sample. A total of seven emotion classes are used for this study. Along with Ekman's

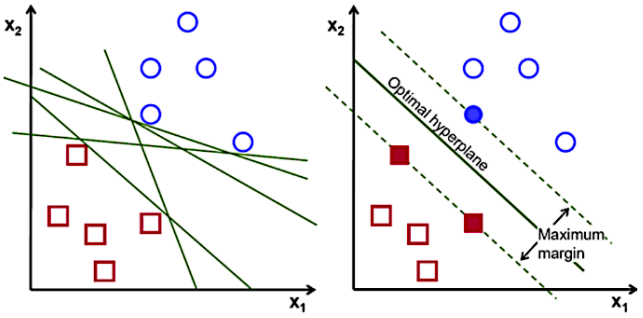

Fig. 2. The first plot shows several possible hyperplanes that can be used to classify the two classes. The second plot shows a single hyperplane created by the Support Vector Machine, that gives the maximum margin [23].

list of emotions of happiness, sadness, surprise, anger, fear, and disgust, an extra emotion class is used – neutral.

III. Feature Extraction and Classification

For this work, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is used. It is a multi-modal database containing audio and video recordings in North American English. The corpus was created by researchers of SMART Lab at Ryerson University. It is a public database under a Creative Commons Attribution license, thus being accessible to anyone. A total of 24 professional actors (12 males, 12 females) participated in the recording process. RAVDESS contains 7,356 files (24.8 GB of data). The audio-only recordings include 1,440 files - 60 trials per actor x 24 actors (215 MB of data). The audio was recorded at a bit depth of 16 bits, at a sampling rate of 48,000 Hz, in WAV format. Two separate sentences were recorded per actor – "*Kids are talking by the door*," and "*Dogs are sitting by the door*." Eight different emotions were captured in the speech recordings - happy, sad, neutral, surprise, calm, angry, fearful, and disgust [25]. Fig. 3 shows a flow diagram of the steps that are followed in this work.

1) Extracting features from the corpus: The first step in the Python code involves reading each utterance file from the corpus directory and extracting the Mel-frequency cepstral coefficients (MFCCs). The files are read serially, and the MFCCs are extracted for each frame. A sampling rate of 16,000 Hz is used for each WAV file. The default frame length of libROSA is used for this work. The MFCCs of each frame are stored in a Numpy array, and the mean of each of the thirteen MFCCs are calculated. This yields thirteen utterance-level MFCC features for each utterance file. This process is repeated for all 1,440 files of the corpus. Since the emotion label is specified in each file's name, it is easily extracted and stored in the same Numpy array. Once all the files have been processed, the features along with their corresponding labels, are stored in a Pandas dataframe.

2) Removing the "calm" class: Since the emotion "calm" is not being classified in this work, it is completely removed in the data processing phase. After removing the "calm" class, the number of utterances remaining is 1,248.

3) Upsampling the "neutral" class: There are eight unique emotion classes in the RAVDESS corpus. They are happy, sad, neutral, surprised, calm, angry, fearful and disgust. Each class has 192 utterances, except for the "neutral" class, which has only 96. This class imbalance is a huge problem in machine-learning since majority classes can influence the training and biasing of the model. Thus, the "neutral" class utterances are increased (upsampled) to match the other six classes.

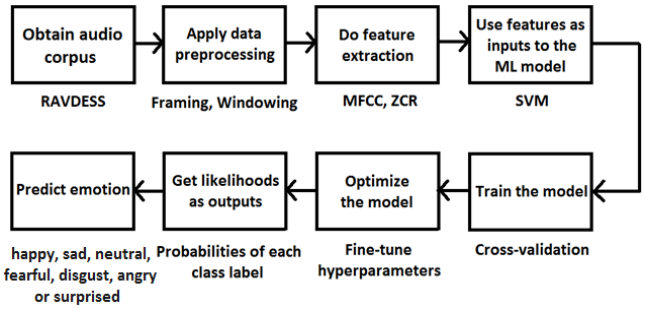

Fig. 3. Project flow diagram of the speech emotion recognition model.

