# Syllable Based Deep Learning for Multi-Dialect Speech Recognition

[1]Rohan Gupta, [2]Rohan Kumar
[1]Student Researcher, [2]Student Researcher
The International School Bangalore

***Abstract*** **- As a by-product of India's linguistically diverse culture, it is estimated that dialects and accents change noticeably every 100km radius. Considering India's burgeoning technological industry as well as our own project to aid the visually impaired, in this paper, we explore methods to effectively apply Automatic Speech Recognition techniques across multiple dialects of Indian languages. We make use of a novel, syllable-based Hidden Markov Model in conjunction with a multi-layered, modified Artificial Neural Network to decode and analyze speech, specifically formal commands. Our research suggests that syllable-based speech recognition presents a lucrative mode of sequential pattern recognition, a consequence of linguistic properties within dialects and phonology of Indian languages.**

***keywords*** **- Deep Learning, Automatic Speech Recognition, Syllabification, Multi-Dialect**

## I. INTRODUCTION

India is home to world's largest visually impaired population, housing over 15 million people (out of the 37 million across the globe according to the 2018 census). It also tops the chart for the largest population of corneal blind in the world.

As part of our volunteering at ESHA - foundation for the blind - we visited several blinds schools to speak with the students and staff to learn about their lifestyles. We learned that their only source of learning resources was the braille library, which was limited in books due to their high cost and rarity. We further discovered that Braille literacy in India is less than 1% and falling. What's worse is that a visually impaired person not affiliated to an institution gets little to no access to knowledge and is also probably not Braille literate.

Since most knowledge resources being in English and print, we created an android application called CLABIL - Central Library of Audio Books in Indian Languages - to be an accessible source of audiobooks. We built our prototype using Android Studio with java, Django Rest Framework, and Google Cloud Speech-to-text. We linked it to a MySQL Database, where the content is stored, and created a Material User Interface for the app.

From autobiographies to school textbooks, CLABIL's open-sourced online library has access to learning material recorded in English, Hindi, Bengali, Gujarati, Marathi, Kannada, Tamil, Telugu, Urdu, and Sindhi.

With CLABIL, we wanted to give visually impaired students access to study materials like case studies, solved question papers and textbooks available to let them pursue a career of their choice. By creating an extensive library, they would have access to a more significant number of books allowing them to learn with greater depth and breadth. Lastly, it was made to be free of cost and accessible to everyone.

About 70 percent of disabled persons are found to be illiterate in rural areas as against 46 percent in urban areas. Only 4 percent of people with disability in rural India have an educational level "secondary and above" as against 12 percent in urban areas. We created CLABIL so that those who are only print illiterate can also have access to resources.

Our primary aim is to make learning resources accessible to people who do not have access, be it those belonging to the differently-abled community or marginalized groups.

***Features***

➢ Ability to download audiobooks giving offline access to listen at anytime, anywhere.

➢ Voice activated commands (Download, Pause, Play nth book, etc) allowing blind and low-vision users will be able to control functions, without the need to touch their screens.

➢ VoiceOver mode provides speech, output and screen magnification for the blind or low vision user, and refreshable Braille displays can be connected and used as well.

➢ CLABIL is easily customizable by the user, and one has access to many features and functions that can be turned on or off, such as sound alerts and verbosity settings.

## II. *Accessibility*

➢Blind Children

➢Older people who lose sight as they age.

➢People in semi urban / rural areas who are print disabled (not educated enough).

➢Girl children (and older women) who are not enrolled in schools.

➢Children who drop out of formal education for any reason and have no access to knowledge or literary resources.

➢Slum and economically disadvantaged children who do not have access to libraries or learning resources

We had to work on accessibility keeping the diversity of users in mind.
Firstly, we needed two different modes for the visually and non-visually impaired users. So we started by adding a disability status in the settings menu. This allows the visually challenged members to get access to copyright-free material granted by their disability status.

