ORIGINAL RESEARCH

Cross-language phoneme mapping for phonetic search keyword spotting using multiple source languages

Ella Tetariy[1], Yossi Bar-Yosef[2], Michal Gishri*[1], Ruthi Alon-Lavi[1,2], Vered Aharonson[1], Irit Opher[1,2], 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: November 22, 2015 **Accepted:** January 12, 2016 **Online Published:** February 3, 2016
DOI: 10.5430/air.v5n2p24 **URL:** http://dx.doi.org/10.5430/air.v5n2p24

ABSTRACT

Performing Phonetic Search Keyword Spotting (PS KWS) in new languages when language resources are scarce is an interesting and challenging task. In a previous paper we reported a methodology that enabled PS KWS under these conditions utilizing cross-language phoneme mappings from another sufficiently resourced and well-trained source language. We performed phoneme recognition in the new target language with the acoustic model of the source language. The keyword search was performed over a phoneme lattice of the target language phonemes following a mapping from one language to the other. In the present work we extend this method and its capabilities by mapping two source language phoneme sets into one target language set and performing a combined lattice search. Testing the technique on English and Arabic as source languages yielded a 50% Detection Rate (DR) and a False Alarm Rate (FAR - measured in number of false alarms per hour per keyword) of 2 when Spanish was the target language, a DR of 36% and FAR of 4 when Dari was the target language and a DR of 35% and FAR of 6 with Farsi as the target language. These results indicate that combining two source languages is better than using a single language since the acoustic space is better represented. Searching in a combined lattice while employing adequate phoneme transformations significantly improves performance. Such a system can be used as an initial version of a PS KWS system in a new language when sufficient language resources are not available.

Key Words: Cross-language phoneme mapping, Keyword spotting, Spoken term detection, Phonetic search, Multi-lingual Keyword Spotting, Parallel lattice search

1. INTRODUCTION

There is a growing demand for Keyword Spotting (KWS) systems that enable specific words to be identified out of a stream of continuous speech.[1] This demand is manifested in international evaluation efforts that led to significant advances in KWS research in recent years.[2,3] The applications based on KWS are many and diverse: from call classification and speech database search for call centers and security-intelligence organizations, to multi-media search applications in the internet and enterprise markets. These applications are relatively easy to develop for languages which are rich in Language Resources (LRs). However, when a new language is concerned, the process of collecting LRs, such as large speech and text databases for training acoustic and language models and a large vocabulary pronunciation lexicon,[4–7] is both long and costly.

Our previous works[8,9] reviewed existing technologies that utilize cross-language phoneme mapping to enable the use

*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.

of statistically representative acoustic models from a well-resourced language in order to perform KWS in an under-resourced language. Examples for various implementations of this concept were proposed in several recent studies.[10–12] The studies vary in both method for modeling[13–16] and in the mapping techniques employed.[17–21] Studies also differ in the amount of source LRs available, as well as, the languages concerned. A comparison of the recognition performance between the various studies reported in the literature is difficult to assess due to the differences in data and methods used. Moreover, most studies target their methods to Automatic Speech Recognition (ASR) (*e.g.* Ref.[4,22]) rather than KWS. In our previous study, we compiled a methodology for rapidly introducing Phonetic Search KWS capabilities in a new language for varying quantities of LR availability. Our methodology was based on using source language acoustic models for producing a string of recognized phonemes in the target language and then searching keywords. Hence there was no need for a full set of LRs or for training dedicated acoustical models in the target language.

