ORIGINAL RESEARCH

Cross-language phoneme mapping for phonetic search keyword spotting in continuous speech of under-resourced languages

Ella Tetariy[1], Yossi Bar-Yosef[2], Vered Silber-Varod[1], Michal Gishri*[1], Ruthi Alon-Lavi[2], Vered Aharonson[1], Irit Opher[2], Ami Moyal[1]

[1] *Afeka Academic College of Engineering, Afeka Center for Language Processing, Tel Aviv, Israel*
[2] *NICE Systems Ltd., Ra'anana, Israel*

Received: March 29, 2015 **Accepted:** May 19, 2015 **Online Published:** June 25, 2015
DOI: 10.5430/air.v4n2p72 **URL:** http://dx.doi.org/10.5430/air.v4n2p72

Abstract

As automatic speech recognition-based applications become increasingly common in a wide variety of market segments, there is a growing need to support more languages. However, for many languages, the language resources needed to train speech recognition engines are either limited or completely non-existent, and the process of acquiring or constructing new language resources is both long and costly. This paper suggests a methodology that enables Phonetic Search Keyword Spotting to be implemented in a large speech database of any given under-resourced language using cross-language phoneme mappings to another language. The phoneme mapping enables a speech recognition engine from a sufficiently resourced and well-trained source language to be used for phoneme recognition in the new target language. The keyword search is then performed over a lattice of target language phonemes. Three cross-language phoneme mapping techniques are examined: knowledge-based, data-driven and phoneme recognition performance-based. The results suggest that Phonetic Search Keyword Spotting based on the cross-language phoneme mapping approach proposed herein can serve as a quick initial solution for validating keyword spotting applications in new, under-resourced languages.

Key Words: Cross-language phoneme mapping, Keyword spotting, Spoken term detection, Phonetic search

1 Introduction

Speech indexing and retrieval tools have become increasingly crucial in coping with the constant accumulation of massive amounts of digital audio and video data. In particular, speech recognition technology is frequently used in Keyword Spotting (KWS)-based applications to enable specific words to be identified out of a stream of continuous.[1] KWS-based applications, in turn, are often used by call centers and security-intelligence organizations for categorizing calls or searching speech databases, or by companies offering multi-media search applications on the internet or in enterprise markets. Such applications can be developed quickly for languages with sufficient available Language Resources (LRs). Supporting an under-resourced language, however, generally requires a long and costly preliminary process of collecting speech and text databases in order to train acoustic and language models, in addition to compiling a large vocabulary pronunciation lexicon. Yet, in spite of these challenges, there seems to be a growing demand for providing rapid support for under-resourced languages,

*Correspondence: Michal Gishri; Email: michalg@afeka.ac.il; Address: Afeka Academic College of Engineering, Afeka Center for Language Processing, 38 Mivtsa Kadesh St. Tel-Aviv, 6998812, Israel.

as evidenced by the emergence of international evaluations, such as the Open Keyword Search sponsored by IARPA and organized by NIST.[2]

The majority of KWS solutions employ Large Vocabulary Continuous Speech Recognition (LVCSR) engines. One of the main problems with an LVCSR-based system is that keyword searches are restricted by the vocabulary used in the transcription process, that is, the system cannot handle Out-of-Vocabulary (OOV) keywords.[3] Phonetic Search KWS (PS KWS) provides a solution for OOV keywords, as it performs the keyword search on a string of recognized phonemes, rather than words. Furthermore, because phonetic search is not dependent on a given vocabulary, the need for a word-based Language Model (LM) is eliminated. Thus, compared with LVCSR-based search, PS KWS is much less dependent on LRs.

The research presented here focuses on the rapid introduction PS KWS capabilities for a new language by using cross-language phoneme mapping techniques that do not rely heavily on LR availability. With this method, strong acoustic models from a well-resourced language can be used for phoneme recognition in a separate, under-resourced language. There are several approaches to cross-linguistic phoneme recognition. In one approach, acoustic models are produced from a large global inventory of phonemes constructed from the phoneme sets of several languages.[4–6] The premise of this method is that all languages have some phonemes in common, and that a large enough phoneme pool should be able to represent the acoustic models needed for any new language. In another approach, acoustic models are mapped from a single source language or from a small set of source languages to a new target language. This mapping is performed either by manual knowledge-based methods or by semi-automatic data-driven methods.[7] Naturally, the phonemes available from the source languages may not provide ideal coverage of the target language phoneme set; however, the need for LRs in the target language is greatly reduced or eliminated compared to standard KWS methods.

When only a small amount target language audio is accessible, acoustic adaptation techniques can be applied on the available acoustic models from the source language[8] or bootstrapping techniques that use well-trained models from several source languages to generate unsupervised transcriptions for training acoustic models in the target language.[9] Both these methods require some target LRs and have been attempted with continuous speech recognition using both small and large vocabularies and with Language Identification (LI), but not with PS KWS.

The research goal was to consolidate a methodology for supporting PS KWS in a target language with few or no LRs available. The method used employs phoneme mappings between a source language with adequate LRs for training acoustic models and an under-resourced target language. The source language acoustic models are used to produce a string of recognized phonemes in the target language. Then, PS is performed on the resulting phoneme strings in order to locate keywords from the target language. Neither a full set of LRs nor dedicated acoustical models are required in the target language.

1.1 Keyword spotting overview

Most KWS procedures are carried out using one of the three following methods:[10]

- LVCSR-based KWS: where an LVCSR engine produces a transcription of the entire speech database, and the KWS-based application searches the resulting text for the designated keywords.
- Acoustic KWS: where the KWS engine operates on the speech signal and the recognition vocabulary consists only of the designated keywords, represented as sequences of phonemes.
- Phonetic Search (PS) KWS: where a phoneme recognition engine produces a phonetic representation of the entire speech database, and a phonetic search engine searches the resulting phoneme sequence or lattice for the designated keywords.

