Transfer learning in audio-based emotion detection: surprising generalizability and limitations

Stanford CS224N Custom Project

Shunyu Yao
Stanford Institute for Theoretical Physics
Stanford University
alfred97@stanford.edu

Abstract

Emotion detection from audio data appears in many real world senario, and ususally it contains more information than its textual transcript. Due to lack of high quality supervised data, in this work, we focus on tranferring pre-trained, or even pre-fined-tuned model on audio emotion detection task with limited amount of data. Surprisingly, a feature extractor on speech-to-text task, after fine tuning on our task, shows a great performance and generalizability. This might imply some knowledge from speech-to-text gets transferred to emotion detection task. We also discuss limitations on transferring the knowledge to different language and other future directions.

1 Key Information to include

- Mentor: Chaofei Fan
- External Collaborators (if you have any): None
- Sharing project: None

2 Introduction

Emotion detection based on audio data has been boosted by the development of deep learning tools recentlyBertero and Fung (2017); Badshah et al. (2017); Schuller et al. (2004); Singh et al. (2023). It appears in many real life senario and contains more information than its textual counterparts. Normally, the classifier need two steps, first a feature extractor to transform raw data into a version that can be processed. The second part will be a neural network classifier. On one hand, the richness of raw data makes the task harder, because it will need a well-designed feature extractor to get as much essential information as possible from raw data. On the other hand, this means we could extract more information from raw data to gain better classification performance compare to only textural data.

There are a lot of well-designed feature extractor from information processing literature Aldeneh and Provost (2017); Mohan et al. (2023). However, due to a rising development of pre-trained model, we could use a feature extractor from a pre-trained modelBaevski et al. (2020); Conneau et al. (2020), and fine-tuned it on our taskPepino et al. (2021). However, the performance of the model can be highly constraint by limited amount of supervised data. In this work, we imagine the following senario where we have enough data on a different task, here the speech-to-text task, to pre-finetune the model. And surprisingly, that pre-finetuned model(we will call it **finer model**) outperform on vannilla pre-trained model(we will call it **vanilla model**) on emotion detecction task with limited amount of data.

We also try to further test the transfer leanring abilities to different language, where we imagine a senario that we have a well trained model for a language with sufficient amount of data(English), but try to further fine-tune on a much smaller dataset(Chinese). The performance happen to be heavily depends on whether we have access to previous training set(English).

3 Related Work

Emotion detection from audio data is a task that has been widely studied in the fieldBertero and Fung (2017); Badshah et al. (2017); Aldeneh and Provost (2017). More specifically, strategies based on pre-trained modelPepino et al. (2021); Chen and Rudnicky (2023) arised recent years due to various audio pretrained modelsBaevski et al. (2020); Conneau et al. (2020).

In this work, we will compare our result with a baseline model(although as we will explain later, the task in this work is different in detail) provided in the ESD dataset paperZhou et al. (2022), where they use the OpensmileEyben et al. (2010) as the feature extractor and a LSTM as the classifier.

3.1 credit claim

The code of the audio based model is modified based on a colab notebookm3hrdadfi (2021), credit should be given to the original author for most part of the code. Besides some details, I modified the trainer in order to enable the adaptive learning rate and early stop function. Extra dataloader was also created. Based heavily on this notebook, I also wrote the code for the text based classifier. GenAI tools were used during coding.

The finer model was based on a pre-finetuned model by Grosman (2021), in fact, his model parameter is already very close to the best model from experimental data.

4 Approach

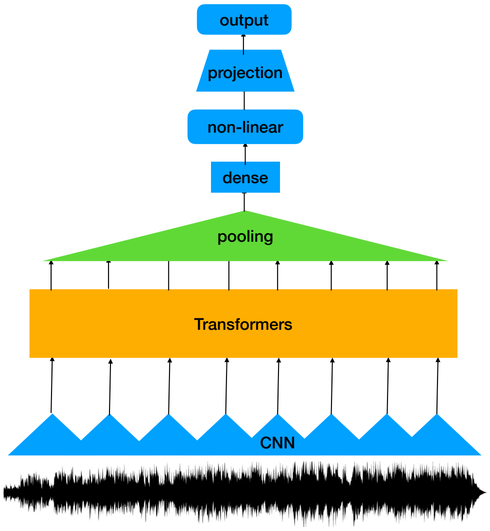

Figure 1: The structure for our feauture extractor+classifier

For the audio emotion detection task, we first need a feature extractor to transform raw data to vectors that can be fed into the classifier. For which, we use a pre-trained model xlsr-53Conneau et al. (2020). This is a multi-language transformer model based on previous model Wav2Vec2.0Baevski et al. (2020). The architect is first a convolution layer taking 16kHz audio data into a latent layer $\mathbf{z_t}$ then undergoes a quantization layer $\mathbf{q_t}$. Afterwards, feed into the transformer layer to output contextual vector $\mathbf{c_t}$. Each $\mathbf{c_t}$ represents a 25ms signal with a stride of 20ms. This model is pretrained on Commonvoice, Babel and multilingual Librispeech dataset. [1]

[1] As we mentioned before, the **finer model** is pre-finetuned on a speech-to-text task.