We employed PS for performing KWS since it is more suitable in cases where vocabulary flexibility is required together with fast search in large speech databases. The PS method is also less dependent on linguistic constraints, and requires training a phoneme-level, rather than a word-level Language Model (LM). Furthermore PS KWS has an advantage when phoneme mappings between languages are applied since PS works mainly on the acoustic-phonetic feature space, rather than any other space (*e.g.* words) and employs a fuzzy search mechanism that may compensate for inaccurate mappings.[9,23]

Our cross-language PS KWS system consists of the following two central components:

(1) 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.
(2) 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 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:

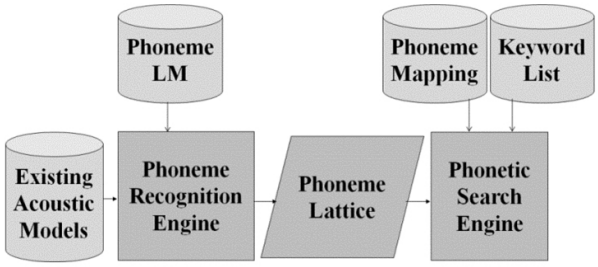

Figure 1. Cross-Language Phonetic Search KWS System

The design of the system allows these two parts of the system to function independently. The phoneme recognition stage uses the source language acoustic models and phoneme LM as input, to produce a phonetic lattice. 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 the source and target language phonemes.

The focus of our previous work was the use of acoustic models from a single source language to a target language. We studied three different mapping methods: manual knowledge-based mapping; data-driven mapping, and performance-based mapping. The data-driven method uses limited target speech data to train coarse acoustic models in the target language and then measures the distance between these coarse acoustic models in the target language and well-trained acoustic models from the source language to automatically generate the best-matched mapping. The performance-based mapping is used to improve the accuracy of the knowledge-based or data-driven mappings by automatic learning from the phoneme recognition statistics. The training set for the recognition statistics was a small amount of speech in the target language, selected randomly.

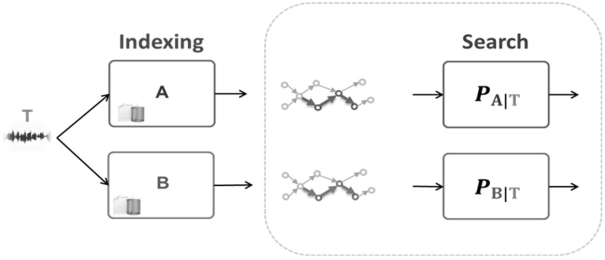

Figure 2. Two independent systems of Cross-Language Phonetic Search KWS

In the present work, we extended the scope of this methodology to study the effect of several source languages mapped into a single target language. Figure 2 illustrates two separate cross-language systems, built for a given target language **T**, where each system uses different source language models,

symbolized as **A** or **B**. Specifically, we sought to assess the effect of acoustic similarity between one or both source languages to the target language and to compare the benefits of either method in terms of performance. The ultimate goal of this research is to improve the accuracy of KWS in a given under-resourced target language by fusing the knowledge extracted from two (or more) independent cross-language KWS systems.

2. METHODS

2.1 Speech databases

The evaluation of our cross-language PS KWS methods was performed using American English (En) and Levantine Arabic (Ar) as the source languages, and Spanish (Sp), Dari (Da) and Farsi (Fa) as target languages.

As is customary for ASR systems, speech audio was used to train acoustic models, and to test the quality of these models. The data was divided into one part used for training, and another for the testing, such that each part contains disjoint sets of the speakers. For the current work, the English acoustic models where trained using 157 hours from the Wall Street Journal portion of Macrophone[24] that contains a collection of read sentences; Arabic models were trained using a total of 115 hours from Levantine Arabic Conversational Telephone Speech[25] and Fisher Levantine Arabic Conversational Telephone Speech;[26] The experimental test sets for Spanish, Dari and Farsi included one hour of speech for each language. Spanish tests were performed on a portion of the SpeechDat Spanish database for fixed telephone network,[27] Dari tests were performed on a portion of DAR_ASR001 Appen Dari audio database, and Farsi tests were performed on a portion of FAR_ASR001 Appen Farsi audio database. The development database used for estimating the confusion matrices (to be used by the weighted Levenshtein distance during the phonetic search stage) included an additional hour of speech for each language from the same databases. The remaining Spanish audio was used to generate a Spanish reference experiment that required well-trained acoustic and language models. The search was performed on a list of keywords of three syllables or more for each language.

