AI for Speech Recognition using Pytorch

[1]S.NAVEENKUMAR, [2]V.NIVETHA, 3DR.D.THIYAGARAJAN

[1]Student, K S Rangasamy College Of Technology, Tiruchengode Tamilnadu.
[2]Assistant Professor, CSE Dept, K S Rangasamy College Of Technology, Tiruchengode Tamilnadu.

ABSTRACT

AI IS THE STUDY OF THE ABILITIES FOR COMPUTERS TO PERFORM TASKS, WHICH CURRENTLY ARE BETTER DONE BY HUMANS. AI HAS AN INTERDISCIPLINARY FIELD WHERE COMPUTER SCIENCE INTERSECTS WITH PHILOSOPHY, PSYCHOLOGY, ENGINEERING AND OTHER FIELDS. HUMANS MAKE DECISIONS BASED UPON EXPERIENCE AND INTENTION. THE ESSENCE OF AI IN THE INTEGRATION OF COMPUTER TO MIMIC THIS LEARNING PROCESS IS KNOWN AS ARTIFICIAL INTELLIGENCE INTEGRATION. WHEN YOU DIAL THE TELEPHONE NUMBER OF A BIG COMPANY, YOU ARE LIKELY TO HEAR THE SONOROUS VOICE OF A CULTURED LADY WHO RESPONDS TO YOUR CALL WITH GREAT COURTESY SAYING "WELCOME TO COMPANY X. PLEASE GIVE ME THE EXTENSION NUMBER YOU WANT" .YOU PRONOUNCES THE EXTENSION NUMBER, YOUR NAME, AND THE NAME OF THE PERSON YOU WANT TO CONTACT. IF THE CALLED PERSON ACCEPTS THE CALL, THE CONNECTION IS GIVEN QUICKLY. THIS IS ARTIFICIAL INTELLIGENCE WHERE AN AUTOMATIC CALL-HANDLING SYSTEM IS USED WITHOUT EMPLOYING ANY TELEPHONE OPERATOR.

KEYWORDS: *ASR -Automatic Speech Recognition, CNN - Convolutional Neural Network, RNN- Recurrent Neural Network*

I. INTRODUCTION

Speech processing is a unique discipline of signal processing. Study of speech signal and its processing method are the principles of speech processing. The speech processing application plays a major part in day-to-day life of commercial applications like Bank, Travel, Telecommunications and Voice Dialling. Some of the major growing applications are Language Identification, Speech Enhancement, Spoken Dialog System, Speaker Recognition and Verification, Speech Coding, Emotion and Attitude Recognition, Speech Segmentation and Labelling, Speech Recognition, Prosody, Text-to-Speech Synthesis, and Audio Visual Signal Processing. Input speech is given to the machine which accepts the command and translates into text format known as Speech Recognition System or Automatic Speech Recognition or Computer Speech Recognition or Speech to Text. Speech recognition systems analyze and train an individual speech that exploit to tune the recognition of specific voice which produces a more accurate result. . Speech recognition consist of Vector Quantization, Feature Extraction, Dynamic Time Warping, Hidden Markov Models, Gaussian Mixture Model, Decision-tree based Clustering, training with Expectation Maximization (EM), Language Models, Speaker Adaptation and Finite-State Formulation. In this paper, continuous and connected words are considered and MFCC features were extracted from the speech corpus. The extracted speech signal is trained by HMM model. Finally, the output result is compared with connected and continuous speech.

PROBLEM DEFINITION

The problem of face recognition can be stated as follows : Types of Speech Utterance An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences. The types of speech utterance are: 1) Isolated Words Isolated word recognizers usually require each utterance to have quiet on both sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. This is fine for situations where the user is required to give only one word responses or commands, but is very unnatural for multiple word inputs. It is comparatively simple and easiest to implement because word boundaries are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The disadvantage of this type is choosing different boundaries affects the results. 2) Connected Words Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them. 3) Continuous Speech Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation. It includes a