Overall, our goal was to create an obvious and simple way to enter, use, and exit the app for every CLABIL user. We then deployed it to a few blind schools for beta testing.

From the beta test results, we learned that both low-vision and blind users have very different needs, and we should consider them both as separate groups when designing app experiences. There was also difficulty navigating through the app, even with the gesture system in place. There were also requests to introduce a dark mode to be used in dimly light surroundings.

Incorporating these changes, we added sound and tactile feedback to communicate messaging in the form of multi-finger gestures. We also introduced VoiceOver mode where there is voice guidance to confirm when choosing something or switching pages (e.g. Switching from My Library page to Recently Added page). To allow it to be used in the dark, we added a light and contrast mode enabling users to get a much better print contrast when needed. For example, the double-tap gesture was not intuitive and required the user to be taught the command before using our app, which was not the accessibility we wanted for the long term. After some brainstorming, we decided that tactile and sound feedback was not the ideal approach for our overarching purpose of accessibility, so we turned to alternative methods.

## III. Speech Recognition

We decided to create and implement our own Speech Recognition software that converts speech-to-text that supports regional languages. With this, we could solve our accessibility problems since it allows for Voice Activated commands that can make navigation through app much more fluid and easy. Yet, we struggled with recognition accuracy due to lack of training data and differing accents. Furthermore, as we sought to expand beyond our regional approach, we encountered the fact that, as a by-product of India's linguistically diverse culture, dialects and accents differ significantly every 100km radius. We, therefore, had to develop a novel Speech Recognition procedure to tackle this problem, and it remains the focus of our research work.

Speech Recognition, in recent years, has grown to be a prominent software feature. Known technically as Automatic Speech Recognition (ASR), this feature describes the idea of computer software decoding the human voice and converting it to machine-intelligible instruction or text. ASR is commonly used to operate a device, perform commands (most pertinent to our research), or write without having to use a keyboard, mouse, or press any buttons.
Speech recognition is a crucial subfield within computational linguistics, with current research into this field-oriented towards the development of methodologies that enable recognition and translation of speech with greater accuracy and computational efficiency.
As described in earlier sections, the current and pressing problem for our research lied in the statistic of dialect changing noticeably every 100km radius. Given that India measures about 3000 km north to south and 3000 km east to west, and comprises around 22 major languages, it is clear that the number of dialect shifts is significant enough to adopt measures to deal with them. It is estimated that, as a by-product of India's linguistically diverse culture, there are over 720 dialect-accent variations.

## IV. Speech Recognition Methods

Speech Recognition has been a historically relevant problem. The adaptation of neural networks for Speech Recognition had been conceptualized around the 1980s, but scientific methodology for the same has indeed come a long way in improving techniques. Today's systems can indeed recognize millions of words with astounding accuracies, and there exist free, open-source applications for the same.

Hence, this section gives a brief overview of how the current Speech Recognition paradigm works and its relevance and applicability towards applications most pertinent to the diverse linguistic culture in India.
To capture speech, the primary software used is an Analog to Digital converter. Since speech in itself is a biological construct, it is naturally analog. A microphone or equivalent first reads the sound-wave and constructs an Amplitude over Time graph. Time intervals are approximated in blocks of fractions of seconds (depends on microphone quality). The height of the blocks determines, in some senses, its state, which is then given a number.

