Verification of Correct Pronunciation of Mexican Spanish using Speech Technology[1]

Ingrid Kirschning and Nancy Aguas

Tlatoa Speech Processing Group, ICT, CENTIA[2],
Universidad de las Américas- Puebla. Sta. Catarina Mártir s/n,
72820 Cholula, Pue., México.

Abstract

This paper presents a new method for the verification of the correct pronunciation of spoken words. This process is based on speech recognition technology. It can be particularly useful when applied to the field of SLA (Second Language Acquisition) in learning environments or Computer-Aided Language Learning (CALL) systems, where the students can practice their pronunciation skills. This method uses an artificial neural network plus a specific grammar for each utterance to compare the text of the expected utterance with the sequence of phonemes recognized in the speech input, in order to detect the pronunciation errors.

1 Introduction

Due to the latest developments in computer technology regarding processing speed, audio devices and storage, the quickly evolving speech technology can now be used in many different applications, able to run on personal computers. This fact makes the newest speech technologies accessible to anyone who can use a computer. Among those applications that seem most suitable for the integration of speech interfaces are the language learning systems. There is a large variety of speech interfaces integrated into computer aided language courses.

There are several commercial products already on the market but none provides an explicit feedback about the pronunciation errors. They usually provide an overall score for the complete utterance. Others display the soundwave of the "correctly" pronounced word and the soundwave of the word uttered by the user. The user should then be able to detect his/her own errors by comparing both soundwaves. None of the products found, showed an efficient way to find out exactly why and where an utterance is not accurately pronounced.

This paper presents a method that enables an automatic speech recognition system to perform an explicit analysis of the correct pronunciation of words and/or phrases. The prototype shown here is part of an undergraduate thesis of one of the members of the TLATOA Speech Research Group at the UDLA (Universidad de las Américas-Puebla. It is also a part of a larger project of TLATOA concerning the development of a second language learning environment applying speech technology.

The following sections introduce the motivation for this work and some examples of software products created for second language learning. Then it presents a new method using neural network based speech recognition to check pronunciation.

[1] Research funded by CONACyT, project No. I28247-A
[2] Research Center for Information and Automation Technologies

The verification method is implemented in a prototype, where a simple interface has been designed to test the recognizer's ability to spot right and wrong pronunciations. The results obtained so far are presented in the last section followed by the conclusions and further work for the near future.

2 Computer Asisted Language Learning (CALL)

Computer Assisted Instruction (CAI) is a means through which computers aid users in the process of learning. Over the past 40 years, there has been an exponential growth in the use of computers assisting in instruction. During this rapid technological revolution, CAI has become more refined as a result of the paradigmatic shifts from behaviorist thinking to cognitivism to constructivism. Computers now serve as a sophisticated medium through which instruction can be delivered.

Symbolic AI proposed interesting schemes in ICAI, but the use of neural networks has been proposed very few times. In general the knowledge for CAI is rather represented explicitly, therefore neural network architectures have rarely been used in this area [Ayala, 99].

CAI gives way to the discovery mode of teaching and learning whereby the student is involved much more freely in self-directed intellectual exploration of the knowledge or the skill domain incorporated in the system. It serves as a tool through which learning can occur. In the last years more and more CAI systems have incorporated multimodal interfaces, i.e. sound, images, video and interactive programs to support various applications. More recently, the incorporation of speech interfaces into the area of CALL or ICALL (Intelligent Computer Aided Language Learning) creates new possibilities for the development of learning tools [Bull, 94].

3 Speech Technology and Education

In the various fields of language education, different groups have developed interesting and useful applications. For example, there is a software product called Pronunciation Power [PronunciationPower], which presents to the users a collection of tools that permit them to learn about the pronunciation of English language. It presents the written form of the sounds, lets the users record their own speech and compare every sample to a "correct" version of the soundwave.

