11-785 Final Report: Domain Adaptation for Accented Speech

Anam Iqbal
Carnegie Mellon University
anami@andrew.cmu.edu

Faizan Khan
Carnegie Mellon University
fakhan@andrew.cmu.edu

Rosanna Vitiello
Language Technologies Institute
Carnegie Mellon University
rvitiell@andrew.cmu.edu

Justin Yoo
Carnegie Mellon University
justinyoo@andrew.cmu.edu

Abstract

Transfer learning has become an important tool to use in various domains of Artificial Intelligence to reproduce results on problems with less or inconsistent data. Domain adaptation is a subset of transfer learning, that deals with different data distributions within the same task. The aim of this paper is to transfer a Speech Emotion Recognition (SER) model from standard English to Singaporean-accented English. We report on baselines and transfer experiments for the speech emotion recognition task on the MSP-Podcast and NSC speech datasets. These results set a foundation for future experiments to better understand how to use domain adaptation to increase performance in accented speech domains. We include code from this work via the following GitHub repository: https://github.com/rosavitiello/11-785-Fall2022-Project

1 Introduction

Speech Emotion Recognition (SER) is a task that aims to detect human emotions from audio. SER has many practical applications in human-machine interaction, especially in tasks where the interaction only happens through voice such as conversational AIs and call centers (Bromuri et al., 2020). However, Speech Emotion Recognition suffers from the problem of lacking large quantities of reliable data with many small datasets differing on labeling schemes, types of emotions annotated, and whether the emotion was simulated or from the natural conversation, leading many SER research to focus on one dataset (Milner et al., 2019). There have been many studies on cross-language(Agarla et al., 2022), cross corpus(Milner et al., 2019) transfer learning, or different data augmentation techniques (Qu et al., 2022) to solve this problem. However, there has been a lack of focus on building a comprehensive model of a single language with multiple accents, which we believe is paramount to building a sufficient working model in English.

With over 1.35 billion English speakers around the world and 67 countries with English as the official language, English is one of the most popular languages in the world. However, this pervasiveness also leads to a large number of accents in the English language. Accents have a non-negligible effect on how other people interpret one's speech. Based on different accents, studies have shown that a listener's response time changes significantly due to issues with comprehension (Valles, 2015) and that differently accented words can even change how others perceive emotion in speech (Seppi et al., 2010), (Sun et al., 2022). With these differences, it's likely that when training a model for SER, it is necessary to train the model with data from diverse English accents for it to perform robustly in real-world applications. However, with almost 160 different English accents and a lack of comprehensive datasets for even the most common accents, training a model in this manner is currently impractical. Looking at the effectiveness of Transfer Learning in Speech Recognition tasks (Dubey and Shah, 2022), our project experiments with transfer learning and domain adaptation techniques in order to create a model which performs well for both English and Accented English.

Transfer Learning is an idea where we use the knowledge gained in one task to help with a different but related task. Abstractly this comes from the idea that a person who has had experience playing the clarinet has an easier time learning to play the saxophone since some knowledge can be transferred between the tasks. As long as there is some connection between the two tasks, transfer learning allows the models to learn from a limited amount of data (Zhuang et al., 2019). Domain adaptation is a subcategory of transfer learning where we leverage a related similar model with a large amount of labeled data to create a model that can perform well across varying domain distributions. In the case of SER, this technique allows us to use Speech Recognition task datasets that are well labeled and diverse (Vergyri et al., 2010) to train an emotion recognition model. In our work, we apply domain adaptation with a well-labeled accented Speech Recognition dataset and compare this model to baselines created with naive transfer learning which is trained in an English SER dataset and tested on an accented SER dataset.

2 Related Work and Background

For this work, we performed an initial literature review to explore different aspects of speech recognition, emotion recognition, and transfer learning techniques for speech-based tasks. Emotion Recognition has been an ever-growing field that is being studied and explored more and more every year. In basic SER, solely the speech modality is used to identify emotions on fixed labels of angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral. However, recently it has become increasingly common to use multiple modalities for emotion recognition with approaches as explored first in Liu et al. (2016). More recent work in the domain is described in Franceschini et al. (2022).

