IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 2, Ver. II (Mar-Apr. 2016), PP 46-51
www.iosrjournals.org

Automatic Speech Recognition For Task Oriented IVRS In Marathi

Manasi Ram Baheti[1] , Bharti W. Gawali[2], S.C. Mehrotra[3]
[1]*Department of CSIT , Dr.B.A.M.University, Aurangabad, (M.S.), India*
[2]*Department of CSIT , Dr.B.A.M.University, Aurangabad, (M.S.), India*
[3]*Department of CSIT , Dr.B.A.M.University, Aurangabad, (M.S.), India*

Abstract: *Speech recognition has become a foundation of self-service interactive voice response (IVR) user interfaces. The speech recognition and IVR applications found to be cost-effective and the user friendly, speediest self-service alternative to speaking with a contact centre agent. The concept of making advanced systems reach up to rural level and solve the problem of communication gap is the main problem addressed in this paper. With improvements in "Speech Recognition" and IVRS providing 24/7 services, objective is to develop an IVRS in Marathi Language that can be useful for agricultural purpose, for rural area people. This can be done by considering specific, pre-decided Marathi Language sentences for IVR based agricultural commodity price information retrieval system developed mainly for the illiterate semiliterate or farmers.*
Keywords: *ASR, DTW,HCI, IVRS, MFCC.*

I. Introduction

Throughout the human history, speech has been the most dominant and convenient means of communication between people. With the rapid development of communication technologies, a capable speech communication technique for human-to-machine interaction has become essential. Today, even though much work in done regarding HCI with respect to Speech Recognition; it is limited up to urban area only. The development of IVRS is for handling online banking queries, customer care services, LPG Cylinder booking, etc. But actually, IVRS can also be used for solving the problems and queries of the farmers, which will develop the agricultural system on a high ratio[1.] The concept of making these advanced systems reach up to rural level and solve the problem of communication gap is the main problem handled over here. Speech recognition has often been suggested as a key to universal information access, as the speech modality is a "natural" way to interact, does not require literacy, and relies on existing telephony infrastructure.

II. Motivation and objective

A successful speech interface is one that supports an application based on local content created by local providers, as the information needs of rural communities include news, events, and innovations happening. A speech driven application for developing communities must address all of these issues in order to effectively extend IT access to the developing world.[2]

- Rural customers lack real-time access to vital information such as commodity pricing, weather reports, local news, entertainment, agro machinery, agro products etc.
- Hence, we need to create "voice" based services on Interactive Voice Response (IVR) platform to reach out to such users.
- Although many interactive software applications are available, the uses of these applications are limited due to language barriers. Hence development of speech recognition systems in local languages will help anyone to make use of this technological advancement.
- In India, where 70% of the country's population is involved in the agriculture industry, speech technology has started been playing a critical role through user friendly speech solutions to rural farmers.
- The interactive voice response (IVR) system will serve as a bridge between people & computer. The telephone user will be able to access the information from anywhere at any time simply by dialing a specified number.[3]

We aim to make the farmer an equal trading partner with his buyers and his suppliers during the entire agricultural cycle thereby connecting the dots in the information flow of the agriculture cycle. [5] It will also help women farmers, having lower educational and literacy levels in Rural India , which discourages female ownership of productive assets due to technical literacy issues. [6]

i) To develop a Speaker independent system that will recognize the continuous sentence and also respond accordingly in Marathi language.
ii) To create an application for the concerned work using IVRS.

2.1 Significance/ Likely End Users

- Farmers and other users interested in obtaining the latest information/prices of agricultural commodities by just speaking over the telephone.
- This system will be especially useful for those who have no access to computers and Internet, and who may not have the required computer skills and reading/writing skills.

Speech recognition technologies, in addition to bringing down operational costs, also support the livelihoods of farmers through a simple and convenient solution.

