Human vs. Machine Recognition of Malayalam Dental, Alveolar & Retroflex Voiceless Stops

Written by Rosemary Lach & Grace Walters

I. Introduction

Three-way place of articulation contrasts between dental, alveolar, and retroflex sounds are relatively rare in the languages of the world. This threefold distinction is attested in some Australian languages and in some Dravidian languages of South India, such as Malayalam (see Ladefoged & Johnson 2014, p. 176) . While Proto-Dravidian is reconstructed with three coronal gestures, the contrast present in Malayalam is not preserved in most other Dravidian languages, which retain instead a two-way dental-retroflex distinction (see Bhadriraju 2003). The disappearance of the Proto-Dravidian alveolar stops likely owes to the perceptual difficulty in distinguishing this three-way contrast, which then resulted in the gradual absorption of alveolars into other phonemes in the Dravidian languages, leaving the two coronal gestures with the furthest articulatory and perceptual distance between them behind. Nevertheless, Malayalam still does retain this contrast, despite the overall diachronic trends present in its language family.

This paper employs computational methods to test just how similar these three coronal gestures really are. In order to investigate whether these sounds are merely perceptually similar to the human ear or acoustically similar overall, the abilities of human subjects as well as different machine learning techniques to distinguish between the three coronal gestures in isolation were tested. The goal of this study is to determine whether human Malayalam speakers or a machine learning algorithm demonstrates greater proficiency in distinguishing dental, alveolar and retroflex stops when produced isolated in nonsense VCV sequences; if the computer were to outperform human native speakers of Malayalam, it might suggest that some acoustic differences exist between the three sounds that humans are not as privy to, as machine learning algorithms often clue in to distinctive features not noticed by humans to identify two separate entities.

Code for this project can be found at https://github.com/rosemarylach/ling430project.git.

II. Background: A Brief Note on Malayalam

Malayalam is a South Dravidian language spoken by approximately 35 million people within India, according to the 2011 census. Below is the language's consonant inventory as described by Asher & Kumari (1997):

	bilab	dent	alv	retro	pal	vel	glott
-voi -asp stop	p	**t**	**t̲**	**ṭ**	c	k	
-voi +asp stop	ph	th		ṭh	ch	kh	
+voi -asp stop	b	d		ḍ	j	g	
+voi +asp stop	bh	dh		ḍh	jh	gh	
fricative	(f)		s	ṣ	ś		h
nasal	m	n	n̲	ṇ	ñ	ṅ	
tap/trill			r, r̲				
lateral			l	ḷ			
approximant				ẓ			
glide	v				y		

Figure 1 – consonant inventory of Malayalam.

Of primary concern for this paper are the three stop consonants bolded in figure 1. First is the voiceless dental stop /t/, as in *kutira* ('horse'). Second is the voiceless alveolar stop /t̲/, as in *kut̲tam* ('fault'). Last is the voiceless retroflex stop /ṭ/, as in *kūṭe* ('with'). The dental and retroflex stops, when produced intervocalically, have been described as having 'slight friction'; likewise, when preceded by a nasal, they are realized as voiced stops at their respective places of articulation (see Asher & Kumari 1997, p. 412-413). The alveolar stop is unique in that it only occurs when geminated in medial position; singletons of this consonant are realized as /r̲/, and when preceded by a nasal, it is realized as /d/.

While these Malayalam stops were studied as a basis for the creation of the experiment described below, the speech sounds utilized in the recognition test were not intended to be exact replications of the Malayalam consonants, especially in their phonotactics (i.e., none of these sounds occur in Malayalam intervocalically as singletons, but rather appear as voiced allophones). Nevertheless, Malayalam speakers' knowledge of these three sounds as contrastive likely gives them an edge in recognizing each of the sounds in isolation, as they think of these sounds phonemically.

III. Methods

In order to test speaker and machine proficiencies in recognition of dental, alveolar and retroflex stops, each of the three stops was produced between 6 different vowels present in both

Malayalam and English – /u/, /o/, /i/, /e/, /a/, and /ə/ – yielding 18 total different VCV sequences that were then recorded. Each of the 18 unique recorded sequences was duplicated, then the resulting 36 recordings were randomized using an RNG to create the recognition testing set.