There has been a significant development in the domain of speech recognition and some of the latest models such as XLSR (Babu et al., 2021b) and HuBERT (Hsu et al., 2021), and most recently, Whisper (Radford et al., 2022) are large models trained on multiple languages and large amounts of data. The idea of creating speech representations using techniques like wav2vec 2.0 (Baevski et al., 2020) are central to the ability of these models to perform well on multiple tasks and be “transferable”. The commonly used tasks to test these baselines are automatic speech recognition (ASR), language identification and machine translation.

In addition to this, there has also been a significant development in emotion recognition. In Mohamed and Aly (2021), it is explained how wav2vec 2.0 and HuBERT have been utilized for emotion recognition in dialogues from Arabic speech. In a similar vein, a more recent paper (Wang et al., 2021) detailed how models like wav2vec 2.0 and HuBERT have been used in emotion recognition with results contrasting between speaker-independent setting and speaker-dependent setting. It is found that there is 73.01% weighted accuracy in speaker-independent settings and 79.58% weighted accuracy in speaker-dependent settings for recognizing emotion on IEMOCAP as highlighted in Table 1. This approach is further demonstrated in (Wongpatikaseree et al., 2022) where there are two main divisions, namely, a front-end network with wav2vec 2.0 used for speech recognition and a back-end network with XLSR used for recognizing emotions in the speech.

Table 1: Weighted Accuracy in Percentage

Speaker-Independent	**Speaker-Dependent**
73.01	79.58

In terms of transfer learning and domain adaptation, techniques like adversarial training, feature projections into common spaces, domain invariant feature learning, and contrasting learning have been popular. Domain Adversarial Training (DAT) for the task of cross-corpus speech emotion recognition has been studied and explored in Milner et al. (2019). Techniques like Domain Invariant Feature Learning are explored in works like Lu et al. (2022). Another interesting approach of Contrastive Loss is shown in the paper, Franceschini et al. (2022). In addition, works like Yang et al.

(2021) explore self-supervised learning(SSL) on emotion recognition from multiple modes as well with some success.

With respect to Domain Adversarial Neural Networks, work has been done specifically with regards to finding distinct representations for emotion recognition in speech Gao et al. (2021) with the help of multi-head attention. The paper explains how this approach has helped in minimizing the discrepancy that exists between different domain distributions. Specifically, it mentions that the model beats the state-of-the-art as it performs better on the unweighted accuracy and improves the same by 4.15%. Similar to this, Li et al. (2021) also details work on a bi-hemisphere domain adversarial neural network which also focuses on improving the model performance by minimizing the discrepancy between domains, such as target versus source domains.

3 Data

For our experiments, we train and collect evaluation metrics on the MSP-Podcast database (Lotfian and Busso, 2019) and National Speech Corpus (Koh et al., 2019). In this section, we outline the qualities and characteristics of the dataset. We include a summarization of the subset we use of each dataset in Table 2.

3.1 MSP-Podcast (Lotfian and Busso, 2019)

In the field of speech emotion recognition, current benchmark datasets often suffer from limitations in size and speakers, inconsistent emotional descriptors, lack of naturalistic interactions, and unbalanced emotional content (Douglas-Cowie et al., 2003; Koolagudi and Rao, 2012). For example, popular approaches in creating speech emotion recognition datasets rely on the use of actors (Cao et al., 2014), in which emotions are often over-emphasized and tend to capture more prototypical behavior. Other databases that sample from naturalistic audio from various domains (McKeown et al., 2012; Ringeval et al., 2013) are often unbalanced and biased to context.

To address such limitations in speech emotion recognition datasets, Lotfian and Busso (2019) proposed the MSP-Podcast database. The dataset samples from naturalistic English podcast audio. To select candidate audio in a balanced manner, machine learning models were used to retrieve examples with target emotional behaviors. Each audio is cleaned and annotated with emotion labels (angry, sad, happy, surprised, fear, disgust, contempt, neutral, other) by Amazon Mechanical Turk. In total, the dataset contains over 100 hours from more than 80 speakers. For the purpose of this project, we limit MSP to five emotions: anger, sad, happy, disgust, and neutral. This subset reduces the number of train samples and validation samples to 30K and 5K.

3.2 National Speech Corpus (Koh et al., 2019)

Designed for speech recognition tasks, the National Speech Corpus (NSC) contains over 2000 hours of orthographically transcribed read speech data from 1,379 native accented Singaporean-English speakers. The dataset contains conditions of phonetically balanced scripts, scripts containing words pertinent to the Singapore context, and conversational speech. While NSC does not contain emotion labels, our project mentors provided emotion annotations for examples in this dataset.

