CLASSIFICATION OF DEPRESSION FROM SPEECH: A TRANSFER LEARNING APPROACH

Joshua Trower[1]

[1] Ontario Tech University, Oshawa, Ontario, Canada

joshua.trower@ontariotechu.net

ABSTRACT

Depression is the most common mental health disorder in the developed world, where 1 in 5 adults will be diagnosed with mood disorder within their lifetimes. This places a growing burden on the already stressed healthcare system. The symptoms of depression affect an individual's life but they also manifest physical tells, notably in speech. Machine learning may help to alleviate this problem by offering an automated pre-assessment.This study investigates the efficacy of a model pre-trained on emotional speech recognition data compared to a model solely trained on the limited amount of depression speech data. The novelty of this research lies within the application of transfer learning to link the fields of Speech Emotion Recognition and Depression Speech Recognition.

INTRODUCTION

At present, depression is the most common mental disorder in many developed nations. Depression has been proven to reduce quality of life and life expectancy. The United States national library of medicine estimates that ⅔ of all cases of depression go undiagnosed [1], all while being one of the most treatable mental health disorders. With all of this in mind, depression is currently diagnosed by a clinician administering questionnaires over the course of multiple weeks. If the patient exhibits symptoms of depression nearly everyday for those multiple weeks they can be diagnosed with depression and treatment can begin.

While depression is a mental disorder and has psychological symptoms, depression also manifests in physical symptoms such as dramatic weight gain or loss. As well as a change in psychomotor activities most notably; speech. Individuals with depression often have a more submissive quality to their voice causing that individual to have a slower speed and softer inflection. As well for individuals with depression, speech is more laboured meaning a slower rhythm, and speed. As proposed by Silverman in the 2000 study[2], the more depressed a person is the more they exhibit these vocal phenomena. Which means that by measuring these features of speech you could measure someones suicide risk. All of these aspects can be measured and through machine learning we can predict whether or not someone has depression solely through hearing them speak.

As mentioned, depression very often goes undiagnosed. From this the field of depression speech recognition(DSR) was created. The classification of depression is a field intrinsically connected to the field of emotion recognition. There are many databases that have been created to train a machine learning model to classify the emotions of people. It has been proposed that the human voice also conveys emotional states[3], multiple studies have been done building classifiers that can read the emotion from a person's speech.

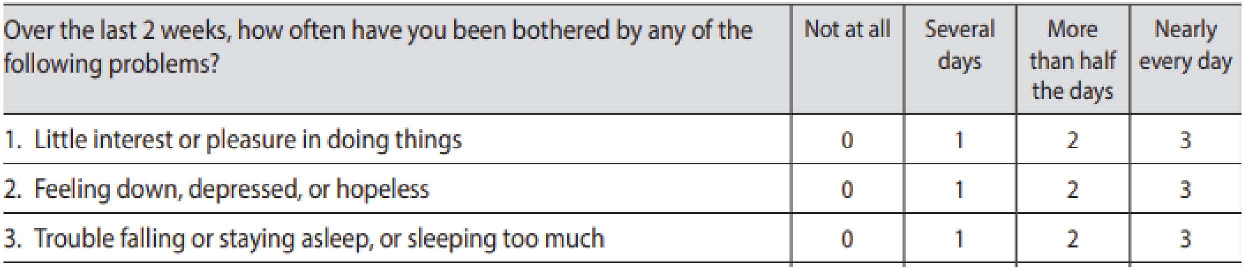
The purpose of this study is to determine whether a model pre-trained on basic emotion speech data would be more effective than a classifier trained entirely on the depressed speech data. This question is important because there is a limited amount of depressed speech data, with the only common dataset being the Distress Analysis Interview Corpus/Wizard-of-Oz (DAIC-WOZ) dataset. The reason for such a small amount of data within this field is due to the privacy concerns around propagating medical details of individuals. The difference within this study and the related works mentioned below is the use of transfer learning to assess whether or not a pre-trained model trained on emotion recognition audio is a more effective classifier than a model trained entirely on the depressed speech data.

BACKGROUND

The field of depression classification from speech is a growing field as more and more people are diagnosed with depression[4]. Due to the time commitment required for testing whether or not someone has depression being so extensive the field of classifying depression from speech has gained more attention.