Each of these methods has benefits in comparison to the others under different circumstances.[10–14] For example, when searching large speech databases, rapid search capabilities are essential. LVCSR and PS methods are applicable to such settings because they perform a one-time transformation of the speech into a textual representation and then index the engine output (as words or as phonemes, respectively), thereby facilitating a quick search process. In the case of acoustic KWS, in contrast, it is necessary to re-run the audio for each new search list. This makes acoustic KWS irrelevant for most modern applications, which generally deal with large amounts of audio data. Indeed, most research in the field today focuses on LVCSR and PS methods.

Comparing the two main KWS methods, LVCSR is at a disadvantage compared with PS when it comes to keyword flexibility.[15–17] Such flexibility is crucial for KWS-based applications that deal with a constant flow of new data and frequent changes in search terms. Often, these search terms are names of people or places, which are in many cases of foreign origin. These types of keywords are not necessarily part of an LVCSR recognition vocabulary (*i.e.*, OOV words). To accommodate such keywords, it is necessary to re-run the LVCSR recognition engine with an updated recognition vocabulary for each new OOV keyword. With PS KWS-based applications, in contrast, it is possible to search for any term, provided the phonetic transcription is available, thus eliminating the OOV problem. Although the PS method offers users total freedom in changing the designated keywords, since the textual transformation of the

speech into phonemes is not restricted by a vocabulary, this flexibility comes at a cost of higher computational complexity in the search phase and often a drastic decrease in performance for in-vocabulary words, compared with the LVCSR method.

In order to resolve the tradeoff between flexibility and performance, much recent research has focused on hybrid systems that combine the two approaches. Some merge LVCSR and PS KWS engine results in a combined lattice or in post-processing procedures,[18–21] while others use hybrid LMs that blend in-vocabulary words and sub-word units (phones, triphones, fragments) into a unified LM.[21–23] The hybrid LM enables the system to produce phoneme strings in place of the OOV words, making them accessible in search results. Bulyko *et al.*[24] performed sub-word recognition alone without the use of a word LM.

A basic assumption of cross-linguistic systems is that few or no LRs may be available in the target language. Thus, the focus of this research is on PS KWS, which, compared with LVCSR, is much less dependent on linguistic constraints and, unlike LVCSR, requires LM training at the phone-level, but not at the word-level. In order to compensate for the possibility that phone-level LM is also unobtainable in the target language, the option of using a source-language phone LM is also explored.

PS is also the most suitable KWS method for applying phoneme mappings between languages because it employs a fuzzy search mechanism that can compensate for inaccurate mappings.

1.2 Keyword spotting in under-resourced languages

Owing to international evaluations such as those sponsored by the DARPA RATS and IARPA BABEL programs, it is now commonly recognized that current KWS solutions provide acceptable results for well-resourced languages recorded under relatively sterile conditions, but that they fall short under real-life conditions with limited language resources. For target languages that are rare or spoken only in regions where collection is difficult or even impossible, LRs can be scarce or non-existent, imposing major, often irresolvable, constraints on training acoustic models in these languages. In dealing with under-resourced or zero-resource languages, research is focused on finding robust KWS solutions using techniques such as subspace-GMM acoustic modeling,[25] multiple system combination and score normalization,[26] and bootstrapping techniques utilizing multi-language acoustic models and neural networks.[9,27,28]

When adapting a PS KWS system to process data in an under-resourced language, it is possible to bypass the long and costly training process by utilizing phoneme acoustic models from accessible and well-trained languages. Specifically, it is possible to incorporate a cross-language phonetic mapping between the target and source language phonemes either prior to the phoneme recognition stage or during the phonetic search stage—in the latter case, the mapping is based on the phoneme sequence or lattice generated during the recognition stage.

Two major issues to consider prior to performing cross-language KWS are what languages to use for the source acoustic models and what type of mapping scheme to employ. Source acoustic models can originate from either a single language[29–32] or multiple languages.[27,33–36] When it comes to selecting the mapping scheme, one major factor to consider is how much speech data is actually available in the target language. When no target language speech data are available, a knowledge-based mapping can be generated. Knowledge-based mappings are produced manually and take into account known phonetic similarities between the source and target language phonemes,[29,31,32] as defined, for example, by the International Phonetic Alphabet (IPA).[37] A knowledge-based mapping can be avoided if phonetic transcriptions of all source and target languages use the same phoneme set (*e.g.* SAMPA) and source language phoneme provide coverage of target language phonemes.[34] When a limited amount of target language speech is available (a few minutes to a few hours), data-driven methods can be used. Data-driven methods include producing mappings based on acoustic distance measurements between source and target phonemes, acoustical model adaptation methods such as maximum likelihood linear regression (MLLR), which can be based on an initial knowledge-based mapping or unsupervised techniques employing DNNs for example.[34]

This paper focuses on the use of acoustical models from a single source language using three different mapping methods. The first is a knowledge-based mapping performed by a linguist. The second is a data-driven mapping that uses a small quantity of audio to train coarse acoustic models in the target language. The distance between the coarse acoustic models in the target language and well-trained acoustic models from the source language is calculated, and the best-matched mapping is generated using the distance matrix. In the third method, source language acoustic models are used to recognize a small amount of data in the target language, after which the recognition statistics produced are leveraged in order to automatically generate a mapping between the languages. In all three mapping paradigms, the source language acoustic models that are used for recognizing speech in the target language remain unaffected. Only their labels are changed to reflect the target language phoneme set.

Phonetic search is particularly suitable for such a method for several reasons: 1) The phoneme lattice produced represent the acoustic content of the target language speech, even when acoustics models from another language are used; 2) The search is performed through a series of soft decisions depending on likelihoods and can easily take mapping costs into account; and 3) Large amount of textual data in the tar-

get language is not needed for producing a word-level LM which is irrelevant for PS.

