Cross-Language Phoneme Recognition for Under-Resourced Languages

Noam Lotner, Ella Tetariy, Vered Silber-Varod,
Vered Aharonson, Ami Moyal
ACLP – Afeka Center for Language Processing
Afeka Academic College of Engineering
218 Bney Efraim Rd.
Tel Aviv 69107

Yossi Bar-Yosef, Irit Opher,
Ruth Aloni-Lavi
NICE systems
22 Zarhin Street
P.O. Box 4122
Ra’anana 43622

Abstract—In the present research, we explore several methods for transforming phoneme models from a language with acoustic models that have been trained (source language) to another, untrained language (target language). One approach uses acoustic distance-measures to automatically define the mapping from source to target phonemes. This is achieved by training basic models for the target language using a limited amount of training data and calculating the distance between the source models and target models. Naturally, this approach requires some data from the target language. Another approach, which also requires some data from the target language, is to use acoustic adaptation for augmenting the source language acoustic models to better match the acoustic properties of the data in the target language. Phoneme recognition results of these approaches are compared to a reference recognizer that is well-trained on the target language.

I. INTRODUCTION

As Automatic Speech Recognition (ASR) based services become ubiquitous, there is an increasing need to support more and more languages. One of the main challenges this introduces is the need to overcome the lack of representative language resources (speech and text databases) available in many languages. In order to resolve this problem, accessible and well-trained models from other languages are sometimes used for training and recognition in the under-resourced language.

This paper presents on-going work in the context of using phonetic search for a keyword-spotting application for under-resourced languages.

In our research, we explored several methods for transforming phoneme models from a language with well-trained Acoustic Models (AM), referred to as a source language, to another, untrained language (target language). The aim of such a process is to find a transformation between the source-language phonemes and the target-language phonemes using only a small amount of speech data.

The set of source acoustic models can comprise a single language (monolingual) or two or more languages (multilingual). A monolingual source model-set was studied in [9], and a multilingual source model-set was explored in [11][12][14][16]. In this paper we describe our current work using monolingual source models.

A major factor in choosing which method to use for the source to target transformation is the amount of speech data that is available in the target language. When no speech data is available, the primary approach is to define a phonetic mapping from the phoneme set of the source language to the phoneme set of the target language, based on similar phonetic properties between phonemes [9][13].

When some speech data is available, other methods become relevant. Here we explore two such methods – the first is to use the (limited) target speech data to train coarse acoustic models for the target language and use them to calculate the distance between each pair of target and source phonemes. This distance matrix can then be used to automatically generate the phonetic mapping, thus alleviating the need for expert knowledge. This method is described in section II B. The second method, described in section II C, is to use the target language data for learning an acoustic transformation that will adapt the source language acoustic model to better match the target language data.

II. METHODS

A. Languages and resources

Three languages were investigated in this study: American 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 [1]; Arabic – 38 phonemes based on the Buckwalter Transliteration [2]; Spanish – 31 phonemes based on SAMPA [3]. Results are also compared to standard monolingual acoustic modeling in Spanish (trained on 80 hours of speech).

Speech databases that were used for this work consist of the following: American English – Macrophone [4] that contains a collection of read sentences; Levantine Arabic - Levantine Arabic Conversational Telephone Speech [5] and Fisher Levantine Arabic Conversational Telephone Speech [6]; Spanish - SpeechDat(II) FDB-4000 [7].
Model-training and recognition experiments were performed using the HTK toolkit.

1) Acoustic model configuration

The source languages – English and Arabic – had ample speech and language resources. Hidden Markov Models (HMMs) were trained for both languages using standard HTK tools. The feature-set used was 39 features per frame (energy plus 12 Mel-Frequency Cepstral Coefficients, with the first and second derivatives), calculated over 25 millisecond-frames with 15 millisecond overlap. The acoustic model configuration was standard 5-state HMM (with 3 emitting states) with state output probabilities modeled using mixtures of 16 Gaussians with diagonal covariance matrices. Both context-independent (monophone) and context-dependent (triphone) models were trained. The triphone HMM states were clustered and tied using decision-tree clustering.

Target language models were trained using 1 hour of speech from the development set of the SpeechDAT(II) database; therefore a full triphone model cannot be reliably trained, and only monophone models with 4-Gaussians per state were trained. In order to calculate the distance-measures (see next section), a similar model was also trained for the source languages.