Because training on over 2000 hours of data was infeasible for this project, we limit the training transcripts and audios to 20K examples. Additionally, as NSC is an automatic speech recognition dataset, emotion labels for audio are not provided. Consequently, we annotate 1700 examples with emotion labels for testing. To produce these annotations, each audio in the NSC data was passed through a speech emotion recognition model, and then manually validated by one annotator. In future work, we hope to acquire more annotated examples with multiple annotator agreement. Despite this limitation in size and potential bias, we believe that these annotations will show sufficient proof of concept for our techniques.

Table 2: Comparison of subsets data for domain adaptation in speech recognition

Dataset	Language	# of samples	Emotions represented	Year
MSP-Podcast	English	30K samples for emotion recognition	angry, sad, happy, disgust, neutral	2016
NSC	Accented Singaporean English	20K subset for ASR, 1769 examples for emotion recognition	angry, happy, sad, neutral (annotations provided by mentors)	2019

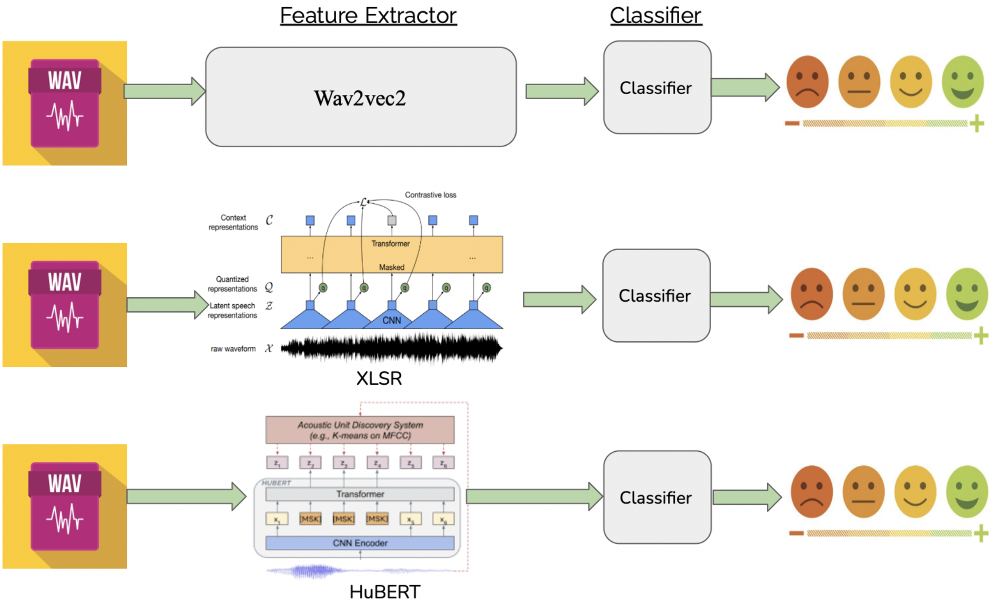


Figure 1: Baseline Architectures

4 Baselines and Proposed Methods

4.1 Transfer Learning: SER Baselines

For our baseline experiments, we adapted three state-of-the-art models in speech recognition for SER as shown in Figure 1. In each approach, we leverage the state-of-the-art pretrained architectures as feature extractors and finetune classification with a speech emotion recognition head. We train each baseline on the subset of English-accented MSP-Podcast train data for 10 epochs and evaluate on the English-accented MSP-Podcast validation and Singaporean-English accented NSC emotion test set.

In the following subsections, we detail each state-of-the-art architecture and our implementation of the baselines.

4.1.1 wav2vec 2.0 (Baevski et al., 2020) for Speech Emotion Recognition

wav2vec 2.0 is a framework for self-supervised learning of representations from raw speech data, and has popularly been used in transfer to a variety of downstream speech tasks. Trained on 53.2K hours of audio, wav2vec 2.0 learns to encode latent speech representations with contrastive tasks and masking techniques similar to masked language modeling. For our baseline, we explore the multiple approaches taken by papers such as (Wang et al., 2021) and (Pepino et al., 2021) where the authors have explored finetuning the wav2vec2.0 feature extractor with a speech classification head. In the vein of these works, we implement a SER classification head on top of the wav2vec 2.0 architecture. As such, the wav2vec2.0 architecture acts as an effective feature extractor, which takes in raw audio files, encodes features and passes it on to a speech classification pipeline.

