Spoken Proper Name Retrieval for Limited Resource Languages Using Multilingual Hybrid Representations

Murat Akbacak, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***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., phone 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 (e.g., different amounts of acoustically matched/mismatched data). In this paper, 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. Proposed algorithms employ confusion-embedded hybrid pronunciation networks, and lattice-based phonetic search within a proper name retrieval task. We use Latin-American Spanish as the target language by intentionally limiting available resources for this language. After searching for queries consisting of Spanish proper names in Spanish Broadcast News data, we demonstrate that retrieval performance degradations (due to data sparseness during automatic speech recognition (ASR) deployment in the target language) are compensated by employing English acoustic models. It is shown that the proposed algorithms for developing rapid transition of rich languages to underrepresented languages are able to achieve comparable retrieval performance using 25% of the available training data.**

***Index Terms*—Hybrid pronunciation, limited resource languages, multi-lingual speech systems, robust automatic speech recognition, spoken document retrieval, weighted parallel lattice search.**

Manuscript received July 15, 2007; revised October 22, 2009. First published November 03, 2009; current version published July 14, 2010. This project was funded by the AFRL through a subcontract to RADC, Inc., under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by John Hansen. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force or The University of Texas at Dallas. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

M. Akbacak was with the Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Richardson, TX 75080. He is now with SRI International, Menlo Park, CA 94025 USA.

J. H. L. Hansen is with the Center for Robust Speech System (CRSS), The University of Texas at Dallas, TX, 75080, USA (e-mail: john.hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2035785

I. Introduction

MULTILINGUAL information search in audio recordings (e.g., audio broadcasts, archives from digital libraries, meetings, audio content on the internet, etc.) is expanding at an increasing rate as more audio data becomes available in different languages. In many of these languages, data sparseness is an important issue since the available training material (e.g., audio and text) needed to train ASR systems, generally speaking, is limited and diverse (e.g., different recording conditions, speaking styles, accents/dialects, etc.). In addition to this, there is limited automatic transcription support and few linguists are experienced, leading to a considerable shortfall in transcription and/or lexicon creation efforts. When these limited-resource languages become high-interest languages (e.g., military applications, or rescue applications abroad), it is important to build speech systems rapidly with tiered resources [1], [2]. For example, when a natural disaster takes place in a country where the spoken language has limited resources, rapid development of audio search applications might become crucial as these applications can be used to find important portions of local broadcast news that need to be translated to foreign rescue teams. Rapid transition of resource-rich language-based automatic speech recognition systems to new dialects of the same language is also of significant interest to the research field [3]. Being able to build a system with the available resources (both limited resources from the target language and existing resources from other resource-rich languages) requires an understanding of how these different knowledge sources can be leveraged in an audio search application.

Large-vocabulary multilingual speech recognition has been an area of intensive work at many research centers (e.g., [4]–[6]) for resource-rich languages such as English, German, French, Spanish, etc. where there are linguistic knowledge and text resources 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

[1]For larger SIR applications where the query contains multiple keywords to represent a topic to be searched, it has been shown that the relationship between ASR performance and SIR performance is not linear mainly because not all query terms are of same importance during search (e.g., stop words), and there is some semantic dependency in the query formulation making it possible to recover some ASR errors by employing query-expansion methods. In this paper, we consider the impact of ASR errors on SIR performance still a major issue because the SIR application we focus on is term (more specifically proper name) retrieval, making the ASR and SIR performances more correlated.

the text provided in the target language using the source language acoustic models mapped via either knowledge-based [4], [5] or data-driven based [7]–[11] phone mapping. Depending on the available amount of aligned speech data in the target language, source language acoustic models are adapted towards the mapped target language acoustic models, or target language acoustic models are directly trained. Spoken information retrieval (SIR) for these languages is not a major 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 automatic speech recognizer becomes more erroneous and this has a major impact on the performance of spoken information retrieval.[1] Existing retrieval methods employ language dependent representations, and they do not provide a formal solution that employs existing acoustic knowledge from resource-rich languages along with the limited acoustic knowledge in the target language to have robust retrieval in the target language.

For resource-rich languages, the main sources of errors during recognition are changing acoustic conditions, different speaking styles/stress/emotion [13], speaker traits (e.g., accent [14], dialect [15], whisper/shouted [16]), and others.[2] The problem of searching for spoken information through a noisy audio stream has been considered in previous studies for English [19], [20]. In some studies such as [21] and [22], 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 out-of-vocabulary (OOV) words, and misrecognized in-vocabulary (In-Vocab) words as well, sub-word (e.g., syllable, n-grams of phones, etc.) based SIR has been employed in many systems [25]–[27]. Although shorter units are capable of representing OOV words, longer units capture more discriminative information, and compared to shorter units are less susceptible to false matches during retrieval. In order to move towards solutions that address the problem of misrecognition (both In-Vocab word and OOV word errors) during SIR, previous studies have employed fusion methods [21], [22], [28], [29] to recover from recognition errors during retrieval, but at the same time to keep false alarm rates low. In these methods, fusion weights[3] for word based and subword based retrieval systems are optimized for specific tasks. More importantly, these methods do not fully benefit from subword-based retrieval due to the fact that fusion weights are optimized globally over all documents, and test data might be acoustically heterogeneous. Results from these studies show small improvements over word-based-only retrieval approaches as the number of documents increases. In [30] and [31], an error correction scheme at the phone level was implemented via a confusion matrix, and phonetic retrieval based on the probabilistic formulation of term weighting using phone confusion data presented.

