Examining the Relationship Between Emotion Recognition Models and Language Properties

Veena Kommu, Randy Gu

June 15, 2023

Abstract

We wanted to investigate the relationship between language properties and the capability of a model trained on language to produce accurate results with other language data. We explored this relationship using a speech transformer with a HUBERT model architecture. We then trained our model on an English language database, and tested accuracy on a German and Urdu database to see how well our model could generalize across different language branches/subfamilies. Our results showed that there was no significant relationship between language properties and model generalization.

1 Group

Veena Kommu
First-Year Undergraduate CSE student
Contact at vkommu@g.ucla.edu
UID: 105975943

Randy Gu
Undergraduate CS student
Contact at randygu@g.ucla.edu
UID:305592076

2 Introduction and Motivation

The goal of our project was to build an emotion recognition model that would generalize well across different languages. To better understand the role language plays in emotion expression, we checked to see if there was any helpful linguistic relationship between language and emotion expression. For this, we used the concept of language branches. Linguists group languages into separate branches based on shared features in their structure or history. Such features include compound word usage, word order, and more. We hypothesized that because of this similarity in language structure, models trained on languages under some language branch will generalize well to other languages in that same branch.

3 Data

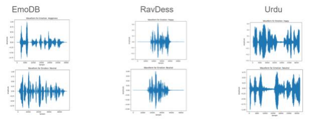


Figure 1: Some Waveforms from Each Dataset

In our experiments, we used three Speech Emotion Recognition(SER) datasets: RAVDESS, EMO, and the Urdu databases. RAVDESS is an English dataset of 1440 audio files from 24 actors classifying between 8 different emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each expression was produced at an emotional intensity of normal and strong, and an additional neutral expression is included. EmoDB is a German dataset consisting of 535 audio files from 10 speakers, classifying anger, boredom, anxiety, happiness, sadness, disgust, and neutral emotions. The Urdu dataset consists of 400 audio recordings from 38 speakers in talk shows with 4 emotions Appy, Happy, Neutral, and Emotion. Between all three datasets the common emotions were Angry, Happy, and Neutral.

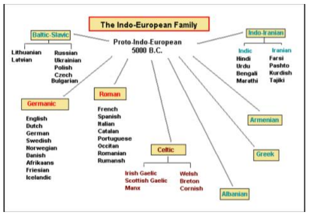
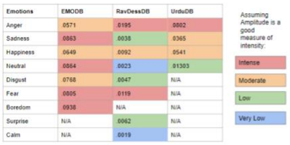


Figure 2: Language Family Model

The English and German languages belonged to the same linguistic subfamily: Germanic. Urdu falls under the Indo-Iranian branch. All three languages belong to the Indo-European family.



Emotions	EMODB	RavDessDB	UrduDB
Anger	.0571	.0195	.0802
Sadness	.0863	.0038	.0365
Happiness	.0649	.0092	.0541
Neutral	.0864	.0023	.01303
Disgust	.0768	.0047	N/A
Fear	.0805	.0119	N/A
Boredom	.0938	N/A	N/A
Surprise	N/A	.0062	N/A
Calm	N/A	.0019	N/A

Assuming Amplitude is a good measure of intensity:

🟥 Intense

🟧 Moderate

🟩 Low

🟦 Very Low

Figure 3: Average Waveform Amplitude Across Different Emotions

One of the statistics that we extracted from the data was amplitude. Assuming that amplitude is a good measure of the intensity of emotion, we were able to see that Anger and Fear generally had a higher intensity across all data-sets, and happiness was moderate. We were not able to see significant differences between the language branch data.

4 Experimental Setup

The model architecture we use for experiments is a speech transformer with the same architecture as the base HuBERT model. Specifically, the model has a 7-layer CNN encoder, a 12-layer BERT-like transformer encoder, and finally we add an additional output layer specific for our emotion recognition classification task. All experiment comparisons will be based on this model. To compare the effects of how different language emotions transfer each other, we have three different setups. First, we have the baseline model with the HuBERT pretrained weights (pre-trained for speech representation) trained on the target language dataset. Then, as an additional source of comparison to see the impact of speech learning, we train on the target dataset entirely from scratch without the pretrained weights. Finally, our transfer learning model would be the HuBERT pretrained model fine-tuned on the RAVDESS english dataset, then tested on the target dataset. For training of all models, we used a mini-batch size of 4, the Adam optimizer with default parameters except a learning rate of 0.0001, and training loss is computed from the cross-entropy loss of the classification. The input data are raw audio waveforms sampled at 16kHz, as this is the rate at which HuBERT is trained on. To see exactly how well the transfer model generalizes to emotion recognition for a different language, we purposefully limited the target training data size to 20% of the original dataset size. This way, the model won't be able to overfit to the target language and the impact of the transfer learning can be properly expressed. The training process and the three comparisons are done over the two target datasets (Urdu and German) in the same manner in order to compare how well the English training transfers to a language from the same linguistic family or an entirely unrelated language. There is also some additional experimentation where the source training language is Urdu and results are tested on the English data.