A. Speaker Testing

Two groups of human subjects were tested: 3 native Malayalam speakers (that is, Malayalam-English bilinguals) as well as 3 monolingual English speakers, all of whom are Rice University students. Malayalam speakers were tested in order to observe recognition proficiency in human speakers familiar with three-way coronal contrasts, while English speakers were tested to observe recognition proficiency in human speakers unfamiliar with three-way coronal contrasts; the scores of these two groups provided benchmarks by which the CNN and ensemble models could be evaluated. Each speaker performed the task in individual one-on-one meetings with the investigator.

The recognition test for Malayalam & English speakers was largely the same. Each participant was primed by the investigator with a description of the sounds they were tasked with identifying, identified as "sound #1," "sound #2," and "sound #3." For Malayalam speakers, example words were provided demonstrating each of the sounds in context (the same example words provided in section II, except *pa<u>tt</u>i* 'about' was provided rather than *ku<u>tt</u>am*, as the latter is a rather uncommon word). For English speakers, rough approximations of the three sounds in Malayalam were selected. For "sound #1" (dental), the example word was *birth*. For "sound #2" (alveolar), *stop* was provided; a /t/ in a cluster was selected to eliminate the aspiration present in English word- & syllable-initial alveolar stops, which is not present in Malayalam. "Sound #3" (retroflex) wss trickier; the word *heart* was eventually selected, as the /r/ preceding the /t/ would simulate the perceptual effect of a lowered F3 in retroflex consonants (see Ladefoged & Johnson 2014, p. 177).

Following this explanation of the sounds in question, the 36 recordings of the VCV sequences were played in the same order for each participant. Participants identified each sound as 1, 2 or 3 based on their intuition. They could ask for the sounds to be repeated multiple times. Following completion of the recognition test, participants were asked to characterize the three sounds based on their own intuitions, as well as prompted to share the cues they used to identify the three sounds. Finally, the investigator tallied up the number of correct responses for each of the three sounds and presented the results to the participant.

B. Machine Testing

In addition to the 18 test samples given to the listeners, we also recorded 10 additional samples of each sound between each vowel to create a total of 180 training sound files. These files were used to train 2 separate machine learning algorithms: an ensemble learning model prioritizing speed, based on features extracted from the sound waveform, and a convolutional neural network prioritizing accuracy, based on images of Mel frequency spectrograms. There are

2 folders with all of the testing and training sound files. One note about the transcription in the codebase - since Windows files are not case sensitive, to distinguish T from t in the filesystem, we write T as tt. There are four Python files that handle all of this analysis, and each of their functionalities are outlined below.

feature_extraction.py:

Running machine learning algorithms directly on sound files or images that contain information about the entire recording tend to be slow and complex because they look at every single data point (including irrelevant ones) in the sound file. In order to simplify and speed up the learning process, we decided to extract features directly from the sound file using the librosa library to create much smaller training data samples (only 91 elements) and save these to a csv.

The code in feature_extraction.py was structured to build a pandas dataframe of all of the desired features in both the training and test data. Additionally, the training dataframe included the phonemes to be used as labels during the learning. The dataframes contained one row for every sound sample, and a column for each feature. We chose this structure because it provided the simplest way to load this data into our model since it was the same general format required by the learning models in the scikit-learn library. After constructing the dataframes row by row, we exported them to separate .csv files that were saved to our main directory. We did this feature extraction process on both the labeled training data and the unlabeled test data to create a separate csv file for each.

Short descriptions of our 88 selected time and frequency domain features are outlined below. The starred signals indicate that our stored value was just a mean and standard deviation that represented the measurements taken at multiple frames in the data.

- **Central Moments:** mean, standard deviation, skewness, and kurtosis of the signal
- **Root Mean Squared Energy*:** Defined as $\sum_{n=1}^{N} |x(n)|^2$.
- **Mel-Frequency Cepstral Coefficients*:** a representation of the short-term power spectrum of a sound**.** We calculated 20 coefficients to take mean and standard deviation of to provide 40 features.
- **Chroma*:** A vector corresponding to the total energy of the signal in each of the 12 pitches (C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, B). We calculated the means and standard deviations to provide 24 features
- **Spectral Centroid*:** The frequency around which most of the energy is centered.
- **Spectral Contrast*:** Measures the decibel difference between spectral peaks and valleys in each frequency subband. We calculated the means and standard deviations at 7 frequency bands to provide 14 features
- **Spectral Rolloff*:** The frequency where a certain percentage of the total spectral energy (in our case, 85%) lies below.

