WIZARD OF OZ EVALUATION OF A DIALOGUE WITH COMMUNICATOR SYSTEM IN CHILE

Néstor Becerra Yoma, Angela Cortés, Mauricio Hormazábal and Enrique López

Dept. of Electrical Engineering/University of Chile

Av. Tupper 2007, P.O.Box 412-3, Santiago, CHILE

nbecerra@cec.uchile.cl

ABSTRACT

Results of Wizard of Oz experiments with Communicator system in Chile are presented in this paper. The task was to book a domestic flight and was used to evaluate the user response to this type of system and, more specifically, to a diphone TTS implemented with speech samples from a Chilean speaker. Two synthesizers were employed, with and without pitch contour. The results suggest that the overall user attitude toward the system was positive. Moreover, a proper pitch curve seems to improve the intelligibility of the synthesized speech and the response to the system.

I. INTRODUCTION

Communicator (Pellom et al., 2000)(Ward & Pellom, 1999) is a system architecture designed to facilitate development of natural dialogue systems and to support basic research on system modules in the context of working systems. Communicator incorporates all of the component technologies needed to develop, deploy and evaluate advanced dialogue systems. In the experiments reported here a version of the CU (Colorado University) Communicator was modified and adapted according to a local task.

The CSLR (Center for Spoken Language Research) Communicator architecture is shown in Figure 1. The system uses a DARPA Hub-compliant architecture based on MIT's Galaxy II architecture (Seneff et al., 1998), which means that the system is composed of a number of servers that interact through a standardized communications interface known as the Hub. Besides the Hub, the CU Communicator system is also composed of eight servers:

Audio Server – receives signals from telephone and sends to recognizer, and sends synthesized speech to telephone;

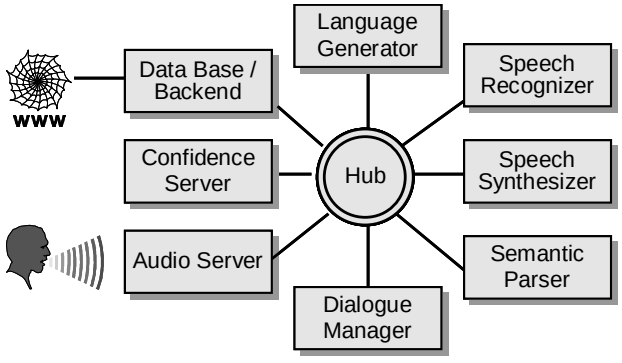
Speech Recognizer – Takes signals from audio server and produces a text string;

Natural Language Parser – Takes the text string from recognizer and produces the "best" interpretation of the recognized utterance;

Dialogue Manager – Resolves ambiguities in the interpretation; estimates confidence in the extracted information; clarifies with user if required; integrates with current dialogue context; builds database queries (SQL); sends data to natural language generation for presentation to user, prompts user for information;

Confidence Server – Takes hypothesis and semantic parse from the speech recognizer and parser as input and annotates the words, and concepts with levels of confidence;

Figure 1: Communicator System.



Database / Backend Server – Receives SQL queries from Dialogue Manager; interfaces to SQL database and returns data; and, retrieves data from the web;

Natural Language Generator – Constructs sentences to speak to the user based on the current dialog action;

Text-to-Speech Synthesizer – Receives word strings from the natural language generation; and, synthesizes them to be sent to the audio server.

Servers interact with each other by passing frames to the Hub. The Hub works as a router, which send the frames to the receiving server. CU Communicator in CSLR (http://cslr.colorado.edu) is running the Travelocity task that allows users to dial up the system using a telephone and natural language to interact with the automated travel agent. The agent then accesses the Internet for flight, hotel, and car rental information.

In this paper the CU Communicator was modified for a local task in Spanish in order to run usability tests in Chile. Due to slow Internet access to Travelocity WWW site, the original task was replaced with one concerning domestic flight information. The contributions of this paper concern: a) Wizard of Oz evaluation of Communicator in Chile; b) the design of a dialogue system adapted to Chilean reality; c) modification of the Speech Recognizer (ASR) server to run Wizard of Oz experiments; and, d) a TTS implemented based on a Chilean speaker.

II. COMMUNICATOR AT UNIVERSITY OF CHILE

The experiments were carried out with a dialogue system for domestic flight information. Fourteen Chilean cities and one or two airports per city were considered in the database. The following servers were modified: Parser; Dialogue Manager; Database/Backend; Language Generator; Speech Recognition; and TTS. The Confidence server was not used in this implementation.

