Cross-Language Phonetic-Search for Keyword Spotting

Yossi Bar-Yosef, Ruth Aloni-Lavi, Irit Opher
NICE systems
Ra'anana, Israel
Yossi.Bar-Yosef;Ruth.Aloni-Lavi;Irit.Opher@nice.com

Noam Lotner, Ella Tetariy, Vered Silber-Varod, Vered Aharonson, Ami Moyal
ACLP – Afeka Center for Language Processing
Afeka Academic College of Engineering
Tel Aviv, Israel
Noaml;ellat;veredsv;vered;amim@afeka.ac.il

***Abstract*— Phonetic-search is a method used to enable fast search of spoken keywords within large amounts of audio recordings. The phonetic search process consists of two stages – the indexing phase, where a phonetic lattice is constructed, and the search phase, where keywords are searched in this lattice. The performance of phonetic search systems is highly sensitive to the accuracy of the phonetic recognition, and therefore acoustic model training requires substantial amounts of audio and linguistic resources. Recently, there is a growing demand for applications that require support for keyword spotting in many different languages, including under-resourced languages. Supporting such languages, however, poses a substantial challenge for phonetic-search, since achieving merely reasonable performance requires a lot of training data. In the current research presented here, we propose methods for supporting a new language (the target language), while coping with limited resources, by using existing acoustic models of another language (the source language). In the indexing phase, acoustic models of the source language are used to generate phonetic lattices. Then, the search for keywords in the target language is performed over the recognized lattices. The search is performed by using a cross-language phonetic mapping between the target and source language phonemes. This paper presents methods for cross-language phonetic-search configurations, which depend on the amount of target language available data. Phonetic-search experiments were performed on Spanish as a target language and using American-English and Levantine Arabic as source languages. Results are compared to standard monolingual acoustic modeling in Spanish and show that it is possible to achieve reasonable applicable accuracy for retrieval of spoken words using different combinations of phonetic mappings.**

Keyword-spotting; phonetic-search; under-resourced languages

I. INTRODUCTION

There is a growing demand for supporting new languages in KeyWord Spotting (KWS) and other Automatic Speech Recognition (ASR) based applications. Supporting a new language requires a long and costly process of data collection and training of new acoustic models. Moreover, in some cases, and particularly for KWS in "exotic languages", sufficient training data is not available, which altogether impedes the development of the application.

Since the 90's, research has focused on two different approaches for coping with this challenge. One approach uses multilingual phoneme sets and modeling [1], and the second employs adaptation of acoustic models from existing languages to new ones [2]. The multilingual approach involves a construction of a global phone inventory suitable for a large group of languages [3]. The second approach either generates or adapts new acoustic models either by using manual or semi-automatic phoneme mappings [2], or by performing acoustic adaptation using a small corpus of the new language [4]. A recent study suggested using existing well-trained models from a few source languages for unsupervised transcription generation for training the under-resourced target language [5].

The methods of the latter two studies ([4] and [5]) involved using source language acoustic models for recognition in a target language, where some adaptation has been applied after the initial mappings and alignments. However, all such attempts were aimed at Large Vocabulary continuous Speech Recognition (LVCSR) or Language ID applications and not at keyword spotting.

Our research seeks to consolidate a methodology for supporting phonetic search (PS) in a new target language using mappings between a well-explored source language and the target language. A system employing this methodology will be able to use acoustic models of the source language in order to find keywords in the target language, without the need for large language infrastructures (i.e. databases) and without training or adaptation of acoustic models. PS is particularly suited for such a configuration for the following reasons: (1) the phonetic lattice represents the acoustic content of the speech; (2) the search is carried out through a series of "soft" decisions, depending on likelihoods into which mapping costs can be easily incorporated; (3) a word-level language model is not required.

This paper presents two aspects of our methodology: phonetic-mappings – where both linguistic based and statistically derived mappings are examined; and the contribution of a phonetic language model in the lattice generation stage.

II. METHODS

Our cross-language method uses acoustic models of a well-resourced language to process recordings in another language in the indexing phase of phonetic search. The keyword search in the target language is performed over the recognized

This research is part of a grant (#45828) provided by the Chief Scientist of the Israeli Ministry of Commerce for developing Phonetic Search in New Languages Based on Cross-Language Transformations. The research was carried out as part of the Magneton program which encourages the transfer of knowledge from academic institutions to industrial companies – in this case ACLP – Afeka Center for Language Processing and Nice Systems Inc.

phonetic lattices generated, using a cross-language phonetic-mapping between the target and source language phonemes.

