Context-sensitive Probabilistic Phone Mapping Model for Cross-lingual Speech Recognition

Khe Chai Sim and Haizhou Li

Institute for Infocomm Research, Singapore
{kcsim,hli}@i2r.a-star.edu.sg

Abstract
This paper presents a probabilistic phone mapping model (PPM) that makes possible automatic speech recognition using a foreign phonetic system. We formulate the training of the phone mapping model in the framework of maximum likelihood estimation. The model can be learned automatically from the reference phonetic transcript and the phonetic transcription resulting from a foreign phonetic recogniser using the Expectation Maximisation algorithm. This paper also compares the use of temporal and spatial contexts to enhance the phone mapping performance. A decision tree clustering technique is used to tie unseen contexts for robustness. We evaluate the PPM method on cross-lingual phone and isolated word recognition tasks, using the Hungarian and Russian phone recognisers to recognise Czech speech. Consistent improvement is obtained by using context-dependent phone mapping.

Index Terms: cross-lingual, speech recognition, hidden Markov model

1. Introduction
Multilingual speech processing has gained increasing popularity over the past few years. One of the main cost factors for developing multilingual speech recognition system is the large amount of transcribed speech data needed for acoustic modeling. On the other hand, the development of speech technologies is affected by factors such as academic and commercial considerations, and the size of speaker population. As a result, some languages are better studied than the others. This motivates a strong interest to rapidly adapt well-established acoustic systems towards less computerized, underresourced languages [1].

We formulate the novel framework to model the probabilistic phonetic mapping (PPM) function between the phone sets of two languages in a probabilistic manner, which allows for easy incorporation of the model into the acoustic decoding process during speech recognition. We devise an automatic learning methodology for estimation of model parameters, which does not require re-training of the acoustic models. The model is also extended to incorporate the phonetic contexts, a technique known as context-sensitive phone mapping. Two types of contexts are considered in this paper, namely the temporal and spatial contexts. The former accounts for coarticulation effects in human speech while the latter attempts to 'interpolate' the phones from different (foreign) languages which provides a better defined acoustic space for phone mapping. The proposed PPM framework is evaluated on cross-lingual phone and isolated word recognition tasks for Czech.

The remaining of this paper is organised as follows. Section 2 establishes the cross-lingual phone recognition framework. Section 3 formulates the PPM model as a discrete hidden Markov model (HMM) and its training and decoding mechanism. Section 4 presents both temporal and spatial context-sensitive phone mapping. Finally, experimental results are reported in Section 5.

2. Cross-lingual Phone Recognition
Cross-lingual phone recognition may be formally expressed as follows:

\[
\hat{Y} = \arg\max_{\theta} P(Y|O, X, \theta(X)) = \arg\max_{\theta} \sum_{X} P_{M}(Y|X)P(X|O, X, \theta(X))
\]

where \(O\) is the observation sequence, \(X\) and \(Y\) denote the source and target phone sets, \(\theta(X)\) denotes an acoustic model for \(X\) and \(P_{M}(Y|X)\) is the probability of mapping phone sequence \(\hat{X}\) to \(Y\) given a mapping model \(M : X \rightarrow Y\). This can be approximated as a simple two-stage process:

\[
\hat{X} = \arg\max_{\theta} P(X|O, X, \theta(X)) \quad (1)
\]

\[
\hat{Y} \approx \arg\max_{\theta} P_{M}(Y|\hat{X}) \quad (2)
\]

The first stage is simply to decode the speech utterance, \(O\) using the foreign phone recogniser, \(\theta(X)\) to yield the best phone sequence, \(\hat{X}\), using the \(X\) phone set. The second stage then maps \(\hat{X}\) to \(\hat{Y}\) using the mapping model \(M\). In the next section, a generative probabilistic phone mapping model for \(M\) will be introduced. The formulation and training of this model will also be presented.