[2]Here, we consider all these sources of errors as noise within the context of speech recognition.

[3]Fusion means system combination where individual system scores are weighted and summed to output a single retrieval score.

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. Different tiers might result from changing lexicon coverage[4] (poor coverage at the beginning, with an evolving lexicon as more resources are employed to obtain better coverage), or data sparseness or mismatch during acoustic model training. In recent studies, multilingual information retrieval systems, in the form of spoken term detection, are deployed [12]–[18] in different languages such as Arabic, Chinese, and Turkish, but these studies do not focus on data sparseness and language leveraging since there are enough resources for these languages. In this paper, we focus exclusively on the problem of data sparseness during acoustic model training in the target language. We consider this problem as language leveraging where acoustic knowledge from source and target languages is leveraged at the retrieval level by modeling inter-language phonetic confusions as well as intra-language phonetic confusions. Inter-language confusions are used to decrease the language mismatch problem by capturing acoustic similarities between source and target languages, whereas the intra-language confusions are used to decrease the impact of less-accurate acoustic models resulting from data sparseness during acoustic model training in the target language. Although the main focus of this paper is on language leveraging, we present a discussion on the applicability of the proposed algorithms on other acoustic leveraging problems in Sections II and VI. In the algorithm formulation, we perform recognition and retrieval at the phone level using different representations: source language phones, target language phones, and broad-class phones, and generate lattices for each utterance. Utterance lattices are indexed via weighted finite state transducers (WFSTs) as explained in [22]. We consider the system employing decision fusion with fixed weights as our baseline system. 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 multilingual hybrid pronunciation network through a hybrid lattice where all representations coexist. In comparison to the former two, in this scheme language leveraging (or fusion) is done at the phone level rather than at the query level during retrieval.

This paper is organized as follows: In Section II, we introduce the concept of tiered resources with sample scenarios towards which the proposed algorithms can be applied. We consider language leveraging as a specific case of acoustic leveraging where the goal is to use acoustically different knowledge sources (e.g., acoustic models trained on different acoustic conditions, in this paper different languages) at the search level to do fusion-based or hybrid-based spoken information retrieval. In Section III, we present our formulation of several retrieval algorithms using multilingual pronunciation representations. Recognition results and evaluation of the proposed retrieval algorithms for Latin-American Spanish are presented in Section IV. It is im-

[4]Lexical variation over time (e.g., 1890s to present) for resource-rich languages are out of the scope of this paper [19].

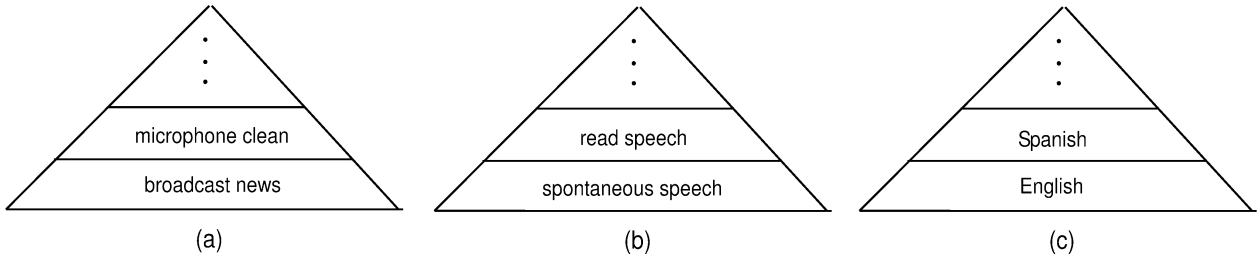

Fig. 1. Scenarios for different tiered structures constructed according to the available data size in different (a) acoustic conditions, (b) speaking styles, or (c) spoken languages. Bottom slices in the pyramids correspond to type of data which is more in quantity but might be a poor fit to the acoustic condition of the task (corresponding to one of the higher tiers).

portant 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 V, with conclusions presented in Section VI.

II. Tiered Resources

Speech recognition technology is based on statistical pattern recognition methods that require a training phase to generate statistical models to represent acoustic events. The accuracy of these acoustic models will depend on the *quality/level* of the transcription, and also the *amount* of available training data. The level of transcription depends on the linguistic knowledge which includes the set of acoustic units used for recognition/transcription, pronunciation dictionary, and the qualifications of the transcriber (e.g., being a speaker of the language or dialect, being a linguist, having linguistic background in languages similar to the target language, etc.). Data resources include available speech data for acoustic model training, transcription text, and text for language model training.

Performance of a multilingual speech application will depend on the availability and accuracy/quality of these resources. For example, errors in transcription (e.g., the wrong phone or word) introduce a major outlier in HMM training. Token weighting [23], and bias normalization [24] have been shown to reduce this impact. Here, we consider this concept as a "tiered structure" in a new target language since accuracy/quality and amount of available resources will be variable. Within this concept, different tiers will correspond to different operating (recognition and retrieval) performances. It is crucial to develop algorithms that work robustly across tiers, and to provide solutions to compensate for performance differences by using other knowledge sources when the system moves from one tier to a lower tier. We realize that there might be different scenarios, as shown in Fig. 1, depending on what criteria (e.g., different amounts of data in multiple acoustic conditions) is used to construct the tiers. In this paper, we construct different tiers by using different amounts of training data from the source and target language as shown in Fig. 1(c). Our main focus is the case where very limited amount of training data (e.g., speech data with sentence transcriptions) is available in the target language. However, in our algorithm development we provide sufficient flexibility to easily move towards an alternative scenario [e.g., leveraging resources from different acoustic conditions, Fig. 1(a), or leveraging resources from different speaking styles, Fig. 1(c)] which allows the same framework to be used as well. It will be easier to make this generalization if language leveraging is considered as a specific case of acoustic leveraging in general. After presenting proposed algorithms and evaluating them on language leveraging for retrieval, Fig. 1(c), we will have a discussion on applying the proposed algorithms on other acoustic leveraging scenarios in Section V.