III. Experimental work

This proposed work aims to create a recognition system in Marathi language to recognize continuous sentence. For this, a database of ten sentences in Marathi language is created for storing the recordings and for matching purpose. The system will recognize sentence spoken by the user and match it with the sentence stored in the database and will answer accordingly. The user can ask a question and the system will answer it by matching with the sentence stored.

3.1 Database Creation

Results of ASR systems are highly dependent on database, i.e., the results obtained in ASR are insignificant if recording circumstances are not known. The factors affecting the performance are recording conditions, gender, literacy of the person , speaker characteristics, etc. Success rates obtained in an ASR system are insignificant if the recording conditions are not known. In this work, a methodology and the typical experimental setup used for building up speech database for speaker-independent ASR system is presented.[7]

Automatic recognition of spoken sentences is a challenging task in the field of ASR. For accuracy of the speech recognition, we need the collection of utterances, which are required for training and testing. The collection of utterances in a structured mode is called the database.

3.1.1 Selection of Sentences/text for database

As the objective is to develop the system for rural agriculture purpose, it was decided to develop the system for agricultural commodity price retrieval system. In pursuit of this, field visit was made to the "Krushi Utpanna Bajar Samiti" at Bhokardan Taluka, District Jalna. After talking to actual farmers and traders, and observing the daily transactions of the trade, the frequently asked questions (FAQ's) of this 'mandi' were taken into consideration . These are the queries those are most frequently asked by the farmers in order to sale their goods. These are the questions related to the highest and lowest price of the commodity, the purchase and sale of the commodity, commodity storage by rent and its price , shares index of this 'mandi' etc.[8] It was observed that, although these questions are simple but are very important in this 'mandi'. So the ten questions and their appropriate answers were shortlisted.

3.1.2 Database Creation

Before recording is started in order to create the database, some important characteristics of the sentences has to be considered. The sentences should be meaningful and natural. They should be easy to read; so sentences should be simple and short. The number of sentences in a set should not be large, For example not more than 10, as the number of samples are considered here which are the utterances of various speakers. The sentences should contain phonemes in various phonetic contexts. Each person speaks one such compact set of sentences. This section describes the formation of various samples of such sets of Marathi sentences [9].

3.1.3 Acquisition setup

We describe here the measures taken for collecting speech data, to develop a robust speech recognition system as a part of voice interface for agricultural information retrieval.To achieve a high audio quality, the recording was done in the ordinary room without noisy sound and effect of echo. The sampling frequency for all recordings was set to be 16000 Hz at the room temperature. The speakers were asked to sit in front of the microphone with the distance of about 12-15 cm. The recording and preprocessing was done with the help of computerized Speech Laboratory (CSL). The CSL is one of the analysis systems for speech and voice. CSL is an input/output recording device for a PC, which has special features for reliable acoustic measurements. The figure 1 shows the setup of database creation using Computerized Speech Lab.

Fig 1: The setup of Computerized Speech Lab for database recording

The figure 2 shows the waveform for the sentence : " आज कोणत्या मालाची सर्वात जास्त विक्री झाली ? "

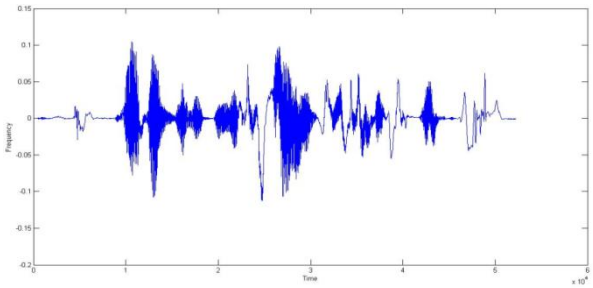

Fig 2: Waveform for the given sentence.