great deal of "co articulation", where adjacent words run together without pauses or any other apparent division between words. Continuous speech recognition systems are most difficult to create because they must utilize special methods to determine utterance boundaries. As vocabulary grows larger, confusability between different word sequences grows. 4) Spontaneous Speech This type of speech is natural and not rehearsed. An ASR system with spontaneous speech should be able to handle a variety of natural speech features such as words being run together and even slight stutters. Spontaneous (unrehearsed) speech may include mispronunciations, false-starts, and nonwords. B. Types of Speaker Model All speakers have their special voices, due to their unique physical body and personality. Speech recognition system is broadly classified into two main categories based on speaker models namely speaker dependent and speaker independent. 1) Speaker dependent models Speaker dependent systems are designed for a specific speaker. They are generally more accurate for the particular speaker, but much less accurate for other speakers. These systems are usually easier to develop, cheaper and more accurate, but not as flexible as speaker adaptive or speaker independent systems. 2) Speaker independent models Speaker independent systems are designed for variety of speakers. It recognizes the speech patterns of a large group of people. This system is most difficult to develop, most expensive and offers less accuracy than speaker dependent systems. However, they are more flexible. C. Types of Vocabulary The size of vocabulary of a speech recognition system affects the complexity, processing requirements and the accuracy of the system. Some applications only require a few words (e.g. numbers only), others require very large dictionaries (e.g. dictation machines). In ASR systems the types of vocabularies can be classified as follows. ***)** Small vocabulary - tens of words ***)** Medium vocabulary - hundreds of words ***)** Large vocabulary - thousands of words ***)** Very-large vocabulary - tens of thousands of words ♣ Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word Apart from the above characteristics, the environment variability, channel variability, speaking style, sex, age, speed of speech also makes the ASR system more complex. But the efficient ASR systems must cope with the variability in the signal.

II. TRAINING USING PYTORCH

RNNs are learning machines that recursively compute new states by applying transfer functions to previous states and inputs. Typical transfer functions are composed by an affine transformation followed by a nonlinear function, which are chosen depending on the nature of the particular problem at hand. It has been shown by Maass et al. that RNNs possess the so-called universal approximation property, that is, they are capable of approximating arbitrary nonlinear dynamical systems (under loose regularity conditions) with arbitrary precision, by realizing complex mappings from input sequences to output sequences . However, the particular architecture of an RNN determines how information flows between different neurons and its correct design is crucial in the realization of a robust learning system. In the context of prediction, an RNN is trained on input temporal data x(t) in order to reproduce a desired temporal output y(t). y(t) can be any time series related to the input and even a temporal shift of x(t) itself. The most common training procedures are gradient-based, but other techniques have been proposed, based on derivative-free approaches or convex optimization. The objective function to be minimized is a loss function, which depends on the error between the estimated output ˆ y(t) and the actual output of the network y(t). An interesting aspect of RNNs is that, upon suitable training, they can also be executed in generative mode, as they are capable of reproducing temporal patterns similar to those they have been trained on. The architecture of a simple RNN is depicted in

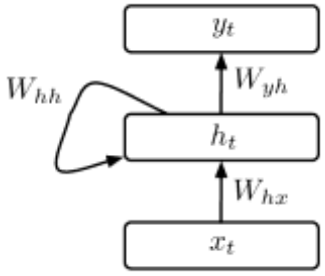

Fig. 1. In its most general form an RNN can be seen

o h h -1
as W h W h W I y h x z
Input layer
Hidden layer

Output layer
Figure 1: Schematic depiction of a simple
RNN architecture. The circles represent
input x, hidden, h, and output nodes, y,

h , W h o are the matrices which represent inputs, respectively. The solid squares W i h and W h hidden and output weights respectively. Their values are commonly tuned in the training phase through gradient descent. The polygon represents the non-linear transformation performed by neurons and z is the unit delay operator. A weighted, directed and cyclic graph that contains three different kinds of nodes, namely the input, hidden and output nodes. Input nodes do not have incoming connections, output nodes do not have outgoing connections, hidden nodes have both. An edge can connect two different nodes which are at the same or at different time

n

instants. In this paper, we adopt the time-shift operator z to represent a time delay of n time steps between a source and a destination node. Usually n = −1, but also lower values are admitted and they represent the so called skip connections. Self-connecting edges always implement a lag operator with |n| ≥ 1. In some particular cases, the argument of the time-shift operator is positive and it represents a forward-shift in time. This means that a node receives as input the content of a source node in a future time interval. Networks with those kind of connections are called bidirectional RNNs and are based on the idea that the output at a given time may not only depend on the previous elements in the sequence, but also on future ones. These architectures, however, are not reviewed in this work as we only focus on RNNs
with n = −1. While, in theory, an RNN architecture can model any given dynamical system, practical problems arise during the training procedure, when model parameters must be learned from data in order to solve a target task. Part of the difficulty is due to a lack of well established methodologies for training different types of models. This is also because a general theory that might guide designer decisions has lagged behind the feverish pace of novel architecture designs. A large variety of novel strategies and heuristics have arisen from the literature in the past the years and, in many cases, they may require a considerable amount of expertise from the user to be correctly applied. While the standard learning procedure is based on gradient optimization, in some RNN architectures the weights are trained following different approaches, such as real-time recurrent learning, extended Kalman filters or evolutionary algorithms, and in some cases they are not learned at all