2 Methods

For the cross-language PS evaluations, either American English (English) or Levantine Arabic (Arabic) were used as the source language and Spanish was used as the target language.

2.1 Phonetic search keyword spotting

To perform cross-language PS KWS, we implemented a system consisting of the following two central components:

- A Phoneme Recognition Engine: Phoneme recognition was performed using acoustic models of English or Arabic as source languages with several options for a phoneme-level LM: ergodic (equal transition probabilities), target LM or source LM.
- A Phonetic Search Engine: Phonetic search was performed over the resulting source language phoneme lattice while employing a mapping scheme between the source language used and the Spanish phonemes. The Levenshtein Distance measure[38,39] was used for sequence matching, where all hypotheses with a distance lower than a pre-defined threshold were declared as recognized keywords.

A block diagram of the PS KWS system using cross-language mapping is shown in Figure 1:

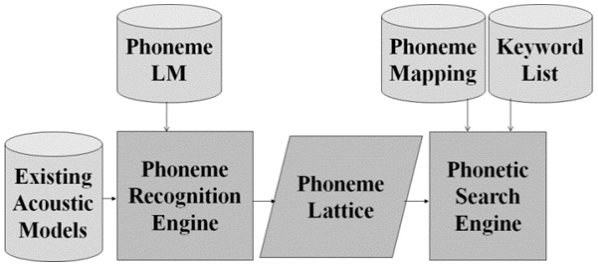

Figure 1: Cross-Language Phonetic Search KWS System

The two parts of the system (phoneme recognition and phonetic search) are designed to function independently. The phoneme recognition stage uses the source language acoustic models and phoneme LM as input, and the phonetic search stage uses the phonetic lattice produced in the phoneme recognition stage and any given phonetic mapping (represented by a mapping matrix) between source and target phonemes.

2.2 Cross-language phoneme mapping for phonetic search

This paper introduces three mapping paradigms: knowledge-based, data-driven and performance-based. The knowledge-based mapping requires no speech data in the target language, whereas the data-driven and performance-based mappings require a small amount of data. In order to evaluate the quality of each of the mapping schemes, phoneme recognition results in Spanish were produced using well-trained Spanish acoustic and language models as a base-line for comparison. The topology used to train the Spanish models was the same as that used to train the source models.

2.2.1 Knowledge-based phonetic mapping

The knowledge-based mapping procedure was designed to compensate for the gap between the source language phoneme set and the target language phoneme set. Take, for example, the English to Spanish mapping. The Spanish vowel set consists of the five cardinal vowels: /a/, /e/, /i/, /o/ and /u/. These vowels are shared by the English sound system which contains a much broader set of vowels. However, one-to-one mappings may not be precise. For the knowledge-based mapping, all Spanish phonemes were mapped to the closest-known counterpart in the source language. So, for example, the Spanish vowel /a/, as in [paDres] "parents" was mapped to /aa/, as in [paad] "pod" in English, even though the Spanish phoneme is pronounced closer to the front of the mouth. A second mapping employed a one-to-many technique, allowing each target phoneme to be mapped to more than one source phoneme. In this case, the Spanish /a/ was mapped not only to the English /aa/, but also to /ae/, which is more fronted than /aa/. In the mappings based on the latter one-to-many approach, some Spanish phonemes were mapped to multiple phonemes in the source language, whereas others were mapped to only one. Table 1 shows some examples of the mappings between Spanish and English.

Table 1: Sample of one-to-many mapping

Spanish	<>	English
a	<>	aa, ae
i	<>	iy, ih
m	<>	m
tS	<>	ch, sh

2.2.2 Data-driven phonetic mapping

For the data-driven mapping, one hour of speech data in Spanish was used to train coarse target-language acoustic models. Naturally, these models were not sufficient for robust recognition, but they were sufficient to calculate the acoustic distance between source and target acoustic models. This process resulted in a data-driven mapping between the source and target phonemes. Previous studies suggested several Distance Measures (DMs) between GMMs. The DMs used in the evaluation follow those suggested by Sooful and Botha,[32] and include: Kullback-Leibler, Bhattacharyya, Mahalanobis, Euclidean, and Jeffreys-Matusita.

Acoustic models were trained using a large amount of data available in each of the source languages and only one hour of audio data in Spanish. The distance measures between the source and target acoustic models were then calculated using only one mixture per state (multi-mixture models were employed for recognition using the source language acoustic models), and a distance matrix was produced from each DM calculation. Each distance matrix was transformed into a mapping matrix, where each matrix element represented the similarity between a phoneme in the source language and a phoneme in the target language.

2.2.3 Phoneme recognition performance-based mapping

A performance-based mapping was created to improve the accuracy of the knowledge-based or data-driven mappings. One hour of Spanish speech and corresponding word-level transcriptions (no time-alignment was required) and a pronunciation lexicon were used to produce the mapping. The recognized phoneme sequence (obtained from the best path in the lattice), using the source language acoustic models (and either the knowledge-based or data-driven mapping), was compared to the correct phoneme sequence (obtained by aligning the orthography with the lexicon transcriptions). Then, a learning mechanism that utilizes the resulting confusion matrix was developed. The confusion matrix reflects the probability of identifying a certain source-phoneme given that a certain target-phoneme was actually pronounced.

This mechanism estimates $p(s_i|t_j)$ for each of the phonemes in both languages, where s_i represents a source language phoneme, and t_j a target language phoneme. A dynamic-programming algorithm was then employed to achieve the best alignment between the two phoneme sequences—the first sequence being the lexical transcription and the second sequence being the recognized sequence mapped into the target language. At this stage, the recognized phoneme sequence resulting from the knowledge-based mapping was used as a bootstrap to the learning process.