In addition to extracting these features, another entirely separate method we used to summarize each sample was to create a Mel spectrogram. The Mel spectrogram was chosen over a standard spectrogram, because it is designed to model human hearing perception which is more ideal for phoneme classification. Additionally, in the ideal case where we would have recorded our training data with multiple speakers, the Mel spectrograms would be normalized, so phonemes could still be identified instead of speakers. We also used librosa to generate and plot each Mel spectrogram, and then saved each to a separate png file in either trainspecs or testspecs depending on which dataset the sample was from. These pngs could then be used later to classify phonemes based on the images instead of the signal features described above.

ensemble.py

This is the file where we generate an ensemble learning model to predict the coronal phonemes of the test data. Ensemble learning refers to using any combination of machine learning algorithms and aggregating the results of each to create a more accurate prediction. To train this model, we first extract all training and test features from the csvs generated by feature_extraction.py. We then normalize the data and run it through sklearn's built-in functions for k-nearest neighbors (KNN), logistic regression, a multilayer perceptron (MLP) model, a state vector machine (SVM), and random forest. For each of these models, instead of outputting a single prediction, it outputs the probabilities of predicting each phoneme. We then aggregate these probabilities with different weights depending on how effective the models performed individually during initial testing (1.3 for KNN, 1 for logistic regression, 1.1 for MLP, 1 for SVM, and 2 for random forest). The final phoneme prediction comes from the highest probability guess after performing this weighted sum of the individual probabilities. The predictions are then stored in a csv with filename in the first column and predicted phoneme in the second. This algorithm was designed to be less accurate, but run very quickly.

img-neural-nets.py

This is the file where we run the MEL frequency spectrograms generated in feature-extraction.py through a convolutional neural network (CNN). We chose to do the CNN on the spectrograms instead of the original sound file because it is a different approach than we used in class, but also because it allows us to then analyze the CNN by physically highlighting on the image which features were useful in distinguishing the phonemes from one another.

We first read in all of the training and testing spectrogram images into numpy arrays that can pass into a keras CNN model like we read in class. After doing some research on what CNN layers are useful for image classification, and experimenting with different batch sizes and epochs, we saved the model that was best at predicting the test data. In tuning the parameters, we had to be careful to give the computer enough time to learn (otherwise it would predict the same phoneme for all spectrograms), but not too much time that we overfit our training data. Once we generate the predictions, we store them in a csv with the filename in the first column and predicted phoneme in the second. This algorithm was designed to be more accurate despite

running slower, and it also provides a stepping point to learn more about the features of the sounds that aid in distinguishing them from each other.

After the model was generated, we generated a heatmap for the last convolutional layer of the CNN. We determined the name of this layer using model.summary(), but this name may have to be updated in the code in a new instance of the kernel. After normalizing and rescaling the heatmap, we overlay it on the original image. Essentially, this heatmap shows in yellow the parts of the spectrogram that were most useful for this layer in classifying the image. All that needs to be changed in this portion is updating the name of the file to create the heatmap for.

accuracy.py

This is a quick file to score the performance of both the ensemble learning and spectrogram CNN algorithms. Each just outputs its predictions to a csv file which can be seen in the github. We then went through the testing data and generated another csv file with the correct labels for each file. This script compares the predicted labels to the correct ones, and prints out the number correct, number incorrect, and percent accuracy.

IV. Results

Malayalam speakers outperformed English speakers in the phoneme recognition test, as was expected. The CNN outperformed all except for one human participant (a native speaker of Malayalam) in the recognition test, and the ensemble model outperformed two human participants. More detailed results are provided in the following sections.

A. Speaker Results

The performance of Malayalam native speakers and English monolinguals is shown below in figure 2 (see next page).

Surprisingly, the difference between average Malayalam and English speaker performances was not as broad as one might expect. The average performance of a Malayalam speaker on the recognition was 23.3/36 – a 64.7% proficiency rate. The average English speaker performance was 19/36, yielding a 52.7% proficiency score. A t-test for two independent means was performed to determine if the difference between Malayalam and English speaker mean proficiency scores was statistically significant, and it was not at $p > 0.05$. This likely owes to our small sample size, a limitation that could not be avoided considering the small scale of this project. Still, Malayalam speakers did perform better than English speakers overall on the recognition task.