The phoneme sets used for each language were as follows: En – 39 phonemes based on the DARPA phonetic alphabet; Ar – 43 phonemes based on the Buckwalter transliteration;[28] Sp – 31 phonemes based on the SAMPA phonetic alphabet for Spanish;[29] Da – 31 phonemes based on the SAMPA phonetic alphabet; Fa – 31 phonemes based on the SAMPA phonetic alphabet.

2.2 Acoustic model training

Acoustic models were trained for both source languages using the standard Hidden Markov Model Toolkit (HTK).[30] An MFCC based, 39-dimensional feature vector was used (13 Mel-Frequency Cepstral Coefficients, with the first and second derivatives), calculated over 25-millisecond frames with a 10 millisecond step. Tri-phone modeling was used with Hidden Markov Models (HMMs) containing 3 emitting states, each state's emission probability modeled by a mixture of 16 diagonal-covariance Gaussians. The development set for estimating the confusion matrices included another hour of speech in the target language. Phoneme recognition was performed using HTK. The search was performed on a list of keywords per language having three syllables or more.

2.3 Phoneme mappings

PS KWS using cross-language mapping techniques was described in our former paper. The current research employed the same three mapping paradigms to map phonemes from two source languages to the target language phonemes. An example of the former, single-source cross-language phoneme mapping is illustrated in Table 1: The mappings, between English and Spanish, were one-to-many on the phoneme level. In the present study, phonemes of two source languages: English and Arabic, were mapped into the target language Spanish phonemes. An example of this mapping is illustrated in Table 2.

Table 1. Examples of one-to-many phoneme mapping between Spanish and English

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

Table 2. Examples of phoneme mapping between Spanish and two source languages: English and Arabic

Spanish	< >	Arabic	English
o	< >	u	ow
b	< >	f	v
rr	< >	r	r

2.4 Lattice construction

Phoneme recognition using different source models was performed similarly to the previous study where only one source language was used. As the two source languages method is based on producing two separate lattices in the two source languages and then performing a combined search in these lattices, the phonetic search phase had to be significantly altered, in order to meet the goals of the current study. The main challenge was to allow transitions between both lattices

in order to increase the flexibility of the phonetic search, without excessively increasing the degrees of freedom in the search. This challenge was met by penalizing the cross-lattice transitions by a proportion that depends on the time gap between the two connected nodes. Next we describe the joint lattice search method.

Assume two cross-language transformations from two different source languages into a single target language denoted by $A \rightarrow T$ and $B \rightarrow T$, where A and B symbolize the source languages and T the new target language. For each transformation, a probabilistic mapping matrix $P_{T|A}$ and $P_{T|B}$ is computed. Each entry (t, s) in the matrix is a conditional probability of the form

$$P_{T|A}(t, s) = p(w_t|a_s) \tag{1}$$

$$P_{T|B}(t, s) = p(w_t|b_s) \tag{2}$$

where w_t is a target phoneme of T, a_s is an acoustic model of A, and b_s is an acoustic model of B. If we denote $|A|$ and $|B|$ as the size of the phoneme set in A and B respectively, and $|T|$ as the number of target phonemes in T, then the dimensions of matrices $P_{T|A}$ and $P_{T|B}$ are $|T| \times |A|$ and $|T| \times |B|$ respectively.