We connect the output of this feature extractor, after mean pooling, with a classifier. This classifier has a linear layer connected to a non-linear layer with tanh function and finally a projection layer(with dropout) to output logits for multilabel classification. We are allowed to fine-tune all parameters, however strategies on fine-tuning matters for the final result, which we will discuss in more detail in the experiments section.

As a parallel cross check, we also trained a text-based emotion detection model, where we keep the classifier to be the same, but change the feature extractor into a pretrained multi-language BERT model.

5 Experiments

5.1 Data

We employ the Emotion Speech DatasetZhou et al. (2022). This dataset include audio data spoken by 10 native English speaker and 10 native Mandarin speakers, with 5 emotional states(angry, sad, happy, surprise and neutral). For each speaker and each emotion, there are 350 examples. Official transcripts are provided for all these data.

For our purpose, we only use data from 5 English speakers and 2 from Chinese speakers. For most part of the work, we will use data from 4 English speakers as our training and validation set, and 1 English speaker as the test set. For transfer learning to Chinese part, we only take one speaker as the training and validation set, while the other one as the test set.

For the purpose of explaining the importance of audio data compare to just its transcripts, we also trained a text based model on the transcripts of the same data set for English, and shows a much worse performance.

We should also mention the **finer model** was pre-finetunedGrosman (2021) on the CommonVoice6.1 datasetArdila et al. (2020). This is a dataset consists of 7335 validated hours of audio data.

5.2 Evaluation method

We evaluate our results use several different metrics. For each emotion, we evaluate its precision, recall and f1-score. The definition for precision is the ratio of right classification(or true positive) among all positive labels predicted under that class(true positive+ fake positive), in short:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Similarly, recall denotes the ratio of true positive among all example that should be positive under such class(true positive + fake negative)

$$\text{Precision} = \frac{TP}{TP + FN} \tag{2}$$

The f1 score is the Harmonic mean of these two. We also present the mean value of these scores.

Besides this, we also present the accuracy of all classes, which is defined by total number of correctly predicted instances divided by total number of instances

$$\text{Precision} = \frac{TP + TN}{ALL} \tag{3}$$

This is actually the loss function we used during training as well, and also we are going to compare this with the baseline provided in Zhou et al. (2022).

For qualitative visualization, we will also present the confusion matrix for this multilabel classification problem.

5.3 Experimental details

For the first part of our experiments, we compare model fine-tuned on a pre fine-tuned model(**finer model**) to fine-tuning on a vanilla pre-trained model(**vanilla model**). For the finer model, we found the best performance is achieved via first freeze the parameters in the feature extractor and only train the classifier for 1 epoch, then unfreeze the feature extractor and train on the all parameters for 1 epoch. For both part, we choose a batch size of 8, and an initial learning rate of 10^{-4}. Adaptive learning rate is implemented, so whenever the loss on evaluation set increase for more than two rounds(one rounds includes 10 steps), the learning rate will multiplied by 0.1.

For the vanilla model, we fine-tuned with a start learning rate 10^{-4} and batch size 8. After 2 epoches, we manually decrease the learning rate by a factor of 0.1 and increase the batch size to 10 to smooth out the learning curve, and fine-tune for another 2 epoches.

To present the feature contained solely in textual data, we also trained a text based classifier, whose feature extractor from BERT and classifier are trained simulatanously.

For transferring to Chinese, we choose two different procedure. For senario where we do not have initial dataset on English, we fine-tune our **finer model** on the Chinese speaker dataset with learning rate 10^{-4} and batch size 8. For senario where we do have access to initial dataset on English, we further fine tuned the **finer model** on all 4 English speakers and 1 Chinese speaker.

