MULTILINGUAL SPEECH RECOGNITION: THE 1996 BYBLOS CALLHOME SYSTEM

Jayadev Billa[1,2] *Kristine Ma*[1] *John W. McDonough*[1] *George Zavaliagkos*[1] *David R. Miller*[1]
Kenneth N. Ross[1] *Amro El-Jaroudi*[1,2]

[1] BBN Systems and Technologies, Cambridge MA 02138. USA
[2]University of Pittsburgh, Pittsburgh PA 15261. USA

ABSTRACT

This paper describes the 1996 Byblos Callhome speech recognition system for Spanish and Egyptian Colloquial Arabic. The system uses a combination of Phoneticly Tied-Mixture Gaussian HMMs and State-Clustered Tied-Mixture Gaussian HMMs in a multiple pass decoder. We focus here on the aspects of the system which are language specific and demonstrate the adaptability of the Byblos English system to new languages. Language related issues arising from both dialectal differences as well as differences between transcribed and spoken language are discussed. This system gave the lowest error rates in both Egyptian Colloquial Arabic and Spanish in the October 1996 NIST Callhome evaluation.

1. INTRODUCTION

This paper describes the 1996 Byblos Callhome speech recognition system for Spanish and Egyptian Colloquial Arabic (ECA). Byblos is the BBN large vocabulary speech recognizer. Full details of the current system are given in [1, 2, 3].

Callhome corpora consist of about 20 hours of spontaneous natural speech over long distance telephone lines mostly from and to family and friends. This results in more difficult corpora than say, WSJ and Switchboard, with higher perplexities and OOV rates. The 1996 Byblos Callhome speech recognition system focussed on improving both core automatic speech recognition performance and major language specific issues such as, dialectal issues (notably that of /s/ deletion), multiple-corpora language modeling in Spanish and, transcription issues arising from differences between transcribed and spoken ECA.

This paper first gives an overview of the primary Byblos Callhome speech recognition system and then proceeds to describe in detail the language specific systems with appropriate results. Finally, in the last section, we provide results for both Arabic and Spanish systems in the October 1996 NIST Callhome evaluation.

2. GENERAL SYSTEM DESCRIPTION

This section describes the core Byblos system on which the Callhome system was developed.

2.1. Frontend feature extraction

The Byblos front end analysis generates base features consisting of 14 mel-warped cepstral coefficients and normalized energy. The cepstra are calculated from a mel-warped 125-3750Hz bandlimited spectrum of a 20ms windowed speech frame at 100 frames/sec. Vocal tract length normalization (VTLN) is performed to reduce speaker variability due to differing vocal tract length [4]. The current VTLN implementation involves stretching the spectra by the ratio of the median third formant of each speaker to the median third formant of all speakers prior to computation of cepstra. Causal cepstral means removal is performed on the base cepstral coefficients by subtracting a decaying local causal mean from each cepstral frame. Normalized energy is calculated as a peak following decay of frame energy. The base coefficients along with the corresponding first and second order difference features forms the input features to the system.

2.2. Models and Training

The basic Byblos acoustic models are context-dependent phoneme HMMs. These HMMs are 5-state with minimum 3 frames duration, diagonal covariance, continuous density Gaussian mixture models.

There are two HMM configurations in use in Byblos: Phoneticly Tied-Mixture (PTM) and State-Clustered Tied-Mixture (SCTM). PTM models are coarser models of speech and serve to provide a lattice for subsequent decoding by the finer SCTM models [1, 2, 3]. The resulting PTM system consists of 256 Gaussian components per mixture density for each of the phonemes and of 32 Gaussian components per mixture density for each of 1,000 clusters in the State-Clustered Tied-Mixture (SCTM) system. Clusters are generated by means of top-down, linguistically driven likelihood based decision trees.

HMMs in Byblos are trained with the vocal tract length normalized 45 cepstrum/energy features in a single parameter stream. Prior to training, these features are normalized to unit variance for numerical stability. Training consists of an initial model estimation using k-means on state-aligned speech followed by 3 Expectation-Minimization (EM) passes over the available data [1]. MAP smoothing is performed to eliminate any resulting singularities. Further training by means of Speaker Adapted Training (SAT) [2, 3] generates the SAT models. SAT models are speaker independent (SI) models that are generated by adapting an initial SI model to each speaker using maximum-likelihood linear regression (MLLR) matrices [5] followed by reestimation of the SI model. Two transformations per speaker are used to train these final SAT models. SAT models are typically sharper than the SI models.