IV. Feature Extraction

Using Mel Frequency Cepstrum Coefficient (MFCC)

A feature is a parameter that can be estimated through processing signal. There are three basic steps in the ASR, (1) parameter estimation (in which the test pattern is created) (2) parameter comparison and (3) decision making.[10]

Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used for a variety of problems in signal processing. Different temporal and spectral analysis is done on the sound signals to extract the use- full features, the most important of them being the Mel Frequency Cepstrum Coefficients (MFCC). MFCC is considered to be the closest possible approximation to the human ear. MFCC is generated from the sound signal by passing through high band pass filters which results in higher frequencies becoming more distinct than the lower frequencies. The MFCC extraction process works with the frequency decomposition of a sound file. Extraction of MFCC: Transferring the input waveform into a sequence of acoustic feature vectors. Each vector represents the information in small time window of the signal. The resulting features are the MFCCs, which are treated as a single vector and are typically computed for every frame of 20 ms. These feature vectors form the input to the training and recognition systems[11].

The Table 1 shows the thirteen features extracted from ten frames. Also, the values for Mean, Variance and Standard Deviation are given. This figure 3 shows the wave, plot and format regions of the given speech signal. This process of parameter estimation is usually called training. This is done for all questions (Q), subjects(S), Utterances (U).

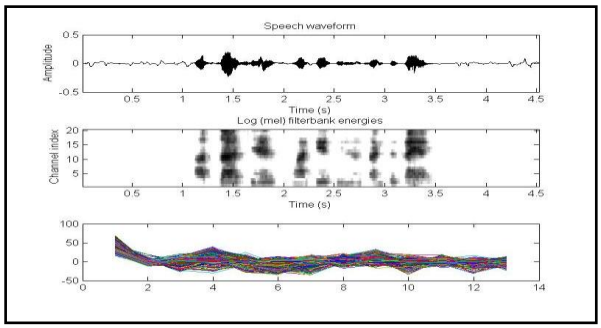

Fig 3: MFCC features plot with waveform and spectrogram for energy

Table 4.6: 13 features extracted from 10 form for S1Q1U1

Features	Frm1	Frm2	Frm3	Frm4	Frm5	Frm6	Frm7	Frm8	Frm9	Frm10
1	15.86	18.85	24.65	19.41	19.48	20.79	19.07	18.81	17.70	15.86
2	-10.79	-6.96	-5.30	-12.68	-12.53	-12.16	-11.36	-13.38	-13.63	-10.79
3	0.03	2.01	1.57	1.28	1.19	0.56	4.93	-0.11	-5.53	0.03
4	-2.08	-0.70	-0.65	1.24	-4.75	-8.43	-7.22	-2.05	-5.41	-2.08
5	-2.65	0.72	1.67	0.10	-2.14	-6.85	-8.06	-2.60	2.45	-2.65
6	0.10	0.62	0.71	-5.68	-6.10	-4.56	-0.87	1.82	1.18	0.10
7	-1.78	-0.53	-1.12	4.73	-6.57	-12.33	-7.77	-4.21	-3.41	-1.78
8	0.61	-0.04	-0.21	0.70	2.96	-1.72	-4.95	-7.90	-11.55	0.61
9	-0.80	-2.82	0.32	0.67	-9.86	-8.70	2.47	9.14	7.25	-0.80
10	-2.85	2.32	-0.88	-5.52	0.77	1.56	-0.69	9.57	10.34	-2.85
11	-0.78	2.25	1.57	3.47	7.31	2.61	2.55	-1.92	0.15	-0.78
12	-2.42	-0.51	0.27	2.98	3.93	6.11	2.77	1.70	-0.75	-2.42
13	-3.34	1.78	-2.51	-0.61	-4.49	-0.75	5.79	2.47	-6.03	-3.34

Mean>>4.7718
Variance>>2.7491
SD>>1.658

V. Recognition : Matching using Dynamic Time Warping (DTW)

Once feature vectors generated using MFCC, the next step is to find the optimal match. For this, the technique is DTW techniques has been used.The simplest way to recognize sentence sample is to compare it against a number of stored templates and determine the best match . DTW is an instance of the general class of algorithms and known as dynamic programming. The algorithm makes a single pass through a matrix of frame scores while computing locally optimized segment of the global alignment path. The dynamic time warping algorithm provides a procedure to align in the test and reference patterns to give the average distance associated with the optimal warping path.[12] Consider two sequences A and B, composed respectively, of n and m feature vectors. When Q1 With Q2 DTW Following Graph showing the optimal path alignment using the Euclidian distance is shown in the figures 1.1 (a) and 1.1 (b) respectively.