This study is centered around depressed speech data. For this purpose, the DAIC-WOZ is currently the largest source of depression speech data, although it only contains 189 clips. This data is an extension of the DAIC dataset, which used speech files to aid in PTSD diagnosis. This is why the DAIC-WOZ files began labeling at patient 300. A single user folder from this dataset contains the audio file of the interview in mp4 format and a transcript of the conversation with speaker lines in a csv format.

The DAIC-WOZ dataset itself has metadata which details a speaker ID(300-489), a gender binary, a “is depressed” binary, and a PHQ-9 score for each interviewee. The Patient Health Questionnaire appendix 9 (PHQ-9) is a nine question depression diagnosis quiz, where each question is graded from "not at all” to “nearly everyday”, 0-3, and has a maximum overall score of 27. The DAIC-WOZ dataset marks someone as "depressed" if they score a 10 or higher. This questionnaire, however, is not used as a sole determining factor of depression, but rather used alongside a clinician's assessment. As reported within [5], the phq-9 questionnaire has an 88% sensitivity rate when determining if a client is (or is not) majorly depressed. The DAIC-WOZ being the only notable dataset is evidence enough that there is simply not enough data when it comes to building a machine learning model for detecting depression.



Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3

Figure 1. Portion of PHQ-9 Questionnaire

The way depression affects speech overlaps acoustically with emotional speech signals that denote sadness. This points to the broader field of Speech Emotion Recognition(SER) being able to support depression classification.The field of SER has emerged as a key component in the field of Human-Computer Interaction(HCI) as it allows a computer to understand more nuance of speech than the words alone. An SER system needs a classifier that has been trained on many data points for each possible classifiable emotion. However, as mentioned by Bo-Hao et al.[6] , each emotion speech database has a different group of speakers that may have similar dialects, given their current location or background. This makes it harder to train a model to classify when given a voice/accent it has never heard before.

The first half of this study is training a model on entirely standard speech emotion data. For this study I chose two large datasets to make up the data for the first model, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Emotion Speech Database (ESD).

I chose to include two datasets rather than one because all emotion databases are not uniform and they each have their own unique list of classifiable emotions[6]. However, for the purpose of this project, all classifiable elements can be manipulated into a binary, which will be elaborated upon below. The benefits of using two datasets is that there is not only more data, but utilizing a more diverse dataset of speakers with their own dialects will create a stronger model, with a greater chance of being able to classify speakers even when the model has never been presented with them before.

The RAVDESS[7] is one of the largest English speech emotion datasets, containing 7360 clips from 24 actors (12 male and 12 female). This means that the model will be trained on male and female voices, including a variety of different pitches and tones, resulting in a stronger classifier when the model is transferred. The RAVDESS has 8 possible classifiable targets, with each clip conveying a certain emotion, including neutral, calm, happy, sad, angry, fearful, surprise, or disgust.

The ESD[8] contains 350 identical spoken phrases by 10 native English speakers and 10 native Chinese speakers. The reason why the ESD was chosen was because it has different dialects within the speakers. The ESD has 5 classification targets, with each phrase being spoken with each target emotion including neutral, happy, angry, sad, and surprised. This dataset was created to address a gap in the field of speech emotion recognition, that gap being a lack of multicultural speakers.

The OpenSmile API was chosen as the preprocessing method to change the data from raw audio files to a csv containing 88 vocal features from each audio recording. OpenSmile is a library created at the Technical University of Munich within the SEMAINE research project, the goal of which was to create a virtual agent with emotional and social intelligence. The technology of Opensmile was extended in 2013 branching out of the SEMAINE project and becoming a toolkit for anyone looking to extract hyper-detailed and common features found in every person's speech. Opensmile is the most common toolkit due to the mass amount of data it scans for. Opensmile calculates low-level features of audio by collecting the average data from frame by frame analyses. This method used in conjunction with different mathematical functions allows Opensmile to collect metrics such as pitch, Mel-frequency cepstrum coefficient data, etc.

