Development of a Spoken Dialogue System for accessing Agricultural Information in Telugu

Gautam Varma Mantena, S. Rajendran, Suryakanth V. Gangashetty,
B. Yegnanarayana, Kishore Prahallad
Speech and Vision Lab,
Language Technologies Research Center,
IIIT - Hyderabad, Hyderabad, India.
gautam.mantena@research.iiit.ac.in, su.rajendran@gmail.com
{svg,yegna,kishore}@iiit.ac.in

Abstract

In this paper we describe the development of Mandi Information System, a Telugu spoken dialogue system for obtaining price information of agricultural commodities like vegetables, fruits, pulses, spices, etc.. The target users of MIS are primarily the farmers in rural and semi-urban areas. Speech recognition is error prone and it is necessary for the dialogue system to make minimum number of errors while acquiring information from a user and also to detect errors (if not correctable) and adopt appropriate strategies. In this paper we suggest an approach to improve the performance and usability of the system by using multiple decoders and contextual information.

Index Terms: spoken dialogue system, speech recognition, multiple decoders, contextual information

1 Introduction

Human-computer interaction plays a significant role for literate/illiterate and visually challenged users to access information. The mode of human-computer interaction could be speech, text, gestures, symbols etc., or a combination of these. An interaction could be in the form of a conversation including statements, questions, answers and expressions. A uni-modal or multi-modal conversational type of human-computer interaction is often referred to as a conversational system. A conversational system with speech as an input mode assumes significance as speech is the most natural means of communication for human-beings. The goal of a speech-based conversation (SBC) system is to provide information by conversing with a human-being in a natural fashion. Our objective is to develop speech based conversational systems for information access in Telugu, spoken in Andhra Pradesh state in India.

As shown in Fig. 1, a simplistic view of a speech-based conversational system consists of: Automatic speech recognition (ASR) which converts speech to text, natural language understanding (NLU), dialogue manager (DM), natural language generation (NLG) and text-to-speech (TTS) (Jurafsky and Martin, 2008; McTear, 2002). When a user utters a query to a SBC system, the speech is converted to text by an ASR. This text is parsed by an NLU module to extract the information or concepts relevant to the state of the conversation. These concepts are passed to the DM which is the core component of a SBC system. DM determines the necessary actions to be performed and response to be given to the user. The required response is provided to the user via NLG in generating the appropriate sentences which are then synthesized by a TTS module. In this work, a preliminary version of speech-based conversational system is demonstrated for accessing price of agricultural commodities by farmers in Andhra Pradesh. We refer to this conversational system as Mandi Information System (MIS).

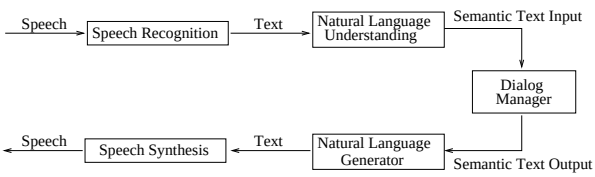

Figure 1: *Architecture of a Speech Based Conversation System.*

Proceedings of ICON-2011: 9th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2011

MIS is a telephone-based conversational system designed to facilitate the farmers, traders and other consumers of the agricultural products in rural and semi-urban areas to obtain the prices of commodities (vegetables, fruits, pulses, spices ,etc.) that are being sold in the markets across the Andhra Pradesh state. The prices of agricultural commodities are obtained from `http://agmarknet.nic.in/` on a daily basis and is provided by the Ministry of Agriculture, Government of India.

The following are the major issues involved in the development of MIS.

- *Noisy environment:* The target user of MIS are primarily the farmers in rural and semi-urban areas. The user can call from anywhere, from farm fields or while driving a powered vehicle or in a busy urban locality. The farmers call MIS through their mobile phones or landline. The quality of speech signal is affected by the distance of microphone, mobile/telephone handsets, speech codecs and communication channel. MIS system is expected to work in noisy environments, including background speech.

- *Dialect/Pronunciation variation:* Linguistically there are four distinct dialects of Telugu in Andhra Pradesh, namely Telangana, Rayalaseema, Kalinga and Coastal. Each dialect differs from the other at phonetic, phonological, morphological, grammatical and lexical level. Multilingualism adds further to linguistic variations in Telugu. The user can speak in any style.

- *Unstructured conversation:* The target audience of the MIS may not have interacted with a computer based information access system. Hence, the conversation is typically unstructured and will be filled with disfluencies including repeats and false starts. Another issue is eliciting the speech data from farmers to capture the acoustic and pronunciation variations for building an ASR.

- *Personalization:* Often, a frequent caller expects his/her preferred query to be automatically answered. Such personalization requires identification of the user from his/her voice samples. It would also require mining user's previous calls to predict his/her preferred query during the next call.