4) Using MFCCs as features: In this work, the first thirteen MFCCs were initially selected as the acoustic features. This is because the higher coefficients represent fast changes in the filterbank energies, which end up degrading the model performance [26]. However, a thorough analysis provides a different result for this speech emotion classification model. Fig. 4 shows the classification accuracies for different number of MFCCs for the RAVDESS corpus data. The accuracies are obtained using default scikit-learn SVM settings (i.e. C = 1.0, kernel = 'rbf', gamma = 'scale'). The highest test accuracy (61 %) is achieved using 26 MFCCs. To verify this finding, the Principal Component Analysis (PCA) technique was used on 40 MFCCs. Fig 5 shows the contribution of each feature (MFCC). It is evident from Fig. 5 that using more MFCCs will not make a significant difference in the test accuracy since the first 26 princial components (MFCCs) contain majority of the information (i.e. variance). Furthermore, since the model will be used in real-time speech emotion recognition, using minimum number of features will reduce processing power and yield faster results.

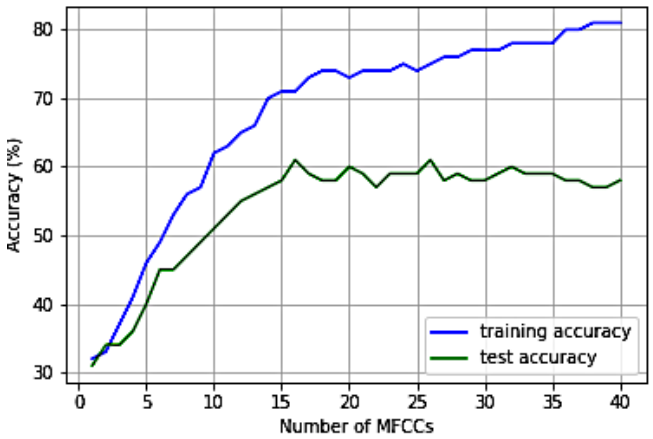

Fig. 4. Training and test accuracies for different number of MFCCs (Sampling rate = 16,000 Hz).

5) Using more features: The zero-crossing rate (ZCR) is defined as the rate of sign changes along a signal. ZCR is calculated using (2), where s is a signal of length T and $1_{R<0}$ is an indicator function.

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}_{<0}}(s_t s_{t-1}) \quad (2)$$

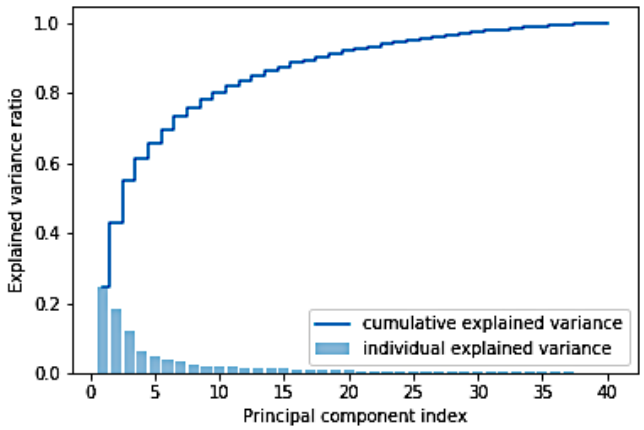

Fig. 5. Principal Component Analysis on MFCCs.

The ZCR is extracted for each frame and then an average is taken over all frames. This gives a total of 27 features per utterance. The result is a test accuracy of 79%. The standard deviations per frame are also used for the 26 MFCCs, causing the total number of features per utterance to increase to 53 – 26 MFCCs (mean), 26 MFCCs (standard deviation), and ZCR (mean). Even though the test accuracy drops to 77%, it is later seen that this additional feature set increases the test accuracy of RAVDESS utterances when noise is introduced to them.

6) Using grid search to find best hyperparameter values: Grid search is a technique that performs a parametric sweep to determine the best possible combination of hyperparameters within a specified range. Grid search is used multiple times in this work, with the parameter range being made narrower each time. A cross-validation split of ten was used for each hyperparameter combination. After running grid search for the SVM model, the combination that provides the highest score (74.3%) consists of the Radial Basis Function (RBF) kernel, with a C-parameter value of 4.88, and a γ of 0.0991.

