The Impact of Face Mask and Emotion on Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER)

Qi Qi Oh
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
ohqiiqii@gmail.com

Chee Kiat Seow
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
Cheekiat.seow@glasgow.ac.uk

Mulliana Yusuff
Learning Vision Team
Panasonic R&D Center Singapore
Singapore, Singapore
mullianayusuff.jamal@sg.panasonic.com

Sugiri Pranata
Learning Vision Team
Panasonic R&D Center Singapore
Singapore, Singapore
sugiri.pranata@sg.panasonic.com

Qi Cao
School of Computing Science
University of Glasgow
Glasgow, United Kingdom
Qi.cao@glasgow.ac.uk

Abstract— **Meetings are common in the workplace, and it's critical to understand employees' feelings during a meeting since emotions influence how individuals talk and behave. Furthermore, due to Covid-19, individuals are required to wear a face mask in meetings, and there is not much research on how wearing a face mask impact the underlying emotion through speech. Although past research has shown that emotions and face masks have some form of impact on ASR, respectively, as of the time of writing, there are no findings on which attribute (emotion or mask) impacts the ASR system more and contributes more significantly to the speech processing error. Experiments were conducted, which aligned with the previous research that the face masks have a relatively small impact on ASR System performance at the sentence level. In addition, emotion affects speech significantly, especially sadness. Therefore, emotion contributes more significantly to speech processing error than a face mask. The experiments proved that the augmentation technique is useful as the F1 Score improved, with the greatest improvement being a 10% increase. The experiment also proved that using a face mask does not have much impact on emotion, making SER a feasible solution for detecting the underlying emotion through speech.**

Keywords—speech emotion, asr, ser, nlp

I. INTRODUCTION

Meetings are an everyday activity in a workplace where people come together to discuss and make decisions. Furthermore, it is necessary to have meeting transcriptions as it helps to serve as a record for future references. However, it can be very tedious to jot down all the conversions in a meeting [1]. Therefore, audio transcription tools are commonly adopted to transcribe the speech to text to document the discussions in text form [1] [2]. It is also important to understand employees' feelings in a meeting as emotions play a crucial role in a successful meeting [3]. Moreover, emotions affect how people speak and affect the overall pitch and intensity [4]. Furthermore, emotions also affect how people make decisions, making it even more essential for everyone in the meeting to be aware of each other emotional state as it may affect the decision-making process [5]. Hence, techniques to understand the impact of emotion on facial expression, speech and language processing have been researched over the past few years [6].

II. BACKGROUND

Research has shown that emotion is a significant contributor to Automatic Speech Recognition (ASR) mistakes, with neutral speech being recognised far better than emotional speech [7]. Furthermore, emotional speech can significantly reduce speech recognition accuracy by 5% to 7% [8]. As a result, emotion should be seen as a noise that obstructs understanding of what the user says, degrading the entire engagement with conversational technology [9]. However, as most research is performed on public emotion datasets recorded in a controlled environment such as a recording studio, the emotion's impact on speech in a meeting environment remains relatively unknown.

As the COVID-19 pandemic gripped the world, one of the primary mitigation measures employed by governments was the use of masks [10] to limit the spread of the virus. Therefore, employees are still required to wear a cloth or surgical face mask during physical face-to-face meetings [11]. However, with the use of non-transparent masks such as cloth or surgical masks, people have trouble distinguishing emotion during face-to-face conversations [12]. Furthermore, it causes confusion among people when discerning emotions from facial expressions due to face masks [13].

One common technique used to understand emotion is Facial Emotion Recognition (FER) technology, which analyses and predicts emotions based on facial expression [14]. Unfortunately, FER faces accuracy issues in predicting emotions due to the occlusion of the mouth by the face mask, hence reducing the overall effectiveness [15].

Due to the impact of face masks on facial features, more research projects were carried out to investigate the effect of face masks on speech which is another possible modality. It was observed that face masks had little effect and still retained speech understanding in a minimal background noise environment [16] [17] [18]. In addition, face masks have a relatively small impact on Automatic Speech Recognition (ASR) System performance at the sentence level as the system can still capture the content and transcribe it to text with a low Word Error Rate (WER) [19].

However, it was proven that wearing a face mask affected the speech rate and changed the acoustic speech signal as it affected the power distribution in frequencies above 5kHz for both cloth and surgical masks [20]. Moreover, it was observed that due to the usage of face masks, people generally would raise their voice deliberately or subconsciously hence increasing the overall voice intensity and changing the speech signal [21]. Efforts have been made to investigate the impact of face masks on speech through the Interspeech 2020 COMputational PARalinguistics challengE (ComParE). ComParE has been a running series of challenges since 2009 that focus on traits and states of speakers' speech signal properties [22]. One challenge in the Interspeech 2020 ComParE was the Mask Sub-Challenge, which focused on the face mask classification based on audio recordings [23]. This challenge proves that wearing a mask does affect acoustic features due to the change in frequencies, affecting the performance of the models depending on the type of acoustic features used [24]. Another technique used to understand emotion is Speech Emotion Recognition (SER), which predicts emotion based on speech signals [25]. Although past research proves that face masks affect speech features, the impact of face masks on SER remains unknown [26].

III. OBJECTIVES

There are two objectives. The first objective is to investigate the impact of face masks and emotion on ASR in a meeting environment. Although previous research was done for ASR, it was conducted either to determine the effects of wearing a face mask on ASR or the impact of emotion on ASR. As of the time of writing, there are no findings on which particular attribute (mask or emotion) serves as a more significant contributor to speech recognition error. The second objective aims to analyse the impact of face masks on SER in a meeting environment. So far, most research projects or competitions only focus on the impact of face masks on speech recognition systems, as all past studies focused on the impact of face masks on speech but not the underlying emotions through speech [26].

IV. METHODOLOGY

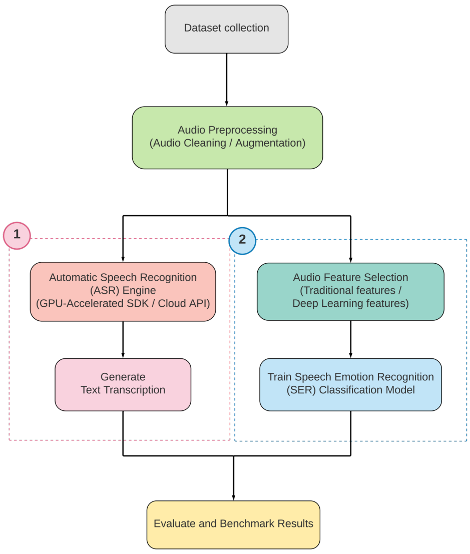

Dataset collection
Audio Preprocessing (Audio Cleaning / Augmentation)
1
2
Automatic Speech Recognition (ASR) Engine (GPU-Accelerated SDK / Cloud API)
Generate Text Transcription
Audio Feature Selection (Traditional features / Deep Learning features)
Train Speech Emotion Recognition (SER) Classification Model
Evaluate and Benchmark Results

Figure 1 Proposed steps

The proposed steps to be carried out are illustrated in Figure 1. Currently, there are not many databases for speech emotion with masks. Therefore, the first step to be carried out is data collection. The Emotional Speech Dataset (ESD) was used as a baseline for the dataset recording. ESD is a publicly released dataset containing 350 parallel utterances from 10 native English and Mandarin speakers along with the transcripts [27]. In addition, it only has recordings of five different emotions (Angry, Happy, Neutral, Sad, and Surprise).