Preliminary experiments indicated that learning the empirical phoneme mapping by applying this standard process, leads to a significant degradation in PS KWS performance. A deeper inspection revealed that taking acoustic mismatches and recognition errors into account could improve the alignment mechanism, which performed poorly with the knowledge-based mapping. An additional bootstrapping method was thus applied to the mapping in order to account for acoustic variations detected in the development set, as well as additional a-priori anticipated phoneme recognition errors. Furthermore, a more robust phoneme-to-phoneme mapping that reduces the impact of phoneme confusions between phonemes belonging to the same major phonetic natural class (*e.g.*, plosives or fricatives) was employed.

In the resulting bootstrapped mapping, less-probable confusions were given lower weights. For example, the mapping of the Spanish /b/ into three Arabic phonemes incorporated different weighting for each Arabic phoneme, as presented in Table 2:

Table 2: Sample of one-to-many mapping

Spanish	<$_{\text{weight}}$>	Arabic
b	<1.0>	b
b	<0.3>	p
b	<0.1>	d, k, t, g

The mapping of /b/ to /b/ is conventional and the mapping of /b/ to /p/ takes into account probable recognition errors. The mapping of /b/ to other plosives is reflective of errors anticipated by acoustic phonetic natural class theory. This approach was able to "fix" the alignment of the series and enabled a robust statistical process of learning the mappings between target and source phonemes.

2.2.4 Phoneme-level language model

To improve phoneme recognition results, it was also necessary to use a phoneme-level LM that assigns probabilities to potential phoneme sequences. Ideally, the phoneme LM should be in the target language; however, estimation of such an LM requires a representative text database with its phonetic transcription. If such a database is unavailable in the target language, it is possible to use a small amount of transcribed speech data and a phonetic lexicon in the target language, or use a source language LM. Of course, it is also possible not to use a LM at all, but this is expected to yield inferior phoneme recognition results. In order to evaluate the impact of the LM on cross-language phoneme recognition, several LM topologies were examined.

3 Evaluation

The three phonetic-mapping techniques, knowledge-based, data-driven and performance-based, were each examined separately. Furthermore, for each mapping paradigm, we evaluated the influence of using various phoneme LMs in the phoneme recognition stage. As noted, the source languages used were English and Arabic, while the target language was Spanish. All evaluations were performed on both phoneme recognition and PS KWS.

A base-line reference was produced using fully-trained Spanish acoustic and language models. To eliminate any bias resulting from Spanish-specific traits, PS KWS experiments were later performed on Russian, using the same two source languages. All mapping schemes provided full coverage of the target language phoneme set.

The following summarizes the experiments performed:

(1) Cross-language phoneme recognition evaluation

- Base-line Spanish phoneme recognition results using Spanish acoustic models for comparison to the cross-language recognition.
- Cross-language phoneme recognition using acoustic models from either English or Arabic to recognize phonemes in Spanish. For this test a phoneme-to-phoneme mapping between the source and target phoneme sets was first defined. Various mapping schemes were implemented.

(2) PS KWS evaluation

- Baseline Spanish KWS results using Spanish acoustic and language models for comparison to the cross-language recognition.
- PS over phoneme lattices generated by the source language acoustic and language models using the knowledge-based mapping into Spanish.
- PS using various cross-language phonetic mapping techniques during the search process.
- Experiments using the knowledge-based mapping on three LM configurations: 1) source acoustic models and an ergodic topology (no LM); 2) source acoustic models with a phoneme LM estimated from the source language; 3) source acoustic models and a target language phoneme LM estimated from a large-scale lexicon and textual database. These experiments were intended to test the influence of the LM type in the phoneme recognition stage.
- Experiments using an additional target language – Russian.

3.1 Speech databases

The English acoustic models were trained using 157 hours from the Wall Street Journal portion of the Macrophone database.[40] The Arabic acoustic models were trained using a total of 115 hours from the Appen Levantine Arabic Conversational Telephone Speech database[41] and the LDC Fisher Levantine Arabic Conversational Telephone Speech database.[42]

The experimental test sets for Spanish and Russian included one hour of speech for each language from the ELRA SpeechDat(II) FDB-4000 database[43] and Appen's Russian Conversational Telephony database[44] respectively. The development database used for estimating each of the resulting confusion matrices contained an additional hour of speech for each language, extracted from the same databases. The remaining audio in the Spanish database (173 hours) was used to generate the Spanish base-line reference results that required well-trained acoustical and language models.

The phoneme sets used for each language were as follows: 39 English DARPA phonemes;[45] 43 Arabic Buckwalter transliteration based (`http://www.qamus.org/translit eration.htm`); 31 Spanish SAMPA phonemes;[46] 49 Russian SAMPA phonemes.[47]

Searches were performed on lists of keywords of three syllables or more. The keyword lists in Spanish consisted of 124 search terms, with average length of 9.2 phonemes each. The Russian keyword list consisted of 25 search terms, with an average length of 9.6 phonemes each.

3.2 Experimental setup

Phoneme recognition was performed using the HTK speech recognition engine. Feature extraction was MFCC of order 39 with first- and second derivatives. The acoustic models were three state tri-phone HMMs with additional models for speaker noises and non-speech events. The phonetic search process was performed over a phoneme lattice following implementation of one of the various mapping schemes. The Levenshtein Distance was used to measure the distance between the keywords and partial phoneme sequences on the lattice. A given phoneme sequence was considered to match the keyword if the distance between them was below a predefined threshold.

