CREATING A MEXICAN SPANISH VERSION OF THE CSLU TOOLKIT

Ben Serridge, Alejandro Barbosa, Ron Cole,
Nora Munive and Alcira Vargas

Universidad de las Américas, Puebla, México, and the
Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology

ABSTRACT

The CSLU Toolkit is designed to facilitate the rapid development of spoken dialogue systems for a wide variety of applications, as well as to provide a framework for conducting research in the underlying speech technologies. This paper describes the creation of a Mexican Spanish version of the CSLU Toolkit (both synthesis and recognition) undertaken at the *Universidad de las Américas* in Puebla, México.

Based on the Festival Speech Synthesis System of the University of Edinburgh, we have developed a complete concatenative text-to-speech system for Mexican Spanish, which is currently incorporated into the toolkit and includes both a male and female voice. In the area of recognition, we have created a set of task-specific Spanish recognizers for continuous digits, spelled words, and yes/no phrases, as well as a "general-purpose" phonetic recognizer suitable for arbitrary sub-domains. Using the Rapid Application Developer (RAD) component of the CSLU Toolkit, it is now possible to quickly prototype spoken dialogue systems in Spanish. The Spanish components of the CSLU Toolkit are freely available for non-commercial use from the following web page: http://info.pue.udlap.mx/~sistemas/tlatoa.

1. INTRODUCTION

The CSLU toolkit is designed to facilitate the rapid development of spoken dialogue systems for a wide variety of applications, as well as to provide a framework for conducting research in the underlying speech technologies [1, 2]. Freely available for non-commercial use, the CSLU Toolkit enables research groups with modest budgets to quickly begin conducting research and developing applications. This arrangement is particularly pertinent in Latin America, where the financial support and experience otherwise necessary to support such research is not readily available.

This paper describes the creation of a Mexican Spanish version of the CSLU Toolkit, undertaken at the *Universidad de las Américas* in Puebla, México. The first part of the paper describes the creation of two voices (one male and one female) for Spanish synthesis. The development of each new synthetic voice involved several steps. First, a corpus of diphone units was designed, recorded, and labeled by hand. Next, letter-to-phoneme, syllabification, and accentuation rules were created for the new language. Finally, the existing pause-prediction, duration, and intonation modules of Festival were modified to account for Mexican Spanish prosody.

The second part of the paper describes the creation of a task-specific recognizer for continuous digit strings. The digit recognizer, although trained on microphone speech, achieved a word error rate of 3.07 percent on read telephone speech. Finally, we describe the creation of a general-purpose phonetic recognizer and, for comparison, provide results obtained by this recognizer on the same test set of read digits.

2. SYNTHESIS

Text-to-speech in the CSLU Toolkit is based on the Festival Speech Synthesis System of the University of Edinburgh [3]. In Festival, synthesis is based on the concatenation of units selected from a pre-existing corpus; thus the creation of such a corpus was the first step in adding a voice for a new language.

2.1. Creating a Corpus for Speech Synthesis

The first step in designing a corpus for synthesis was to define the phonemic units to be used for concatenation. In Spanish the set of phonemes is fairly small (relative to English) and well-defined. Table 1 lists the Worldbet symbols for the phonemes used for synthesis.

Worldbet Symbol	Word	Worldbet Symbol	Word
p	**p**unto	n~	ba**ñ**o
b	**b**aile	N	man**g**o
t	**t**ino	l	**l**ago
d	**d**iga	r(	pe**r**o
k	**c**asa	r	pe**rr**o
g	**g**ato	w	**hu**eso
f	**f**alda	j	ma**y**o
s	ca**s**a	i	p**i**so
x	**j**ota	e	m**e**sa
tS	**ch**ato	a	c**a**so
dZ	**ll**anta	o	m**o**do
m	**m**ano	u	c**u**ra
n	**n**ada		

Table 1: Worldbet [4] symbols and examples for the phonemes defined for synthesis.

