International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.3, No.6, June 2014 DOI:10.15693/ijaist/2014.v3i6.43-47

The Comprehensive Review of Arabic Speech Recognition Systems

Bidoor Noori[1], Priyanka Abhang[1], Bharti Gawali[1]
Department of CS and IT, Dr. B.A.M.University, Aurangabad

Abstract
Speech technology and systems in human computer interaction have witnessed stable and remarkable advancement over the last two decades. Recent research concentrates on developing systems that would be most robust against variability in environment, speaker and language, focusing mainly on automatic speech recognition systems. Arabic is the world's fifth largest most spoken language in terms of number of speakers, still has not receive much attention from the traditional speech processing research community. Thus this paper attempts to focus on the development of Arabic Speech Recognition system using various techniques.

Keywords: Arabic Speech Recognition, MFCC, LDA, PCA, techniques

1- Introduction

Natural and efficient form of exchanging information among human is speech. Speech processing is one of the exciting areas of signal processing. The goal of speech recognition area is to develop technique and system for speech input to machine. Based on major advances in statistical modeling of speech, automatic speech recognition today find widespread application in task that require human machine interface such as automatic call processing[1]. Since 1960's computer scientists have been researching ways and means to make computer able to record interpret and understand human speech. Throughout the decades this has been a daunting task. Even the most rudimentary problem such as digitalizing (sampling) voice was a huge challenge in the early year. Until, the 1980's before the first systems arrived which could actually decipher speech. These early system was very limited in scope and power.

Arabic is a semantic language, and it is one of he oldest languages in the world. Speech recognition technology has advanced to the point where it is being used by more and more individuals in a wide variety of industries and professional carriers every day. Arabic automatic speech recognition tasks mainly addressed Arabic digits, broadcast news, command and control, the Holy Qur'an, and Arabic proverbs . The major work been done in Arabic speech recognition is to recognize the distinct Arabic phonemes, the phonetic features are also discussed and the general recursive neural network is used for the accurate implementation of Arabic phonemes [2]. The most popular statistical method used in speech recognition is the hidden Markov Model which is also used to estimate the probabilities for each phoneme [3]. The study of Arabic Text to speech synthesis system uses an automatic tool based on Diaphone concentration with MBROLA synthesis system uses an automatic tool based on analyzing and estimating the voice source into different types[4].

2- Characteristics of Arabic language:

Arabic language does not have a normalized form that is used in all circumstances of speech and writing. The characteristics of Arabic language are described below :

2.1Phonetic Features

The standard Arabic language has 34 phonemes, of which six are vowels and 28 are consonants. A phoneme is the smallest element of speech units that indicates a different meaning, word or sentence. Arabic phoneme contains two distinctive classes, which are pharyngeal and emphatic phonemes where are found in semantic language [5]. The Arabic language has fewer vowels than the English language. It has three long and three short vowels that are investigated and the differences and similarities between the vowels explored using Consonant –Vowels-Consonant (CVC) utterances. Standard Arabic is distinct from Indo-European language because of its consonantal nature. The allowed syllable structures in Arabic are CV, CVC and CVCC where V indicates a (long or short) vowel while C indicated a consonant. Arabic utterances can only start with a consonant [6, 7]. Arabic sound can be divided into macro classes such as stop consonants, voiceless fricatives, voiced fricatives, consonants, liquid consonants and vowels. The originality of Arabic phonetics is mainly based on the relevance of lengthening in vocalic system and on the presence of emphatic and geminated consonants. These particular features play a fundament role in the nominal and verbal morphological development [8,9].

2.2 Arabic Phoneme Set

Many languages have numerous dialects that differs in pronunciation. The Arabic language is more properly describe continuation of verities. Table 1 describes phoneme set and Table 2 describes pronunciation of Arabic digit [10]. Automatic recognition of foreign accented Arabic speech is a challenging task since it involves a large number of non-native accents. The non-native speech data available for training are generally insufficient. Moreover, as compared to other language, the Arabic language has relatively small number of research efforts [11,12,13].

Table 1: The complete phoneme set for daily life

Table 2: The Pronunciation of Digit in Arabic

Digit	***Arabic writing***	***Pronunciation***
1	واحد	Wahed
2	اثنين	Aathnay
3	ثلاثة	Thalathah
4	لربعة	Aarbaah
5	خمسة	Kaamsah
6	ستة	Settah
7	سبعة	Subaah
8	ثمانية	Thamaneyeh
9	تسعة	Tesah
0	صفر	Sefer

3. Arabic Speech Recognition System

Speech recognition at its most elementary level comprises a collection of algorithms drawn from a wide variety of disciplines, including statistical pattern recognition, communication theory, signal processing and linguistics among others .