3.3 Scoring paradigms

Phoneme recognition and PS KWS were carried out on the Spanish and Russian test sets. Phoneme recognition performance was measured using a Phoneme label correct rate (%Correct), estimated based on the Levenshtein distance between the phoneme recognition results and the reference pronunciation and using the following calculation:

$$\%Corect\ labels = \frac{number\ of\ correct\ labels}{total\ number\ of\ labels} \times 100 \quad (1)$$

As for PS KWS results, the Detection Rate (DR) and False Alarm Rate (FAR) were estimated for various values of a decision threshold (θ). The calculations of DR and FAR are given by (1) and (2) respectively:

$$DR(\theta) = \frac{1}{K}\sum_{k=1}^{K}\frac{N_{Detect}(k,\theta)}{N_{True}(k)} \quad (2)$$

$$FAR(\theta) = \frac{1}{Q}\sum_{k=1}^{Q}\frac{N_{FA}(k,\theta)}{T_{Speech}} \quad (3)$$

where:

Q = total of keywords;

K = of keywords with 1 or more reference occurrences;

N_{Detect} (k,θ) = of detections of keyword k using threshold θ (calculated only for keywords with a reference);

N_{FA} (k,θ) = of false alarms of keyword k (calculated for any keyword);

N_{True} (k) = of reference occurrences of keyword k;

T_{Speech}= the total duration of evaluated speech in the test data (in hours).

In a parallel analysis we also provide the Maximum Term-Weighted Value (MTWV), a metric officially used by NIST for the OpenKWS evaluations. The term-weighted value (TWV) is 1 minus the weighted sum of the term-weighted probability of missed detections $P_Miss(\theta)$ and the term-weighted probability of false alarms PFA(θ).

$$TWV(\theta) = 1 - [P_{Miss}(\theta) + \beta \cdot PFA(\theta)] \quad (4)$$

To compute the TWV, we assume β=999.9 (NIST, 2014). The MTWV is the maximum possible TWV over all possible threshold values, θ.

4 Results

The main purpose was to evaluate PS KWS performance under different cross-language mappings. Thus, the KWS evaluation was the primary evaluation. However, phoneme recognition produces the phoneme sequence on which the PS KWS is performed and thus has a strong impact on the PS KWS results. Consequently, before carrying out the KWS experiments, we performed phoneme recognition experiments in order to evaluate the phoneme mapping. In a consecutive step, the results of the PS KWS are presented. Since, phonetic-mapping evaluations for both source languages (Arabic and English) produced similar results, for the sake of brevity, only the results obtained using Arabic as the source language are presented (for both Spanish and Russian as targets). The concluding results, which integrate the target language LM, were carried out in English and are therefore presented last (see Figures 5 and 6).

4.1 Phoneme recognition

Table 3 shows the phoneme recognition results obtained using the knowledge-based mapping and each of the data-driven mappings that resulted from the various DMs tested. These can be compared to the base-line Spanish results. The discrepancy in the number of mapped phonemes is related to the one-to-many phoneme mapping between the target and source phonemes. All mappings schemes fully cover the target phoneme set, but not all source phonemes are utilized.

Although none of the mappings produced results that approached the base-line Spanish results, the results obtained using data-driven mapping were similar to those obtained using knowledge-based mapping. The best recognition results were achieved using the Euclidean distance measure and Arabic source models. However, using the Kullback-Leibler measure produced the highest average recognition rate (averaged over the two source languages).

Table 3: Cross-language Phoneme Recognition Evaluation Results – Percentage of Correct Phoneme Labels

Mapping Technique	Source Language	#Mapped Phonemes	%Correct Labels
Base-line Spanish	Spanish	31	68.42%
Knowledge-based	English	27	39.85%
	Arabic	24	55.63%
Data-driven	English	24	44.03%
Kullback-Leibler	Arabic	21	55.03%
Data-driven	English	31	37.65%
Bhattacharyya	Arabic	21	42.14%
Data-driven	English	23	39.48%
Mahalanobis	Arabic	23	54.88%
Data-driven	English	26	37.99%
Euclidean	Arabic	26	56.71%
Data-driven	English	31	38.21%
Jeffreys-Matusita	Arabic	31	40.41%

4.2 Keyword spotting performance

KWS performance under various conditions is represented in Figures 2-6, in which the DR is plotted as a function of the FAR. Figure 2 shows KWS performance on the Spanish test set, using lattices generated by the English and Arabic source acoustic and language models, with knowledge-based mapping. The Spanish base-line results (PS KWS in Spanish using well-trained Spanish acoustic and language models) are also presented for comparison. The results show that, although significantly lower than the performance obtained in the Spanish base-line experiment, the performance of Spanish KWS using Arabic as a source language was quite similar to that obtained using English as a source language.

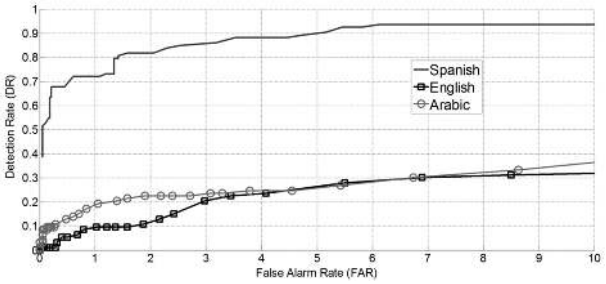

Figure 2: Spanish KWS Evaluation Results: English and Arabic Source Models Using Knowledge-Based Mapping

This is surprising, given the fact that Arabic performed better in the phoneme recognition resulting from the knowledge-based mapping (see Table 3). In fact, Figure 2 shows that the KWS performance resulting from the Arabic-Spanish mapping was indeed superior to that resulting from the English-Spanish mapping up to a FAR of approximately 5. At this working point, it seems that the DR reached a maximum regardless of the language used as the source.

Figure 3 compares Spanish PS KWS performance using Arabic acoustic models and the various mapping

schemes: (a) knowledge-based mapping, (b) data-driven mapping, (c) performance-based mapping initialized from the knowledge-based mapping, and (d) performance-based mapping initialized from the data-driven mapping. All configurations used a source language phoneme-level LM. The MTWV results for the same Arabic source conditions are given in Table 4.

