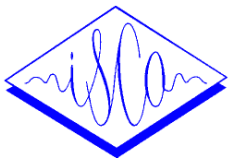

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

REAL-TIME, SPEAKER-INDEPENDENT, CONTINUOUS SPANISH SPEECH RECOGNITION FOR PERSONAL COMPUTER DESKTOP COMMAND & CONTROL

Kamil A. Grajski and Kurt Rodarmer
Apple Computer, Inc.
One Infinite Loop
Cupertino, CA 95014

ABSTRACT

We analyzed a speaker-independent, continuous Spanish speech recognition system for desktop command & control of a Macintosh personal computer. A word error rate of 0.6% was obtained for the System 7.1 Finder navigation task with "clean" speech recorded under semi-controlled conditions in Cupertino, California, and 2.8% for "real world" speech recorded in an open office environment in Mexico City. In obtaining these results, we demonstrated the utility of cepstral normalization as a means of pooling data across speech data collection sites. Finally, data was obtained showing that for the measured tasks, acoustic, phonetic and linguistic data could be leveraged across Spanish dialects.

I. INTRODUCTION

In 1993, Apple began shipping a line of speech products named PlainTalk™, which forms the basis for spoken language interaction with the Macintosh and PowerMacintosh lines of personal computers. This paper presents laboratory and initial real-world testing of a version of the PlainTalk™ speech recognition subsystem localized for the Spanish language.

Initial laboratory studies were performed on a 150-speaker speech database recorded in Cupertino, California. A 50-speaker speech database, recorded in Mexico City at the Universidad Nacional Autonoma de Mexico (UNAM) under the direction of Dr. Enrique Daltabuit, provided the basis for initial real-world performance testing.

Best performance on the real-world training set was obtained by combining Cupertino and UNAM data to generate a new recognizer. Consistent with previous experience, real-world word error rates increased between two- and four-fold over clean, laboratory speech.

II. METHODS

Data Collection

Speech data collection was carried out in Cupertino, California and in Mexico City at the UNAM. The emphasis was on obtaining Mexican, Latin and South American Spanish dialects. Stereo recordings were made using several samples of the PlainTalk™ Desktop microphone and a headset mounted close-talking microphone. Each speaker read up to 300 short phrases and sentences selected at random from a large text corpus. Script material was partitioned into several categories consistent with specific recognition tasks, e.g., Macintosh Finder. Approximately 30% of sentences were from a general-purpose Spanish text corpus designed to give broad phonetic coverage. All script material and read speech were verified by a native Spanish speaker. All speech donors were screened to ensure strong Spanish written and verbal abilities.

The Cupertino recording site was acoustically "clean", but not sterile, e.g., thick door, carpeting, walls. The UNAM recordings were "real-world" in that they were obtained with a recording station setup in an open environment used by a departmental secretary. Table I lists the distribution of Cupertino and UNAM data into independent training and testing sets.

Data	Speakers	Sentences
Cupertino Train	113	34K
Cupertino Test	35	7.8K
UNAM Train	20	3.6K
UNAM Test	29	4.6K

Table 1. Spanish speech data available for training and testing.

Speech Processing

Figure 1 gives an overview of the PlainTalk™ speech recognition architecture [1,4]. For the experiments reported in this study, all word pronunciations were obtained by orthography-to-sound rules developed in collaboration with a linguist and phonetician.

Sound input processing followed standard methods for short-time cepstral analysis and multiple-vector quantization. Acoustic modeling was achieved using phonetic, discrete hidden Markov

models, trained using standard Baum-Welch training, and decision tree methods for allophonic clustering [2]. A beam-search Viterbi recognition method was driven by recursive finite state grammars developed for several specific tasks [3].

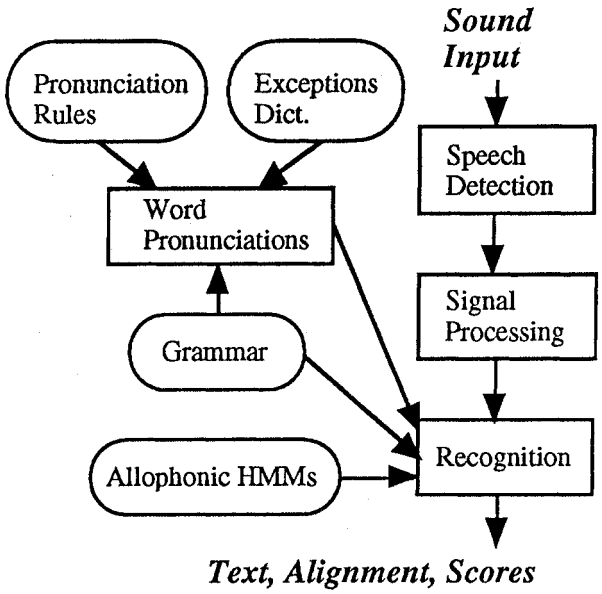

Figure 1. PlainTalk™ system architecture.