Three languages were investigated in this study: English and Levantine Arabic as *source* languages and Spanish as the *target* language. The phonemic inventory of each language was set according to the following: English – 39 phonemes based on the DARPA phonetic alphabet [6]; Arabic – 43 phonemes based on the Buckwalter Transliteration [7]; Spanish – 31 phonemes based on SAMPA [8].

A. Training of acoustic models

The source languages – English and Arabic – had ample speech and language resources. Acoustic models were trained for both languages using standard HTK tools. The feature-set used was 39 features per frame (energy plus 12 Mel-Cepstrum Coefficients, with the first and second derivatives), calculated over 25 milliseconds frames with 15 milliseconds overlap. Model's configuration was 5 tied-state HMM for each context dependent phoneme (with 3 emitting states), where each state's output probability was modeled by a mixture of 16 Gaussians with diagonal covariance matrices.

B. Cross-language phonetic mapping

The cross-language phonetic mapping is a transformation from the source language's phoneme set to the target language's phoneme set. Different types of cross language mappings can be used, depending on the availability of audio data in the target language: Pure linguistic knowledge is applicable when no sufficient target language data is available, while statistical learning of mappings is possible when a small amount of acoustic data in the target language is available. In our methodology, the linguistic mapping was required in both tests, since it was also used as a bootstrap for the statistical learning.

The linguistic mapping between the language pairs was based on phonologic similarities, a notion expressed in [9] as follows: "languages share the same basic architectural blueprint and features: similar components, types of rules, units, constraints of these units, functions of the system, and cognitive basis." [9: p. 2]. The mapping procedure is thus designed to "close the gap" between the phoneme sets of two languages, the complexity of which is reflected, for example, in the vowel mapping of English to Spanish: The Spanish vowel set is like most Latin languages, relying on five vowels /a/, /e/, /i/, /o/ and /u/. These vowels are represented in the English sound system. However, English also employs an additional eight vowels on average (a total of 13). A bilingual research [10] showed that a Spanish-speaker trying to speak English would be expected to create additional vowel sounds that are not native to her. On the contrary, an English-speaker would be expected to compress her speech to rely on less than half of the normal number of vowels used." [10: p. 5]. The same "gap" between the two languages exists in the phonetic mapping.

In our initial mapping each target phoneme was mapped to a single source phoneme, representing the closest acoustic counterpart in the source language, according to phonetic and articulation manners. For example, the Spanish [a] vowel, as in [paDres] (Spanish for 'parents') was mapped into the Arabic [a], as in [baEdayn] بعدين (Arabic for 'then,' 'later'). This Spanish phoneme was also mapped to the English [AA], as in [PAAD] 'pod', or 'father' (in some dialects), although the Spanish phoneme is pronounced closer to the front of the mouth and thus may be interpreted by most English speakers as /æ/ or /ɑ/ [9], which correspond respectively to [AE] and [AO] in the DARPA transcription method. Our second, broader transformation added the possibility of a target phoneme mapping into more than one source phoneme. The two mapping methods led to different scenarios in the recognition and search phases. The results presented in this paper use the latter transformations, since our preliminary experiments showed their advantage over the one-to-one transformation.

C. Keyword search on a phonetic lattice

Keyword search over a given phonetic sequence is a pattern matching problem. We use the following notations:

$O = \{o_1, \dots, o_T\}$ is a series of T observation vectors.

$R = \{r_1, \dots, r_S\}$ is a recognized sequence of S phonemes.

$W = \{w_1, \dots, w_P\}$ is a searched pattern of P symbols.

A keyword search is based on the likelihood $p(O, R|W)$ - the probability of observing O and recognizing R, given that a particular keyword W was pronounced. Using the simple Bayes' rule we obtain,

$$p(O, R|W) = p(O|R, W)p(R|W), \quad (1)$$

and applying the Markov chain relation, $O \leftarrow R \leftarrow W$, yields,

$$p(O, R|W) = p(O|R)p(R|W). \quad (2)$$

Conveniently, the result in Eq. (2) is composed of two independent types of conditional probabilities. The left term, $p(O|R)$ is the "acoustic" probability, and the right term, $p(R|W)$ can be considered as the "cross-phoneme (series)" probability. The major advantage of this solution is that the acoustic probabilities can be pre-calculated and stored as a phonetic lattice in the indexing phase, regardless of the searched keywords. The search process thus requires only the calculation of the "cross-phoneme" probabilities over the various paths in the recognized lattice.