Phoneme Recognition Test Data - Human Participants

			Malayalam-English Bilingual Participant Responses			English Monolingual Participant Responses		
Sound File #	Vowel Used	Correct Answer	Speaker #1	Speaker #2	Speaker #3	Speaker #4	Speaker #5	Speaker #6
1	a	alveolar	1	1	2	2	3	1
2	i	dental	2	1	1	1	2	1
3	i	retro	2	3	3	3	2	3
4	e	dental	2	1	1	1	2	1
5	e	retro	3	3	3	3	1	3
6	i	alveolar	2	1	1	2	2	1
7	a	dental	1	1	1	1	3	1
8	u	dental	1	2	1	1	1	1
9	u	retro	2	3	3	3	2	3
10	uh	dental	1	1	2	2	2	2
11	e	alveolar	1	2	2	1	1	1
12	i	alveolar	1	2	1	1	2	2
13	e	dental	1	1	2	1	1	1
14	o	dental	3	1	1	2	3	2
15	o	retro	2	3	3	3	2	3
16	u	alveolar	2	2	2	1	2	2
17	i	retro	2	3	3	2	2	3
18	o	dental	1	1	2	1	3	1
19	o	alveolar	3	2	1	2	3	1
20	a	dental	1	1	2	1	3	2
21	uh	retro	3	3	3	3	2	3
22	a	alveolar	1	2	2	2	1	2
23	a	retro	1	3	2	2	3	3
24	u	retro	3	3	3	3	2	3
25	a	retro	1	3	1	3	2	3
26	uh	alveolar	3	1	2	1	3	1
27	uh	retro	2	3	1	2	1	3
28	e	retro	2	3	3	2	2	1
29	o	retro	3	3	3	3	3	3
30	uh	alveolar	3	2	2	1	1	2
31	u	alveolar	1	2	2	2	2	2
32	u	dental	2	1	1	1	1	1
33	e	alveolar	2	1	2	1	3	1
34	o	alveolar	2	1	2	1	1	2
35	i	dental	1	1	2	2	2	2
36	uh	dental	1	1	2	1	3	1
		TOTAL CORRECT:	16/36	30/36	24/36	23/36	9/36	25/36

Figure 2 – results of phoneme recognition test with human participants. Correct answers are shown in black text, incorrect ones are shown in red. Dentals are ‘1,’ alveolars are ‘2,’ and retroflexes are ‘3.’

Detailed Scores

Speaker #1	dental	8/12
	alveolar	4/12
	retroflex	4/12
Speaker #2	dental	7/12
	alveolar	11/12
	retroflex	12/12
Speaker #3	dental	5/12
	alveolar	10/12
	retroflex	9/12
Speaker #4	dental	9/12
	alveolar	6/12
	retroflex	8/12
Speaker #5	dental	3/12
	alveolar	4/12
	retroflex	2/12
Speaker #6	dental	8/12
	alveolar	6/12
	retroflex	11/12

Participant Commentary

says the sounds in his head before guessing. tries to put the sounds in sentences in malayalam words to test.
retroflexes are "more bobo," dentals and alveolars are "more kiki" (in reference to bouba-kiki effect). says dentals seem like they move, while alveolars feel like a pause. says retroflexes feel "like a pot. like filling up a pot."
focuses on the shape of his tongue when trying to identify, repeats the sounds in his head. says alveolars are "softer," retros are "dull."
says dentals are a "th sound," retroflexes have an "r in front" and are "softer." says alveolars are "none of the above."
about a retroflex: "that's just an r." also on retros: "I'm trying to do where my tongue goes & that's just not working." tries to identify sounds my mimicking them. if it "sounds like d," goes with alveolar.
says alveolars are "sharper," dentals are "softer." says dentals sound like d, alveolars sound like t, retroflexes sound "roly-poly" and "more round." also says retroflexes have "a bit of an r."

Figure 3.1 – detailed scores of human participants on the phoneme recognition test.
Figure 3.2 – records of participant commentary when asked to characterize each of the three sounds.

Mistake/Confusion Stats

	1 as 2	1 as 3	2 as 1	2 as 3	3 as 1	3 as 2
Sp #1	3	1	5	3	6	6
Sp #2	1	0	5	0	0	0
Sp #3	6	0	3	0	2	1
Sp #4	3	0	7	0	0	4
Sp #5	4	5	4	4	2	7
Sp #6	4	0	5	0	1	0
TOTAL:	21	6	29	7	11	18

Figure 4 – Which sounds were most commonly mistaken as which others by individual speakers as well as overall.

