Speech Recognition Using Dynamic Grammar and Parallel Listening Events

Muhammad Sharif, Syed Zia Uddin Bukhari, Aman Ullah Khan, Waseem Khan, and Mudassar Raza

***Abstract*—This paper presents a method for speech recognition using hybrid of dynamic vocabulary through dynamic grammar and parallel recognition events in runtime environment. The key idea is to use different recognition operations in such a way that speech can be recognized accurately. The method makes the recognition of the speech independent of old static vocabulary through a dynamic grammar loading process. The proposed approach is divided into different tasks which are Dynamic grammar upload and Parallel listening events for recognition process. This algorithm can quickly and correctly detect the speech. Under the experimental database which was taken from real speech inputs, 247 out of 250 words were successfully recognized. The average accuracy of speech recognition is, therefore, 85.0%.**

***Index Terms*—Dynamic grammar, parallel listening events, speech recognition, parallel recognition events, runtime environment.**

I. Introduction

In the current modern era, the use of speech activated systems is becoming more and more common. The speech recognition (SR) technology has attracted great attention and many systems are being developed and applied all over the world. It has many applications in almost all fields which include human interaction through the use of speech for command and control systems and controlling the hardware as well as the software security issues.

The most crucial and difficult part of an SR system is the detection and extraction of words which directly affects the system's overall performance and accuracy. The presence of noise in the speech sound, uneven frequencies [1] and wrong pronunciation of specific language words e.g., English makes the task even more difficult. The proposed system is a detailed and novel method for accurately detecting the spoken word.

II. Related Work

The problem of speech recognition has been studied for many years. In the conventional approach, sound is listened by a single process which gives it to the SR engine for comparison on the basis of static grammar [2] loaded before the execution of SR system. The listener finds the probability of the recognized word and gives the result on the basis of more probability in single attempt.

Manuscript received September 25, 2013; revised November 16, 2013.

Muhammad Sharif, Syed Zia Uddin Bukhari, Waseem Khan and Mudassar Raza are with the Department of Computer Sciences, COMSATS Institute of Information Technology, Wah Cantt., Pakistan (e-mail: muhammadsharifmalik@yahoo.com, muhammad.wasim1@gmail.com, mudassarkazmi@yahoo.com).

Aman Ullah Khan is with the Department of Computer Science and Engineering, Air University Multan Campus, Pakistan (e-mail: auk_pk@yahoo.com).

III. Defects in Conventional Design

A. Misfit for Pakistani Culture Due to Urlish (Urdu and English)

In Pakistan we usually deal with Urlish language, for example:

ADD CUSTOMER "FAZAL DIN"

In the above example words like "ADD" and "CUSTOMER" are English words but word "FAZALDIN" is not an English word. Hence when we talk about conventional speech recognition module "FAZALDIN" is an unrecognizable word. So the recognition system fails here in Pakistan or in any NON-ENGLISH society.

B. It Is General, Not Specific

We often work in specific environments with limited words like in telephone dialing, for example:

DIAL ISLAMABAD PIA ACCOUNT OFFICER

This example shows that on commercial basis this technique is more useful. The reason is that it is using number of specific words i.e,

- "DIAL"(Any command to be executed)
- "ISLAMABAD" (reference to any country or city code)
- "PIA" (reference to any organization number)
- "ACCOUNT OFFICER" (any particular officer extension number)

SR Engine should focus more on these words than thousands of other words used in the conventional approach.

Similarly, the use of static grammar makes the user unable to modify it accordingly. Also the insertion of new words in the language is almost impossible.

C. Larger the Vocabulary, Greater the Probability of Errors

Although the conventional model has immense scope i.e., it has thousands of words in the vocabulary but along with this quality it has a very considerable defect of the loss of accuracy. Accuracy is of great importance because time saving is the basic reason of speech recognition usage and if we compromise on errors of the system it wastes our time due to bad accuracy level. Because of all the above defects the accuracy level of conventional model is about 55%.

DOI: 10.7763/IJFCC.2014.V3.275

IV. The Proposed System

The SR system proposed is mainly based on Loading Dynamic Grammar [3] along with Dynamic Vocabulary in runtime environment. Fig. 1 and Fig. 2 show the block diagrams of proposed system for Urlish language and for banking environment.