3.1Speech classification

The Speech recognition system can be separated by different classes on the basis of utterances that can recognize the following :

a)Isolated word Recognition system

This type of system accepts single utterances at a time. Example: Kasra, Damma etc.

b)Connected word Recognition system

It is same as isolated word recognition system but allow separate utterances to be run together minimum pause between them. Example: (9955442233) nine nine five five four four two two three three.

c)Continuous speech recognition system

It allow user to speak almost naturally. Example: I am writing a research paper.

3.2 Production of Speech

The production of speech sound is through the air flow from the lungs to the glottis to open and then to the throat and mouth. Depending on these sound speeches, the signal can excited in three possible ways.

1.Voiced Excitation: Here the glottis is closed. The sir pressure forces the glottis to open and close periodically thus generating a periodic pulse train. The fundamental frequency usually lies in the range from 80Hz to 350Hz.

2.Unvoiced Excitation: Here the glottis is opened and air passes a narrow passage in the throat or mouth. This results in a turbulence which generates a noise signal. The spectral shape of the noise is determined by the location of the narrowness.

Write form	***pronunciation***	***Meaning***
تلفاز	Telefaz	TV
كتب	Katab	he word
ذهب	Thahab	Gold
يتكلم	Yitkallim	he speaks
صيف	Saif	Summer
باب	Bab	Door
بنت	Bent	Girl
ولد	Walad	Boy

3.Transient Excitation: A closure in the throat or mouth will raise the air pressure. By suddenly opening the closure the air pressure drops down immediately.

3.3 Speech Recognition Techniques

The goal of speech recognition techniques is to analyze, extract, characterize and recognize information about speech identity. From the literature since 1939 various techniques are robust and dynamic for speech recognition analysis. The speech recognition technique viewed working in three stages:-

i.Feature Extraction Technique

ii.Clustering Technique

iii.Classification Technique

i. Feature Extraction Technique :

Transforming the input data into the set of features is called feature extraction. The features extracted are carefully chosen . It is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced

representation instead of the full size input. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. Following are the most frequently used feature extraction techniques**:**

a)Mel Frequency Cepstral Coefficient(MFCC)

MFCC is based on human hearing perceptions which cannot perceive frequencies over 1KHz. In other words, MFCC is based on known variation of the human ear's critical bandwidth with frequency [14,15].MFCC has two filters which are spaced linearly at low frequency below 1000Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech MFCC's are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone.

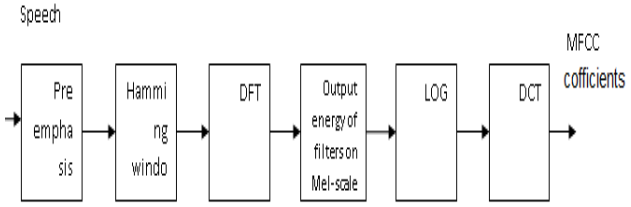

Figure1. Step involved in MFCC feature extraction

The basic property of MFCC is that the power spectrum is computed by performing Fourier Analysis. MFCC also lies in its ability for compact representation of amplitude spectrum. The Limitations of MFCC includes Quick MFCC algorithm reduces the run time while main training recognition accuracy of the system?

b)Linear Predictive Coding:

The basic idea behind the Linear Predictive Coding(LPC) analysis is that a speech sample can be approximated as linear combination of past speech sample. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. LPC of speech has become the predominant technique for estimating the basic parameters of speech. The Linear prediction method provides a robust, reliable and accurate method for estimating the parameters that characterize the linear time-varying system representing vocal tract. These coefficients form the basis for LPC of speech. 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 [16,17,18]. The main property of LPC , Acceleration and Delta coefficients. The strength includes providing alternative interpretation of N400 and LPC old/new effects in terms of memory strength and decisional factor. Analyze the LPC residual error of natural speech and try to reveal the limitation of LPC model.

ii) Clustering Technique:

Clustering is the process of grouping similar objects together. The resulting groups are called clusters. Clustering algorithms group points according to various criteria. Unlike most classification methods, clustering handles data that has no labels or ignores the labels while clustering.

a)Principal Component Analysis(PCA)

Principal Component Analysis is an old technique of multivariate statistical analysis, consisting of computing the eigenvectors of D*D covariance matrix, then sorting them according to the corresponding eigen values, in descending order and finally building the projection matrix A (called Karhuen-Loeve Transform, KLT) with the largest K eigenvectors(i.e. the K directions of greatest variance). Each feature vector X is then pre-processed according to the expression Y=A(X-μ), where μ represents the mean feature vector. KLT decorrelates the features and provides the smallest possible reconstruction error among all linear transforms, i.e. the possible mean-square error between data vectors in the projection k-feature space[19]. The basic property includes non linear feature extraction method supervised linear map.PCA finds a set of mothogonal vectors that a count for the greatest variance in the data. Dimension reduction can only be the original variable were correlated, PCA is not scale invariant.