III. Algorithm Development

As mentioned before, our research goal is to focus on spoken information retrieval applications in limited-resource target languages, and to leverage information between resource-rich languages and the target language to compensate performance differences between different tiers (different amounts of training data) in the target language. We refer to this as language leveraging or fusion which can be considered as a specific case of acoustic leveraging in general. The following subsections provide the details on system development towards language leveraging for a retrieval task.

We propose three algorithms.[5] In the baseline algorithm, we present the system components that include: multilingual acoustic model training and phone recognition for a new target language, finite-state transducer (FST) based lattice indexing and search, confusion-network based pronunciation representation, and decision fusion. In the baseline system, we employ different retrieval systems with different language-dependent representations, and merge the retrieval results. In the second algorithm, we merge the retrieval systems' outputs in a dynamic fashion where the fusion weights depend on the query pronunciation. In other words, we employ query-dependent dynamic weights during decision fusion. In our third algorithm, instead of merging language-dependent system outputs, we employ multilingual hybrid representations to represent query pronunciation and spoken documents.

A. Algorithm $\mathcal{A}$—Weighted Parallel Lattice-Based Search

This algorithm can be considered as our baseline system. Here, we present system components such as acoustic model training and recognition, lattice indexing, phonetic confusion networks, and decision fusion of language-dependent retrieval systems.

1) Acoustic Model Training and Sub-Word Unit Recognition: We employed a multilingual acoustic modeling approach that is similar to the previously proposed methods ([4]–[6]) to port an

[5]The first algorithm is considered as our baseline system.

English based ASR system to a new language.[6] In our system, knowledge-based (e.g., IPA mapping [32]) and data-driven (e.g., confusion or phonetic distance based) phone mapping are employed consecutively during bootstrapping and iterative training steps.

Initial Spanish[7] acoustic models are trained using the alignment generated with IPA-mapped English phone models. Next, these initial Spanish acoustic models are used in the alignment step. This procedure is repeated until the phone recognition error rate (PER) converges to a minimum. In our experiments, five iterations are used. Spanish alignments from the final alignment step are used to generate confusion-based phone mapping by running recognition on the Spanish training set with English acoustic models. Most frequent phone confusion pairs are used as mapping. This data-driven phone mapping is used during bootstrapping and iterative alignment/training steps. Performance improvements obtained by using knowledge-based and data-driven phone mappings consecutively are presented in Section VI.

We perform recognition at the phone level using source language phone models (adapted via maximum-likelihood linear regression-based adaptation to have better-matching acoustic models to the target language), target language phone models, and language-independent broad-class phone models (e.g., vowel, nasal, glide, etc.) with voicing/un-voicing (v/u) distinction, and generate lattices for each utterance. Parallel phonetic recognizers represent different tiers which need to be leveraged to obtain optimal retrieval performance.

We assume that recognition at the word level is not feasible due to a lack of resources (e.g., text data for language model training, pronunciation dictionary with a good coverage) for acoustic and language model training for a limited-resource language. To be able to observe an upper-bound on performance levels and to make comparisons with phonetic-based retrieval approaches, we also provide word-based recognition and retrieval results in Section VI by using all available Spanish audio and text to build a Spanish word recognizer and word-based retrieval engine.

2) Lattice and Query Indexing via Weighted Finite-State Transducers: As we mentioned before, state-of-the-art audio search systems employ lattice-based search to lower the impact of recognition errors on retrieval performance by considering the fact that the recognition lattice provides additional information where the correct hypothesis could appear. In our system, we implemented the lattice-based indexing and search scheme presented in [22] using AT&T's Finite State Machine (FSM) Toolkit [33].

Lattice output of the phone recognizer for each speech utterance u_l $(l = 1, \ldots, n)$ is considered as automata where the path weights correspond to joint probabilities. General weight-pushing algorithm in the log semiring is applied to this automata to convert these weights to the desired (negative log of) posterior probabilities, more generally log-likelihoods as done in [22]. From this weighted automaton, we create unique initial and final states, and also create two different types of transitions: 1) a transition from the new initial state to every existing state with epsilon input/output labels and weights being negative-log of forward probability to reach the existing state; and 2) a transition from every existing state to new final state with epsilon input label, utterance ID number l as the output label, and weights being the negative-log of the backward probability to reach the final state. To summarize, except the final transitions in transducer T_l, input symbols are the phone labels, and the output symbols are epsilons, with transducer probabilities. In the final transitions, the input symbol is epsilon and the output symbol is l with transition probability equal to 1. Transducer index T is constructed by taking the union of all utterance transducers $T_l, l = 1, \ldots, n$. The response to a query x is computed using the general algorithm of composition of weighted transducers [34]:

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

P_x is the pronunciation network for the query x, and is represented as a transducer with inter-language and/or intra-language phonetic confusions. $S(x)$ 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.

