Machine Learning in Linguistics and Speech Recognition

Project Option 1

DAWSON DAMUTH and DANIEL VIOLA

Abstract - This research explores the effectiveness of various machine learning techniques in accent recognition from audio samples. Specifically, the study evaluates the performance of K-Nearest Neighbors (kNN), Support Vector Machines (SVMs) with different kernels, and Convolutional Neural Networks (CNNs) on Mel-Frequency Cepstral Coefficient (MFCC) features derived from speech recordings. In addition, the direct audio files by which the MFCC was derived will be used on facebook's wav2vec. The transformer model is based on the code in reference [23].

Motivated by the persistent challenges that accents pose in speech recognition systems, this work aims to assess how modern classification models handle accent variability and whether the extra computational time of a model such as wav2vec is warranted over the more classical and less computationally heavy models. Knowing that automatic speech recognition (ASR) model struggle when accents of input data is not uniform, we hope that a low compute and sparse data classifier could be a solution as a preprocessing tool for an ASR. A comprehensive review of related literature highlights the advantages and limitations of traditional and deep learning approaches, while also emphasizing emerging trends such as transformer-based models. The project contributes to the growing field of accent-aware speech recognition by benchmarking model accuracy, examining architectural implications, and addressing common biases, ultimately paving the way for more inclusive and robust language technologies.

1 PROJECT OBJECTIVE

This project aims to evaluate the effectiveness of various machine learning techniques in distinguishing speaker accents from audio samples. We will implement and compare K-Nearest Neighbors (kNN), One-vs-All Support Vector Machines (SVMs) with RBF kernel, Convolutional Neural Networks (CNNs), and transformer-based models such as Wav2Vec 2.0. Our goal is to investigate how advancements in deep learning, particularly in transformer architectures, improve performance on the complex task of accent recognition, in both raw accuracy, and accuracy compared to computational effort. Often is the case in the real world that the computational power to retrain a large transformer model for specific data does not exist, and that the data is often to sparse. By looking at a pretrained transformer with classification head training as well as unfreezing only the last layer we believe this represents a reasonable real world scenario for an accent classifier that could be fed into an ASR to improve ASR stability.

Using the classical methods provide a good benchmark as to what can be achieved with more limited processing power. As part of our methodology, we will collect audio data from speakers with different accents, extract Mel-Frequency Cepstral Coefficients (MFCCs) and other relevant features, and train our models on these representations. the transformer model will be trained on the direct audio files. Ultimately, we aim to determine which approach offers the most robust and accurate performance for accent classification, while taking measure of the computational requirements for each model.

Authors' address: Dawson Damuth, dawsonda@buffalo.edu; Daniel Viola, dviola@buffalo.edu.

2 RELATED WORKS

Several related works in accent recognition and language models are compiled. These works are relevant to our models, either in their application methods or in their consideration of shortcomings and challenges faced in their processes.

First, we have several articles which are related to the machine learning methods implemented in our research specifically and their inherent advantages/disadvantages in comparison to traditional models, including the use of k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Recurrent Neural Networks (RNN), Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Multi-later Perceptron (MLP), and Long Short Term Memory (LSTM) models.

This 2015 paper [4] explores the utility of Mel-Frequency Cepstral Coefficients (MFCCs) as the primary feature extraction technique, a method well-suited to modeling human auditory perception due to its transformation of raw audio signals into a more compact, perceptually meaningful representation. One of the notable challenges addressed by the authors is the high dimensionality of raw time-domain audio signals; MFCCs effectively mitigate this issue by condensing thousands of signal samples into a manageable number of coefficients, preserving critical acoustic features while reducing computational load. The study compares several classification models, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machines (SVMs) with both radial basis and polynomial kernels, and the k-Nearest Neighbors (k-NN) algorithm. Interestingly, the results highlight the k-NN classifier as not only the most accurate but also the fastest. However, the paper does not shy away from emphasizing the limitations faced when using MFCCs in noisy environments, where the feature extraction process can suffer due to the sensitivity of MFCCs to signal corruption. Additionally, the older statistical models like LDA, though foundational, are shown to be less competitive in terms of accuracy and adaptability compared to more modern, flexible algorithms of the time like SVM and k-NN. The paper is relevant not only for its empirical comparisons but also for its methodological clarity, offering a comprehensive framework for researchers interested in accent recognition systems, especially those looking to balance computational efficiency with classification accuracy.

Continuing with the use of Neural Networks, we have an article which focuses on the use of RNN and DNN models, trained on long-term and short-term features, respectively. [5] The objective of the authors research is to train a model which can reduce the impact of background noise by using the combination of both RNN and DNN models. They make mention of the model's ability to surpass the baseline Neural Network, as well as a SVM model. However, the article also articulates the struggles faced in their research, providing evidence of the models struggling to identify languages which are geographically close, having a greater error rate among them. They

give examples of these issues, specifically among Chinese, Korean, and Japanese, as well as grouping error among Hindi and Telugu. They propose a remedy to this issue by implementing a hierarchical classifier to consider groups of languages so that the aforementioned hybrid model may make more fine-grained decisions.

Another article makes use of deep learning and machine learning models to effectively identify speech accents. [6] This article found that the most effective model was the decision tree, surpassing all other machine and deep learning models. This conclusion in the scope of our research suggests that the CNN being implemented may not be the best model in terms of accuracy, contrary to our hypothesized outcome. A similar article tested a similar set of models, which suggested that the CNN model performed best. In their article, Sheng and Edmund utilized the same approach as what we will be utilizing for data decryption for use in the model, MFCCs, implementing four main models for comparison: CNN, MLP, Random Forest, and Gradient Boosting. As mentioned, the CNN model performed best, in contrast to the previous article's results which suggested the decision tree model performed best. Both articles suggest improvement for future research, such as including gender-specific training for improved model accuracy across all models. Bridging to more advanced approaches within Neural Network implementation for accent recognition, we have an article which discusses methods of improvement to accent recognition models through a novel descriptor referred to as "Grad-Transfer", which serves the purpose of transferring knowledge from CNN models to traditional machine learning models. The results from implementing Grad-Transfer in the modeling process is shown to outperform baseline approaches and self-supervised learning models by a significant margin. This approach could be considered as a potentially eager next step in our research, though it is computationally expensive.

The next article introduces VFNet [7], a CNN architecture designed for accent classification. The model captures a hierarchy of features by applying variable filter sizes along the frequency band of audio utterances, outperforming previous benchmarks in accent classification tasks.

Another paper presents a deep learning framework for accent recognition [8], employing a Convolutional Recurrent Neural Network as the front-end encoder and integrating local features using a RNN to create an utterance-level accent representation. The model achieved significant improvement in accent classification, demonstrating the effectiveness of discriminative training methods.

The next article implements ensemble methods along with deep learning approaches to determine the best singular and ensemble models. [9] Their study also utilized MFCCs in their implementation, concluding that the best single model was LSTM model in deep learning, and the best ensemble model was a vote of average probabilities between a Random Forest and LSTM. This gives another insight as to what future research with our models may include, utilizing ensemble methods to further improve our model's predictive power in accent identification.

Our final article of this section, and perhaps most closely associated to our research, is a study which explores the use of SVM and kNN models. [10] The study is done on a commonly referenced open-source dataset titled, "Speaker Accent Recognition." As the name suggests, the data is intended for use with accent recognition models and includes a distinct six classes of accents, as opposed to our two-class identifier. Through the model building and evaluation process, it is found that the SVM utilizing a Radial Basis Function performed best, in terms of accuracy. The author also suggests exploring deep learning methods and feature reduction techniques as potential next steps. This gives us a good reference to what we can expect when implementing our SVM and k-NN models in our research, with the SVM being anticipated to perform better.

Furthering our claims, we will investigate the concerns and challenges presented in this line of research. The following three articles discuss model-specific challenges, bias mitigation, and the need for diversity in datasets.

This next paper's research addresses the challenge of accent recognition using a CNN-based model [11] trained on audio features extracted from the Speech Accent Archive dataset. The study demonstrates that incorporating time-frequency and energy features, such as spectrogram, chromogram, spectral centroid, spectral roll off, and fundamental frequency, alongside MFCCs, enhances accent classification accuracy. The model achieved recognition accuracies ranging from 96.4% to 98.7% for English with Germanic, Romance, and Slavic accents.

Our first article discusses the consequences of accent bias, specifically in models trained on data which may have overly general labels for accent recognition. [12] The discussion is centered around machine learning models which incorporate these generic and non-specific labels for accent identification, emphasizing the importance of an accurate and well-trained model for accent recognition. We see through their analysis that, even when accents are not of concern, the identification of accents cannot be overlooked when attempting to implement speech recognition models.

Second, we have an article which analyzes the consequences of gender bias in accent recognition and the various ways in which we may mitigate the concerns. [13] The initial research concern for this issue was when investigating accent recognition models and realizing a discrepancy in accuracy ratings among male and female-specific samples, with the latter performing better in testing. To mitigate these concerns, gender-specific models were trained, significantly improving performance. Following these results, the models can be applied in tandem with the original model to improve overall accuracy and decrease bias.

The next two articles investigate transformer-based speech models, identified to be among the most cutting-edge algorithms for accent recognition.

Our first article presents CommonAccent [14], a benchmark for accent classification using transformer-based models, specifically Wav2Vec 2.0/XLSR and ECAPA-TDNN, evaluated across English, German, Spanish, and Italian using the Common Voice dataset. The transformer-based Wav2Vec 2.0 model consistently outperforms traditional architectures like ECAPA-TDNN, achieving up to 97.1% accuracy in English accent recognition and demonstrating superior

generalization with minimal architectural changes. The research confirms that large pretrained transformer models, when fine-tuned, are highly effective and scalable tools for recognizing diverse accents across languages.

The second article explores the internal phonological structure of Wav2Vec 2.0 embeddings [15] by using probing models and association rule mining to assess whether phonetic features align with linguistic theory. The findings reveal that Wav2Vec 2.0, despite being trained without explicit phonetic supervision, captures a structured and linguistically coherent organization of speech features. These results underscore the efficiency and interpretability of transformer models in speech processing tasks, supporting their strong potential for accent recognition and other phonetics-aware applications.

For the final two articles, our first investigates another more current, but very computationally demanding approach to the speech recognition issue which doesn't utilize an accent-specific layer, and the second demonstrating the need for this domain of research as a consequence of perceptual learning and human error.

As mentioned, the first article discusses the inner-workings of a systematic Unsupervised Data Augmentation (UDA) approach for cross-domain speech recognition, labeled CASTLE, or Causal Structure Learning. [16] This approach seeks to improve neural network generalizations by learning causal structure within the data, making this an unsupervised approach, in comparison to the previous supervised approaches. Notably, this method as written in the research article does not utilize an accent recognition mechanism. Rather, this approach analyzes similarity in structure of the words in their recordings to categorize the samples indirectly. Although this approach has a quite significant accuracy rating, even when compared to some of the most powerful algorithms and networks, even in ensemble methods, the model itself is almost impossible to interpret and translate to any audience, even someone with a vast knowledge in machine learning and deep learning.

Our final article discusses human ability for perceptual learning, focusing on the factors which influence the learning of human-varied speech. [17] Even though the article contains a plethora of knowledge in perceptual learning, what we're concerned with for the sake of comparison is the difficulty and time needed to advance human perceptual learning. The article states that the process of perceptual learning involves long-lasting changes in the perceptual system and can be a very lengthy process for many. This is also seen as an essential developmental step for humans, as the ability to adapt to unusual dialects and accents is a necessary component of our response to environmental input. Having accounted for the cost and importance of accent recognition in humans, we can identify a need-based market for our research in accent recognition software, and in-turn, language learning models. As is a common theme in the last half-century, technological integration is becoming more and more prevalent in our everyday lives and is recently considered to be another fundamental point of our development. With that, we can already see accent recognition, language learning models, and live-transcription models are becoming increasingly popular.

There are an additional three articles which deserve recognition in their contributions, especially in the more modern aspects of their research associated with transformer models. Their contributions specifically highlight the hardships with the more recent modeling approach, being the transformer model.

This article demonstrates that many automatic speech recognition (ASR) models struggle when working with accented speech data [19]. This article highlights locational differences as even slight changes in the German accent (German spoken in Germany, versus outside of Germany) had substantially different effects in ASR model performance. Even small deviations can substantially change the target distribution p(y).

This article highlights the use of transformers in somewhat low resource environments [20],though only data size was an issue here, not processing power. The best results were paired with wav2vec + 100 frames of segmentation of the data + logistic regression. This approach did not train the classifier head of wav2vec, rather used it as a feature extractor. This then fed into frame segmentation of the MFCC transformation of the data which fed into the Logistic regression. Other less complex models saw accuracies ranging from the mid 60%'s to the 70%'s. This highlights that even Complex setups for state of the art models without computational limitations seems to max out at what traditionally would be considered 'low accuracies.'

This article highlights the use of a pretrained model such as wav2vec in the classification of accents with substantial data [21]. It recognized that accent recognition is a very difficult task with low accuracy. In some cases some accents received no recognition at all, some were single digit. The overall f-1 score was 34% within an 11 percent margin. This highlights how accent classification even with a strong pretrained model can often be terrible.

3 PROJECT APPROACH

The baseline methods we will be using will be K-Nearest neighbors, Support Vector Machines, and Convolutional neural networks. While CNN's are only a few years out of being state of the art, and in many cases preform very well, the use of wav2vec as a strong state of the art option will also be included.

The K Nearest neighbors supports multi classification out of the box so it will provide no complications. Support Vector Machines are inherently binary machines so what we will do is use the One Versus all setup commonly used to adapt SVM's into multi classification problems. We will be testing a variety of common kernels to prevent that from being a large limiting factor.

Next we will use a CNN with multiple hidden layers to test its accuracy. We know from our previous research these methods all have the ability to preform moderately well, with the CNN being the favorite to win in the baseline methods. Finally, as previously mentioned, wav2vec will be run with the training of the classification head as well as the unfreezing of the final layer of the pretrained model. The unfreezing of a greater amount of layers is not cohesive to our goals as many low compute scenarios will likely be unable to handle the unfreezing of more layers.

Because data is often sparse in the real world, we will be using the kaggle accent data and using a process of drawing multiple subclips from the greater sentences to augment the data in a way without using the usual perturbation methods which in this case could heavily convolute the true classification of the accent. Normalization factors such as L2 of weights and label smoothing in the softmax loss will also be applied

Overfitting was an issue that was encountered as well. To address this issue, we first trained a few epochs only with the classification head active for the wav2vec model. then after it had baseline stability we unfroze layer 11, the last layer of the transformer, and allowed that to tune to the data as well. As the model was run, test set accuracies were being stored, the running best was stored with the model parameters at its time. We stopped training when after many epochs of training percentage going up without the test set following. From here, a roughly .51 test set accuracy was achieved. The regularization we used for this model was to use label smoothing in the softmax loss function of .1, as well as regularization of L2 norms with a coefficent of .15. This helped us achieve a higher test accuracy than without regularization. In its absence we were only able to achieve 30%.

4 DATASET DESCRIPTION

The main dataset that we will use is the Global Speech Accent Recognition dataset available on Kaggle. A large part of the challenge will be that these data are available in audio format, meaning Mel-frequency cepstral coefficients will need to be generated from them so that we can use the data in our models. In addition, the audio files are in MP3 format, so some additional preprocessing will be needed to properly feed it into the transformer.

Beyond these baseline challenges, we are also presented with an issue of data size, while there are 2140 audio samples, many of the accents are sparse in their sample count, with many having only a single sample. As a result, to make this feasible, we have centered on 5 main accents to train for of which the data size is reasonable. Those accents are Arabic, English, French, Mandarin and Spanish.

An additional challenge is that english and Spanish present with very high amounts of samples, and the other 3 of our 5 accents are still far less abundant than those 2. It would be not of great use for us to allow english and Spanish to have highly abundant representation in the training batches, where the others were sparse. Especially given that english and Spanish are already somewhat well trained accents, not only would it not properly represent real classification to have starkly different numbers of data points, it would also in many ways defeat the purpose of the project to achieve great results only in english and Spanish and poor results in the others.

To address this concern, we will be limiting the amount of data points of any one type of accent in a given training batch for the transformer. This worked to curb over training on english specifically.

That said, this did not fully address the global problem that some accents were sparse in data points. To solve this especially in the case of the transformer where the data size is more impaction, each audio file will be clipped to 4 seconds, and 4 deterministically determined audio samples will be drawn from each data point. We have chosen to treat this as a form of data augmentation, creating a larger amount of faster training data points for the model.

The challenge to this approach in this context is that wav2vec has good recognition of sentences and works well with longer audio clips. However, having spoken to Ms.Pezouvanis (ongoing masters in Speech Pathology) who conducts research in speech pathology, accents themselves are human classifiable in the context of a few words clipped from a sentence. The entirety of the sentence is not needed to recognize the patterns in speech for sake of specific accent classification. She agreed that it was reasonable to classify an accent off of a short clip and that based off of speech theory alone there should be minimal difference compared to a longer clip. Once a baseline amount of phonetics was expressed the accent should be accessible. As a result the creation of sub clips is a valid technique for augmenting the data. The breaking of the sentence structure is not a vital factor given that accents alone can be classified even from just a few words. This worked with good success.

It is a natural question to ask why some other data augmentation techniques were not used, ones that are less about clip sampling and more about perturbations. The small perturbations in the data often used for augmenting some in the form of speech rate and speech pitch. Having again spoken to Ms. Pezouvanis we find that such types of perturbations are in fact not 'small' changes that they would be intended to be to help generalization. Rather, changes in pitch and speed can outright alter the classification of the accent, as accents are dependent on speed and pitch very directly. One accent could legitimately 'become' another through speed or pitch changes. Specifically, the groupings of French and Arabic, English and Arabic and Spanish and French could become far less distinguishable with even slight changes to pitches and speeds. This would not cause the small changes we expect, but rather a large departure from the truth of the dataset by having a chance to in some cases strongly misrepresent the accents.

The MFCC process does not generate features that are directly interpretable given the nature of the transformations required to generate them. However, it is understood that these features capture patterns in the audio data relating the pitch, loudness, timbre, and other sonic qualities. The large challenge of our project is meaningfully interpreting the classical baseline models as it is hard to parse models under these circumstances.

A sample of the data can be seen, "ES",7.0714,-6.5128,7.6507,11.1507,-7.6573,12.4840,-11.7097,3.4265,1.4627,-2.8127,0.8665,-5.2442, As is evident, the MFCC format lacks interpretability in the data format. The First two letters are the true accent of the data point. The following 12 numbers are the numeric representations of the MFCC transformation, shortened to 4 decimal places for sake of demonstration in this paper.

A t-SNE plot of the MFCC data can be seen, this allows us to recognize that there is not any clear clustering in our data nor is there any simple linear separability between the classes. The one thing that we can see if that arabic accents tend to stay to the right

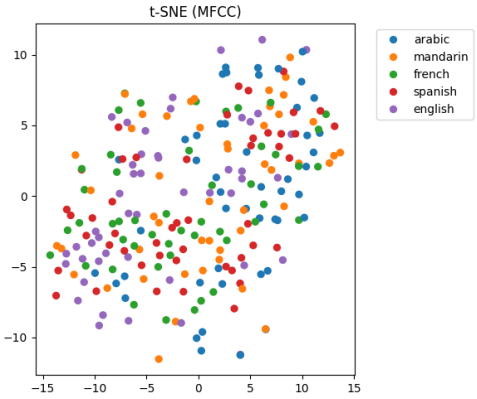
side of the graph, and this is effectively the only thing of note from this graph.

As briefly mentioned before, the transformer model, wav2vec uses audio files directly. Likewise though, due to the nature of deep networks and transformers, interpretability is once again limited. It is not exactly clear the pattern being found in black box methods.

5 RESULTS

When comparing the transformer to the classical models, it should be noted that transformer accuracy for any given sample was calculated with taking the average of 4 individual clips from that single sample, because samples are deterministic, the sum of 4 clips effectively works as the entire sample. We believe this keeps the compatibility of the transformer accuracy to the classical method accuracy.