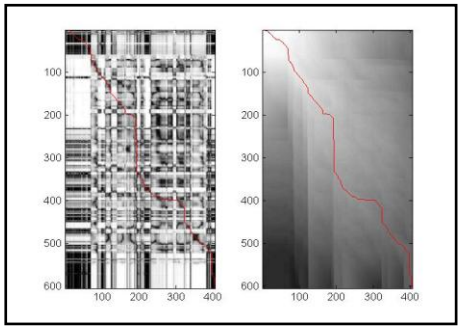

Figure 1.1 (a) DTW optimal path alignment

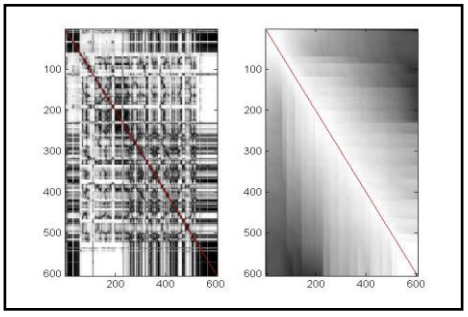

Figure 1.1 (b) DTW optimal path alignment.

The diagonal values in the above table indicate the match while non diagonal values show the nonalignment. This can be more understood by the following figure showing the red line of the proper alignment of the match. While the figure 1.1 (a) and (b) shows the red line with improper alignment indicating the mismatch[13].

Working of DTW: The recognition process, consists of matching the incoming speech with stored templates. The template with the lowest distance measure from the input pattern is the recognized word. The best match (lowest distance measure) is based upon dynamic programming. This is called a Dynamic Time Warping (DTW) recognizer.

In order to understand DTW, two concepts need to understand,

- Features: the information in each signal has to be represented in some manner.
- Distances: some form of metric is used in order to obtain a match path. There are two types:
- Local: a computational difference between a feature of one signal and a feature of the other.
- Global: the overall computational difference between an entire signal and another signal of possibly different length[14] .

5.1 Performance of the system :

The performance of the Marathi IVRS system is calculated on the basis of accuracy. The real time factor is the time required for recognition in response to the operation. The accuracy is calculated on the basis of

confusion matrix in which Number of token was passed randomly. A sentence at some exceptional case makes system confused about speech From Confusion Matrix we derived overall accuracy from following equation:

$$Accuracy = \frac{N - C}{N} * 100 \quad (1)$$

Where N is a Number of Token Passed and C is a Number of Token Confuse. The complexity of the system is calculated on the basis of time required for training and testing.

In order to test for speaker independent of the system, some of the subjects who participated in the creation of the testing corpus had not participated in the creation of the database. To further test the system on live data and also again test its speaker independence, the system was tested by running it live.

5.1.1 Training and testing of Marathi IVRS system

- The Marathi IVRS system was trained using MFCC 13 feature of the individual sentence of 10 speakers.
- The system was tested using the Euclidian distance based approach. If the sentence is recognized, then the respective answer was played.
- Total ten sentences were tested in 40 trials using MFCC feature extraction. The system is tested for two modes and the corresponding performance of this is illustrated in the Ttable1[15].

