Speech Emotion Recognition with wav2vec 2.0

ELEC-E5510

Feifan Wang, Nora Raud and Priyanshi Pal

15 December 2023

Contents

1 Introduction

1.1 Problem

Emotion recognition in speech is a valuable problem to solve, given that it associates to emotional intelligence in humans, a skill they use daily but that machines do not [1]. Machine learning is more difficult to apply to audio due to low accessibility of labelled data, large file sizes, temporality, and periodicity. In addition, there is no wide agreement on what structure emotion possesses [1], based on which to decide what sort of an output domain SER should have.

One of these problems, the lack of large amounts of labelled data, is remedied by utilizing a wav2vec 2.0 model, which is pretrained on large amounts of unlabelled data, after which it can be fine-tuned to a specific problem [2]. However wav2vec 2.0 could still be biased based on its pretraining dataset [3].

We intend to investigate the following on the dataset CREMA-D [4]:

1. How distinct are the emotion clusters based on audio features and to what degree?
2. How well can Speech Emotion Recognition (SER) be performed using classical classification approach?
3. How does the selection of a pretrained wav2vec 2.0 model and tuning hyperparameters affect outcomes?
4. How well does wav2vec 2.0 work with speech emotion recognition? What factors into that?

1.2 Speech Emotion Recognition: Literature Study

Emotion, despite having no agreed upon definition, is often characterised in terms of activation and valence [5]. The amount of energy required to express a certain emotion is known as activation while valence is typically defined as pleasantness or unpleasantness associated with a emotion, such as happiness, sadness or anger. Emotional arousal is one of the most discernible aspect in vocal communication. Other aspects that highly correlate with certain emotions are pitch, rhythm, voice quality [6].

Being able to automatically recognize human emotions and emotion-related states from speech can have significant implications for improving human intelligence, rational decision-making, social interaction, perception, memory, learning, and creation. Additionally, SER can help improve the accuracy and efficiency of various speech-related technologies and services, such as speech recognition, speech synthesis, and human-robot interaction [7].

There are three components to emotion recognition, of which speech emotion recognition is a sub task. These are linguistic content, paralinguistic content, and non-linguistic content. Emotion recognition is a type of intelligence that machines need to perform many tasks like Human-Computer Interaction (HCI), and mental state and mental health recognition. There is no single theory for understanding where and why emotions occur or what its structure is. However, the Dimensional and Discrete thoeries of emotion, in addition to the evolutionary view, are they main ways of understanding emotion [1].

Approaching the automatic recognition of emotion requires an appropriate emotion representation model. Hence, prior to recognising emotions in speech, a very important task is the need to identify the set of important emotions that must be classified with an emotion recogniser.

This brings us to an important question, how many emotions should be classified?
Representing emotions in an adequate way to ensure proper fit with the psychology literature while also selecting a representation that can well be handled by a machine, two models are usually found in practice. The first one is called "palette theory" of emotions. Various researches agree to this Dimensional theory, which states that any emotion can be composed using primary emotions, similar to the color theory in vision science. These primary emotions are anger, disgust, fear, joy, sadness and surprise [8]. The second approach a value "continuous" dimension approach instead of discretised versions.

Ekman, a renowned psychologist, also argues that these primary emotions are experienced universally in all cultures [9].

There also exists certain challenges that make Emotion Recognition in speech a complicated problem. Factors such as the speaker's culture, environment and mother tongue can influence how they express emotions [10]. Cross cultural studies have been quite sparse in the past and robustness across cultural differences is still an active research area [11].

1.2.1 Methods

The classical approach to speech emotion recognition (SER) involves four main steps: 1) data collection and preprocessing, 2) feature extraction, 3) feature selection, 4) and classification.

Classical algorithms used in Speech Emotion Recognition are Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Decision Trees, with SVM performing the best. In recent years, neural networks, especially deep neural networks (DNNs) have been shown to overcome the limitations of handcrafted features and attain better results than classic classifiers, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Deep Belief Networks (DBNs). Combinations of the two, such as the combination of a deep neural network and GMM or a DBN and a SVM, have also been shown to be successful [12]. Transfer learning can also be used to develop robust SER models, which reduces the time and effort required for training [13].