5.4 Results

5.4.1 results for fine-tuning on test sets

Here we present the **finer model**, which performs best on our task. The result on the test set is

	Precision	Recall	F1-score	Support
Angry	0.99	0.92	0.96	357
Happy	0.98	0.84	0.90	350
Neutral	0.86	1.00	0.93	353
Sad	1.00	0.91	0.95	350
Surprise	0.87	1.00	0.93	350
Accuracy			0.93	1760
Macro avg	0.94	0.93	0.93	1760
Weighted avg	0.94	0.93	0.93	1760

Table 1: The result for finer model on test set

and the corresponding confusion matrix is:

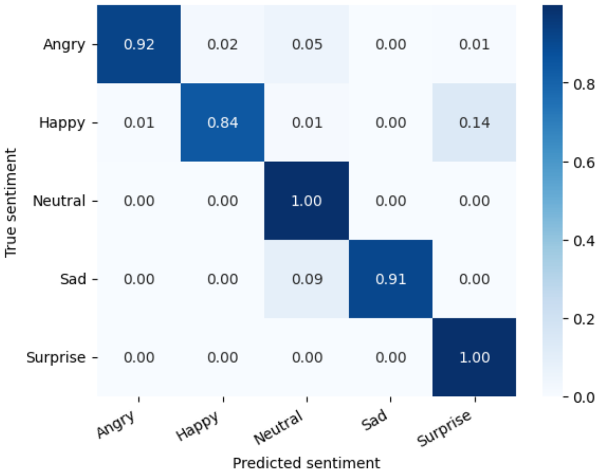

Figure 2: The confusion matrix for finer model

For reader who want to know the performance on evaluation set, see appendixA.

As a comparasion, we show the same thing for the vannila model

	Precision	Recall	F1-Score	Support
Angry	1.00	0.59	0.74	357
Happy	0.98	0.90	0.94	350
Neutral	0.50	1.00	0.66	353
Sad	1.00	0.44	0.61	350
Surprise	0.95	0.98	0.96	350
Accuracy			0.78	1760
Macro Avg	0.88	0.78	0.78	1760
Weighted Avg	0.88	0.78	0.78	1760

Table 2: result for vanilla model on the test set

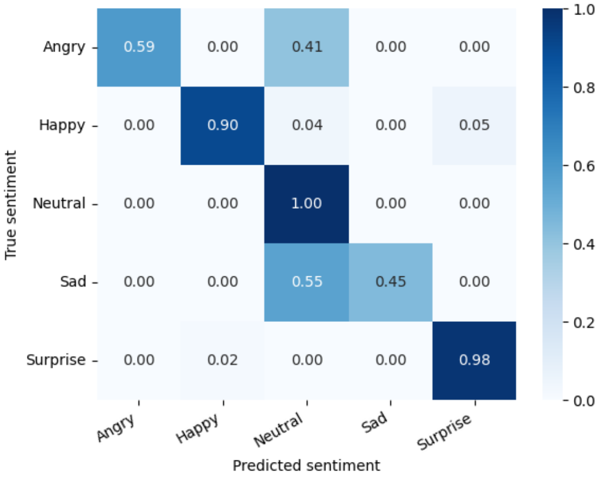

Figure 3: The confusion matrix for vanilla model

one can see the finer model significantly out-perform the vanilla model, in generalizing the emotion detection from example from four speakers to other unseen speaker. For more analysis and discussion, see next section.

As a comparasion to our baseline provided in original ESD paperZhou et al. (2022). The accuracy they got for English dataset is 0.89. However, we would like to mention that their training is done on all 10 speakers, so all test data are coming from speakers the model has **already seen** during training. For our experiments, the test data is coming from speaker that has never been seen by the model.

5.4.2 transfer learning to the Chinese dataset

We now try to transfer what the model learnt from English to Chinese. As a controlling baseline, we show our finer model without training on Chinese and test it on Chinese dataset

	Precision	**Recall**	**F1-score**	**Support**
Angry	1.00	0.02	0.04	350
Happy	0.32	0.91	0.48	350
Neutral	0.00	0.00	0.00	350
Sad	0.40	0.64	0.49	350
Surprise	0.08	0.05	0.06	350
Accuracy			0.32	1750
Macro avg	0.36	0.32	0.21	1750
Weighted avg	0.36	0.32	0.21	1750

Table 3: The result for finer model on the Chinese test set without any dine-tuning

One can see, certain amount of knowledge get transferred to Chinese dataset, but not a lot. Since we have 5 different emotions, a random guess will get a accuracy of 0.20, while without fine-tuning on Supervised Chinese dataset, we get 0.32, which is only slightly better than random guess.