The apparent disadvantage of this method is that when there is a pronunciation error this is not stated explicitly. It requires practice to analyze a soundwave and determine if the variation should be considered an error. It might be relatively easy for an adult user, but it might be tedious. Additionally, it is not known under which criteria the "correct" waveform was chosen. A correct utterance can vary in its appearance enough due to the natural differences in the human vocal tract to make the user believe that something might be wrong.

Another interesting example is from *Language Connect* [Language Connect], which uses IBM's ViaVoice. The software "listens" to each word and phrase, even complex sentences, then responds whether the user would be understood by a native speaker, also giving a general score for the pronunciation of the whole word or phrase. The speech recognition software is very powerful and has a good accuracy; however, this score the student receives also doesn't state explicitly which part of the utterance was not correctly pronounced, nor how it can be corrected.

In another application field, the Center for Spoken Language (CSLU) [CSLU] has been collaborating with educators at the Tucker Maxon Oral School [Tucker Maxon] in a joint effort focused on speech training with profoundly deaf children. They have developed a Toolkit that incorporates speech recognition and synthesis facilities, as well as an animated conversational agent, called Baldi [CSLU] [Cole et al.,98] [Cole et al.,99]. The agent is represented by an animated 3D face that produces visual speech: facial expressions, eyebrows, lip, jaw and tongue movements during speech production using a synthesized voice [Black&Taylor,97]. The children can play with the speech interface where a lesson presents different questions and a correct answer permits them to continue (see figure 1).

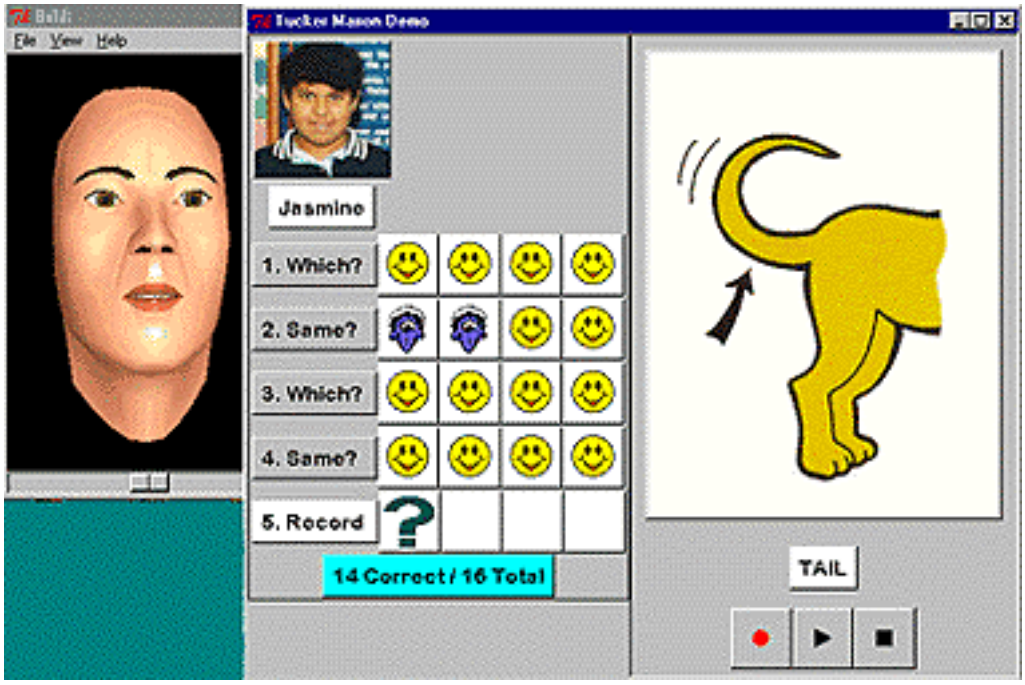

Figure 1: An example of the animated agent and a language training lesson [Cole et al.,98] [Cole et al.,99]

Apart from the software mentioned above there exist a large variety of tools, programs and games that intend to help people to learn a language. Most of these tools are quite impressive and they adapt to different needs of the users.