5 Results

We use classification accuracy as our evaluating metric.

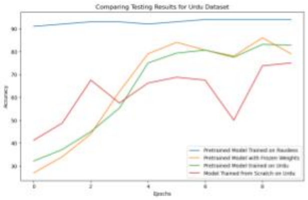
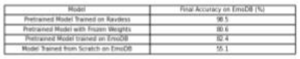


Figure 4: Urdu Results



Model	Final Accuracy on EmoDB (%)
Pretrained Model Trained on Ravdess	98.5
Pretrained Model with Frozen Weights	80.6
Pretrained Model trained on EmoDB	82.4
Model Trained from Scratch on EmoDB	55.1

Figure 5: German Results

Additional Experimentation Results:
With Urdu only having 4 emotion labels and far less data than the English data-set, we achieved a testing accuracy (80% of data-set) of 65%.

6 Conclusion

From these results, it is hard to come to any conclusive generalizations about the relationship between languages. While the German dataset achieved an approximately 5% better accuracy, it is not enough to say that English transferred better to German than compared to Urdu. There could be many factors affecting the accuracy improvement, such as similarity of emotion labels, or the data itself is just of a higher quality. However, there are some interesting observations to be drawn from the results. From the significant improvement between the pretrained model and the transfer model from English for both target languages, it can be said that emotion recognition generalizes very well across languages in the general sense. The relationship between languages doesn't play that big of a factor, as Urdu also significantly improved in accuracy with the transfer of learning from English. On the contrary, the speech representation learning from HuBERT itself was not sufficient enough to generalize well in the target language, which could either be due to a difference in the language HuBERT was trained on, or just the lack of training data in general. Either way, it can be said that emotion recognition in speech as a task itself is generally language agnostic. Since usually data for SER tasks are very expensive to collect, most benchmark datasets are limited in size and all vary in the way they're collected (language, sampling rate, modality, labels,etc). However, observations from our results show that utilizing different datasets actually may improve generalization of SER tasks significantly rather than hinder performance.

7 Challenges and Future Prospects

Data-set availability was a huge challenge for us. Currently, we were only able to find publicly available data for languages in the same family, but under different branches. While our results showed weak correlation between language structure and emotion classification, we are unsure if this generalization would hold true across different language families, which have more drastic differences in language structure. We also could not find a data-set other than Urdu that was in the Indo-Iranian subfamily. If we could, to better support our results, we would have liked to test the same process by finetuning on data from the Indo-Iranian subfamily then checking the results on the other three data-sets.

In addition, originally we hoped to test the data on a CNN with two different forms of data: the waveform itself and then spectrograms, to show that any possible relationship we saw wasn't completely architecture dependent. However, we struggled to find a relevant pretrained model to use for our task that had weights or data publicly accessible, so regardless of the data-set we were unable to get a good accuracy and extract meaningful results.

8 References

[1] M. E. A. ElShaer, S. Wisdom, and T. Mishra, "Transfer learning from sound representations for anger detection in speech," arXiv.org, https://arxiv.org/abs/1902.02120

[2] Phukan, O. C., Buduru, A. B., amp; Sharma, R. (2023, April 22). A comparative study of pre-trained speech and audio embeddings for speech emotion recognition. arXiv.org. https://arxiv.org/abs/2304.11472

[3] Ram, S., amp; Aldarmaki, H. (2023, January 3). Supervised acoustic embeddings and their transferability across languages. arXiv.org. https://arxiv.org/abs/2301.01020

[4] Wei-Ning Hsu, B. Bolte, Y. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," https://arxiv.org/abs/2106.07447

[5] Zhang, S., Liu, R., Tao, X., amp; Zhao, X. (2021, November 29). Deep Cross-corpus speech emotion recognition: Recent advances and perspectives. Frontiers in neurorobotics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8666588/

[6] Zhou, S., amp; Beigi, H. (2020, August 15). A transfer learning method for speech emotion recognition from automatic speech recognition. arXiv.org. https://arxiv.org/abs/2008.02863v2

Datasets: [1] https://github.com/alantanlc/torchemotion
[2] https://www.kaggle.com/datasets/bitlord/urdu-language-speech-dataset
[3] https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb
[4] https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio