Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5351–5358
Marseille, 11–16 May 2020

AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition

Afroz Ahamad*†, Ankit Anand*, Pranesh Bhargava
BITS Pilani, India
{afrozsahamad, ankit0905anand}@gmail.com
pranesh@hyderabad.bits-pilani.ac.in

Abstract

Modern Automatic Speech Recognition (ASR) technology has evolved to identify the speech spoken by native speakers of a language very well. However, identification of the speech spoken by non-native speakers continues to be a major challenge for it. In this work, we first spell out the key requirements for creating a well-curated database of speech samples in non-native accents for training and testing robust ASR systems. We then introduce AccentDB, one such database that contains samples of 4 Indian-English accents collected by us, and a compilation of samples from 4 native-English, and a metropolitan Indian-English accent. We also present an analysis on separability of the collected accent data. Further, we present several accent classification models and evaluate them thoroughly against human-labelled accent classes. We test the generalization of our classifier models in a variety of setups of seen and unseen data. Finally, we introduce the task of accent neutralization of non-native accents to native accents using autoencoder models with task-specific architectures. Thus, our work aims to aid ASR systems at every stage of development with a database for training, classification models for feature augmentation, and neutralization systems for acoustic transformations of non-native accents of English.

Keywords: Speech Resource/Database, Prosody, Speech Recognition/Understanding

1. Introduction

In Sociolinguistics, accent is a manner of pronouncing a language. Anyone who speaks a language, does so in an accent. The way the native speakers of a language speak that language defines the standard pronunciation, and is generally considered to be the standard or reference accent for that language. When the non-native speakers of a language speak that language, say an Indian person speaking English, the phonological requirement of the non-native language, in this case English, interacts with the phonological knowledge of their first language, say Hindi. This influences their manner of speaking, giving rise to what is considered as the non-native accent.

Accents *per se* are interesting because they refer to a wide variety of social issues such as the acceptance of speakers into a community, indication of class in society, and linguistic issues such as those pertaining to the phonology of languages. This in itself warrants a better understanding of accents. However, there is another fundamental reason for studying accents. Speakers always have a manner of speaking and the speech always has accent. Since spoken communication is an important form of communication, studying accents becomes important to design technologies built to interact with human speech.

1.1. Indian Accents in English

Internet has led to English language becoming the linguafranca for conveying information about science, culture, sports and society in the world. The continued advancements in technologies supporting speech, in the form of audio and video media, has led to an increase in the usage of spoken English on the web. Since these speakers come from various different linguistic backgrounds, English language happens to be spoken in many different accents across the world.

English has become an important language of communication among the younger generation of India because of its status as the language of formal education. A large number of young Indians is bilingual, i.e. they speak one of the 22 Indian languages as their first language, alongside English. An implication of this is that when speaking English, the intervention from the phonology of their first language, e.g. Malayalam, gives rise to an accent in the speech of Indian speakers of English. This accent is generally very distinct and is readily identifiable, for example, as the Malayalam English accent, the Telugu English accent, the Bangla English accent, etc.

Interestingly, this younger generation of India is also a large and growing group of users of speech-based technology through hand-held devices and voice assistants. These voice assistants have become very good at identifying English spoken in a native accent. However, non-native accented speech continues to be a challenge for them. If the automatic speech recognition (ASR) systems of the voice assistants have *apriori* knowledge that the speaker is going to speak with a certain accent, the voice assistant may be primed to listen to certain features in the voice, which would lead to a greater performance accuracy. For the success of this technology, it becomes pertinent then to identify and process accents, apart from the semantic content of the speech. Due to the large number of speakers, and vast varieties of accent, English spoken within India is an excellent resource for creating and testing technology whose success is contingent on detecting, identifying and understanding the native and non-native accents.

* denotes equal contribution from the authors.
† currently at Google.

2. The Database

A key requirement for developing speech-based technology is the access to a well-curated database of speech samples. Some of the widely used datasets for specific ASR tasks are very well labelled, either manually or through automation. For example, Google's AudioSet (Gemmeke et al., 2017) is a massive dataset for audio event detection, that includes more than 2 million manually-labelled 10-second sound clips belonging to over 600 classes. Similarly, VoxCeleb (Nagrani et al., 2017) is a speaker identification dataset which contains audio clips extracted from interviews of celebrities.

In this section, we first establish certain key requirements for constructing an accent database that could be well-suited for ASR tasks. Then we survey a few existing accent datasets. Further, we discuss our approach and setup for collecting our database, AccentDB[1]. Finally, we present an analysis of the distribution of speech samples that constitute AccentDB.

2.1. Key Requirements

The following are some of the key requirements for an accent database suitable for ASR systems.

1. **Variety of Speakers:** In order to represent the speaker differences, the database should ideally contain spoken material from a wide range of speakers.
2. **Words vs. Sentences:** The pronunciation patterns for words spoken in isolation are different from when they appear in connected speech, due to the suprasegmental phenomena such as elision and assimilation (Ladefoged, 1993). Therefore, for the purposes pertaining to the processing of spoken sentences, the database should contain sentence-length material.
3. **Uniformity of Content:** For the sake of isolating and identifying accents, it is necessary to have uniformity in the speech material across speakers. One way to address this is to have all the speakers speak the same sentences, preferably at the same speed. A related requirement is for the speech material to be phonetically balanced, so that no specific phonemes get over-represented in the database.
4. **Semantic Requirement:** If the sentences are meaningful, it avoids semantic factors affecting the pronunciation of the sentences.

2.2. Existing Accent Databases

Various attempts have been made in the past at creating accent focused speech databases with varied data sources, speakers, accents and corpora. Teixeira et al. (1996) created a word database with 20 speakers for each accent from a total of 6 countries. They used a small corpus of around 200 isolated English words spoken twice in a row by each speaker. Hernandez et al. (2018) presented a collection of British and American accents in the form of utterances from non-playable characters of the video game, "Dragon Age: Origins (BioWare 2009)", with manual labelling of the accents done by three individuals.

Two of the most popular datasets used for accent-related tasks are: the Foreign Accented English (FAE) corpus (Lander, 2007), and the Speech Accent Archive (Weinberger and Kunath, 2011). FAE data comprises 4925 telephonic utterances by native English speakers of 22 different languages. The subjects spoke about themselves for 20 seconds and the recordings were rated on a 4-point scale to determine the strength of accent. The Speech Accent Archive is a crowd-sourced collection of speech recordings of readings of a passage (colloquially referred to as *"Please call Stella.")* in English. Information about speakers' demographic and linguistic background is publicly available [2]. The passage has been spoken by more than 2000 speakers covering over 100 accents and 30 languages, but a significant number of samples are not tagged with the correct accent. This is because the database is crowd sourced, and there is no independent supervision on the accent label that is assigned to a recorded audio sample. For instance, a speaker whose first language is Bengali/Bangla, might mark his samples as belonging to the Bangla accent, even if his Bangla accent is neutralized after living in the UK for many years. Another drawback of using such crowd sourcing approaches for collection of accent data is that neither the recording environment, nor the recording hardware are consistent across speakers. This leads to the introduction of significant noise in samples. The lack of correct label for each sample adds to the difficulty of using any supervised learning algorithm for speech recognition tasks.

The birch canoe slid on the smooth planks.
Glue the sheet to the dark blue background.
It's easy to tell the depth of a well.
These days a chicken leg is a rare dish.
Rice is often served in round bowls.

Table 1: First 5 sentences of the Harvard Sentences dataset.

The CMU Festvox Project has a dataset titled CMU-Arctic (Kominek and Black, 2004) which contains speech samples in native English accents. In CMU-Indic, another dataset in the Festvox project, the content across the samples is not uniform as they are spoken not in one language with different accents, rather in different languages altogether. The samples here incorporate certain manifestations of an accent as well, as is evident from samples in any Indian language such as Gujarati, but the task of accent classification now entails modelling two attributes - the difference in utterances and the accent itself.

2.3. Introducing AccentDB

To fulfill the aforementioned key requirements and to avoid the issues faced by some existing databases, we created a multiple-pair parallel corpus of well structured and labelled data of accents. The database, AccentDB, contains speech recordings in 9 accents, split across 4 non-native accents of Bangla, Malayalam, Odiya and Telugu; 1 metropolitan Indian accent referred as "Indian" and 4 native accents namely American, Australian, British and Welsh. The number of samples, duration of all samples and the number of

[1] https://accentdb.github.io/

[2] Speech Accent Archive, George Mason University.

	Accent	Number of Samples	Duration	Number of Speakers
AccentDB	Bangla	1528	2 h 13 min	2
	Malayalam	2393	3 h 32 min	3
	Odiya	748	1 h 11 min	1
	Telugu	1515	2 h 10 min	2
	Total	**6184**	**9 h 6 min**	**8**
Amazon Polly	American	5760	5 h 44 min	8
	Australian	1440	1 h 21 min	2
	British	1440	1 h 26 min	2
	Indian	1440	1 h 29 min	2
	Welsh	720	0 h 43 min	1
	Total	**10 800**	**10 h 43 min**	**15**
Total		**16 984**	**19 h 49 min**	**23**

Table 2: The details of total 9 accents: 4 collected by the authors and 5 compiled using Amazon Polly.

Speaker Code	Native Language	Age of Speaker	Highest Qualification	English Usage
Ban-1	Bangla	24	Masters	19 yrs
Ban-2	Bangla	25	Masters	21 yrs
Mal-1	Malayalam	25	Masters	20 yrs
Mal-2	Malayalam	25	Masters	20 yrs
Mal-3	Malayalam	26	Masters	21 yrs
Odi-1	Odiya	33	Ph.D.	28 yrs
Tel-1	Telugu	26	Masters	21 yrs
Tel-2	Telugu	32	Ph.D.	17 yrs

Table 3: Demographic details of the speakers of speech samples in AccentDB database.

speakers per accent are listed in Table 2.
AccentDB is collected by employing the Harvard Sentences (IEEE, 1969) which are phonetically balanced sentences that use specific phonemes at the same frequency as they appear in English language. The sentences in this dataset are neither too short nor too long, making them suitable for proper manifestation of accents in sentence-level speech. Harvard Sentences dataset contains 72 sets, each consisting of 10 sentences. The first five sentences from this dataset are listed in Table 1. We ensure that the corpus is also parallel by recording a minimum of the same 25 sets across all 4 of the non-native accents. Additionally, we compile recordings of all the 72 sets across rest of the 5 accents.

2.4. Collection of Speech Data

The data for the 4 non-native accents, namely Bangla, Malayalam, Odiya and Telugu, was collected by the authors. For the task of recording speech samples, we recruited volunteers whom we identified to have strong non-native English accents in their daily conversations. Another requirement for these speakers was for them to be the native speakers of at least one Indian language since childhood. The demographics of the speakers can be found in Table 3. The data was collected in the form of audio recordings made inside a professionally-designed soundproof booth. The text of the sentences was presented to the participants on a computer screen through a web-app[3] designed specifically for this purpose. The participants were asked to read the text of the sentences aloud. The speech samples were recorded using the following equipment:

- Microphone : Audio Technica AT2005USB Cardioid Dynamic Microphone
- Recorder: Tascam DR-05 Linear PCM Recorder

Each set was repeated thrice to account for the speech variations in each sentence spoken by the same speaker.
For the 4 native accents, namely British, Welsh, American and Australian, and the metropolitan Indian accent, we generated speech samples by using Amazon Polly's Text-to-Speech API [4]. The API was used with a special speech synthesis markup formatted file[5] containing the Harvard Sentences.

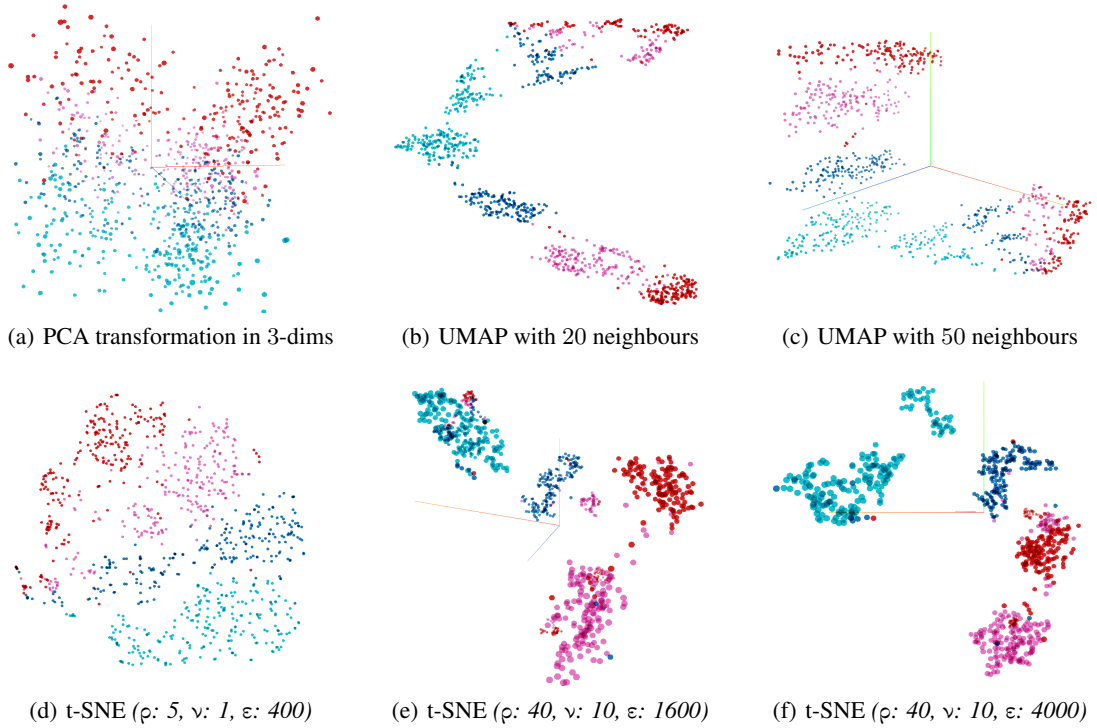
2.5. Cleaning and Post-processing

Any noise or other unwanted events (sneeze, giggle etc.) that were introduced while recording were sliced out using Audacity (Mazzoni, 1999) software. The cleaned audio files consisting of more than an hour-long recordings from each speaker were split on a pre-computed silence threshold to make one audio file per sentence. A split was created wherever the energy level was below 1.0 % for a duration of atleast 2 seconds. We then also trimmed silence slices at the beginning and the end of each sample to create richer data. These processed audio files were structured into directories tagged with the accent of the speaker.

2.6. Separability of AccentDB: An Analysis

Understanding the distribution of AccentDB speech recordings provides more insight into the quality of the collected data. To use the speech samples for any computational task or mathematical representation, they must first be converted to feature vectors. Mel-Frequency Cepstral Coefficient

[3] http://speech-recorder.herokuapp.com/
[4] https://aws.amazon.com/polly/
[5] HarvardSentences.ssml



(a) PCA transformation in 3-dims (b) UMAP with 20 neighbours (c) UMAP with 50 neighbours

(d) t-SNE *(ρ: 5, ν: 1, ε: 400)* (e) t-SNE *(ρ: 40, ν: 10, ε: 1600)* (f) t-SNE *(ρ: 40, ν: 10, ε: 4000)*

Figure 1: Projection of MFCC features for 4 non-native Indian Accents; (a)(3D): Principal Component Analysis (PCA) of feature vectors in 3-dimensions; (b)(2D), (c)(3D): Uniform Manifold Approximation and Projection (UMAP) with 20 and 50 neighbours; and (d)(2D), (e)(3D), and (f)(3D): t-distributed Stochastic Neighbor Embedding (t-SNE) projections with different perplexity*(ρ)*, learning rate*(ν)* and number of epochs*(ε)*.

(MFCC) extraction is a very widely used technique to represent audio files as vectors. The MFCC extraction of audio clips generally produces very high-dimensional vectors (for example, Guo et al. (2017) use 40 MFCC dimensions per audio frame). We concatenated the MFCC features of each frame to obtain high-dimensional acoustic vectors for the full-length of a clip. Since modelling the distribution of high dimensional data is difficult, we performed dimensionality reduction to obtain a set of principal variables and reduce the number of random variables under consideration. Dimensionality reduction techniques, when used for speech, learn projections of high-dimensional acoustic spaces into lower dimensional spaces.

The Principal Component Analysis on the acoustic vectors shows that the recordings from each accent in our collected database follows a definite convexity (Fig. 1(a)). We also performed Uniform Manifold Approximation and Projection with 20 and 50 neighbours to show that the speech samples from an accent are closer to each other (Fig. 1(b) & Fig. 1(c)). Further, t-SNE projections of the data (Figures 1(d), 1(e) & 1(f)) show the separability of the accents, establishing that the speech samples collected in AccentDB model their respective accents distinctively and are well-suited for use in machine learning tasks.

3. Accent Classification

Accent classification is an important step for tasks such as speech profiling and speaker identification. The current state-of-the-art ASR systems are already within the striking range of human-level performance with word error rates (WER) as low as $5.5\,\%$ (Saon et al., 2017). Accent classification can also be used to enhance ASR systems for better generalization towards unseen data by augmenting the training dataset with more relevant features (Ko et al. (2015); Park et al. (2019)). One such very relevant feature is present in human communication in the form of accent and hence, the task of accent classification has been crucial in the combined modelling of speech.

Over the years, multiple approaches have been used to tackle the tasks related to the accent classification for speech recognition. These include classical methods of Gaussian mixture models (GMMs) and Hidden Markov models (HMMs), machine learning models using Support Vector Machine (SVM), and very recently, deep neural architectures like Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM).

An early architecture that was proposed for this task by Teixeira et al. (1996) used parallel ergodic nets with context-dependent HMM units for word-level accent identification. Their system obtained a global accuracy score of 65.48% on their word-level speech data comprising 6 different accents. Ge et al. (2015) used purely acoustic features to build a GMM based accent classifier optimized using Heteroscedastic Linear Discriminant Analysis (HLDA). They used the FAE dataset (Lander, 2007) and achieved a success rate of 51% on 7 accents.

The recent advancements in deep learning architectures have proven to be a great success in a variety of speech recognition tasks including accent classification. In the work by Yang et al. (2018), the authors highlighted the im-

Task	Type	MLP	CNN	CNN (with attention)
Indian vs. Non Indian	2-class classification	100.0%	100.0%	100.0%
Non-native Indian Accents	4-class classification	98.3%	98.6%	99.0%
All accents	9-class classification	98.4%	99.3%	99.5%

Table 4: Classification accuracy of various models on three classification tasks.

Samples Used		Bangla,	Bangla,	Telugu,
Training Set	**Testing Set**	**Telugu**	**Malayalam**	**Malayalam**
$A_1P_1 + A_2P_1$	$A_1P_2 + A_2P_2$	82.86%	**90.06%**	79.75%
$A_1P_1 + A_2P_2$	$A_1P_2 + A_2P_1$	70.38%	74.68%	81.30%
$A_1P_2 + A_2P_1$	$A_1P_1 + A_2P_2$	73.35%	83.08%	**95.73%**
$A_1P_2 + A_2P_2$	$A_1P_1 + A_2P_1$	**96.18%**	76.15%	69.55%

Table 5: Model accuracy when training on speech samples from one speaker and testing on unseen samples from other speaker for 3 different accent pairs. A_iP_j denotes all speech samples from *j*-th speaker of the *i*-th accent.

portance of accent information for acoustic modeling and presented a joint end-to-end model for multi-accent speech recognition that achieves significant improvement on word-error rates. They used a bi-directional LSTM model with average pooling, and trained it with a Connectionist Temporal Classification (CTC) loss function. Bird et al. (2019) explored a variety of different techniques for accent classification on diphthong vowel sounds collected from speakers from Mexico and the United Kingdom. They achieved a classification accuracy of 94.74% using an ensemble model of Random Forest and LSTM.

3.1. Experiments

We ran classification experiments on our database using two standard baseline neural network architectures - a multi layer perceptron (MLP) and a CNN model. We evaluated the classification models in three different setups - (i) classifying amongst Indian accents collected in our database and non-Indian accents obtained from AWS Polly, (ii) classifying amongst the 4 collected Indian accents in our database, (iii) and finally classifying amongst all the 9 accents in AccentDB.

3.1.1. Preprocessing

Each audio file was divided into 10ms segments with a 1ms overlap between the segments. All the samples were less than 5 seconds in duration and hence padded to a standardized input dimension of 499. For each of these segments, we extracted 13 MFCC features. Hence, our final vector input for *n* audio files is of the dimension (n, 499, 13). This two-dimensional image-like vector for each audio file was used as the input to the first convolutional layer in all the CNN-based models. For MLP models, the input vector was created by flattening the image to one dimension.

3.1.2. Model Architecture and Training

The MLP model consists of multiple fully connected layers stacked together. The CNN model uses a combination of 1D Convolutional and Max Pooling layers, followed by multiple dense fully connected layers. For calculating the class probabilities, softmax activation was used in the final layer of each model. We used Adam (Kingma and Ba, 2014) and RMSProp (Tieleman and Hinton, 2012) optimizers, with a learning rate of 0.001 using cross-entropy loss function. Dropout was used in dense layers for regularization. A variety of batch sizes were tried during training to achieve the best results. As part of the evaluation, we used 20% of the total data present as test set for evaluation.

As the next step, we augmented the CNN network with attention. Attention mechanism has been successfully applied in machine translation (Bahdanau et al., 2014) and image captioning (Xu et al., 2015). Promising results have also been obtained for speech based tasks, e.g., in (Chorowski et al., 2015), where the authors solved the task of acoustic scene classification using a Convolutional, Long Short-term memory, Deep Neural Network (CLDNN) network and several attention-based LSTM models. We took this motivation further to apply attention mechanism onto the accent data to analyze the segments of the audio that are given more importance by our classification model. We used multiple variations of attention, firstly 1D and 2D variations based on the number of dimensions used. In the 1D version, attention vector is shared across the input dimensions, which correspond to the number of MFCC features used (13 in our case). For the 2D version, separate attention probability vectors were learnt for each input feature dimension. We also varied the layer to the output of which attention is applied.

3.2. Results

We evaluate our MLP, CNN and attention-CNN models on three different classification tasks, as described in the section 3.1. The accuracy results are summarized in Table 4. As is observable in the results, all the models performed exceptionally well, with the CNN models having a slight edge in accuracy as expected. Particularly in the binary classification setup, these models were able to detect the correct class with 100% accuracy. These instances of high accuracy can also be attributed to the presence of a quality dataset with good separability as discussed in section 2.6.

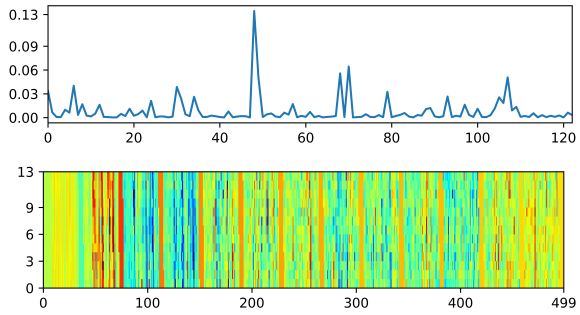


Figure 2: Time aligned attention scores with MFCC features corresponding to the sentence: *"Four hours of steady work faced us."*

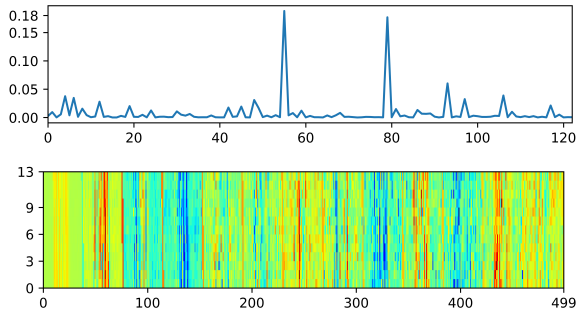


Figure 3: Time aligned attention scores with MFCC features corresponding to the sentence: *"It's easy to tell the depth of a well."*

3.3. Train on One, Test on Other

Speech classification models tend to overfit if they have a large number of trainable parameters but the training data is not extensive enough. This leads to poor generalization of models from training samples to unseen samples. To test if the models described previously, perform well even on unseen data, we evaluated our models in a challenging setup. Three accents in AccentDB - Bangla, Telugu and Malayalam were chosen for this experiment and the data for each accent was split into two, based on the speaker. We trained binary classifier models on sets of two accents by feeding them with only one half of the data of each accent (i.e. data of one speaker per accent). The models were tested on the unseen half data of each accent (i.e. the other speaker). Table 5 shows that our classifier models generalized well on the test data consisting of samples from other speakers even without seeing them during training.

3.4. Interpreting Attention

The attention scores that were obtained were analyzed by plotting them against the corresponding MFCC features for two audio files of Malayalam accent. In Figure 2, we observe a clear spike around the word *"Four"*, while in figure 3, the spikes correspond to timestamps around the words *"depth"* and *"well"*. These can be attributed to the different pronunciations of a particular phoneme sequence. For example, "depth" has the sounds that don't occur next to each other in the phonetics of Indian languages. So, each participant looks up to their own phonology to pronounce the word.

4. Accent Neutralization

State-of-the-art ASR systems often do not perform well on rare non-native accents, primarily due to the non-availability of good quality data for training such systems. We present our dataset on Indian Accents to augment training data for existing ASR systems to help make them more robust. ASR systems that perform well on native accents can further be improved for rare accents by performing accent neutralization. This means processing non-native audio file to make it sound like that of a native accent that the ASR system performs well on. The accent neutralization performed here involves extracting and transforming the para-linguistic and non-linguistic features of a source accent into those of a target accent while preserving the linguistic features. Acoustic feature conversions have been explored in other speech processing tasks as well. Toda et al. (2016) devised a challenge to better understand transformations of voice identity among speakers. For accent conversion, Kitashov et al. (2018) proposed a method to create accented samples of words by leveraging the difference between a dialect and General American English. Their model learns generalizations that would otherwise be created using rules written manually by phonologists.