TLATOA is collaborating since 1997 with CSLU in the construction of the Mexican Spanish version of the CSLU Toolkit [Serridge et al.,98]. At the UDLA, as in many other institutions, foreigners spend their summer in language courses to learn Spanish. A large percentage of these students come from the United States. The ever-increasing demand of Spanish courses has prompted us to think of other means to provide suitable learning environments for the students. Thus a project for the development of a Mexican Spanish learning environment has been initiated.

As a part of this project that is still in an early stage, we focused on the problem of verifying the correct pronunciation of a word or phrase. Since language learning is about to be able, not only to write it, but also to speak, the training of a correct pronunciation is obviously important.

4 Verifying Correct Pronunciation

4.1 System Architecture

The main objective of this work is to test how far a neural is capable of detecting mispronunciations in a stream of speech. In a first step the recognizer has to recover the exact sequence of phonemes as the user pronounced them, without neither correcting any nor fitting them to the correct word model. A normal speech recognition system processes the speech input and produces as an output one of the words or phrases of the vocabulary with the highest similarity score. For this purpose the speech recognition system uses a vocabulary and a grammar to try to map the soundwave to one of the options in the grammar, rejecting results with low recognition confidence (below a determined threshold). These systems usually try to overcome noise and some errors to find the closest words - although not always corresponding to the input steam. The more robust a system is, the better, because it is able to produce the correct output most of the times.

Applying this to a language training system, a student can be asked to try and pronounce a set of words according to the level of difficulty of the lesson. A speech recognition system can easily be used to try to recognize the student's speech and give a value of certainty of the recognition. We believe that it would be extremely useful to be able to tell the

student also which part of the word was correctly said and where an error occurred, strong enough to be noted. This would help to tailor the lesson to the particular weaknesses of each learner, and support them in the practice of those parts they wish to improve.

However, in our case we wouldn't want the system to reject an utterance as "wrong", just because it is not a valid word in the vocabulary. We neither want it to find something that sounds similar (its best match). The task is to check every utterance phoneme by phoneme and compare what was said and what should have been said. The part of what should have been said is easy to know, since the system manages the lesson being practiced. The problem of detecting what was said needs a recognizer that was trained to detect all kinds of sounds (phonetic contexts) and transcribes them correctly without searching the full words or phrases out of a vocabulary. If the utterance contains pronunciation errors, most probably the word won't be in the vocabulary, and we don't want it to be adapted to the existing vocabulary nor rejected as "out of the vocabulary".

4.2 Using the Speech Recognition Approach

In a first attempt we experimented with a neural network trained for Mexican Spanish speech recognition and defined a vocabulary consisting only of the letters of the alphabet and their phonetic transcriptions. The grammar allows any combination of these letters. For this approach the recognizer had to be extremely accurate. It should preferably be speaker dependent and tested in a noise free environment, avoiding even breathing noise into the microphone. Otherwise it would insert a large quantity of garbage between the correctly recognized phonemes.

A way to solve this would be to restrict the grammar to only the words necessary for the application, but this would cause the system to reject the wrongly pronounced words. Another option would be that the vocabulary includes an extensive list with every word and all it's possible mispronunciations. This increases the search space too much, causing again an increased error rate, particularly a high number of elimination errors.

To solve the problem we adopted some ideas from a technique we use for labeling speech corpora automatically. This technique is known as forced alignment [Young et al.,97] [Olivier,99] [Olivier&Kirschning, 99] [Hosom, 99]. This technique aligns a soundwave to its textual transcription, marking the beginning and ending of each word and/or phoneme within the soundwave. In order to do so, forced alignment requires a previously trained neural network (or another recognizer), the soundfile, it's textual transcription and a vocabulary. It then takes the transcription, creates a specific grammar fitted only to the text being aligned, and then processes the sound, frame by frame, classifying and mapping them to the small specific grammar, annotating the position in milliseconds where each phoneme starts and ends.