SPEECH RECOGNITION FEATURE EXTRACTION

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. Because every speech has different individual characteristics embedded in utterances. These characteristics can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as: ♣ Easy to measure extracted speech features ♣ It should not be susceptible to mimicry ♣ It should show little fluctuation from one speaking environment to another ♣ It should be stable over time ♣ It should occur frequently and naturally in speech The most widely used feature extraction techniques are explained below. A. Linear Predictive Coding (LPC) One of the most powerful signal analysis techniques is the method of linear prediction. LPC [3][4] of speech has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech [10]. The analysis provides the capability for computing the linear prediction model of speech over time. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients. The following figure 2 shows the steps involved in LPC feature extraction. ( ) ( ) (/) argmax (/) argmax P X P W P X W W P W X w w = = Speech Input Text Output Feature Extraction Language Model Acoustic Model Dictionary Decoding Search WCSIT 2 (1), 1 -7, 2012 4

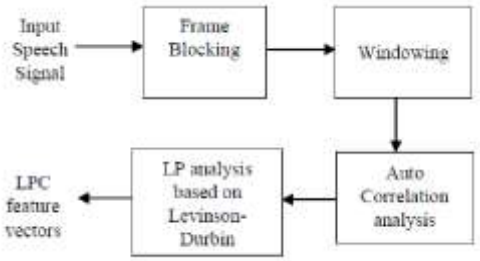

Figure 2. Steps involved in LPC Feature extraction

B. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC [3] [4] is the most evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC [6], it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation [3] [4]. It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by using the formula (2). Mel(f)= 2595*log10(1+f/700) (2) The following figure 3 shows the steps involved in MFCC feature extraction.

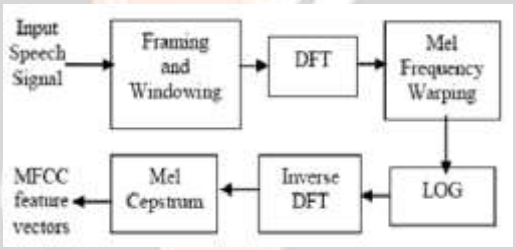

Figure 3. Steps involved in MFCC feature extraction

SPEECH RECOGNITION APPROACHES

In the earlier years, dynamic programming techniques have been developed to solve the pattern-recognition problem. Subsequent researches were based on Artificial Neural Network (ANN) techniques, in which the parallel computing found in biological neural systems is mimicked. More recently, stochastic modeling schemes have been incorporated to solve the speech recognition problem, such as the Hidden Markov Modeling (HMM) approach. At present, much of the recent researches on speech recognition involve recognizing continuous speech from a large vocabulary using HMMs, ANNs, or a hybrid form . These techniques are briefly explained below.

A. Template-Based Approaches

Template based approaches to speech recognition have provided a family of techniques that have advanced the field considerably during the last two decades. The underlying idea of this approach is simple. It is a process of matching unknown speech is compared against a set of pre-recorded words (templates) in order to find the best match (Rabiner et al., 1979). This has the advantage of using perfectly accurate word models; but it also has the disadvantage that the pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical. Template preparation and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred words. This method was rather inefficient in terms of both required storage and processing power needed to perform the matching. Template matching was also heavily speaker dependent and continuous speech recognition was also impossible.

B. Knowledge-Based Approaches

The use of knowledge/rule based approach to speech recognition has been proposed by several researchers and applied to speech recognition (De Mori & Lam, 1986; Alikawa, 1986; Bulot & Nocera, 1989), speech understanding systems (De Mori and Kuhn, 1992). The "expert" knowledge about variations in speech is hand-coded into a system. It uses set of features from the speech, and then the training system generates set of production rules automatically from the samples. These rules are derived from the parameters that provide most information about a classification. The recognition is performed at the frame level, using an inference engine (Hom, 1991) to execute the decision tree and classify the firing of the rules. This has the advantage of explicitly modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully, so this approach was judged to be impractical, and automatic learning procedures were sought instead.

C. Neural Network-Based Approaches

Another approach in acoustic modeling is the use of neural networks. They are capable of solving much more complicated recognition tasks, but do not scale as excellent as Hidden Markov Model (HMM) when it comes to large vocabularies. Rather than being used in general-purpose speech recognition applications they can handle low quality, noisy data and speaker independence. Such systems can achieve greater accuracy than HMM based systems, as long as there is training data and the vocabulary is limited. A more general approach using neural networks is phoneme recognition. This is an active field of research, but generally the results are better than HMMs. There are also NN-HMM hybrid systems Input Speech Signal Frame Blocking Windowing Auto Correlation analysis LP analysis based on LevinsonDurbin recursion LPC feature vectors Input Speech Signal Framing and Windowing DFT Mel Frequency Warping MFCC feature vectors Inverse DFT Mel Cepstrum LOG WCSIT 2 (1), 1 -7, 2012 5 that use the neural network part for phoneme recognition and the HMM part for language modeling.