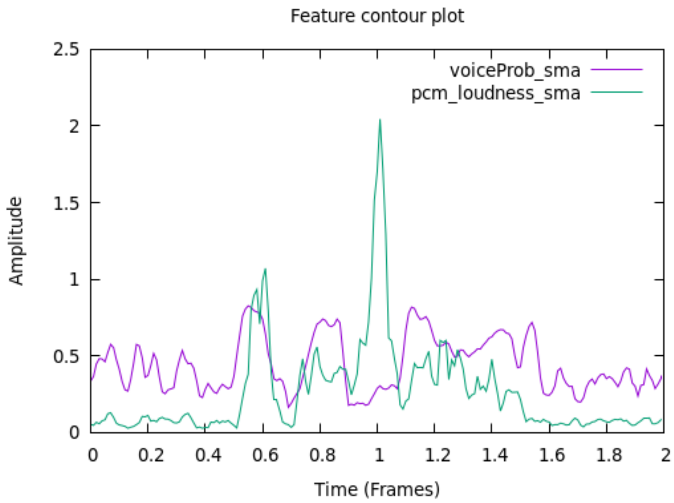


Figure 2. Opensmile collecting amplitude data(Sourced from Opensmile)

As mentioned earlier there is a very limited selection of data when it comes to depression classification, and there is an abundance of data for speech emotion classification. I made the decision to attempt to apply transfer learning to leverage the large amount of speech emotion data to support a classifier built for depression classification. Transfer learning is the technique of taking the convolutional layers of a model after they have been trained on a task, and plugging them into a new model with new input data and using them in conjunction with a new set of dense layers.

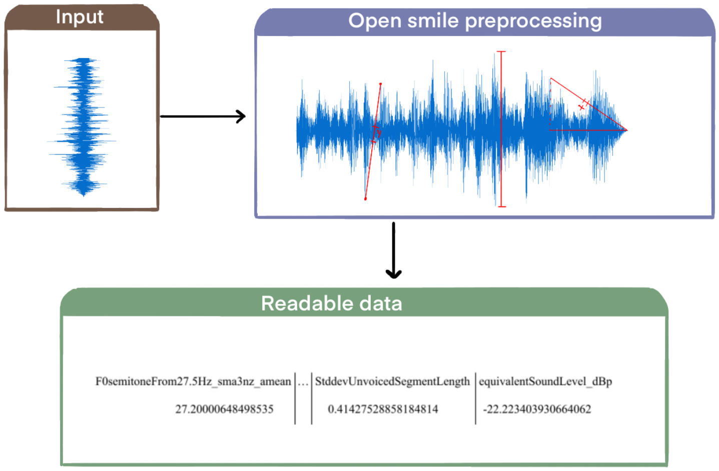
METHODOLOGY

A.DATA MANIPULATION AND OPENSMILE

The target data within the RAVDESS and the ESD databases are designed for a multi-classifier that will determine a speaker's emotion from 8 possible emotions. As the goal of this study is to create a binary classifier, the target data of both datasets was altered to be a binary of either "sad" speech or "not sad" speech. The reason the target data was changed is because the ESD database only has 5 possible targets while the RAVDESS has 8 so in order to combine them and keep the integrity of the data a common target was chosen. As for the reason that the target data became specifically “sad” and “not sad” rather than “anger” and “not anger” is because the acoustic properties of a “sad” voice and the acoustic properties caused by symptoms of depression are very similar. As described in the Analysis of Speech Affected by Depression[9], when a person with depression speaks their range of pitch and volume drop, causing them to sound flatter and softer.

The conversion from multi-classifier to binary classifier has an added benefit of making the Emotion Classifier Model more accurate, as using 2 classification targets rather than 6-8 classification targets decreases the difficulty of the classifying task by ⅛.

The data from the 3 datasets was in .mp4 and .wav file formats which while possible to feed directly into a model, it would remove the ability to look into the links between speech and depression and what about speech truly changes between depressed and non-depressed speakers. As well, “raw data” can be difficult to interpret and often requires much more data to achieve desired results, while preprocessed data is easier to interpret and identify patterns. With this in mind the decision was made to use preprocessed data.