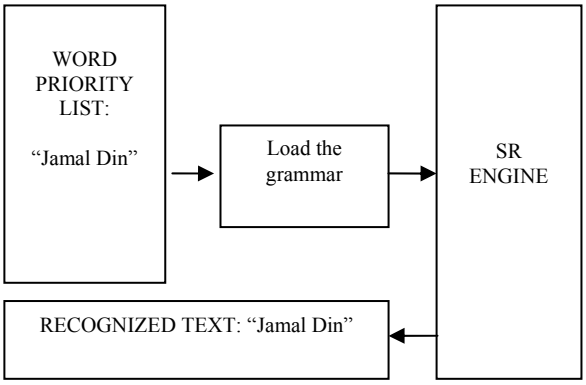

Fig. 1. Block diagram showing dynamic grammar along with dynamic vocabulary

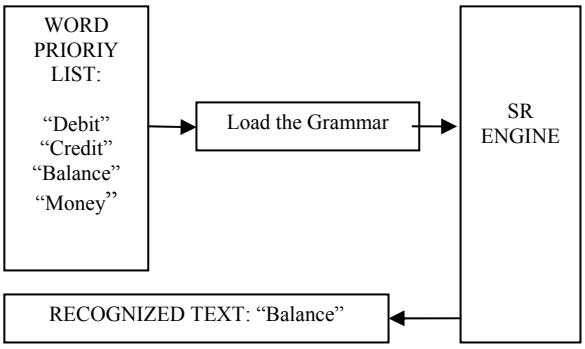

Fig. 2. Block diagram showing dynamic grammar update example for banking environment

We now discuss the above mentioned steps of proposed system in detail:

Conventionally dynamic grammar is used for runtime entry of data. For example:

"Send new e-mail to NAME"

In this example, the sentence can be divided into two separate parts:

- Static part, "Send new e-mail to"
- Dynamic part, "NAME"

Static part will be handled through "Static Grammar Interface" while dynamic part will be handled through "Dynamic Grammar Interface".

In our proposed system "Dynamic grammar interface" is used with a different perspective. The preprocessing step includes the updating of existing grammar through a dynamic process. This is a speaker focused technique. The main idea behind this technique is the fact that the speakers may have different environments.

For example, if the speaker is working in a banking environment he would speak words like "credit", "debit", "profit", "loss", "money", "percentage", etc. with more frequency. Now the SR system grammar should focus on these words. The conventional approach, which is of course a generalized approach, cannot cope with this environment. Hence the accuracy rate is very low for this specific environment. On the other hand, our proposed method provides the ease to dynamically load the grammar according to our environmental needs and then process the speech with more accuracy.

As we know speech recognition is always language dependent [4], in this case speech recognition is English language dependent. In conventional approach SR engine recognizes none other but English words. But if we focus at our own Pakistani environment, we use "URLISH" language, a mixture of two languages that are: URdu and engLISH. For example, if speaker speaks:

"Update Account Zia 5 2 3 4 at Twenty Thousands"

In the above speech, "zia" is not English word; hence it cannot be recognized.

The algorithm we proposed provides speaker an opportunity to add words like "Jamal Din", "Samandar Khan", "Sakeena Bibi" and so on in the grammar so that they can be recognized.

The system takes the word as a priority list of words based on the environment and then whenever the speaker speaks the sentences of "URLISH" they can be recognized. Hence the dynamic grammar helps the user to modify and use the SR system according to the environment he is working in.

V. Future Recommendation

When a speaker speaks a word, listeners listen to the sound and process it. The proposed system compares the word in the dictionary along with grammar and finds the probability of occurrence; more probability has more chance to be a fit candidate for the output as a recognized word [5]. Now if there is more than one listener at a time, the accuracy level of recognition goes high.

When we talk about parallel processes we first think of "THREADING PROCESS". Although threading is a good technique, in a single processor the concept of threading is actually "TIME SHARING". As it is understood that speech recognition is a very time critical and fine process so it cannot be handled accurately through a "PARALLEL-LIKE" process of threading rather it needs pure parallel processing units which listen to the sound. Table I shows Parallel-Like process for speech recognition whereas results of Parallel-Like process are presented in Table II.