Given the two transformations described above, we propose a simple method to construct a new multi-language transformation, $AB \rightarrow T$, as follows. First we define a new probabilistic mapping $P_{T|AB}$ by concatenating $P_{T|A}$ and $P_{T|B}$ (assuming the same phoneme order of T inherent by the row order of both matrices), such that

$$P_{T|AB} = [P_{T|A}\ P_{T|B}] \tag{3}$$

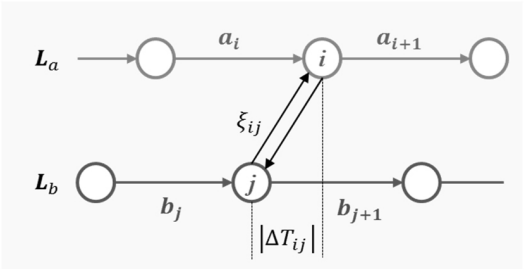

Figure 3. Cross-lattice bi-directional transition between node i of lattice L_a and node j of lattice L_b, with transition costs $\xi_{ij} = \xi_{ji}$

The next operation was implemented per speech utterance during the search process. Each utterance was indexed by two phonetic lattices, L_a and L_b, from which a new lattice L_{ab} that allows cross-language transitions from one lattice to another and vice versa was produced in a constrained manner. The transition method can be visualized as follows: the phonetic lattice is modeled by a graph, where the graph nodes indicate time stamps and the arcs indicate the recognized phonemes. Transition arcs are added to the graph to connect nodes of different source languages, while a cost rule is implemented for each transition. The cross-lattice connections are illustrated in Figure 3.

The cost is a log likelihood one, based on $|\Delta T_{ij}|$ - the time gap between node i of lattice L_a and node j of lattice L_b (see Figure 3), allowing for a bi-directional transition between the two nodes. We define the probability of crossing, $P_{cross}(\Delta T_{ij})$, as an exponential family density function having the form

$$P_{cross}(\Delta T_{ij}) = K \exp(-|\Delta T_{ij}| - \gamma) \tag{4}$$

where K is a positive normalization constant and γ is a positive constant bias (set in order to have some minimal penalty on any transition).

Then the cost, denoted by ξ_{ij} can be written as

$$\xi_{ij} = \log[P_{cross}(\Delta T_{ij})] = -\epsilon_0 - \epsilon_1|\Delta T_{ij}| \tag{5}$$

where ϵ_0 and ϵ_1 are positive constants.

The cost rule in equation (5) implements a penalty of cross-language transitions including the case where the time gap $|\Delta T_{ij}|$ equals zero, in order to prevent loopbacks in the search. Therefore a small penalty, $\epsilon_0 > 0$ is set. The second constant in equation (5), $\epsilon_1 > 0$, can be calibrated to optimize search results. In practice ΔT_{ij} is given in frame-step units (typically quantized to 10 msec time-gap between adjacent frames). Our experiments showed that the transition cost should be set to be significantly large within a few frame steps, say between 5 to 10 frames.

In order to reduce the computational cost during the search, we found it efficient to prune cross-lattice transition arcs in the graph that entail very low probability. Preliminary experiments also indicated that a reduced form of lattice fusion can be implemented to save search computations: namely, it was sufficient to connect cross-language nodes only within a time-gap of 30 msec (up to 3-frame distance) with a minimal transition cost of $\xi_{ij} = -\epsilon_0$ (where $\epsilon_0 = 0.001$ and $\epsilon_1 = 0$). This approach led to a negligible decrease in accuracy.

2.5 Evaluation

To assess the contribution of the proposed lattice fusion approach, we conducted cross-language phonetic search experiments using several configurations. The experiments

were performed on 3 target languages (separately): Spanish, Dari and Farsi, using phonetic lattices indexed by the original acoustic models of two source languages, English and Arabic.

The one-to-one (single source) configurations (acoustic models of a single source language mapped to the new target language), En→ Sp (for English-To-Spanish), and Ar→Sp (for Arabic-To-Spanish) for Spanish and for Dari, En→Da and Ar→Da accordingly, were calculated for reference.

The new lattice fusion method was also compared to a reference of combined results of the two corresponding single lattice searches that were independently generated. We tested several approaches for combining results, in the post-decision stage. The reference chosen was the approach that yielded the best KWS performance using a combination of results of En→Sp and Ar→Sp mapping for Spanish as a target language. The most simplistic approach was found to yield the best results: pooling together all spotting results, from both searches, using a score normalization that is source-language-dependent. "Z-normalization" (Znorm), given by equation 6, was applied for each cross-language search:

$$z = (s - \mu)/\sigma \tag{6}$$

where s is the non-normalized score, and μ and σ are the mean and standard-deviation of true-detection scores, computed over a small development set (less than half an hour).

