Machine Learning for Non-native Human Spoken-word Recognition: Literature Review

Introduction

Speech recognition is the capacity of a machine, database or program to recognize phrases, expressions or words in the spoken language and translate them to a machine-readable layout. The detection of non-native speakers' dialog is, by itself, a very stimulating task. The discourse of speech recognition has been around for years, but the question in mind should be why it is now salient. The motivation is that deep learning has finally made speech recognition accurate enough to be useful outside of a carefully controlled environment.

Machine learning is the awareness that there are universal algorithms that can tell something fascinating and attention-grabbing about a set of data shorn of you having to inscribe any custom code particular to the problem. As an alternative of writing codes, data is inputted to the generic algorithm, and it constructs its logic grounded on the data. It suffices to mention that with actual world disposition, nonetheless, the miscellany of users, stresses serious considerations. Though the prerogative that all users should be presented identical access to speech recognition is not solid. This is synonymous to the analogy that people with pitiable reading skills do not have the equal access to newspapers as the vastly literate. In simple terms, Machine learning is a canopy term that shelters different kinds of generic algorithms.

Moreover, non-native recognition plays a key need for boundary control security systems. Such recognition systems help security officials to identify immigrants with a counterfeit and forged permit or ID by spotting actual country of articulated foreign accent. Also, it does appear that speech recognition applications are on a flight to become a default interface for information spreading systems. Housing users whose language use are somehow compromised are not just a research problem but also a momentous useful concern.

Methodology

There are few considerations, however that seem principally expedient for encoding non-native speech. Modes, choice, lexical, syntactic soundness, accent, and fluency are facets of spoken English that can both label disparity in native speech and be used to differentiate it from native speakers. "Accent commonly come from articulation habit of the speaker in his or her language". As it is widely known that beginners of a language are mostly exposed to preliminary grammar in the premature stage of their study, yet, imperfect mastery of syntax is

one of the sorts that can make an even vastly proficient speech as non-native. Lately, perception between native and non-native speech has been battered by means of binary classification structures. These frameworks essentially depend on prosodic, cepstral, speech recognition based or N-gram language sorts, and engage support vector machines (SVMs) for classification.

Automatic Speech Recognition (ASR)

Efforts to construct Automatic Speech Recognition (ASR) systems were first made in the 1950s. These initial speech recognition systems tried to relate a set of grammatical and syntactical rules to pinpoint speech. The system could only identify the word if the spoken words stick to a certain rule set. An Automatic Speech Recognition (ASR) module forms the root of virtually all the spoken Language evaluation systems. An ASR front-end component for most state-of-the-art evaluation systems gives word speculations about the responses given by the person available for the assessment. Consequently, it can be predicted that a huge amount of data, more precisely a pool of non-native speech, and careful transcriptions of each piece of that speech, would be required to train this type of ASR module. Moreover, there are no doubts that this would implicate human effort in transcribing the whole speech collection. Despite the advancement of Automatic Speech Recognition (ASR) systems that have led to supporters is still an issue in developing robust ASR systems that deliver high performance across diverse user groups.

The problem of the present ASR systems that these are working mostly with the native speech only, and the correctness affectedly lessens when words are articulated with an unusual pronunciation (foreign accent). However, human language has copious concessions to its guidelines. The way words and phrases are articulated can be enormously altered by dialects, assents, and mannerisms. First, there is a disparity in what is said by the speaker. For open vocabulary systems, there is no way to gather training data for every conceivable utterance or even every possible word. Second, there is dissimilarity due to differences between speakers. Different people have different voices and accents and ways of speaking. Third, there is variation in noise conditions. Anything in the acoustic data that is not the signal is noise, and thus noise can include background sounds, microphone specific artifacts, and other effects. Hence, to accomplish Automatic Speech Recognition, we make use of Deep Learning Algorithm. Therefore, for this study, Deep Learning Algorithm will be considered as our methodology. It may also interest one to know that Deep learning canvassers who know almost nothing about language translation are knitting together comparatively simple machine learning resolutions that are thrashing the best expert-built language translation systems in our world today.

Experiment and result

In machine learning, a neural network (Deep learning) is a construction especially used for grouping or regression tasks when the extraordinary dimensionality and non-linearity of the data make these tasks unlikely to accomplish. In the case of a visual data, the standard is to engage Convolutional Neural Networks (CNN). CNNs are directly inspired by the hierarchy of the cells in visual neuroscience.

It is important to note that the Neural network itself is not an algorithm, but rather a charter for many other machine learning algorithms to work collectively and process multifarious and complicated data inputs. Siniscalchi, et al. , (2013) already established that manner and place of articulation attributes can efficiently characterize any spoken language along the same lines as in the automatic speech attribute transcription (ASAT) model for Automatic Speech Recognition.

Conclusion and discussion

Non-native human speech is multifarious; thus, it constrains quite a few studies in their research to a selected language speaking group or nation. Programmed assessment of some sides of spoken language adeptness, including grammar, content appropriateness, vocabulary, and dialog rationality, profoundly on how correctly the input speech can be accepted. While state-of-the-art acoustic models constructed on deep neural networks have meaningfully upgraded recognition performance of native speech, correct recognition results are still puzzling to achieve when the input is unstructured non-native speech. This is due, in large part, to the point that non-native spoken reactions tend to encompass substantively complex amounts of pronunciation errors and flawed phrases. In other to determine the DNN model parameter for maximum accurateness, a good number of experiments have been carried out before extracting our data from its source. In order to measure the performance of the system of any data set and parallel the models generally, one could use the DET (Detection Error Trade-offs) curves approach or the AvgEER. Also, when the CNNs was compared with the DBN-DNNs, the CNN provides a reduction producing 18. 8 percent for the manner and 10. 3 percent for the place relative AvgEER improvement. It is not flawless that speech recognition technology has gotten to the point at which it can make judgments as to the exactness of pronunciations that corresponds to human judgments at an acceptable level. However, the truth is that our present world, with the constantly evolving technological knowledge base, we can now confidently say we are exposed to highly effective speech recognition systems like Google, Amazon, and so on, with minimal speech detection errors (even for non-native spoken-words) and somewhat close to being perfect.