2.3. Decoding

The Byblos decoding system employs a multi-pass recognition strategy using progressively more detailed models to reduce candidate word sequences before arriving at the

best word sequence. The first three passes use PTM non-crossword triphone acoustic models with a bigram language model to determine the most likely candidate word beginnings and endings. At the end of this search, surviving words are connected together to form a lattice. SCTM crossword triphone acoustic models with a trigram language model are then used to score this lattice and generate an N-best list of most probable word sequences. The top hypothesis can be chosen as the final recognition result.

In order to further improve performance the PTM and SCTM models can be adapted to yield models that are better suited to the test speakers. Speaker adapted SAT models are generated by MLLR adaptation driven by the transcripts generated in the initial decode. A second multi-pass decode, identical to the initial decode except for the use of speaker adapted SAT models, generates the final result.

3. CALLHOME SYSTEMS

Callhome systems are in general characterized by much higher error rates than systems built for other prevalent corpora. The major differences that cause this degradation are:

- Small size. Callhome corpora size varies from 15-20 hours of recorded and transcribed speech. In comparison, most other corpora, such as Wall Street Journal (WSJ) and Switchboard (SWBD), are composed of about 150 hours of recorded and transcribed speech.
- Language Modeling. Due to the small corpora size the amount of language modeling data is also reduced. For example, Callhome Spanish has about 150k words available for language modeling where as SWBD has several million words available, a ten-fold disparity and WSJ has over 200 million words available.
- High language perplexity. As Callhome corpora are collected from conversations amongst familiar parties, there is significant increase in vocabulary size mostly due to incomplete words, and ill formed sentences and hesitations. Both of these contribute significantly to increased language perplexity.
- Familiarity amongst speakers results in highly spontaneous speech with many references to people and places which in turn increases the out-of-vocabulary (OOV) rates. For comparision, WSJ speech is planned and read, and in SWBD conversations, the concerned parties are unknown to one another resulting in spontaneous yet more careful speech.
- The non-domestic channel typically has higher noise levels as well as higher variations compared to domestic channels.
- Dialect also plays an important part in the overall degradation. For example, Callhome Spanish has several dialects, e.g., Columbian, Mexican etc., resulting in different pronunciations for the same words.

We have explored methods to address the large vocabulary size and high perplexity seen in these systems as well the small data problem. Our base system was primarily one that we had previously used for Switchboard English. It is important and interesting to note the language independence of the underlying system. Our experience here shows that the same core speech recognition techniques can be used in both Spanish and ECA and possibly applied to other languages and that these techniques display much of the same behavior across all languages.

3.1. Spanish and Arabic Corpora

The corpus for Egyptian Colloquial Arabic (ECA) consists of 80 telephone conversations each 10 minutes long, for a total of 14 hours of speech. The speech data was transcribed using roman representations of 39 Arabic phonemes. It is important to note that ECA is not a written language and hence there was no possible transliteration from Arabic script to roman representation. The conversations contained about 20,000 utterances with approximately 16,000 unique words. A separate set of 20 conversations were used for testing during system development.

The Callhome Spanish corpus consists of spontaneous telephone conversations from the United States to family and friends in Spanish speaking countries such as Puerto Rico, Mexico, Chile, Colombia, Spain, etc. Training data has about 18 hours of speech, with more than 10K unique words, from a total of 100 conversations and 252 speakers. Also used in training was 7.5 hours of telephone speech from the Ricardo corpus. Ricardo is an evoked monologues speech corpus collected by Dragon Systems.

The major differences between the Arabic and Spanish corpora can be summarized as:

- For similar size Spanish corpus, Arabic has more unique words that Spanish (16k instead of 10k). This may be due to the fact that in Arabic compounding is a frequently used method to create new words.
- The training set perplexity of the ECA model is twice that of Spanish. The difference can be mostly accounted for by the increased number of unique word tokens in ECA.
- ECA has twice as many foreign words in its lexicon as in Spanish ($\sim$1k instead of $\sim$0.5k).
- ECA acoustic training has twice as many incomplete words as Spanish ($\sim$2k instead of $\sim$1k).
- ECA has 40% more occurrences of background noise and other non-speech artifacts.

3.2. Spanish and Arabic Acoustic training

Acoustic Modeling in the Byblos Callhome System incorporates several improvements to the base system. The improvements are due to the incorporation of PTM-SCTM interpolation and changes to the SAT implementation. Also phoneme loop adaptation was attempted on ECA to provide a comparison with transcription mode adaptation.