We further simplify the process by assuming that the cross-phoneme probabilities are context-independent. This leads to a naive derivation of the likelihood computation such that

$$p(O, R|W) = p(O|R) \prod_{i \in \boldsymbol{B}} p(r_i|w_i), \quad (3)$$

where $i \in B$ is the examined path. Notice that w_i in the conditional probabilities, $p(r_i|w_i)$ can accommodate both insertion and deletion events. These phoneme-to-phoneme probabilities, $p(r_i|w_i)$ are pre-defined similarities in the system and are used by the search mechanism to compute the pattern matching costs. In practice, $p(R|W)$ is computed through a dynamic-programming algorithm searching for the best matching path using $p(r_i|w_i)$ for the likelihood scoring.

D. Cross-language phonetic search

Assuming that the acoustic model parameters of the source language remain fixed, we address two higher level issues of the cross-language PS. The first issue is the influence of a phonetic language model used by the phonetic recognizer. This

relates to phonotactic constraints (realized as bi-gram transition probabilities in the phonetic recognizer), as reviewed in the results section. The second issue is the phonetic mapping used in the search module.

E. Learning statistical phonetic mapping

In addition to using mappings determined by a linguist, we implemented a learning mechanism which we hypothesized could improve the accuracy of the mappings when little target language acoustic data is available. Namely, the learning mechanism estimates $p(s_i|t_j)$ where s_i represent source language phonemes, and t_j are the target language phonemes. This technique requires a small amount of target language data (in our case only one hour of speech), with the corresponding word-level transcriptions and lexicon, but without time-aligned segmentations.

First, we constructed a confusion matrix to accumulate the confusions between the correct phonetic series of the target language (obtained from the lexicon) and the recognized phonetic series (obtained from looking at the best path in the lattice) of the source language. To achieve the best alignment between the two series, we use the dynamic-search algorithm with the linguistic phonetic mapping as similarity measures (so the mapping is used for bootstrapping the learning process). However, our initial experiments showed that applying this standard process for learning the empiric phonetic mapping resulted in a significant degradation in the performance of the KWS search. Performance was poor for both types of linguistic mapping – one-to-one or one-to-many. A deeper inspection revealed that the series' alignment mechanism performed poorly since the initial linguistic mapping consisted of "pure" phoneme transformations and did not take into account acoustic mismatches and recognition errors. We therefore extended the mapping possibilities to include acoustic variations that were detected in the development set, as well as, other a-priori anticipated phonetic recognition errors. Moreover, a broader phone-to-phone mapping was added to cover confusions between phonemes belonging to the same Natural Class (NC). The NC mapping consists of three NC classifications: *Plosives*, quasi-periodic signals – *Sonorants* and voiced and voiceless *Fricatives*, which are considered as basic forms of speech signals. In our system, the linguist can provide mapping weights to the different possibilities, where less probable confusions are given lower weights. An example of this procedure is as follows: To map Arabic phonemes to the Spanish phoneme /b/ three weighted mapping options are given:

b < 1.0 > b
b < 0.3 > p
b < 0.1 > d, k, t, g.

The first two pairs are conventional transformations, but the mapping in the third reflects possible recognition errors. We used additional NC mappings (with lower weights) among plosives, vowels, nasals, and fricatives. The weighting of a phoneme-to-phoneme mapping followed a simple principle: The initial cross-lingual mappings were given a default weight of <1.0>, while possible in-class errors were given a small weight of <0.1> and probable errors between closely-articulated phonemes (for example /b/ and /p/ that only differ in voicing) were given the weight of <0.3>. This approach proved capable of fixing the series' alignment and enabling a robust statistical learning of the mapping.

F. Languages and resources

Phonetic-search experiments were performed on Spanish as a target language using American English and Levantine Arabic as source languages. Results are also compared to standard monolingual acoustic modeling in Spanish (trained on 80 hours of speech).

