Stream-based Context-sensitive Phone Mapping for Cross-lingual Speech Recognition

Khe Chai SIM, Haizhou LI

Human Language Technology Department, Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
{kcsim, hli}@i2r.a-star.edu.sg

Abstract
Recently, a Probabilistic Phone Mapping (PPM) model was proposed to facilitate cross-lingual automatic speech recognition using a foreign phonetic system. Under this framework, discrete hidden Markov models (HMMs) are used to map a foreign phone sequence to a target phone sequence. Context-sensitive mapping is made possible by expanding the discrete observation symbols to include the contexts of the foreign phones in which they appear in the sequence. Unfortunately, modelling the context dependencies jointly results in dramatic increase in model parameters as wider contexts are used. In this paper, the probability of observing a context-dependent symbol is decomposed into the product of probabilities of observing the base symbol and its contexts. This allows wider contexts to be modelled without greatly compromising the model complexity. This can be modelled conveniently using a multiple-stream discrete HMM system where the contexts are treated as independent streams. Experimental results are reported on TIMIT English phone recognition task using the Czech, Hungarian and Russian foreign phone recognisers.

Index Terms: cross-lingual phone recognition, phone mapping, noisy-channel model

1. Introduction
Cross-lingual speech recognition involves using one or more foreign recognisers to recognise speech of a target language. Several work on cross-lingual speech recognition can be found in [1, 2, 3]. Such an approach has several benefits over rebuilding a system from scratch if a good one is not readily available. Firstly, it may be costly and time consuming to obtain transcribed speech data for the language of interest to reliably train the acoustic models, particularly for languages which are under-resourced. Under such circumstance, it may be better to perform a fast adaptation of an existing well-trained foreign system [1, 2]. Secondly, cross-lingual speech recognition enables one system to be used for multiple languages. This motivates the idea of language independent acoustic modelling [4], which allows data from different languages to be pulled together to train better acoustic models.

One of the major aspects of cross-lingual speech recognition is phone mapping [5, 2]. In our previous work, we proposed a context-sensitive Probabilistic Phone Mapping (PPM) model [6] to facilitate cross-lingual speech recognition. This model is essentially a noisy channel model represented by discrete Hidden Markov Models (HMMs). The observation features are the foreign phone sequences extracted by the foreign recogniser. Context-sensitive mapping is made possible by expanding the discrete observation symbols to include the contexts of the foreign phones in which they appear in the sequence. Unfortunately, modelling the context dependencies jointly results in dramatic increase in model parameters as wider contexts are used. To overcome this problem, parameter tying schemes such as decision tree clustering has been investigated in our previous work to reduce model complexity and to cope with unseen contexts [7]. In this paper, the probability of observing a context-dependent symbol is decomposed into the product of probabilities of observing the base symbol and its contexts [6, 7]. This allows wider contexts to be modelled without dramatic increase in model complexity.

The remaining of this paper is organised as follows. Section 2 establishes the cross-lingual phone recognition framework. Section 3 briefly describes the PPM model as a discrete hidden Markov model (HMM). Section 4 presents both joint and stream-based context-expansion for context-sensitive phone mapping. Finally, experimental results are reported in Section 6.

2. Cross-lingual Speech Recognition
As described in [6], cross-lingual phone recognition may be formally expressed as follows:

$$\hat{Y} = \arg\max_Y P(Y|O, X, \theta(X))$$

$$= \arg\max_Y \sum_X P_M(Y|X)P(X|O, X, \theta(X))$$

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 $X$ to $Y$ given a mapping model $M : X \rightarrow Y$. This can be approximated as a simple two-stage process:

$$\hat{X} = \arg\max_X P(X|O, X, \theta(X)) \quad (1)$$

$$\hat{Y} \approx \arg\max_Y P_M(Y|\hat{X}) \quad (2)$$

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 $M$.

3. Probabilistic Phone Mapping Model
Previously, we have investigated the use of Probabilistic Phone Mapping (PPM) models for phone recognition [6, 7] and isolated word recognition [6] on the Czech database. The PPM
model is essentially a discrete HMM model as depicted in Figure 1. Each target phone is modelled by a single emitting state HMM with a self transition. The source sequence is generated by a foreign recogniser and this sequence is expanded by replicating each symbol \( n \) times, where \( n \) is the duration of the phone in frames as generated by the foreign recogniser. This is important to ensure that there is sufficient symbols in the source sequence to be aligned to the target model sequence. Furthermore, it also allows the duration of the phones to be modelled by the transition probabilities. Training and decoding of the discrete HMMs can be carried out efficiently using the well-known Baum-Welch and Viterbi algorithms respectively [6].

### 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. There are two forms of contexts, namely the temporal and spatial contexts to improve the performance of cross-lingual phone recognition [6].

In continuous speech, the sound of a phone is influenced by its preceding 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 bigphone and triphone contexts are used. An example of triphone context expansion is illustrated in Figure 2. 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. a triphone context), many of the context-dependent phones may be unseen in the training data. To ensure robustness, decision tree clustering may be used to control the complexity of the mapping model [7].

