VOCABULARY-INDEPENDENT RECOGNITION OF AMERICAN SPANISH PHRASES AND DIGIT STRINGS

Yeshwant K. Muthusamy and John J. Godfrey

Speech Recognition Branch
Media Technologies Laboratory
Texas Instruments, Dallas, Texas USA.
E-mail: {yeshwant,godfrey}@csc.ti.com

ABSTRACT

We describe the development of an R&D recognizer for several Spanish applications, starting from an existing recognition system for American English and modest language-specific resources. The experiments emphasize achieving phonetic accuracy on telephone speech without vocabulary specific training. We use our basic recognition engine, and simple grammar-building tools for predicting word sequences. Only the read sentences from two telephone speech corpora (Voice Across Hispanic America (VAHA) and a smaller TI corpus) are used for training. Word error rates (WER) of 1.9% on telephone service command phrases, 5.5% on telephone numbers, and 12% on continuously spoken sentences are achieved with the newly ported system.

1. INTRODUCTION

Given the high cost of speech data collection, it is important that vocabulary-independent, speaker-independent recognition systems be developed for use in a variety of task vocabularies, reducing or eliminating the need to collect data for new task domains. This is especially true for telephone speech, and for languages such as Spanish, in which there are few public speech corpora available.

Recent work on vocabulary-independent Spanish speech recognition includes Torres and Casacuberta, [1], Varona *et al* [2] and Bonafonte *et al* [3], who experiment with building effective subword-size phonetic units to cope with context variation. These studies begin with relatively small training corpora of microphone quality (16 kHz) Castilian speech from as few as 10 speakers. They report good phonetic recognition results for test sentences of the same type as training. For telephone quality speech, Villarrubia *et al* [4] report results on vocabulary-independent recognition of Castilian surnames with monophone, biphone and triphone models, using telephone speech from the VESTEL corpus [5].

In this paper, we describe experiments in vocabulary-independent, speaker-independent recognition of American Spanish telephone speech. Using only read sentences, i.e., domain-independent speech, we build and train Spanish acoustic models for recognition of two application vocabularies, telephone service command phrases and connected telephone digit strings. as well as for test sentences from the training domain. We use only phoneme-size units, but test various subtypes of phonetic models, including mixtures and clustered triphones.

2. SPEECH DATA

In American Spanish, we have two telephone speech corpora available: a 408-speaker corpus collected internally at Texas Instruments (the "*TI corpus*"), and the Voice Across Hispanic America (VAHA) corpus [6], sponsored by the Linguistic Data Consortium as part of the POLYPHONE project.

The two corpora are similar in design and content, featuring a combination of telephone service command phrases (from domains such as voice-activated dialing, voice-messaging, etc.), telephone and credit card numbers, spelled words and phonetically rich sentences taken from Spanish newswire texts. Overall, the TI corpus contains 48,824 utterances from 408 speakers (237 female and 171 male), while the VAHA corpus has 38,740 utterances from 915 speakers (570 female and 345 male). For purposes of speech recognition training and testing, the principal difference is that the TI corpus contains about three times as many utterances per speaker as VAHA (138 vs. 45), and also a much greater proportion of telephone-number-style digit strings. The TI corpus has about 60 digit strings and 60 command words and phrases per speaker, where VAHA has 12 each of command phrases and numeric items of all kinds, including money, dates, etc.

These corpora provide just the material needed for our vocabulary- and speaker-independent recognition experiments: training material in the form of read sentences; testing material in the form of digit strings and command phrases; and enough speakers so that test and training voices do not overlap.

2.1. Test Data

For the initial round of testing, we selected 49 speakers from the TI corpus, and from their utterances created three test sets:

- **CP**: 2490 tokens of 60 command phrases,
- **TEL**: 2123 4-digit, 7-digit and 10-digit telephone numbers, and
- **PS**: 826 phonetically rich sentences.

Note that for the **TEL** experiments, only digit strings of "American style" phone numbers were used for testing recognition – i.e., where "3512" was said as *tres cinco uno dos* instead of *treinta y cinco doce*.

Another complete ensemble of three test sets, identical in size and structure, was set aside for future use.