b)Linear Discriminate Analysis(LDA)

Linear Discriminate Analysis(LDA) and the related Fisher's Linear Discriminate are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier or more commonly, for dimensionality reduction before later classification. LDA seeks to reduce dimensionally while preserving as much of the class discriminatory information as possible[20].Non linear feature extraction method supervised linear map. It is important to motivate the use of a correlation discriminate approach for estimating feature space transformations. The limitation is in dealing with the small sample size problem LDA impose a significant performance.

iii) Classification Techniques

The result of recognition system is depending on decision from classification system. The classification method uses a set of parameters to characterize each object.

a)DTW (Feature Matching)

DTW algorithm is based on Dynamic Programming techniques as it describes distance measurement between time series which is needed to determine similarity between

time series and for time series classification [21]. The algorithm is for measuring similarity between two time series which may vary in time or speed. This technique is also used to find the optimal alignment between two time series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two series can then be used to find corresponding regions between the two series or to determine the similarity between the two time series. Various energy measurement techniques are simple, fast, portable and inexpensive. It is a time domain based method and easy to be embedded into real-time devices, is also able to capture the modulation characteristics with high accuracy. Height-diameter growth is not a direct response to mechanical failure DTW is the limitation in general.

b)Hidden Marko Models(HMM)

ASR systems area based on the Hidden Markov Model(HMM) started to gain popularity in the mid 1980's[22].HMM is well-known and widely used statistical method for characterizing the spectral features of speech frame. The underlying assumption of the HMM is that the speech signal can be well characterized as a parametric random process and the parameters of the stochastic process can be predicated in a precise, well defined manner. The HMM method provides a natural and highly reliable way of recognizing speech for a wide range of applications [23]. Spectral properties of the speech waveform within the window. The reasons for this method to be popular are the inherent statistical (mathematically precise) framework, the ease and availability of training algorithms for estimating [24].The limitation are in modules directly without any intervening structures such as phoneme lattice the general problem is of completely fluent. For Arabic language phonemes are suitable sub-word because a small set of them can cover the entire Arabic language and it is easy to collect many examples of each phonemes in a database of a small size as this is the most famous type of stochastic process modeling **.**

a)Neural Networks

Artificial Neural Networks (ANN's) have been investigated for many years for the desire of achieving human-like performance in the field of ASR. These models are composed of many nonlinear computational elements operating parallel in patterns similar to the biological neural networks [25]. ANN has been used extensively in ASR field during the past two decades. The most beneficial characteristics of ANN's for solving ASR problem are the fault tolerance and nonlinear property. ANN models are distinguished by the network topology, node characteristics, and training or learning rules. One of the important models of the neural networks is the multilayer perceptions (MLPs), which are feed-forward networks with zero, one, or more hidden layers of nodes between the input and output nodes [26]. The capabilities of the MLP stem from the nonlinearities used with its nodes. Any MLP network must consists of one input layer (not computational, but source nodes), one output layer (computational nodes), and zero or more hidden layers (computational nodes) depending on the network sophistication and the application requirements. Many Arabic ASRs were designed using ANN techniques [27]. In the first research a spoken Arabic digits recognizer was designed to investigate the process of automatic recognition process [28]. The system was operated in two different modes, multi-speaker mode and speaker independent mode. The overall system performance was 99.47% in the first mode and 96.46% in the second model [29].It is used for identification of non-linear systems are proposed. Most important advantages are that the resulting neural model can be easily lineared around different operating points, allowing application of classifiably stability theorems from the linear systems domain to this class. The algorithm form can be used for any regression problem in which an assumption of linearity is not justified [30].

4. Arabic Speech Database

There are very few databases available for Arabic speech recognition .spoken Arabic data set is created by machine learning repository this dataset contains time series of MFCC corresponding to spoken Arabic digits . It includes data from 44 female native Arabic speakers .

Data collected by the laboratory of automatic and signals ,University of Badji- Mokhtar. Annaka,Alegria.

Each line of this database represents 13 MFCC, coefficients .the sampling rate was :11025Hz ,16bit with hamming window .

The another database contains 4740 utterances from six speakers (three males and three females)there are 620 statements for training and 171 statement for each speaker . there are 3622 words ,with 27725 triphones ,where 5034 of them are unique.