Finally, for reference, a full acoustic model of the target language (with topology similar to the source triphone AMs) was also trained using the full train-set of the SpeechDAT(II) database.

2) Language model

All recognition experiments utilized a bigram language model trained on the development data set (1 hour of speech) of the target language. The language model in this case is bi-phones, phoneme transition probabilities.

B. Experimental Setup

The test phase is demonstrated in Fig. 1, where recognition is carried out for the target language with the acoustic models of the source language.

![The Test System](image)

The upper part of Fig. 1 represents the mapping process that uses the acoustic Distance Measure (DM), where the mapped phonemes are those target phonemes with matched acoustic models of the source language. The bottom part of Fig. 1 represents the phoneme recognition process, whose input is: target speech, acoustic models from the mapping process and language models (phoneme level) of the target language that were created from the development set. The output is a phoneme sequence in the target language.

C. Acoustic Distance Measures

In the past, a number of distance measures (DMs) between GMMs have been suggested. The selected distance measures are based on the work presented in [9] which examined mappings between a new target language to a source language based on the following distance measures: Kullback-Leibler (KL) distance, Bhattacharyya distance, Mahalanobis distance, Euclidean distance, and Jeffreys-Matusita distance. For a detailed definition of each DM, see the cited work.

A basic training process (as described in the previous section) was carried out for the source languages, as well as for 1 hour from the development set of the target language, in order to serve in the calculation of the distance measure between the acoustic models (see Fig. 1).

Although the Gaussian models that were trained were multi-mixture, the distance between two phonemes was calculated in the experiments using only one mixture per state and using multi-state (three states).

A distance metric of the target phoneme set versus the source phoneme set is produced from each DM calculation. The phoneme-to-phoneme mapping is automatically calculated according to minimum distance (fixed threshold) between two acoustic models, while the goal is to cover as many target phonemes as possible with phonemes of the source language. Since more often than not there is no exact match between the languages, two options are taken into account: either we are left with unmapped source phonemes (leaving redundant phonemes in the model); or we do not get full target phoneme-set coverage, a problem that can be solved by multilingual source models.

D. Acoustic Model Adaptation

One of the primary factors contributing to low recognition performance is the mismatch between the speech data used for training the acoustic models, and the actual speech that is used for testing. This includes differences in microphone, environmental conditions (e.g. noise) and channel noise among others. Additionally, the models are trained using data from many different speakers, yielding a model that fits an ‘average’ speaker, but which is sub-optimal for specific speakers. In the context of cross-language acoustic modeling, this situation is even more pronounced, as there is also a mismatch between the acoustic properties of the trained model and the speech, stemming from the difference in the phonetic content between the source and target languages.

When there is some speech data that matches the real test conditions, it can be used to further tune the acoustic model to better match the test conditions – a process termed acoustic model adaptation. In [13], several methods for acoustic model adaptation for cross-language recognition are surveyed. In this work we used Maximum-Likelihood Linear Regression (MLLR) for adapting the source-language acoustic models.
In short, MLLR finds a linear transformation matrix that is applied to the model parameters, such that the likelihood of the adaptation data is maximized given the transformed model. This work presents results obtained with MLLR applied only to the models’ mean vectors (in HTK this type of transformation is referred to as MLLRMEAN); in the future, we also intend to try adaptation of the covariance matrices. Both global MLLR transform and class-based transforms were used. Global MLLR applies the same transformation matrix to all Gaussian means in the model, while class-based MLLR uses a different matrix for each class of models. The classes are determined using regression-tree clustering of the HMM states such that two states with closely matching output probabilities are clustered together.

For cross-language adaptation, each source language model was labeled with the target phoneme that it is mapped to. The mapping used was the one supplied by the expert linguist. Source language models that were not mapped to target phonemes were omitted, and for source phonemes that are mapped to more than one target phoneme the model was duplicated so that each target phoneme had a copy.

Another problem that is specific to the cross-language case is that the phonetic contexts of the source and target languages differ, sometimes significantly. This fact makes it difficult to use context-dependent modeling (as some of the contexts that appear in the target data are missing from the source model). Therefore, at this stage, we only applied the adaptation to context-independent models.

