DEVELOPMENT OF MULTIPLE AUTOMATIC SPEECH RECOGNITION SYSTEMS IN THE GALAXY FRAMEWORK

Muhammad Qasim, Aneek Anwar, Tania Habib*, Sarmad Hussain
Center for Language Engineering,
Al-Khwarizmi Institute of Computer Science, UET, Lahore, Pakistan
Firstname.lastname@kics.edu.pk

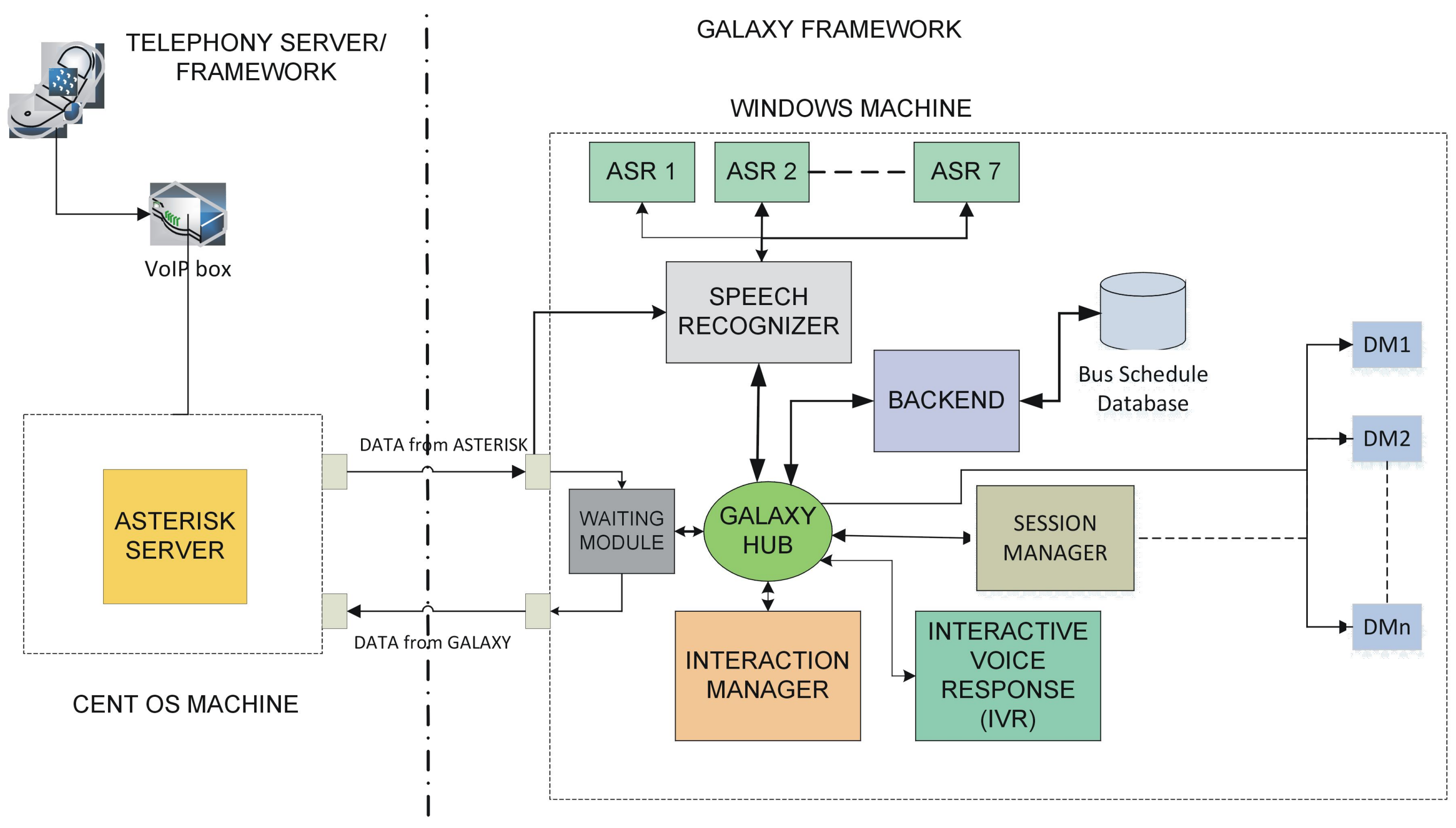
INTRODUCTION

Spoken dialog systems provide speech interface to users in order to access information. Most spoken dialog systems use a single Automatic Speech Recognizer (ASR) to understand the user's response. This paper presents a bus reservation system built to be used for travel reservation from Lahore city to 44 other cities of Pakistan. It uses multiple ASRs depending on the dimension of the user's response. Currently, the state-of-the-art in speech recognition are far from being perfect which results in high error rate in case of large vocabulary. Therefore, it makes sense to use separate ASRs for each dimension of user input as it reduces the vocabulary size for each ASR which in turn can lead to better performance.