3) Embedding Confusion Pairs Into the Query: We use a previously proposed approach [30], [31], namely phonetic confusion modeling, to mimic recognition errors in our query formulation. This is similar to a lattice-based indexing approach where we have alternative document representations, except this time we also consider alternative representations on the query side. We try to introduce confusion patterns in our query representation so that we can match queries in the recognition lattice. In other words, we try to mimic recognition errors in the query representation to decrease the impact of language mismatch (resulting from using source language acoustic models on the target-language audio documents) and also to decrease the impact of using less-accurate target-language acoustic models (resulting from having sparse training data). Different from the previous work on phonetic-confusion modeling, in addition to modeling intra-language phonetic confusions, we also model inter-language phonetic confusions as well, since our motivation is to effectively leverage the source and target language knowledge sources at the retrieval level.

Depending on how much audio data is available, different methods can be used to calculate confusion pair probabilities by using either data-driven approaches or linguistically motivated approaches, or both. In the query representation, instead of rule-based phone mappings, we use data-driven phone mappings, which also allows us to construct probablistic query representations using transducers. 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 [35] database for this purpose since TIMIT is phonetically transcribed. Since there is a sufficient amount of audio data, we can calculate class-context-based trigram confusion probabilities such as the probability of English phone AE being recognized as AX when it is followed by the phone class STOP and preceded by the phone class FRICATIVE, which is represented as

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

[6]Different from these studies, data sharing in a joint acoustic model training fashion is not employed since it is not the focus of this paper.

[7]Spanish and English are considered as target and source languages, respectively.

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 phone pairs exist in the target language are mapped to the target language using the phone mapping developed in Section III–A1.
- Confusion statistics are inherited at the class level from the source language by using the classes from phonetic decision-tree questions[8] (e.g., is the phone on the right of the current phone a *vowel*?) which are used in ASR during acoustic model training. Target language phones not having a mapped source language phone share the confusion probability with a same phonetic class (e.g., alveolar-stop) target language phone.

The resulting confusion matrix in the target language is normalized to have row values that sum to 1. Occurrence probabilities (e.g., unigram probabilities) of the target language phones can be used to scale confusion probabilities, but the impact of this kind of approach on the final retrieval metric is not significant since there is already enough redundancy in the query representation. 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 ith representation. Here, i refers to the index of the recognizer (e.g., $i = 1$ is source language phone recognizer, $i = 2$ is target language phone recognizer, $i = 3$ is language-independent broad-class phone recognizer) among all parallel recognizers.

4) Decision Fusion: We form a new retrieval score by linearly combining the individual retrieval scores obtained from different representations:

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

where w_i is a tunable weight parameter.[9] In our experiments, N is set to 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.

B. 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.

[8]Decision tree questions are used in acoustic model training. Since there is not enough training data to model context-dependent phones separately, these phones are clustered using decision trees and the resulting cluster of phones are modeled jointly.

[9]A broad-class (BC) representation has lower weight since it provides less discriminative information during the retrieval task.

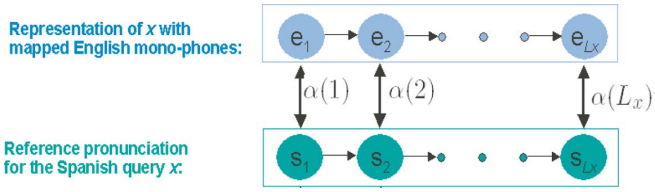

Fig. 2. Interpretation of similarity measure $\alpha(k)$ in (4).

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 ith 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 Algorithm $\mathcal{A}$, L_x is the number of phones in the reference query pronunciation, and I_k is a parameter used to take linguistic similarity (in addition to acoustic similarity calculated on real data) between phones into consideration. Depending on which representation the w_i weights are calculated for, $\alpha(k)$ can be interpreted as a *similarity measure* (e.g., first representation where source language phones are used during recognition) or a *model confidence measure* (e.g., second representation where target language phones are used during recognition) for the kth phone in query x. In Fig. 2, the similarity measure, $\alpha(k)$, is shown when the source-language phone representation is used for the query x. Here, we employ phone recognition accuracy (PRA) to assign values to $\alpha(k)$.

- When we decode the target language development test set with the source language acoustic models, PRA represents the *similarity measure* between the target language phone (kth phone in x) and its mapped entry in the source language phone-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 kth phone in query x.

I_k in (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 phones in addition to the acoustic similarity. The power terms 0.5 and 1.0 are not the optimized values to get the best retrieval performance, in other words these values are not tuned using a development set. Here, we want to emphasize the fact that both linguistic and acoustic similarities are being used in our system. When $\alpha(k)$ is calculated for the second representation, I_k has the value 0.5 for every k since the pronunciation is represented with the same set of phones. Using w_{xi}, we quantify how well

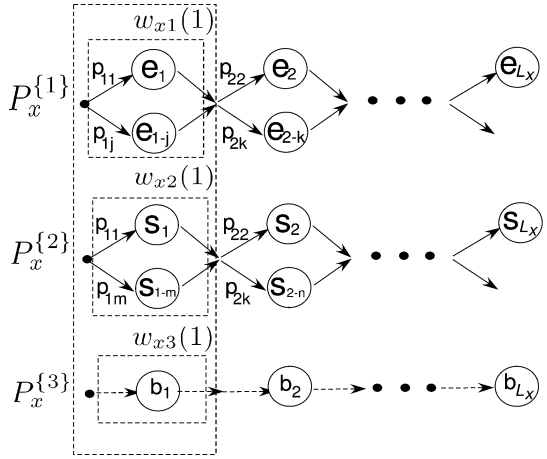

Fig. 3. Hybrid pronunciation construction.

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.