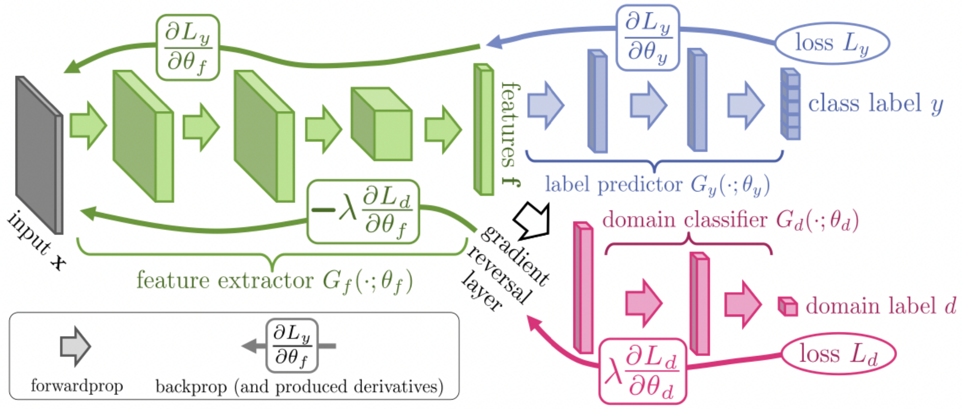

$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features $\mathbf{f}$
label predictor $G_y(\cdot;\theta_y)$
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
input x
feature extractor $G_f(\cdot;\theta_f)$
gradient reversal layer
domain classifier $G_d(\cdot;\theta_d)$
domain label d
$\lambda \frac{\partial L_d}{\partial \theta_d}$
loss L_d
forwardprop
$\frac{\partial L_y}{\partial \theta_f}$
backprop (and produced derivatives)

Figure 2: Domain Adversarial Neural Network

4.1.2 XLSR (Babu et al., 2021b) for Speech Emotion Recognition

The XLSR architecture as proposed in Babu et al. (2021a) by Facebook has been shown to perform well on certain downstream tasks (Wongpatikaseree et al., 2022). In contrast to wav2vec 2.0 which is trained on monolingual data, XLSR learns cross-lingual speech representations by pre-training from raw speech audio in multiple languages. The resulting model has been shown to significantly outperform previous monolingual models. We hypothesize due to the variety of data in training, XLSR will generalize the most effectively across accented speech emotion recognition. We experiment with the performance of XLSR on speech emotion recognition by using it as a feature extractor, freezing the weights and adding a speech classification head to perform the classification.

4.1.3 HuBERT (Hsu et al., 2021) for Speech Emotion Recognition

HuBERT presents another architecture for self-supervised learning of speech representations. In particular, HuBERT utilizes techniques in offline clustering to provide aligned target labels for a BERT-like prediction loss. Wang et al. (2021) presents approaches for partial and complete fine-tuning of HuBERT for several downstream speech tasks. The paper shows that a partially fine-tuned HuBERT architecture gives the best results on Speech Emotion Recognition (SER). Again, we propose utilizing the HuBERT architecture as a feature extractor, taking in the raw audio files, encoding representations of audio, and passing it to the speech classification head. We freeze the weights of the complete architecture except the projection and classifier layers.

4.2 Domain Adaptation: Domain Adversarial Neural Network (DANN)

Domain adaptation specializes in improving model performance in cases where there is a difference in distributions amongst different data sets. For instance, this difference could look like a data shift, a shift in the variance of the variable or a shift in the mean of the variable due to various factors such as the training and test data sets originating from varying sources Farahani et al. (2020).

Domain Adversarial Neural Networks (DANN) is an approach within domain adaptation that has labeled data at the source and unlabelled data at the target Ganin et al. (2015). During the training of a DANN, certain features appear that do not differentiate as we move between domains. As shown in Figure 2 the Domain Adversarial Neural Network architecture produces features for label and domain classifier and then predicts the class label and domains.

