TLATOA: Developing Speech Technology & Applications for Mexican Spanish

Ingrid Kirschning
Grupo TLATOA (CENTIA), Universidad de las Américas-Puebla
Sta. Catarina Mártr, Cholula, Pue., MEXICO
ingrid @ mail.udlap. mx

Abstract. This paper presents a summary of the activities and projects under way at “TLATOA” the speech technology group at Universidad de las Américas in Puebla, whose focus is on the research and development of speech technology and tools for Mexican Spanish. The main approaches addressed in our work is the improvement of recognition and synthesis of speech in Spanish. Also the development of applications that support education and information access are important issues on which our students do their research and theses work. Important in this process is always the collaboration with researchers and research centers, such as CSLR in Colorado University and the SLS at MIT, as well as companies investing in the formation of experts in speech technology.

1. Introduction

The rapid growth of the Internet has produced an extremely large network of information, interconnecting countries and cultures. This global connectivity opens many opportunities for people to have access to data and exchange information in order to solve different problems. However, it is still not there for everybody. Current technology provides ever-improving human-computer interfaces, but there exist large communities of citizens in different countries that cannot use a computer. They do now have access to computers or internet connection, nor do they have the know-how to find information that might be useful. This is caused mainly by economical, educational and technological recessions found in many areas of many countries.

Some of the fundamental problems faced by researchers in this area are the representation, sharing, access, retrieval and translation of information already present on the internet in other languages. Also, the heterogeneous information networks, which differ in design, reliability and performance can pose a difficulty to the communication. However, even when the digital data is available, our concern are the channels by which the access should be provided. Our work is focused on speech technology to provide services (information and education) in Mexican Spanish. As computers are not available anywhere, but a telephone is more commonly found, we base our work on this device to communicate to one central system as the gateway to the world of digital information, trying to breach the digital divide [1].

2. Development of Speech Technology for Mexican Spanish

Alongside the various advances in fundamental research on improving the speech technologies is the consideration of Human-Computer interaction factors, specific to every user’s culture and language that need to be included in information systems. Our work has been to perform basic research in the different speech processing techniques, improving the performance of speech recognition (based on both, Artificial Neural Networks or Hidden Markov Models and synthesis in Spanish (using dipphone concatenation and Unit Selection). We also work on the analysis of the Spanish language, dialogue structure and perception for the development of conversational speech interfaces and language learning and language acquisition systems.

3. Speech Recognition

In the area of speech recognition our research focuses on the improvement of the methods for training and development of speech recognizers, as well as training and testing new versions of our recognizers for Mexican

Footnotes:
1 Projects supported by the Mexican National Council of Science & Technology (CONACyT) grants No. CERVO30050-A and No. 35804-A, and a collaboration with CSLR, Colorado University (NSF grant No. IIS-0086107).
Spanish [2, 3]. We have recently trained a general-purpose recognizer based neural networks and another with HMM’s using the CSLU Toolkit and with a speech corpus containing more than 500 speakers. The performance has been tested on specific purpose corpora with excellent recognition rates [4]. The following table shows some of the results. The corpus used here contains 50 speakers (at 8 KHz) uttering sequences from 2 to 6 digits and the number 10.

<table>
<thead>
<tr>
<th>Recognizer</th>
<th>Accuracy</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Network</td>
<td>98.2%</td>
<td>Phonemes modeled as context-independent units</td>
</tr>
<tr>
<td>Neural Network</td>
<td>98%</td>
<td>Phonemes modeled as context-dependent units</td>
</tr>
<tr>
<td>HMM</td>
<td>71.19%</td>
<td>Phonemes modeled as context-independent units</td>
</tr>
<tr>
<td>HMM</td>
<td>92.14%</td>
<td>Phonemes modeled as context-dependent units</td>
</tr>
</tbody>
</table>

Table 1 Recognition results for the digits corpus.

The reason why the tests were done only with specific purpose corpora instead of all the available data, is that it was too large to test all together.

However, the corpus used to train these recognizers does not reflect all the different dialectal zones of Mexico, thus there is still room for improvement, once more data from different zones of the country becomes available.

4. Speech Synthesis

Until now speech the applications we have developed have used a Mexican Spanish voice developed by one of our members and provided now with the CSLU Toolkit for use with Festival. Although this speech synthesis was later improved by adding a duration estimation module [5], the voice still sounds robotic. Now we have started to develop a new voice with a better quality, using a Unit Selection approach for the synthesis. In this approach a speech corpus, recorded with high quality by a professional speaker is searched to extract large units of speech to concatenate them. These units of speech can be entire phrases, parts of phrases, words, or if none can be found, syllables. The result is a much more natural sounding speech. The program developed can use any pre-recorded corpus, which needs to be completely transcribed and labeled (word- and phoneme-level) [6].

5. Conversational Interfaces

The term “conversational” can have different meanings depending on the context, but in general it refers to an interactive system which works in a restricted domain. Although many speech interfaces are considered conversational, they largely differ in one main aspect, and this is the degree in which the system takes a more active role in the conversation [7].

Conversational systems allow users to interact with an automated system and recover information, perform transactions, or other tasks through a more or less flexible dialogue with the user.

