Volume 3, Issue 4, April-2016, pp. 190-193 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

AUTOMATIC SPEECH RECOGNITION-A SURVEY

Julna Nazer, Sajeer K. (Asst prof)

MES College of Engineering

Kuttipuram

Abstract: Speech recognition is the next big step that the technology needs to take for general users. An Automatic Speech Recognition (ASR) will play a major role in focusing new technology to users. Applications of ASR are speech to text conversion, voice input in aircraft, data entry, voice user interfaces such as voice dialing. Speech recognition involves extracting features from the input signal and classifying them to classes using pattern matching model. This can be done using feature extraction method. This paper involves a general study of automatic speech recognition and various methods to generate an ASR system. General techniques that can be used to implement an ASR includes artificial neural networks, Hidden Markov model, acoustic –phonetic approach

Keywords: Automatic speech recognition, feature extraction, neural networks, hidden Markov model, acoustic phonetic approach.

———————————— ◆ ————————————

I. INTRODUCTION

Automatic speech recognition system can be generally defined as a computer driven translation of spoken words into a machine readable format, i.e. speech is converted 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 [1]. 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. Phonemes carry the textual content of an utterance; prosodic information gives valuable support to understand a spoken utterance. In short, prosody is the rhythm, stress and intonation of continuous speech, and is expressed in pitch, loudness and formants. Prosody is an important mean of conveying non-verbal information. Utterances can be lengthened or shortened; the relative length carries prosodic information. The bottom of the human vocal tract is the vocal cords, or glottis. For unvoiced speech, the glottis remains open, for voiced speech it opens and closes periodically. The frequency of the opening is called the fundamental frequency or pitch. It can be calculated from the spectrum and its contour over the utterance reveals useful information.

II. LITERATURE SURVEY

a. Neural network model

Neural network models are powerful speech recognition engines [2]. Their ability to classify data and ability in parallel processing pave the way for speech recognition. A typical neural network consists of input layer, hidden layer, output layer. Input layer receives the input signal and transfer 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, multi-layer perceptron model as well as radial bias functions can be used to classify the data.

b. Support Vector Machine model

Support Vector Machines (SVM) supervised learning models with associated learning algorithms. They are used for classification and regression analysis. They analyses the data and recognize patterns. Text independent speaker recognition uses as their features a compact representation of a speaker utterance, known as i-vector [5]. Rather than estimating an SVM model per speaker, according to the one versus all discriminative paradigms, the Pair wise 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 large amount of data, however, is a memory and computational expensive task, because the number of training pairs grows quadratic ally 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 hyper plane 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 hyper plane is similar to the optimal hyper plane. 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.

c. Generalized variable Hidden Markov model

Generalized variable Hidden Markov model (GVP-HMM) is used for speech recognition in a noisy environment [4]. A crucial task of automatic speech recognition systems is to robustly handle the mismatch against a target environment introduced by external factors such as environment 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.

d. F1 score method

Maximum Fl-score Criteria (MFC) is a discriminative training objective function for Goodness of Pronunciation (GOP) based automatic mispronunciation detection that makes use of Gaussian Mixture Model-Hidden Markov model as acoustic models [5]. The formulation of MFC seeks to directly optimize Fl score by converting the non-differentiable F1-score function into a continuous objective function to facilitate optimization.

Mispronunciation detection experiments show MFC based model-space training and feature-space training is effective in improving F1-score and other commonly used evaluation metrics. It is also shown MFC training in both the feature-space and model space outperforms either model space training or feature-space training alone. Then review and compare mispronunciation detection results with the use of MFC and some traditional training criteria that minimize word error rate in speech recognition. Using of GOP based mispronunciation detection method; use GMM-HMM based acoustic models to compute GOP scores. In this approach, GMM-HMM based acoustic models are often trained with maximum likelihood (ML) criterion. In ASR, discriminative training (DT) of the acoustic models has been widely used and has proved to give significant improvement over traditional ML estimation method.

e. Harmonic feature extraction

Voiced speech signals are built from harmonic components which are periodic and can be resynthesized easily [6]. Many algorithms to extract harmonic components from speech signals have been proposed where the motivation for the intensive research is its wide usability within the field of engineering, such as: speech analysis, speech coding, pitch tracking and estimation, speech enhancement using harmonic regeneration, post-speech enhancement, and bandwidth extension and voice activity detection. Recently, the periodicity of harmonic structure has been used to compensate noise and to identify the corrupted time-frequency regions. A new clean signal based on its likely characteristics estimated from the distorted signals has been resynthesized instead of filtering. The resulting signals show an increased naturalness and perceptual quality in speech. Another new application of the harmonic model has been an add-on module with several popular noise reduction methods. The add-on module utilizes the harmonic plus noise model (HNM) of speech to retrieve damaged speech structure. An improved sinusoidal modeling method based on perceptual matching pursuits computed in the bark scale has been proposed by for parametric audio coding applications.

III. OBSERVATION AND ANALYSIS

The performance analysis of various speech recognition systems is evaluated. Speech recognition has a big potential in becoming an important factor of interaction between human and computer. A successful speech recognition system has to determine features not only present in the input pattern at one point in time but also features of input pattern changing over time. Most common performance measure is word error rate. This measure is computed by comparing a reference transcription output by the speech recognizer. From this comparison it is possible to compute the number of errors, which is typically belong to three categories:

Insertions I (when in the output of ASR it is present a word not present in reference).

Deletions D (a word is missed in ASR output).

Substitutions S (a word is confused with another word).

IV. PROBLEM DEFINITION

Automatic Speech Recognition system takes a human speech utterance as an input and requires a string of words as output. It is based on feature extraction together with N-gram language model. The problem of automatically recognizing speech with the help of a computer is a difficult problem, and the reason for this is the complexity of the human language. Lack of linguistic corpora for dialect language models seems to be a difficulty in implementing an ASR. This stems from difficulty in collecting sentences with dialects. Scope of the problem should be broadened into larger vocabularies, continuous speech.

V. CONCLUSION

Automatic Speech Recognition is seen as an important part of human-computer interfaces that are envisaged to use speech, among other means, to attain natural, pervasive, and omnipresent computing. The state of ASR lacks robustness, to channel and environment noise continues to be a major impediment. A method is developed for recognizing mixed dialect utterances with multiple dialect language models by using small parallel corpora of the common language and a dialect and a large common language linguistic corpus.

REFERENCES

[1] Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G. Okuno,”Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models” IEEE Transactions on Audio, speech and language processing , VOL. 23, NO. 2, FEBRUARY 2015.

[2] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, Automatic estimation of dialect mixing ratio for dialect speech recognition, IEEE in Proc. Interspeech 13, 2013.

[3] N. Hirayama, S. Mori, and H. G. Okuno, Statistical method of building dialect language models for ASR systems,” IEEE in Proc. COLING , 2012.

[4] Rongfeng Su,Xunying Liu,Lan Wang ”Automatic Complexity Control of Generalized Variable Parameter HMMs for Noise Robust Speech Recognition,” IEEE Transactions on Audio, speech and language processing, 2015.

[5]Cumani S, Laface P.”Large-Scale Training of Pair wise Support Vector Machine for Speaker Recognition” IEEE transactions on audio, speech and language processing 2014.

[6]Sajeer Karattil,”A novel approach of implementation of speech recognition using neural networks for information retrieval”, International journal in Science and Technology vol 8 issue 33 Dec 2015.