All results were found using 5 classes of accents. The results for our models are as follows:



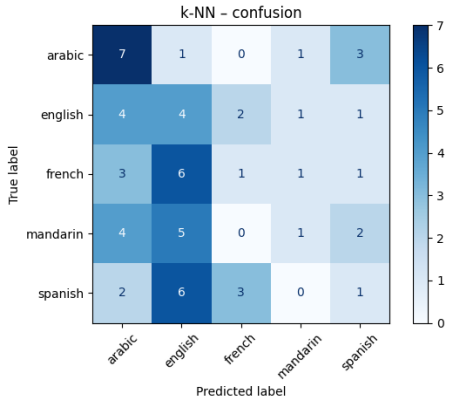
kNN, 3 Neighbors

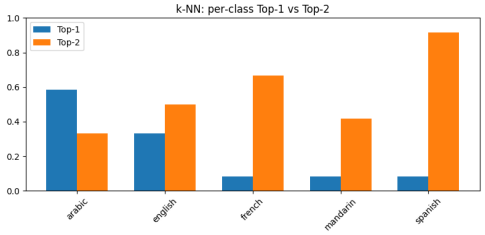
Accent	Precision	Recall	f1-score
Arabic	0.35	0.58	0.44
English	0.18	0.33	0.24
French	0.17	0.08	0.11
Mandarin	0.25	0.08	0.12
Spanish	0.12	0.08	0.10

Accuracy - 23.3%
Top-2 Accuracy - 56.7%

AUC Metric - kNN

Arabic	0.69
English	0.56
French	0.63
Mandarin	0.47
Spanish	0.57





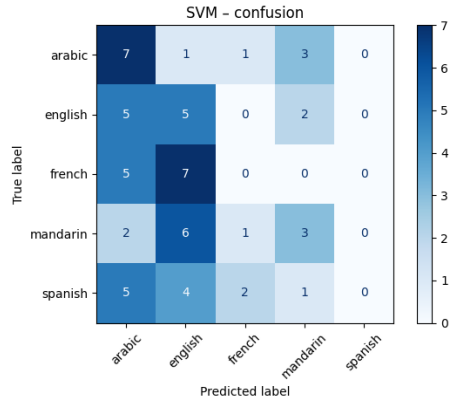
SVM - RBF Kernel - One versus All approach

Accent	Precision	Recall	f1-score
Arabic	0.29	0.58	0.39
English	0.22	0.42	0.29
French	0.00	0.00	0.00
Mandarin	0.33	0.25	0.29
Spanish	0.00	0.00	0.00

Accuracy - 25.0%
Top-2 Accuracy - 55.0%

AUC Metric - SVM

Arabic	0.70
English	0.61
French	0.67
Mandarin	0.59
Spanish	0.48



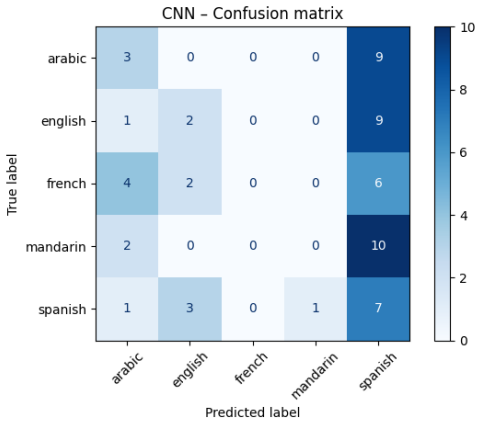
CNN - 2 Dense layers of 128 neurons

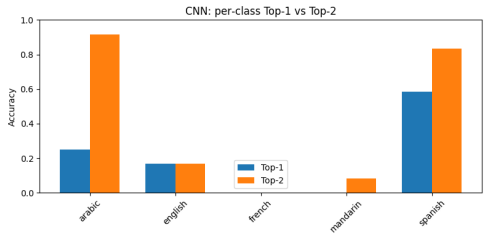
Accent	Precision	Recall	f1-score
Arabic	0.27	0.25	0.26
English	0.29	0.17	0.21
French	0.00	0.00	0.00
Mandarin	0.00	0.00	0.00
Spanish	0.17	0.58	0.26

Accuracy - 20.0%
Top-2 Accuracy - 40.0%

AUC Metric - CNN

Arabic	0.66
English	0.52
French	0.64
Mandarin	0.51
Spanish	0.54





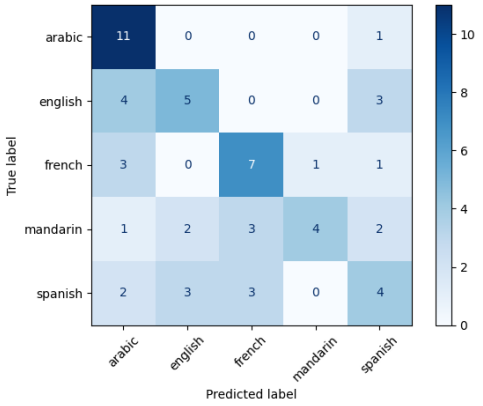
wav2Vec - Classifier Head and Unfrozen Final Encoder Layer