Experiments discussed below test operation of the Signal Processing and Recognition modules, and bypass the Speech Detection module, as all data was presegmented at the sentence level at the time of speech data collection.

III. RESULTS

Clean Speech

The Cupertino clean speech training set consisted of 113 speakers and 34K sentences. The training data was recorded with the PlainTalk™ Desktop microphone and contained speakers of both genders; country of origin was heterogenous. An independent testing set consisted of 35 speakers and 7.8K sentences divided into five tasks. Table 2 lists the vocabulary size, grammar perplexity, number of sentences and percentage word error rate for the Spanish 7.1 Macintosh Finder, Numeric, Calendar, ClarisWorks and FileMakerPro tasks. For the Numeric task, the word error rate was relatively high; approximately 1/3 of errors were due to function word deletions. For the remaining tasks, word error rates were comparable to their English equivalents [4].

The Spanish language is spoken by more than 300 million people in more than fifteen countries. *How big a job is it, therefore, to localize PlainTalk to Spanish?* Table 3 lists word error rates for the Macintosh Finder task by country of origin for speakers in the Cupertino testing set. The speaker counts were low to moderate, and due mainly to logistical constraints, it was not possible to systematically control speakers' language ability, length of residence in country of origin, USA, etc. Taken together, however, these error rates suggested that, as evaluated for personal desktop command & control, acoustic, phonetic and linguistic data could be shared across Spanish dialects. Further support was obtained by analyzing a real-world speech testing dataset recorded in Mexico City.

Task	Voc.	Pplx.	N	Wd Err
Macintosh Finder	150	28	2961	0.6%
Numeric	78	38	591	6.6%
Calendar	48	24	576	2.0%
ClarisWork	655	15	1876	1.4%
FileMaker	510	32	1871	1.8%

Table 2. Word error rates (%) for training and testing with Cupertino speech data.

Country	Spk	Sent	Wd Er
Argentina	3	260	1.5%
Bolivia	2	170	0.2%
Colombia	3	256	0.3%
El Salvador	2	173	0.2%
Mexico	13	1091	0.4%
Peru	5	444	0.8%
Puerto Rico	3	244	0.3%
Venezuela	4	323	0.7%

Table 3. Word error rates (%) as a function of country of origin.

Real-world Speech

There was a risk factor in approaching PlainTalk Spanish localization using clean speech recorded in Cupertino, California. Namely, *how would the acoustic models handle real world data from target countries?* To address this key question, we undertook recording of an independent testing set at the UNAM in Mexico City. Table 4 lists word error

rates for the UNAM testing set using speech recorded with the PlainTalk™ Desktop microphone. The same five tasks were used as for the Cupertino testing set. Speaker count was 29 and the number of sentences in each task are shown in the second column. The column marked "Raw" lists word error rates obtained when the clean speech recognizer was used - without cepstral normalization. Error rates were unacceptably high. Differences in recording site, ambient noise, and speaking style were clearly evident in the database. The column marked "ENV" lists word error rates obtained when the UNAM data was subjected to a signal-dependent cepstral normalization procedure [5] - on a speaker-dependent basis. Error rates were significantly reduced.

Task	N	Raw	ENV	TRAIN
Macintosh Finder	2K	17.4	3.6	2.8%
Numeric	397	29.0	10.1	10.1%
Calendar	423	13.0	6.9	5.4%
ClarisWork	1K	20.4	4.3	3.8%
FileMaker	1K	21.8	5.3	4.1%

Table 4. Word error rates (%) for testing data collected at the UNAM in Mexico City and tested with Cupertino data trained recognizer (Raw) without cepstral normalization, (ENV) with normalization and combined Cupertino/UNAM trained recognizer (TRAIN).

Finally, the column marked "TRAIN" in Table 4 lists word error rates obtained using a newly trained recognizer. The recognizer was trained on a combined Cupertino-UNAM database. UNAM training data was subjected to the speaker-dependent signal-dependent cepstral normalization procedure, as above. With the addition of only these 20 UNAM speakers, word error rates were further reduced. Table 5 lists results of the control experiment of re-recognizing Cupertino testing data with the new recognizer.

Table 6 lists the distribution of word error rates by speaker for the "ENV" and "TRAIN" experiments. We anticipate further improvements with additional UNAM training data. Nevertheless, at these levels, the real-world data is within range of our previous experience that error rates increase between two- and four-fold over clean speech.

Task	CUP Only	CUP + UNAM
Macintosh Finder	0.6%	0.6%
Numeric	6.6%	5.9%
Calendar	2.0%	2.3%
ClarisWorks	1.4%	1.5%
FileMaker	1.8%	1.8%

Table 5. Hybrid Cupertino/UNAM trained recognizer control. Word error rates Cupertino data not significantly affected. (See Table 2, also.)