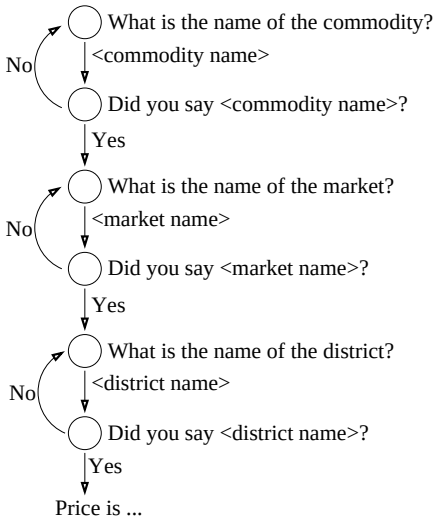

Figure 2: *State diagram for Mandi Information System.*

This paper describes our current efforts towards reducing the number of turns in a conversation with MIS. The organization of this paper is as follows: Section 2 describes the development of a baseline version of MIS. Section 3 proposes the use of multiple decoders and contextual information for reducing the number of interactions and obtaining a better dialogue flow for MIS. Section 4 discusses the evaluation of the baseline and improved version of MIS.

2 A Baseline Mandi Information System

2.1 Data Collection and building Acoustic Models

Speech data was collected via telephone (digital line) by the following three methods:

1. Users were asked to read out the names of the commodities, place names, etc. from the text provided to them.

2. Users were shown a series of pictures of agricultural commodities and were asked to say the names of the commodities shown in the picture. This is to collect the dialectal variations of the commodities.

3. Users were asked a series of questions related to agriculture and the places around their locality. This is to record conversational speech.

The reported work is based on the speech data collected from 96 speakers, where each speaker

spoke approximately 500 words. The total duration of the collected speech data is about 17 hours. We refer to this data as *Mandi data*.

Table 1: *Vocabulary size used in Mandi Information System.*

Word Category	Vocabulary Size
Commodity	72
Markets	348
Districts	23

Approximately 15 hours of *Mandi data* was used to train a set of acoustic models referred to as *AM-1*. These are context dependent tri-phone HMM models built with 8 Gaussian mixtures per state using Sphinx recognition system (Sphinx3,). These models were tested against 20 speakers of 5718 test utterances and the accuracy was found to be 81.24%.

2.2 MIS Dialogue Structure

MIS expects three concepts or inputs from the user, which are commodity, district and market names. A finite state dialogue manager built to extract these three concepts in the basline MIS is as shown in Fig. 2. When a user provides some information to MIS, the goodness of recognition hypothesis needs to be checked. This can be accomplished using explicit/implicit or no confirmation strategy. As shown in Fig. 2, the DM in baseline MIS would ask for an explicit confirmation. The user is required to respond either yes/no. This is for MIS to make sure that the input query is right as recognition is error prone. The system needs to identify the errors and come up with appropriate strategies to overcome the errors. A sample conversation recorded by the MIS is as shown in Table 2.

Table 2: *Sample conversation where the system asks for an explicit confirmation from the user.*

System:	What is the name of the commodity?
User:	*Orange*
System:	Did you say *orange*?
User:	Yes

The strategy of explicit confirmation is hardly convenient, as the user has to confirm to all the information he has provided to MIS queries. Hence, we propose a technique in Section 3, where the number of explicit confirmations are minimized by using recognition hypothesis from multiple ASR systems and the contextual information pertaining to that state.

3 Improving the Dialogue Flow

An efficient spoken dialogue system would provide accurate information to a user in less number of turns (or interactions). Speech recognition being error prone, it is difficult to avoid confirmations from users. However, the objective would be to limit these confirmations. An approach would be to associate a confidence score to the recognition output of an ASR. Examples of confidence scoring techniques include normalized likelihood scores, counts from N-best hypothesis, language model scores, parsing related etc (Jiang, 2005). It is also necessary for the system to use features from various levels of dialogue system. In (Carpenter et al., 2001) decoding, parsing and dialogue features have been used to measure the confidence score. Confidence scores could also be calculated on the N-best lists using semantic and pragmatic features along with the acoustic scores (Timothy J. Hazen and Seneff, 2000; Gabsdil and Lemon, 2004; Thomson et al., 2008). The obtained score is typically subjected to a threshold to decide the correctness of the recognized hypothesis. Most often, these thresholds are data driven, i.e., obtained on a held-out test set.

3.1 Multiple Decoders

In this work, we propose a confidence scoring technique based on multiple ASR decoders. The key idea is that one could build multiple ASR decoders, where each decoder tries to capture complementary information about the speech data. This could be attempted through training multiple decoders using different training datasets, or different features such as Mel-frequency cepstral coefficients and linear prediction cepstral coefficients. Given these multiple decoders, if a majority of these decoders agree on an hypothesis, i.e., their recognized output is same, then dialogue manager could choose to avoid an explicit confirmation from the user.

