SPOKEN PROPER NAME RETRIEVAL IN AUDIO STREAMS FOR LIMITED-RESOURCE LANGUAGES VIA LATTICE BASED SEARCH USING HYBRID REPRESENTATIONS

Murat Akbacak and John H.L. Hansen

Center for Robust Speech Systems
University of Texas at Dallas
Richardson, TX, 75083, U.S.A
{murat.akbacak,john.hansen}@utdallas.edu Web: http://crss.utdallas.edu

ABSTRACT

Research in multilingual speech recognition has shown that current speech recognition technology generalizes across different languages, and that similar modeling assumptions hold, provided that linguistic knowledge (e.g., phoneme inventory, pronunciation dictionary, etc.) and transcribed speech data are available for the target language. Linguists make a very conservative estimate that 4000 languages are spoken today in the world, and in many of these languages, very limited linguistic knowledge and speech data/resources are available. Rapid transition to a new target language becomes a practical concern within the concept of tiered resources. In this study, we present our research efforts towards multilingual spoken information retrieval with limitations in acoustic training data. We propose different retrieval algorithms to leverage existing resources from resource-rich languages as well as the target language using a lattice-based search. We use Latin-American Spanish as the target language. After searching for queries consisting of Spanish proper names in Spanish Broadcast News data, we obtain performance (max-F value of 28.3%) close to that of a Spanish based system (trained on speech data from 36 speakers) using only 25% of all the available speech data from the original target language.

1. INTRODUCTION

Multilingual information search in audio archives is expanding at an increasing rate as more audio data becomes available in different languages. In many cases, there are only a few or no linguists in high-interest languages leading to a considerable shortfall in transcription efforts within the task of acoustic model training for a multilingual speech recognition system. Therefore, portability to these new target languages becomes a practical concern. To improve the universality of current speech technologies, effective methods for rapid transition to new target languages using tiered resources are required and this presents new research challenges.

Large vocabulary multilingual speech recognition has been an area of intensive work at many research centers (e.g., [1, 2, 3]) for resource-rich languages such as English, German, French, Spanish, etc. where there is linguistic knowledge for language modeling and lexicon construction, as well as sufficient training data to train acoustic models. These studies employ an initial bootstrapping step to align acoustic data with the text provided in the target language using the source language acoustic models mapped via either knowledge-based or data-driven based phoneme mapping. Spoken information retrieval (SIR) for these languages is not a big challenge since they achieve reasonable performance at the recognition level. When resources are limited in a target language (e.g., Dari, Pashto, Somalian, etc.), the text hypothesis of the speech recognizer becomes more erroneous and this has a major impact on the performance of spoken information retrieval. Existing methods in spoken information retrieval would fail for this problem.

The problem of searching for spoken information through a noisy audio stream has been considered in previous studies for English [4]. In some studies such as [5, 6], the search is done through the recognition lattice or N-best list rather than being applied to 1-best word strings by considering the fact that the lattice structure provides additional information where the correct hypothesis could appear. For the purpose of searching for OOV words, sub-word (e.g., syllable, n-grams of mono-phones, etc.) representation based SIR has been employed in many systems [7, 8]. These studies investigate a decision fusion method to merge the retrieval results from systems using different representations in a weighted scheme. In [9, 10], an error correction scheme at the phoneme level is implemented via a confusion matrix, and phonetic retrieval based on the probabilistic formulation of term weighting using phoneme confusion data is presented.

Here, we will present our solution to the problem of multilingual spoken information retrieval with tiered resources knowing that there will be high error rates during recognition. We perform recognition at the phone level using different representations: source language mono-phones, target language mono-phones, and broad-class phones, and generate lattices for each utterance. Utterance lattices are indexed via weighted finite state transducers (WFSTs) as explained in [6]. We propose two novel retrieval algorithms. In the first algorithm, we use query-dependent dynamic weights during decision fusion. These weights show how well the pronunciation in the target language is represented with a given representation. Our second algorithm searches for a hybrid pronunciation network through a hybrid lattice where all representations coexist.