To adapt the Parser server the empiric grammar rules were translated to Spanish. This task involved the translation of words, sentences and expressions according to the local use of the language. Some sentences like "It's all right" cannot be replaced just by "*Todo esto es correcto*" because answers such as "*esta bien*", "*correcto*", "*perfecto*", "*me parece*" o "*bueno*" are more common. The Dialogue Manager was also modified according to the task and the prompts were translated to Spanish. A dialogue sample is shown below:

Communicator: *Bienvenido al sistema communicator de la Universidad de Chile, para ingresar al sistema presione la tecla gato (#). Si usted se siente confundido, en cualquier momento puede decir "empezar de nuevo", cancelando asi el itinerario actual. Responda, ¿Está usted llamando de un teléfono celular?*

User: *No*

Communicator: *¿Cuáles son sus planes de viaje?*

User: *Quisiera ir de Santiago a Arica.*

Communicator: *¿En que día y mes del año desea partir de Santiago?*

User: *El 2 de marzo.*

Communicator: *¿A que hora desea viajar desde Santiago?*

User: *En la mañana.*

Communicator: *Si usted tienen una aerolinea de preferencia porfavor digala ahora o diga sin preferencia.*

User: *Sin preferencia.*

Communicator: *Entonces usted desea viajar desde Santiago a Arica el día dos del mes de marzo, ¿es esto correcto?*

User: *Si*

Communicator: *Por favor, déjeme verificar disponibilidad y ver qué puedo encontrar. Una opcion fue devuelta. Opcion uno, el vuelo siete treinta y cuatro de la aerolinea Lan Chile que sale de Santiago a las diez am y llega a Arica a las doce cero ocho pm, ¿le parece bien?*

User: *Si*

Communicator: *¿Desea volver de Arica?*

User: *No*

Communicator: *¿Es Arica su destino final?*

User: *Si*

Communicator: *Espere mientras cotizo su itinerario. El total de todo lo pedido es cuarenta y cuatro mil cuatro cientos ochenta y seis pesos. ¿Desea un resumen de su itinerario?*

User: *Si*

Communicator: *Aqui está su itinerario, el valor total de su pasaje es de cuarenta y cuatromil cuatro cientos ochenta y seis pesos, el día dos del mes de marzo, la aerolinea Lan Chile vuelo siete treinta y cuatro, sale de Santiago a las diez am y llega a Arica a las doce cero ocho pm. ¿Debo enviarle por email este itinerario?*

User: *No*

Communicator: *¿Obtuvo usted la informacion que estaba buscando?*

User: *Si*

Communicator: *¿Realmente va a tomar este viaje?*

User: *Si*

Communicator: *Gracias por utilizar el sistema de informacion de vuelos de la Universidad de Chile, que tenga un buen viaje.*

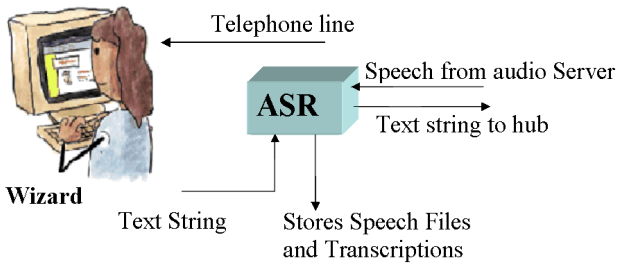
2.1. TTS

The TTS used in the CU Communicator uses a combination of variable-sized units concatenation and Festival, which provides synthesis with diphones. In the experiments reported here only Festival was employed in order to test the influence of pitch contour in the dialogue. The synthesizer used speech samples from a Chilean male speaker aged 24.

2.2. Speech Recognition Server

The Wizard of Oz configuration required the modification of the Speech Recognition (ASR) server according to Figure 2. The main idea was to replace the recognizer with the wizard who listens to the speech from the telephone line and types the text string corresponding to the speech signal transcription. Abbreviations were employed to substitute the more common words in the dialogue to reduce the time needed to type the user's answers. Finally, the modified ASR server stores the speech signals (from the Audio server) and transcriptions in files.

Figure 2: Speech Recognizer (ASR) server modified to run Wizard of Oz evaluation.