Consider a set of decoders $\{d_1, d_2\} \in D$. For a given acoustic signal, let the corresponding hypothesis of D be $\{h_1, h_2\} \in H$. Let C_i is the contextual information for the dialogue state i. The following are possible cases and the correspond-

ing actions incorporated into MIS using multiple decoders and contextual information:

Case 1 : The hypotheses h_1 and h_2 are same and are present in the contextual information: $h_1 = h_2$ and $h_1 \in C_i$

Action: The recognition output is most likely to be correct and the system would jump to subsequent dialogue states.

Case 2 : $h_1 = h_2$ and $h_1, h_2 \notin C_i$

Action: The recognition output is most likely to be correct, but input is of no help as it is not present in the contextual information. The system would prompt the user, saying that no such information is available pertaining to that given input and would ask the user to provide some other query.

Case 3 : $h_1 \neq h_2$ and $h_1/h_2 \in C_i$

Action: In such cases, system would try to consider the hypothesis that is present in the contextual information and discard the other. To make sure that the recognition is correct, the system would ask for an explicit confirmation from the user.

Case 4 : $h_1 \neq h_2$ and $h_1, h_2 \notin C_i$

Action: Mis-recognition might have occurred and the system would prompt the user to provide the information again.

It should be noted that the above discussion presented for two decoders can be easily extended for $n(> 2)$ decoders.

3.2 Role of contextual information

In the current version of MIS, the scheme of multiple decoders is implemented using two decoders. Each decoder differs from the other in its acoustic models. The first decoder uses the acoustic models *AM-1* which is built on Mandi data as described in Section 2.1. The second decoder uses the acoustic models referred to as *AM-2*. These acoustic models are built on a speech corpus consisting of 15 hours of continuous speech data collected from news domain (Anumanchipalli et al., 2005) and 15 hours of *Mandi data*. On a test data set of 5718 utterances from 20 speakers, the word level accuracy of *AM-1* and *AM-2* is shown in Table 3.

Fig. 3 shows an example dialogue flow of the MIS system using multiple decoders. The MIS system would opt for an explicit confirmation only when the multiple decoders are not in agreement. Hence, it is important to understand as to how often these decoders are not in agreement with each other and the role of contextual information under this condition.

Table 3: *Recognition score of the speech recognition module using the built acoustic models. 5718 test utterances of 20 speakers were taken to obtain the recognition accuracy.*

Acoustic Models	Speech Data	Accuracy
AM-1	*Mandi*	81.24%
AM-2	*Mandi + Continuous speech*	76.86%

Table 4 shows the number of times *AM-1* and *AM-2* agree/disagree on their hypotheses. This evaluation was done on test data set. From Table 4, it could be observed that both the decoders agree on the same hypothesis in 71 out of 100 attempts. When the decoders are not in agreement, then the contextual information (expected set of concepts for the given state of a dialogue) plays a role in accepting/rejecting the hypothesis. This is stated as Case 3 and Case 4 in Section 3.1. For the cases of ($\sim$A, B) and (A,$\sim$B) in Table 4, if one of the hypothesis is present in the contextual information, then it could be asserted through an explicit confirmation from the user. Table 5 shows an example dialogue for Case 3 in the MIS.

Table 4: *Comparing AM-1 and AM-2 on test data set. Here, A indicates AM-1 being correct in its hypothesis, while B indicates AM-2 being correct. $\sim$A indicates AM-1 being incorrect. $\sim$B indicates AM-2 being incorrect in its hypothesis.*

	B	$\sim$B
A	4092 (71.56%)	541 (9.46%)
$\sim$A	303 (5.29%)	782 (13.67%)

Another scenario in the MIS is that the user provides non-contextual information as input at a given state. For example, the user would utter a market name that is not present in the district or not selling a particular commodity. In such cases, it is important to identify that a non-contextual information has been provided. Hence, for a given dialogue state, models representing context and a non-context should be used. In the current version

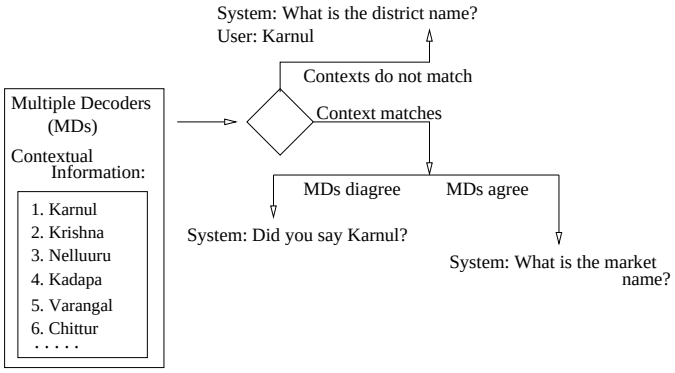

Figure 3: *Use of Multiple Decoders and Contextual Information in determining the dialogue flow.*