2.2. Training Data

Training was done only on the phonetically rich sentences in the two corpora. In the first series of experiments, we used just the 4993 sentences from the TI corpus, read by 310 speakers, 159 male and 151 female. In the second series, we doubled the amount of training material to 10404 sentences by adding 5411 VAHA sentences from 796 speakers, 303 male and 493 female. None of the 49 test speakers appeared in any of the training data.

3. EXPERIMENTS

3.1. Spanish Phonetic Models

Wheatley et al. [7] showed that effective acoustic models for a new language can be created by cross-language adaptation from existing models of similar phonetic units. They built Japanese phonetic HMMs from pre-existing English models, and compared several methods of doing this. Having both their Japanese and English models available, we thought the adaptation would be simpler from the Japanese models, because of greater isomorphism with the phoneme inventory of American Spanish. In fact, parametric experiments revealed that Spanish phonetic models bootstrapped from trained Japanese telephone speech phonetic models performed better than those derived from English. The bootstrapping was done using cross-language adaptation techniques described in [7]. The baseline model set had 46 (23 male and 23 female) context-independent (CI) finite-duration Spanish phonetic models. We also used two gender-specific silence models to account for silence and background noise. Table 1 shows the 23 Spanish phones, using Worldbet symbols [8]. /r/ refers to the trilled or double 'r', while /r(/ is the tapped single 'r'.

Table 1. Spanish Phone Set

Phone Type	*Phone List*
Vowels	a e i o u
Stops	b d g p t k
Fricatives	f s x
Affricates	tS
Nasals	m n *n* ~
Semi-vowels	l j r r(w

The baseline acoustic models were continuous density, single-mixture, gender-specific Gaussians. The acoustic features were 16 principal components of 34 LPC-based filterbank and delta-filterbank parameters.

3.2. Training

The Spanish phonetic models were trained using a Viterbi alignment algorithm. The pronunciations were derived using Spanish letter-to-sound rules. The top-level grammars used for recognition depended on the test set vocabulary; all were constructed with standard tools normally used for English. A length-constrained telephone number grammar that allowed only 4-, 7- and 10-digit strings was used for the **TEL** test set. An exact grammar that allowed only the 60 command phrases was used for the **CP** test set. A non-probabilistic word-pair grammar was used for the **PS** set.

3.3. Experiment Series I

The objective of these experiments was to evaluate different model types, using just the TI corpus training data:

- CI finite-duration models (baseline set)
- CI infinite-duration models
- multiple "mixture" models with 2, 3, 4 and 6 mixtures
- finite duration (unclustered) triphones, and
- infinite duration (unclustered) triphones

Our "mixtures" were implemented as a selection of one of N Gaussians per state based on minimum acoustic distance, rather than weighting and combining of N Gaussians. We use the term "Viterbi mixtures" to distinguish them from the more common usage of mixtures. Recognition performance was evaluated only on the **CP** and **TEL** sets. Since the triphones were not clustered, there were too many unseen triphone contexts in the **PS** test set for a fair test to be made at this point.

3.4. Results of Series I

Table 2 shows the recognition results on the two development test sets. The *#Models* count excludes the two silence models. In the table, the following terms need further explanation:

- **mix_non-mix**: a condition in which 6-mixture models were used only for those phones in which each mixture had at least 50 training exemplars. For phones with less than 50 exemplars in any mixture, the corresponding single mixture (baseline) models were used. We found that single-mixture models had to be used for /g/, /n /, /r/, /tS/, /w/ and /x/ as they fell below the threshold.
- **inf_triphones**: refers to infinite duration triphones.

From these experiments, it is clear that

- infinite duration models perform slightly worse than finite duration ones,
- multiple mixture models perform best on the CP task (3.7% word error with 2 mixtures), and

Table 2. Experiment Series I results (% word error) on the test vocabularies

Experiment	*#Models*	*CP*	*TEL*
CI_finite	46	4.7	9.7
CI_infinite	46	4.9	10.5
2-mix	92	3.7	9.9
3-mix	138	4.0	9.9
4-mix	184	3.9	10.0
6-mix	276	4.7	9.0
mix_non-mix	210	5.0	9.3
triphones	5424	4.9	5.9
inf_triphones	5424	5.1	6.9

- the triphone-based digit recognition performances are superior to the best monophone and mixture results by around 30-40%. This indicates the importance of contextual information for distinguishing among connected digits, which have a substantial number of inter-digit contexts.