F0semitoneFrom27.5Hz_sma3nz_amean	…	StddevUnvoicedSegmentLength	equivalentSoundLevel_dBp
27.20000648498535		0.41427528858184814	-22.223403930664062

Figure 3. Sample of OpenSmile Data

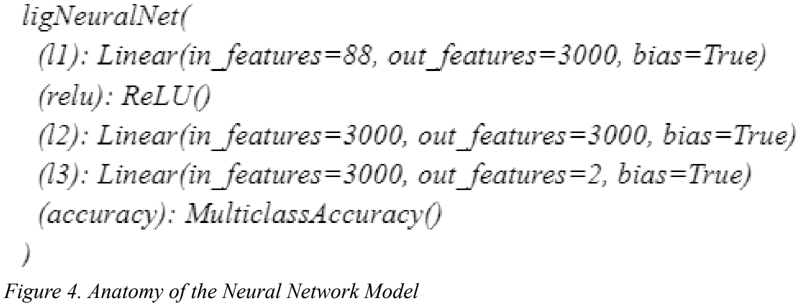
As detailed in Figure 3, the Opensmile API takes an audiofile and returns an array of 88 features that detail a miarade of qualities within the voice it analyses. These features range from the speed at which the voice stops, to how long it takes for the voice to reach its maximum pitch. The data manipulation required for the DAIC-WOZ dataset was on a larger scale. The DAIC-WOZ is a Wizard of Oz style interview meaning the audio file of the interview contains 2 voices: the interviewer and the interviewee. The portions of the audio file where the interviewer is talking are functionally useless for our purposes. As such using the transcript as a guide the large audio file was cut into dozens of different pieces where only the interviewee is speaking. The process of this is broken down as for each participant, scanning the transcript and noting the start and end point whenever the participant is talking and when the interviewer starts talking. With that data and the Pydub library cutting out the sections of the total interview recording where the participant is talking. One of the hurdles was that the transcript restates the speaker whenever there is a pause meaning that often the start and end times would be milliseconds apart because the interviewee took a breath, to rectify that I only noted the end point if the next speaker was not the current speaker. The result of this process was a subfolder for each participant containing 30-80 short audio clips, the next step was running each file through Opensmile and adding that data to a large csv file. In addition each row of Opensmile data had an additional column containing the user that speech belonged to. After completing that process for each participant the final step of pre production was to add target data to the completed csv, this was done by using the metadata of the DAIC-WOZ that contains whether or not each participant is depressed by a binary. If the participant is depressed the target data is 1, while a participant who is not depressed will be marked by a 0.

B.MODEL CHOSEN, LIGHTNING AND PARAMETERS

The first model chosen was a neural network model. The structure of a neural network is based on the organic structure of the brain. All layers within this model are dense layers. A dense layer is one where every neuron is connected to every neuron from the previous layer. Each of those connections is a weight from 0 to 1, while this does increase the workload, and complexity of the model. It also allows the model to capture more complex patterns.[11]

The code for the model was written using python applying Pytorch Lightning which is a wrapper for Pytorch. Pytorch is an open-source framework for building machine learning models. Pytorch Lightning extends the utilities of base Pytorch simplifying the model building process and allowing for hot swapping models.

When a Lightning model is stored it contains each layer within that model.



```
ligNeuralNet(
  (l1): Linear(in_features=88, out_features=3000, bias=True)
  (relu): ReLU()
  (l2): Linear(in_features=3000, out_features=3000, bias=True)
  (l3): Linear(in_features=3000, out_features=2, bias=True)
  (accuracy): MulticlassAccuracy()
)
```

Figure 4. Anatomy of the Neural Network Model

Figure 4 details the anatomy of the neural network model used in the DAIC-WOZ only classifier. It has a primary layer which creates weights between the inputs and a hidden layer of 3000 neurons. The second layer is a hidden layer that creates weights between 2 hidden layers. The final layer creates weights between 3000 neurons and the 2 output conditions, which in our case is depressed or not depressed. Each layer is classified as a *Linear Layer* which means it takes an input tensor and performs a linear operation utilizing the weights and biases that are constantly evolving as the model is trained, the layer returns an output tensor. The output tensor for this model would be *[x, y]* where *x* is the confidence the model has that the input data comes from a depressed person and *y* is the confidence the model has that the input data comes from a non-depressed person. The ReLU layers are ones that apply the ReLU function to the values given from the previous layer. The purpose behind the ReLU function is to remove any negative weights and instead any weight below 0 is treated as 0 and by doing that killing neurons that have no value.

C. TRANSFER LEARNING

Transfer learning is a machine learning technique that uses a model pre-trained for one task, and applying it to another related task. Here the task that the primary model is trained to do is classify emotions. Specifically the binary classification between “sad” and “not sad”. The reason for attempting transfer learning on this task is due to a lack of data within this field. However I theorize that the task of emotion recognition through speech which has a colossal amount of data and resources is similar enough that a pre-trained model created to classify emotions could be used to classify depression from speech.