The cross-language PS was evaluated using four corpora: the Wall Street Journal portion of the Macrophone [11] that contains a collection of read sentences; Levantine Arabic Conversational Telephone Speech [12]; Fisher Levantine Arabic Conversational Telephone Speech [13] and Spanish SpeechDat(II) FDB-4000 [14]. The experimental test set includes one hour of speech from the Spanish SpeechDat corpus, and the search was performed on a list of keywords with three syllables or more. The development set for estimating the confusion matrices included another hour of speech. Phoneme recognition was performed using the HTK toolkit.

III. RESULTS

The initial experiment was performed over lattices generated by the original models (without any resource manipulation and using only the original source phonetic language model), where the search module used the linguist's phonetic mapping. Fig. 1 shows similar detection rates for both English and Arabic lattices, which were significantly lower than the monolingual reference in Spanish.

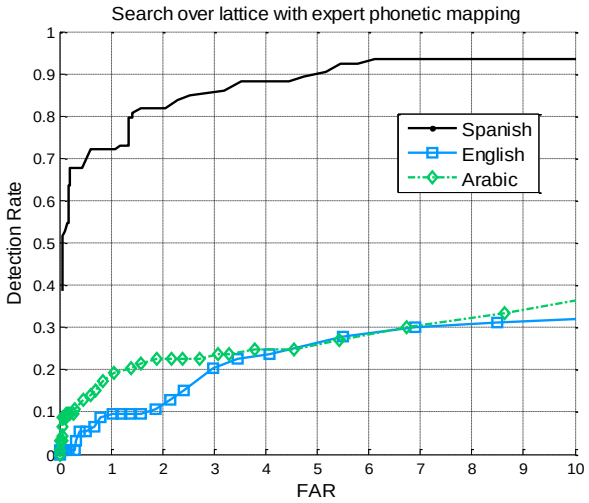

Figure 1. Keyword spotting results on Spanish test set over lattices generated by different models. Spanish is the monolingual reference.

Next we examined the influence of the phonetic LM in the recognition phase. Three configurations were tested: (1) source models with ergodic topology, (2) source models with their original phonetic LM, (3) a combination of the "optimal" Spanish phonetic LM with the acoustic models of the source language. By "optimal" we mean that a large scale lexicon and textual database can give us a robust phonetic LM for the target

language. Fig. 2 demonstrates the KWS performance of the described configurations for English as the source language. The results clearly indicate that using a good language model is imperative for improving performance. Furthermore, the results indicate that any given phonetic LM, even one from the source language, is better than a flat ergodic model. The configuration of acoustic English models combined with the "optimal" Spanish phonetic LM showed the best results.

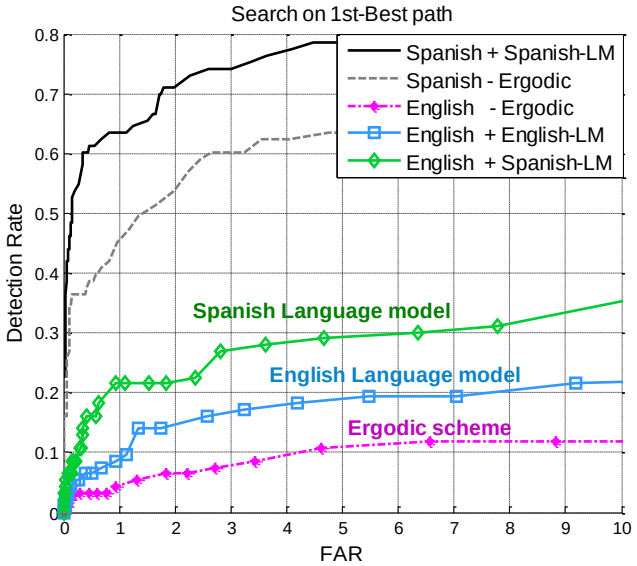

Figure 2. Keyword spotting results on Spanish test set on the best recognized path with different language model schemes.

Fig. 3 presents the results obtained with the statistical phonetic mapping vs. the linguistic mapping. It is obvious that the monolingual system also benefits from using a more accurate mapping in the search phase. A more substantial improvement is observed in the cross-language configurations for English and Arabic. Indeed, when we inspected the statistical confusion matrices, the cross-language matrices were much less diagonal, reflecting the true probabilistic mapping. Hence, using this information in the search phase increased the detection rate by up to 20% absolute.