1.2.2 wav2vec 2.0

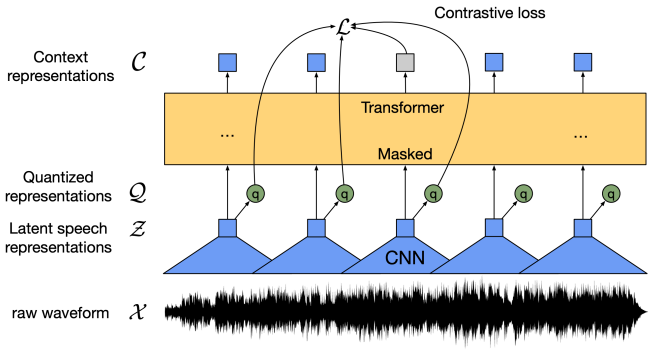


Figure 1: Illustration of wav2vec2.0 framework[2]

The wav2vec 2.0 model is especially beneficial in the upstream of transfer learning because it allows for pretraining on large amounts of unlabeled data. It is a self-supervised learning model. The wav2vec 2.0 model improves on the previous model by learning context and speech units end-to-end using a Transformer network. This allows the model to be pretrained to produce quantized latent speech features, which can then be trained and fine-tuned for a particular downstream task, for example, speech emotion recognition. This can be done on small-data sets and puts less strain on resources, making it particularly useful for speech problems with little data. This approach has been shown to outperform previous semi-supervised methods while being conceptually simpler [2].

Wav2vec 2.0 includes a Transformer architecture in its context network. Transformer is an encoder-decoder model that can be applied in finding latent speech features through self-attention. Transformers are similar to RNNs and CNNs, which have contained Transformer architecture for attention, but the paper shows that they work on their own as well. Compared to RNN, this means that a Transformer is less dependent on previous output and can be more easily parallelized. Compared to CNN, the Transformer is better suited for finding non-local speech dependencies [14]. Slightly different from the original one, the positional embedding in wav2vec 2.0's context network is encoded as relative using a convolutional layer [2].

For self-supervised training, the quantization model in wav2vec 2.0 converts the continuous output of the feature encoder to a finite set of discrete space for speech representations via product quantization[2]. The number of

phonemes and pairs of them in a language are finite, which means they can be represented by the same latent speech representation.

The objective of Wav2vec 2.0 is a weighted sum of contrastive loss and diversity loss. Contrastive loss identifies the true quantized latent speech representation. It is then augmented by diversity loss to encourage the equal use of codebook entries.[2]

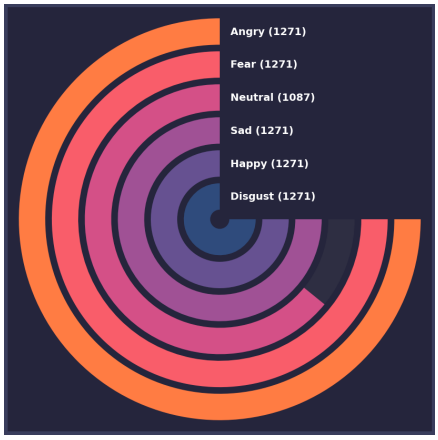
2 Dataset: CREMA-D

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [4] is considered for our project, due to its diversity of subjects, ranging from race to age. It consists of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified).

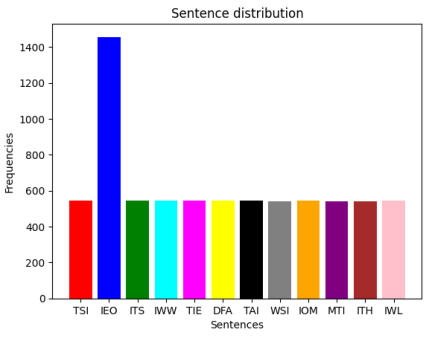
Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

The sentences were as follows: It's eleven o'clock (IEO), That is exactly what happened (TIE), I'm on my way to the meeting (IOM), I wonder what this is about (IWW), The airplane is almost full (TAI), Maybe tomorrow it will be cold (MTI), I would like a new alarm clock (IWL), I think I have a doctor's appointment (ITH), Don't forget a jacket (DFA), I think I've seen this before (ITS), The surface is slick (TSI), We'll stop in a couple of minutes (WSI).