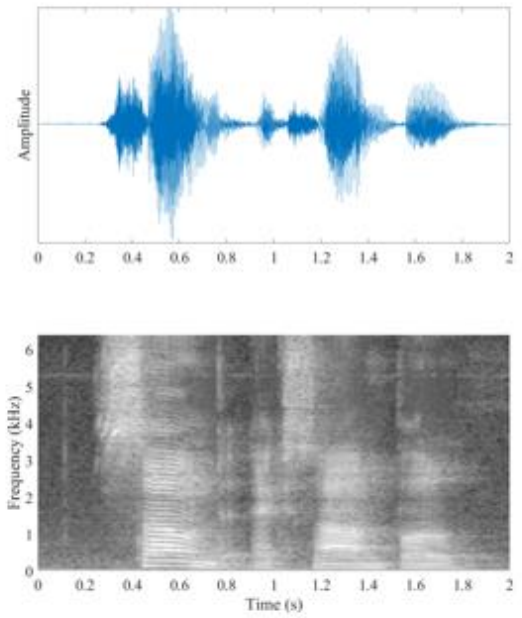


**Fig 1: A compartmentalized, classical amplitude/time graph and a spectrogram, for the same graph.**

This enables a conversion of analog-based sound to digital sound, where each time block is associated with a corresponding, numerical (can be represented digitally) state.

For successful speech recognition, however, one needs three elements of sound: frequency, intensity, and time taken. To glean the frequency from this graph, we use the Fast Fourier Transform, an elaborate mathematical method, to convert this graph to what is known as a *spectrogram*, which represents Frequency over Time.

Having established the digital bases, we now delve into the linguistics and computational machinery, which contributes to the decoding and recognition of speech.

## V. PHONEME BASED SPEECH RECOGNITION

In linguistics, a phoneme is a unit of sound, as small as 20-40 milliseconds, that distinguishes one word from another in a particular language. For example, in most dialects of English, the sound patterns are two separate words that are distinguished by the substitution of one phoneme for another phoneme. An example of a phoneme and its distinguishing properties are the two words “dumb” and “thumb,” where “d” and “th” respectively are distinguishing phonemes, allowing the words to be different.

Phonemes can be spoken differently by people of differing accents, gender, position within the word, and emotional states.


**Fig 2: All 44 phonemes in the English Language**

Hence, phonemes form the very basic building blocks of speech, which speech recognition software use to arrange in order to form a word, sentence, etc. Consequently, they are used in Speech Pattern Recognition models such as the one we used below:

***Hidden Markov Model (HMM)***

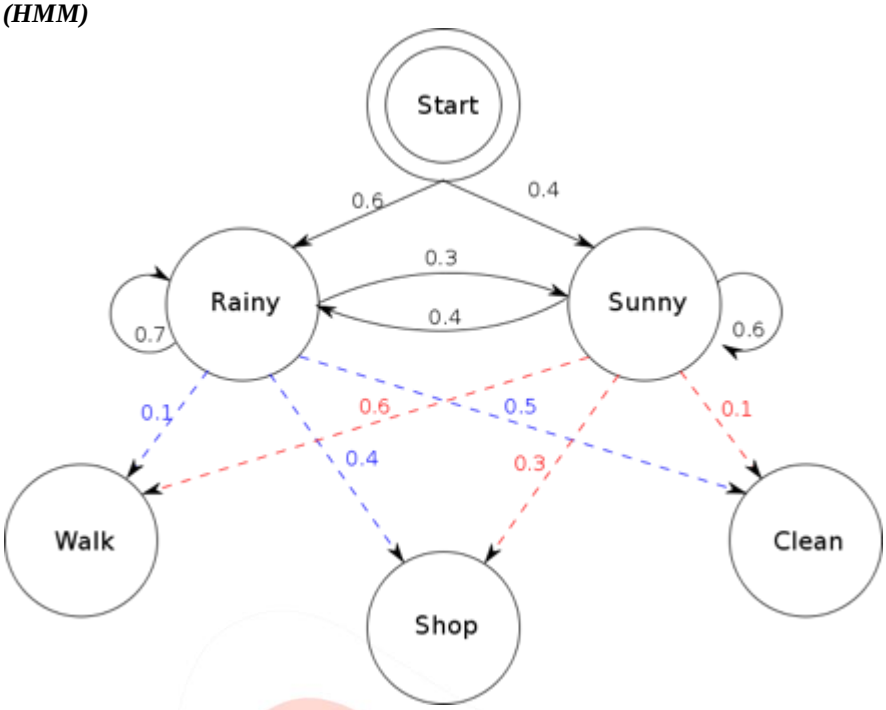


**Fig 3: A sample illustration of an HMM, describing high level actions**

The Hidden Markov Model (HMM) acts sequentially on phonemes by computing statistical probabilities for phoneme clustering i.e. the probability that one phoneme follows the other.