This reference was used for all the following experiments. Results were evaluated on the test part of Spanish, Dari and Farsi.

Phoneme recognition was performed using the HTK speech recognition engine. The feature vector, of order 39, consisted of MFCC 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 the mapping schemes described above. The weighted Levenshtein distance was used to measure the distance between the keywords and partial phoneme sequences on the lattice. True and false detections were estimated for various thresholds and presented in a graph showing DR as a function of FAR, that were calculated as follows:

$$DR = 100 * N\_true/N\_total \tag{7}$$

where *N_true* is the number of true detections and *N_total* is the total number of all occurrences of keywords in the audio.

$$FAR = N\_false/(Dur * N\_kw) \tag{8}$$

where N_false is the number of false detections (false alarms), Dur is the audio duration in hours, and N_kw is the number of keywords in the list of words we are searching.

This means that the DR is measured in percentages while FAR is a real number.

3. RESULTS

The following figures show comparative results for different multi-language configurations. As described in the methods section, the target languages in the experiments were Spanish, Dari, and Farsi. In all figures, the notations "En" and "Ar" correspond to English and Arabic, respectively, as single source languages, whereas "En+Ar" corresponds to their combination, which can be performed in two distinct methods. As described above, we examined reference combination derived by a post-decision technique, which is denoted in the figures as "En+Ar: union + Znorm"; and the new lattice fusion method, proposed in this study, which is denoted in the figures as "En+Ar: lattice fusion".

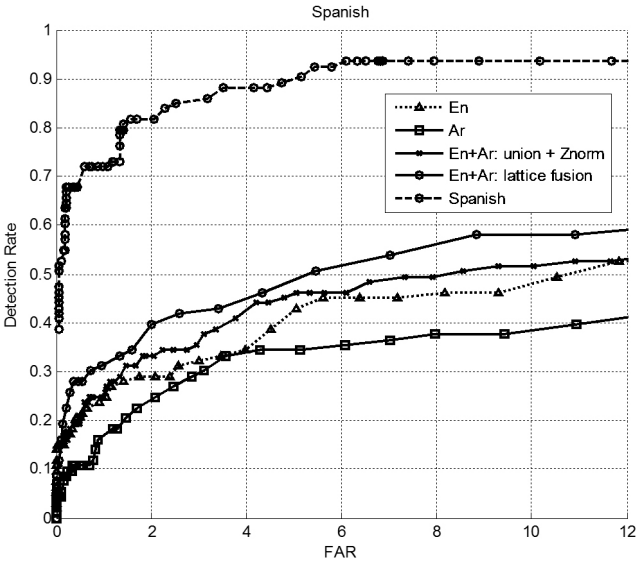

Figure 4. Spanish results in four cross-language configurations

Figure 4 demonstrates results for Spanish as the target language using the four search methods. It can be observed that with the simple approach of result pooling and score normalization, using two source languages already improves performance compared with the single source language methods. Additional improvement can be seen for the proposed lattice fusion method.

Figure 5 presents a somewhat different behavior with Dari as the target language. First, we note that the Ar→Da configuration yielded significantly better results than En→Da. The graphs further demonstrate degradation in performance when the two source languages are pooled together using Znorm normalization, compared to using Arabic language alone. The lattice fusion method, however, provided a modest (but obvious) improvement over the performance of the Ar→Da system.

A possible explanation to the different behavior we see when using Dari as the target language (compared with the Spanish case) is that even though Dari & Arabic do not belong to the same linguistic families, Arabic is much better suited to span the acoustic space of Dari speech. When combining the KWS results in post processing, many FAs are introduced due to the English phoneme recognition. Employing the combined lattice search makes the system less prone to these errors, as the English part is "chosen" only when needed.