The set of phonemes used in Mexican Spanish is practically a subset of the phonemes used in English; therefore, a recognizer trained for Mexican Spanish will not detect phonemes outside this phoneme set. Thus, to be able to detect mispronounced phonemes the recognizer has to be trained all the possible sounds and their contexts.

Using the approach mentioned before, the verification process has to generate first the specific grammar for the expected utterance. This is based on the vocabulary, which contains all the words that can be practiced together with their phonetic transcriptions. However, these transcriptions also need to include all of the other possible pronunciations of every phoneme in the word, correct and incorrect. For example:

Table 1. Example of the transcriptions of some letters (correct and incorrect from the Spanish Language point of view)

Letter	Phonetic Transcription
A	{ a | ay | e | ey | uh} ;
B	{ bc b | B } ;
E	{ e | ey | i } ;
J	{ x | h } ;

The system can take the word that is expected to be uttered by the learner. For example if the user is expected to say "ABEJA" the system first creates a grammar using a transcription chart as in table 1 for the possible pronunciation of the word including the *incorrect ones*. This specific grammar looks like the following example:

ABEJA = {{ a | ay | e | ey | uh}{ bc b | B }{ e | ey | i }{ x | h }{ a | ay | e | ey | uh};

In the next step the neural network analyzes the input speech frame by frame trying to mach them to the specific grammar, recording the best matching sequence. If the user mispronounced the 'j' as an 'h' the network should output the sequence {a bc b e h a}. This result can then be used to compare it to the correct one, which is {a bc b e x a}. Then the system can detect if they differ and pinpoint exactly where they differ.

5 Training and Testing of the Prototype

With the above-mentioned method a prototype was implemented, programming the interface in Jacl [Scriptics] that invokes the recognizer and performs the pronunciation check. The reason for using Jacl, a not yet very robust version of Java, able to call functions written in Tcl, was that it provides the advantages of Java to program visual interfaces easily. It also allows the interaction with the scripts of the CSLU Toolkit, which are written in Tcl and C.

5.1 Verification Tool

The neural network for this experiment was trained with the samples of 400 speakers (8 MHz) plus samples from another corpus of 10 speakers containing the vowels, diphthongs and consonants in English. It is a feedforward-, fully connected neural network with 130 input units, 200 hidden and 366 output units. The network was trained as a general purpose-, context dependent recognizer, using standard backpropagation. It reached an accuracy of 95% after the 30th iteration in the training phase. The vocabulary for this experiment consisted in phonemes only and a grammar that permits any sequence of these phonemes with their various pronunciations. For example, the letter 'o' can have the pronunciation of the 'o' in Spanish or the 'ow' in English (as in "only"). This is expressed in the vocabulary as: o { o | ow } ;

Like this example all the other phonemes were expressed in terms of their possible pronunciations in order to have the recognizer pick the correct one. The main difficulty is to determine the possible pronunciations a user could give each letter, in order to be able to recognize them.

It has been found that the most common pronunciation mistakes a native English-speaker makes when learning Spanish are those sounds that are pronounced in English with an aspiration, like 'p', 't', 'k', 'b', 'd', and 'g'. Also the vowels and the consonants like 'r', 's', 'z', 'y' and the Spanish 'll', are difficult at first, specially when they don't exist in English, like the 'ñ' [Dalbor, 69].

A separate database contains a list of the all the words in Mexican Spanish the student can train with, their translation into English, and its correct pronunciation. It also includes a short text giving a recommendation on the correct pronunciation of each letter in the word. Additionally we recorded a sample pronunciation of each word and an image that illustrates the concept of the word; both identified by the word they are related to.

The system allows the user to navigate through menus to select the topic and the word for practice. It then allows the student to record his/her speech and when he/she chooses, it will verify the pronunciation. This means that the recognizer classifies the sequence of phonemes in the uttered word, and compares the result to the word the user was supposed to say, marking the differences.