The HMM does this by using three different layers; the first layer checks, on an acoustic level, the probability that the phoneme discerned is the phoneme spoken.

Once the model has reached a decent probability (decent denoted by an arbitrary threshold), it moves on to the second layer, where the model checks the probability of phonemes being next to each other and chooses the most likely option, based on the results of the first layer.

For example, the sound (in English) "st" is very likely to be followed by a vowel sound like "a" and not a pure consonant such as "m."

Now, the model has grouped phonemes together and moves on to the third layer, where the software checks on the word level. The software first checks whether words standing next to each other are probable and if they conform to the grammatical rules of a particular language. For example, two verbs cannot be directly next to each other. Once again, this model works primarily based on probabilities and adapts sequentially to new words being processed, discarding improbable and incorrect assumptions as it progresses.

It is, therefore, understandable how the HMM is able to create intelligible sentences from given spectrograms. Since the HMM acts sequentially, it is aptly suited to the temporal pattern recognition problem present in speech recognition. However, the model is not particularly flexible, and there exists an exponential number and variety of phonemes. (about 40 that make up the English language).

***Artificial Neural Network (ANN)***

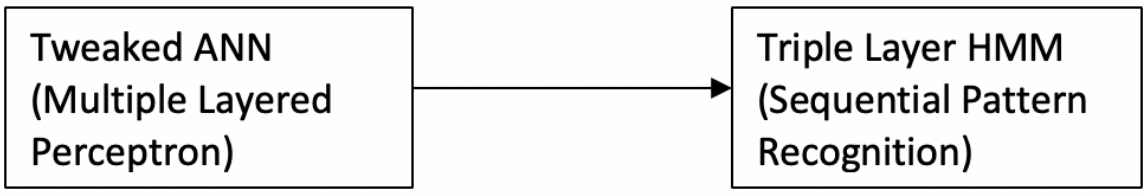


**Fig 4: A typical hybrid model of ANN and HMM**

These problems were partially solved by creating a hybrid model with an Artificial Neural Network. For this particular purpose, we use a standard neural network, with an input layer, hidden layer(s), and output layer. As in standard neural networks, outputs are governed by weights, optimized continually using a loss function, backpropagation, and large quantities of training data. However, for our purpose, we found it difficult to procure precise training data for our neural network and hence primarily rely on our HMM.

While a neural network might have been ideal for the task of identifying dialects, due to the lack of training data, we weight our HMM higher. Still, for the purposes of accuracy, we scoured literature within linguistics, finally arriving at syllables to solve our long-standing problem of dialect normalization.

## VI. Syllable Based Speech Recognition

### *Dialects*

A dialect is the variety of language that signals the geographic origin of a particular person. While there exist dialects of non-geographic nature, such as 'class dialect' and 'occupation dialect,' those are more or less suppressed in the Indian landscape. A dialect is chiefly distinguished from other dialects of the same language by features of its linguistic structure – i.e. grammar (specifically morphology and syntax) and vocabulary. A few key differences between dialects include differences in phonemes, vowel reduction in unstressed syllables, and phonological features (vowels, consonants, and intonation). Some linguists treat accents as separate, but for the purpose of our computational model, they can be encompassed under the umbrella of dialect.

### *Rationale*

Now, since dialects entail differences in phonemes, sentence structures, and even vocabulary choices (probabilistic frequencies), it is ostensibly too computationally (and procedurally) cumbersome to include a veritable database of all possible phonemes, their probabilities and so on.
However, having read linguistic journals, we note that differences in dialects, especially Indian dialects, are mostly seen in their

- ➢ Phonology (speech sounds)
- ➢ Condensed grammar
- ➢ Lexicon

Lexicon is too difficult to map and store in a database, and our purpose serves a finite, small set of commands, so we wish to tackle only differences in phonology and condensed grammar.