In Section 2, we present our formulation of several retrieval algorithms. Recognition results and evaluation of the proposed retrieval algorithms for Spanish are presented in Section 3. It is important to note that sufficient resources clearly exist for Spanish based ASR development. Our goal here is to intentionally limit the available resources to see what performance can be achieved as further data/resources are made available. Discussion and future work are presented in Section 4. Conclusions are presented in Section 5.

This research was sponsored in part by RADC under contract SC-05-02/S27410, and the state of Texas under Project EMMITT.

2. RETRIEVAL ALGORITHMS

2.1. Baseline System: Algorithm $\mathcal{A}$ – Weighted Parallel Lattice-based Search

2.1.1. Acoustic Modeling and Sub-word Unit Recognition

In our system, knowledge-based (e.g., IPA mapping [11]) and data-driven (e.g., confusion based) phoneme mapping are employed consecutively during bootstrapping and iterative training steps. Initial Spanish acoustic models are trained using the alignment generated with mapped English phoneme models. Next, these Spanish acoustic models are used in the alignment step. This procedure is repeated until the recognition error rate converges to a minimum. Alignments from the final alignment step are used to generate confusion-based phoneme mapping by running recognition on the Spanish training set with English acoustic models. The resulting data-driven phoneme mapping is used during bootstrapping and iterative alignment, and training steps are repeated. We perform recognition at the phoneme level using source language mono-phones, target language mono-phones, and broad-class phones, and generate lattices for each utterance. We assume that recognition at the word level is not feasible due to a lack of resources for acoustic model training and language model training.

2.1.2. Lattice Indexation via Weighted Finite State Transducers

We implemented the indexation and search scheme presented in [6] using AT&T's Finite State Machine (FSM) Toolkit [12]. A phone lattice for each speech utterance u_l ($l = 1, ..., n$) is generated using a phonetic recognizer, and represented with the transducer index T_l. Transducer index T is constructed by taking the composition of all utterance transducers T_l, $l = 1, ..., n$. The response to a query x is computed using the general algorithm of composition of weighted transducers [13]:

$$T = T_1 \circ T_2 \circ T_3 \circ ... \circ T_n$$
$$S(x) = P_x \circ T \quad S(x,u) = \iota_u(P_x \circ T). \tag{1}$$

$S(x)$ is the list of all utterance indices and their corresponding log likelihoods of containing query x. Applying the operator ι_u to this list gives the log likelihood of having query x in utterance u.

2.1.3. Embedding Confusion pairs into the query

Depending on how much audio data is available, one can use different methods to calculate confusion pairs. For the source language, we can perform a recognition test where we use trained acoustic models on a development test set to calculate the phonetic confusion matrix. Here, we use the TIMIT [14] database for this purpose since TIMIT is phonetically transcribed. Because there is a sufficient amount of audio data, we can calculate class-context-based trigram confusion probabilities such as the probability of English phoneme AE being recognized as AX when it is followed by the phoneme class STOP and preceded by the phoneme class FRICATIVE, which is represented as follows:

$$Prob(\mathrm{AX}|\mathrm{fricative} - \mathrm{AE} - \mathrm{stop}).$$

For the target language, the phonetic confusion matrix generated in the same way would not be reliable since the confusion statistics would be calculated from a small amount of data. To overcome this problem, a confusion matrix in the target language can inherit confusion statistics from the source language in two ways:

- Source language confusion matrix entries where confusable phoneme pairs exist in the target language are mapped to the target language using the phoneme mapping developed in Section 2.1.1.
- Confusion statistics are inherited at the decision-tree-class level from the source language. Target language phonemes not having a mapped source language phoneme share the confusion probability with a same decision-tree-class (e.g. alveolar-stop) target language phoneme.