7) Using different sampling frequencies: The sampling frequency is the number of samples of the audio signal collected per second when converting the analog signal to a digital signal. The higher the sampling frequency, the more the digital signal resembles the analog signal. TABLE I lists some sampling frequencies along with their resulting testing accuracies. These results are obtained using fourteen features – thirteen MFCCs (mean) and ZCR (mean).

TABLE I. TEST ACCURACIES FOR VARYING SAMPLING FREQUENCIES

Sampling frequency (Hz)	Test Accuracy
16,000	63.9 %
22,050	65.7 %
44,100	67.1 %
96,000	64.6 %

It should be noted that the initial analysis was done using a sampling rate of 16,000 Hz. Using this sampling rate, the highest test accuracy that can be achieved (using all 53 features) is 76%. However, the audio recordings in the RAVDESS corpus have originally been recorded and sampled at 48,000 Hz. After using this native sampling rate to process the utterances, the test accuracy increases to 77%. Since the audio recordings were previously being resampled at a sampling rate, which was lower than the native sampling rate, a large portion of the audio samples were not used for feature extraction that resulted in poorer learning performance and a lower classification accuracy. It can also be seen from TABLE II that sampling frequencies above 48,000 Hz produce poor classification accuracies. Therefore, it is better to use the native sampling frequency of recorded audio in data processing.

8) Using audio with background noise: The audio recordings in the RAVDESS corpus are noise-free since the recordings were done in an indoor environment. To properly evaluate the performance of this model, it has to be tested with audio recordings containing noise. For simulating an outdoor scenario, a sample audio track containing a recording of a city center is obtained from [27]. The record involves sounds of people talking, walking by (footsteps), and birds chirping. It is then mixed (combined) with the audio recordings from the RAVDESS corpus to produce audio recordings with background noise. Fig. 6 shows the waveforms of the three types of audio. The first plot is for a RAVDESS audio recording of a female voice actor (actor 22) saying the phrase "*Dogs are sitting by the door*" with strong surprise emotion. The second plot shows the recording of the city center recording. The third plot shows the mixed (combined) audio. The model performed quite well in noisy environment data.

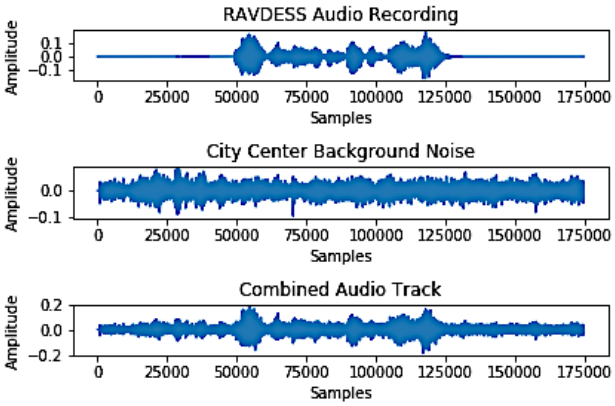

Fig. 6. Waveform of three different audio tracks – an utterance waveform (top), the background noise waveform (middle), and the combination of both (bottom).

9) Plotting the Learning Curves: Fig. 7 shows the learning curve for the best SVM model (kernel=RBF, C=4.88, γ=0.0991), using a sampling frequency of 48,000 Hz, and 53 features per audio file (26 MFCCs (mean), 26 MFCCs (standard deviation), and ZCR (mean)). Fig. 8 shows the learning curve for the same model but with background noise in the RAVDESS audio recordings. It is clear from the learning cuves that increasing the number of training samples increases the validation accuracy. However, the model is not complex enough to learn to distinguish each class with high accuracy, which causes a high variance between the training and validation accuracies. Using prosodic features such as pitch and intensity might increase the model performance. Also, the model is trained with only 1,344 utterances. Adding more data from other speech corpora is expected to fix the overfitting problem.