III. RESULTS

The methods presented above were evaluated on cross-language phoneme recognition tests. American-English (AE) and Levantine Arabic (LA) were used as source languages, and Spanish was used as the target language. These results are compared to a baseline result of using well-trained Spanish models for Spanish (monolingual) phoneme recognition.

A. Baseline results

In order to achieve baseline results we ran a phoneme recognizer of the target language with its original acoustic models that were trained in exactly the same way as the source language models.

Results were evaluated on the test part of the Spanish database and the performance was measured by the Phoneme Error Rate (PER), which is the Levenshtein distance (the minimum number of insertion, deletion and substitution operations required to transform one sequence into the other, also referred to as edit-distance) between automatic transcription results and reference pronunciation divided by the number of phonemes in the reference pronunciation.

```
Table 1: Baseline Spanish Recognition (Monolingual)

<table>
<thead>
<tr>
<th>#Total</th>
<th>#Sub</th>
<th>#Del</th>
<th>#Ins</th>
<th>#Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>283930</td>
<td>60964</td>
<td>28698</td>
<td>30306</td>
<td>194268</td>
</tr>
<tr>
<td>PRR</td>
<td>PER</td>
<td>%Sub</td>
<td>%Del</td>
<td>%Ins</td>
</tr>
<tr>
<td>57.75%</td>
<td>42.25%</td>
<td>21.47%</td>
<td>10.11%</td>
<td>10.67%</td>
</tr>
</tbody>
</table>
```

The recognizer correctly identified 68.42% of the Spanish phonemes with PER of 41.25%.

B. Acoustic Distance Measures

Each of the five distance measures that were mentioned above was calculated for every AE-Spanish phoneme pair (a 39-by-31 distance matrix), and a minimum-distance phoneme mapping was automatically generated for each of the suggested DMs. The results of applying this mapping for recognizing the target data (using the source models), are presented in Table 2.

These results are compared to the phonetic mapping procedure carried out by a linguist (rightmost column). A similar procedure was carried out for the LA-Spanish language pair. The results are presented in Table 3.

In the experiments that were conducted, the following performance criterion is used:

\[%\text{Correct labels} = \frac{\text{number of correct labels}}{\text{total number of labels}} \times 100\%

Table 2: Results for recognizing Spanish phonemes (target) using American English models (source)

<table>
<thead>
<tr>
<th>Distance measure</th>
<th>Kullback-Leibler</th>
<th>Bhattacharyya</th>
<th>Mahalanobis</th>
<th>Euclidean</th>
<th>Jeffreys-Matusita</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Mapped Phones</td>
<td>24</td>
<td>31</td>
<td>23</td>
<td>26</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>#Correct labels</td>
<td>118414</td>
<td>106878</td>
<td>98519</td>
<td>102351</td>
<td>108463</td>
<td>106243</td>
</tr>
<tr>
<td>#Total Labels</td>
<td>268949</td>
<td>283855</td>
<td>249524</td>
<td>269408</td>
<td>283855</td>
<td>266606</td>
</tr>
<tr>
<td>%Correct labels</td>
<td>44.03%</td>
<td>37.65%</td>
<td>39.48%</td>
<td>37.99%</td>
<td>38.21%</td>
<td>39.85%</td>
</tr>
</tbody>
</table>

Table 3: Results for recognizing Spanish phonemes (target) using Levantine Arabic models (source)

<table>
<thead>
<tr>
<th>Distance measure</th>
<th>Kullback-Leibler</th>
<th>Bhattacharyya</th>
<th>Mahalanobis</th>
<th>Euclidean</th>
<th>Jeffreys-Matusita</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Mapped Phones</td>
<td>21</td>
<td>31</td>
<td>19</td>
<td>19</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>#Correct labels</td>
<td>123810</td>
<td>123635</td>
<td>111970</td>
<td>84812</td>
<td>122081</td>
<td>111998</td>
</tr>
<tr>
<td>#Total Labels</td>
<td>218765</td>
<td>283930</td>
<td>204045</td>
<td>149522</td>
<td>283930</td>
<td>201333</td>
</tr>
<tr>
<td>%Correct labels</td>
<td>55.03%</td>
<td>42.14%</td>
<td>54.88%</td>
<td>56.71%</td>
<td>40.41%</td>
<td>55.63%</td>
</tr>
</tbody>
</table>
As can be seen from the results for both languages, the mapping produces automatically produces similar results to those achieved by expert-provided mapping. Note, however, that the results vary significantly both in terms of the actual coverage of target phonemes and in recognition performance. Using the KL distance-measure seems to produce the best combination of phoneme-set coverage and recognition performance.