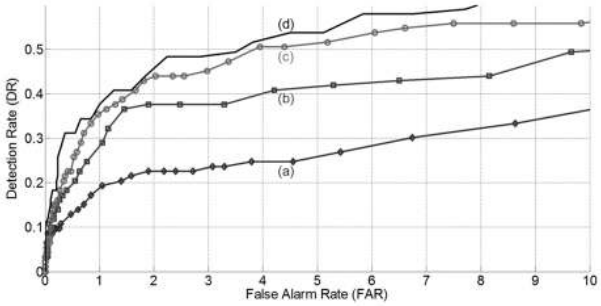

Figure 3: Spanish KWS Evaluation Results: Arabic Source Models Using Different Mapping Schemes

Table 4: Arabic Source MTWV for Spanish Target

Arabic source	(a)	(b)	(c)	(d)
MTWV	0.071	0.088	0.115	0.210

The results suggest that data-driven mapping yields better PS KWS performance than the knowledge-based mapping (assuming a small development set is available for the target language) than does knowledge-based mapping, and that further refinement of either mapping using a phoneme recognition performance-based learning approach improves the KWS results substantially.

The same methodology and source languages were tested on the Russian data, yielding similar results. Figure 4 compares Russian PS KWS performance using Arabic acoustic models and the various mapping schemes: (a) knowledge-based mapping, (b) data-driven mapping, (c) performance-based mapping initialized from the knowledge-based mapping and (d) performance-based mapping initialized from the data-driven mapping. All configurations used a source language phoneme-level LM. The MTWV for the same Arabic source conditions is given in Table 5. Results indicate that the Arabic models do not provide an adequate representation of the Russian phonemes. However, some statistical performance-based mapping learning may alleviate the problem if the user agrees to accept larger numbers of false alarms.

Table 5: Arabic Source MTWV for Russian Target

Arabic source	(a)	(b)	(c)	(d)
MTWV	0.0	0.005	0.045	0.015

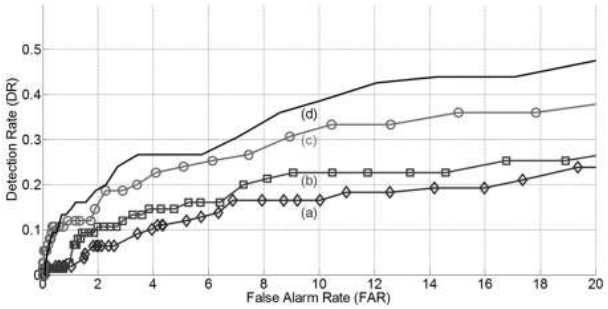

Figure 4: Russian KWS Evaluation Results: Arabic Source Models Using Different Mapping Schemes

Figure 5 presents the results obtained for each of the three LM options used in the recognition phase, under knowledge-based mapping, with English as the source language and Spanish as the target language. Results are obtained on the one-best recognized phonetic path: (a) no LM, (b) English LM and (c) Spanish LM. These are compared to the base-line Spanish results (Spanish KWS using Spanish acoustic models), specifically, (d) the Spanish base-line with no LM, and (e) the Spanish base-line using a Spanish LM. The corresponding MTWV results are provided in Table 6.

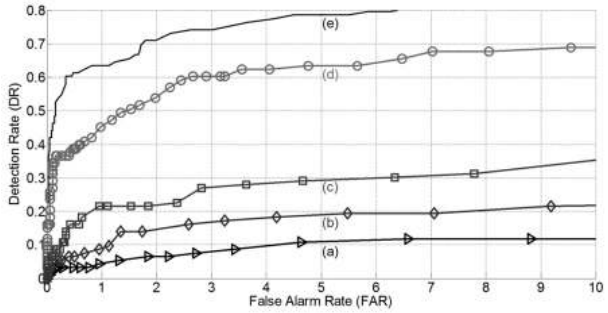

Figure 5: Spanish KWS Evaluation Results: English Source Models; Knowledge-Based Mapping, Search Performed over One-Best Phoneme Recognition Results using Various LM Schemes

Table 6: English Source MTWV for Spanish Target Compared to Spanish Base-Line

English source	(a)	(b)	(c)	(d)	(e)
MTWV	0.011	0.015	0.045	0.51	0.51

Again, the performance using cross-language mappings was considerably poorer than the Spanish base-line performance results. However, among the cross-language mapping schemes, using English acoustic models in combination with a Spanish phoneme-level LM produced best results. This clearly indicates that using a target language LM can greatly improve performance. More importantly, even using a phone-level LM from the source language is superior to using no LM when performing cross-language PS.

Figure 6 shows results obtained for several LM options, but this time with different target-to-source phoneme mapping schemes, and this time, the search is performed on a phonetic lattice: (a) knowledge-based mapping and no LM, (b) knowledge-based mapping and Spanish LM, (c) performance-based mapping and no LM, and (d) performance-based mapping and Spanish LM. The performance-based mapping used was initialized from the knowledge-based mapping and enriched by the natural class groupings. These results can be compared to the Spanish base-line using a (e) Spanish LM and (f) Spanish LM and performance-based mapping between Spanish phonemes (automatic mapping of Spanish phonemes to Spanish phonemes). The corresponding MTWV results are provided in Table 7.

The results also show that a monolingual base-line system can also benefit from a performance-based mapping during the search, although the improvements for the cross-language configurations are more substantial. Moreover, a performance-based mapping is more effective in improving performance than integrating a LM in the target language. When examining the phoneme recognition confusion matrices, it was observed that each of the target phonemes was mapped to numerous source phonemes with substantial probability, producing much more smeared matrices in comparison to the monolingual configuration. Hence, utilizing this information for the PS led to a significant improvement.