The resulting confusion matrix in the target language is normalized to have row values that sum to 1. One can calculate the occurrence probabilities (e.g., unigram probabilities) of the target language phonemes, and use them to scale confusion probabilities. In the following sections, we will denote the resulting pronunciation network for the query x as $P_x^{\{i\}}$, and denote the resulting lattice index as $T^{\{i\}}$ respectively, for the i^{th} representation.

2.1.4. Decision Fusion

We form a new retrieval score by linearly combining the individual retrieval scores obtained from different representations,

$$S_i(x,u) = \iota_u(P_x^{\{i\}} \circ T^{\{i\}})$$
$$S(x,u) = \sum_{i=1}^{N} w_i \, S_i(x,u) \tag{2}$$

where w_i is a tunable weight parameter[1]. In our experiments, N is 3 since we use three sets of representations. The optimum value for the weight parameter w_i is found by performing retrieval tasks on a development set. In the baseline algorithm, the w_i values are fixed for every query.

2.2. Algorithm $\mathcal{B}$ – Dynamically Weighted Parallel Lattice-based Search

Within the concept of tiered resources, we consider the fact that depending on the reference pronunciation of a query, w_i values might be optimum in the global sense, but not locally. In this algorithm, we dynamically change the weight values for each query:

$$S(x,u) = \sum_{i=1}^{3} w_{xi} \, S_i(x,u) \tag{3}$$

where w_{xi} is a metric that shows how well the i^{th} acoustic unit set represents the reference pronunciation of the query word x. To be able to define this mathematically, we use the following definition:

$$w_{xi}(k) = w_i \, \alpha(k)^{I_k}$$
$$w_{xi} = \sum_{k=1}^{L_x} w_{xi}(k) = w_i \sum_{k=1}^{L_x} \alpha(k)^{I_k} \tag{4}$$

where w_i is the optimum value assigned in Alg. $\mathcal{A}$, and L_x is the number of phonemes in the reference query pronunciation. Depending on which representation these weights are calculated for, $\alpha(k)$ can be interpreted as a *similarity measure* (e.g., 1^{st} representation where source language mono-phones are used during recognition) or a *model confidence measure* (e.g., 2^{nd} representation where target language mono-phones are used during recognition) for the k^{th} phoneme in query x as shown in Figure 1. Here, we employ phoneme recognition accuracy (PRA) to assign values to $\alpha(k)$.

[1] A Broad-Class (BC) representation has lower weight since it provides less discriminative information during the retrieval task.

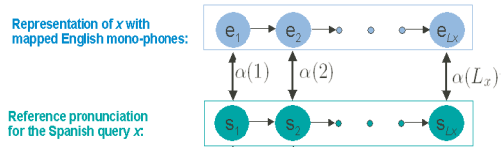

Fig. 1. Interpretation of similarity measure $\alpha(k)$ in Eq. 4.

- When we decode the target language development test set with the source language acoustic models, PRA represents the *similarity measure* between the target language mono-phone (k^{th} phoneme in x) and its mapped entry in the source language phoneme-set.
- When we decode the target language development test set with the target language acoustic models, PRA represents the *model confidence measure* for the target language acoustic model corresponding to the k^{th} phoneme in query x.

I_k in Equation 4 is set equal to either 0.5 or 1.0 depending on whether the mapping shares the same IPA symbol or not, respectively. In this way, we consider the linguistic similarity between the phonemes in addition to the acoustic similarity. When $\alpha(k)$ is calculated for the 2^{nd} representation, $I(k)$ has the value 0.5 for every k since the pronunciation is represented with the same set of phonemes. Using w_{xi}, we quantify how well the given target language pronunciation is represented with different acoustic model sets. This helps to weight our retrieval results dynamically based on the query pronunciation.