However, out of the 350 lines, only 100 English lines will be recorded. Every participant will be required to finish uttering each line within 5s with and without a face mask repeated for all five emotions. This dataset will be used as the test dataset to evaluate the automatic speech recognition and speech emotion recognition models.

After data collection, the next step is to do audio preprocessing. Data preprocessing is essential in preparing the raw data for further processing. Different audio augmentation can be carried out on the ESD to create more synthetic data and improve model generalisation [28]. After cleaning and amplifying the recorded test dataset, it will be fed as an input to the ASR engine to convert the speech to text transcription. The transcription will be evaluated based on the Word Error Rate (WER) metric shown in Equation 1. The calculation of WER is based on the measurement metric Levenshtein distance, which measures the differences between two string sequences [29]. The lower the WER, the better the ASR engine.

$$WER = \frac{S + D + I}{N} \qquad (1)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of words in the reference. For SER, feature selection will be performed on the audio data. The feature type affects the speech processing and accuracy of the models [30]. The extracted audio features will then be fed as input to a simple feedforward neural network to train the classification model. Finally, the evaluation will be performed, and the metric will be based on the model accuracy of the recorded test dataset. The result will be used to understand and derive the impact of face masks on SER.

V. DATASET COLLECTION

There are no publicly available emotional speech datasets recorded with the participants wearing a face mask. Therefore, data collection was carried out to build our emotional speech dataset recorded with participants with and without a face mask. This dataset will be termed the Emotional Speech Meeting Dataset (ESMD) and will be used as the test dataset to evaluate the automatic speech recognition and speech emotion recognition models. The recording was done in a medium-size meeting room with background sounds (e.g. aircon condenser sound). Figure 2 depicts the recording setup in the meeting room.

The Emotional Speech Dataset (ESD) was used as a baseline. ESD is a publicly released dataset containing 350 parallel utterances from 10 native Mandarin speakers and ten native English speakers and transcripts [27]. However, only 100 English lines will be recorded out of the 350 lines. Every participant must finish uttering each line within 5s with and without a face mask repeated for all five different emotions (Angry, Happy, Neutral, Sad, Surprise). Figure 3 depicts the wave plot for each emotion. It was observed that the loudness of speech with a mask is relatively comparable to the loudness

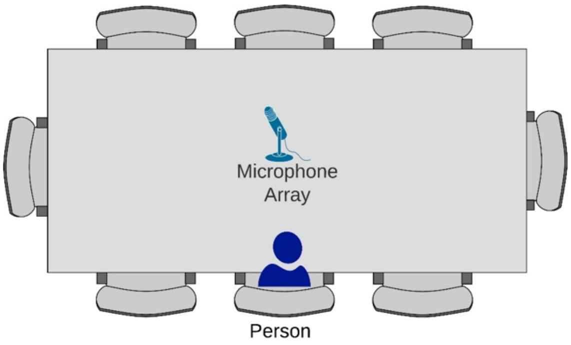


Figure 2 Recording Setup for Emotional Speech Dataset

of speech without a mask. Figure 4 depicts the log frequency power spectrogram for each emotion. There is not much difference between the speech with and without a mask from the spectrogram. However, it was observed that the tonal and pitch for each respective emotion are different. For example, the tonal for sadness is relatively flat compared to the other emotions.

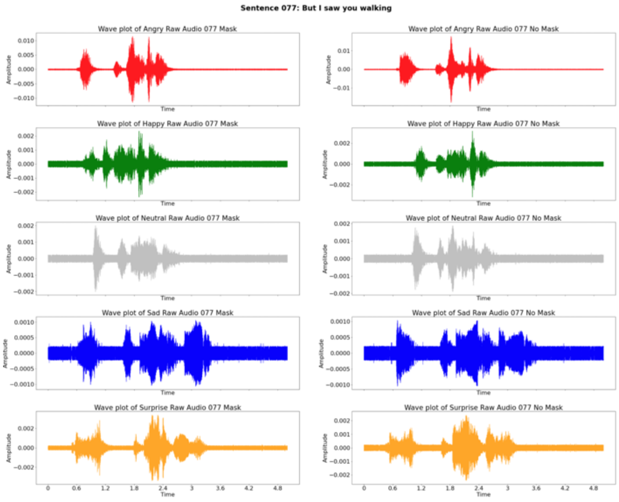


Figure 3 Wave Plot Mask vs No Mask

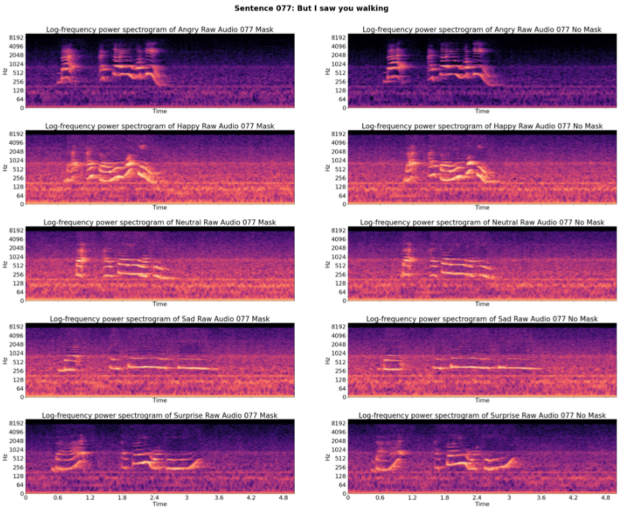


Figure 4 Log Frequency Power Spectrograms Mask vs No Mask

VI. Dataset Pre-Processing

Data preprocessing is essential in preparing the raw data for further processing.