III. EXPERIMENTS

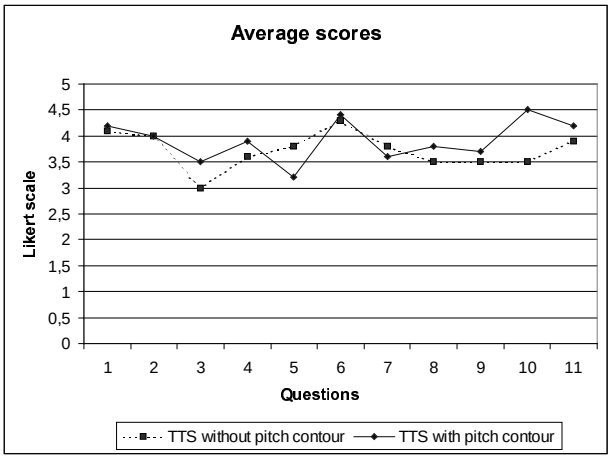
A Wizard of Oz experiment procedure was followed considering a continuous speech recognition system with 100% word accuracy. Twenty subjects, undergraduate students (15 males and 5 females) from the Faculty of Physical and Mathematical Sciences at University of Chile, were recruited for the experiments. Two TTS engines were compared: with and without pitch contour. Both TTS systems were tested with 10 subjects each. All the 20 subjects were exposed to the same task, which in turn was to book a flight between two Chilean cities.

The user satisfaction questions were all stated in terms of a 5 point multiple choice Likert scale (Nielsen, 1993**).** Each survey response is then mapped into a range of one to five: *Completo acuerdo* (strongly agree)-5 points; *acuerdo* (agree)-4 points; *neutro* (neutral)-3 points; *desacuerdo* (disagree)-2 points; and, *completo desacuerdo* (strongly disagree)- 1 point. To evaluate the user satisfaction, the following questions were employed:

Q1. *Usar el sistema fue fácil* (Using the system was easy).

Q2. *Comunicarse con el sistema fue fácil* (Communicating with the system was easy).

Figure 3: Average scores to Likert questionnaires.



Q3. *Entender lo que el sistema dijo fue fácil* (Understanding what the system said was easy).

Q4. *La tarea fue fácil* (The task was easy).

Q5. *Es fácil saber qué decirle al sistema* (It is easy to find out what to answer to the system).

Q6. *No hay nada irrelevante en el sistema, ninguna pregunta fuera de lugar o contexto* (There is nothing irrelevant in the system, no out-of-context questions).

Q7. *Es rápido completar la tarea deseada* (The task is completed fast).

Q8. *Tuve una buena compreensión de la estructura del diálogo* (I understood well the dialogue structure).

Q9. *Es fácil avanzar en el sistema* (Going through the dialogue is easy).

Q10. *Me gustó usar el sistema* (I liked to use the system).

Q11. *Siento que el sistema es eficiente* (I have the impression that the system is efficient).

The subjects were instructed about the task before calling the system. The questionnaires were filled immediately after the call. The results are presented in Figures 3 that shows the averaged response score achieved in each question within the subjects involved in the experiments.

IV. DISCUSSIONS AND CONCLUSIONS

Examining the overall user attitude towards the system with both TTS engines as calculated by averaging scores of each of the 11 Likert statements completed by subjects, the following picture emerges: firstly, the results suggest that the system showed a reasonably high acceptability from the user point of view, although a diphone TTS engine was employed; secondly, the introduction of a pitch curve to emphasize the questions improved the averaged overall score from 3.7 and 3.9.

Compared to other questions, Q3 provided a low score. However, the results show that the proper use of prosodic elements can improve the intelligibility of the generated speech and the response to the system.

V. ACKNOWLEDGMENT

The authors wish to thank Dr. Ron Cole (Director), Dr. Wayne Ward and Dr. Bryan Pellom from CSLR/U. of Colorado for the license and support to use the CU Communicator system. The research described here was supported by Fondecyt-Chile.

REFERENCES

Nielsen, J. (1993). *Usability Engineering.* Academic Press, 1993.

Pellom, B., Ward, W., & Pradhan, S. (2000). *The CU Communicator: An Architecture for dialogue Systems.* Proceedings of International Conference on Spoken Language Processing, Beijing, China, October, 2000.

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., Zue, V. (1998). *Galaxy-II: A Reference Architecture for Conversational System Developmen.* Proc. ICSLP-98, Sydney Australia, Vol. 3, pp. 931-934, 1998.

Ward, W., Pellom, B. L. (1999). *The CU Communicator System*. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Keystone, Colorado, December 1999.