5.1.2 Final Performance of the IVRS System

On running and testing the system against the testing data, the following performance statistics were obtained as shown in the Table 2

Table 2:Performance Statistics

1	Speaker Dependant	Offline	93.43%
		Online	90.93%
2	Speaker Independent	Offline	88.12%
		Online	85.62%

From the above table it is observed that, the recognition rate for speaker dependent system is more because the pattern is tested against stored pattern in the database. On the other hand, in the speaker independent system the recognition rate decreases due to noise in the live recording. If this noise in live recording is reduced by using good sound quality of microphone, certainly the recognition rate will be increased. The confusion occurs due to same length of the utterance, accent as well as acoustic parameters. The variations in the performance rate depends upon the distance of speaker from or to the microphone, speed of the utterance, level of literacy etc. This observation leads to make use of high quality of microphone as to give input Also, if the speakers are trained in recording or they have clearly understood the purpose of the system , it will definitely increase the performance rate of the system.[16]

VI. Conclusion

Speech technology can increase user satisfaction and retention. Allowing users to use the most natural human interface, speech, to communicate, instead of forcing them to navigate entering multiple digits into a key pad. The goal of speech-enabled applications has always been to allow callers to obtain information and perform transactions simply by speaking naturally. This system described in the paper will allow farmers to access the commodity prices by using spoken dialogue system.

The work to develop an IVR system in Marathi language that will serve the need of rural agricultural queries regarding selling and buying of Agri goods, will be useful to the rural area where connectivity, literacy and access to information is less than the urban area.

Acknowledgement

We are thankful to Prof.R.R.Deshmukh, Head of the department, for his support and encouragement.

References

[1]. Rabiner, L. R., and Schafer R.W. "Introduction to digital speech processing." Foundations and trends in signal processing 1, no. 1 (2007): 1-194.
[2]. Meng, Y. (2004). Speech recognition on DSP: Algorithm optimization and performance analysis (Doctoral dissertation, The Chinese University of Hong Kong).
[3]. Carroll, J. M. (1997). Human-computer interaction: psychology as a science of design. Annual review of psychology, 48(1), 61-83.
[4]. Rabiner, L. R., & Schafer, R. W. (2007). Introduction to digital speech processing. Foundations and trends in signal processing, 1(1), 1-194.
[5]. https://www.hcii.cmu.edu/courses/speech-recognition-and-understanding
[6]. Kemble, K. A. (2001). An introduction to speech recognition. Voice Systems Middleware Education-IBM Corporation.

[7]. Laxminarayana P. , Ramana A.V., Mythilisharana P. , Srikanth A., Sandeep Kumar B. , Mounika J., "Automatic Speech Recognition, Tutorial & Lab Manual", Research and Training Unit for Navigational Electronics Osmania University, Hyderabad.
[8]. Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.
[9]. http://ai.ato.ms/MITECS/Entry/rabiner.html
[10]. Kumar, R., Kishore, S., Gopalakrishna, A., Chitturi, R., Joshi, S., Singh, S., & Sitaram, R. (2005). Development of Indian language speech databases for large vocabulary speech recognition systems. In International Conference on Speech and Computer (SPECOM) Proceedings.
[11]. Patil, H. A., & Basu, T. K. (2008). Development of speech corpora for speaker recognition research and evaluation in Indian languages. International Journal of Speech Technology, 11(1), 17-32.
[12]. Godambe, T., & Samudravijaya, K. (2011, June). Speech data acquisition for voice based agricultural information retrieval. In Proc. Of 39th All India DLA Conference, Punjabi University, Patiala, June.
[13]. Samudravijaya, K., & Gogate, M. R. Marathi speech database. In Proc. of Int. Symp. on Speech Technology and Processing Systems and Oriental COCOSDA-2006, Penang, Malaysia (pp. 21-24).
[14]. Plauché, M., & Nallasamy, U. (2007). Speech interfaces for equitable access to information technology. Information Technologies & International Development, 4(1), pp-69.
[15]. Khan, S., Basu, J., Bepari, M. S., & Roy, R. Field Trial, Evaluation and Error Correction methods of an IVR based Commodity Price Retrieval System.
[16]. http://iitg.vlab.co.in/?sub=59&brch=164&sim=613&cnt=1