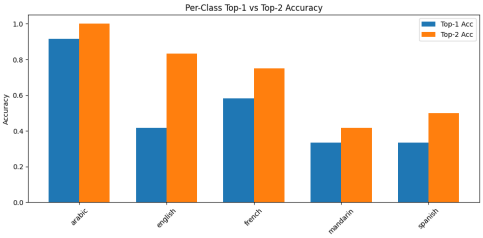
Accent	Precision	Recall	f1-score
Arabic	0.52	0.92	0.67
English	0.50	0.42	0.45
French	0.54	0.58	0.56
Mandarin	0.80	0.33	0.47
Spanish	0.36	0.33	0.35

Accuracy - 51.7%
Top-2 Accuracy - 70.0%

AUC Metric - wav2Vec

Arabic	0.91
English	0.71
French	0.72
Mandarin	0.73
Spanish	0.73





The results show interesting numbers. It is clear that no one model did particularly well in averaged accuracy, though there were some individual accuracy scores that were quite good. Given that other low resource papers have shown only slightly better accuracies with substantially more data, this falls in line with what was to be expected in a lot of ways.

It seems as though for the classification of accents MFCC data is not necessarily representative enough on small data sets to paint the picture properly. Even if we do not expect very high numbers, it is clear that all the classical models fell behind the transformer in the range of nearly 20% averaged accuracy.

To analyze these models through a lens of comparative analysis we will look at multiple metrics. First, overall accuracy of the model, secondly the top 2 accuracy. Thirdly, we will focus on the ceiling and floor of individual classes. Finally, we will use the AUC and F1 scores.

5.1 Average Accuracy

The convolutional neural network was the worst of the models in terms of average accuracy with a 20% accuracy. Followed by KNN with three neighbors at 23.3%, then the RBF Kernel Support vector machine with a 25% accuracy. Leading up to the victor in main accuracy of 51% with the transformer model.

5.2 Top-2 Accuracy

Using top 2 accuracy the results do slightly differ. Top 2 can allow us to see when a model was not totally right but was in the ballpark. Again is the convolutional neural network at the bottom with 40%. then the support vector machine at 55%. Followed by KNN with 56.7%. Once again lead by the transformer model at 70%.

5.3 Floor and Ceiling Accuracies

Moving past the main accuracy and top 2 accuracy we can then use the confusion matrices, and the floor and ceiling of the accuracy for individual classes as a byproduct, of the models to further compare and rank them.

It is clear that the convolutional network and the support vector machine rank at the lowest for their inability to classify all 5 accents in a first guess, meaning floors of 0. the support vector machine did not recognize french or Spanish at all. The convolutional network failed to recognize Mandarin and French. Since the convolutional network had a lower accuracy ceiling for any single class, and since both failed the same amount of classes, by this metric we see the SVM ranked higher.

K nearest neighbors is then ranked in second place as it is the first model to not have complete failure in any class, having a floor above zero. Though its classification power was quite limited as it achieved very low accuracy in three out of five classes. KNN did feature a higher individual class accuracy ceiling of .44, substantually higher than that of the CNN or the SVM.

In this particular realm of measurement, the transformer model once again reigns supreme. Having the highest accuracies not only on the overall average, but the highest classification floor for any one class at .33 for mandarin and Spanish, whereas Knn's lowest was .08 and the other two obtained lowest of 0.

When considering peak accuracy, the transformer is first with a peak single class accuracy of.92 for arabic. this far outshines all of the other models by a very large amount.

5.4 AUC and F-1 Scores

To then compare the models using AUC (area under curve) and F1 score, a brief explanation of what the AUC and F1 represents will be explained first. The AUC represents a models ability to distinguish an accent in a one versus all scenario. In our case, KNN's .63 for

french tells us that when distinguishing french versus all other accents combined into one 'other' class, it ranks a randomly chosen french accent higher than any other random option 63% of the time. Even when AUC is high, accuracy can still be low. Wether or not the model knows something about the accent is represented by the AUC, its choice to act on it and classify based on it is what is seen in the F1-score.

To take these metrics into account for sake of comparing the models, AUC and F1 will be taken as a whole, as AUC alone does not substantially describe the models.

The CNN was the worst based on AUC and F1. with many AUC values being very close to .5, its ability to recognize and accent in a one versus all scenario is marginally better than random chance. The very low F1 scores suggest that even if knowledge of the accent existed, the model rarely acted upon it correctly.

K nearest neighbors and the support vector machine were next in line. The performance difference between them is small but significant by this measure. The average F1 scores for KNN was .202, for the SVM was .194. While the AUC for the SVM ranked higher than that of the KNN for nearly every single class, the lower overall F1 score, as well as the two zero's suggest that even if it has a better awareness of the accents in one versus all scenarios, it failed to act upon it.

As with all of the previous metrics, we see the transformer take the greatest performance once again. Having no AUC score for any class under .71, we can see that in the one versus all scenario it is able to recognize any of the accents individually with reasonable power. The relatively high F1 scores also indicate that compared to the other models it was able to act upon this information to a far greater degree.

To clarify, a model may have a high AUC and a low accuracy for that given class because AUC measures that one class in isolation versus all others, but does not garuntee that it will perform in the multiclass problem. This is why the F1 scores are relevant for comparison with great importance. Generally if AUC is high but F1 is low, we can asses that a model is poorly calibrated. This would suggest to us that the transformer is the best calibrated model of the bunch,

In the case of KNN, the distance between points is likely not a good indicator as it preformed strictly worse than everything else and observed dubious results with its top 2 prediction accuracy (more on this soon). The high dimensional space of the MFCC data and numeric audio interpretations in general likely do not lend well to basic distance measurements as highlighted with the curse of dimensionality. The concept of 'closeness' becomes meaningless.