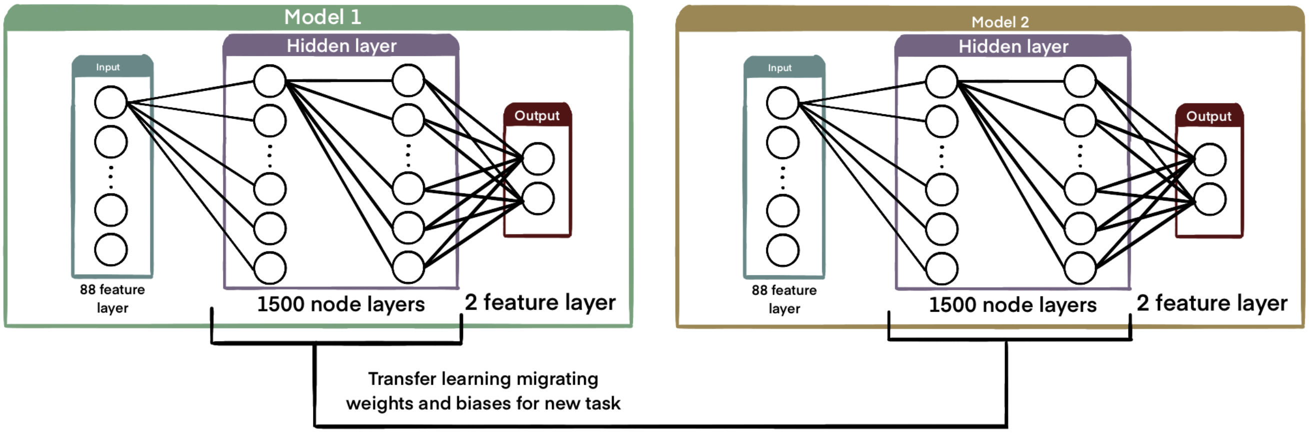


Figure 5.Transfer learning breakdown

As shown in figure 5, the way transfer learning is implemented in pytorch lightning is that “model 1” is stored as an object, and all of its weights and biases it developed while being trained on the primary data are carried over to “model 2”. These weights and biases are used as a new training start point for the secondary model.

RESULTS:

Here, we study the aforementioned method experimentally and compare the effect transfer learning had on the model.

A. Speech Emotion Recognition:

The classifier achieved results well above chance. With consideration that this speech emotion classifier only classifies a binary of “sad” or “not-sad” data which makes the process of finding patterns far simpler. Regardless, within figure 6 we can see that the ESD/RAVDESS database creates a strong base for the speech emotion recognition model. As well, the parameters used for the SER model create a classifier capable of performing well above expectations. The parameters for this model are: a hidden layer size of 1500, a batch size of 128, as well as a learning rate of

0.001. Finally the model has yet to reach a plateau so given more than 30 epochs the model will reach an even greater accuracy.

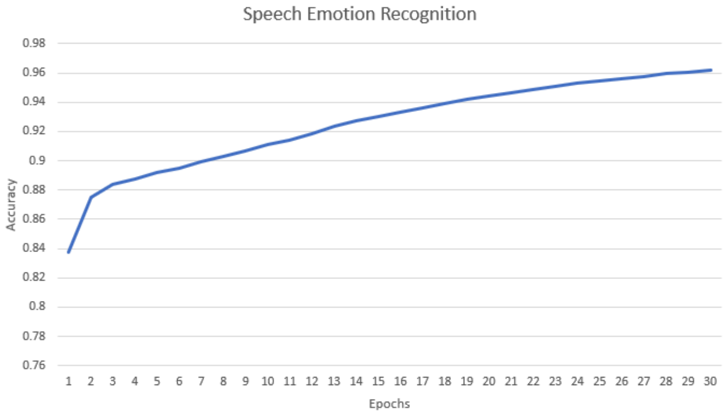


Figure 6. Accuracy Graph of Speech Emotion recognition

B. Depression Speech Recognition

The DAIC-WOZ database has a total of 10632 data points of varying quality. Some are Opensmile data from clean voice clips, while others are Opensmile data from a patient taking a breath. Figure 7 shows the difference between transfer learning and solely using the available depression classification data. As shown the DAIC-WOZ classifier plateaus at 90% accuracy. While the Transfer learning model is far more volatile bouncing from 90% to an amazing high of 96% accuracy. This data proves that the use of transfer learning can create a classifier that is marginally better than the sole depression data classifier, at the 100th epoch the transfer learning model is trending up while the DAIC-WOZ data remains plateaued. This implies that given more training time the transfer learning model could reach a far greater accuracy than the DAIC-WOZ classifier alone.