3.5. Experiment Series II

The use of just 5000 phonetically representative sentences and a few hundred speakers produces recognition results in the 5 to 10% range. The better scores are on the larger vocabulary CP task with shorter utterances, but the dramatic improvement provided by triphones shows only in the digit strings, whose greater length offers more cross-word as well as within-word contextual effects.

We have had success in English recognition using acoustic phonetic decision trees (ADTs)[9] to cluster triphones at the state level [10, 11] as a means of modeling phonetic context. Thus in the second series of experiments, our objectives were: (i) to determine whether and by how much the addition of the VAHA training data improves performance, and (ii) to evaluate the effect of using ADTs to cluster triphones at the state level.

Based on the results of the first series of experiments, where some of the 3-mixture models had less than 100 tokens each, we decided not to include 4- and 6-mixture models in this series. We did include a phonetic sentence (PS) test condition, in view of the fact that the clustered triphones provide coverage for them. The following conditions were thus evaluated on CP, TEL and PS tasks:

- Baseline: 46 male and female monophone models
- 2Mix: monophone models with 2 Viterbi mixtures
- 3Mix: monophone models with 3 Viterbi mixtures
- TriClust: Triphones clustered using a decision tree

3.5.1. Triphone Clustering

We use the following procedure to cluster the Spanish triphones by tying the HMM states. In the following steps, we use the term "frame" or "acoustic frame" to refer to an acoustic observation. Note that in a finite-duration model, there is a many-to-one mapping between model states and acoustic frames.

1. An initial set of 2798 male and 2831 female finite-duration Spanish triphone models are created from the trained monophones.
2. The acoustics of these triphones are trained for 2 passes over the training set, followed by one pass of training for both the acoustics and the HMMs.
3. The HMMs of the trained triphones are now pooled into monophone HMMs, to compensate for the low training counts of many triphone models and to prevent missing transitions in the triphone HMMs, given that they are finite-duration models.
4. The pooled monophone HMMs, which *are* well-trained, are then cloned into triphone HMMs. Each of these "new" triphones will have the same HMM information as the corresponding monophone, but different acoustics.
5. A fourth pass over the training data with these new triphones is followed by yet another pooling and cloning procedure, as described in step 3.
6. A fifth and final pass is then made over the training data, using the triphones from step 5. At the end of step 6, we get triphones with well-trained acoustics *and* well-trained HMMs.
7. The triphone frames are then clustered using ADTs. The decision-tree questions for Spanish were formulated using Spanish phonological rules and constraints. Based on our experience with English, we used 1/3rd sub-model clustering, where the total number of frames in each triphone is partitioned into 3 groups with comparable numbers of frames. For example, a 3-frame 'b' model will have 1 1 1 as its partition, while a 4-frame 'r' will have 2 1 1 as its partition. A decision tree is formed for each group of frames in the partition.

After clustering, the frames were used directly for recognition, without retraining, which in our experience helps only marginally. We mapped the triphone frames in our recognition grammars to the clustered frames before running the recognition tests. Although this was a moot issue for the CP and TEL tasks, all of whose triphone contexts were already present in the training set, for the PS task this mapping insured that unseen triphone contexts in the test set were mapped to the nearest cluster.

3.6. Results of Series II

3.6.1. Triphone Clustering Results

We varied the number of clustered frames output from 2,000 to 12,000, obtaining a set of decision trees for each value. The objective was to determine the best performing cluster size for the different tasks. Table 3 below shows the results on the three tasks for different values of the cluster size. The two numbers in the *#Clusters* column represent the number of clusters requested and the number formed based on the decision trees. For each task, the number in parentheses indicates the number of frames actually used during recognition of the task vocabulary. The 12,000 triphone clusters were only evaluated on the PS task, as the performance on the CP and TEL tasks was already asymptotic for much smaller sizes.