For tied mixture systems, good gains have been obtained for small data problems by smoothing the triphone-state model probability distribution functions (pdf) with the pdf of more general models, such as left and right biphones and phoneme models. This kind of smoothing is present in our PTM system but does not generalize to our SCTM system, where each cluster only covers a few triphone-states. In order to apply a similar strategy to the SCTM system we considered a simple smoothing of SCTM densities with context independent PTM phoneme models. Table 1 shows the results using a global interpolating weight.

The incorporation of the Ricardo corpus results in an interesting problem for our SAT paradigm. Ricardo consists of many speakers each with very little data resulting in

	% Word Error Rate
no interpolation	71.8
interpolated model	71.4

Table 1. SCTM-PTM interpolated results on Spanish

numerical problems for the SAT paradigm as well as a disproportionate increase in computation. To address this we attempt to cluster the Ricardo speakers into clusters based on the median third formant. The Callhome speakers are kept as is. The results for this speaker clustering are shown in Table 2. The first two rows of Table 2 show the results

		% Word Error Rate
SI	No adaptation	70.2
	SI adaptation	68.6
SAT	Callhome only	69.1
	CH Spanish + 6 clusters	68.5
	CH Spanish + 10 clusters	67.9

Table 2. SAT on Callhome Spanish with clustered Ricardo speakers

of using only the SI models both with and without adaptation. The next three rows show the results of using the SAT models during the adapted decode. The "Callhome only" row refers to results obtained with SAT models generated only from Callhome Spanish. The last two rows are results with combined Callhome Spanish and Ricardo trained SAT models with differing number of Ricardo speaker clusters.

Adaptation in our Spanish system involves two decoding passes where the first pass yields putative transcriptions (hence *transcription mode adaptation*) which are used to supervise adaptation of the SI models. The second pass then uses these adapted models. Adaptation therefore depends on errorful transcriptions. Another approach using a context independent phoneme loop to align speech may result in better performance. The main idea here being that a labeling error made by the phoneme loop will atleast be with an acoustically similar phoneme, in contrast to a full decoding which may produce strings of totally erroneous labels. Table 3 summarizes our results. These results indicate that more constraints improve performance: a phoneme loop with a unigram is better than one with no grammar, and transcription mode is better than any phoneme loop adaption. However, phoneme loop adaptation has the advantage of speed as there is no need for a full decoding.

		% Word Error Rate
	No adaptation	70.2
Phoneme loop	No grammar	69.6
	phone-1gram	69.3
	phone-2gram	69.4
	transcription mode adaptation	68.6

Table 3. Phoneme loop adaptation results on Spanish

3.3. Arabic and Spanish System Language modeling

The ECA dictionary consists of about 16,000 entries including incomplete words, foreign (to Arabic) words and hesitations. English words were spelled out phonetically using a mapping from English phonemes to Arabic. Incomplete words were also spelled phonetically using an automatic generator we developed.

Specific to ECA were two issues the article "il" and the processing of the "teh marbuta".

In ECA, the article "il" (the) is written attached to the next word. Its pronunciation also depends on the first letter of this next word. In the Callhome ECA corpus, "il" is transcribed with a "+" attaching it to the next word. Consequently, many words appear twice in the dictionary, e.g.,

HAgaB and *il+HAgaB* (*the thing*).

To reduce the vocabulary and hence perplexity, we separated the article from the word and modified its spelling appropriately, e.g.,

il+HAgaB becomes *il HAgaB*

and

il+salAm becomes *is salAm* (*the peace*).

The resulting language model shows a 25% reduction in perplexity with a 7% drop in vocabulary size ($\sim$1000 words). The resulting system shows a 1% absolute reduction in word error rate (WER).

In ECA, the feminine ending can have one of two pronunciations depending on the following word, this is referred to as the "teh marbuta". The corpus was tagged to indicate which pronunciation was actually used, giving us the means to directly incorporate the pronunciation information in the training of the acoustic models. This yielded a net increase of 0.2% absolute in WER. We feel that this is a result of the preexisting allowance for alternate pronunciations in the Byblos system and the ability of the EM algorithm to find which of the two alternatives was used.