The distribution of emotions, sentences and races are as follows:



(a) Class distribution for various emotions in CREMA-D



(b) Class distribution for sentences in CREMA-D

Actors' Race/Ethnicity Breakdown

Ethnicity / Race	Not Hispanic	Hispanic	Total
Caucasian	53	8	61
African American	21	1	22
Asian	7	0	7
Unspecified	0	1	1
Total	81	10	91

(c) Class distribution for Race and Ethnicity [4]

Figure 2: CREMA-D data description

The emotion intensity distributions are as follows:

- (Unspecified) XX: 6077, (Low) LO: 455, (High) HI: 455, (Medium) MD: 455

An interesting aspect to consider evaluating is, **are there emotions which are not so easily recognised during crowd-sourcing?**
Along with the dataset is also a compilation of agreement score for each audio file corresponding to a score for a given emotion. We found mean agreement score found for various emotions are as follows: Anger (ANG): 0.684, Disgust (DIS): 0.638, Fear (FEA): 0.637, Happiness (HAP): 0.792, Neutral (NEU): 0.763 and Sadness (SAD): 0.619. This revealed that the emotions such as Sadness, Fear and Disgust were less unanimously recognised and are more prone to recognition errors in humans.

3 Experiments

3.1 Methodology

3.1.1 How distinct are the emotion clusters based on audio features and to what degree?

Two parts are involved to our approach for this problem. To extract audio features, 13-dimensional MFCCs are obtained for each file, correspondiong to 7,332 audio files. Then, the MFCC features are standardized by removing the mean and scaling to unit variance.

As we know, the dataset is contains audio clips specific to 6 emotions and to observe how distict or closely related they are, K-means clustering algorithm is deployed for 6 clusters on the standardised MFCC features, followed by Principal Component Analysis with two components for visualisation. The results for each cluster is as follows:

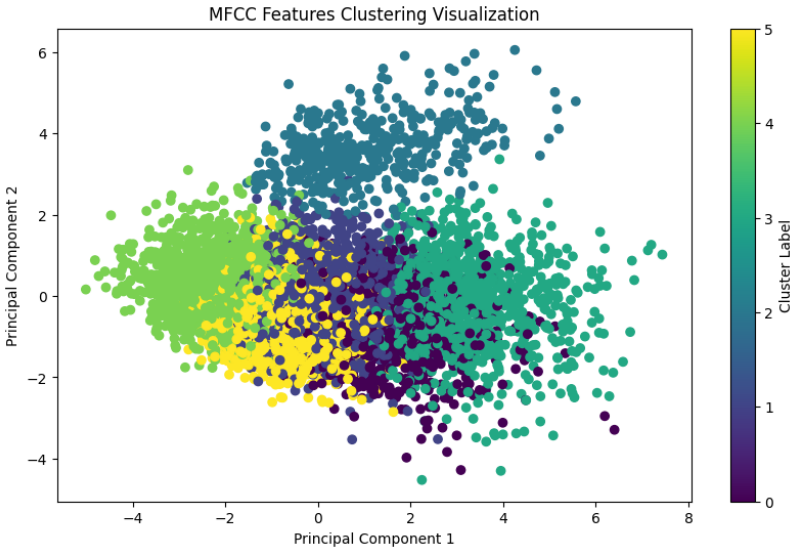


Figure 3

Table 1: Cluster Results

Cluster	ANG	DIS	FEA	HAP	NEU	SAD
0	274	101	106	279	127	33
1	197	306	260	347	326	141
2	75	73	75	80	60	71
3	573	71	147	178	6	6
4	27	333	436	136	222	694
5	125	387	247	251	346	326

Interpretation: It can be observed that the emotions aren't well separable based on audio features solely, such as MFCCs. Although some emotions are a bit more distinct [(Cluster-3, anger), (Cluster-4, Sadness)], rest seem to be somewhat related. The Cluster-2 was supposed to be Neutral but it can be similar to any other Emotion in terms of MFCCs. The indicates the recognising emotions solely based on speech based features may be quite complex for the given dataset and it is further confirmed that the clusters aren't that easil well formed (given the Silhouette score was found to be lower than 0.2, indicating low cluster quality)

3.1.2 How well can Speech Emotion Recognition (SER) be performed using classical classification approach?

We consider using some classical methods for classification such as Random Forest and Support Vector Machines (SVM), for the sake of curiosity and comparison purposes. Also, Random Forest and SVMs are conventionally used for various classification tasks in general. The data is split into training (5,209) and test set (2,233).