A. Audio Cleaning

The audio recordings recorded in the meeting room with the miniDSP microphone array consisted of background noises and were very soft as the recordings were recorded in raw mode. The raw mode of the miniDSP provides more customisation but requires external amplification and noise reduction techniques [31]. Therefore audio amplification and noise reduction techniques were required. Two different ways to amplify the audio have been experimented with namely Gain Algorithm [32] and Librosa Normalisation Algorithm [33]. The gain amplification algorithm converts the linear volume to a logarithm scale based as given below

$$2^{\frac{\sqrt[8]{V}\times 198.0-192.0}{6.0}} \qquad (2)$$

where V is the Volume Factor. Meanwhile, the Librosa library performs amplification by normalising the given array along the chosen axis. The result can be seen in Figure 5, where the technique used by Librosa amplified the vocal tract region more, resulting in clipping of the audio. In comparison, the Ardour amplification technique amplified the audio equally, increasing the audio gain without clipping. Based on the result, the Ardour gain method provides a more desirable outcome and will be adopted.

There are two ways to amplify the audio. The first way is to amplify with a constant V. The second way is to amplify with a dynamic V depending on the audio intensity. If the audio intensity is lower, the V will be higher. If the audio intensity is higher, the value of V will be smaller.

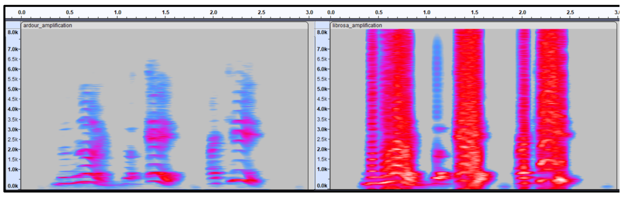


Figure 5 Gain Result vs Librosa Normalisation Result

For the reduction of noise, a python library was used. Noisereduce [34] is a python-based noise reduction technique for time-domain signals such as voice. It uses a technique known as "spectral gating" and operates by calculating a signal's spectrogram and predicting a noise threshold for each frequency band of that signal/noise [34]. Two different algorithms of the Noisereduce python library were experimented with to reduce the background noise in the audio namely Non-Stationary and Stationary. Per-Channel Energy Normalization is a modern bioacoustics approach that inspired the non-stationary algorithm [34]. This algorithm continuously updates the estimated noise threshold over time [34]. Conversely, stationary noise reduction maintains the estimated noise threshold throughout the signal at the same level [34]. Therefore, the data must first be calculated for each frequency channel to determine a noise gate [34]. The calculated noise gate is then applied to the input signal [34]. The main difference between both methods is that the non-stationary method allows the noise gate to alter with the passage of time [34].

Based on the result seen in Figure 6, the stationary method allows lesser low-frequency sounds through its filter than the non-stationary method, rendering a cleaner audio result. Furthermore, the stationary method may be more effective because the audio is recorded in a meeting room with

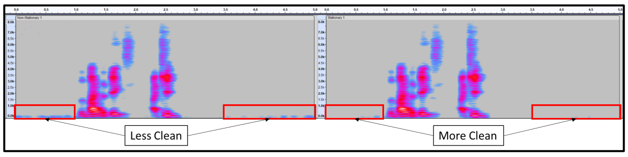


Figure 6 Gain Result vs Librosa Normalisation Result

persistent background noises, such as the air conditioner condenser sound. Hence, the stationary method will be used to reduce background noise from the audio.

After evaluating which suitable technique to use for amplification and noise reduction, the next step was to determine the order sequence to apply the selected techniques to the audio. An investigation was conducted to determine the best sequence of actions in carrying out the steps for audio cleaning:

1) Amplification > Noise Reduce
2) Noise Reduce > Amplification
3) Noise Reduce > Amplification > Noise Reduce

Figure 7 illustrates the result for each sequence of actions, indicating that the first sequence proved to be the most effective. Thus, the audio will be amplified first before performing noise reduction to reduce the background noises.

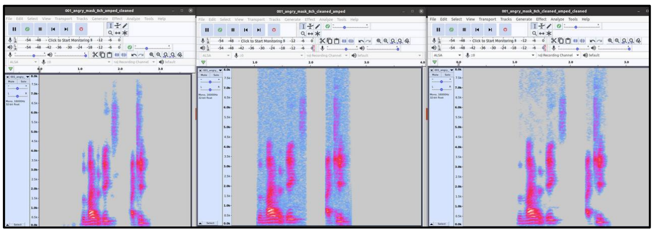


Figure 7 Result for the 3 different sequences of actions

B. Audio Augmentation

Data augmentation aids in generating synthetic data from existing data sets, improving the model's generalizability [28]. Therefore, data augmentation was performed on the public ESD to increase the model generalisation capability in the context of the meeting room.

Background noises (e.g. keyboard typing, mouse-clicking) that often occur in the meeting room were recorded and overlaid into the original audio file. Changing the audio intensity by making it softer (-24db) was also performed on the original audio file, considering that actual voice recording done in a meeting room will not be able to meet studio-level sound quality.

VII. Dataset Allocation

Two different datasets will be used. The first dataset is the ESMD, which consists of the audio recorded in the meeting room. The dataset used the publicly available ESD as a baseline and chose 100 lines out of the 350 lines to record. Four participants took part in the recording session where every participant must finish uttering each line within 5s with and without a face mask repeated for all five different emotions (Angry, Happy, Neutral, Sad, Surprise).

This dataset has two different versions due to the different ways of amplification. The audio dataset can either be amplified with a constant V or a dynamic V using the ardour gain method. The amplified dataset with a constant V will be termed “amplified equally”. In contrast, the amplified dataset with a dynamic V will be termed “amplified dynamically”. Both versions will solely be used for testing purposes for both ASR and SER experiments. Table 1 illustrate the breakdown of the ESMD for both versions.

Table 1 Emotional Speech Meeting Dataset (ESMD) breakdown

Participant	Gender	Age	Emotions	Channel	Mask	No Mask	Total Number of Audio
1	Male	25	5 Emotions (Angry, Happy, Neutral, Sad, Surprise)	1	100	100	1000
2	Male	24					1000
3	Female	24					1000
4	Female	22					1000
Total number of audio recordings in the dataset:					400	400	4000

The second dataset is the publicly available ESD which contains 350 parallel utterances from 10 native English and Mandarin speakers [27]. This dataset will only be used for SER model training. Many different versions of the ESD were created during the preprocessing step, but only two different versions of the dataset will be used. ESD6 is the original dataset, while ESD11-v3 is the augmented version of ESD6. Both versions will be used for training the SER classification model and evaluated with ESMD. Table 2 shows the breakdown for each different version.