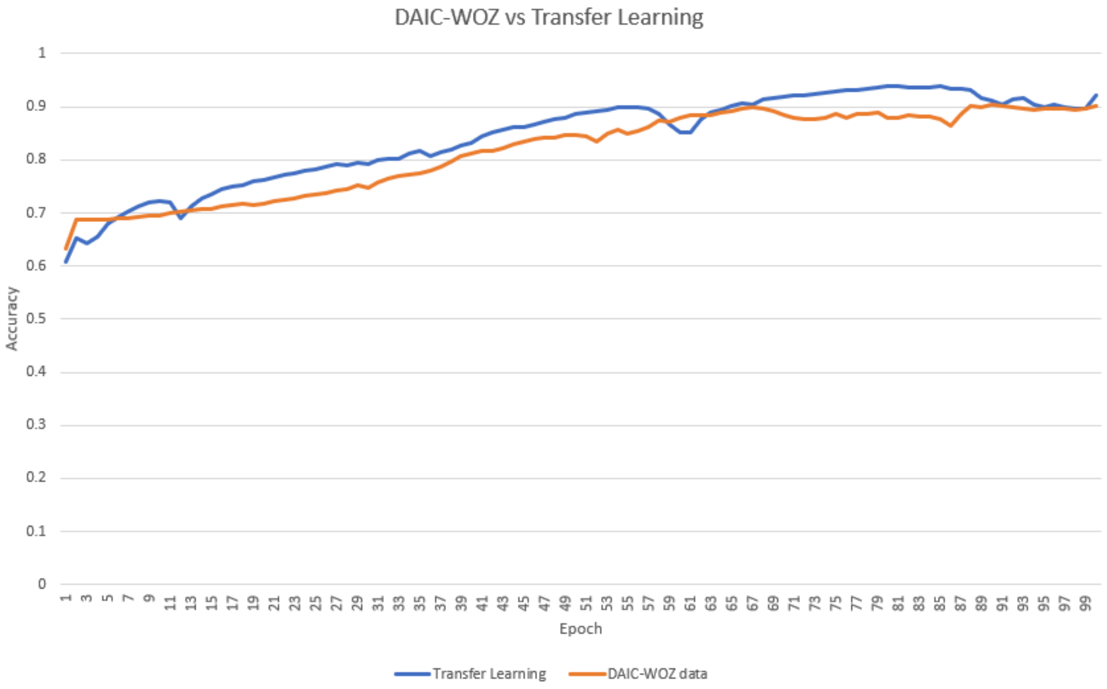


Figure 7. Accuracy Graph of Depression Speech Data

DISCUSSION:

What these results prove is that transfer learning does have an effect on the accuracy of the data although marginal, within the medical field that small amount of additional accuracy means a lot. While more depression speech data would result in a better classifier regardless of the implementation that data is not easily accessible or creatable. That is why this research is important if it means that a model can be pre-trained on SER data to reduce the amount of Depression speech data that it needs to reach an accuracy that is well above chance.

The potential next steps of this project are expanding on the DSR model adding additional layers and analysing the results with more epochs. Additionally while the implementation of transfer learning did yield a greater result it would be a valuable metric to analyze the weights and biases of the model for each feature of the data to get more concrete data on how depression affects speech. Adding additional speakers with different dialects from different countries would be a fantastic next step in creating a usable piece of technology, since as of now the model has very little data with many different accents and there is no way to know how the model would react if it was asked to classify an accent it has never been exposed to before. This technology may prove useful in speeding healthcare access by providing an automated pre-screening for healthcare professionals.

CONCLUSION:

In this work, we proposed a novel implementation of SER as a base that could elevate the field of Depression Speech Recognition to a higher level through the use of transfer learning. Taking a neural network trained on speech emotion recognition databases such as the ESD and RAVDESS, and using that model as a base to perform transfer learning onto a model trained with depression speech data. While this does not remove the need for depression speech data it reduces the amount needed to obtain a medically viable model to use alongside clinical diagnosis. This is quite important as depression speech data is quite rare and difficult to create due to the ethical considerations that come with obtaining that data.