Classification using Random Forest and SVM

Table 2: Random Forest Classification Results

Emotion	Precision	Recall	F1-score	Support
ANG	0.56	0.72	0.63	386
DIS	0.38	0.26	0.31	384
FEA	0.28	0.14	0.18	379
HAP	0.37	0.37	0.37	390
NEU	0.32	0.42	0.36	318
SAD	0.45	0.57	0.50	376

Table 3: SVM Classification Results

Emotion	Precision	Recall	F1-score	Support
ANG	0.61	0.69	0.65	382
DIS	0.39	0.35	0.37	381
FEA	0.37	0.22	0.28	381
HAP	0.40	0.41	0.40	382
NEU	0.40	0.43	0.42	326
SAD	0.46	0.60	0.52	381

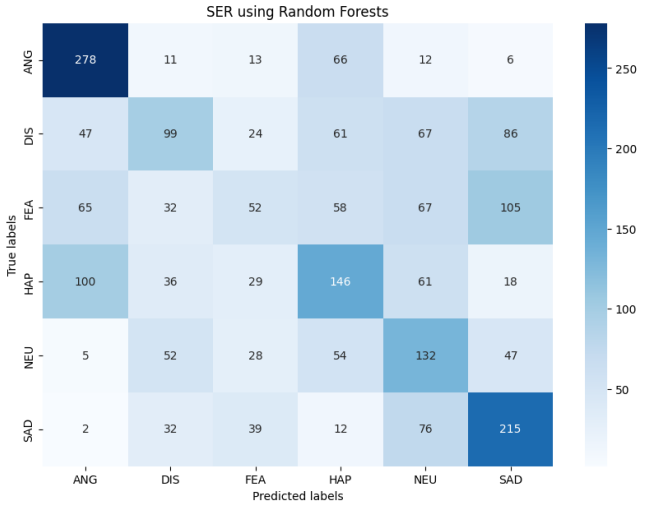


Figure 4: Confusion Matrix for Random Forest

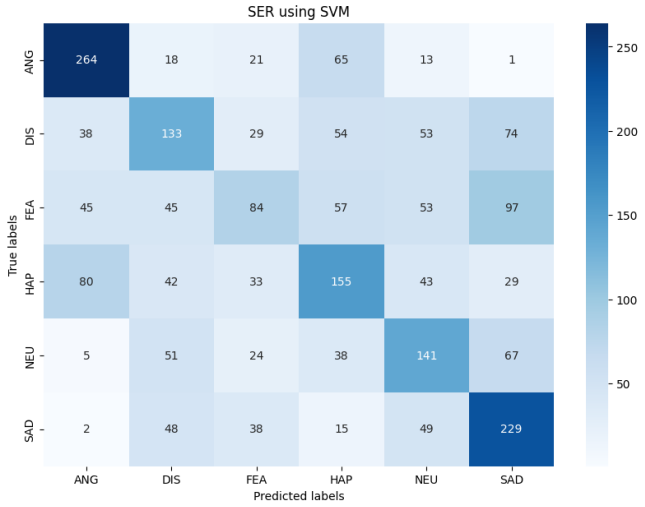


Figure 5: Confusion Matrix for SVM

Selection of Hyperparameters for Random Forest:

For random forests, it is essential to find the best parameters for number of estimators, max depth, min samples split and min samples leaf.

A Simple grid search was implemented to get optimum hyperparameters among the following-
'n_estimators': 100, 200, 300,
'max_depth': 10, 20,
'min_samples_split': 2, 5, 10,
'min_samples_leaf': 1, 2, 4

To check if it is overfitting or not, 5 fold cross valdiation (CV) was performed and results for each fold were 0.416, 0.401, 0.445, 0.450 and 0.4448 where Average CV Score for Random Forest was: 0.431

Relatively, the best performance (0.45 accuracy) was achieved through the 300 estimators, max_depth of 10, min_samples_leaf=4, min_samples_split=10.