Table 2 Emotional Speech Dataset (ESD) breakdown for each version

Dataset Version	Total Participants	Type	Emotions	Channel	Per Participant		Total
					Train	Val	
ESD6	20	Raw Audio	5 Emotions (Angry, Happy, Neutral, Sad, Surprise)	1	300		50
ESD11-v3	20	• Reduce noise • Softer (-24db) Added office background sounds			3600		600

VIII. Automatic Speech Recogntion (ASR)

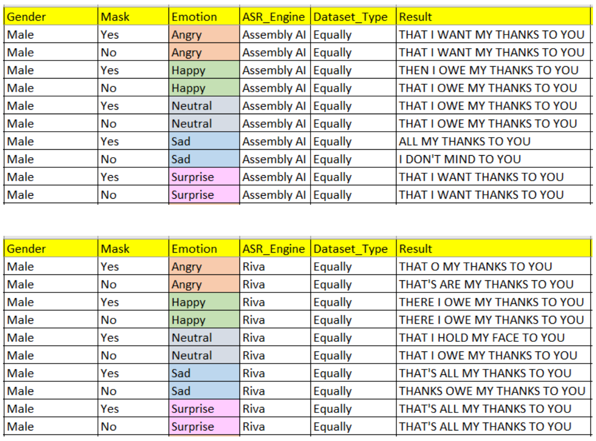
ASR is the process of converting human speech to text using ML technology [35]. AssemblyAI and NVIDIA Riva ASR engines will be used as a tool to investigate the impact of face masks and emotion on ASR to find out which particular attribute (mask or emotion) serves as a more significant contributor to speech recognition error in a meeting environment setting.

AssemblyAI is a cloud service provider that provides speech-to-text via an API call. Based on their benchmarking result, their speech-to-text model has a better accuracy rate than Google and Amazon Web Services (AWS) cloud speech-to-text model [36]. The model was conceptualised based on Transformer and Convolution Neural network (CNN) capabilities.

Transformers are known to capture speech’s global features (e.g. whole audio features), whereas CNN excels in modelling local features (e.g. chunks of individual audio features) [37]. Therefore, their approach by interleaving CNN layers between the Transformer layers allows the model to focus attention on both local and global features of speech, resulting in a better understanding of the audio speech pattern and thus generating accurate predictions [37].

Nvidia Riva is a GPU-accelerated world-class speech SDK that provides speech-to-text models [38]. NVIDIA Riva 2.0 engine that uses the Citrinet ASR model. Citrinet is one of the finest autoregressive transducer models based on an end-to-end convolutional Connectionist Temporal Classification (CTC) [39]. CTC is an approach that uses a single Recurrent Neural Network (RNN) to directly label unsegmented data sequences, eliminating the necessity for segmented training data and post-processing [40].

After preprocessing the ESMD by performing both amplification ways (equally and dynamically) and noise reduction, the audio data is then fed to each respective ASR engine to generate the transcription. Figure 8 shows an example of the transcription generated by AssemblyAI and Nvidia Riva 2.0.



Gender	Mask	Emotion	ASR_Engine	Dataset_Type	Result
Male	Yes	Angry	Assembly AI	Equally	THAT I WANT MY THANKS TO YOU
Male	No	Angry	Assembly AI	Equally	THAT I WANT MY THANKS TO YOU
Male	Yes	Happy	Assembly AI	Equally	THEN I OWE MY THANKS TO YOU
Male	No	Happy	Assembly AI	Equally	THAT I OWE MY THANKS TO YOU
Male	Yes	Neutral	Assembly AI	Equally	THAT I OWE MY THANKS TO YOU
Male	No	Neutral	Assembly AI	Equally	THAT I OWE MY THANKS TO YOU
Male	Yes	Sad	Assembly AI	Equally	ALL MY THANKS TO YOU
Male	No	Sad	Assembly AI	Equally	I DON'T MIND TO YOU
Male	Yes	Surprise	Assembly AI	Equally	THAT I WANT THANKS TO YOU
Male	No	Surprise	Assembly AI	Equally	THAT I WANT THANKS TO YOU

Gender	Mask	Emotion	ASR_Engine	Dataset_Type	Result
Male	Yes	Angry	Riva	Equally	THAT O MY THANKS TO YOU
Male	No	Angry	Riva	Equally	THAT'S ARE MY THANKS TO YOU
Male	Yes	Happy	Riva	Equally	THERE I OWE MY THANKS TO YOU
Male	No	Happy	Riva	Equally	THERE I OWE MY THANKS TO YOU
Male	Yes	Neutral	Riva	Equally	THAT I HOLD MY FACE TO YOU
Male	No	Neutral	Riva	Equally	THAT I OWE MY THANKS TO YOU
Male	Yes	Sad	Riva	Equally	THAT'S ALL MY THANKS TO YOU
Male	No	Sad	Riva	Equally	THANKS OWE MY THANKS TO YOU
Male	Yes	Surprise	Riva	Equally	THAT'S ALL MY THANKS TO YOU
Male	No	Surprise	Riva	Equally	THAT'S ALL MY THANKS TO YOU

Figure 8 AssemblyAI and Nvidia Riva 2.0 Transcription Result

The transcription result was then evaluated based on the WER. The calculation of WER is based on the measurement metric Levenshtein distance, which measures the differences between two string sequences [29]. The lower the WER, the better the ASR engine.

Table 3 illustrate the overall WER result for both AssemblyAI and Nvidia Riva 2.0. Nvidia Riva 2.0 has a lower WER than AssemblyAI by 10-12%. In addition, it was observed that the trend is consistent where both ASR engines performed better between a range of 3-5% on amplified dynamically dataset compared to amplified equally dataset.

Therefore, from the result, it can be concluded that the way of amplification affects the WER and Nvidia Riva 2.0 proves to be the better ASR engine compared to AssemblyAI. The next step was determining whether face masks significantly affect the ASR system performance. From the result seen in Table 4. It was observed that the WER only have a marginal difference range of 3-4% between the face mask and no face mask.

Table 3 Overall ASR Result based on WER

ASR Engine	**Amplified Equally ESMD (%)**	**Amplified Dynamically ESMD (%)**	**Delta (%)**
AssemblyAI	47.43	43.76	3.67
RIVA 2.0	36.88	32.12	4.76
Delta (%)	10.55	11.64	1.09

This result aligns with the previous research, which proves that the face masks have a relatively small impact on ASR System performance at the sentence level as the system can still capture the content and transcribe it to text with a low Word Error Rate (WER) [19].

Table 4 Face Mask vs No Face Mask ASR Result based on WER

Type	**Mask**	**NVIDIA RIVA 2.0**	**AssemblyAI**
Equally (%)	Yes	38.94	49.17
	No	34.82	45.69
Delta (%)		4.12	3.48
Dynamically (%)	Yes	33.98	45.59
	No	30.26	41.93
Delta (%)		3.72	3.66

Table 5 shows the WER result evaluated based on the five respective emotions for both ASR engines. It is evident that sadness affects speech recognition the most as it has the worst WER compared to the other emotions.