TABLE I: Showing Parallel-Like Process for Speech Recognition

Threads	**T1**	**T2**	**T3**
Words	Pais	ata	kian
Reco. Result	Pays	ate	kin
Probability Of Recognition	80%	92%	76%
Final Result	"Ate" with 92% accuracy		

TABLE II: Showing Results of Parallel-Like Process

Threads	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
Time Division	**0.2 Sec**	**0.2 Sec**	**0.2 Sec**	**0.2 Sec**	**0.2 Sec**	**0.2 Sec**
Sound Input	pa	a	Ki	is	Ta	an

Modern speech recognition systems are generally based on Hidden Markov Models (HMMs)[6], [7]. This is a statistical model which outputs a sequence of symbols or quantities. Having a model which gives us the probability of an observed

sequence of acoustic data given one or another word (or word sequence) will enable us to work out the most likely word sequence [8].

By the application of Baye' rule:

$$\Pr(word \mid accoustics) = \frac{\Pr(acoustics \mid wrd)}{\Pr(acoustics)} \quad (1)$$

For a given sequence of acoustic data (think speech input), Pr (acoustics) is a constant and can be ignored. Pr(word) is the prior probability of the word obtained through language modeling (a science in itself; suffice it to say that Pr(mushroom soup) > Pr(much rooms hope)); Pr(acoustics word) is the most involved term on the right hand side of the equation obtained from the Hidden Markov Models.

In this paper we emphasize on the need of more than one listener to recognize the word. So in the Bayes' rule Pr (acoustics word) gets value from the dynamic grammar and each listener follows a different grammar [9]. If we just provide each listener with a single grammar based vocabulary then it is useless because of the following reasons:

- Single input
- Same process on each event
- Same probability of word recognition
- Same output

Hence each listener needs different set of vocabulary priority listening.

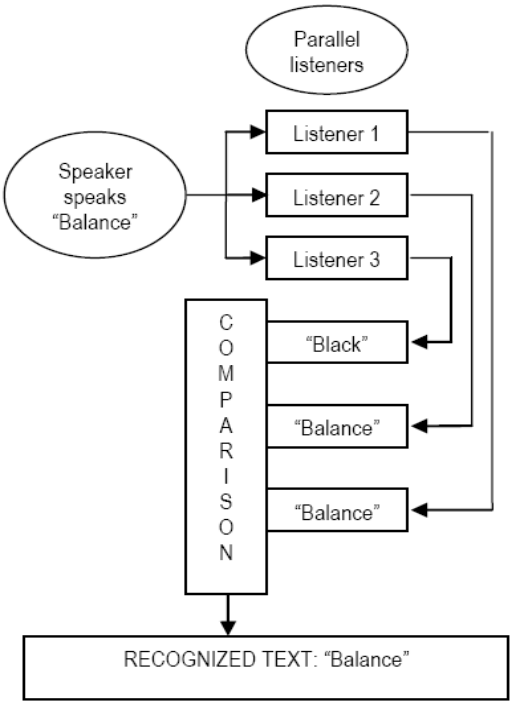

Fig. 3. Block diagram of parallel listeners along with dynamic grammar update example for a specific environment based system

Due to this technique, if we use three listeners then they all independently recognize the word spoken and at the end of their process if we compare the output from all of them we can achieve high accuracy. Fig. 3 shows block diagram of parallel listeners along with dynamic grammar update for a specific environment based system.

In the above example there are three listeners working in parallel. When speaker speaks "Balance", listener 1 recognizes "Black", listener 2 recognizes "Balance" and listener 3 recognizes "Balance". Now all these outputs are compared and output with high probability the final recognized text i.e., "Balance".

When each listener has its own predefined grammar containing a list of specific environment based words included then the scope of the system widens and we get the accurate recognition results more efficiently[10].

Grammar must have different sets of vocabulary to deal with number of specific modules in the respective system. Each vocabulary listing must be matured and focused on the nature of environment.

VI. Experiments

Experiments have been performed to test the efficiency and accuracy of the proposed system. 150 words vocabulary is used for testing. For improving the complexity and universality of the test database, the words were acquired from different environments and work fields. The words of different pitches and variable sizes were taken. These results reported a high accuracy rate of above 85% as shown in Table III.

