Cross-Lingual Speech Emotion Recognition using Unsupervised Domain Adaptation

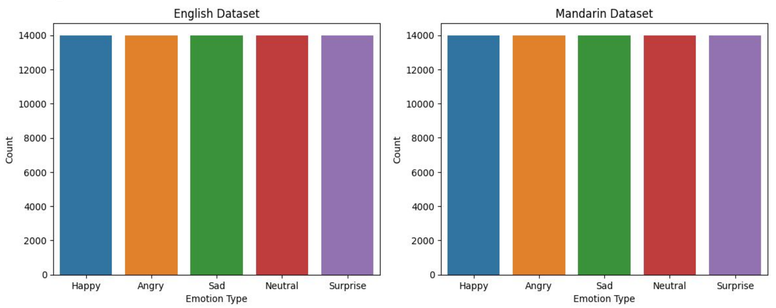
Purushothaman Saravanan[1], Iype Eldho[2], Grace Vincent[3]
psarava3@ncsu.edu[1], ieldho@ncsu.edu[2], gmvincen@ncsu.edu[3]
ECE 542-001: Neural Networks

Background

Cross-lingual speech emotion recognition poses unique challenges due to variations in language, culture, and expression. The task requires understanding emotional states conveyed in speech and translating them across different languages. Traditional supervised models struggle to generalize well due to the lack of labeled data in different languages. Unsupervised domain adaptation techniques can overcome this challenge by adapting models to new languages using only unlabeled data. In this study, we explore the effectiveness of UDA techniques for cross-lingual speech emotion recognition between English and Mandarin. We focus on five emotions: happy, sad, angry, neutral, and surprise, and evaluate our models on a dataset of speech recordings from both languages.

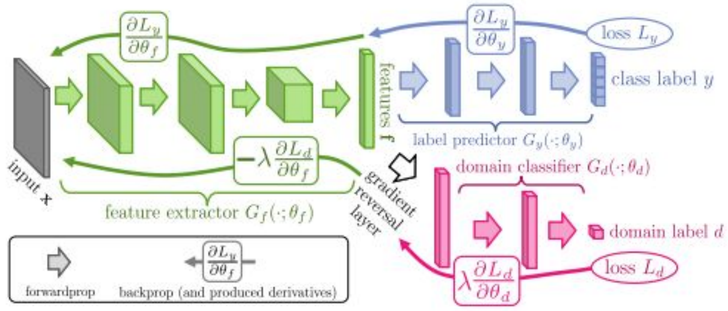
Dataset

- The dataset contains 350 parallel utterances spoken by 10 native Mandarin speakers, and 10 English speakers with 5 emotional states (neutral, happy, angry, sad and surprise). All the speech data are samples at 16 kHz and saved in 16 bits.
- Audio files were preprocessed by extracting features:
  - **Zero crossing rate:** the rate at which a signal transitions from positive to zero to negative or negative to zero to positive
  - **MFCC**: describes the overall shape of a spectral envelope
  - **Melspectrogram**: The mel scale is a scale of pitches that human hearing generally perceives to be equidistant from each other. As frequency increases, the interval, in hertz, between mel scale values (or simply mels) increases. The name mel derives from melody and indicates that the scale is based on the comparison between pitches. The mel spectrogram remaps the values in hertz to the mel scale.
  - **Spectral contrast:** considers the spectral peak, the spectral valley, and their difference in each frequency subband.
- Figure 1 shows the distribution of both the datasets

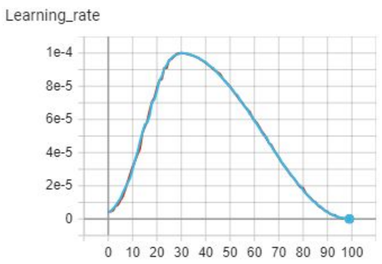


Methods

- We have decided to train a Domain Adversarial Neural Network, (DANN) [1] for this UDA problem.
- DANN is a deep learning technique that aims to learn domain-invariant features from source and target domains by using a gradient reversal layer that maximizes the discrepancy between the two domains.
- We experimented with different backbone feature extraction methods for the two classification heads.Including a ResNet-50 encoder and a 1D Convolutional Neural Network.
- The DANN architecture is depicted in Figure 2.



- Hyperparameter tuning was conducted to attempt to optimize the performance of the models. We tuned parameters such as optimizer, learning rates, and regularization strength to achieve the best results. A learning rate scheduler was used to improve convergence and accuracy (Fig 3)..



Mandarin speaker - Angry

English speaker - Happy

English speaker - Neutral

References
[1] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," J. Mach. Learn. Res., vol. 17, no. 1, p. 2096–2030, jan 2016

Results

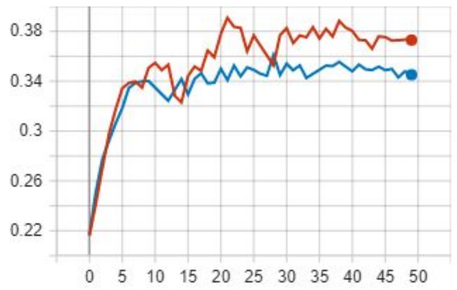
- We note that despite hyperparameter tuning and adjusting the domain classifier DANN has low performance in the target class label predictor.
- The ResNet-50 encoder does, however, improve results over the 1D Convolutional encoder.
- Our table of results also highlights that a domain shift from Mandarin to English has lower performance than vice-versa.

Model	Encoder	English to Mandarin		Mandarin to English	
		Acc	F1	Acc	F1
DANN	ResNet	0.373	0.375	0.345	0.346
DANN	1d Conv	0.357	0.59	0.334	0.330
DANN (Supervised)	1d Conv	0.974	0.974	0.951	0.951

Conclusions

- Cross-lingual speech emotion recognition has growing importance as globalization drives multi-lingual societies and a need for language learning.
- We will need to further update our domain classification head to further improve the model.
- Future work includes potentially developing a GAN model for cross-lingual speech recognition

Target Domain Accuracy



Target Domain F1-Macro