2.3. Algorithm $\mathcal{C}$ – Lattice-based search via hybrid pronunciation networks

In this algorithm, we construct a hybrid representation for the recognition lattice and the query pronunciation. In other words, the recognition lattice and the query pronunciation network contain source language mono-phones, target language mono-phones, and broad-class mono-phones at the same time,

$$T^{\{\text{hybrid}\}} = T^{\{1\}} \circ T^{\{2\}} \circ T^{\{3\}} \quad (5)$$
$$P_x^{\{\text{hybrid}\}} = (w_{x1} \cdot P_x^{\{1\}}) \circ (w_{x2} \cdot P_x^{\{2\}}) \circ (w_{x3} \cdot P_x^{\{3\}}).$$

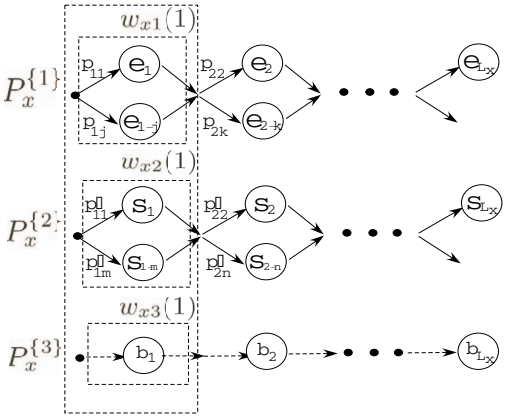

Fig. 2. Hybrid pronunciation construction in Eq. 5.

In Eq. 5, hybrid pronunciation network $P_x^{\{\text{hybrid}\}}$ is a composition of weighted pronunciation transducers. Each node of the pronunciation network is weighted with the corresponding $w_{xi}(k)$ as shown in Figure 2. w_{xi} can be considered as a vector consisting of $w_{xi}(k)$ values:

$$w_{xi} = [w_{xi}(1)\ w_{xi}(2)\ \ldots\ w_{xi}(L_x)]. \quad (6)$$

The resulting retrieval score is calculated by applying the operator ι_u to the composition of $P_x^{\{\text{hybrid}\}}$ and $T^{\{\text{hybrid}\}}$:

$$S(x,u) = \iota_u(P_x^{\{\text{hybrid}\}} \circ T^{\{\text{hybrid}\}}) \quad (7)$$

3. EVALUATION

To demonstrate performance for the proposed algorithms, we used Spanish as the target language, and focused on a proper name retrieval task within the broadcast news domain. While other languages (e.g., Dari, Pashto, Somalian, etc.) are possible, we selected Spanish to be able to intentionally limit the available resources to see what performance can be achieved as further data/ resources are available? In other words, we could select the tier level of resources (e.g., amount of training data) of interest for the algorithms.

The acoustic model development was based on the Latino-40 database [14] with the aid of English Wall Street Journal (WSJ) acoustic models via bootstrapping as explained in Section 2.1.1. The Latino-40 comprises about 5,000 utterances, 125 utterances from each of 40 different speakers (20 male, 20 female). We trained two continuous density, context dependent (CD), gender dependent (GD) Latino-40 models using data from 36 speakers and 10 speakers for training: $\mathbf{AM}_{\text{SPN}_{36}}$ and $\mathbf{AM}_{\text{SPN}_{10}}$, respectively. We performed mono-phone recognition experiments on two test sets: (1) $\text{test}_{\text{latino40}}$- 0.5 hour of speech (4 speakers, open set) from Latino-40, and (2) $\text{test}_{\text{SPN-BN}}$- 1 hour of speech from Spanish Broadcast News (SPN-BN). We note that Spanish Broadcast News (SPN-BN) corpus is held out and employed only for recognition and retrieval experiments rather than being used for acoustic model training. As you can see in Table 1, four sets of acoustic models are used during monophone recognition experiments: $\mathbf{AM}_{\text{SPN}_{36}}$, $\mathbf{AM}_{\text{SPN}_{10}}$, $\mathbf{AM}_{\text{ENG}}$ and $\mathbf{AM}_{\text{BC}}$. When ENGlish acoustic models ($\mathbf{AM}_{\text{ENG}}$) are used during recognition, depending on the test set ($\text{test}_{\text{latino40}}$ or $\text{test}_{\text{SPN-BN}}$), either English WSJ (ENG-WSJ) models or ENG Broadcast News (ENG-BN) models, respectively, are adapted via Maximum Likelihood Linear Regression (MLLR). Spanish acoustic models are adapted via MLLR as well during recognition experiment $\text{test}_{\text{SPN-BN}}$. For $\mathbf{AM}_{\text{BC}}$, we used 10 Broad Classes (e.g., vowel, nasal, glide, liquid, plosive (v/u), affricates (v/u), fricative(v/u)) with voicing/unvoicing (v/u) distinction. In all experiments, we used trigram phone language models that are trained from Latino-40 phone transcripts and mapped phone transcripts.