With the success of neural networks in speech modelling, recent works have attempted end-to-end accented speech recognition. The experiments performed by Bearman et al. (2017) and Viglino et al. (2019) are on datasets that consist primarily of native accents (such as American, British, Australian, Canadian) and Indian as well, but the performance of models are underutilized due to the absence of non-native accents data. As reported in the next sections, we utilized the data collected in AccentDB to train and test deep neural networks on the task of non-native accent neutralization. We propose these transformation models to be used as an inference-time pre-processing step for ASR systems in order to overcome challenges associated with low resource accents.

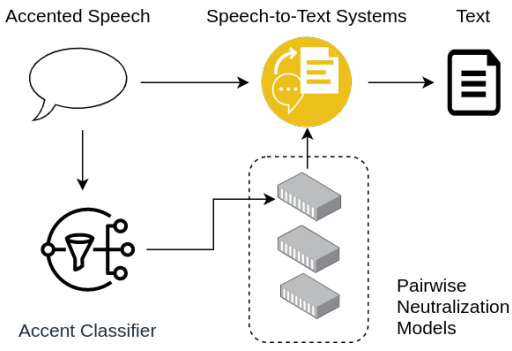


Figure 4: Use of pairwise accent neutralization in speech-to-text systems.

4.1. Pairwise Neutralization

A pairwise accent neutralization system consists of a set of individual models which can convert MFCC feature vectors of samples belonging to a source accent to those of samples belonging to a target accent. This set of individual converter systems can be used in conjunction with an accent classification system. An input audio file is routed to the converter corresponding to its predicted accent class from the classifier. The selected converter would be the one which can convert files belonging to this predicted accent to the given target accent. (Figure 4). The pre-processing step for this experiment is the same as that described in section 3.1.1.

4.1.1. Model Architecture

We trained a stacked denoising autoencoder network (Vincent et al., 2010) consisting of a series of convolution and pooling layers followed by deconvolutional (deconvolutional layers are required for upsampling) and pooling layers. The output of each layer is passed through a *tanh* activation function. Further, we train another similar network for evaluating reconstruction on reversed pairs where the source and target files are swapped. The autoencoder network's loss function is defined by feature-wise mean squared error between the input and output vectors. We used RMSProp optimizer (Tieleman and Hinton, 2012) with a learning rate of 0.001. The convolutional layers act as feature extractors for the input MFCC feature vectors and learn to encode them into a dense representation. The deconvolutional layers learn transformations on this dense representation for reconstruction into MFCC features of the target accent.

Source Accent	Target Accent	Accuracy	Accuracy (reverse)
Bangla	American	**99.21%**	98.67%
	Australian	99.02%	85.52%
	British	95.17%	97.36%
	Welsh	98.96%	**98.73%**
Indian	American	**98.77%**	95.23%
	Australian	98.63%	**95.77%**
	British	95.27%	92.64%
	Welsh	97.27%	90.48%
Odiya	Malayalam	**87.84%**	**93.11%**

Table 6: Classification accuracy on pairwise neutralization of 9 accent pairs.

4.1.2. Results

The reconstructed feature vectors obtained from the autoencoder model were evaluated on classification accuracy metric using CNN classifiers trained in section 3. We performed this experiment for neutralizing Bangla and Indian accents into 4 native accents. Our model performed very well on the non-native to native accent neutralization achieving an accuracy of $>95\%$ on all 8 experiments. We obtained an accuracy of $>85\%$ when converting from native to non-native accents as well. The model can also be used to neutralize a non-native accent into a different non-native accent as shown through the Odiya-Malayalam pair. The results in Table 6 show that the transformations learnt through our model can be used effectively as a preprocessing step in ASR systems enabling them to work well on non-native accents.

Model	1 source 2 targets	2 sources 2 targets
CNN Autoencoder + Skip Connections	52.15%	66.73%
LSTM Autoencoder + Skip Connections	53.46%	70.08%

Table 7: Classification accuracy on multi-source multi-target accent neutralization.

4.2. Multi-source Accent Multi-target Accent Neutralization

Extending the neutralization task for a set of n accents requires training $2 \times {}^{n}C_{2}$ pairwise neutralization models. Moreover, any device using this system would also require the source accent to be identified first before choosing a pairwise trained model to perform neutralization. To overcome both of these challenges, we present a single model that can be trained over multiple accents to neutralize samples from S_n number of source accents to T_n number of target accents.