Table 2: An axample of the transcriptions and pronunciation mistakes

Expected Word	Transcription	User said
ABEJA	a bc b e x a	a bc b **ey** x a
CUATRO	kc k w a tc t r o	kc k w a tc t r **ow**

In the above example, the first column contains what the user sees on screen and is going to pronounce. The transcription in the second column is the internal representation the recognizer uses as the correct version (there can be more than one). The third column is an example of what a user could have said with some of the common pronunciation mistakes. The correct phonetic transcription and the one obtained from the users utterance are compared phoneme by phoneme trying to detect the existence of an error. The result is stored in a file. It should be noted here that the plosives, like p, t, k, or b, d and g, which are divided into two parts, the closure and then the phoneme, are treated as one phoneme when comparing the pronunciation results, i.e. as a 'k' and not as 'kc k'.

The types of errors that can be encountered are insertion, elimination and substitution errors. Insertion errors occur when the user says more than he should or when an external noise causes the recognizer to insert more information. An elimination error means that a phoneme was skipped by the same reasons. At this point the system handles only the substitution and simple insertion and elimination errors by checking the recognition result and the correct transcription phoneme by phoneme. When a mismatch is encountered it looks forward one phoneme to see if it can recover the sequence. When a point is reached where nothing matches the system asks for the student to try again from the beginning.

To test the method we focused on the development of a system able to help students whose native language is American English in their Mexican Spanish language learning course. The domain for the vocabulary was chosen to be animal names, food and digits.

5.2 Interface Design

This section shows the interface designed to test the verification method. Figure 2 shows two examples of the screens of the prototype. When the user enters the first screen he/she can choose among different topics. For every topic there will appear a list of words that the database contains for practice. When a certain word is chosen the system plays it's pronunciation example and displays the image illustrating the words meaning, at the same time showing it's translation in a separate field. The lower text field presents a recommendation for the pronunciation of the selected word. The left side of figure 2 shows the screen when a user has chosen the topic "ANIMALES" and then the word "ABEJA".

From the "Options" menu above, the user can choose to record a speech sample, where he/she will be prompted with a set of buttons to control when to start and stop the recording, as well as play it.

Using the same menu the user can choose the option of verifying the recorded sample. At that point the system presents the screen shown on the right side of figure 2, where the text field at the bottom now contains the verification results.