3. A Generative Phone Mapping Model
This section describes the formulation of a generative model for phone mapping, where the source sequence, \(\hat{X}\), is the observation and the target sequence, \(Y\), is the model sequence. This model is known as the Probabilistic Phone Mapping (PPM) model. The model parameters are given by \(P_{M}(x|y)\). Since \(x\) and \(y\) are both discrete, \(P_{M}(x|y)\) is essentially a \(|Y| \times |X|\) matrix, where \(|X|\) and \(|Y|\) are the cardinality of the sets \(X\) and \(Y\) respectively. For simplicity, \(P_{M}(x|y)\) will be referred to as the phone mapping matrix. Since we cannot guarantee that \(|X| = |Y|\), it is necessary to model each target phone as a discrete Hidden Markov Model (HMM) [2]. Furthermore, the generative model also requires \(|X| \geq |Y|\). This may be achieved by expanding the original phone sequence \(\hat{X}\) where each phone \(x\) in the sequence is repeated \(d_x\) times and \(d_{\max}\) is the duration of \(x\). This duration can be obtained since \(X\) is generated by a phone recogniser (c.f. Equation (1)). An ex-
3.1. Decoding using PPM Model

Decoding of discrete HMMs can be realised efficiently using the well-known Viterbi algorithm. There are two possible decoding modes:

- **Tandem Mode**
  In tandem mode decoding, the input to the Viterbi algorithm is the expanded sequence, similar to that in training. Therefore, the generative HMM model as shown in Figure 1 can be used directly. An insertion penalty may be applied to control the length of the decoded target phone sequence.

- **Mapping Mode**
  The second decoding mode for the PPM model is called the mapping mode. To perform the decoding in a mapping mode, the original source sequence is used as the input to the Viterbi algorithm. The self-transition is removed from the generative model, forcing each state to generate only one observation. The resulting model simplifies to a discrete Markov model (MM), as shown in Figure 2, since it is not capable of modelling variable length sequences anymore. Therefore, each phone in the source sequence is mapped to only one target phone and the decoded phone sequence has the same length as the source sequence.

3.2. Parameter Estimation

The Maximum Likelihood (ML) estimation of the discrete HMM parameters can be obtained efficiently through the Baum-Welch algorithm [3] in an iterative manner. At each iteration, the following auxiliary function is maximised:

\[
Q(M, \hat{M}) = \sum_{n=1}^{N} \sum_{y \in Y} \gamma_y(n) \log P_M(x_n | y)
\]  

where \( \gamma_y(n) \) is the posterior probability of \( y \) being mapped to \( x_n \), evaluated based on the current model parameters, \( M \). The optimum \( P_M(x_n | y) \) that maximises Equation (3) is given by:

\[
P_M(x | y) = \frac{\beta(x, y)}{\beta(y)}
\]

where the sufficient statistics are given by:

\[
\beta(x, y) = \sum_{n \in Y \cup y} \gamma_y(n) \quad \text{and} \quad \beta(y) = \sum_{x \in X} \beta(x, y)
\]

Note that the alignment between the source and target phone sequence is implicitly defined via \( \gamma_y(n) \). If explicit alignment is available (corpus annotated with phone boundaries), it is possible to incorporate this information directly into the estimation formulae by setting \( \gamma_y(n) = 1 \) if \( x_n \) is aligned to the target phone, \( y \). Otherwise, set \( \gamma_y(n) = 0 \).

This paper also proposes an alternative training method called Augmented Maximum Likelihood (AML). This method maximises the posterior probability indirectly by augmenting the discrete feature space such that the prior distribution of the target models is uniform. Since the prior is uniform, maximising the likelihood is the same as maximising the posterior probabilities. To derive the AML estimation formulae, we augment the input space with a latent symbol, \( x^0 \), such that the ‘effective’ number of frames \( y \) appears in the training data becomes

\[
\hat{\beta}(y) = \beta(x^0, y) + \beta(y)
\]

We need \( \beta(y) = K \) (a constant) to yield a uniform prior for \( y \). This can be achieved by adjusting the values of \( \beta(x^0, y) \). Furthermore, to minimise the effect of \( x^0 \), the value of \( K \) should be as small as possible subject to the constraint that \( \beta(x^0, y) \geq 0 \) (or \( K \geq \beta(y), \forall y \)). Therefore, \( K = \max_{y \in Y} \beta(y) \) and the AML estimation formula is given by

\[
P_M(x | y) = \frac{\beta(x, y)}{K}
\]

Note that this becomes the maximum likelihood estimation in equation (4) when \( \beta(x^0, y) = 0 \).