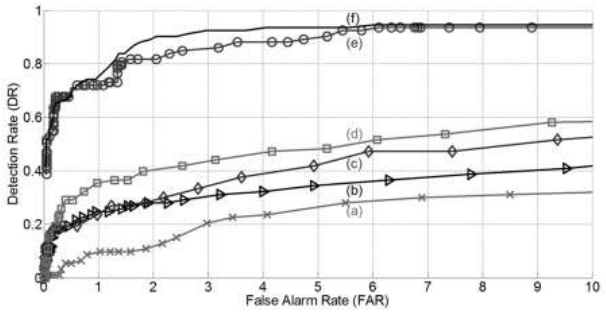

Figure 6: Spanish KWS Evaluation Results: English Source Models, Search over Phoneme Recognition lattice using Various Mapping and LM Schemes

Table 7: English source MTWV for Spanish target compared to Spanish Base-Line

English Source	(a)	(b)	(c)	(d)	(e)	(f)
MTWV	0.011	0.121	0.132	0.181	0.61	0.63

5 Conclusions

The research presented here proposes a methodology for implementing PS KWS in under-resourced languages without the need to train new acoustic models. Three cross-language mapping techniques and phonetic language model configurations were examined. Whereas previous works on cross-language phoneme mapping have mostly concentrated on LVCSR, the research described here focused on PS KWS. The best results were obtained using a phoneme recognition performance-based mapping initialized by a previously learned data-driven mapping. This topology significantly improved target language KWS performance in comparison to the knowledge-based or data-driven mappings alone. Although these methods still yielded substantially poorer performance compared with fully-trained target language used for the base-line results, they were relatively reasonable when accounting for the fact that they were attained with very limited use of language resources in the target language.

Incorporation of a phoneme-level LM in the recognition phase enhanced KWS performance substantially. Our results suggest that, if obtainable, a language model estimated from target language data produces the best results; however, importantly, even a language model estimated from the source language is superior to using an ergodic, non-restrictive topology.

The approach presented provides a rapid means of supporting new languages with very limited resources (both language and human resources) and at virtually no cost. This initial, rapid, low-cost version of a KWS application can be later upgraded to incorporate updated acoustic and language models estimated from target language data that are logged from the application.

Future research directions include combining searches in two (or more) source lattices, as well as using data-driven methods to correctly span the acoustic space using several source languages.

Acknowledgements

This work was supported by grant #45828 provided by the Chief Scientist of the Israeli Ministry of Economy as part of the Magneton program which encourages the transfer of knowledge from academic institutions to industrial companies.

References

[1] Wilpon JG, Rabiner LR, Lee C, *et al*. Automatic recognition of keywords in unconstrained speech using hidden Markov models. IEEE Transactions on Acoustic Speech Signal Processing. 1990; 38 (11): 1870-1878. http://dx.doi.org/10.1109/29.103088

[2] Harper MP. IARPA Babel Program; 2013. Available from: www.iarpa.gov/Programs/ia/Babel/babel.html

[3] Manos AS, Zue VW. A segment-based wordspotter Using phonetic filler models. Proceedings of the IEEE International Conference on

Acoustics, Speech, and Signal Processing (ICASSP); 1997. 21-24 Apr; Munich, Germany.

[4] Gokcen S, Gokcen JM. A multilingual phoneme and model set: Toward a universal base for automatic speech recognition. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU); 1997 8-12 Dec; Olomouc, Czech Republic.

[5] Schultz T. Globalphone: a multilingual speech and text database developed at Karlsruhe University. Proceedings of the 7th International Conference on Spoken Language Processing (INTERSPEECH – ICSLP); 2002 16-20 Sept; Denver, Colorado.

[6] Schultz T, Waibel A. Fast bootstrapping of LVCSR systems with multilingual phoneme sets. Proceedings of the Fifth European Conference on Speech Communication and Technology (EUROSPEECH); 1997 22-25 Sept; Rhodes, Greece.

[7] Wheatley B, Muthusamy Y, Kondo K, *et al*. An evaluation of cross-language adaptation for rapid HMM development in a new language. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP); 1994; Adelaide, Australia.

[8] Fung P, Ma CY, Liu WK. MAP-based cross-language adaptation augmented by linguistic knowledge: from English to Chinese. Proceedings of Sixth European Conference on Speech Communication and Technology (EUROSPEECH); 1999 5-9 Sept; Budapest, Hungary.

[9] Vu NT, Kraus F, Schultz T. Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training. Proceedings of 12th Annual Conference of the International Speech Communication Association (INTERSPEECH); 2011 27-31 Aug; Florence, Italy.

[10] Szöke I, Schwarz P, Matejka P, *et al*. Comparison of keyword spotting approaches for informal continuous speech. Proceedings of the 9th European Conference on Speech Communication and Technology (EUROSPEECH); 2005 4-8 Sept; Lisbon, Portugal.

[11] Shen W, White CM, Hazen TJ. A comparison of query by-example methods for spoken term detection. Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH); 6-10 Sept 2009; Brighton, United Kingdom.

[12] Šmídl L, Psutka J. Comparison of keyword spotting methods for searching in speech. Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH – ICSLP); 2006 17-21 Sept; Pittsburgh, Pennsylvania.

[13] Wang D, Tejedor J, Frankel J, *et al*. A comparison of phone and grapheme-based spoken term detection. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2008 30 Mar – 4 Apr; Las Vegas, Nevada.

[14] Moyal A, Aharonson V, Gishri M, *et al*. Phonetic Search Methods for Large Speech Databases; 2013; Springer, New York.

[15] Burget L, Černocký J, Fapšo M, *et al*. Indexing and search methods for spoken document. Text, Speech and Dialogue, 4188/2006 of Lecture Notes in Computer Science; 2006. p. 351–358.

[16] Cardillo PS, Clements M, Miller MS. Phonetic searching vs. LVCSR: How to find what you really want in audio archives. International Journal of Speech Technology. 2002; 5 (1): 9-22. `http://dx.doi.org/10.1023/A:1013670312989`

[17] Wallace R, Vogt R, Sridharan S. A phonetic search approach to the 2006 NIST Spoken Term Detection Evaluation. Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH); 27-31 Aug 2007; Antwerp, Belgium.