The next step was to record one example each of all the possible diphones in the language. (In this context, a diphone is a sequence of two successive phonemes.) An attempt was made to record each diphone within a similar, neutral context. For example, the diphones "a-l" and "a-m" were recorded within the nonsense words "atala" and "atama," respectively.

The recordings were made in a professional recording studio in Portland, Oregon. In addition to the voice signal, a simultaneous recording was also made of each speaker's fundamental frequency, using a laryngograph.

Once recorded, the resulting corpus was then segmented and phonetically labeled by hand using the labeling tool from the CSLU Toolkit. Since a diphone is defined as beginning in the center of the first phone and terminating in the center of the second, the labeling process, somewhat unconventionally, consisted in placing boundaries in the *middle* of phones.

2.2. Letter-to-Phoneme, Syllabification, and Accentuation Rules

Spanish orthography is very regular, such that letter-to-phoneme rules are fairly straightforward and accurate. The dictionary of exceptions, compared to that in English, is relatively small.

Once translated into phonemes, each word is divided into syllables according to the following algorithm [5]. Traversing from right to left, each phoneme is checked against its neighbor to the left. A fixed lookup-table (Table 2) determines whether or not the left phoneme can belong to the same syllable as the current phoneme. If so, or if the current syllable contains only a single consonant, the phoneme to the left is added to the current syllable and the algorithm proceeds to the left. Otherwise, a syllable boundary is placed between the two phonemes and the left phoneme becomes the seed of a new syllable.

Phoneme	Allowable Left Neighbors
a, a*, e, e*	i u o b tS dZ d f g j k l m n N n~ p r(r s t w x
o, o*	i u b tS dZ d f g j k l m n N n~ p r(r s t w x
i, u	a a* e e* o o* i i* u u* b tS dZ d f g j k l m n N n~ p r(r s t w x
i*, u*	b tS dZ d f g j k l m n N n~ p r(r s t w x
w	b tS dZ d f g j k l m n N n~ p r(r s t x
l	b d f g k p t
r(	b d f g k p t
other	none

Table 2: Lookup table used for syllabification. Vowels tagged with an asterisk (*) are produced by the text-to-phoneme rules in the case of an explicit accent mark.

If the stressed syllable is not explicitly marked in the orthography, the following standard accentuation rules apply:

1. If there is only one syllable in the word, it remains unstressed.
2. Otherwise, if the last syllable ends in /n/, /s/, or a vowel, the penultimate syllable is stressed.
3. In all other cases, the last syllable of the word carries the stress.

2.3. Duration and Intonation Modules

The duration of a given phoneme is predicted heuristically based on syllabification information. The basic assumption is that phonemes are shortened within long syllables and lengthened within short ones, thus maintaining a relatively constant duration for each syllable [6]. Table 3 lists multiplication factors applied to the average phoneme durations depending on the number of phonemes in the syllable.

# Phonemes	Duration Factor
1	1.12
2	0.82
3	0.68
4	0.66
5	0.54
6 or more	0.50

Table 3: Multiplication factors for phoneme durations, according to the number of phonemes in the syllable.

Intonation models are applied at both the syllable and phrase level. At the syllable level, the intonation is defined by a fixed F0 curve which depends on whether the syllable is stressed or not (see Table 4).

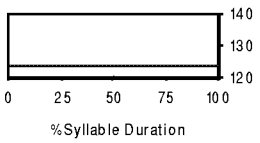
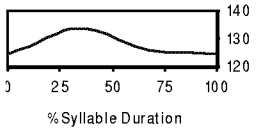
Type of Syllable	F0 Contour	Example
unstressed		ter**mi**nó
stressed		reu**nión**

Table 4: F0 contours for stressed and unstressed syllables.

At the phrase level, the intonation is defined by a similar curve which is multiplied by the F0 contour resulting from the previous analysis applied at the syllable level. Table 5 illustrates the phrase-level contours for each of four predefined phrase types.