D. Dynamic Time Warping

(DTW)-Based Approaches Dynamic Time Warping is an algorithm for measuring similarity between two sequences which may vary in time or speed [8]. A well known application has been ASR, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of HMM. In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. This technique is quite efficient for isolated word recognition and can be modified to recognize connected word also [8].

E. Statistical-Based Approaches

In this approach, variations in speech are modeled statistically (e.g., HMM), using automatic learning procedures. This approach represents the current state of the art. Modern general-purpose speech recognition systems are based on statistical acoustic and language models. Effective acoustic and language models for ASR in unrestricted domain require large amount of acoustic and linguistic data for parameter estimation. Processing of large amounts of training data is a key element in the development of an effective ASR technology nowadays. The main disadvantage of statistical models is that they must make a priori modeling assumptions, which are liable to be inaccurate, handicapping the system's performance.

III. CONCLUSION

Speech Recognition has been in development of more than 60 years. The various speech recognition methodologies and approach is available to enhance the recognition system. The fundamentals of SR system, various approaches existing for developing an ASR system are explained and compared in this paper. In recent years large vocabulary independent continuous speech has highly enhanced. From the review, it is concluded that HMM based MFCC feature is more suitable for speech recognition requirements and produces more good results than other models. In this paper, MFCC feature is extracted and the speech is trained by HMM model which is implemented for both connected and continuous speech. In order to improve the accuracy, other modeling techniques will be implemented in future. Researchers, working on the very promising and challenging field of automatic speech recognition, are collectively heading towards the ultimate goal i.e. Natural Conversation between Human beings and machines, are applying the knowledge from areas of Neural Networks, Psychoacoustics, Linguistics, Speech Perception, Artificial Intelligence, Acoustic-Phonetics etc.. The challenges to the recognition performance of ASR are being provided concrete solutions so that the gap between

recognition capability of machine and that of a human being can be reduced to maximum extent. An attempt has been made through this paper to give a comprehensive survey and growth of automatic speech recognition over the last six decades through the never ending efforts of researchers in countries like China, Russian, Portuguese, Spain, Saudi Arab, Vietnam, Japan, UK, Sri-Lanka, Philippines, Algeria and India.

REFERENCES

1] Davis, K., Biddulph, R., and Balashek, S., "Automatic Recognition of Spoken Digit," J.Acoust. Soc. Am. 24: Nov 1952, p. 637.

[2] Hemdal, J.F. and Hughes, G.W., A feature based computer recognition program for the modeling of vowel perception, in Models for the Perception of Speech and Visual Form, Wathen-Dunn, W. Ed. MIT Press, Cambridge, MA.

[3] Watcher, M. D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., "Template Based Continuous Speech Recognition", IEEE Transaction on Audio, Speech, & Language Processing, 2007.

[4] Samoulian, A., "Knowledge Based Approach to Speech Recognition", 1994.

[5] Tripathy, H. K., Tripathy, B. K., Das, P. K., "A Knowledge based Approach Using Fuzzy Inference Rules for Vowel Recognition", Journal of Convergence Information Technology Vol. 3 No 1, March 2008.

[6] Savage, J., Rivera, C., Aguilar, V., "Isolated word speech recognition using Vector Quantization Techniques and Artificial Neural Networks", 1991.

[7] Debyeche, M., Haton, J.P., Houacine, A., "Improved Vector Quantization Technique for Discrete HMM speech recognition system", International Arab Journal of information Technology, Vol. 4, No. 4, October 2007.

[8] Hatulan, R. J. F., Chan, A. J. L., Hilario, A. D., Lim, J. K. T., and Sybingco, E., "Speech to text converter for Filipino Language using Hybrid Artificial Neural Network and Hidden Markov Model", ECE Student Forum December 1, 2007 De La Salle University.

[9] Sendra, J. P., Iglesias, D. M., Maria, F. D., "Support Vector Machines For Continuous Speech Recognition", 14th European Signal Processing Conference 2006, Florence, Italy, Sept 2006.

[10] Jain, R. And Saxena, S. K., "Advanced Feature Extraction & Its Implementation In Speech Recognition System", IJSTM, Vol. 2 Issue 3, July 2011.

[11] Aggarwal, R.K. and Dave, M., "Acoustic Modelling Problem for Automatic Speech Recognition System: Conventional Methods (Part I)", International Journal of Speech Technology (2011) 14:297–308.

[12] Aggarwal, R. K. and Dave, M., "Acoustic modelling problem for automatic speech recognition system: advances and refinements (Part II)", International Journal of Speech Technology (2011) 14:309–320.

[13] Ostendorf, M., Digalakis, V., & Kimball, O. A. (1996). From HMM‟s to segment models: a unified view of stochastic modeling for speech recognition. IEEE Transactions on Speech and Audio Processing, 4(5), 360–378.

[1].