4. Context-sensitive Phone Mapping

Although different languages share many common sounds, it may not be possible to find a direct mapping between phone sets of two languages. Mapping accuracy may be improved by adding contexts to the source phone sequence to yield a context-sensitive phone mapping. Therefore, the original phone sequence is converted to a context-dependent phone sequence. Subsequently, the converted source sequence may be used as the feature to train and decode a context-sensitive phone mapping model using exactly the same training and decoding strategies as described in the previous section. In this paper, two forms of contexts will be examined, namely the temporal and spatial contexts.

4.1. Temporal Context

In continuous speech, the sound of a phone is influenced by its preceeding and succeeding phones, a phenomenon known as co-articulation. Depending on its neighbours, a source phone may be mapped to different target phones. In this work, the left or right biphone and triphone contexts are used. These are the typical temporal contexts used in speech recognition to distinguish allophones [4]. An example of triphone context expansion is illustrated in Figure 3. The top and bottom sequences represent the original and triphone expanded phone sequences respectively. The triphones are indicated by \( x+y+z \), where \( y \) is the centre phone; \( x \) and \( z \) are the left and right contexts respectively. “pau”, the silence phone, is not modelled with contexts as it is normally not affected by the neighbouring phones.
When rich contextual information is used (e.g. triphone), many of the context-dependent phones may be unseen in the training data. To ensure robustness, decision tree clustering [5] may be used to control the balance between the complexity of the mapping model and the amount of training data available. This was studied and shown to yield promising improvements with explicit time alignment training in our previous work [6].

### 4.2. Spatial Context

While temporal contexts are derived from a single phone sequence, spatial contexts are obtained from multiple phone sequences. The phone sequences may be generated by multiple language specific recognisers. Spatial contexts are then derived by merging phones which occur at the same time segment. This can be viewed as an attempt to ‘interpolate’ phones from different languages to obtain a better defined acoustic space for phone mapping. An example of spatial context expansion is shown in Figure 4. The top two sequences represent the time-aligned output sequences to be combined. The bottom sequence is the resulting sequence after spatial-context expansion. Note that the phone boundaries after spatial-context expansion is the union of the phone boundaries of the original phone sequences.

![Figure 3: An example of triphone context expansion](image)

![Figure 4: An example of spatial context expansion](image)

### 5. Experimental Results

This section presents the experimental results of phone and isolated word recognition tasks on the Czech SpeechDat-E database [7]. There are in total 21.40 hours of training data. For phone recognition experiments, 5.32 hours of test data was used. In addition, 798 isolated word segments (0.94 hours) of test data was used for isolated word recognition. There are 692 unique words in this test set. In this work, three NN/HMM hybrid phone recognisers [8] were used, namely the Czech (CZ), Hungarian (HU) and Russian (RU) 1. These are monophone (context-independent) phone recognisers trained on the SpeechDat-E database. The baseline CZ phone recogniser gave Phone Error Rate (PER) performance of 30.5% and 33.7% on the training data and test data respectively.

The HU → CZ and RU → CZ phone mapping models were trained for cross-lingual speech recognition. Two models were built for each case using the ML and AML estimation methods described in Sections 3.2. The PER results using the mapping and tandem modes (see Section 3.1) are shown in Tables 1 and 2 respectively. In Table 1, it was found that AML estimation consistently outperformed ML estimation. On the other hand, by comparing Table 2 with Table 1, it was found that the tandem decoding mode is consistently better than the phone mapping mode. This is not surprising because the tandem mode allows a more flexible mapping, instead of a one-to-one mapping in the case of phone mapping mode. Similarly, the AML estimation is consistently better than ML estimation in all cases.

### 5.1. Implicit vs. Explicit Time Alignment

The PPM model, as introduced in Section 3, obtains the alignment between the source and target phones during training implicitly using an EM algorithm. This eliminates the need of manually labelling the phone boundaries of the training corpus or relying on having a target phone recogniser to do so explicitly. In our previous work [6], explicit time-aligned corpus was used to learn the phone mapping. The explicit time alignment was obtained by forced-aligning the training data with a C2 phone recogniser. Therefore, it is interesting to compare the performance of PPM model using implicit and explicit alignments. This comparison is summarised in Table 3. As expected, better phone recognition performance is obtained when explicit alignment was used. The performance difference between implicit and explicit alignments is about 7.0–10.0% relative. This shows that implicit alignment using the discrete HMM gives a reasonable approximation although there is room for potential improvement for the implicit alignment approach.