We learned that while there were differences in phonemes, differences in intonation or stress on those specific (replaced) phonemes varied less. Additionally, while pronunciation of words differed, specific syllables (that were used in place of one another) had similar linguistic properties, including their phonology.

To correct for this, we adopted a syllable based Hidden Markov Model (HMM), with a fourth layer focusing on syllable-matching. We compiled the most probable syllables, with higher weight to their intonation and verbal stress, only in the base languages, which was still computationally feasible, and hence created another layer of our HMM to solve this problem.

Additionally, within our Neural Network, we relaxed the weights on grammatic accuracy and soundness as well as syntactical structures, to facilitate for grammatical changes in differing dialects.

Hence, our hybrid model was complete and yielded better results for the purpose of speech recognition than a standalone Neural Network or HMM.

## VII. Conclusion

In conclusion, we note that the inclusion of syllable-based HMM, coupled with a suitably tweaked ANN results in much better multi-dialect speech recognition efficacy. Yet, there is still room for improvement. In most of our testing, albeit on a short dataset (not included here due to statistical biases), we achieved an accuracy of 89% as compared to 76% on our raw HMM. Some pressing shortcomings include the ever-changing linguistic culture in India (even influenced by globalization), which leads to accents and dialects changing. Since our project is aimed at a comprehensive improvement of India, we also aim to target marginalized sectors, whose dialects and language have not been appropriately studied yet.

In extension, having procured more data and knowledge of linguistics within India, one could adopt a Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) Network to improve efficacy even further. Certain, well optimized, networks present across the literature have been able to achieve success rates of 94%, albeit for the Polish language.

For the CLABIL project, since our work primarily centers around the functioning of a set list of commands, spoken in formal tone, our tweaks to the ANN and HMM should be read with caution. Additionally, while we have been able to map likely syllable formations in the base language reasonably, we have not been able to map vocabulary choices and all grammatical/syntactical changes (we have simply relaxed our ANN) due to sheer complexity of the task. Still, given the absence of data, we have been able to tackle phonological and grammatical changes to a suitable extent.

As India's flourishing linguistic culture continues to evolve, we hope and believe that it will be rightly accompanied with appropriate technology to procure data and gain a better understanding of an element crucial to India's welfare and pride – its plethora of languages. We look forward to improvements to our very own hybrid model, as well as revolutions in the ASR community towards a better, smarter, and linguistically intelligent future.

## VIII. References

[1] Sinha, K., & Tnn. (n.d.). India has largest blind population: India News - Times of India. Retrieved from https://timesofindia.indiatimes.com/india/India-has-largest-blind-population/articleshow/2447603.cms.
[2] Punani, B., & Rawal, N. (1993). Visual handicap: hand book.

[3] Prakash, Anusha & Jeena, Prakash & Murthy, Hema. (2016). Acoustic Analysis of Syllables Across Indian Languages. 327-331. 10.21437/Interspeech.2016-1127.
[4] Wolfram, W., & Schilling-Estes, N. (2006). American English: dialects and variation. Malden, MA: Blackwell.
[5] Gumperz, J. J. (1958). Phonological Differences in Three Hindi Dialects. Language, 34(2), 212. doi: 10.2307/410824
[6] Botros, N., Siddiqi, M., & Deiri, M. (n.d.). Automatic speech recognition using hidden Markov models and artificial neural networks. IEEE International Conference on Neural Networks. doi: 10.1109/icnn.1993.298825
[7] Aslyan, R. (2011). Syllable Based Speech Recognition. Speech Technologies. doi: 10.5772/16307
[8] Atherton, H. E., & Gregg, D. L. (1929). A Study of Dialect Differences. American Speech, 4(3), 216. doi: 10.2307/452359