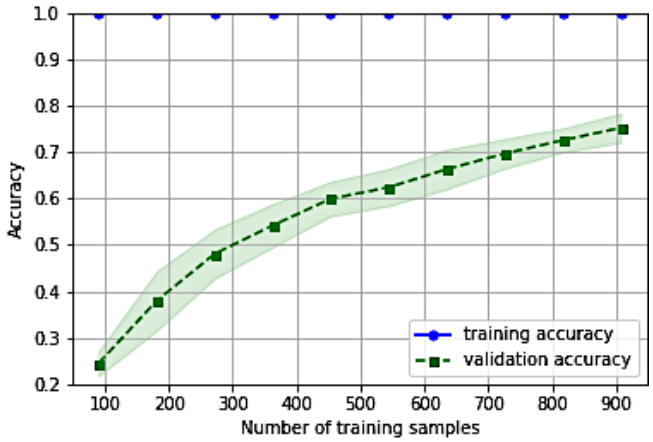

Fig. 7. Learning curve for the best SVM model (no background noise).

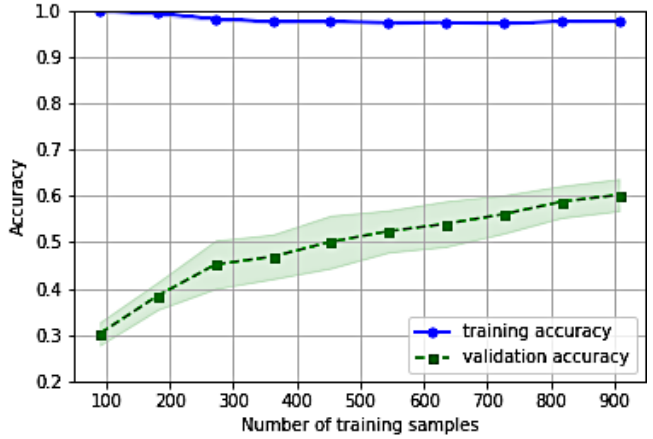

Fig. 8. Learning curve for the best SVM model (with background noise).

10) Using different learning algorithms: Besides SVM, some other popular supervised learning algorithms are also tested for this emotion classification model. TABLE II lists the grid search results for some of these machine learning classifiers. For all the classifiers, thirteen MFCCs (mean) and one ZCR (mean) per audio file are used as features. The SVM performs best out of the four classifiers.

TABLE II. TEST ACCURACIES FOR VARIOUS MACHINE LEARNING CLASSIFIERS USING GRID SEARCH

Classifier	Test Accuracy
SVM	67.1 %
Logistic Regression	42.6 %
Perceptron	32 %
Decision Tree	42.5 %

IV. PREDICTION RESULTS

TABLE III lists the training and test accuracies obtained for different models. The best classification accuracy obtained in this study is 77%. It is achieved using the RBF kernel of the SVM classifier, with C = 4.88 and γ = 0.0991. The 53 features used include the means per frame of the lower 26 MFCCs, along with their standard deviations per frame and the mean of ZCR. Additionally, the model is resampling the utterances at 48,000 Hz, which is the native sampling rate of the utterances belonging to the RAVDESS speech corpus. Upsampling is used to match the "neutral" class count to the other six classes. The confusion matrix of this model is given in Fig. 9. Authors in [28] used the Logistic Model Tree (LMT) classifier and reached a classification accuracy of 67.14% when using the RAVDESS dataset to train and test their model. Authors in [29] managed to get a classification accuracy of 64.48% using the GResNet classifier on RAVDESS.

The model is also trained and tested by modifying the utterances, where background noise is added to each of the corpus recordings. The model achieves 98% training accuracy and a 64% test accuracy. The confusion matrix of this model is shown in Fig. 10. In the original RAVDESS speech recordings, there are silent, non-speech frames in each file. Once the noise is added to the recordings, there are no more silent frame regions. Since the acoustic features are extracted for each audio frame, the frames containing the background noise also contribute to the final feature values. This is the reason the training and test accuracy of the model decreases when noisy data is introduced for classification.