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 recognizers. 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 3. The top two sequences represent the time-aligned output sequences to be combined. The bottom sequence is the resulting sequence after spatial-context expansion.

### 5. Context Expansion Schemes

There are two ways of modelling the probabilities of observing these context-expanded symbols, which we refer to as the Joint Context Expansion and Stream-based Context Expansion respectively. In the following, these two types of context expansion schemes will be described.

#### 5.1. Joint Context Expansion

The Joint Context Expansion scheme models the probability of observing a context dependent symbol jointly as

\[
P_M(\tilde{x}|y) = P_M(x, c_1(x), c_2(x), \ldots | y) \tag{3}
\]

where \( \tilde{x} \) denotes the context-sensitive discrete symbol. \( x \) denotes the base symbol while \( c_i(x) \) represents the \( i \)th context of \( x \). Note that if \( x \in \mathcal{X} \) and \( |\mathcal{X}| \) denotes the number of discrete symbols \( x \) can take, the size of the discrete observation space increases to \(|\mathcal{X}|^2\) and \(|\mathcal{X}|^3\) for biphone and triphone expansions respectively. This dramatically increases the model complexity and causes the problem of unseen observations. To overcome the problem, parameter tying using decision tree clustering may be used to control the model complexity by allowing different context-dependent symbols to share the same observation probabilities [7]. However, it becomes cumbersome to perform decision tree clustering for contexts wider than triphone.

#### 5.2. Stream-based Context Expansion

This paper proposes an alternative approach of modelling the observation probabilities for the context-dependent symbols by decomposing the joint probability defined in Equation (3) into

\[
P_M(x, c_1(x), c_2(x), \ldots | y) = P_M(x|y)P_M(c_1(x)|y)^{\gamma_1} P_M(c_2(x)|y)^{\gamma_2} \ldots
\]

where the probabilities of observing the base symbol and its contexts are assumed to be independent. Therefore, the correlations between the base symbols and their contexts are ignored. On the other hand, the independence assumption allows wider contexts to be used without dramatically increasing the model complexity. For example, stream-based biphone and triphone expansions only increase the size of the observation space to \(2|\mathcal{X}|\) and \(3|\mathcal{X}|\) respectively. The base symbol sequence and each of the contexts can be treated as independent streams and a multiple-stream discrete HMM systems may be used to model stream-based context expansion. \( \gamma_i \) denotes the stream weight.
for the \( i \)th context, which can be adjusted to reflect the importance of each stream. Stream-based context expansion can be applied to temporal and spatial contexts:

- **Stream-based Temporal Context Expansion**

<table>
<thead>
<tr>
<th>Stream 1:</th>
<th>pau</th>
<th>pau-a*b</th>
<th>a-b*c</th>
<th>b-c*d</th>
<th>pau</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stream 2:</td>
<td>pau</td>
<td>pau</td>
<td>a</td>
<td>b</td>
<td>pau</td>
</tr>
<tr>
<td>Stream 3:</td>
<td>pau</td>
<td>b</td>
<td>c</td>
<td>pau</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: An example of triphone stream-based context expansion

An example of the triphone stream-based context expansion is given in Figure 4, where \( \otimes \) denotes a stream-based expansion. This expansion will be referred to as a \( \pm 1 \) stream-based expansion as it uses one context to the left and one to the right. Similarly, \( \pm 2 \) denotes a quinphone stream-based expansion. For temporal context expansion, the stream weights, \( \gamma_i \), are computed as

\[
\gamma_i = \frac{1}{1 + |i|} \quad \text{for} \quad i = 0, \pm 1, \pm 2, \ldots
\]

Therefore, the stream weight decreases exponentially for contexts further away from the centre phone. For \( \pm 2 \) contexts, the stream weights are \( \frac{1}{3}, \frac{1}{2}, 1, \frac{2}{3}, \frac{3}{4} \).

- **Stream-based Spatial Context Expansion**

Stream-based context expansion can also be applied to spatial contexts, where the output from each recogniser is treated as a separate stream. An example of the spatial stream-based context expansion is given in Figure 5. In this work, stream weights for spatial context expansion were set to 1.0 for spatial context expansion, i.e. all the phone recognisers receive equal weights.

\[
\begin{align*}
pau & \pm a \pm b \pm c \pm pau \\
pau & \pm p \pm q \pm pau \\
pau & \pm a \pm b \pm b \pm c \pm pau \\
pau & \pm p \pm p \pm q \pm q \pm pau
\end{align*}
\]

Figure 5: An example of stream-based spatial context expansion

5.3. Hybrid Context Expansion

To improve the context-sensitive phone mapping, it is possible to combine both joint and stream-based context expansion. For example, joint context expansion may be used to obtain the left and right biphone context-dependent symbols. Then, stream-based context expansion is used to combine these contexts together. This is illustrated in Figure 6. Note that this way of context expansion uses the same context information as a triphone expansion scheme. It will be shown later in Section 6 that the hybrid expansion scheme is superior to both the joint and stream-based context expansions alone.