4.2.1 Domain Adaptation for Accented Speech Emotion Recognition

As we previously discussed in our project proposal and as titled for our project, the heart and future impactful contributions of this project are to explore domain adaptation of SER on accented speech. While speech emotion recognition on accented and multilingual speech is highly desired to expand the applicability of emotion recognition across many users, not many contributions have been made in this area likely due to a lack of dataset resources in the domain. Consequently, transfer learning poses a promising avenue in extending SER across accented domains.

To study emotion recognition transfer, we require an accented English database. For this project, we will be using the NSC (Koh et al., 2019), which consists of accented Singaporean-English audio across different contexts. While this data does not provide emotion labels, our mentors in this project will be providing annotations for our group to work with from this data.

For future experimentation, we propose using fine-tuning methods on our architectures to explore the initial extent of the transfer of our models to accented data. We hypothesize that these preliminary baselines will perform inadequately on accented Singaporean-English speech, as a direct transfer from a non-accented pre-trained model may be poor. After collecting these initial results, we intend to explore domain adaptation techniques that have been promising in speech recognition applications, such as Domain Adversarial Training (Milner et al., 2019), to improve along these results and track ablations. From these experiments, we intend to provide contributions in understanding the limitation of domain adaptation and transfer learning techniques in speech emotion recognition.

4.2.2 Domain Adversarial Neural Network for Accented Speech Transfer

A Domain Adversarial Neural Network (DANN) contains a deep network which is the Feature Extractor, and another deep network which is the Classifier or the label predictor. Along with this, an additional Domain Classifier is connected to the Feature Extractor through a Gradient Reversal Layer which aims to ensure that the feature distributions over the two domains are made similar, as visualized in Figure 2.

At the time of training, we use MSP Podcast Speech and Emotion labels for training the Domain Classifier and the Emotion Classifier. We also use NSC to train the domain classifier and the Automatic Speech Recognition module.

During each batch of training, we will combine the mixture of NSC and MSP Podcast data and pass it through the speech encoder. Then, we extract only the MSP Podcast section of the batch and use it to retrieve the emotion logits. Subsequently, we use the NSC-only portion of the batch and get ASR logits with CTC. Using these logits, we compute the corresponding losses, scale parameters and then compute a final loss.

5 Results

5.1 Metrics

For each model, we report the following metrics

- **accuracy the :** total number of correctly classified examples over the total number of examples
- **weighted F1:** mean of all per class F1 scores weighted by actual occurrences of each class in the dataset

  The F1-Score for each class captures the tradeoff between recall and precision of each class and is calculated as follows:

$$\frac{\mathtt{TP}}{\mathtt{TP}+\frac{1}{2}(\mathtt{FP}+\mathtt{FN})}$$

- **macro F1:** computes F1 such that each metric is calculated independently for each class and computes the average. That is macro F1 weights each class equally. More precisely, the metric is the harmonic mean of the average precision and the average recall of all the classes. Due to imbalance of our validation and test sets, we highlight this metric as an important indicator of performance, since other metrics, such as accuracy, may report inflated performance due to a model attributing high recall to the most common label.

5.2 Transfer Learning Results

We report hyperparameter fine-tuned baseline results on MSP-Podcast across all three proposed models: Wav2Vec 2.0, HuBERT, and XLSR. For each model, we report accuracy, weighted F1, and macro on the MSP validation and NSC test splits. In Table 4, we outline a comparison of each model. We train over 10 epochs. Each model took around 10-15 training hours for 10 epochs on a 40GB NVIDIA GTX A6000. Performance reported is from the best-performing model over 10 epochs.

Methods	Metrics		
	Accuracy ↑	Weighted F1 ↑	Macro F1 ↑
MSP Validation			
Wav2Vec 2.0 (Baevski et al., 2020)	48.5	31.7	13.1
HuBERT (Hsu et al., 2021)	**61.1**	**56.5**	37.8
XLSR (Babu et al., 2021b)	60.9	56.1	**38.2**
NSC Test			
Wav2Vec 2.0 (Baevski et al., 2020)	90.4	85.8	23.7
HuBERT (Hsu et al., 2021)	90.7	88.4	40.9
XLSR (Babu et al., 2021b)	**90.8**	**89.0**	**45.3**

Table 3: Validation performance (accuracy, weighted F1, and macro F1) on MSP-Podcast and NSC across implemented baseline models

5.3 DANN Results

In the following table, we report results from our DANN experiments. We train the architecture for 10 epochs on a 40 GB A100 Google Colab GPU. We report results on MSP and NSC test data from the highest performing checkpoint over training.