C. Acoustic Model Adaptation Results

Tables 4 and 5 present recognition results on the Spanish test data using American English and Levantine Arabic models, respectively. In each table, the rows correspond to the following experiments: the line denoted ‘src’ shows the results for running the original source acoustic models; ‘glob’ shows results after applying a global MLLR transform; ‘rc’ shows performance after further applying class-based MLLR transformations. The last line shows the results of running the Spanish reference model.

Columns show the Phone Error Rate (PER), substitution, insertion and deletion rates (%Sub, %Ins, %Del respectively), and the percentage of correctly recognized phonemes (%Corr). The rightmost column shows the reduction in PER of the relevant test relative to the previous test.

As can be seen, using global MLLR adaptation reduced the PER by ~4% for both languages. Class-based MLLR improves the performance significantly, achieving a reduction of close to 10% relative, when applied to the Arabic models, and over 15% for the English models.

Table 4: Adaptation of AE models

<table>
<thead>
<tr>
<th></th>
<th>PER</th>
<th>%Sub</th>
<th>%Del</th>
<th>%Ins</th>
<th>%Corr</th>
<th>ERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>src</td>
<td>74.65%</td>
<td>46.44%</td>
<td>15.08%</td>
<td>13.13%</td>
<td>38.47%</td>
<td></td>
</tr>
<tr>
<td>glob</td>
<td>71.69%</td>
<td>44.76%</td>
<td>15.21%</td>
<td>11.71%</td>
<td>40.02%</td>
<td>3.97%</td>
</tr>
<tr>
<td>rc</td>
<td>60.46%</td>
<td>33.96%</td>
<td>13.61%</td>
<td>12.39%</td>
<td>52.43%</td>
<td>15.66%</td>
</tr>
<tr>
<td>ref</td>
<td>41.34%</td>
<td>23.39%</td>
<td>10.11%</td>
<td>7.84%</td>
<td>66.50%</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Adaptation of LA models

<table>
<thead>
<tr>
<th></th>
<th>PER</th>
<th>%Sub</th>
<th>%Del</th>
<th>%Ins</th>
<th>%Corr</th>
<th>ERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>src</td>
<td>73.36%</td>
<td>43.27%</td>
<td>11.03%</td>
<td>19.06%</td>
<td>45.70%</td>
<td></td>
</tr>
<tr>
<td>glob</td>
<td>70.34%</td>
<td>41.01%</td>
<td>11.15%</td>
<td>18.17%</td>
<td>47.83%</td>
<td>4.12%</td>
</tr>
<tr>
<td>rc</td>
<td>63.72%</td>
<td>33.94%</td>
<td>10.39%</td>
<td>19.39%</td>
<td>55.67%</td>
<td>9.41%</td>
</tr>
<tr>
<td>ref</td>
<td>41.34%</td>
<td>23.39%</td>
<td>10.11%</td>
<td>7.84%</td>
<td>66.50%</td>
<td></td>
</tr>
</tbody>
</table>

Moreover, when the results of the phoneme recognizer are used as input for further processing that takes into account additional information on the target language – for example for doing Keyword Spotting using phonetic search as shown in [17], the effects of reduced performance are further mitigated.

Naturally, for source-target language pairs that are phonetically similar, this approach would yield better results, as can be seen from the fact that Levantine Arabic models perform better than American English ones when used for recognizing Spanish data.

V. FUTURE WORK

One of the primary issues that lead to degraded performance is the phonetic mismatch between the source and target languages. This problem can be addressed by using a larger variety of phonemes as source models. The most straight-forward way is to simply use phonemes from more than one language. Thus our next step is to combine the phoneme-sets of the source languages into a multilingual phoneme-set to use as source models. Another direction we plan to take is to further adapt the acoustic models for the target data using MLLR for adapting the covariance matrices as well as the means, and exploring other methods for adaptation and combination of such methods (e.g. MLLR followed by MAP as suggested in [13]).

ACKNOWLEDGMENT

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 Magnet 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.

REFERENCES