6. Experimental Results

This section presents the experimental results of cross-lingual phone recognition task on the English TIMIT database. The speaker independent (SI) portion of the training data, which contains approximately 3.12 hours of speech data, was used to train the models. There are altogether 3696 utterances spoken by 326 male and 136 female speakers. The test data consists of 1344 utterances collected from 112 male and 56 female speakers. This amounts to 1.14 hours of test data. The baseline continuous HMM phone models, the discrete HMM phone mapping models and the decision trees for clustering of triphone contexts were all trained using the HTK [8]. For cross-lingual phone recognition, the speech waveforms were first decoded into token sequences using the high quality NN/HMM hybrid phone recognisers [9] to obtain the HMM state posterior probabilities. These probabilities are then used in the Viterbi algorithm to perform recognition. Three such foreign recognisers were used, namely the Czech (CZ), Hungarian (HU) and Russian (RU)\(^1\). These are monophone (context-independent) phone recognisers trained on the SpeechDat-CZ database [10]. In this work, the speech data are downsampled to 8000 Hz sampling rate to be compatible with the foreign acoustic models which were trained on telephony data. The baseline ML and MMI trained 8-component HMM systems yielded PER performance of 45.92% and 42.24% respectively.

6.1. Stream-based Spatial Context Expansion

\[
\begin{array}{cccc}
\text{Foreign} & \text{PER (%)} \\
\text{Recognisers} & \text{mono} & \text{lc} & \text{rc} & \text{tri} \\
\hline
\text{CZ} & 65.49 & 66.30 & 64.24 & 63.38 \\
\text{HU} & 63.84 & 62.49 & 62.82 & 62.04 \\
\text{RU} & 68.50 & 66.78 & 66.45 & 66.62 \\
\text{CZH\textcircled{RU}} & 56.07 & 54.38 & 53.88 & 52.07 \\
\end{array}
\]

Table 1: PER performance comparing various joint temporal context expansions and stream-based spatial context expansions

Table 1 summarises the Phone Error Rate (PER) performance of various PPM models using joint context expansion for temporal contexts and stream-based context expansion for spatial contexts. mono denotes PPM models without temporal context expansion while lc, rc and tri denote the left-biphone, right-biphone and triphone contexts respectively. CZ\textcircled{HU\textcircled{RU}} denotes stream-based spatial expansion combining the CZ, HU and RU outputs.

The baseline context-insensitive PPM models gave PER performance of 63.84–68.50%. With joint biphone expansion, the performance improved to 62.49–66.78%. Finally, with joint triphone expansion, the performance improved further to 56.07–52.07%.

\(^{1}\)These recognizers are available for download from http://speech.fit.vutbr.cz/en/software/phoneme-recognizer-based-long-temporal-context

3021
6.2. Stream-based Temporal Context Expansion

Next, the effect of stream-based context expansion was investigated and the results of using the ±1 and ±2 contexts are shown in Table 2. In general, the performance improves with wider contexts. The ±1 and ±2 contexts gave 2.97–4.38% and 3.41–5.30% absolute PER reductions respectively over the context-independent baseline. Note that for the individual phone recognisers, using the ±1 context outperforms the joint triphone contexts (c.f. last column of Table 1). This shows that robust estimation of the joint context mapping probabilities can not be obtained and factoring the joint probabilities through the independence assumption leads to better results. However, when the three recognisers are combined, the CZ⊗HU⊗RU system using the joint triphone context gave slightly better performance (52.07%) compared to the ±1 context (52.29%).

6.3. Hybrid Context Expansion

Finally, the results of combining the joint and stream-based context expansion as described in Section 5.2 are given in Table 3. For individual recognisers, the ±1 contexts and the hybrid 1c+rc contexts gave similar performance, both better than the joint triphone contexts by about 2.0% absolute. After combining the three recognisers to form the CZ⊗HU⊗RU system, the stream-based combination of the joint left- and right-biphone expansions (1c+rc) is clearly superior, with the best performance of 50.85%.

7. Conclusions

This paper has proposed a stream-based context expansion scheme for context-sensitive cross-lingual speech recognition using the probabilistic phone mapping model (PPM). This scheme models the probability of observing a context-dependent source symbol as a product of the probabilities of observing the base symbol and its contexts, making the assumption that these probabilities are independent. This scheme allows wider context expansion without greatly increasing the model complexity, at the expense of ignoring the correlation between the base symbols and their contexts. It is also possible to combine the stream-based and joint context expansion schemes to further improve the recognition performance. Experimental results on the TIMIT English database showed that the stream-based context expansion schemes provides a simple way of combining the temporal and spatial contexts to yield further improvement for the context-sensitive cross-lingual phone recognition.

8. References