#### 5.1.1. Temporal-context Mapping

Next, the effect of temporal-context mapping is examined. Table 4 shows the performance of the HU → CZ and RU → CZ models with and without triphone context expansion. Decision tree clustering was used to reduce the number of triphones and to handle unseen triphones. There is a consistent absolute improvement of 4.4–5.4% and 9.0–10.4% from using triphone context expansion for the HU → CZ and RU → CZ mapping respectively.

<table>
<thead>
<tr>
<th>Phone Recogniser</th>
<th>Triphone Context</th>
<th>No. of phones</th>
<th>Phone Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>66.1</td>
<td>57.0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1052</td>
<td>57.8</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>66.1</td>
<td>66.1</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1411</td>
<td>57.0</td>
</tr>
</tbody>
</table>

Table 4: Temporal-context tandem Czech phone recognition results using AML trained models

5.1.2. Spatial-context Mapping

With multiple phone recognisers, it is also possible to combine multiple phone recognition outputs using spatial context expansion. Table 5 shows the phone recognition performance using spatial-context mapping. Combining the outputs from the HU and RU recognisers results in 2305 spatial-context-dependent phones. The HU+RU → CZ model improved the HU → CZ model by about 14.1% absolute on both training and test data. The gain is greater than those obtained using triphone-contexts (4.4–5.4%). Therefore, spatial contexts play a more important role in phone mapping compared to temporal contexts. This is because spatial-context mapping combines multiple phone sets to diversify the coverage of sound units.

<table>
<thead>
<tr>
<th>Phone Recogniser</th>
<th>No. of phones</th>
<th>Phone Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1052</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>1411</td>
</tr>
<tr>
<td>HU+RU</td>
<td>2305</td>
<td>43.4</td>
</tr>
</tbody>
</table>

Table 5: Spatial-context tandem Czech phone recognition results using AML trained models

5.2. Cross-lingual Isolated Word Recognition

Finally, the phone mapping models were applied to Czech isolated word recognition. This can be achieved easily by replacing the phone-loop decoding network with a word-loop. In our experiment, the word-loop consists of 692 words (without out-of-vocabulary words). Therefore, the phone mapping model, together with the Czech pronunciation dictionary, can be used to decode the input foreign phone sequence into Czech words. Table 6 summarises the results using context-sensitive phone mapping for cross-lingual isolated word recognition. The CZ → CZ model gave 17.4% Word Error Rate (WER). The WERs using the HU → CZ and RU → CZ models are much higher, 40.2% and 45.1% respectively. Using triphone mapping models gave about 5.0–8.2% absolute improvements. On the other hand, the HU+RU → CZ model gave 25.3% WER (14.8% absolute error reduction over the HU → CZ model). Again, this gain is substantially larger than that obtained using triphone context mapping.

6. Conclusions

This paper has proposed a probabilistic phone mapping model for cross-lingual speech recognition. It models the mapping between the source and target phone sequences using a discrete hidden Markov model, where efficient algorithms exist to estimate the model parameters and to decode the target phone sequence. This model does not require explicit time aligned training data and it can be easily extended to word recognition. Furthermore, with simple modification to the source sequence, context-sensitive phone mapping can be achieved with this model. Temporal context-sensitive mapping with decision tree clustering was reported to yield a better and robust phone mapping. In addition, this paper proposes a spatial context-sensitive mapping which allows multiple phone sequences generated by different phone recognisers to be combined to improve phone mapping performance. The proposed probabilistic phone mapping method was applied to cross-lingual phone and isolated word recognition tasks, where one or more foreign language phone recognisers were used to decode the phone sequence and later mapped to the target phone or word sequence. In general, context-sensitive phone mapping gave consistent improvements over context independent mapping. Given the flexibility of the probability phone mapping model, as our future work, we would like to apply this model to continuous speech recognition and pronunciation modelling.

7. References