3. RECOGNITION

As is the case with most modern speech recognition systems, the architecture of the CSLU Toolkit is language-independent. Nevertheless, in developing recognition for a new language, a certain amount of bootstrapping effort is inevitable. The following sections describe the development of a continuous digit recognizer, from the definition, recording, and labeling of the corpus through the training of the neural net recognizers. We then present some results of experiments on an additional test corpus of digits read over the telephone. Finally, we describe the creation of a general-purpose phonetic recognizer and present results of that recognizer on the same test data.

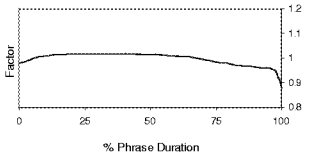
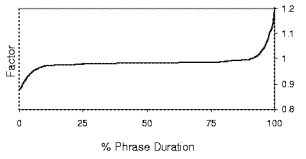
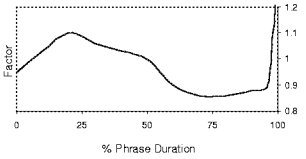
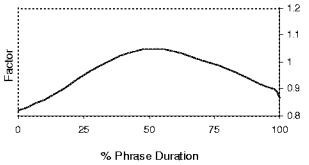
Type of Phrase	F0 Contour
default	
comma	
question	
exclamation	

Table 5: Relative F0 contours for four types of phrases.

3.1. Corpus Description

The initial corpus consisted of 50 speakers, 25 male and 25 female, each of whom recorded the same 40 sequences of six digits. The sequences were designed such that each of the possible across-word phoneme combinations occurs with approximately the same frequency. Recordings were made using a close-talking microphone connected to a PC and sampled at 8000 Hz.

3.2. Training and Testing

The corpus was divided (randomly) into training, development, and test sets of 30, 10, and 10 speakers, respectively. Of the training set, the first 10 speakers were phonetically labeled by hand, while the remainder were labeled by forced-alignment, using the neural-network recognizer trained from the first 10 speakers. A new recognizer was then trained using all 30 speakers of the training set. Using this new recognizer, the data in the training set, including the first 10 speakers originally labeled by hand, were re-labeled using forced-alignment. Finally, a third-generation recognizer was trained from these new labels and evaluated on the test set, the results of which are presented in Table 6.

Later, a separate test corpus of 20 speakers was recorded by telephone. Each speaker recorded 20 random sequences of digits, ranging in length from one to nine digits. (In the original test set, the digit sequences were the same as those pronounced by speakers in the training set.) With the added difficulty introduced by the mismatch between training and testing conditions, it was felt that these new data would more fairly reflect the performance of the recognizer in real systems. The same recognizer described in the previous paragraph was also tested on the telephone data, and the results are presented in the third row of Table 6.

Data Set	% Error	% Sub	% Ins	% Del
dev	0.33	0.05	0.05	0.24
test	0.76	0.09	0.27	0.40
telephone	3.07	0.59	2.24	0.24

Table 6: Recognition results for the digits recognizer.

3.3. General-Purpose Recognition

In addition to the digits recognizer, task-specific recognizers for spelled words and yes/no phrases have also been developed, following similar procedures. However, in order to be able to build arbitrary applications, a general-purpose phonetic recognizer is essential. (The recognizer is general-purpose in the sense that the phonetic training data are not drawn from any particular sub-domain. Combined with particular domain-specific vocabularies and grammars, the general-purpose recognizer can be used to create arbitrary task-specific sub-dialogues without the need for re-training.)

A simple, context-independent general-purpose recognizer was trained from Spanish data in CSLU's Multi-Language Telephone Speech (MLTS) corpus [7]. The corpus consists of 81 one-minute recordings of unconstrained spontaneous speech, phonetically labeled by hand at the Center for Spoken Language Understanding. The callers are U.S. residents who are native speakers of Spanish, from both Latin America and Spain.