Methods	Metrics		
	Accuracy ↑	Weighted F1 ↑	Macro F1 ↑
MSP Validation			
Domain Adversarial Neural Network	49.3	34.3	15.1
NSC Test			
Domain Adversarial Neural Network	88.1	85.2	27.9

Table 4: Domain Adversarial performance (accuracy, weighted F1, and macro F1) on MSP-Podcast and NSC

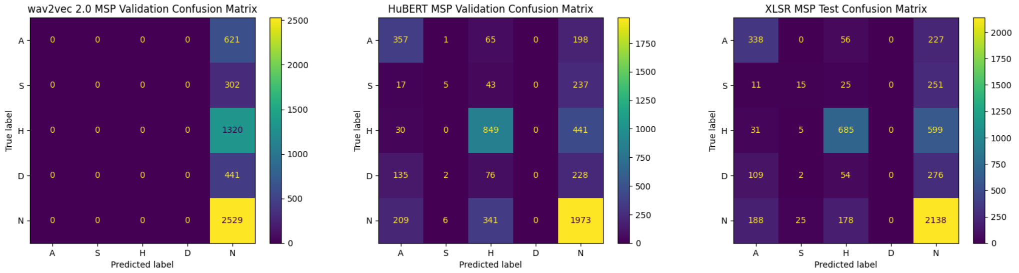
6 Analysis

6.1 Transfer Learning Analysis

6.1.1 Baseline Comparison

Our baseline experiments reveal that HuBERT performs the best in accuracy and weighted F1 across the three baselines at 61.1% and 56.5% with XLSR baseline performing comparably at 60.9% and 56.1%. The worst performing baseline by a larger margin is wav2vec 2.0, performing at 48.5% and 31.7% on accuracy and weighted F1, respectively. To better analyze these results, we plot a confusion matrix 3 across the subset of MSP-Podcast emotions: anger, sad, happy, disgust, and neutral.

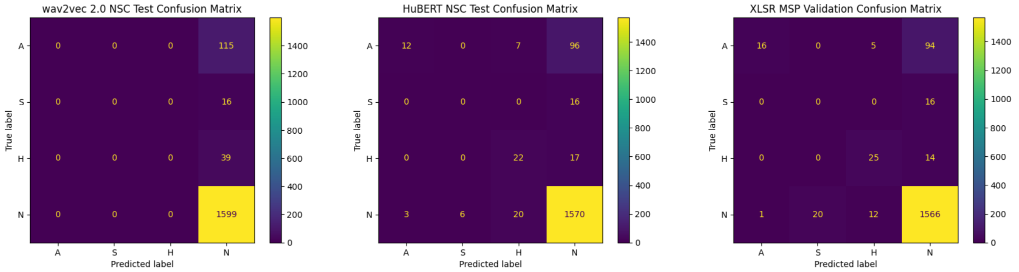
In particular, Figure 3 reveals that the poor performance of wav2vec 2.0 may be due to overfitting on MSP-Podcast, in which it incorrectly leverages the imbalance of neutral labels in the dataset by only classifying examples with neutral labels. This is reflected in its low macro F1 score. Moreover, upon closer examination, XLSR outperforms all baselines on macro F1, revealing it may generalize better because higher macro F1 performance signifies that the model is not as incorrectly leveraging the imbalance of the validation set.



(a) (b) (c)

Figure 3: Confusion matrix on MSP validation set for each baseline: (a) wav2vec 2.0, (b) HuBERT, and (c) XLSR

XLSR's success in macro F1 on MSP-Podcast is further corroborated by its higher performance across-the-board on the accented NSC test dataset, beating both wav2vec 2.0 and HuBERT on all reported metrics. The NSC confusion matrices in Figure **??** support this finding and show that while XLSR labels fewer overall neutral examples correctly, it outperforms other baselines for other emotions, such as anger and happiness. We hypothesize that XLSR tends to generalize better than other baselines in accented speech domains because it is trained on more hours of audio and across multiple languages whereas wav2vec 2.0 and HuBERT are trained monolingually on less data. This variety in audio training likely allows the model to better generalize across non-standard English accent domains.