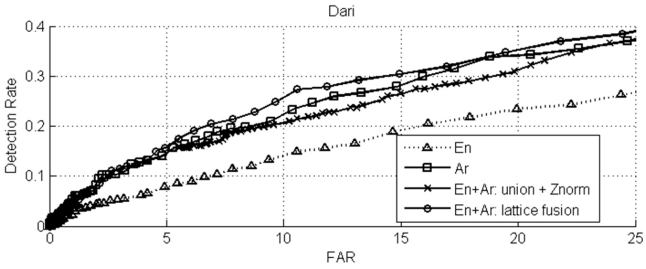

Figure 5. Dari results in four cross-language configurations

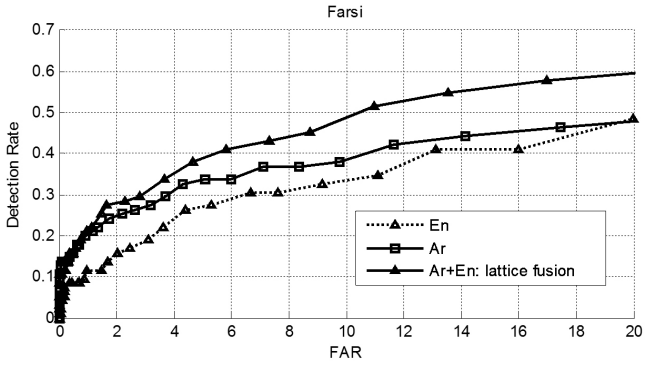

Figure 6. Farsi results of the lattice fusion method (solid line denoted as "Ar+En: lattice fusion") compared to single lattice configurations, from English (En) and Arabic (Ar).

In order to further validate the lattice fusion method, we tested it on a new target language - Farsi. Figure 6 presents PS results using the lattice fusion approach on Farsi as a new target language and compares them to the single source language En and Ar results. Again, the fusion leads to a substantial improvement.

4. CONCLUSIONS

This study presented a lattice fusion approach for applying PS KWS in a new target language given acoustic models of two different source languages. Our experiments compared four different methods: Two separate single-source into a single target cross-language transformations and a post-processing approach that collects all results and applies Znorm score normalization and a combination of the two source language searches, using a new lattice fusion method. Our results indicated two distinct cases. In one case, the two source languages (En and Ar) were quite different (acoustically) from the target language (Sp). In this case, both the post-lattice search score normalization and the fused lattice method yielded better results than the single language cross-language mapping and search. The fused lattice performed better than all other methods, including the post-processing method.

The major advantage of the fused configuration, however, was demonstrated in the second case in the Dari experiments. In this case, the asymmetry between the two single cross-language mappings, Ar→Da, and En→Da is substantial. Not surprisingly, the Ar→Da configuration out-performed the En→Da mapping. This imbalance poses a difficulty when attempting to exploit the weaker system (En→Da) in order to improve the results of the stronger system in a post decision approach. Indeed, results demonstrate that the Znorm operation up-scaled the scores of the weaker system, and thus inserted additional false detections, which led to a degradation in the overall performance. The lattice fusion method, however, provided a modest (but obvious) improvement that exceeded the performance of the single language Ar→Da system.

We believe that the robustness of the suggested fusion method lies in the probabilistic approach of the combined fused search, which is based on the mapping matrix, $P_{(T|AB)}$ (equation (3)). In the Dari case, for example, the search path will make a transition to an English route only when the phonetic mapping likelihood is high enough compared to other Arabic options. Thus in most cases only strong (in a probabilistic sense) English-Dari matches could affect the results, while the others are essentially neglected.

To further validate this assumption we experimented with a similar case, using Farsi as the target language instead of Dari. Although an asymmetry between English and Arabic mappings into Farsi is noted, here again, the lattice fusion method provided superior results.