REFERENCES:

[1] Williams SZ, Chung GS, Muennig PA. Undiagnosed depression: A community diagnosis. SSM Popul Health. 2017 July 28;3:633-638. doi: 10.1016/j.ssmph.2017.07.012. PMID: 29349251; PMCID: PMC5769115.

[2]Shiavi, R., Silverman, S. E., Silverman, M. K., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering (Print)*, *47*(7), 829–837. https://doi.org/10.1109/10.846676

[3]Kamiloglu, R. G., Fischer, A. H., & Sauter, D. (2020). Good vibrations: A review of vocal expressions of positive emotions. *Psychonomic Bulletin & Review*, *27*(2), 237–265. https://doi.org/10.3758/s13423-019-01701-x

[4]Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., & Sun, M. (2022). Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*, *8*(3), 701–711. https://doi.org/10.1049/cit2.12113

[5]Kroenke, K., Spitzer, R.L. and Williams, J.B.W. (2001), The PHQ-9. Journal of General Internal Medicine, 16: 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[6]B. -H. Su and C. -C. Lee, "Unsupervised Cross-Corpus Speech Emotion Recognition Using a Multi-Source Cycle-GAN," in IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 1991-2004, 1 July-Sept. 2023, doi: 10.1109/TAFFC.2022.3146325.

[7]Livingstone, S. R., & Russo, F. (2018). 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*, *13*(5), e0196391. https://doi.org/10.1371/journal.pone.0196391
[8]Zhou, Kun, et al. “Emotional Voice Conversion: Theory, Databases and ESD.” *Speech Communication*, vol. 137, Feb. 2022, pp. 1–18, https://doi.org/10.1016/j.specom.2021.11.006.
[9]Cummins, Nicholas, et al. “Analysis of Acoustic Space Variability in Speech Affected by Depression.” *Speech Communication*, vol. 75, Dec. 2015, pp. 27–49, https://doi.org/10.1016/j.specom.2015.09.003. Accessed 9 June 2021.
[10]Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56–76. https://doi.org/10.1016/j.specom.2019.12.001
[11]De Lope, J., & Graña, M. (2023). An ongoing review of speech emotion recognition. *Neurocomputing (Amsterdam)*, *528*, 1–11. https://doi.org/10.1016/j.neucom.2023.01.002
[12]:S. Harati, A. Crowell, H. Mayberg and S. Nemati, "Depression Severity Classification from Speech Emotion," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 2018, pp. 5763-5766, doi: 10.1109/EMBC.2018.8513610.
[13]Chen, J., Guo, X., Guo, Y., Zhang, J., Zhang, M., Yao, Q., & Yao, J. (2021). Deep Neural Network-based approach for breakdown voltage and specific on-resistance prediction of SOI LDMOS with field plate. *Japanese Journal of Applied Physics*, *60*(7), 077002. https://doi.org/10.35848/1347-4065/ac06da
[14]Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, *87*(1), 93–98. https://doi.org/10.1016/j.biopsycho.2011.02.010
[15]Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256. https://doi.org/10.1016/s0167-6393(02)00084-5
[16]Chen, J., Wang, C., Wang, K., Yin, C., Zhao, C., Xu, T., Zhang, X., Huang, Z., Liu, M., & Yang, T. (2021). HEU Emotion: a large-scale database for multimodal emotion recognition in the wild. *Neural Computing & Applications*, *33*(14), 8669–8685. https://doi.org/10.1007/s00521-020-05616-w
[17]Cowen, A., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(38). https://doi.org/10.1073/pnas.1702247114
[18]Barrett, L. F. (2016). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, nsw154. https://doi.org/10.1093/scan/nsw154
[19]Eyben, F., Scherer, K. R., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GEMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202. https://doi.org/10.1109/taffc.2015.2457417
[20]Douglass, M. (2020). Book Review: Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, 2nd edition by Aurélien Géron. *Physical and Engineering Sciences in Medicine/Physical and Engineering Sciences in Medicine*, *43*(3), 1135–1136. https://doi.org/10.1007/s13246-020-00913-z