Table 3. Triphone Clustering Results (% word error)

#Clusters	*CP*	*TEL*	*PS*
2000/1943	1.9 (1169)	5.5 (270)	12.9 (1938)
4000/3932	2.0 (1612)	5.6 (286)	12.3 (3901)
6000/5922	1.9 (1838)	5.6 (291)	12.1 (5809)
9000/8937	2.0 (2021)	5.6 (292)	12.7 (8437)
12000/11726	-	-	12.0 (11726)

Table 4. Experiment Series II Overall Results (% word error)

Experiment	*#Models*	*CP*	*TEL*	*PS*
Baseline	46	3.1	9.5	26.2
2Mix	92	3.1	11.6	26.6
3Mix	138	2.8	9.3	21.7
TriClust (6k)	5629	1.9	5.6	12.1

3.6.2. Overall Results

The overall results for Series II are shown in Table 4. (For ease of comparison, just the figures from the 6k cluster condition are repeated on the line representing the triphone clustering results.) Doubling the amount of available training data does improve performance substantially. On the CP task this is true across all conditions:

- 32% improvement with the baseline monophone models (3.1% vs. 4.7%),
- 16% improvement with 2-mixture models (3.1% vs. 3.7%), and
- 30% improvement with 3-mixture models (2.8% vs. 4.0%).

On the TEL task, the improvement is less dramatic, and in one condition (2-mixture models) the WER actually increases from 9.9% to 11.6%, an anomalous result which needs further investigation. Whether, or by how much, the addition of more data would improve performance on each task is also unknown at this point.

For the amounts of training data available, the clustered triphones, not surprisingly, provide the best performance in all three task vocabularies, though of course at a cost in terms of increased numbers of models. Compared to 3-mixture monophone models. for example, they provide improvements of 32%, 40% and 44% on the CP, TEL and PS tasks, respectively. They thus appear to do the best job of accounting for phonetic context on this scale.

Substitution errors contributed significantly to the error rates for the CP and TEL vocabularies in both series of experiments. For example, for the Spanish digits, we found that *tres* and *seis* were the most confusable (around 250 confusions), while among the command phrases, recognition performance was most affected by the confusion between *diferido* and *diferida*, and between *operador* and *operadora*.

4. CONCLUSION

Creating a vocabulary-independent speech recognition system for a new language with modest training resources is a challenging but realistic task. For a system being ported from English to Spanish and tested on different tasks, the performance levels in these first experiments are, if not commercially competitive, nevertheless encouraging as a point of departure for the development of useful technology. Further improvements can undoubtedly be achieved, especially for the small vocabulary tasks, with the ADT-clustered triphone techniques, with more targeted or adaptive training, and with the availability of more representative data.

REFERENCES

[1] I. Torres and F. Casacuberta. Spanish phone recognition using semicontinuous hidden markov models. In *ICASSP*, Minneapolis, MN, 1993.

[2] A. Varona, I. Torres, and F. Casacuberta. Discriminative-transitional/steady units for Spanish continuous speech recognition. In *Eurospeech*, Madrid, Spain, 1995.

[3] A. Bonafonte, R. Estany, and Eugenio Vives. Study of subword units for Spanish speech recognition. In *Eurospeech*, Madrid, Spain, 1995.

[4] L. Villarrubia, L. H. Gomez, J. M. Elvira, and J. C. Torrecilla. Context-dependent units for vocabulary-independent Spanish speech recognition. In *ICASSP*, Atlanta, GA, 1996.

[5] D. Tapias, A. Acero, J. Esteve, and J. C. Torrecilla. The VESTEL telephone speech database. In *ICSLP*, Yokohama, Japan, 1994.

[6] Y. Muthusamy, E. Holliman, B. Wheatley, and J. Picone. Voice Across Hispanic America: A telephone speech corpus of American Spanish. In *ICASSP*, Detroit, MI, 1995.

[7] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy. An evaluation of cross-language adaptation for rapid HMM development in a new language. In *ICASSP*, Adelaide, Australia, 1994.

[8] J. L. Hieronymus. Ascii phonetic symbols for the world's languages: Worldbet. Technical report, AT&T Bell Laboratories, Murray Hill, NJ, USA, 1994.

[9] L. R. Bahl, P. V. deSouza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny. Context dependent modeling of phones in continuous speech using decision trees. In *Proceedings of the DARPA Speech and Natural Language Processing Workshop*, Pacific Grove, CA, 1991.

[10] S. J. Young and P. C. Woodland. The use of state tying in continuous speech recognition. In *Eurospeech*, Berlin, Germany, 1993.

[11] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the ARPA Workshop on Human Language Technology*, Merrill Lynch Conference Centre, 1994.