Selection of Hyperparameters for SVM:

For support Vector machines, the important hyperparameters worth tuning are C, gamma and choosing the kernel. Although there can be various combinations, the following for implemented.
'C': 0.1, 1, 10, 100,
'gamma': 1, 0.1, 0.01, 0.001,
'kernel': 'rbf', 'poly', 'sigmoid'

The results of 5 Fold cross-validation are as follows:
0.446, 0.437, 0.478, 0.465, 0.455, where the Average CV Score for SVM was: 0.456.

Best accuracy (0.478) was achieved with C=10, gamma as 0.01 and kernel as Radial basis function (rbf).

Using the chosen configurations, the classification task was performed for 6 emotions based on the MFCC features.

Interpretation: In both classification models, "Anger" and "Sad" is more easily recognisable. It can also be observed that "Sad" can be similar to "Fear", as seen in the confusion matrices. The hardest emotion to recognise was "Fear".

Lastly, although the classical models did not have very good accuracy, it still helped us understand more about which emotions can be close to which other ones, the easily and the hard to recognise ones. Therefore, certainly models with better accuracy are required, be it through much thorough fine-tuning , using hybrid models made of the classical models or through pre-trained models.

A small note worth mentioning is that, Although Fear and Disgust seemed harder to recognise, similar to human crowdsourcing, Sad was contrarily recognised better.

3.1.3 How to select a pretrained wav2vec 2.0 model and find best hyperparameters?

wav2vec 2.0 was trained multiple times with different hyperparameters using different pretrained models. This allowed for the comparison of the effects they had on the task. Throughout, the same split for training, evaluation, and testing datasets was used, including the same seed for random selection. The split was 0.7 for training, and 0.15 for evaluation and testing. All actors were present in all splits, however the spread of actors was not balanced, meaning the model had seen some actors more than others in training.

Algorithm 1 Learning rates beam search

$L_{best} \leftarrow L_{init} \leftarrow$ initial lr ▷ Usually 0.0001
$L_{min} \leftarrow \frac{L}{10}$ ▷ lower bound of lr search range
$L_{max} \leftarrow L \times 10$ ▷ upper bound of lr search range
$K \leftarrow$ epochs $\leftarrow 5$
$F \leftarrow$ beam search factor $\leftarrow 0.5$
$S \leftarrow$ score $\leftarrow 0$
$S_{max} \leftarrow$ max score $\leftarrow$ -1
while $S > S_{max}$ **do**
    $S_{max} \leftarrow S$
    for $L_{cur} = [L_{min}, L_{best}, L_{max}]$ **do**
        Train model with L_{cur} learning rate and K epochs
        $S_{cur} \leftarrow$ Compute current score metric
        if $S < S_{cur}$ **then**
            $S \leftarrow S_{cur}$;
            $L_{max} \leftarrow L_{cur}$;
        end if
    end for
    $K \leftarrow K + 5$
    $L_{min} \leftarrow L_{best} - L_{init} \times F$
    $L_{max} \leftarrow L_{best} + L_{init} \times F$
    $F \leftarrow \frac{F}{2}$
end while

During exploration of pretrained models and training hyperparameters, the selection was evaluated using F1 score. At first different pretrained wav2vec 2.0 models were applied to the problem in search of more successful models. After that, hyperparameters were the focus for improving results. While different hyperparameters were tested for all models, the focus was on the more successful ones. The main hyperparameter of interest was learning rate due to its large influence on the outcomes. Different learning rates were observed, first in a wider range, then in smaller

range, depending on the previous step. There were also different approaches to changing the learning rate after the starting point: linear decrease over time, cosine decrease over time, and loss-dependent decrease.

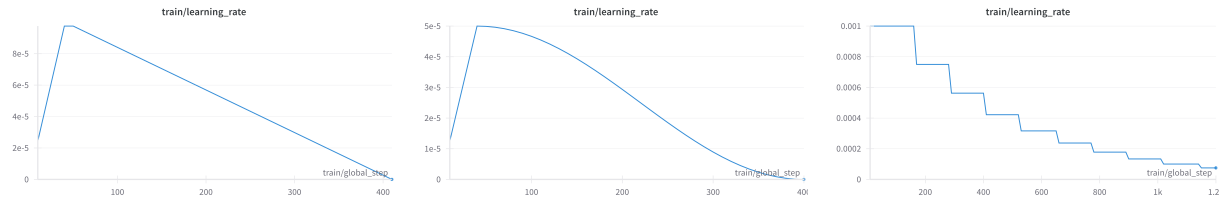


Figure 6: Schedulers (linear, cosine, reduce on plateau)

To reduce training time, we implemented an Early Stop strategy when the model encounters a plateau. Training is stopped if there is no improvement in model performance for 3 epochs. This means that the model does not need to complete the entire predetermined number of epochs once it reaches a bottleneck.