System Architecture

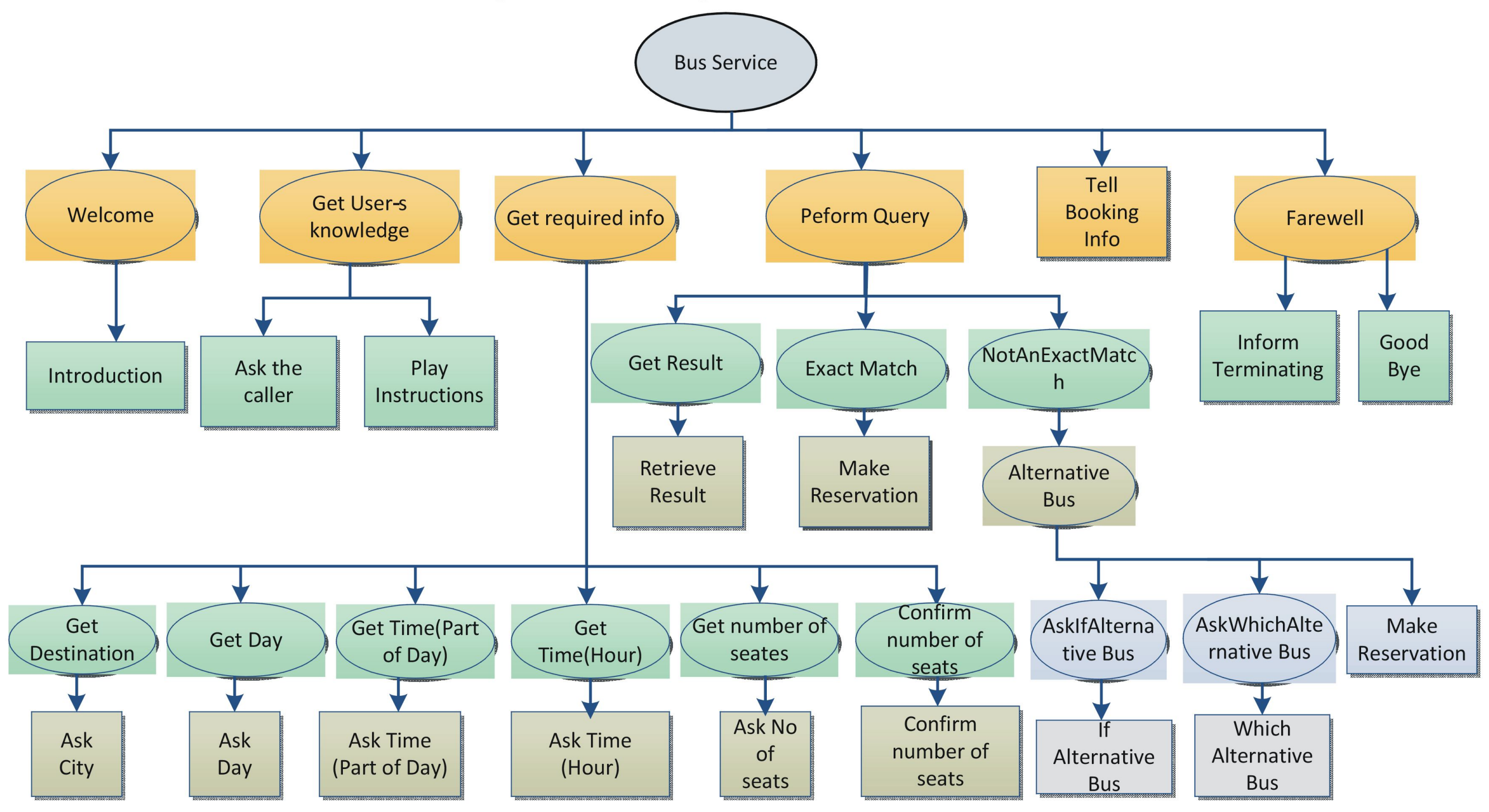
Bus Reservation System is developed by utilizing two open source elements; Galaxy framework and RavenClaw. The communication protocols of Galaxy architecture are used but all the individual modules have been developed from scratch.

- ***Telephony Framework/Asterisk Server*** – responsible for initiating the calls, getting user inputs and playing back the system response
- ***Speech Recognizer*** – decodes the utterance spoken by the user
- ***Backend Module*** – populates database on its startup and performs user's query
- ***Interactive Voice Response (IVR)*** – generates the system response
- ***Session Manager*** – keeps track of multiple Dialog Managers parses Dialog Manager messages for Hub and other modules
- ***Interaction Manager*** – parses Dialog Manager's messages for Hub and other



Dialog Flow

RavenClaw dialog manager is used in the dialog system which controls the entire dialog. It is programmed using the dialog task tree. The dialog task tree is traversed from left to right, starting from the left most node.



Speech Corpus

- Recorded from speakers of Punjab province
- Recorded in office environment over telephone channel

Recorded data duration	Number of Speakers	
	Male	Female
18 hours	418	300

Experiments and Results

Lab Testing of system

Type of ASR	Vocabulary size	Training Utterances	Testing Utterances	Correct Decoded	Accuracy (%age)
Destinations ASR	44	1543	584	563	96.40
Reservation Day ASR	23	805	307	291	94.78
Reservation Time ASR (Part of Day)	5	170	31	29	93.54
Reservation Time ASR (Hour)	19	659	219	204	93.15
Number of Seats ASR	10	385	150	146	97.33
ASR for choice of Bus	2	70	20	20	100
Confirmation ASR	2	70	26	26	100
Overall	86	3043	1118	1075	96.15

Field testing of system

Testing Utterances	Correct Decoded	Incorrect Decoded	Accuracy (%age)
222	201	21	90.54054

Conclusion

- System performs reasonably well in low noise
- Use of multiple ASRs has certainly improved the recognition of user input
- Error handling capabilities of the system make it very user friendly

Future Work

- Interaction between user and system can be modified to be more flexible
- Work is being done to reduce the overall call time by merging reservation time (part of day) and reservation time (hour) fields

Acknowledgment

This work has been conducted through Enabling Information Access through Mobile Based Dialog Systems and Screen Readers for Urdu project supported through a research grant from National ICT RnD Fund, Pakistan.

National ICT R&D Fund

مرکز تحقیقات لسانیات