For senario one where we only have access to new data(Chinese), if we further fine tune on that Chinese speaker, we will unfortunately see very little progress[2]

	Precision	**Recall**	**F1-score**	**Support**
Angry	1.00	0.02	0.04	350
Happy	0.32	0.91	0.48	350
Neutral	0.00	0.00	0.00	350
Sad	0.40	0.65	0.49	350
Surprise	0.09	0.05	0.06	350
Accuracy			0.33	1750
Macro avg	0.36	0.33	0.21	1750
Weighted avg	0.36	0.33	0.21	1750

Table 4: Fine tuning the finer model more on the Chinese dataset, without access to the English training data

This is quite unexpected, however, for senario two, which means if we are allowed to fine tune on both English and Chinese dataset, we can see the it can significantly improve the result on Chinese dataset[3]

	Precision	**Recall**	**F1-score**	**Support**
Angry	1.00	0.45	0.62	350
Happy	0.47	0.97	0.63	350
Neutral	0.99	0.46	0.63	350
Sad	0.59	0.99	0.74	350
Surprise	0.32	0.12	0.17	350
Accuracy			0.60	1750
Macro avg	0.68	0.60	0.56	1750
Weighted avg	0.68	0.60	0.56	1750

Table 5: Fine tuning the finer model more on the Chinese dataset, with access to the English training data

6 Analysis

For first part of our work, we can see from the experiment results that the performance of the finer model is better than the vanilla model in almost all aspects, except for "happy". I would interpretate this result as not only the performance is better, but also a level of generalizability. This is because in our experiment, training data(and validation data) are coming form distinct speakers compare to test data. As one can see through appendixA, both model have comparable performance on evaluation data. However, when coming to test data with a new speaker, finer model completely out-performs.

Another point to clarify is, one might think the fine-tuning on vanilla model might not be enough, so we can achieve higher performance if we do enough fine-tuning. However, we could give a argument based on the performance of the baseline model, that model is trained on more data, and dealing with a easier task(identifying emotion from a speaker **it has seen**), and that model only achieve a accuracy of 0.89. We would suspect any improvement on fine-tuning the vanilla model can not exceed this.

One interesting thing to understand is why this model pretrained on text to speech task could perform that well. One could suspect if that is because most of the emotional knowledge for our dataset is actually in the textual data. However, we can test that for our dataset because we have official

[2] however the result on validation set is pretty good, so I will interpret this result as saying the model do not have generalizability on Chinese dataset.

[3] however the performance on English training set decreased, we show numerical results in appendixB

transcript for the dataset, and we trained a text based emotion detector, one can see these data are actually featureless:

	Precision	Recall	F1-score	Support
Angry	0.20	0.25	0.22	350
Happy	0.00	0.00	0.00	350
Neutral	0.20	0.75	0.31	350
Sad	0.20	0.00	0.01	350
Surprise	0.00	0.00	0.00	350
Accuracy			0.20	1750
Macro avg	0.12	0.20	0.11	1750
Weighted avg	0.12	0.20	0.11	1750

Table 6: Classification Report

To justify this is not because our text based model is not good, we trained our text based model on other dataset, where reasonable performance is shown in appendixC.

7 Conclusion

We have two main results in this project:

1. Knowledge from speech-to-text supervised data can transfer to speech-emotion-detection task, and help improving performance and generalizability.

2. Knowledge from English only generalize a little to Chinese. When fine-tuning on Chinese data, whether or not we have access to original training data from English matters for generalizability.

The first feature we observed is very surprising to me. As we explained in the main text, this two tasks has no direct relation to each other. This is extremely true for the dataset we choose. However, having seen more supervised data from other speakers seems to help. Are these coincidence, or there is some reason behind this? We leave these for future work, where we plan to test on more datasets and also, output the attention scores to study qualitatively why this happens.

There are quite a few limitations for current work, including a relatively small dataset, and a specific pre-finetuned model. All these need to be double-checked in the future to justify the correctness of this work.

8 Ethics Statement

One risk related to transfer learning is, there might be bias comming from previous training set. As an example, in our model we transfer Speech-to-text to Speech-emotion-detection. But there might be some word, for example, great, that as a text is more positive. But in audio siuition, it could be satire depending on the tone. These cases might only appears a few times so one can not identify them during training and validation. But it might cause issue, for example, if people is going to use this classifier to identify hateful words, then they will miss these due to the bias coming from STT task. And since we do not have access to those dataset, it will be hard to identitfy.

Another social risk is if people want to use this classifier for medical purpose, then even if the correctness is more than 90 percent, there are still finite chance to make a wrong diagnose. This could be important for medical purpose.

One way we can address this risk is instead of let the model output just one result on the prediction, we let it exam the confidence on that result. For example, if the probability of that result is lower than 80 percent, we will let the model output a warning sign, and ask human to do further investigation.