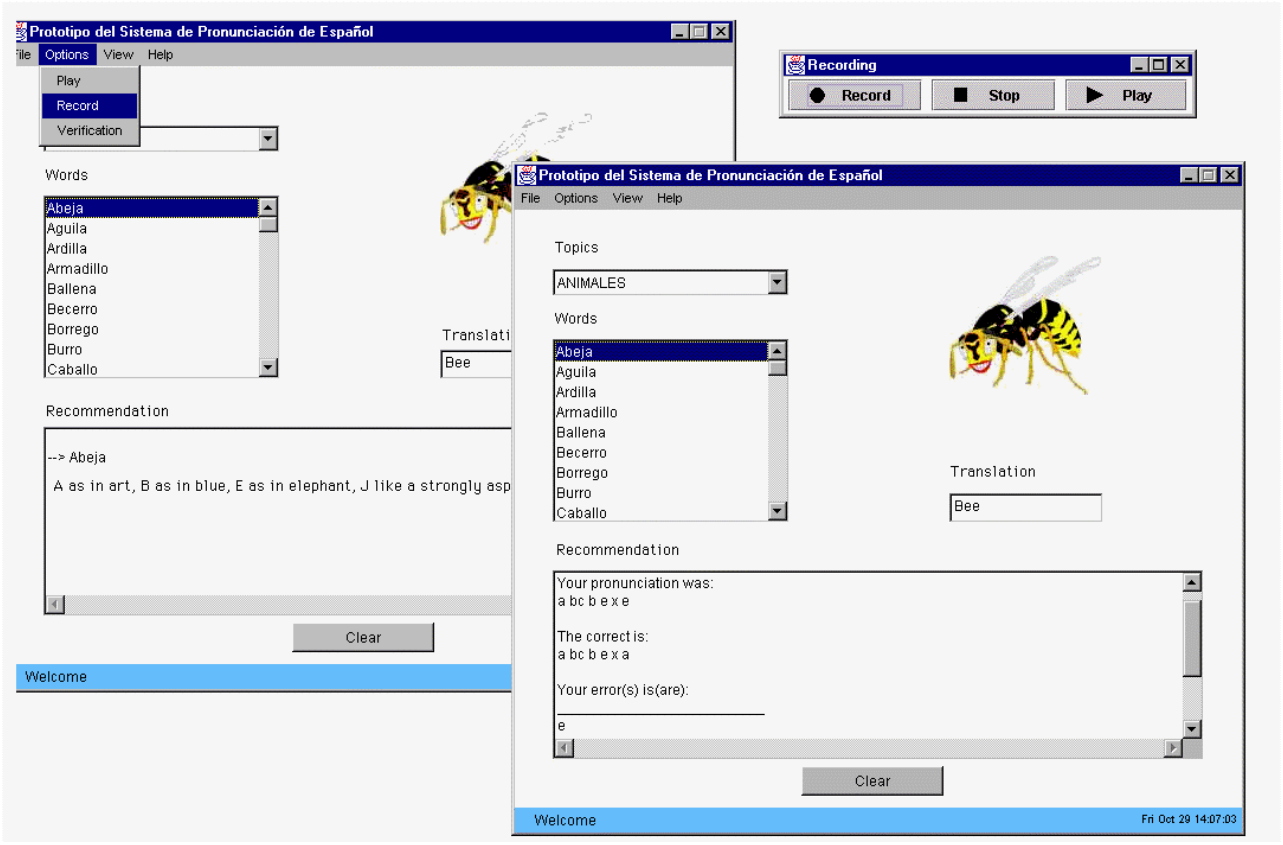

Figure 2: The prototype's screens, showing the information the system can provide to the user

5.3 First Results

The previously explained interface calls the necessary processes to run the neural net-based verification and then analyzes the outcome. At this point the system is able to cope with simple errors, mainly one-character substitution, insertion and elimination errors.

To test the system a group of five foreign students were asked to record each a set of 12 sequences of words. First, these recordings were evaluated by 15 native Spanish speakers, who marked every error at the phoneme-level. Next the system was run to analyze these recordings as well and the differences recorded in a log.

The speech samples recorded by the students were sequences of words belonging to various topics (food, animals, etc.). The recordings contained a total of 256 phonemes each. The human evaluators detected a total of 86 errors (6.72%), whereas the system found only 62 errors. Not all the errors marked by the system matched those of the evaluators. The results of the automatic verification matched only a 98.28% of the human evaluations. There was a 5.16% of mispronounced phonemes not detected by the system. Further 3.44% of the phonemes were marked by the system as mispronounced but not by the human evaluator [Aguas, 99].

6 Conclusions

This paper presented a new approach to verify the pronunciation of spoken language. It uses a neural network based speech recognizer and a vocabulary of the letters of the alphabet, which include all the possible correct and incorrect pronunciations of every letter. The neural network was trained as a general purpose-, context dependent recognizer, with samples from more than 400 speakers and it has an error rate of 5% during the training phase.

So far the results are satisfactory, which means that when intentionally mispronouncing a word in the vocabulary the system detects it. The difference between th human and the automatic evaluation should be treated with care, as this kind of evaluation is highly subjective. However, we discovered that the neural network was not trained with enough erroneous phonemes in all the possible contexts of Mexican Spanish. Thus it is necessary to find more samples and re-train the network.

The evaluation done by native Spanish speakers and the automatic pronunciation check matched a 98.28% of the cases (verified phonemes). These results obtained so far are promising but still need much more testing with a larger number of foreign students and a larger vocabulary.