3.1.4 How well does wav2vec 2.0 work with speech emotion recognition on CREMA-D? What factors into that?

After it became more difficult to find better pretrained wav2vec 2.0 models and better hyperparameters, we were ready to observe its ability to detect emotions based on the Crema-D dataset. The approach was both quantitative and qualitative and done based on the metadata features for audio clips, demographics, and ultimately the prediction. At first, confusion, correlation of features, and F1 differences over all possible groupings of a single feature were observed. Then, the data was analyzed more closely to assess why wav2vec 2.0 was more or less successful. The following features were created for this task: prediction correctness (0/1), age binning (steps of 10), and one-hot encoding for nominal features. The only objective audio measure was duration. Finally, combinations of different features were observed with a decision tree classifier using one-hot encoding for nominal values.

3.2 Pretrained wav2vec 2.0 Models

A base size model contains 90 million parameters, while a large size model has 300 million parameters. The table4 shows details about the pretrained model we use.

Model	Description
facebook/wav2vec2-large-960h	**Baseline** model pretrained on 960 hours of Librispeech
facebook/wav2vec2-base-960h	The same as baseline except for a **smaller parameter size**.
facebook/wav2vec2-xls-r-300m	Pretrained on 436k hours of unlabeled speech in **128 languages**
jonatasgrosman/wav2vec2-large-xlsr-53-english	The same as facebook/wav2vec2-xls-r-300m but **fine-tuned on English**
facebook/wav2vec2-large-100k-voxpopuli	Pretrained on the 100k hours unlabeled subset of VoxPopuli[15] corpus that contains different **accents**

Table 4: Model description

3.3 Results

3.3.1 Assessment of pretrained models for Crema-D

According to the training results, we can divide the model into two groups: one comprising the wav2vec2-large-960h and wav2vec2-base-960h, and the other consisting of wav2vec2-xls-r-300m, wav2vec2-large-xlsr-53-english, and

Model	Learning rate	epochs	scheduler	warmup ratio	F1 score
facebook/wav2vec2-large-960h	0.00008	15	linear	0.1	0.7196
facebook/wav2vec2-base-960h	0.00009	20	linear	0.1	0.7343
facebook/wav2vec2-xls-r-300m	**0.000125**	**30**	**cosine**	**0.1**	**0.7932**
jonatasgrosman/wav2vec2-large-xlsr-53-english	0.00014	10	cosine	0.1	0.7914
facebook/wav2vec2-large-100k-voxpopuli	0.0005	15	cosine	0.1	0.7754

Table 5: Best model results

wav2vec2-large-100k-voxpopuli. There is almost no performance difference among the models in the first group, suggesting that the model size has minimal impact on the task. The 90 million parameter model demonstrates sufficient capability to extract audio feature from the dataset.

Similarly, there is a small performance gap among the models in the second group, indicating that the number of languages in the pretraining data and whether fine-tuning is done on the target language have a limited impact on the task. When comparing the two groups of models, it is observed that the pre-training datasets for the second group include VoxPopuli. This implies that datasets with different accents or languages are more suitable for this task. Rajoo[16] suggested that emotions expressed by native speakers have higher accuracy rates. Given that actors in CREMA-D come from diverse backgrounds and have different accents, the performance of fine-tuned models in speech emotion recognition may be influenced by the cultural background embedded in the language content of the original pretrained models.

3.3.2 Assessment of wav2vec 2.0 for Crema-D

The model selected for assessment was wav2vec2-xls-r-300m with loss changes on plataeus and an initial learning rate of 0.0001. It was trained for 30 epochs and reached an F1 score of 0.7932. The same model was used again with a different learning rate setup that resulted in a slightly better evaluation split F1 score, however, there was no meaningful difference in results.

as seen on figure 7, sadness, fear, and disgust were the least recalled classifications. Disgust was confused for anger, fear for sadness, and sadness for fear, disgust and neutralness. Sadness was the worst recalled emotion overall. Anger was best recalled, possibly because it distinctly higher in energy when listened to. The opposite might be true for some other emotions, being expressed with low energy, becoming less distinct from each other due to less energy information and and disappearing into the noise.