5.1 Architecture of Conversational Interfaces

The elements required in a conversational system (see figure 1) are:

- A speech recognition module to convert speech to text.
- The Natural Language Understanding (NLU) module that obtains the semantic representation of the recognized speech.
- A language generating module which puts the retrieved information or the questions of the system in understandable and correct sentences / phrases.
- The dialogue manager, which keeps the control of the interaction, and can decide who is to have a more active role in the dialogue, the speaker or the system.

![Figure 1](image-url)
• A mechanism to transmit the information to the user, which can include, apart from speech synthesis, other devices to visually display the data as well.

The last decade has seen an increase in the research on conversational systems in USA as well as in Europe. Different applications have been developed for the retrieval of travel information, traffic information, travel reservations, car rental, where in Europe the focus has been strongly on multilingual systems.

Even though systems are being developed in Spanish (in Spain mainly) few can be used in other Spanish speaking countries. Although TLA TOA is a young group, it has established itself as one of the strongest speech technology groups in Mexico, generating young people with expertise in speech recognition, speech synthesis, natural language processing, conversational agents and speech corpora creation & processing.

Important in this process have been the different grants obtained from the Mexican National Council of Science & Technology (CONACyT) and the National Science Foundation (NSF). Also a critical factor is the amount of support we have received through our collaboration projects with CSLR (UCB), CSLU (OGI), SLS (MIT). Companies interested in the formation of experts in the field have helped to maintain the growth of the group through donation of infrastructure and scholarships for some students.

5.2 Conversational Interfaces for Spanish Language

Based on CSLU’s robust parser “PROFER” we had initially integrated natural language processing into query systems that allowed students to obtain information about their courses via a speech interface [8]. The CSL Toolkit is a good approach for developing in a fairly short time some NLU interfaces.

There are, however much more powerful architectures, that provide better capabilities for mixed initiative conversational interfaces.

In order to provide better access to digital information (Digital Libraries, information on the Internet, etc.) we are beginning to work with the CU-Communicator System [9, 10], which has the architecture shown in figure 2. It is our objective to develop conversational applications that are able to perform a dialogue in Spanish, detecting the degree of expertise of a user with a system and allow a flexible communication where the system can take or relinquish the initiative.

This is in order to be able to obtain enough information from the user to perform the required task and at the same time give the user the chance to provide more information than was specifically asked for but the user already knows, speeding up the process of completing the information. Mixed initiative techniques try to give more flexibility to an expert user and guidance to the newcomer in order for both to reach the same goal in a user-friendly environment. It implies the monitoring of the turns taken in the dialogue between the user and the system, both trying to reach the same goal (depending on the task this can be retrieval of information, flight reservation, etc.).

The implementation of the necessary modules to maintain a mixed-initiative interaction in Mexican Spanish includes natural language analysis and processing [11] and an user agent that monitors the dialogue, registers the elements relevant in the context (filling the slots of required information) and the data given by the user, until it is enough to form a query to consult a database. The development of mechanisms to make digital data accessible to multiple users, together with query languages to access databases implies the definition of task specific vocabularies, dialogue structures and language models.

![Figure 2. GalaxyII architecture (taken from[ 10])](image)

6. Speech Application Development

Several small applications have been developed as demonstration systems like a voice mail, a system to access e-mail via the telephone among others [12, 13]. Additionally we collaborated with SpeechWorks Intl. in the
development of an auto-attendant for the university, as well as a system that allows the students check on their account status with the university, also via the telephone.

One very important application field for speech technology is the education. As part of a larger project to be developed we created a couple of tools for a computer-assisted language-learning environment. These tools are a first prototype for pronunciation verification and a bilingual dictionary, both for Spanish language students, whose native tongue is American English [14,15]. Another very interesting aspect is the use of speech technology based systems to support language acquisition for children with hearing disabilities. We are developing a system based on the CSLU Toolkit as a first prototype for the Jean Piaget Special Education School in Puebla. The main challenge here is to create a tool that can be used also by the children that do not speak yet (or do not modulate correctly) so that they can practice pronunciation and reading, as well as vocabulary in different study fields. We need to design the interface and the content of the lessons in a way that is easy to use for the students and easy to manage by the teachers. As a first step a vocabulary tutor from CSLU has been adapted from English to work in Spanish, as shown the following figure:

Figure 3. Sample screen of the tutor-system that practices the states of the country.

Based on this system as a first prototype, we intend to continue working in two main directions: teaching children in general to read, write and spelling in Spanish and, second, support children with disabilities to acquire vocabulary and speech skills.

7. Corpus Development

Each and every one of the previous projects could not be done without the existence of a sufficiently complete speech corpus. Our group has recorded a corpus consisting of 550 speakers, mainly from the central area and a little from the north of Mexico, speaking a large variety of words (names, numbers, digits, letters) as well as spontaneous speech. This has been one of the most time-consuming tasks, but now this corpus, completely transcribed and labeled, is available to anyone without cost for educational and research purposes (http://mailweb.udlap.mx/~sistemas/tlatoa/corpora/corpora.htm). The recording of speech corpus is a continuing effort, aiming to cover in some near future all the dialectical zones of the country, as well as different ages, including children.

8. Conclusions

Our main interest is first, to produce students with the knowledge and capability to work with speech technology, and the awareness of CHI specific to their environment. Second, we focus on the development of technology that is accessible to Mexican people and all applications that provide a support/aid to the real needs of Mexican society, using speech technology. Among these applications are systems for language learning/acquisition and systems for the access of digital information via devices such as the telephone.

References


---

1 Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.