In addition, the technique of amplifying the audio affects the WER as most emotions WER have improved, with sadness speeches improving the most when using dynamic amplification.

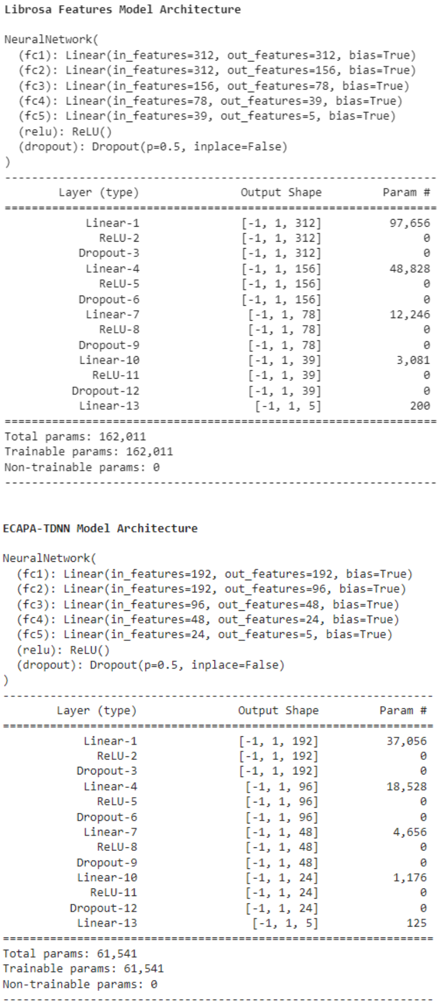
Table 5 Emotion ASR Result based on WER

Emotion	**NVIDIA Riva 2.0**			**AssemblyAI**		
	Equally (%)	**Dynamically (%)**	**Delta (%)**	**Equally (%)**	**Dynamically (%)**	**Delta (%)**
Angry	29.09	29	0.09	37.31	37.23	0.08
Happy	35.79	31.94	3.85	45.42	42.71	2.71
Neutral	29.20	22.24	6.96	41.82	35.45	6.37
Sad	54.18	42.48	11.7	64.34	56.27	8.07
Surprise	36.15	34.94	1.21	48.29	47.12	1.17

IX. Speech Emotion Recogntion (SER)

SER is another technique to understand emotion where it predicts emotion based on speech signals [25]. Although past research proves that face masks affect speech features, the impact of face masks on SER remains unknown [26]. Both versions (original and augmented) of the ESD will be used as the training data.

Traditional audio features will be extracted with Librosa python library [41] and used to train the SER classification model. Traditional audio features include Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate, Chroma, etc. SpeechBrain is an open-source, all-in-one speech toolbox that aims to make neural speech processing research and development more straightforward [42]. SpeechBrain's Emphasized Channel Attention, Propagation and Aggregation based Time Delay Neural Network (ECAPA-TDNN) speaker embedding model will also be used to extract the audio features. The pretrained model is trained on Voxceleb 1+ Voxceleb2 training recordings sampled at 16kHz (single channel) [43]. The extracted features will then be passed as an input into a simple neural network (NN). The model's architecture for traditional features extracted with Librosa and ECAPA-TDNN features is featured in Figure 9. The model saved at each epoch was evaluated on a validation set that the model had not seen before. The best model will then be selected based on the highest F1 Score, which is shown in equation 3. The F1-score takes the harmonic mean of a classifier's accuracy and recall creating a single statistic. It is



```
Librosa Features Model Architecture

NeuralNetwork(
  (fc1): Linear(in_features=312, out_features=312, bias=True)
  (fc2): Linear(in_features=312, out_features=156, bias=True)
  (fc3): Linear(in_features=156, out_features=78, bias=True)
  (fc4): Linear(in_features=78, out_features=39, bias=True)
  (fc5): Linear(in_features=39, out_features=5, bias=True)
  (relu): ReLU()
  (dropout): Dropout(p=0.5, inplace=False)
)
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1               [-1, 1, 312]          97,656
              ReLU-2               [-1, 1, 312]               0
           Dropout-3               [-1, 1, 312]               0
            Linear-4               [-1, 1, 156]          48,828
              ReLU-5               [-1, 1, 156]               0
           Dropout-6               [-1, 1, 156]               0
            Linear-7                [-1, 1, 78]          12,246
              ReLU-8                [-1, 1, 78]               0
           Dropout-9                [-1, 1, 78]               0
           Linear-10                [-1, 1, 39]           3,081
             ReLU-11                [-1, 1, 39]               0
          Dropout-12                [-1, 1, 39]               0
           Linear-13                 [-1, 1, 5]             200
================================================================
Total params: 162,011
Trainable params: 162,011
Non-trainable params: 0
----------------------------------------------------------------


ECAPA-TDNN Model Architecture

NeuralNetwork(
  (fc1): Linear(in_features=192, out_features=192, bias=True)
  (fc2): Linear(in_features=192, out_features=96, bias=True)
  (fc3): Linear(in_features=96, out_features=48, bias=True)
  (fc4): Linear(in_features=48, out_features=24, bias=True)
  (fc5): Linear(in_features=24, out_features=5, bias=True)
  (relu): ReLU()
  (dropout): Dropout(p=0.5, inplace=False)
)
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Linear-1               [-1, 1, 192]          37,056
              ReLU-2               [-1, 1, 192]               0
           Dropout-3               [-1, 1, 192]               0
            Linear-4                [-1, 1, 96]          18,528
              ReLU-5                [-1, 1, 96]               0
           Dropout-6                [-1, 1, 96]               0
            Linear-7                [-1, 1, 48]           4,656
              ReLU-8                [-1, 1, 48]               0
           Dropout-9                [-1, 1, 48]               0
           Linear-10                [-1, 1, 24]           1,176
             ReLU-11                [-1, 1, 24]               0
          Dropout-12                [-1, 1, 24]               0
           Linear-13                 [-1, 1, 5]             125
================================================================
Total params: 61,541
Trainable params: 61,541
Non-trainable params: 0
----------------------------------------------------------------
```