Conclusion

This paper attempts to present review of research efforts for Arabic speech recognition . It also presents the characteristic of Arabic language and technique commonly used for speech recognition tasks the database. As very few databases are available in Arabic language for speech recognition much of research efforts are expected in this area.

Reference:

[1] R.Klevansand, R. Rodman, "Voice recognition ", Artech House, Boston London.

[2] Khalooq y. Al. Azzawi Khaled Daqrouq, "Feed Forward Back Propagation Neural Network Method for Arabic Vowel Recogntiion based on wavelet linear prediction coding ", IJAET ISSN: 2231-1963, Sept 2011.

[3] M. A. Mokhtar, A.Z. El- Abddin, " A model for the acoustic phonetic structure of Arabic language using a single ergodic hidden markov model"

Electrical engineering department faculty of Engineering, Alexandria University, Alexandria, Egypt.
[4] Abdelkader Chabchoub, Adnan Cherif, "An Automatic mbrola tool for high quality Arabic speech synthesis" IJCA volume 36-No-1, Dec 2011.
[5] Yousef Ajami Alotaibi, Mansour Alghamdi, Fahad Alotaiby, "Speech Recognition System of Arabic Digits based on A Telephony Arabic Corpus".
[6] King Abdulaziz City for Science and Technology PO Box 6086, Riyadh 11442, Saudi Arabia.
[7] M.Alghamdi , Y.O. Mohamed El Hadj, Alkanhal, "A MANUAL SYSTEM TO SEGMENT AND TRANSCRIBE ARABIC SPEECH"
[8] Yousef Ajami Alotaibi,Comparative Study of ANN and HMM to Arabic Digits Recognition Systems, JKAU: Eng. Sci., Vol.19 No.1, pp:43-60(2008 A.D. /1429 A. H.
[9]http://www.crulp.org/Publication/theses/2009/Automatic_Speech_Recognition_System_for _Urdu.
[10]Loura Mayfield Tomokiya, Alan W Black and Kevin A. Lenzo," Arabic in my Hand: Small-Footprint Synthesis of Egyptian Arabic" Cepstral LLC 1801 E. Carson St. Pittsburgh, PA 15203 USA.
[11]http://www.slideshare.net.hend.alkhalifa/mom2010-arabic-natural-language-processing.
[12] Speech Recognition System of Arabic Alphabet Based on a Telephony Arabic Corpus http://www.springerlink.com/content/j6207wx494197114/
[13] Yousef Ajami Alotaibi,Mansour Al-Ghmdi, Fahad Alotaiby , " Speech Recognition System of Arabic Digits based on A Telephony Arabic Corpus.IPCV 2008:35-38 .
[14] E.C. Gordon, Signal and Linear System Analysis, John Wiley & Sons LTD., New York, USA, 1998.
[15] Stan Salvador and Philip Chan, "Fast DTW: Toward Accurate Dynamic Time Warping in Linear time and space ", Department of Computer Sciences Florida Institute of Technology, Melbourne.
[16] Corneliu Octavian Dumitru, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continues Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.
[17]DOUGLASO'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September 2003, 0018-9219/03$17.00 © 2003 IEEE .
[18]N.UmaMaheswari,A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 1793-8201.
[19] Jolliffe, I.T. 2002. Principal Component Analysis, Springer.
[20]http://en.wikipedia.org/wiki/linear_discriminant_analyisis
[21] Rabiner, L.R. , " A Tutorial on Hidden Markov Models and Selected Applications in speech recognition ", Proceedings of the IEEE, Vol.77, No. 2, pp:257-286, Feb 1989.
[22] Juang B. , Rabiner L. , "Hidden Markov Models for Speech Recognition", Technometrics, 33(3), Aug, pp:251-272, 1991.
[23] Duda R.O., Hart P.E., Stork D. G. 2000, "Pattern Classification", Wiley Interscience.
[24]luthuli.cs.uiuc.edu/~daf/courses/...HMMs/0.pdf
[25] T.Ganchev, M. Siafarikas, N. Fakotakis, "Evaluation of speech parameterization methods for speaker recognition", Proc. Of the Acoustics, Vol.18-19,pp:105-110,2006.
[26]Ben Milner, "a comparison of front-end configurations for robust speech recognition", ICASSP page 797-800.IEEE,(2002).
[27] Pronunciation Modeling for Dialectal Arabic Speech Recognition
http://www.cs.cmu.edu/~awb/papers/asru2009/AS090074
[28]http://wwwclips.imag.fr/geod/User/laurent.besacier/Publis/icslp06-2.pdf
[29] IJCSES International Journal of Computer Sciences and Engineering System Vol.1,No.3, July 2007.
[30]http://www.hindawi.com/journals/asmp/2008/679831.abs.html