TABLE III. TRAINING AND TEST ACCURACIES FOR DIFFERENT PARAMETERS OF THE SVM MODEL

List of features	Sampling frequency	Utterances	Training accuracy	Test accuracy
13 MFCCs (mean)	16,000 Hz	RAVDESS	96 %	65 %
26 MFCCs (mean) and ZCR (mean)	16,000 Hz	RAVDESS	100 %	74 %
26 MFCCs (mean), 26 MFCCs (standard deviation), and ZCR (mean)	16,000 Hz	RAVDESS	100 %	76 %
26 MFCCs (mean), 26 MFCCs (standard deviation), and ZCR (mean)	48,000 Hz (native)	RAVDESS	100 %	77 %
26 MFCCs (mean), 26 MFCCs (standard deviation), and ZCR (mean)	48,000 Hz (native)	RAVDESS (with noise)	98 %	64 %

V. CONCLUSION

Children with ASD suffer from poor communication skills. In this research, a speech emotion recognition solution has been proposed that would help children with ASD differentiate between different human emotions during conversations. The emotion classifier developed yields decent classification accuracy for audio with background noise. One major issue with this model is the problem of high variance (overfitting). It could be due to the lack of audio recordings used to train the model. Increasing the regularization strength also fixes overfitting problems, but since various values of C have already been tested using grid search, the regularization strength might have already been optimized for this model. Multiple speech emotion corpus could be used to train this model to overcome the overfitting issue.

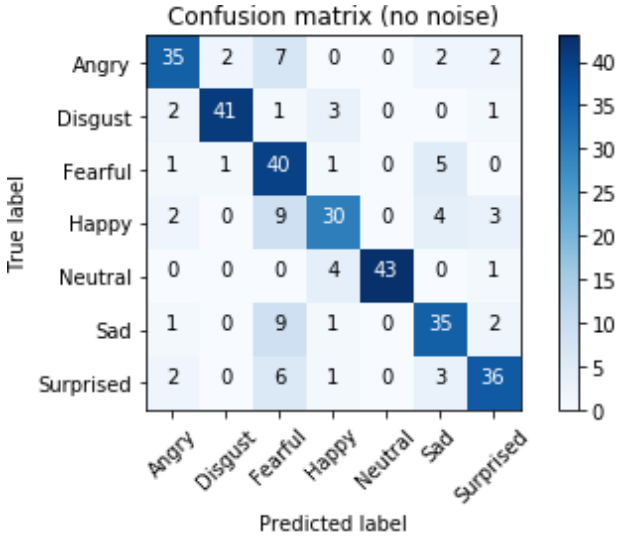

Fig. 9. Confusion matrix for the best SVM model with no background noise.

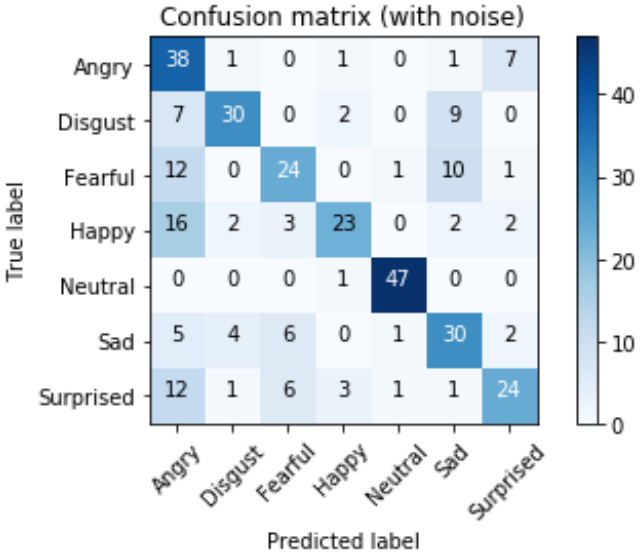

Fig. 10. Confusion matrix for the best SVM model with background noise.

REFERENCES

[1] S. Sadhu, R. Li, and H. Hermansky, "M-vectors: sub-band based energy modulation features for multi-stream automatic speech recognition," *Int. Conf. on Acoustics, Speech and Signal Processing*, May 2019, pp. 6545- 6549.

[2] W. Liu et al., "State-time-alignment phone clustering based language-independent phone recognizer front-end for phonotactic language recognition," *14th Int. Conf. on Computer Science & Education*, August 2019, pp. 863-867.

[3] G. Shanmugasundaram, S. Yazhini, E. Hemapratha, and S. Nithya, "A comprehensive review on stress detection techniques," *Int. Conf. on System, Computation, Automation and Networking*, Mar. 2019.