[18] Akbacak M, Vergyri D, Stolcke A. Open-vocabulary spoken term detection using graphone-based hybrid recognition systems. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2008 30 Mar – 4 Apr; Las Vegas, Nevada.

[19] Miller D, Kleber M, Kao CH, *et al*. Rapid and accurate spoken term detection. Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH); 2007 27-31 Aug; Antwerp, Belgium.

[20] Rastrow A, Sethy A, Ramabhadran B, *et al*. Towards using hybrid word and fragment units for vocabulary independent LVCSR systems. Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH); 6-10 Sept 2009; Brighton, United Kingdom.

[21] Yu P, Seide F. A hybrid word / phoneme-based approach for improved vocabulary-independent search in spontaneous speech. Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH – ICSLP); 2004 4-8 Oct; Jeju Island, Korea.

[22] Szöke I, Fapšo M, Burget L, *et al*. Hybrid word-subword decoding for spoken term detection. Paper presented at the Speech Search Workshop at SIGIR (SSCS); 2008 20-24 Jul; Singapore.

[23] Yazgan A, Saraclar M. Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2004 17-21 May; Montreal, Canada.

[24] Bulyko I, Herrero J, Mihelich C, *et al*. Subword speech recognition for detection of unseen words. Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association (INTERSPEECH); 2012 9-13 Sept; Portland, Oregon, USA.

[25] Zhang X, Demuynck K, Van Compernolle D, *et al*. Subspace-GMM acoustic models for under-resourced languages: Feasibility study. Proceedings of the Third Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU); 2012 7-9 May; Cape Town, South Africa.

[26] Mamou J, Cui J, Cui X, *et al*. System combination and score normalization for spoken term detection. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2013.

[27] Schultz T, Waibel A. Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Communication. 2001; 35 (1): 31-51.

[28] Vu NT, Metze F, Schultz T. Multilingual bottle-neck features and its application for under-resourced languages. Proceedings of the Third Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU); 2012 7-9 May; Cape Town, South Africa.

[29] Le VB, Besacier L. First steps in fast acoustic modeling for a new target language: Application to Vietnamese. Proceedings of the 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2005 Mar 18-23; Philidelphia, Pennsylvania.

[30] Nieuwoudt C. Cross-Language Acoustic Adaptation for Automatic Speech Recognition [Dissertation]. [Pretoria] University of Pretoria; 2000.

[31] Nieuwoudt C, Botha EC. Cross-language use of acoustic information for automatic speech recognition. Speech Communication. 2002; 38 (1): 101-113.

[32] Sooful JJ, Botha EC. An acoustic distance measure for automatic cross-language phoneme mapping. Proceedings of the Twelfth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2001); 2001; Stellenbosch.

[33] Kienappel AK, Geller D, Bippus R. Cross-language transfer of multilingual phoneme models. Paper presented at: the ASR2000-Automatic Speech Recognition: Challenges for the new Millennium ISCA Tutor. and Res. Workshop (ITRW). 2000 Sept 18-20; Paris, France.

[34] Knill KM, Gales MJF, Ragni A, *et al*. Language independent and unsupervised acoustic models for speech recognition and keyword spotting. Proceedings of InterSpeech 2014; 2014 Sept 14-18; Singapore.

[35] Liu C, Melnar L. Training acoustic Models with speech data from different languages. Paper presented at: the ISCA Workshop on Multilingual Speech and Language Processing (MULTILING 2006). 2006 April 9-11; Stellenbosch, South Africa.

[36] Žgank A, Kačič Z, Vicsi K, *et al*. Crosslingual transfer of source acoustic models to two different target languages. Paper presented at: the COST278 and ISCA Tutor. and Res. Workshop (ITRW) on Robustness Issues in Conversational Interact; 2004 Aug. 30-31 Norwich.

[37] The International Phonetic Alphabet. 2005. Available from: `www.langsci.ucl.ac.uk/ipa/IPA_chart_(C)2005.pdf`

[38] Hermelin D, Landau GM, Landau S, *et al*. A unified algorithm for accelerating edit-distance computation via text compression. Proceedings of the 26th International Symposium on Theoretical Aspects of Computer Science; 2009 Feb 26-28; Feiburg. IBFI Schloss Dagstuhl; 2009.

[39] Pucher M, Türk A, Ajmera J, *et al*. Phonetic distance measures for speech recognition vocabulary and grammar optimization. Paper presented at: the 3rd Congress of the Alps Adria Acoustics Association; 2007 Sept 27-28; Graz, Austria.

[40] Bernstein J, Taussig K, Godfrey J. Macrophone: An American English telephone speech corpus. Paper presented at: the Human Language Technology Workshop (HLT-94); 1994 Mar 8-9; Plainsboro, NJ.

[41] Levantine Arabic Conversational Telephone Speech. Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2007S02.

[42] Maamouri M, Buckwalter T, Graff D, *et al*. Fisher Levantine Arabic Conversational Telephone Speech. Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2007S02.

[43] Moreno A, Fonnolosa JA. Spanish SpeechDat(II) FDB-4000, ELRA Catalog No.: ELRA-S0102. 2001.

[44] Russian Conversational Telephony, Appen Pty. Ltd., Appen Catalog No.: RUS_ASR001.

[45] Garofolo JS, Lamel L, Fisher W, *et al*. DAPRPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus. 1993. Available from: `http://perso.limsi.fr/lamel/TIMIT_NISTIR4930.pdf`

[46] Spanish SAMPA Computer Readable Phonetic Alphabet. 1995. Available from: `http://www.phon.ucl.ac.uk/home/sampa/spanish.htm`

[47] Russian SAMPA Computer Readable Phonetic Alphabet. 1995. Available from: `http://www.phon.ucl.ac.uk/home/sampa/russian.htm`