The SVM likely did not do well because the RBF Kernel does not help separate the data very much. Perhaps a Kernel exists that can achieve good separation of MFCC data, but at this time that is not known.

The CNN likely failed because by nature of the convolution and the semblance between different accents, the model may have diluted the signals of the different accents when applying filters to the data. Given that many accents are similar, the filters may have stripped most of the distinctions away. In addition, because MFCC data is flattened and does not have a time element to it, there is little for the CNN to work with regarding its affinity for similar local and translational patterns.

When considering accent classification in the scope of how we recognized the problem, that being disruptions to large language models with unhandled accents, where the classification is in the setting of low computational power for somewhat sparse datasets, the results are not as bad as might initially appear.

Firstly, some accents received such low accuracies that one could easily make the point that if some accents are not remotely distinct in direct accent classification, they may not pose issues to models that this classification model is meant to provide information to.

Secondly, if data is often classified correctly by at least 2 categories, the same can be said. A certain sample may lean towards a given direction, but not by such an intense nature that its classification into either of the 2 neighboring categories would have substantial effects on the model this would be feeding into. By nature of observing its general direction, and not having clear splitting points in that direction, the direction alone may be enough to be useful. This may be especially useful for Accent Specific Models (ASR) which could be a realistic example of an upstream module.

We have illustrated this with our top 2 class percentages. A graph displaying the proportion of data that was correctly classified into the top 2 choices of the model. If we believe that general directions for similar accents may not need perfect single class accuracy, then this may be of great value. We can see that the model was not exact, but very close for the accents of Arabic, english and french. Spanish and mandarin still lag behind. Our averaged top 2 accuracy for the transformer model was 70 percent.

Finally, we offer one versus all AUC metrics for the transformer as a measurement of confidence. While the Top-2 tells us direct classification accuracy, the AUC can tell us the confidence by which the model ranks the class of our data point versus others. This lets us know that even if the model sometimes gets points wrong, by nature of these metrics being high, it tells us that it is predicting the correct general direction for the data. This aids in the premise of general classification being a strength of the transformer in the low compute, sparse data environment when feeding into another speech based module that is accent dependent. While there was not many strengths of the classical models, it should be noted that each of the classical models were able to run in a matter of seconds, even the CNN only took a few dozen seconds. The transformer did take multiple hours to train even the final layer and classifier head. This would be the main strength of the more simple classifiers over the transformer. There may be additional strengths of the classical models if addressed in a situation with less overall classes. It is likely that the diverse accents created too much noise for them to properly handle. With fewer classes it is conceivable that they would preform much better. If a small set of classes known to disrupt an ASR model was identified and the SVM for example was used

as an upstream filter with a large amount of data, it could prove to be quite efficient to use over something like the transformer if computation is a limiting factor.

6 CONCLUSIONS

As established in the landscape of the research, seen in references Aksënova [19], Zeng[20] and Matos[21] the task of multiclass accent classification is rather hard with state of the art models in the best conditions achieving roughly 80 percent classification. In the same realm of research landscape some of the best models will exhibit some classes having total dropout of 0% classification, and some that behave incredibly poorly. As a result, we can not expect our models, though using state of the art methods, with spare data and spare computational power to outperform these. We can however monitor if something comparable can be created.

It seems evident that our models do suffer from the same negative aspects of the larger ones, but was still able to produce decent results. While overall accuracy was not particularly high, we did get some reasonably promising individual class accent recognition. [21] As well as averaging 70% in the top 2 classification. We find this to be especially important because as a very low computationally powered model, with very sparse data general accent directions are recognized and would potentially fulfill our initial goal of helping an upstream ASR module.

We also make the conjecture that some classes being very poorly classified is not as devastating as it might be in other scenarios. Because accent recognition is being suggested as a pre-step to larger models that are audio based and sensitive to accents, if some accents are indistinguishable to classification models, then those accents being fed downstream into others models is unlikely to have strong effects. If they could not be classified at all, cases of 0-10 (percentage classified) then they probably will not confuse other models as they would be non distinct accents. In the case of Arabic and english being classified similarly, generally knowing that it is in one of those 2 classes is likely to aid the upstream model

Generally speaking we believe that we have shed some insight on the importance of achieving at least one of the two important aspects of accent training being computational power or data abundance. The research landscape generally presents with at least one of those two things and already struggled to see results. Our project saw neither and we learned how big of an impact that can have on the outcome.

That said, we are still proud of the top 2 accuracy of the transformer model for the goal we specifically set out to achieve. Significantly simpler models like KNN, SVM's and CNN's did not present with enough power to justify being used in this scenario. However; we believe we have also put forth evidence that low compute and low data partial pretrained transformer accent classifiers can still be worthwhile, even if not as effective as their bigger brothers, for sake of a preprocessing step towards other upstream models.

7 REFERENCES

(1) UC Irvine Machine Learning Repository
"Speaker Accent Recognition." UCI Machine Learning Repository, 2020, https://doi.org/10.24432/C52329.