UNAM Test Data Finder Task	% Spkrs ENV	% Spkrs TRAIN
0 - 1%	6.8	10.3
1 - 2%	10.3	10.3
2 - 3%	0	20.7
3 - 4%	17.2	3.4
5 - 6%	6.8	13.8
6 - 7%	3.4	13.8
> 7%	55.2	27.6

Table 6. Distribution of word error rates across speakers (% of 29 UNAM test speakers) for Macintosh Finder task. Combined environmental correction and training (TRAIN) gave further improvement over environmental correction alone (ENV). Results are expected to improve even further with additional UNAM training data.

IV. DISCUSSION

Summary

This study analyzed the performance of a PlainTalk™ speech recognition subsystem localized for the Spanish language. Both "clean" (Cupertino) and "real-world" (UNAM) data were analyzed. Overall performance was consistent with the shipped English PlainTalk™ line of speech products. Using cepstral normalization as a means of pooling data across multiple recording sites, we were able to significantly reduce the UNAM word error rates. Last, comparable word error rates across speakers' country of origin suggests that universal coverage of

the Spanish language for personal computer desktop command & control can be obtained at low cost.

Sources of Error

Potential sources of error arose in several areas. First, the word pronunciation module was rather simplistic. No special provisions were made for accented vowels, for instance. A more sophisticated orthography-to-sound-rule system could relieve this source of error. Similarly, there were pronunciation effects due to the bilingual speaker population. Particularly in the case of technical words borrowed from the English language, some users used "proper" English pronunciations, while others applied Spanish phonetic rules for pronunciations, e.g., "MULTIFINDER." A multiple-pronunciation dictionary could address this source of error.

Second, acoustic models were derived largely from a mixed population of bilingual speakers. Of the Cupertino-recorded speakers, seventy percent of speakers resided full-time in the Silicon Valley. The other 30% of speakers were individuals affiliated with Apple, but who resided full-time in Mexico, or any of several Latin and South American countries. From a scientific point of view, such a database complicates interpretation of various results, e.g., the country-specific results listed in Table 3. For instance, there was not enough data to complete several obvious control experiments, e.g., compare Mexican test speaker performance with a Mexican-only trained recognizer versus mixed population-trained recognizer. On the other hand, the training database reflected the likely Spanish PlainTalk™ target population and by introducing this heterogeneity at the training stage, we made maximum use of the commonality across populations, at the expense (still unknown) of modelling detailed differences.

Last, as shown in Table 6, for some 25% of speakers in the real-world database, word error rates remained high even after combined cepstral normalization and across-site training. Detailed analysis revealed few general sources of error, however, the most common explanation was that non-stationary environmental noise, simultaneous background speech, music, etc., contributed to increased word error rates. Some pronunciation effects were also noted (see above).

Implications for Desktop User-Interface

Additional spoken language capability on the personal computer introduces new challenges for the desktop user-interface. For instance, it is not uncommon to find desktops with several applications running at once, but in different languages, e.g., Spanish ClarisWorks simultaneously with English QuickMail™. How best to manage system resources to provide simultaneous, real-time, multi-language spoken input? These issues could be addressed, in part, by sensing and dynamically responding to the context of a given desktop environment. Additional challenges occur in truly multilingual contexts, e.g., where the language model or grammar itself is multi-lingual. How to deliver spoken language interfaces for the bilingual, language learning and literacy environments?

Ultimately, all of these issues must be addressed and incorporated in the underlying system software in a form easy to use by developers in creating a new generation of spoken language processing applications.

BIBLIOGRAPHY

[1] Lee, K.F., Hon, H.W., Reddy, R. "An Overview of the Sphinx Speech Recognition System", *IEEE Trans ASSP,* January, 1990.

[2] Hon, H.W. and Lee, K.F. "CMU Robust Vocabulary-Independent Speech Recognition System", *Proc ICASSP,* April, 1992.

[3] Schwartz, R., et al., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *Proc ICASSP,* April, 1985.

[4] Lee, K.F., "The Conversational Computer: An Apple Perspective", *Proc EuroSpeech,* September, 1993.

[5] Liu, F., Acero, A., Stern, R. "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering", *Proc ICASSP,* March, 1992.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the following individuals' support and contributions to this study. In Mexico City: E. Daltabuit, M. Polo. In Cupertino: S. Austin, G. Beauregard, Y. Chow, J. Colosso, A. Fineberg, C. Henton, P. Hernandez-Ramos, H.W. Hon., K.F. Lee, S. Meredith, M. Pallakoff, A. Reeves. We especially thank M. Rodarmer and A. Snodgrass for their technical support in all aspects of data collection.