[4] C. Xue, "A novel english speech recognition approach based on hidden Markov model," *Int. Conf. on Virtual Reality and Intelligent Systems*, Aug. 2018.

[5] S. Jendoubi, S. B. Yaghlane, and A. Martin, "Belief hidden Markov model for speech recognition," *Int. Conf. on Modeling, Simulation and Applied Optimization*, Apr. 2013.

[6] W. Dan, "An audio classification approach based on machine learning," *Int. Conf. on Intelligent Transportation, Big Data & Smart City*, Jan. 2019, pp. 626-629.

[7] F. Rong, "Audio classification method based on machine learning," *Int. Conf. on Intelligent Transportation, Big Data & Smart City*, Dec. 2016, pp. 81-84.

[8] T. Pellegrini, "Densely connected CNNs for bird audio detection," *25th European Signal Processing Conf.*, Aug. 2017, pp. 1734-1738.

[9] A. Lieto et al., "'Hello? Who Am I Talking To?' A Shallow CNN Approach for Human vs. Bot Speech Classification," *Int. Conf. on Acoustics, Speech and Signal Processing*, May 2019, pp. 2577-2581.

[10] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," *9th European Conf. on Speech Communication and Technology*, France, 2005, pp. 1841-1844.

[11] A. K. Oryina and A. O. Adedolapo, "Emotion recognition for user centred e-learning," *40th Annu. Int. Computer Software and Applications Conf.*, in vol. 2, Jun. 2016, pp. 509-514.

[12] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," *Int. Conf. on Artificial Intelligence and Computational Intelligence*, in vol. 1, Oct. 2010, pp. 537-541.

[13] Nico H. Frijda, *The Emotions*, Cambridge, England, UK: CUP, 1986.

[14] J. S. Lerner, Y. Li, P. Valdesolo, and K. Kassam, "Emotion and decision making," in *Annu. Rev. Psychol.,* in vol. 66, Jan. 2015, pp. 799-823.

[15] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," in *J. Pers. Soc. Psychol.*, in vol. 17, no. 2, 1971, pp. 124–129.

[16] A. Mehrabian, *Silent Messages*, Belmont, CA, USA: Wadsworth Pub. Co, 1971.

[17] nimh.nih.gov https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml. (Accessed: Dec. 8, 2019).

[18] S. Schelinski and K. V. Kriegstein, "The relation between vocal pitch and vocal emotion recognition abilities in people with autism spectrum disorder and typical development," in *Journal of Autism and Developmental Disorders*, in vol. 49, iss. 1, 2018, pp. 68-82.

[19] Twitter.com. https://twitter.com/hhsgov/status/1122553066077736960. (Accessed: Nov. 9, 2019).

[20] G. Trigeogis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," *Int. Conf. on Acoustics, Speech and Signal Processing*, Mar. 2016, pp. 5200-5204.

[21] StackExchange.com. https://dsp.stackexchange.com/questions/14067/what-is-a-window-function-in-dsp-and-why-do-we-need-it. (Accessed: Nov. 9, 2019).

[22] A. A. Bashit, "A comprehensive solar powered remote monitoring and identification of Houston toad call automatic recognizing device system design," M.S. Thesis, Eng., Texas State Univ., San Marcos, TX, USA, 2019.

[23] TowardsDataScience.com. https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c. (Accessed: Nov. 9, 2019).

[24] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, 2nd edition, Birmingham, UK: Packt Pub., 2017, pp. 192.

[25] 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," in PLoS ONE, in vol. 13, iss, 5, May 2018, pp. 1-35.

[26] practicalcryptography.com. http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/. (Accessed: Dec. 8, 2019).

[27] soundbible.com. http://soundbible.com/tags-city.html. (Accessed: Dec. 8, 2019).

[28] A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames", *Proc. Int. Conf. Robot. Elect. Signal Process. Techn. (ICREST)*, pp. 281-285, Jan. 2019.

[29] Y. Zeng, H. Mao, D. Peng and Z. Yi, "Spectrogram based multi-task audio classification", *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3705-3722, Feb. 2019.