(2) American Speech-Language-Hearing Association
https://www.asha.org/Practice/multicultural/Phono/

(3) Speech Accent Archive

(4) Z. Ma and E. Fokoué, "A Comparison of Classifiers in Performing Speaker Accent Recognition Using MFCCs," Open Journal of Statistics, vol. 04, no. 04, pp. 258–266, 2014, doi: https://doi.org/10.4236/ojs.2014.44025.

(5) Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features," Interspeech 2016, Sep. 2016, doi: https://doi.org/10.21437/interspeech.2016-1148.

(6) A. Carofilis, E. Alegre, E. Fidalgo, and L. Fernández-Robles, "Improvement of Accent Classification Models Through Grad-Transfer From Spectrograms and Gradient-Weighted Class Activation Mapping," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2859–2871, 2023, doi: https://doi.org/10.1109/taslp.2023.3297961.

(7) A. Ahmed, P. Tangri, A. Panda, D. Ramani, and S. Karmakar, "VFNet: A Convolutional Architecture for Accent Classification," arXiv.org, 2019. https://arxiv.org/abs/1910.06697 (accessed Apr. 18, 2025).

(8) C. Wang, G. Peng, and B. De Baets, "Deep Feature Fusion through Adaptive Discriminative Metric Learning for Scene Recognition," Information Fusion, vol. 63, pp. 1–12, Nov. 2020, doi: https://doi.org/10.1016/j.inffus.2020.05.005.

(9) J. J. Bird, E. Wanner, A. Ekárt, and D. R. Faria, "Accent Classification in Human Speech Biometrics for Native and non-native English Speakers," Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 554–560, Jun. 2019, doi:https://doi.org/10.1145/3316782.3322780.

(10) M. Muttaqi, A. Degirmenci, and O. Karal, "US Accent Recognition Using Machine Learning Methods," 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), pp. 1–6, Sep. 2022, doi: https://doi.org/ 10.1109/asyu56188.2022.9925265.

(11) V. Mikhailava, M. Lesnichaia, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin, "Language Accent Detection with CNN Using Sparse Data from a Crowd-Sourced Speech Archive," Mathematics, vol. 10, no. 16, p. 2913, Aug. 2022, doi: https://doi.org/10.3390/math10162913.

(12) K. Reid and E. T. Williams, "Common Voice and accent choice: data contributors self-describe their spoken accents in diverse ways," Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–10, Oct. 2023, doi: https://doi.org/10.1145/3617694.3623258.

(13) Z. Qiu, "Mitigating Gender Bias in Accent Recognition Through Gender-Specific Models: An Analysis of Performance and Fairness," 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), pp. 230–234, Aug. 2024, doi: https://doi.org/10.1109/icsece61636.2024.10729385.

(14) J. Zuluaga-Gomez, S. Ahmed, D. Visockas, and C. Subakan, "CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice," INTERSPEECH 2023, pp. 5291–5295, Aug. 2023, doi: https://doi.org/10.21437/interspeech.2023-2419.

(15) P. C. English, J. D. Kelleher, and J. Carson-Berndsen, "Searching for Structure: Appraising the Organisation of Speech Features in wav2vec 2.0 Embeddings," Interspeech 2024, pp. 4613–4617, Sep. 2024, doi: https://doi.org/10.21437/interspeech.2024-2047.

(16) H. Zhu, G. Cheng, J. Wang, W. Hou, P. Zhang, and Y. Yan, "Boosting Cross-Domain Speech Recognition with Self-Supervision," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 471–485, 2024, doi: https://doi.org/10.1109/taslp.2023.3301230.

(17) A. G. Samuel and T. Kraljic, "Perceptual Learning for Speech," Attention, Perception, & Psychophysics, vol. 71, no. 6, pp. 1207–1218, Aug. 2009, doi: https://doi.org/10.3758/app.71.6.1207.

(18) R. Tatman, "Speech Accent Archive," Kaggle.com, 2017. https://www.kaggle.com/datasets/rtatman/speech-accent-archive/data (accessed Apr. 22, 2025).

(19) A. Aksënova et al., "Accented Speech Recognition: Benchmarking, Pre-training, and Diverse Data," arXiv.org, May 16, 2022. https://arxiv.org/abs/2205.08014 (accessed Jan. 30, 2024).

(20) Q. Zeng, D. Chong, P. Zhou, and J. Yang, "Low-resource Accent Classification in Geographically-proximate Settings: A Forensic and Sociophonetics Perspective," Interspeech 2022, Sep. 2022, doi: https://doi.org/10.21437/interspeech.2022-11372.

(21) A. Matos, G. Araújo, A. Candido, and M. Ponti, "Accent Classification is Challenging but Pre-training Helps: a case study with novel Brazilian Portuguese datasets." Accessed: Apr. 22, 2025. [Online]. Available: https://aclanthology.org/2024.propor-1.37.pdf?utm_source

Coding References

(22) A. Vaswani et al., "Attention Is All You Need," Jun. 2017. Available: https://arxiv.org/pdf/1706.03762

(23) A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Aug. 2020. Available: https://arxiv.org/pdf/2006.11477

(24) T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Jul. 2020. Available: https://arxiv.org/pdf/1910.03771