	$\mathbf{AM}_{\text{SPN}_{36}}$	$\mathbf{AM}_{\text{SPN}_{10}}$	$\mathbf{AM}_{\text{ENG}}$	$\mathbf{AM}_{\text{BC}}$
$\text{test}_{\text{latino40}}$	18.4	23.2	46.3	11.3
$\text{test}_{\text{SPN-BN}}$	31.8	38.7	54.2	23.1

Table 1. *Phoneme Error Rates (PERs) (%) for Latino-40 and Spanish Broadcast News tasks using different acoustic models.*

Based on SPN-BN transcripts, we segmented the audio data into shorter utterances where the gender information is provided so that gender dependent acoustic models can be used during decoding. For utterances shorter than 2 sec., and ones including overlapping segments and music/filler/commercial portions are discarded

during search. At the end, we have approximately 20K utterances worth of search material. We use 20 queries (variable lengths from 6 to 14 phonemes) that are mostly proper names during our retrieval experiments. Pronunciation for these proper names are generated via Letter-to-Sound rules that are trained using a pronunciation dictionary of 5000 Spanish words from the Spanish Callhome project. The top 1000 hits are considered during retrieval performance calculation. In Table 2 and Table 3, in addition to average precision values, we also compute the F-measure defined in terms of precision and recall[2] values:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

	$\mathbf{AM}_{\text{SPN}_{36}}$	$\mathbf{AM}_{\text{SPN}_{10}}$	$\mathbf{AM}_{\text{ENG}}$	$\mathbf{AM}_{\text{BC}}$
Avg. Precision	26.7	22.4	17.1	5.3
Max F	29.2	23.4	18.8	8.2

Table 2. *SIR performance in broadcast news domain.*

As we can see in Table 2 and Table 3, the spoken information retrieval system trained using a small amount of training material from the target language, yields performance close to $\mathbf{AM}_{\text{SPN}_{36}}$ that is fully trained on Spanish by using Algorithm $\mathcal{C}$ with acoustic models $\mathbf{AM}_{\text{SPN}_{10}}$, $\mathbf{AM}_{\text{ENG}}$ and $\mathbf{AM}_{\text{BC}}$. A key observation here is that a similar max-F and average precision values are obtained by using 75% less acoustic data from the target language (avg. precision of 26.7% and max-F value of 29.2% in Table 2 versus avg. precision of 26.1% and max-F value of 28.3% in Table 3). Another observation is that Algorithm $\mathcal{C}$ improves the baseline system (Algorithm $\mathcal{A}$) by around 2% absolute in terms of max-F measure.

	Algorithm $\mathcal{A}$	Algorithm $\mathcal{B}$	Algorithm $\mathcal{C}$
Avg. Precision	25.6	25.8	26.1
Max F	26.5	27.2	28.3

Table 3. *SIR performance via proposed algorithms using* $\mathbf{AM}_{\text{SPN}_{10}}$, $\mathbf{AM}_{\text{ENG}}$ *and* $\mathbf{AM}_{\text{BC}}$.

When the training set in the target language becomes less phonetically balanced, improvements obtained from the hybrid representation in Algorithm $\mathcal{C}$ become more substantial compared to other algorithms. These results are very promising, and suggest a viable procedure to follow for future advances in spoken information retrieval applications.