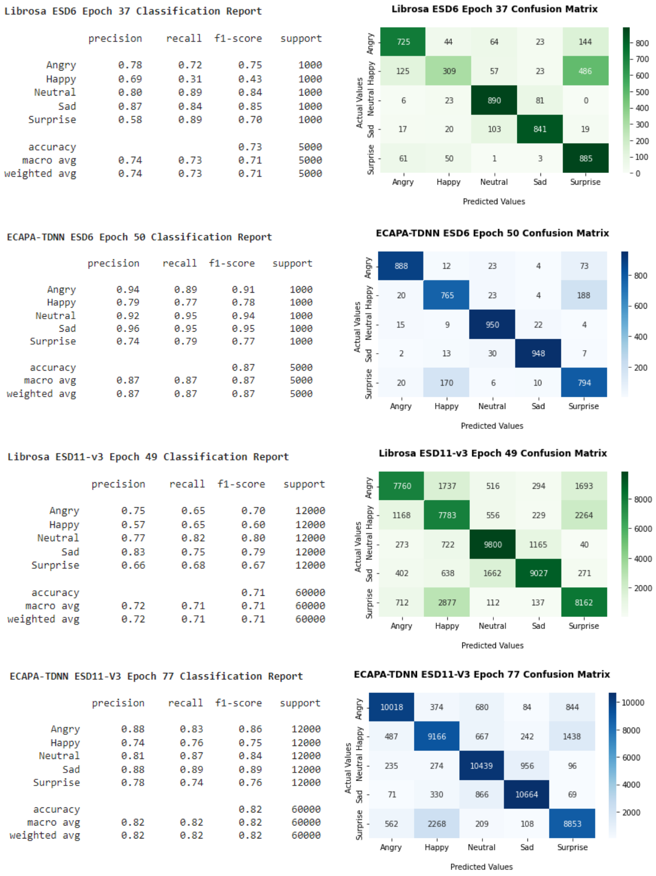
Figure 9 Model summary for traditional features extracted with Librosa and ECAPA-TDNN features

commonly used to compare the results between classifier models [44].

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Figure 10 illustrates the classification reports of the selected best models. The classification report shows that the model train with ECAPA-TDNN features generates a higher F1 score than the models train with traditional features extracted with Librosa. Generally, the model classified most of the emotions correctly, with only a small handful of data classified incorrectly. From the confusion matrix, a general observation was that Happy was classified wrongly as Surprise and vice-versa. Surprise also has some misclassification as Angry. Neutral and sad were also misclassified as one another. Next, the models will be evaluated with the ESMD. The ESMD has gone through some form of preprocessing, such as trimming the leading and trailing silence for each audio, amplifying, and undergoing noise reduction to remove the background sounds (e.g., aircon condenser).

Table 6 summarises the results of each model evaluation on each variation of the ESMD test dataset. The result showed that the ECAPA-TDNN feature was a better selection than traditional features extracted with the Librosa library. Furthermore, it was observed that the F1 Score between a mask and no mask and the F1 Score between the two ways (equally or dynamically) only differ slightly, with the highest difference of 4%. Therefore, using a face mask and how to amplify dynamically or equally do not impact much on SER. In addition, one noticeable observation is the difference between using the original dataset and the augmented dataset to train the classifier model. The model trained using the augmented dataset has a higher F1 score with the highest difference of 10%, which is a significant difference.



```
Librosa ESD6 Epoch 37 Classification Report

              precision    recall  f1-score   support

       Angry       0.78      0.72      0.75      1000
       Happy       0.69      0.31      0.43      1000
     Neutral       0.80      0.89      0.84      1000
         Sad       0.87      0.84      0.85      1000
    Surprise       0.58      0.89      0.70      1000

    accuracy                           0.73      5000
   macro avg       0.74      0.73      0.71      5000
weighted avg       0.74      0.73      0.71      5000
```

Librosa ESD6 Epoch 37 Confusion Matrix

Actual Values \ Predicted Values	Angry	Happy	Neutral	Sad	Surprise
Angry	725	44	64	23	144
Happy	125	309	57	23	486
Neutral	6	23	890	81	0
Sad	17	20	103	841	19
Surprise	61	50	1	3	885

```
ECAPA-TDNN ESD6 Epoch 50 Classification Report

              precision    recall  f1-score   support

       Angry       0.94      0.89      0.91      1000
       Happy       0.79      0.77      0.78      1000
     Neutral       0.92      0.95      0.94      1000
         Sad       0.96      0.95      0.95      1000
    Surprise       0.74      0.79      0.77      1000

    accuracy                           0.87      5000
   macro avg       0.87      0.87      0.87      5000
weighted avg       0.87      0.87      0.87      5000
```

ECAPA-TDNN ESD6 Epoch 50 Confusion Matrix

Actual Values \ Predicted Values	Angry	Happy	Neutral	Sad	Surprise
Angry	888	12	23	4	73
Happy	20	765	23	4	188
Neutral	15	9	950	22	4
Sad	2	13	30	948	7
Surprise	20	170	6	10	794

```
Librosa ESD11-v3 Epoch 49 Classification Report

              precision    recall  f1-score   support

       Angry       0.75      0.65      0.70     12000
       Happy       0.57      0.65      0.60     12000
     Neutral       0.77      0.82      0.80     12000
         Sad       0.83      0.75      0.79     12000
    Surprise       0.66      0.68      0.67     12000

    accuracy                           0.71     60000
   macro avg       0.72      0.71      0.71     60000
weighted avg       0.72      0.71      0.71     60000
```

Librosa ESD11-v3 Epoch 49 Confusion Matrix

Actual Values \ Predicted Values	Angry	Happy	Neutral	Sad	Surprise
Angry	7760	1737	516	294	1693
Happy	1168	7783	556	229	2264
Neutral	273	722	9800	1165	40
Sad	402	638	1662	9027	271
Surprise	712	2877	112	137	8162

```
ECAPA-TDNN ESD11-V3 Epoch 77 Classification Report

              precision    recall  f1-score   support

       Angry       0.88      0.83      0.86     12000
       Happy       0.74      0.76      0.75     12000
     Neutral       0.81      0.87      0.84     12000
         Sad       0.88      0.89      0.89     12000
    Surprise       0.78      0.74      0.76     12000

    accuracy                           0.82     60000
   macro avg       0.82      0.82      0.82     60000
weighted avg       0.82      0.82      0.82     60000
```

ECAPA-TDNN ESD11-V3 Epoch 77 Confusion Matrix

Actual Values \ Predicted Values	Angry	Happy	Neutral	Sad	Surprise
Angry	10018	374	680	84	844
Happy	487	9166	667	242	1438
Neutral	235	274	10439	956	96
Sad	71	330	866	10664	69
Surprise	562	2268	209	108	8853

Figure 10 Classification Reports for each model on ESD Validation Set

Table 6 Overall result evaluated on the ESDM test dataset

Train Dataset Version	Type	ECAPA-TDNN Features		Traditional Features extracted with Librosa	
		F1 Scores (%)		F1 Scores (%)	
		Amplified equally (trimmed)	Amplified dynamically (trimmed)	Amplified equally (trimmed)	Amplified dynamically (trimmed)
ESD6	Mask	51	52	29	33
	No Mask	52	52	30	34
ESD11-v3	Mask	62	60	41	38
	No Mask	58	56	41	39

X. Discussion

Based on the experiment result for ASR, NVIDIA RIVA 2.0 engine is better by 10 - 12% compared to AssemblyAI. Another finding is that the use of face masks will affect the WER by 3% to 4%, which aligns with the previous research, which proves that the face masks have a relatively small impact on ASR System performance at the sentence level as the system can still capture the content and transcribe it to text with a low Word Error Rate (WER) [19].