C. 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 phones, target language phones, and broad-class phones at the same time.

In this method, we employed interpolated phone n-gram language models to merge language dependent language models by using 3-gram context. We do it using a class based language model since the recognition units are different. The amount of language model training data is always the same as what we used for acoustic model training:

$$\begin{aligned} T^{\{\text{hybrid}\}} &= T^{\{1\}} \cup T^{\{2\}} \cup T^{\{3\}} \\ P_x^{\{\text{hybrid}\}} &= \left(w_{x1} \cdot P_x^{\{1\}}\right) \cup \left(w_{x2} \cdot P_x^{\{2\}}\right) \\ &\quad \cup \left(w_{x3} \cdot P_x^{\{3\}}\right). \end{aligned} \tag{5}$$

In (5), the hybrid pronunciation network $P_x^{\{\text{hybrid}\}}$ is a union of the three weighted pronunciation transducers. Each node of the pronunciation network is weighted with the corresponding $w_{xi}(k)$ as shown in Fig. 3. In this hybrid pronunciation network, transitions between different representations are allowed although not shown in Fig. 3. Here, w_{xi} can be considered as a vector consisting of $w_{xi}(k)$ values with k indexing across the L_x phones in the reference query:

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

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

$$S(x,u) = \iota_u \left(P_x^{\{\text{hybrid}\}} \circ T^{\{\text{hybrid}\}}\right). \tag{7}$$

Next, we consider the evaluation of the three proposed algorithms.

TABLE I
PHONE ERROR RATES (PERS) (%) FOR LATINO-40 AND SPANISH BROADCAST NEWS TASKS USING DIFFERENT ACOUSTIC MODELS

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

IV. EVALUATIONS

To demonstrate performance for the proposed algorithms, we used Spanish as the target language, and focus on a proper name retrieval task within the broadcast news domain. While other languages (e.g., Dari, Pashto, Somalian, etc.) are possible, we select Spanish to be able to intentionally limit the available resources to assess the potential performance improvement which could 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 [35] with the aid of English Wall Street Journal (WSJ) acoustic models via bootstrapping as explained in Section 3-A1. The Latino-40 comprises 5000 utterances, with 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 first from 36 speakers, and a second from ten speakers for training: $\mathbf{AM}_{\text{SPN}_{36}}$ and $\mathbf{AM}_{\text{SPN}_{10}}$, respectively. We performed phone recognition experiments on two test sets: 1) $\text{test}_{\text{latino40}}$: 0.5 hour of speech (four 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 the 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 can be seen in Table I, four sets of acoustic models are used during phone recognition experiments: $\mathbf{AM}_{\text{SPN}_{36}}, \mathbf{AM}_{\text{SPN}_{10}}, \mathbf{AM}_{\text{ENG}}$ and $\mathbf{AM}_{\text{BC}}$, where SPN, ENG, and BC subscripts are used to denote Spanish, English, and broad-class phone acoustic models, respectively. When English (ENG) 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) using a single class transformation. Spanish acoustic models are adapted via MLLR as well during recognition experiment: $\text{test}_{\text{SPN-BN}}$. $\mathbf{AM}_{\text{BC}}$ are language-independent broad-class acoustic models. We used ten broad classes (e.g., vowel, nasal, glide, liquid, plosive (v/u), affricates (v/u), fricative (v/u)) with voicing/un-voicing (v/u) distinction. In all experiments, we used 3-gram phone language models that are trained from Latino-40 phone transcripts and mapped phone transcripts. During Algorithm $\mathcal{C}$, interpolated language models are employed by using a phone mapping table as class information.

As can be seen in Table I, there is a 21% (from 18.4% to 23.2%) and 26% (from 31.8% to 38.7%) relative increase in PER when $\mathbf{AM}_{\text{SPN}_{10}}$ is used instead of $\mathbf{AM}_{\text{SPN}_{36}}$ on Spanish Latino-40 and Spanish Broadcast News testsets, respectively. As expected, the performance of English acoustic models ($\mathbf{AM}_{\text{ENG}}$) on Spanish recordings is not very good (e.g., 54.2% when **English WSJ** models are used to recognize Spanish

TABLE II
DESCRIPTION OF DOCUMENT AND QUERY SETS IN SPANISH BN

Number of Documents	5000
Average Length of Documents	9 seconds
Average # of Words per. Documents	11 words
Number of Queries	100
Average Length of Queries	8 phonemes
Number of Relevant Documents	356
Average Relevant Documents per. Query	3.56 doc

TABLE III
SIR PERFORMANCE IN BROADCAST NEWS DOMAIN

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

TABLE IV
SIR PERFORMANCE VIA PROPOSED ALGORITHMS USING $\mathbf{AM}_{\mathrm{SPN}_{10}}$, $\mathbf{AM}_{\mathrm{ENG}}$ AND $\mathbf{AM}_{\mathrm{BC}}$

	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

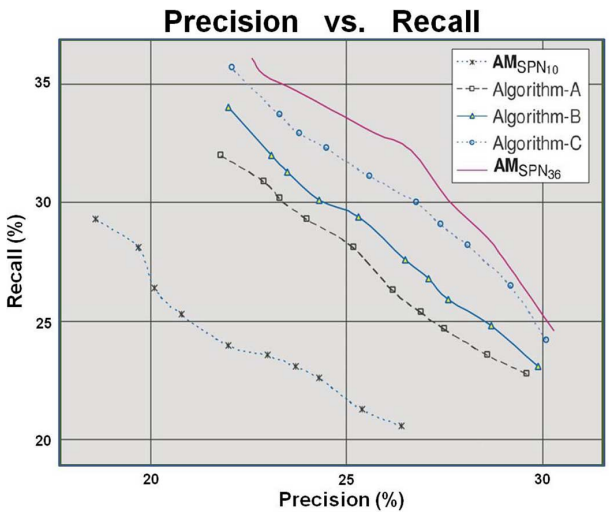

Fig. 4. Precision versus recall curves for $\mathbf{AM}_{\mathrm{SPN}_{10}}$ and $\mathbf{AM}_{\mathrm{SPN}_{36}}$ systems, and proposed algorithms.

