Problems And Perspectives Of Automatic Speech Recognition

Speech recognition is the capacity to listen to talked words and distinguish different sounds exhibit in it, recognize them as words of some known language. In the computer area, the Speech recognition is might be characterized as the capacity of the computer framework to acknowledge talked words in the sound configuration like wav and afterward create its substance in the content arrangement. Automatic speech recognition (ASR) simulates human listening, it transforms speech into text.

This sort of conversion should be independent of vocabulary size, accent, speaker characteristics such as male or female etc. A more technical definition is given by Jurafsky, where he defines ASR as the building of system for mapping acoustic signals to a string of words. He continues by defining automatic speech understanding (ASU) as extending the goal to producing some sort of understanding of the sentence. Speech recognition is basically a pattern recognition problem. This involves extracting features from the input signal waves and classifying them to classes using pattern matching model. Performance of ASR system is measured on the basis of recognition accuracy, complexity, and robustness. Advantages of automatic speech recognition are accessibility for the deaf, cost reduction through automation, searchable text capability.

The problem of ASR is to program a computer to take digitized speech samples and print the words that a human would recognize when listening to the same sound. ASR has grown roughly in proportion to other areas of pattern recognition because of the desire to invent the machine capable of making complex decisions, and, practically, one that could function as swiftly as humans.

Although we have learned a great deal about how to build practical and useful speech recognition systems, there remain a number of fundamental questions about the technology to which we have no definitive answers. Unmistakably the speech signal is a standout amongst the most complex signals that we have to manage. It is created by a human's vocal system and in this way difficult to be described by a straightforward 2-dimensional model of sound spread. While there exist various advanced numerical models which endeavor to simulate the speech production system, their demonstrating capacity is as yet restricted.

In this work, we investigate the effectiveness of using the Bacteria Foraging Optimization Algorithm (BFOA) for optimizing the structure and parameter learning of HMM. The BFOA, proposed by Passino (Passino, 2002), is a newcomer to the family of nature-inspired optimization algorithms. For over the last five decades, optimization algorithms like Genetic Algorithms (GAs) (Holland, 1975) and (Benmachiche et al., 2016), Evolutionary Programming (EP), Evolutionary Strategies (ES) (Rechenberg et al., 1994), which are inspired by evolution and natural genetics, dominated optimization algorithms field. Recently natural swarm inspired algorithms like Particle Swarm Optimization (PSO) (Kennedy and Eberhart, 1996), Ant Colony Optimization (ACO) (Dorigo and Gambardella, 1997) have found their way in this research area and proved their effectiveness. Following the same trend of swarm-based algorithms, Passino proposed the BFOA (Passino, 2002). Furthermore, it is possible to design operators that favor biologically plausible changes to the structure of an HMM. That is, to ensure that modules of the states are kept intact. The objective of this work is to propose algorithms that improve this quality. The criterion used to quantify the quality of HMM is the probability that a given model generates a given observation. To solve this problem, we use as we have already mentioned a BFOA hybridization with HMM.

To our best learning, the point of the machine and human interaction is to utilize the most regular method of communicating, through our speech. The execution of the programmed ASR systems depends on traditional HMMs which depend on maximum likelihood estimation (MLE) systems. Different models have been investigated by specialists like neural and Bayesian systems, discriminative training strategies, state duration modeling and the use of support vector machines with HMM.

The most punctual endeavors to devise systems for ASR by machine were made in the 1950's. In 1952, at Bell Laboratories, Biddulph, and Balashek (Biddulph and Balashek, 1952) fabricated a system depended heavily on measuring spectral resonances during the vowel region of each digit. In 1970's speech recognition research achieved various huge turning points. To start with, the isolated word or discrete utterance recognition became a viable and usable technology based on the fundamental studies by Velichko and Zagoruyko in Russia.

In recognizing syllables or isolated words, the human auditory systems perform above chance level already at -18dB signal-to-noise ratio (SNR) and significantly above it at -9dB SNR. No ASR system is able to achieve performance close to that of human auditory systems in recognizing isolated words or phonemes under severe noisy conditions, as has been confirmed recently in an extensive study by Sroka.

Norris in 2008 presented a Bayesian model of continuous speech recognition, which is based on shortlist and shared many of its key assumptions: parallel competitive evaluation of multiple lexical hypotheses, phonologically abstract pre-lexical and lexical representations, feed-forward architecture with no online feedback, and a lexical segmentation algorithm based on the viability of chunks of the input as possible words.