The next finding is that emotion does affect the speech significantly as the WER for each emotion differs, with sadness having the worst WER out of all the emotions for both amplification ways. This might be because the speech is softer when expressing sadness through speech, and articulation might not be clear. The result shows that the difference between the emotions with the best and worst WER ranges from 20% to 27%.

Furthermore, depending on the amplification technique, the emotion with the best WER may differ because the amplification technique will affect the overall intensity of the audio and hence affect the WER. For example, comparing the improvement from amplifying equally to amplifying dynamically, neutral speech WER has improved between the range of 6% to 7%, while sad speech WER has improved between 8% to 11%.

From the experiment and findings, it can be concluded that emotion contributes more significantly to speech processing error than the use of a face mask. Based on the experiment results for SER, using ECAPA-TDNN features as a training input proves to be a better choice than using handcrafted features. Furthermore, detecting emotions through speech is more feasible than FER, as face mask only minimally affects the classification of underlying emotions through speech.

It was observed that Surprise is often misclassified as Happy or Angry and Neutral as Sad, possibly because the average intensity between those emotions is very similar, as seen in Figure 11. However, another possible factor is that people of different gender express each emotion differently, which affects the overall result.

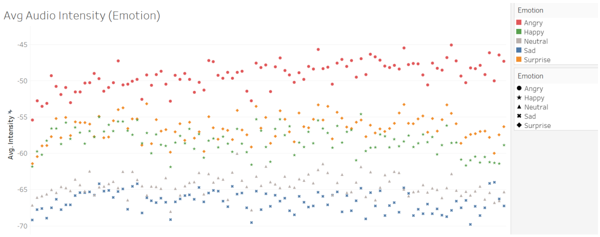


Figure 11 Average Audio Intensity between 5 emotions

XI. CONCLUSION AND FUTURE WORKS

Meetings are common in the workplace, and it's critical to understand employees' feelings during a meeting since emotions influence how individuals talk and behave. Furthermore, due to Covid-19, individuals are required to wear a face mask in meetings, and there is no research on how wearing a face mask impact the underlying emotion through speech. Experiments were conducted, which aligned with the previous research that the face masks have a relatively small impact on ASR System performance at the sentence level. In addition, emotion affects speech significantly, especially sadness. Therefore, it can be concluded that emotion contributes more significantly to speech processing error than a face mask.

Since emotion impacts speech processing, one possible future work is to train an ASR model with a dataset containing emotional elements in the speech and compare it against an ASR model trained on pure speech without factoring in emotion. There might be a possible improvement in WER by including speech with emotion as part of the training dataset for the ASR system.

The experiments also proved that using a face mask does not have much impact on emotion, making SER a feasible solution for detecting the underlying emotion through speech. However, the dataset used to evaluate the dataset only contained 4 participants. One future work is to record more participants and re-evaluate again. It was observed that emotion does affect the intensity of the overall speech. One possible future work is to explore more distinct features other than intensity, as the intensity for some emotions tends to overlap. If the feature used can separate the emotions better, there might be some improvement in determining the emotion.

Little was done in this research on improving the overall F1 Score, except for the augmentation technique, which proved useful in generalising the SER model, with the greatest improvement being a 10% increase in F1 Score. Possible enhancements include investigating more advanced modelling techniques such as using Transformers to improve the model accuracy in classifying emotions. Many state-of-the-art outcomes were attained using pre-trained neural networks composed of self-attention layers (transformers). Transformer-Encoder may improve the overall result assuming that the network can develop the ability to anticipate frequency.

REFERENCES

[1] MOS Legal, "Importance of Recording Meetings and Meeting Transcription," 29 Mar 2016. [Online]. Available: https://www.legaltranscriptionservice.com/blog/importance-of-recording-meetings-meeting-transcription

[2] R. Lennon, "Best meeting transcription software in 2022," Hugo, 2022. [Online]. Available: https://www.hugo.team/blog/best-meeting-transcription-software#

[3] A. Sima, "Managing your emotions for better meetings," 29 Jan 2018. [Online]. Available: https://magazine.vunela.com/https-www-aurorasa-coaching-com-emotions-in-the-workplace-ba85e0b74fa1.

[4] C. Out, M. Goudbeek and E. Krahm, "Do Speaker's emotions influence their language production? Studying the influence of disgust and amusement on alignment in interactive reference," *Language Sciences,* vol. 78, no. 0388-0001, p. 101255, 2020.

[5] E. Larson, "How The Most Common Emotions Affect Business Decision Making And What To Do About It," 21 Mar 2017. [Online]. Available: https://www.forbes.com/sites/eriklarson/2017/03/21/how-common-emotions-affect-team-decision-making-and-what-to-do-about-it/

[6] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura and J. Yamato, "Real-time meeting recognition and understanding using distant microphones and omni-directional camera," in *2010 IEEE Spoken Language Technology Workshop*, 2010.

[7] R. Munot and A. Nenkova, "Emotion Impacts Speech Recognition Performance," in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Student Research Workshop*, Minneapolis, Minnesota, 2019.

[8] A. Zgank and M. S. Maucec, "Influence of Emotional Speech on Continuous Speech Recognition," *2020 ELEKTRO,* pp. 1-4, 2020.

[9] F. Catania, P. Crovari, M. Spitale and F. Garzotto, "Automatic Speech Recognition: Do Emotions Matter?," in *2019 IEEE International Conference on Conversational Data Knowledge Engineering (CDKE)*, 2019.

[10] J. Howard, A. Huang, Z. Li, Z. Tufekci, V. Zdimal and H.-M. v. d. Westhuizen, "An evidence review of face masks against COVID-19," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 118, no. 4, 2021.

[11] . N. Thong, "COVID-19 Rules and Restrictions in Singapore (Nov 2021)," 25 Nov 2021. [Online]. Available:

https://singaporelegaladvice.com/covid-19-rules-restrictions-singapore-2021/.