Table 5: *Recorded MIS conversation with a user. CI is the contextual information and AM-1, AM-2 are the acoustic models. CI is the contextual information and AM-1, AM-2 are the acoustic models.*

System:	What is the commodity name?
User:	*Red Grams*
(*AM-1*, *AM-2*:)	*Red Grams*
System:	*Red Grams* is being sold in 9 districts.
System:	What is the name of the district?
User:	*Karnul*
(CI:	*Karnul, Krishna, Nelluru, ...*)
(*AM-1*, *AM-2*:)	*Karnul*
System:	*Red Grams* is sold in 6 markets in *Karnul*
System:	What is the market name?
User:	*Dhoni*
(CI:	*Adoni, Karnul, Dhoni, ...*)
(*AM-1*:)	*Dhoni*
(*AM-2*:)	*Guti*
System:	Did you say *Dhoni*?
User:	Yes

Table 6: *Handling Non Contextual Information. CI is the contextual information and AM-1, AM-2 are the acoustic models.*

System:	*Rice* is sold in 3 markets in *Karnul*
System:	It is being sold in *Maacharla, ...*
System:	What is the market name?
User:	*Bhimavaram*
(CI:	*Macharla, ...*)
(*AM-1*:)	*Bhimavaram*
(*AM-2*:)	*Bhimavaram*
System:	Rice is not available in *Bhimavaram* market in *Guntur* district

of MIS system, a non-context is simply the list of concepts (i.e., district/market names) not relevant to the current state. Table 6 shows an example dialogue of MIS system for a non-contextual input from a user.

4 Evaluation Results

Experiments were conducted to evaluate the baseline system (MIS version 1) and the improved MIS (multiple decoders and contextual information included and is called MIS version 2). A set of 8 subjects were randomly chosen and were asked to access the MIS systems to retrieve the prices of five commodities of their preference. They were not trained or provided any information on the dialogue flow. The results are as shown in Table 7. The success of each trial was based on whether the system was able to retrieve the required information to the user or not.

Table 7: *Total of* $8 * 5 = 40$ *tests conducted on the system, with 5 male and 3 female speakers.*

System	Accuracy
MIS version 1	57.4% (23/40)
MIS version 2	77.5% (31/40)

From the Table 7, it could be observed that the MIS version 2 performed better than MIS version 1. This is because of using the contextual information on the multiple decoders in obtaining the best hypothesis. This has led to decreased usage (i.e. less number of times an ASR is invoked) , and also reducing the chance of an error input

from the user. Moreover, the number of interactions have been significantly reduced (about 50%) in MIS version 2.

5 Discussion and Conclusions

In this paper we have discussed the development of a spoken conversational system referred to as Mandi Information System. This system is being developed for accessing prices of agriculture commodities in Telugu by farmers in rural and semi-urban areas. We have discussed the baseline MIS system and also an improved version using multiple decoders and contextual information. The goal is to have an effective telephone based service for the farmers to obtain price information on a daily basis. We hope to apply a similar strategy for continuous speech and make the system as natural and robust as possible.

6 Acknowledgements

We would like to acknowledge the help of P. Vishala and B. Rambabu for speech data collection and transcription. This work was partially supported by the project Speech-based Access for Agricultural Commodity Prices sponsored by Ministry of Communication and Information Technology, Government of India in ASR consortium mode.

References

Gopalakrishna Anumanchipalli, Rahul Chitturi, Sachin Joshi, Rohit Kumar, Satinder Pal Singh, R.N.V. Sitaram, and S P Kishore. 2005. Development of indian language speech databases for large vocabulary speech recognition systems. In *Proc. SPECOM*.

Paul Carpenter, Chun Jin, Daniel Wilson, Rong Zhang, Dan Bohus, and Alexander I. Rudnicky. 2001. Is this conversation on track. In *Proc. Eurospeech*.

Malte Gabsdil and Oliver Lemon. 2004. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proc. ACL*, pages 344–351.

H. Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech Communication*, 45(4):455–470, April.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing*. Prentice Hall, 2 edition, May.

Michael F. McTear. 2002. Spoken dialogue technology: Enabling the conversational user interface. *ACM Comput. Surv.*, 34:90–169, March.

Sphinx3. CMU Sphinx, The Carnegie Mellon Sphinx Project. http://cmusphinx.sourceforge.net.

B. Thomson, K. Yu, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, and S. Young. 2008. Evaluating semantic-level confidence scores with multiple hypotheses. In *Proc. Interspeech*.

Joseph Polifroni Timothy J. Hazen, Theresa Burianek and Stephanie Seneff. 2000. Integrating recognition confidence scoring with language understanding and dialogue modeling. In *Proc. ICSLP*, page 2000.