broadcast news documents). On the same testset, the performance of $\mathbf{AM}_{\mathrm{BC}}$ is much higher (e.g., 31.8% versus 23.1%) than the one $\mathbf{AM}_{\mathrm{SPN}_{36}}$ yields mainly due to a reduced number of acoustic units to be recognized during broad-class phone recognition.

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. Utterances shorter than 2 s, and any that include overlapping speaker segments and music/filler/commercial portions are discarded during search to set any environmental acoustic mismatch problems aside. At the end, we have 5k utterances worth of search material.

As can be seen in Table II, we use 100 queries (variable lengths from 6 to 14 phones) that are proper names, for which the annotations are provided by LDC, during our retrieval experiments. Pronunciation for these proper names are generated via letter-to-sound (LTS) rules that are trained using a pronunciation dictionary of 5000 Spanish words from the Spanish Callhome project [35]. For other target languages having more LTS rules than Spanish has, it might be more critical to do a careful manual check for the query pronunciations.

Fig. 4 shows the precision-recall curves for proposed algorithms. In Tables III and IV, in addition to average precision values, we also compute the F-measure defined in terms of precision and recall[10] values:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

As shown in Tables III, IV, and in Fig. 4, the spoken information retrieval system trained using a small amount of training material from the target language, yields performance close to the $\mathbf{AM}_{\mathrm{SPN}_{36}}$ model which is fully trained on Spanish by using Algorithm $\mathcal{C}$ with acoustic models $\mathbf{AM}_{\mathrm{SPN}_{10}}, \mathbf{AM}_{\mathrm{ENG}}$ and $\mathbf{AM}_{\mathrm{BC}}$. A key observation here is that similar max-F and average precision values are obtained using 75% less acoustic data from the target language (average precision of 26.7% and max-F value of 29.2% in Table III versus average precision of 26.1% and max-F value of 28.3% in Table IV). Another observation is that Algorithm $\mathcal{C}$ improves the baseline system (Algorithm $\mathcal{A}$) by around 2% absolute in terms of the max-F measure. Another promising result is that the Max-F metric is improved from 29.2% to 32.6% when algorithm $\mathcal{C}$ is used to leverage systems using acoustic models $\mathbf{AM}_{\mathrm{SPN}_{36}}$ and $\mathbf{AM}_{\mathrm{ENG}}$. These improvements are statistically significant (for Wilcoxon signed-ranks test on the pair of retrieval results using the difference of max-F values of all queries) with $p < 0.05$. With the same statistical significance testing procedure, when we compare the proposed algorithms, we observe that they are statistically different from each other.

We also checked how well the system fusion weights generalize from dev-set to test-set for Algorithm $\mathcal{A}$ and $\mathcal{B}$. For Algorithm $\mathcal{C}$, the parameter space is too large, since weights are not at the query level but instead are at the phone level, and these weights are used to do system fusion in a hybrid fashion, rather than at the system output level. For Algorithm $\mathcal{B}$, although query-specific weights are dynamically calculated from the phone-level weights, we can still re-optimize the query weights on the test-set since fusion is done at the system output level. We ran a brute-force search, and observed that optimum weights on the test-set produced an overall relative improvement of 1.39% and 4.82% on Max-F scores (compared to results obtained with the weights optimized on the dev-set) for Algorithm $\mathcal{A}$ and $\mathcal{B}$, respectively. Both values for the relative improvement are less than 5%, and we can conclude that weights are well generalized between dev-set and test-set.

To see the upper bound on the retrieval performance, we deployed a word-based Spanish retrieval system assuming that we have enough training material to train acoustic and language models for a Spanish LVCSR system. An initial pronunciation lexicon (approx. 45k words) was obtained from the Callhome

[10]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.

Spanish lexicon [35]. During our experiments, we used a larger lexicon (approx. 51k words) with an additional most frequently occurring 5k words from the Spn-BN corpus, and all query words that are used in retrieval experiments. In other words, OOV rate for the query words is zero. We used the Spanish letter-to-sound rules to generate pronunciations for these additional words. N-gram $(\mathrm{N} = 3)$ language models at the word level were trained using Spanish Newswire Text corpus [35] consisting of 5 Million words. 2-grams and 3-grams occurring less than 4 times are pruned during N-gram counting. Sentences having high OOV rates (in our experiments sentences with more than 40% OOV rate) are also discarded in our language model training to prevent spelling errors, as well as high unigram probability for the generic OOV word *UNK*. During word recognition, we also apply single-class MLLR adaptation. Word error rate was calculated as 29.31% and 26.83 for 1-best and oracle, respectively, on the SPN-BN test data which is used in our previous recognition and retrieval experiments above.