(a) (b) (c)

Figure 4: Confusion matrix on NSC test set for each baseline: (a) wav2vec 2.0, (b) HuBERT, and (c) XLSR

6.1.2 Common Errors and Dataset Limitations

Across all baselines, we note models struggle to classify sadness, disgust, and anger. We highlight two factors which may contribute to this error. Firstly, after qualitative investigation, we found models tended to incorrectly label audio with milder emotion. Thus, because these labels are less exaggerated and close to neutral, the models are more prone to incorrectly classify these difficult examples.

And lastly, we highlight that both MSP and NSC suffer from dataset imbalance. In each dataset, the most common emotion is neutral and the least common include sadness and disgust. Consequently, the model is the best at evaluating neutral emotion and the worst at sad emotion, which are respectively the most common and least common examples in the train and test data. This likely contributes to both the high performance and recall on neutral labels and the low performance recall on sad emotion. We highlight this as a limitation of the field, more balanced annotation of datasets must be provided to get a better understanding of SER model performance.

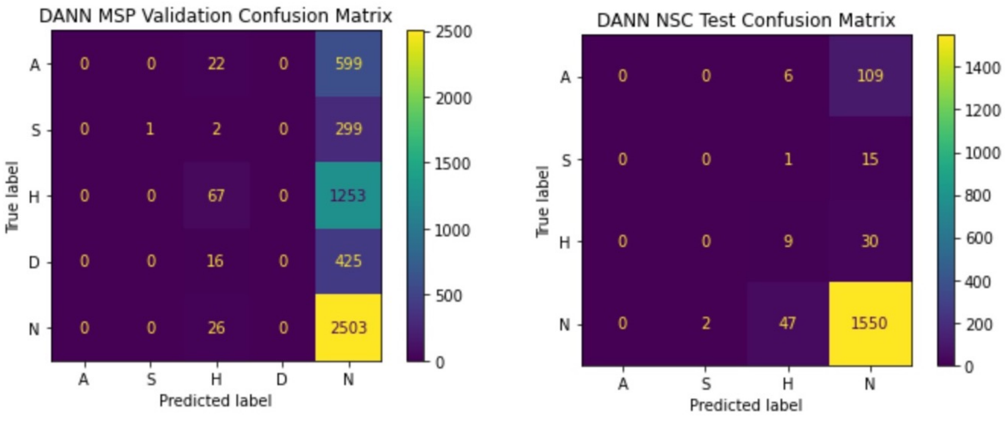
6.2 DANN Analysis

As shown in the confusion matrices below, the Domain Adversarial Neural Network does a decent job at the Speech Emotion Recognition task with a sub-optimal model, achieving 88.1% and 49.3%

on the NSC and MSP data, respectively. While these do not improve upon our transfer learning baselines, we highlight reasons why the DANN may be under performing.

Firstly, given the time constraints, we were only able to train the model for 10 epochs with a base set of hyperparameters. It is likely with longer train times and a larger hyper parameters search, performance will improve.

And lastly, another major limitation is the imbalance in the emotion classes, which results in a sub-optimal domain transfer performance. Consequently, the training tends to overfit on the training data, and over predicts the emotions found more commonly in the data set, such as neutral and happy. This is supported by the confusion matrices in Figure 4, in which these emotions are predicted most often. Moreover, the model reached 99.2% accuracy on the MSP dataset while training. This discrepancy in comparison to the validation performance suggests that the model is still incorrectly leveraging the imbalanced train distribution. This is a similar trend that we mentioned in the baseline analysis.



(a)

(b)

Figure 5: Confusion matrix performance of DANN on (a) MSP and (b) NSC

7 Future Work and Limitations

In conclusion, we report our transfer learning and domain adaptation results. We find that the DANN implementation does not improve against transfer learning baselines. We hypothesize that the DANN architecture did not surpass baselines due to insufficient distributions of train data. In particular, we note that our NSC emotion dataset was small and training with more NSC emotion labels would likely improve transfer more effectively as the model would better train the domain classifier to understand the distribution shift. As such, in future work, we hope to have more even batches between MSP and NSC train data.

Despite the success in the transfer learning baselines, we highlight limitations in imbalance of our datasets, which likely contribute a skewed over exaggerated performance on our accented test split. To address this issue, we hope to evaluate these models on a more balanced and larger speech emotion test dataset. Moreover, if we accept imbalance of distribution as a natural aspect of emotion recognition in general, future work may also focus on techniques to address dataset imbalance.