Some phonemes that were found to be particularly difficult for the system to verify were the /x/, the pronunciation of the letter 'j', and /dZ/ the reading of 'll'. Additionally, not all the possible contexts of the correct and incorrect phonemes were represented in the training set for the neural network. This forced us to eliminate a large portion of the test set.

The presented method offers a way to check and pinpoint exactly the mistakes in pronunciation of a word or phrase. The recognizer is still restricted to fit the recognition to a grammar and a predefined vocabulary, but the method inspired by the one used for forced alignment eliminated the insertion errors in the output. The actual vocabulary of the prototype is small and should be tailored better to the real needs of the Mexican Spanish language learner. However, this first part of the development of a Mexican Spanish language learning environment that uses the CSLU Toolkit shows an interesting potential for automatic pronunciation checking.

References

[Ayala, 99] Ayala, G.: Personal communication (1999)

[Bull, 94] Bull, S.: Student Modelling for Second Language Acquisition,. Computers and Education (1994)

[PronunciationPower] Pronunciation Power, http://www.englishlearning.com/

[Language Connect] Language Connect, http://shop.languageconnect.com/

[CSLU] Center for Spoken Language Understanding, http://cslu.cse.ogi.edu/tm/

[Tucker Maxon] Tucker Maxon Oral School, http://www.oraldeafed.org/schools/tmos/index.html

[Cole et al.,98] Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J., de Villiers, J., Tarachow, A., Massaro, D., Cohen, M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., Soland, C.: Intelligent Animated Agents for Interactive Language Training. STILL:ESCA Workshop on Speech Technology in Language Learning, Stockholm Sweden, (May 1998)

[Cole et al.,99] Cole, R., Massaro, D., de Villiers, J., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, P., Connors, P., Tarachow, A., Solcher, D.:New Tools for Iinteractive speech and language training: Using animated conversational agents in the classroom of profoundly deaf children. Proc.ESCA-Matisse ESCA/Socrates Workshop on Method and Tool Innovation for Speech Science Education, London, UK. (April 1999)

[Black&Taylor,97] Black, A., Taylor, P.:Festival Speech Synthesis System:System documentation (1.1.1.). Human communication Research Centre Technical Report HCRC/TR-83, Edinburgh, (1997)

[Serridge et al.,98] Serridge, B., Cole, R., Barbosa, A., Munive, N., Vargas, A.:Creating a Mexican Spanish version of the CSLU Toolkit. Proc. of the International Conference on Spoken Language Processing, Sydney, Australia (November 1998)

[Young et al.,97] Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P.:HTK Book, (1997) http://ceres.ugr.es/HTKBook/HTKBook.html

[Olivier,99] Olivier, A.:Evaluación de métodos de determinación automática de una transcripción fonética. Undergraduate thesis, Dept. of Computer Systems Engineering, Universidad de las Américas, Puebla, México (May 1999)

[Olivier&Kirschning, 99] Olivier, A., Kirschning, I.:Evaluación de métodos de determinación automática de una transcripción fonética. Proc. of the II. Encuentro Nacional de Computación 1999, ENC'99, Pachuca, Hidalgo, México (September 1999)

[Hosom, 99] Hosom, J.P.: Accurate Determination of Phonetic Boundaries Using Acoustic-Phonetic Infomation in Forced Alignment. Thesis Proposal, CSLU, Oregon Graduate Institute (August 1998)

[Dalbor, 69] Dalbor, J.B.: Spanish Pronunciation:Theory and Practice. Pennsylvania State University, Holt, Rinehart and Wiston, Inc., (1969)

[Scriptics] http://scriptics.com/products

[Aguas, 99] Aguas, N.:Verificación de pronunciación para un ambiente de aprendizaje basado en tecnología de reconocimiento de voz. Undergraduate thesis, Dept. of Computer Systems Engineering, Universidad de las Américas, Puebla, México (December 1999)