In addition, a separate general-purpose recognizer was created by combining the English recognizer included in the toolkit with rules that map each Spanish phoneme to its closest English counterpart. Both recognizers were evaluated on the test corpus of telephone digits, and the results are presented below in Table 7.

The English recognizer outperformed the Spanish recognizer for several reasons. First, the English recognizer is context-dependent and trained on a much larger quantity of speech. More importantly, perhaps, and independent of language, the English recognizer has been trained on a much larger quantity of non-speech sounds, including "silence," breath noise, background noises, etc., which the Spanish recognizer mistook for speech in many cases. With the availability of larger corpora of Spanish data, the general-purpose Spanish recognizer will become context-dependent and its performance should improve significantly.

Recognizer	% Error	% Sub	% Ins	% Del
MLTS	22.30	7.67	14.58	0.05
English	18.30	13.20	4.58	0.53

Table 7: Recognition results for the general-purpose recognizers on read telephone speech (digit strings).

4. DISCUSSION

In this paper we described the creation of a Spanish text-to-speech system, as well as task-specific and general-purpose recognizers for Mexican Spanish. By combining these components, it is now possible to quickly prototype spoken dialogue systems in Spanish using the CSLU Toolkit.

In addition to the development of new systems, the CSLU Toolkit also serves as a platform for research, supporting several projects at both the undergraduate and graduate level. Current research is focused primarily on the recording of larger corpora of telephone speech, both in order to improve recognition performance and to study the phonology of Mexican Spanish. In the area of synthesis, a prosodic corpus of professionally spoken speech is being developed in order to create more sophisticated duration and intonation models for Mexican Spanish. Meanwhile, continuing efforts are being made to further improve the recognition performance of both the task-specific and the general-purpose recognizers. The availability of larger quantities of labeled corpora should permit the creation of context-dependent general-purpose recognizers with much improved performance.

More information about the speech group at the Universidad de las Américas can be found at the following web site: http://info.pue.udlap.mx/~sistemas/tlatoa/.

5. ACKNOWLEDGEMENTS

The project described in this paper would not have been possible without the mentoring provided by the people of CSLU. Alejandro Barbosa participated in a short course in text-to-speech taught by Alan Black and Mike Macon, and thereafter spent an additional week at OGI with Helen van Scoy (the female voice) recording the corpus of diphones and developing the TtS system for Spanish. Nora Munive and Alcira Vargas also visited OGI and both benefited from the help of Andrew Cronk and Ed Kaiser. The technical assistance provided by Stephen Sutton, Jacques de Villiers, and Johan Schalkwyk has been very helpful. Finally, we would especially like to thank Ron Cole, whose guidance and inspiration since the inception of this project have been invaluable.

6. REFERENCES

1. Sutton, S., Novick, D. G., Cole, R., and Fanty, M., "Building 10,000 spoken-dialogue systems." Proceedings of the International Conference on Spoken Language Processing, Philadelphia, PA, Oct., 1996.

2. Schalkwyk, J., de Villiers, J., van Vuuren, S., and Vermeulen, P., "CSLUsh: An Extendible Research Environment", EUROSPEECH, 1997.

3. Black, A. *The Festival Text to Speech System*, system documentation. Technical Report HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, UK, Jan., 1997.

4. Hieronymous, J. L. "ASCII Phonetic Symbols for the World's Languages: Worldbet." Technical report, Bell Labs, 1993.

5. Barbosa, A. *Desarrollo de una nueva voz en Español de México para el Sistema de Texto a Voz Festival*.Master's Thesis, Universidad de las Américas-Puebla, Dec., 1997.

6. Barrutia, R. Fonética y fonología españolas. Ed. John Wiley & Sons, Inc., 1982.

7. Muthusamy, Y. K., Cole, R. A., and Oshika, B. T. "The OGI multi-language telephone speech corpus," Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, Oct., 1992.