TABLE III: Showing Parallel Processes with Dynamic Grammar

Parallel Processors	P1 With Country Names Priority Grammar	P2 With Province Names Priority Grammar	P3 With City Names Priority Grammar
Sound Input	Pakistan	Pakistan	Pakistan
Reco. Result	Pakistan	Balochistan	Paristan
Probability Of Recognition	100%	50%	86%
Final Result	Pakistan With 100% Probability		

VII. Conclusion

The algorithm designed in this paper attains high accuracy level for recognition of speech from the sound using a hybrid of dynamic grammar and parallel recognition events in runtime environment.

The main advantage of the proposed technique is high accuracy of the system working in a particular environment. Therefore this system can be used effectively in any environment of any country. These features can be improved with more optimized models and techniques but this particular model also steps forward the ongoing process of advancement in the field of knowledge with a relatively new perspective of innovation.

Future work is intended to be done on the efficiency of the system so as to make it computationally more efficient.

Acknowledgment

We acknowledge COMSATS Institute of Information Technology, Pakistan to support us in this work.

References

[1] B. Claudio and L. Ricotti, *Speech recognition theory and C++ implementation*, John WILEY&Sons, Ltd, pp. 125-137, 1999.

[2] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. 1990 International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 849-852.

[3] L. R. Bahl *et al.*, "A maximum likelihood approach to continuous speech recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, pp. 179-190, 1983.
[4] E. Bocchieri and J. Wilpon, "Discriminative feature selection for speech recognition," *Computer Speech & Language*, vol. 7, pp. 229-246, 1993.
[5] B. Baldauf and G. Santoni, "Stock price volatility: some evidence from an ARCH model," *Journal of Futures Markets,* vol. 11, pp. 191-200, 1991.
[6] J. H. S. M. Sharif, S. Mohsin, and M. Raza, "Sub-holistic hidden markov model for face recognition," *Research Journal of Recent Sciences*, vol. 2, pp. 10-14, 2013.
[7] A. Rosti and M. Gales, "Factor analysed hidden Markov models for speech recognition," *Computer Speech & Language*, vol. 18, pp. 181-200, 2004.
[8] D. M. Cutler *et al.*, "Speculative dynamics," *The Review of Economic Studies*, vol. 58, pp. 529-546, 1991.
[9] W. Buntine, "A guide to the literature on learning probabilistic networks from data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 8, pp. 195-210, 1996.
[10] R. D. Mori, *Spoken Dialogues with Computers*, Academic Press, Inc., 1997.

Muhammad Sharif has been working under Department of Computer Science COMSATS Institute of Information Technology, Wah Campus, Pakistan. He is a PhD scholar at COMSATS institute of information technology, Islamabad Campus, Pakistan. He has more than 15 years of experience including teaching graduate and undergraduate classes. His areas of interest are parallel and distributed computing, algorithms analysis and image processing.

Syed Zia Uddin Bukhari completed his undergraduate from COMSATS Institute of Information Technology, Wah Campus, Pakistan. His areas of interest are Pattern recognition and Speech recognition

Aman Ullah Khan is working in Air University Multan Campus as Chair, Department of Computer Sciences and Engineering. after receiving an MSc degree in applied mathematics from Multan University, Dr. Khan secured masters and PhD degrees in computer sciences from Wales, UK. Dr. Khan has over thirty years of teaching experience at various prestigious institutions, including King Khalid University, Saudi Arabia. Besides teaching experience, Dr. Khan also brings with him extensive administrative and leadership experience. He has led the departments of Computer Sciences at Bahauddin Zakariya University Mutlan, COMSATS Institute of Information Technology (CIIT) WAH, and Faculty of Information Sciences & Technology, CIIT Wah.

Muhammad Waseem Khan is a student of MS (computer science) at COMSATS Institute of Information Technology, Wah Campus, Pakistan. He has completed his bachelor degree from COMSATS Institute of Information Technology, Lahore in 2010. His areas of interest are Image Processing, Information Security and Mobile Computing.

Mudassar Raza is working as an assistant professor at COMSATS Institute of Information Technology, Wah Campus, Pakistan. He has more than seven years of experience of teaching computer science subjects to undergraduate classes. His areas of research are parallel and distributed computing and image processing