Additional data on the performance of English and Malayalam-speaking participants is provided in figures 3 & 4. While not all of it is directly relevant to the paper's main question, it sheds light on which sounds are perceptually easier to distinguish for both English and Malayalam speakers, as well as which sounds are frequently confused, and how sounds are characterized by linguistically uninformed speakers of both English and Malayalam. Particularly

of note is the frequency with which alveolars were confused with dentals or alveolars, as opposed to dentals and retroflexes being confused with one another, as shown in figure 4. Dental-alveolar confusion happened a total of 50 times, twice as much as retroflex-alveolar confusion, which happened 25 times; each of these is markedly greater than the 17 times retroflexes and dentals were confused. This finding confirms that the perceptual similarity between dentals and retroflexes is much greater than that of alveolars with either one of them, strengthening the argument that alveolars were lost in many modern Dravidian languages in order to enhance perceptual clarity in the coronal series.

B. Machine Results

Ensemble Learning:

Out of the 18 test files, the ensemble learning algorithm got 10 correct and 8 correct for an accuracy of 55.555%. This performance was better than 2 of our 6 speakers and still better than random guessing. The upside of this version is that it runs almost instantaneously and uses very few computational resources. See figure 5 for a comparison of the ensemble predictions relative to the correct labels. We can see that the model was able to very easily recognize retroflex stops, but it had difficulty distinguishing the alveolars and dentals.

Spectrogram CNN:

Out of the 18 test files, the spectrogram-based CNN model got 13 correct and 5 incorrect for an accuracy of 72.222%. This improved score was expected since this is a more complex network. It outperforms all but one of our speakers. The downside of this version is that it takes multiple minutes to run and consumes about 70% of the computer's memory while running. See figure 5 for a comparison of the CNN predictions relative to the correct labels. We can see that this model was also able to easily recognize the retroflex stop; it also did a slightly better job than the ensemble model at distinguishing alveolars and dentals.

filename	correct label	ensemble prediction	CNN prediction
ata-7_testing.wav	t	th	t
atha-1_testing.wav	th	th	th
atta-13_testing.wav	T	T	T
ete-10_testing.wav	t	th	t
ethe-4_testing.wav	th	t	t
ette-16_testing.wav	T	T	T

ithi-3_testing.wav	th	t	t
iti-9_testing.wav	t	th	t
itti-15_testing.wav	T	T	t
otho-5_testing.wav	th	th	th
oto-11_testing.wav	t	th	T
otto-17_testing.wav	T	T	T
uhthuh-2_testing.wav	th	t	th
uhttuh-14_testing.wav	T	T	T
uhtuh-8_testing.wav	t	th	T
uthu-6_testing.wav	th	th	th
uttu-18_testing.wav	T	T	T
utu-12_testing.wav	t	t	t

Figure 5 – results of phoneme recognition test with the 2 machine learning algorithms.

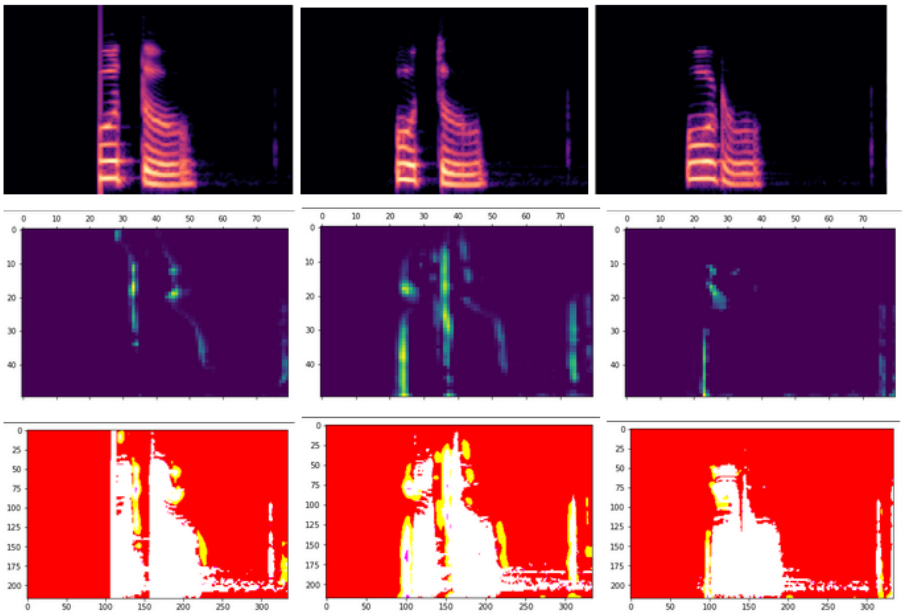

Figure 6 –
from top to bottom: spectrograms, heatmaps, heatmaps overlaid on spectrograms;
from left to right: dental, alveolar, retroflex