In addition to the class imbalance in the datasets we used, due to time restrictions, we were not able to experiment greatly with hyperparameters and train for longer periods of time. We are fairly confident that proper hyperparameter tuning and sufficient number of epochs combined with some basic class balancing techniques would result in a significant jump over the baselines using domain adaptation techniques.

References

Mirko Agarla, Simone Bianco, Luigi Celona, Paolo Napoletano, Alexey Petrovsky, Flavio Piccoli, Raimondo Schettini, and Ivan Shanin. 2022. Semi-supervised cross-lingual speech emotion recognition.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021a. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021b. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

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.

Stefano Bromuri, Alex Henkel, Deniz İren, and Visara Urovi. 2020. Using ai to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*, ahead-of-print.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1):33–60.

Priyank Dubey and Bilal Shah. 2022. Deep speech based end-to-end automated speech recognition (asr) for indian-english accents.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. 2020. A brief review of domain adaptation.

Riccardo Franceschini, Enrico Fini, Cigdem Beyan, Alessandro Conti, Federica Arrigoni, and Elisa Ricci. 2022. Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss. *arXiv preprint arXiv:2207.11482*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2015. Domain-adversarial training of neural networks.

Yuan Gao, JiaXing Liu, Longbiao Wang, and Jianwu Dang. 2021. Domain-adversarial autoencoder with attention based feature level fusion for speech emotion recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6314–6318.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jia Xin Koh, Aqilah Mislan, Kevin Khoo, Brian Ang, Wilson Ang, Charmaine Ng, and Ying-Ying Tan. 2019. Building the Singapore English National Speech Corpus. In *Proc. Interspeech 2019*, pages 321–325.

Shashidhar G. Koolagudi and K. Sreenivasa Rao. 2012. Emotion recognition from speech: A review. *Int. J. Speech Technol.*, 15(2):99–117.

Yang Li, Wenming Zheng, Yuan Zong, Zhen Cui, Tong Zhang, and Xiaoyan Zhou. 2021. A bi-hemisphere domain adversarial neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 12(2):494–504.

Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2016. Emotion recognition using multimodal deep learning. In *International conference on neural information processing*, pages 521–529. Springer.

R. Lotfian and C. Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Cheng Lu, Yuan Zong, Wenming Zheng, Yang Li, Chuangao Tang, and Björn W Schuller. 2022. Domain invariant feature learning for speaker-independent speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2217–2230.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Rosanna Milner, Md Asif Jalal, Raymond WM Ng, and Thomas Hain. 2019. A cross-corpus study on speech emotion recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 304–311. IEEE.

Omar Mohamed and Salah A. Aly. 2021. Arabic speech emotion recognition employing wav2vec2.0 and hubert based on BAVED dataset. *CoRR*, abs/2110.04425.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *CoRR*, abs/2104.03502.

Leyuan Qu, Wei Wang, Taihao Li, Cornelius Weber, Stefan Wermter, and Fuji Ren. 2022. Data augmentation with unsupervised speaking style transfer for speech emotion recognition.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. Technical report, Technical report, OpenAI, 2022. URL https://cdn. openai. com/papers/whisper. pdf.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Dino Seppi, Anton Batliner, Stefan Steidl, Björn Schuller, and Elmar Noeth. 2010. Word accent and emotion.

Yifan Sun, Werner Sommer, and Weijun Li. 2022. How accentuation influences the processing of emotional words in spoken language: An erp study. *Neuropsychologia*, 166:108144.

Benigno Valles. 2015. The impact of accented english on speech comprehension" (2015). open access theses dissertations. *Open Access Theses Dissertations*.

Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. 2010. Automatic speech recognition of multiple accented english data. pages 1652–1655.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *CoRR*, abs/2111.02735.

Konlakorn Wongpatikaseree, Sattaya Singkul, Narit Hnoohom, and Sumeth Yuenyong. 2022. Real-time end-to-end speech emotion recognition with cross-domain adaptation. *Big Data and Cognitive Computing*, 6(3).

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: speech processing universal performance benchmark. *CoRR*, abs/2105.01051.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2019. A comprehensive survey on transfer learning.