Next, word lattices are converted into word-confusion networks using the SRI Language Modeling toolkit [42], and a word index table is created with corresponding time locations and word posterior probabilities. An offset value of 0.5 seconds is used since the word aligment information that the ASR produces might be slightly off. On the same set of queries, a Max-F value of 36.4% is obtained by using word-based retrieval. Word-based retrieval performance is not as high as it would be expected to be since the queries being searched are proper names and are not represented well in the language model and this results in low ASR performance, and accordingly retrieval performance, for these queries.

V. Discussion

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

Although retrieval performance rates are low due to the fact that only 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.

During our algorithm formulation, we did not make any assumption about the query term being a proper name or not. In other words, the proposed algorithms can be applied to any spoken term retrieval task. However, the upper-bound performance which is considered as the word-based system will be higher due to better modeling of frequent terms with the limited amount of language model training data during word-based recognition. This is why we specifically focused on proper name retrieval in this study, and we proposed algorithms that employ phonetic recognition and retrieval, since for proper names word-based systems are not feasible with limited amounts of training data.

Future work could 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 a target language. In addition to the level of acoustic similarity between target and source languages, complexity in letter-to-sound rules in the target language is also an important factor for building proper name retrieval applications where proper names are most likely not in the pronunciation lexicon. In this paper, this problem is minimized since Spanish has a shallow orthography. As part of future work, we also plan to look at different tiers as the pronunciation dictionary in the target language evolves (from poor-coverage towards good-coverage) during time, making it feasible to run word-based recognition and retrieval. One approach would be to employ word-based and phone-based systems in a parallel or hybrid fashion (where the recognition system outputs words and phones during decoding).

Although the proposed algorithms are shown to be effective solutions to remove/decrease the impact of language variation, they could also be employed to remove/decrease the impact of dialect variation (e.g., transition from one dialect to another where limited resources exist), accent (e.g., English documents spoken by non-native speakers), acoustic environment variation (e.g., leverage acoustic knowledge from microphone speech data with the one from limited amount of broadcast news data), speaker traits (e.g., rapid transition from read speech to spontaneous speech).

VI. Conclusion

In this study, 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. Novel aspects include generating a lattice using adapted English phones and broad-class phones, and Spanish phones trained from a limited amount of training data. The 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 fully trained 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 required and expense incurred 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.

References

[1] M. Akbacak and J. H. L. Hansen, "Spoken proper name retrieval in audio streams for limited-resource languages via lattice based search using hybrid representations," in *Proc. IEEE Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Toulouse, France, 2006, pp. 113–116.