Heatmap Generation:

We specifically analyzed the spectrograms and heatmaps for each of the three coronal phonemes between the /u/ vowel, as this was the environment in which both the machine and human participants demonstrated the highest proficiency in distinguishing the different sounds; analyzing these waveforms in particular would thus best show the regions of the spectrogram that were most important to the last convolutional layer of the neural network. Figure 6 shows the original Mel spectrograms for each phoneme, the heatmap (important pixels in green/yellow), and the heatmap combined with the original spectrogram (important pixels in yellow). This illustrates that the CNN primarily looked at the edges between the stops and the surrounding vowel, as well as the effect the stop had on the outlining shape of the vowel spectrum to classify the phoneme.

V. Discussion

As was expected, the machine learning models outperformed human participants in the phoneme recognition task. This owes, of course, to the lack of natural biases and misconceptions inherent to the human condition; the computer, rather, deals strictly with data – in this case, raw images of spectrograms, pictorial representations of sounds – analyzing only what is given to it without external influences or past experiences.

One dissimilarity to note between human and computer recognition is that human participants were provided with audio stimulus, while the computer was fed graphic data in the form of spectrograms. While this was an intentional choice made by the researchers, as computer recognition of audio was covered in LING 430 and this paper aims to expand upon methods touched upon in class, it is possible that this could affect the performance of the computer, giving it an advantage over human participants.

Several interesting continuities are apparent between the performances of human participants and the two machine learning models. First, both human participants and the algorithms found retroflex stops easiest to identify out of the three, and second, both parties also performed best in the identification task when consonants were featured between the vowel /u/. This similarity between human and machine proficiencies illustrates that there likely is acoustic rhyme and reason to which sounds humans find most perceptually distinct, as participants exhibited the least hesitancy when identifying retroflex tokens as well, with one speaker (Speaker #2) even alleging that retroflexes were “a completely different sound.”

This study, however, has some limitations. Most notably, the sample size of human participants was small (n = 6), causing results to lack statistical significance. Additionally, training data for the CNN and ensemble models was limited (only 180 tokens), and these limited tokens were provided only by one speaker, who was also not a native speaker of Malayalam. Lastly, the performance of the Malayalam speakers may have been affected by the presentation of Malayalam sounds in phonotactic positions where they do not naturally occur in natively spoken Malayalam in the phoneme recognition test (i.e., Malayalam voiceless unaspirated stops

become voiced intervocalically in native speech), causing confusion for Malayalam speakers and thus skewing their proficiency scores. In order to improve this study, a larger sample size, more training data from multiple speakers (ideally native speakers of Malayalam) for the machine learning models, and tokens presenting Malayalam phonemes in their native phonotactic environments would be needed.

VI. Conclusion

As was anticipated, machine learning models exceeded human subjects in their ability to distinguish between the threefold dental-alveolar-retroflex contrast present in Malayalam. This likely owes to the lack of external influences and biases in machines that human listeners had to contend with. However, the CNN and ensemble methods did not utilize criteria overlooked by human participants; both speakers and computer found the same environments and consonants easiest to identify. This suggests that there are no acoustic differences between these three perceptually similar sounds that humans are not privy to. The computer found no more salient features or distinguishing characteristics of these three sounds than human participants did, illustrating the keenness of the human speech perception system in differentiating even the most similar of auditory signals.

VII. References

Asher, R. E., and T. C. Kumari. *Malayalam*. London : Routledge, 1997.

Gani, Areeb. "Visualizing Activation Heatmaps Using Tensorflow." Medium, Analytics Vidhya, 13 Jan. 2021, https://medium.com/analytics-vidhya/visualizing-activation-heatmaps-using-tensorflow-5bdba018f759.

Gupta, Vikas, and Anastasia Murzova. "Image Classification Using Cnns in Keras." LearnOpenCV, 5 May 2021, https://learnopencv.com/image-classification-using-convolutional-neural-networks-in-keras/.

Krishnamurti, Bhadriraju. *The Dravidian Languages.* Cambridge University Press, 2003.

Ladefoged, Peter, and Keith Johnson. *A Course in Phonetics*. Cengage Learning, 2014.

Mishra, Prateek. "Multiclass Image Classification Using Keras." Kaggle, Kaggle, 19 Sept. 2019, https://www.kaggle.com/code/prateek0x/multiclass-image-classification-using-keras/notebook.