The Callhome Spanish corpus consists of about 150k words. In order to improve the language models, 770k words from the Spanish portions of the European Corpus Initiative (ECI) corpus, a newswire text corpus, were added to the language model training. The final language models were created by weighting the Callhome and ECI by 0.9 and 0.1 respectively. We have also attempted a "similarity weighting" to combine these corpora. For similarity weighting, each "article" of the non-corpus data is given a weight proportional to its similarity with the within corpus data. The results of the article weighting were negligible, as illustrated in Table 4.

	% Word Error Rate
Callhome Spanish only	73.8%
merging based on simple global weights	73.2%
merging based on similarity weighting	73.1%

Table 4. Article weighting of ECI for language modeling

The Callhome Spanish consists of a wide collection of Spanish dialect, allowing an opportunity to look into modeling dialect dependent information. Specifically, we experimented with modeling dialect dependent pronunciation of the final /s/ syllable in Spanish words. This is motivated by the fact that in some Spanish dialects, the final /s/ syllable of each word can be selectively deleted or turned into an aspired /h/ sound. To address this we assumed that the final /s/ can be either deleted or aspirated (/s/$\rightsquigarrow$/h/) and decoded with three dictionaries (with /s/ pronunciations, without /s/ pronunciations and with both pronunciations). After the fact selection gives a 0.7% absolute reduction in WER with a few speakers strongly preferring the modified dictionaries. However, automatic selection based on likelihood fails even when true transcriptions are used. One possible reason could be that the /s/ rule is not exact and hence not universally applicable.

Also attempted for Spanish is the inclusion of compound words. Starting from a list of most frequent word pairs, realistic pronunciations incorporating coarticulation effects were added to the dictionary. While this resulted in gains on Switchboard, Spanish failed to see any gains with the improved recognition of compound words being offset by insertions of the same words elsewhere in the test. We feel that as the performance of the core Spanish system improves the inclusion of compound words will display a more pronounced improvement. There is also the possibility of interactions between compound words and dialect specifics. We are working with a native speaker to assess specific dialectic differences for a possible combined approach to dialect and compound words.

4. 1996 CALLHOME EVALUATION RESULTS

Our experience with the Byblos Callhome system suggests that it is strongly language independent, with the basic technology being equally applicable to any language. Tables 5 and 6 presents the evolution of the performance and shows (whenever possible) comparisons between comparable ECA and Spanish systems.

System	Development	Evaluation
Unadapted PTM	78.6%	-
SCTM with phone loop	73.3%	75.6%
+ transcription mode adapt	69.6%	-
+ SAT	-	71.1%

Table 5. 1996 Callhome ECA Results as WER

System	Development	Evaluation
Unadapted PTM	73.2%	-
SCTM with phone loop	69.3%	-
+ transcription mode adapt	68.6%	-
+ SAT	67.9%	65.9%

Table 6. 1996 Callhome Spanish Results as WER

Three observations can be made from the tables above:

- The direction and magnitude of all improvements is very close for the two languages.
- Despite the fact that the size of the ECA corpus is considerably less that Spanish, the performance of both is reasonably close. This seems to be a result of ECA being a single dialect as opposed to Callhome Spanish which consists of many dialects.
- The system is mostly language independent, with any variation being accounted by either the size of corpus or language specific idiosyncrasies.

5. CONCLUSIONS

We have described the development of the 1996 Byblos Callhome speech recognition system for Spanish and Egyptian Colloquial Arabic. Approaches to addressing problems inherent to small corpora as well as approaches to increasing training data by combining corpora are described. Language dependent details addressing language idiosyncrasies were presented with our approach to solving them. The resulting Byblos Callhome system represents the current state-of-art in multilingual speech recognition.

REFERENCES

[1] L. Nguyen, T. Anastasakos, F. Kubala, C. Lapre, J. Makhoul, R. Schwartz, N. Yuan, and G. Zavaliagkos, "The 1994 BBN / BYBLOS speech recognition system", in *Proc. SLS Technology Workshop*. 1994, Morgan Kaufmann Publishers.

[2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training", in *Proceeding ICSLP*, October 1996, vol. 2, pp. 1137–1140.

[3] J. McDonough, T. Anastasakos, G. Zavaliagkos, and H. Gish, "Speaker-adapted training on the switchboard corpus", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, April 1997, IEEE, vol. 2, pp. 1043–1046, IEEE.

[4] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1996, pp. 346–349, IEEE.

[5] V. J. Leggetter and P. C. Woodland, "Speaker adaptation using linear regression", Technical Report CUED / F-INFENG / TR.181, Cambridge University, Engineering Department, Cambridge, England, 1994.