IV. Conclusions

This paper introduced methods for applying phonetic search in cross-language conditions when there are insufficient language resources in the target language. We focused on mapping methods between the source and target language phoneme sets, as well as, phonetic language model configurations. We proposed a methodology for adding *Natural-Class* mappings to the linguistic mapping bootstrap, to be used in the process of empirically learning confusion probabilities. Using the statistical learning based mapping in the search module boosted the performance significantly, achieving still lower but reasonable results in comparison to PS performance using the well trained target language models. The incorporation of different LMs in the recognition phase increased the search performance significantly. An important observation is that even using the language model of the source language is better than using an ergodic, non-restrictive topology. Future research in the phonetic search framework will be performed in order to improve results. One possible direction is to investigate acoustic model adaptation for varying amounts of target language data, beginning with either linguistically or statistically derived mappings.

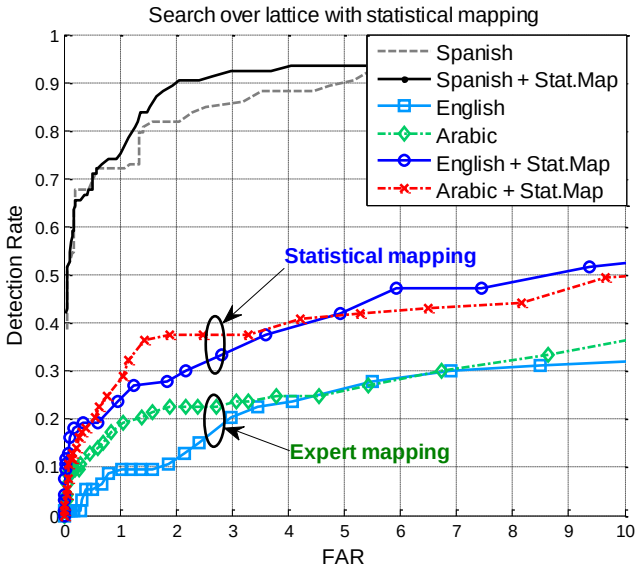

Figure 3. Keyword spotting results on the Spanish test set over lattices using different phonetic mappings for the search process. All configurations include the original language model. "Stat.Map" indicates statistical-mapping. Other configurations use the linguist's mappings.

References

[1] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in: *Proc. Eurospeech*, pp. 371-374, Rhodes, 1997.

[2] B. Wheatley et al., "An evaluation of cross-language adaptation for rapid HMM development in a new language," *ICASSP-94*, 1994.

[3] T. Schultz, "Globalphone: A multiligual speech and text database developed at Karlsruhe University" *ICSLP*, 2002.

[4] P. Fung, C.Y. Ma and W. K. Liu, "MAP-based cross-language adptation augmented by linguistic knowledge: from English to Chinese" *Eurospeech*, 1999.

[5] N. T. Vu, F. Kraus and T. Schultz, "Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training", *Interspeech*, 2011.

[6] J. S. Garofolo, et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *LDC*, Philadelphia, 1993.

[7] T. Buckwalter and M. Maamouri, "Guidelines for the Transcription of Arabic Dialects (EARS)" *Arabic Treebank Project LDC*, University of Pennsylvania, 2004.

[8] www.phon.ucl.ac.uk/home/sampa/spanish.htm

[9] M. S. Whitley, *Spanish/English Contrasts: A Course in Spanish Linguistics*, 2nd ed., Georgetown University Press, 2002.

[10] Bilinguistics, "Typical Development of Speech in Spanish in Comparison to English," Retrieved 1 June 2012 from: www.pediastaff.com/uploads/resources/abad_0707.pdf, 2007.

[11] J. Bernstein, K. Taussig, and J. Godfrey, "MACROPHONE", *LDC*. Philadelphia, USA, 1994.

[12] Appen Pty Ltd, "Levantine Arabic Conversational Telephone Speech," *LDC*, Philadelphia, USA, 1994.

[13] M. Maamouri et al. "Fisher Levantine Arabic Conversational Telephone Speech", *LDC*, Philadelphia, USA, 2007.

[14] A. Moreno and J. A. Fonnolosa, "Spanish SpeechDat(II) FDB-4000 (ELRA-S0102)," *ELRA*, 2001.