As noted previously, this sort of classification might not be the best understanding of how emotions work. This is also visible in the emotion level's relation to its learnability. The F1 score was highest for XX (undefined emotion level) with 0.80, and worse for the levelled voice clips (low 0.77, mid 0.74, high 0.72). Leaving the levelled voice clips out has no significant impact on the confusion matrix. Emotion levels complicated classification. It is therefore not trivial to say that a emphasized version of the emotion is the same as its normal version.

Women's emotions were detected better than men's emotion, F1 scores of 0.81 and 0.77. The test split was slightly biased towards men, meaning the other two together were slightly biased towards women. This could be a reason for the slightly discrepancy. In addition men's voices are on the same frequency as a lot of noise, which was clearly present in this dataset.

Ethnicity had no impact on detection despite bias in the dataset. However, race did have a large impact, with F1s of 0.73 for African American, 0.75 for Unknown, 0.77 for Asian, and 0.76 for Caucasion. Note that the support for African American was 270, and 761 for Caucasian. Listening to the audio, there was an audible difference in speech affection, especially in how actors used energy and pitch. This together with the biasedness in the data, might lead to the worse results.

There was a high variance for sentences themselves, so the sound quality of the sentences was clearly important. DFA had the highest F1 score of 0.87 and IWW had the lowest F1 score of 0.74, which makes sense, given how one sentence has way less vocals than the other.

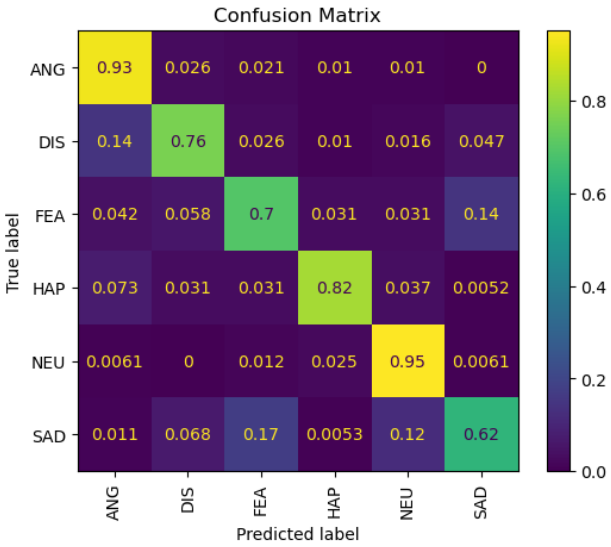


Figure 7: The emotion confusion matrix of the test split for the 6 emotions (including all levels).

Finally there is a large variety for F1 scores of different actors as seen in figure 8. Some of them may be outliers due to low support, however some voices are clearly detected than others. However there was no strong overall correlation between how many clips an actor had in the test split vs the other splits. However, this may still be a heightening factor for rarer speech patterns. In some of the low F1 actors, the environment is clearly noisy, quiet, and poorly clipped, leading to contamination with other voices. In addition some actors did all their recordings in the same environment and other did them across several environments. Given that the same actors were present in both test and train splits, this makes fitting to some quality of the actor's voice harder. However, noise level alone was not indicative of poor predictions, as it was present across actors.

After this, a decision tree was made to predict based on the metadata, whether the model classified the emotion correctly. This informed the following conclusions. The clearest indicator of misclassification was the emotion itself. However, ignoring that, a combination of an actor having less clips in the training split, and having long clips were susceptible to misclassification. Otherwise, whether a sound clip was levelled or spoken by an African American was also indicative of that. Age was a compounding issue in all cases, this is likely because the dataset was biased towards younger people.