[12] "The impact of facemasks on emotion recognition, trust attribution and re-identification," *Scientific Reports,* vol. 11, no. 1, pp. 5577-5590, 2021.

[13] E. Giovanelli, C. Valzolgher, E. Gessa, M. Todeschini and F. Pavani, "Unmasking the Difficulty of Listening to Talkers With Masks: lessons from the COVID-19 pandemic," *i-Perception,* vol. 12, no. 2, 01 Mar 2021.

[14] S. Gupta, "Facial Emotion Detection using AI: Use-Cases," 1 May 2018. [Online]. Available: https://towardsdatascience.com/facial-emotion-detection-using-ai-use-cases-248b932200d6.

[15] L. Zhang, B. Verma, D. Tjondronegoro and V. Chandran, "Facial Expression Analysis under Partial Occlusion: A Survey," *ACM Computing Surveys,* vol. 51, no. 2, pp. 1-49, 18 Jun 2018.

[16] J. C. Toscano and C. M. Toscano, "Effects of face masks on speech recognition in multi-talker babble noise," *PLoS ONE,* vol. 16, no. 2, 24 Feb 2021.

[17] "Face mask type affects audiovisual speech intelligibility and subjective listening effort in young and older adults," *Cognitive research: principles and implications,* vol. 6, no. 1, p. 49, 18 Jul 2021.

[18] D. Watt, "The science of how you sound when you talk through a face mask," The Conversation, 2 Jul 2020. [Online]. Available: https://theconversation.com/the-science-of-how-you-sound-when-you-talk-through-a-face-mask-139817.

[19] S. E. Gutz, H. P. Row and J. R. Green, "Speaking with a KN95 Face Mask: ASR Performance and Speaker Compensation," *Proc. Interspeech 2021,* pp. 4798-4802, 2021.

[20] M. Magee, C. Lewis, G. Noffs, H. Reece, J. C. S. Chan, C. J. Zaga, C. Paynter, O. Birchall, S. R. Azocar, A. Ediriweera, K. Kenyon, M. W. Caverlé, B. G. Schultz and A. P. Vogel, "Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols," *The Journal of the Acoustical Society of America,* vol. 148, no. 6, p. 3562, Dec 2020.

[21] R. Gama, M. E. Castro, J. T. van Lith-Bijl and G. Desuter, "Does the wearing of masks change voice and speech parameters?," *European Archives of Oto-Rhino-Laryngology,* 22 Sep 2021.

[22] compare.openaudio.eu, "Home of the Interspeech Computational Paralinguistics Challenges," 2021. [Online]. Available: http://www.compare.openaudio.eu/.

[23] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," *Proc. Interspeech 2020,* pp. 2042--2046, 2020.

[24] M. M. Mohamed, M. A. Nessiem, A. Batliner, C. Bergler, S. Hantke, M. Schmitt, A. Baird, A. Mallol-Ragolta, V. Karas, S. Amiriparian and B. W. Schuller, "Face mask recognition from audio: The MASC database and an overview on the mask challenge," *Pattern Recognition,* vol. 122, pp. 108361-108371, 4 Oct 2021.

[25] R. A. Khali, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access,* vol. 7, pp. 117327-117345, 2019.

[26] "Affects and Emotions in IEEE Signal Processing Magazine [From the Editor]," *IEEE Signal Processing Magazine,* vol. 38, no. 6, pp. 3-4, Nov 2021.

[27] NUS, "Publicly Available Emotional Speech Dataset (ESD) for Speech Synthesis and Voice Conversion," 2021. [Online]. Available: https://github.com/HLTSingapore/Emotional-Speech-Data.

[28] E. Ma, "Data Augmentation for Audio," Medium, 1 Jun 2019. [Online]. Available: https://medium.com/@makcedward/data-augmentation-for-audio-76912b01fdf6.

[29] F. Chiusano, "Two minutes NLP — Intro to Word Error Rate (WER) for Speech-to-Text," 03 Feb 2022. [Online]. Available: https://medium.com/nlplanet/two-minutes-nlp-intro-to-word-error-rate-wer-for-speech-to-text-fc17a98003ea.

[30] R. Rajak and R. Mall, "Emotion recognition from audio, dimensional and discrete categorisation using CNNs," *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON),* pp. 301-305, 2019.

[31] miniDSP Ltd, "miniDSP," 14 Sep 2018. [Online]. Available: https://www.minidsp.com/images/documents/UMA-8%20v2%20User%20manual.pdf. [Accessed 17 Jan 2022].

[32] Ardour, "slider_gain.h," 2021. [Online]. Available: https://github.com/Ardour/ardour/blob/master/libs/surfaces/tranzport/slider_gain.h.

[33] Librosa Development Team, ".normalize,"2021. [Online]. Available: https://librosa.org/doc/main/generated/librosa.util.normalize.html.

[34] T. Sainburg, M. Thielk and T. Gentner, "Finding, visualising, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology,* vol. 16, no. 10, p. e1008228, 2020.

[35] K. Foster, "What is ASR? A Comprehensive Overview of Automatic Speech Recognition Technology," AssemblyAI, 9 Nov 2021. [Online]. Available: https://www.assemblyai.com/blog/what-is-asr/.

[36] V. Lee, "2021 Benchmark Report," 3 Dec 2021. [Online]. Available: https://www.assemblyai.com/blog/2021-benchmark-report/.

[37] D. Fox, "Releasing our v8 Transcription Model - 18.72% Better Accuracy," 19 Oct 2021. [Online]. Available: https://www.assemblyai.com/blog/releasing-our-v8-transcription-model-major-accuracy-improvements/.

[38] NVIDIA Corporation, "NVIDIA RIVA," 2022. [Online]. Available: https://developer.nvidia.com/riva.

[39] S. Majumdar, J. Balam, O. Hrinchuk, V. Lavrukhin, V. Noroozi and B. Ginsburg, "Citrinet: Closing the Gap between Non-Autoregressive and Autoregressive," 5 Apr 2021. [Online].

[40] A. Ferraioli, "Connectionist Temporal Classification 1," 21 Jan 2021. [Online].Available:https://m0nads.wordpress.com/2020/01/21/introduction-to-connectionist-temporal-classification-part-i/.

[41] Librosa Development Team, "Feature extraction," 2013-2022. [Online]. Available: https://librosa.org/doc/main/feature.html.

[42] M. Ravanelli , T. Parcollet, P. Plantinga , A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori and Y. Bengio, *SpeechBrain: A General-Purpose Speech Toolkit,* 2021.

[43] B. Desplanques, J. Thienpondt and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *INTERSPEECH 2020*, Shanghai, China, 2020.

[44] J. Korstanje, "The F1 score," 31 Aug 2021. [Online]. Available: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.