4.2.1. Preprocessing

To train a single model with pairs of *(source, target)* samples belonging to multiple accents, we added an additional marker in each training pair, similar to zero-shot neural machine translation system proposed in (Johnson et al., 2017). The MFCC feature vectors of source accent samples were prefixed with a 13-dimensional one-hot encoded representation of accent label of target samples. Hence, the transformation of each input vector of dimension $(499, 13)$ is as follows:

$$S_i^{(499,\,13)} \rightarrow T_j^{(499,\,13)} \Rightarrow (L_{T_j} \,.\, S_i)^{(500,\,13)} \rightarrow T_j^{(499,\,13)}$$

where S_i denotes input files of *i-th* source accent, T_j denotes target files of the *j-th* accent, L_{T_j} denotes label of T_j*-th* target accent and (.) represents concatenate operation.

4.2.2. Experiments and Results

We used the prefixed inputs to run experiments in two setups for this task. We started with a set of 3 accents such that all the samples wee from same source accent and wee to be neutralized into two different target accents. We then experimented with the same set of 3 accents but now with samples from two different source accents. The convolutional autoencoder described in section 4.1.1. was augmented with skip connections to propagate target accent information in the form of label vector. This target accent label information is available in each layer up until the last one. We also experimented with a stacked LSTM autoencoder with skip connections. Table 7 compiles our preliminary results.

5. Conclusion and Future Works

We presented AccentDB, a well-labelled parallel database of non-native accents that shall aid in the development of machine learning models for speech recognition. Having a parallel corpus is better suited for tasks such as accent neutralization where each source sample should correspond to a target sample with the same vocabulary such that the differences in accent could be modelled easily. We evaluated accent classification models in a variety of settings and also discussed an interpretation of attention scores for analyzing audio frames. Finally, we showed the applicability of autoencoder models for accent neutralization. Future scope of our work includes enriching the database with more accents, and a larger variety of speakers in terms of age and gender. We would also like to add single-word database, ideally labelled for phonemes to have the data devoid of the effects of suprasegmental features.

6. Acknowledgements

This study was funded by the OPERA grant from BITS Pilani, provided to Dr. Pranesh Bhargava.

7. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bearman, A., Josund, K., and Fiore, G. (2017). Accent conversion using artificial neural networks.

Bird, J. J., Wanner, E., Ekárt, A., and Faria, D. R. (2019). Accent classification in human speech biometrics for native and non-native english speakers.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

Ge, Z., Tan, Y., and Ganapathiraju, A. (2015). Accent classification with phonetic vowel representation. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 529–533. IEEE.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Guo, J., Xu, N., Li, L.-J., and Alwan, A. (2017). Attention based cldnns for short-duration acoustic scene classification.

Hernandez, S. P., Bulitko, V., Carleton, S., Ensslin, A., and Goorimoorthee, T. (2018). Deep learning for classification of speech accents in video games. In *Joint Proceedings of the AIIDE 2018 Workshops 2018*.

IEEE. (1969). Ieee recommended practice for speech quality measurements. *IEEE No 297-1969*, pages 1–24.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Googleâs multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitashov, F., Svitanko, E., and Dutta, D. (2018). Foreign english accent adjustment by learning phonetic patterns. *CoRR*, abs/1807.03625.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *INTERSPEECH*, pages 3586–3589. ISCA.

Kominek, J. and Black, A. (2004). The cmu arctic speech databases. *SSW5-2004*, 01.

Ladefoged, P. (1993.). *A Course in phonetics.* Harcourt,, Firt Worth :, 3rd ed. edition.

Lander, T. (2007). Cslu foreign accented english release 1.2 ldc2007s08.

Mazzoni, D. (1999). Audacity ® software is copyright © 1999-2019 audacity team.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *CoRR*.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L., Roomi, B., and Hall, P. (2017). English conversational telephone speech recognition by humans and machines. *CoRR*, abs/1703.02136.

Teixeira, C., Trancoso, I., and Serralheiro, A. (1996). Accent identification. 01.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z., and Yamagishi, J. (2016). The voice conversion challenge 2016.

Viglino, T., Motlicek, P., and Cernak, M. (2019). End-to-end accented speech recognition. *Proc. Interspeech 2019*, pages 2140–2144.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, December.

Weinberger, S. and Kunath, S. (2011). The speech accent archive: Towards a typology of english accents. *Language and Computers*, 73, 12.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Yang, X., Audhkhasi, K., Rosenberg, A., Thomas, S., Ramabhadran, B., and Hasegawa-Johnson, M. (2018). Joint modeling of accents and acoustics for multi-accent speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.