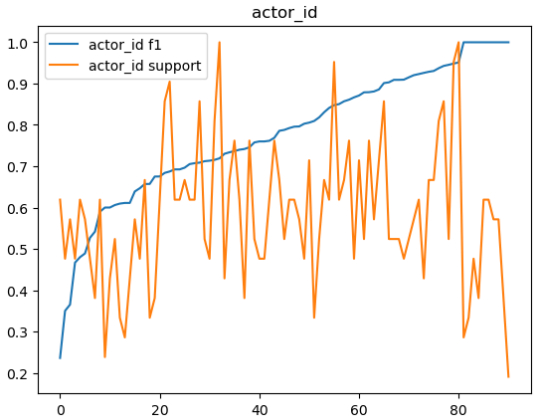


Figure 8: The actors, sorted by F1 score.

4 Conclusion

In recent years wav2vec 2.0 has been a significantly useful approach to many audio problems, and it is no different for emotion recognition in speech. However, it is not a magic bullet, as it is still prone to biases in the data, the quality of the original data, and the problem stipulation. The dimensionality and structure of emotions, especially when intensiveness is considered, is not trivial, and is even subjective, so it is sensitive to bias and smallness of the database. Some emotions appeared to be similar, happiness with anger, and fear, disgust, and sadness with each other. It is important for voices with similar emotion patterns in speech to exist in the database for good classification.

5 Division of labor

Feifan Wang

1. Literature review of wav2vec 2.0, transformers, and SER
2. Selected target pretrained models
3. Builded a complete fine-tuning pipeline on Kaggle
4. Introduced and improved the methods for fine-tuning the learning rate of large models.
5. Trained 4 models: facebook/wav2vec2-base-960h, facebook/wav2vec2-large-960h, facebook/wav2vec2-large-100k-voxpopuli and jonatasgrosman/wav2vec2-large-xlsr-53-english
6. Analysed results between different pretrianed models

Nora Raud

1. Literature review of wav2vec 2.0, transformers, and SER
2. Researched databases and CREMA-D in-depth
3. Trained 1 model: facebook/wav2vec2-xls-r-300m
4. In-depth assessment of wav2vec2 2.0 on CREMA-D

Priyanshi Pal

1. Literature review for Emotion Recognition and Speech in emotion recognition.
2. Performed Exploratory Data Analysis on CREMA-D
3. Performed Clustering on MFCC features using K-means and PCA.
4. Implemented Random Forest and Support Vector Machine based classification models.

6 Acknowledgements

Thanks to the official framework provided by Hugging Face, which has streamlined the training process for fine-tuning our models. Also, appreciation for the introduction of beam search and greedy search in the course, which served as the inspiration for our fine-tuning learning rate approach.

References

[1] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, Mar. 2018. [Online]. Available: http://link.springer.com/10.1007/s10772-018-9491-z

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 2020, arXiv:2006.11477 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2006.11477

[3] L. Maison and Y. Estève, "Some voices are too common: Building fair speech recognition systems using the Common Voice dataset," Jun. 2023, arXiv:2306.03773 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2306.03773

[4] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313618/

[5] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[6] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, no. 1, pp. 1–1, 1990.

[7] G. Liu, S. Cai, and C. Wang, "Speech emotion recognition based on emotion perception," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 22, May 2023. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-023-00289-4

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[9] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[10] R. Jürgens, M. Drolet, R. Pirow, E. Scheiner, and J. Fischer, "Encoding conditions affect recognition of vocally expressed emotions across cultures," *Frontiers in psychology*, vol. 4, p. 111, 2013.

[11] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[12] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9383000/

[13] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23 745–23 812, Jul. 2021. [Online]. Available: https://link.springer.com/10.1007/s11042-020-09874-7

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023, arXiv:1706.03762 [cs]. [Online]. Available: http://arxiv.org/abs/1706.03762

[15] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021, pp. 993–1003. [Online]. Available: https://aclanthology.org/2021.acl-long.80

[16] R. Rajoo and C. C. Aun, "Influences of languages in speech emotion recognition: A comparative study using Malay, English and Mandarin languages," in *2016 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. Penang, Malaysia: IEEE, May 2016, pp. 35–39. [Online]. Available: http://ieeexplore.ieee.org/document/7575033/

Appendix A Codes

Fine-tuning on Kaggle: https://github.com/RootReturn0/SER

Jupyter Notebook used in assessment: https://version.aalto.fi/gitlab/raudf1/Speech_Recognition_Project_Code_Nora/

Clustering and other classification models: https://github.com/pal-priyanshi/SER_SR