To summarize, when two cross-language configurations are constructed, where in each configuration a single source language is mapped onto a new target language with a proper probabilistic mapping matrix, performing a search on our fused lattice (created from the two resulting phonetic lattices) using a unified multi-language mapping matrix is straight forward. A major advantage of the suggested lattice-fusion approach lies in its flexibility in making transitions between the original lattices during the search operation itself. Our results indicate that this method can enrich the phonetic content and context that can be found in a single search path. When this procedure was performed in a constrained manner, as formulated in this study, the lattice fusion approach led to significant improvements.

We have thus introduced a methodology for applying phonetic search in cross-language conditions when there are insufficient language resources in the target language. This approach improves the KWS performance and is also relatively simple and quite generic. It can be further applied in cases where model sets of multiple (more than two) source languages are available.

ACKNOWLEDGEMENTS

This work was supported by grant #45828 provided by the Chief Scientist of the Israeli Ministry of Economy for developing Phonetic Search Keyword Spotting 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 the Afeka Center for Language Processing (ACLP) and Nice Systems Ltd. In addition, Dr. Ruthi Alon-Lavi and Irit Opher were in NICE Systems Ltd., Ra'anana, Israel when the research was carried out.

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): 1870-8. 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] NIST Open Keyword Search 2014 (OpenKWS14) Evaluation Plan. 2014. Available from: http://nist.gov/itl/iad/mig/openkws14.cfm

[4] 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.

[5] Besacier L, Barnard E, Karpov A, *et al*. Automatic speech recognition for under-resourced languages: A survey. Speech Communication. 2014; 56: 85-100. http://dx.doi.org/10.1016/j.specom.2013.07.008

[6] 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 Sept 4-8; Lisbon, Portugal.

[7] Le VB, Lamel L, Messaoudi A, *et al*. Developing STT and KWS systems using limited language resources. Proceedings of Interspeech; 2014 Sept. 14-8; Singapore.

[8] Tetariy E, Bar-Yosef Y, Silber-Varod V, *et al*. Cross-language phoneme mapping for phonetic search keyword spotting in continuous speech of under-resourced languages. Artificial Intelligence Research. 2015; 4(2).

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

[10] 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. 30-1; Norwich.

[11] 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.

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

[13] Nieuwoudt C, Botha EC. Cross-language use of acoustic information for automatic speech recognition. Speech Communication. 2002; 38(1): 101-13.

[14] Liu C, Melnar L. A non-acoustic approach to crosslingual speech recognition performance prediction. Proceedings of INTERSPEECH. Brisbane, Australia; 2008 Sept. 22-8. p. 2719-22.

[15] 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.

[16] 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 Apr 21-4; Munich, Germany.

[17] 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.

[18] Bar-Yosef Y, Aloni-Lavi R, Opher I, *et al*. Automatic learning of phonetic mappings for cross-language phonetic-search in keyword spotting. Proceedings of the 2012 IEEE 27th Convention of Electrical & Electronics Engineers in Israel (IEEEI), 2012 Nov. 14-7; Eilat, Israel. p. 1-5.

[19] Le VB, Besacier L, Schultz T. Acoustic-phonetic unit similarities for context dependent acoustic model portability. Proceedings of 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2006 May 14-9; Toulouse, France.

[20] Pascale FUNG, Yuen MC, Kat LW. MAP-based cross-language adaptation augmented by linguistic knowledge: from English to Chinese. Proceedings of Eurospeech; Budapest, Hungary. 1999 Sept 5-9. p. 871-4.

[21] 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.

[22] 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 22-5 Sept; Rhodes, Greece.

[23] 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

[24] 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.

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

[26] Maamouri M, Buckwalter T, Graff D, *et al*. 2007. Fisher Levantine Arabic Conversational Telephone Speech. Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2007S02.

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

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

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

[30] Young SJ, Odell JJ, Woodland PC. Tree-based state tying for high accuracy acoustic modelling. Proceedings of the Workshop on Human Language Technology; 1994 Mar 8-11; Association for Computational Linguistics; 1994. p. 307-12.