Neural network models are powerful speech recognition engines. Their ability to classify data and ability in parallel processing pave the way for speech recognition. A typical neural network consists of the input layer, hidden layer, output layer. Input layer receives the input signal and transfers the data to the hidden layer. Hidden layer computes the action function and all the necessary calculations are done in this layer. After computations output is transferred to the output layer. Artificial neural networks are directed graph structure with nodes having some weights. Weights are initially random and are updated accordingly. Learning algorithms are used to classify the data. Back-propagation algorithm, iterative learning process, a multi-layer perceptron model as well as radial bias functions can be used to classify the data. Graves in 2013 investigated deep recurrent neural networks (RNNs), which combined the multiple levels of representation, that have proved so effective in deep networks with the flexible use of long-range context that empowers RNNs. When they trained end-to-end with suitable regularization, they found that deep Long Short-term Memory of RNNs achieve a test set error of 17.7% on the TIMIT phoneme recognition benchmark.

Support Vector Machines (SVM) supervised learning models with associated learning algorithms. They are used for classification and regression analysis. They analyze the data and recognize patterns. Text-independent speaker recognition uses as their features a compact representation of a spoken utterance, known as i-vector. Rather than estimating an SVM model per speaker, according to the one versus all discriminative paradigms, the Pairwise Support Vector Machine (PSVM) approach classifies a trial, consisting of a pair of i-vectors, as belonging or not to the same speaker class. Training a PSVM with a large amount of data, however, is a memory and computationally expensive task, because the number of training pairs grows quadratically with the number of training i-vectors. Among the numerous data selection techniques that have been proposed for binary SVMS, the ones that best fit to the problem are presented in but are computationally very expensive. In across training approach is proposed where the training data are split into non-overlapping subsets, which are used for training independent SVMs. The training patterns close to the average margin hyperplane are selected for training the natural SVM. This approach is interesting because the training procedure can be performed in parallel on each subset, but it has several drawbacks. Not only it is difficult to select meaningful non-overlapping subsets of i-vector pairs, but also this technique remains expensive for a large speaker set, and does not offer any guarantee that the average margin hyperplane is similar to the optimal hyperplane. Hierarchical parallel training is proposed in the cascade SVM approach of which is, however, ever more expensive than the formal because all the training patterns have to be scored by each SVM in the tree, and also because the procedure is iterative.

Generalized variable Hidden Markov model (GVP-HMM) is used for speech recognition in a noisy environment. A crucial task of automatic speech recognition systems is to robustly handle the mismatch against a target environment introduced by external factors such as environmental noise. When these factors are of time-varying nature, this problem becomes seven more challenging. To handle this issue, a range of model-based techniques can be used: multi-style training exploits the implicit modeling power of mixture models, or more recently deep neural networks, to obtain a good generalization to unseen noise conditions. An alternative approach to the above techniques is to directly introduce controllability to the underlying acoustic model. It is hoped that by explicitly learning the underlying effect imposed by evolving acoustic factors, such as noise, on the acoustic realization of speech, an instantaneous adaptation to these factors becomes possible.

The used language in this study is Arabic. It is well known that Arabic is the fifth most widely spoken language in the world with approximately 300 million native speakers, cutting across a wide geographical area from North Africa to the Middle East. It is also one of the six official languages adopted in the United Nations and represents the official language in some twenty-two countries, whereas there are substantial Arabic-speaking communities in many countries. Arabic is also the liturgical and worship language for more than one billion and a half Muslims worldwide.

Moreover, many challenges are faced by Arabic speech recognition. For example, Arabic has short vowels which are usually ignored in the text, which add more confusion to the ASR decoder. Additionally, Arabic has many dialects where words are pronounced differently. Elmahdy and Gruhn summarized the main problems in Arabic speech recognition, which include Arabic phonetics, discretization problem, grapheme-to-phoneme relation, and morphological complexity. Bourouba et al. (2006) presented a new HMM/support vectors machine (SVM) (k-nearest neighbor) for the recognition of isolated spoken words. Sagheer in (Sagheer et al., 2005) proposed a novel visual speech features representation system. They used it to comprise a complete lip-reading system. While Muhammad evaluated conventional ASR system for six different types of voice disorder patients speaking Arabic digits. Mel-frequency cepstral coefficients (MFCCs) and Gaussian mixture model GMM/HMM are used as features and classifier, respectively. The recognition result is analyzed for types of diseases.

Atal in Atal and Hanauer (1971), began a series of independent. They used a wide range of sophisticated clustering algorithms to determine the number of distinct patterns required to represent all variations of different words across a wide user population. In the 1980’s a shift in technology from template-based approaches to statistical modeling methods, especially the hidden Markov model approach.