References

Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition. In *2017 IEEE international conference on acoustics, speech and signal processing*

(ICASSP), pages 2741–2745. IEEE.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. 2017. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Dario Bertero and Pascale Fung. 2017. A first look into a convolutional neural network for speech emotion detection. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5115–5119. IEEE.

Li-Wei Chen and Alexander Rudnicky. 2023. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Jonatas Grosman. 2021. Fine-tuned XLSR-53 large model for speech recognition in English. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`.

m3hrdadfi. 2021. Emotion recognition in greek speech using wav2vec2. `https://github.com/m3hrdadfi/soxan/blob/main/notebooks/Emotion_recognition_in_Greek_speech_using_Wav2Vec2.ipynb`.

Meera Mohan, P Dhanalakshmi, and R Satheesh Kumar. 2023. Speech emotion classification using ensemble models with mfcc. *Procedia Computer Science*, 218:1857–1868.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE international conference on acoustics, speech, and signal processing*, volume 1, pages I–577. IEEE.

Jagjeet Singh, Lakshmi Babu Saheer, and Oliver Faust. 2023. Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 20(6):5140.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

A The results for fine-tuned model on evaluation set

The result for **finer model** on evaluation set is

	Precision	**Recall**	**F1-score**	**Support**
Angry	0.98	0.96	0.97	140
Happy	0.96	0.97	0.96	140
Neutral	0.95	0.99	0.97	140
Sad	0.99	1.00	1.00	140
Surprise	0.99	0.96	0.97	140
Accuracy			0.97	700
Macro avg	0.97	0.97	0.97	700
Weighted avg	0.97	0.97	0.97	700

Table 7: The result for finer model on the evaluation set

The result for **vanilla model** on evaluation set is

	Precision	Recall	F1-Score	Support
Angry	0.99	0.99	0.99	140
Happy	0.98	0.96	0.97	140
Neutral	0.97	1.00	0.99	140
Sad	1.00	1.00	1.00	140
Surprise	0.99	0.96	0.97	140
Accuracy			0.98	700
Macro Avg	0.98	0.98	0.98	700
Weighted Avg	0.98	0.98	0.98	700

Table 8: The result for vanilla model on evaluation set

One can see they are comparable to each other, and the vanilla model looks even better. Thus giving more justification that the generalizability of the finer model is better.

B The performance for the English set

	Precision	**Recall**	**F1-score**	**Support**
Angry	1.00	0.96	0.98	357
Happy	0.99	0.86	0.92	350
Neutral	0.66	1.00	0.80	353
Sad	0.99	0.55	0.71	350
Surprise	0.90	1.00	0.95	350
Accuracy			0.87	1760
Macro avg	0.91	0.87	0.87	1760
Weighted avg	0.91	0.87	0.87	1760

Table 9: The Chinese model test on English dataset

One can the performance on English test set decreased after fine-tuning on Chinese task. Is this a general feature or it will go away if we have a large enough model?

C test of our text based model on different dataset

In this appendix, we are going to test our text based model on an extra dataset[4]. The training set contains 16000 examples with six emotions: ['anger', 'fear', 'joy', 'love', 'sadness', 'surprise']. The test set contains 2000 examples. And the result for the validation set and test set are:

This appendix has nothing to do with the experiments in the main text, the only purpose is to show the text based model is working, and the textual data for ESD dataset is really featureless.

[4] This dataset is downloaded from Kaggle: https://www.kaggle.com/datasets/chandrug/textemotiondetection/data

	Precision	**Recall**	**F1-score**	**Support**
Anger	0.87	0.90	0.89	216
Fear	0.84	0.85	0.85	194
Joy	0.95	0.93	0.94	536
Love	0.84	0.87	0.86	130
Sadness	0.93	0.93	0.93	467
Surprise	0.78	0.74	0.76	57
Accuracy			0.90	1600
Macro avg	0.87	0.87	0.87	1600
Weighted avg	0.90	0.90	0.90	1600

Table 10: On validation set

	Precision	**Recall**	**F1-score**	**Support**
Anger	0.88	0.87	0.88	275
Fear	0.84	0.85	0.85	224
Joy	0.92	0.91	0.91	695
Love	0.76	0.75	0.75	159
Sadness	0.91	0.93	0.92	581
Surprise	0.73	0.73	0.73	66
Accuracy			0.88	2000
Macro avg	0.84	0.84	0.84	2000
Weighted avg	0.88	0.88	0.88	2000

Table 11: On test set