[2] M. Akbacak and J. H. L. Hansen, "A robust fusion method for multilingual spoken document retrieval systems employing tiered resources," in *Proc. ISCA Interspeech'06/ICSLP'06*, Pittsburgh, PA, Sep. 2006, pp. 1177–1180.
[3] S. AmudaH. Boril and J. H. L. Hansen, "Limited resource speech recognition for Nigerian English," in *Proc. IEEE ICASSP'10*, Dallas, TX, Mar. 2010, pp. 5090–5093.
[4] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, pp. 31–51, 2001.
[5] J. Kohler, "Multilingual phone models for vocabulary-independent speech recognition tasks," *Speech Commun.*, vol. 35, pp. 21–30, 2001.
[6] U. Uebler, "Multilingual speech recognition in seven languages," *Speech Commun.*, vol. 35, pp. 53–69, 2001.
[7] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and T. Wang, "Toward language independent acoustic modeling," in *Proc. IEEE Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, 2000, pp. 1029–1032.
[8] O. Andersen, P. Dalsgaard, and W. Barry, "Data-driven identification of poly- and mono-phonemes for four European languages," in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Berlin, Germany, 1993, pp. 759–762.
[9] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. Int. Conf. Speech, Lang. Process. (ICSLP)*, Philadelphia, PA, 1996, pp. 2195–2198.
[10] J. J. Sooful and E. C. Botha, "An acoustic distance measure for automatic cross-language phoneme mapping," in *Proc. 12th Annu. Symp. South African Pattern Recognition Assoc.*, Franschhoek, South Africa, 2001.
[11] N. Srinivasamurthy and S. Narayanan, "Language-adaptive Persian speech recognition," in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Geneva, Switzerland, 2003.
[12] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 spoken term detection evaluation," NIST Tech. Rep., 2006.
[13] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, no. 2, pp. 151–170, Nov. 1996, Special Issue on Speech Under Stress.
[14] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 634–646, Mar. 2006.
[15] R. Huang, J. H. L. Hansen, and P. Angkititrakul, ""Dialect/accent classification using unrestricted audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 453–464, Feb. 2007.
[16] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: Whispered through shouted," in *Proc. ISCA Interspeech'07*, Antwerp, Belgium, Aug. 2007, pp. 2289–2292.
[17] D. Miller, M. Kleber, C. Kao, and O. Kimball, "Rapid and accurate spoken term detection," in *Proc. Interspeech*, Antwerp, Belgium, 2007.
[18] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar, "Turkish broadcast news transcription and retrieval," *IEEE Trans. Speech Audio Proc.*, vol. 17, no. 5, pp. 874–883, Jun. 2009.
[19] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 712–730, Sep. 2005.
[20] D. W. Oard, D. Soergel, D. S. Doermann, X. Huang, G. C. Murray, J. Wang, B. Ramabhadran, M. Franz, S. Gustman, J. Mayfield, L. Kharevych, and S. Strassel, "Building an information retrieval test collection for spontaneous conversational speech," in *Proc. Special Interest Group on Information Retrieval (SIGIR) Conf.*, Sheffield, U.K., 2004, pp. 41–48.
[21] D. A. James and S. J. Young, "A fast lattice-based approach to vocabulary independent word spotting," in *Proc. IEEE Conf. Acoust., Speech, Signal Process. (ICASSP)*, Istanbul, Turkey, 2000, pp. 1029–1032.
[22] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata—Application to spoken utterance retrieval," in *Proc. Human Lang. Technol.—North Amer. Chap. Assoc. Comput. Linguist. (HLT-NAACL) Conf.*, Boston, MA, 2004.
[23] L. M. Arslan and J. H. L.. Hansen, "Selective training in hidden Markov model recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 46–54, Jan. 1999.
[24] L. M. Arslan and J. H. L.. Hansen, "Likelihood decision boundary estimation between HMM pairs in speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 410–414, Jul. 1998.
[25] M. Witbrock and A. Hauptmann, "Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents," in *Proc. 2nd ACM Int. Conf. Digital Libraries*, Philadelphia, PA, 1997, pp. 30–35.
[26] G. J. F. Jones, J. T. Foote, K. S. Jones, and S. J. Young, "Retrieving spoken documents by combining multiple index sources," in *Proc. Special Interest Group on Information Retrieval (SIGIR) Conf.*, Zurich, 1996, pp. 30–38.
[27] K. Ng and V. W. Zue, "Subword-based approaches for spoken document retrieval," *Speech Commun.*, vol. 32, no. 3, pp. 157–186, 2000.
[28] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *Proc. Human Lang. Technol.—North Amer. Chap. Assoc. Comput. Linguist. (HLT-NAACL) Conf.*, Boston, MA, 2004.
[29] M. G. Brown, J. T. Foote, G. J. F. Jones, K. S. Jones, and S. J. Young, "Open-vocabulary speech indexing for voice and video mail retrieval," in *Proc. ACM Multimedia*, Boston, MA, 1996, pp. 307–316.
[30] A. Amir, A. Efrat, and S. Srinivasan, "Advances in phonetic word spotting," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, Atlanta, GA, 2001, pp. 580–582.
[31] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2000, pp. 81–87.
[32] J. Hieronymus, ASCII Phonetic Symbols for the World's Languages: Worldbet AT&T Bell Lab., Tech. Rep., 1994.
[33] *AT&T FSM Library*, [Online]. Available: http://www.research.att.com/sw/tools/fsm
[34] M. Mohri, F. C. N. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech, Lang.*, vol. 16, no. 1, pp. 69–88, 2002.
[35] *Linguistic Data Consortium* [Online]. Available: http://www.ldc.upenn.edu
[36] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, "A study of multilingual speech recognition," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 359–362.
[37] B. Logan and J. M. V. Thong, "Confusion-based query expansion for OOV words in spoken document retrieval," in *Proc. Int. Conf. Speech Lang. Process. (Interspeech/ICSLP)*, Denver, CO, 2002, pp. 1997–2000.
[38] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. S. Jones, "Effects of out of vocabulary words in spoken document retrieval," in *Proc. SIGIR*, Athens, Greece, 2000, pp. 372–374.
[39] M. Wechsler, E. Munteanu, and P. Schauble, "New techniques for open-vocabulary spoken document retrieval," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Melbourne, Australia, 1998, pp. 20–27.
[40] M. Wechsler, "Spoken Document Retrieval Based on Phoneme Recognition," Ph.D. dissertation, Swiss Federal Inst. of Technol. (ETH), Zurich, Switzerland, 1998.
[41] B. Logan, P. Moreno, and O. Deshmukh, "Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio," in *Proc. Human Lang. Technol. (HLT) Conf.*, 2002.
[42] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Int. Conf. Speech Lang. Process. (Interspeech/ICSLP)*, Denver, CO, 2002, pp. 901–904.

Murat Akbacak (S'02) received the B.S. degree in electrical engineering from Bogaziçi University, Istanbul, Turkey, in 2001, and the M.S. and Ph.D. degrees in electrical engineering from the University of Colorado, Boulder, in 2003 and 2009, respectively.

He has been working as a Research Engineer at Speech Technology and Research (STAR) Laboratory since 2007. From 2005 to 2007, he was a Staff Member with the Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson. Prior to that, from 2001 to 2005, he was a Research Assistant with the Robust Speech Processing Group, Center for Spoken Language Research, University of Colorado, Boulder. He was a visiting summer intern at the same group in 2000. He held a summer intern position at the Department of Human Language Technologies, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, in 2004. He was a part-time Research Engineer at GVZ Speech Technologies, Istanbul, Turkey, in 2001. His research interests include automatic speech recognition, spoken information retrieval, spoken language understanding, speech-to-speech translation, and digital speech processing.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, NJ, in 1982.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in the fall of 2005, where he is a Professor and Department Head of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as a Professor in the School of Behavioral and Brain Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado, Boulder, (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 50 (22 Ph.D., 28 M.S./M.A.) thesis candidates, was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body, author/coauthor of 352 journal and conference papers and eight textbooks in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), and lead author of the report "The impact of speech under 'stress' on military speech technology," (NATO RTO-TR-10, 2000).

Prof. Hansen was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise," in 2007 and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee (2009–2011; 2006–2008) and Educational Technical Committee (2006–2008; 2008–2010). Previously, he has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/2006), Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–2003). He has also served as guest editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the International Speech Communications Association (ISCA) Advisory Council. He also organized and served as General Chair for Interspeech-2002/ICSLP-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and served as Technical Program Chair and Co-Organizer for the IEEE ICASSP-2010 Conference in Dallas, TX, March 2010