4. DISCUSSION AND FUTURE WORK

During the algorithm evaluations, our main goal was to find efficient algorithms to solve the data sparseness problem: how can we achieve similar performance using less acoustic data? Our proposed algorithms provide sufficient flexibility to leverage different tiers at the search level.

Although retrieval performance rates are low due to the fact that only mono-phones are used during retrieval, this knowledge can be used appropriately to either reject low probability streams, or provide further confidence using combined systems. It should be noted that in a potential bilingual SIR application (e.g., proper names from Somalian appear in English audio documents), results from word-based retrieval for the source language and results from phonetic retrieval for the target language can be merged to achieve higher performance rates.

[2]Precision rate is the percentage of retrieved material actually relevant. Average precision is calculated by averaging the precision values over queries. Recall rate is the percentage of relevant material actually retrieved.

Future work will focus on evaluating the existing framework in other languages, especially ones having far less acoustic overlap with the English acoustic space. Finding a correlation between the degree of acoustic overlap and retrieval performance improvement would be important to estimate how much resource/effort is needed to achieve a desirable performance in the target language. This would be useful when resources from multiple resource-rich languages (e.g., English, French, German, etc.) are leveraged with the resources from target language.

5. CONCLUSIONS

We described the structure and development process of a multilingual speech application using tiered resources. We performed experiments for the task of spoken information retrieval in a Spanish Broadcast News domain. We first generated a lattice using adapted English mono-phones and broad-class phones, and Spanish mono-phones trained from a limited amount of training data. Pronunciation for the query word is represented with a weighted transducer, in which confusable pronunciations are embedded. The current system achieved performance close to that of a Spanish system (trained on speech data from 36 speakers) using only 25% of all the speech data available from the target language. Given the time and expense in collection and transcription of audio materials for new languages, the proposed framework represents an important step towards rapid transition of spoken information retrieval systems to new languages with limited resources.

6. REFERENCES

[1] T. Schultz, et al., "Language-independent and language-adaptive acoustic modeling for speech recognition", Speech Communication, vol. 35, pp. 31-51, 2001.

[2] J. Kohler, "Multilingual phone models for vocabulary-independent speech recognition tasks", Speech Communication, vol. 35, pp. 21-30, 2001.

[3] U. Uebler, "Multilingual speech recognition in seven languages", Speech Communication, vol. 35, pp. 53-69, 2001.

[4] J.H.L. Hansen, et al, "Speechfind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word", IEEE Trans. Speech and Audio Proc., vol. 13, no. 5, Sept. 2005.

[5] D.A. James, et al., "A Fast Lattice-Based Approach to Vocabulary Independent Word spotting", in Proc. IEEE Conf. Acoustics, Speech, and Signal Processing, pp. 1029-1032, 2000.

[6] C. Allauzen, et al., "General Indexation of Weighted Automata - Application to Spoken Utterance Retrieval", in Proc. HLT-NAACL Conf., 2004.

[7] M. Witbrock, et al., "Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents", In Proc. 2nd ACM Int. Conf. on Digital Libraries, pp. 30-35, 1997.

[8] G.J.F. Jones, et al., "Retrieving spoken documents by combining multiple index sources", In Proc. SIGIR Conf., pp. 30-38, 1996.

[9] A. Amir, et al., "Advances in phonetic word spotting", in Proc. 10th Int. Conf. On Information and Knowledge Management, pp. 580-582, 2001.

[10] S. Srinivasan, et al., "Phonetic confusion matrix based spoken document retrieval", In Proc. ACM SIGIR, pp. 8187, 2000.

[11] J. Hieronymus, "ASCII Phonetic Symbols for the World's Languages: Worldbet", Technical report, AT&T Bell Laboratories, 1994.

[12] AT& FSM Library, http://www.research.att.com/sw/tools/fsm

[13] M. Mohri, et al., "Weighted finite-state transducers in speech recognition", Computer Speech and Language, 16(1), pp. 6988, 2002.

[14] Linguistic Data Consortium, http://www.ldc.upenn.edu