ROBUST PHONE SET MAPPING USING DECISION TREE CLUSTERING FOR CROSS-LINGUAL PHONE RECOGNITION

Khe Chai Sim and Haizhou Li

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

ABSTRACT

Recently, research related to multi-lingual and cross-lingual speech has gained increasing popularity. One of the major problems when dealing with multi-lingual speech data is the mapping of the phone sets between different languages. Phone mapping is useful for cross-lingual speech recognition, cross-lingual pronunciation modelling and mixed language speech synthesis, to name a few. In this paper, an automatic context sensitive phone set mapping method is presented to improve the mapping accuracy. A training methodology that allows the mapping to be learned automatically from parallel time-aligned phone transcriptions is also described. In particular, a decision tree clustering technique is used to tie unseen contexts for robustness. The quality of the proposed mapping method is evaluated on a cross-lingual phone recognition task where the Hungarian and Russian phone recognisers are used to recognise Czech speech and produce Czech phone sequences through phone set mapping. The mapping was trained on only a small amount of data. A consistent relative improvement of 5 – 7% is reported when contextual information is added to phone set mapping.

Index Terms— cross-lingual, phone recognition, decision tree clustering, context sensitive mapping

1. INTRODUCTION

Multi-lingual and cross-lingual speech research has gained increasing popularity over the past few years, including multi-lingual speech reocnigtion [1, 2], cross-lingual pronunciation modelling [3] language identification [4] and mixed language speech synthesis [5]. One of the major problems when dealing with multilingual data is the mapping of phone sets between different languages. Most phone mapping techniques are based on learning a phone mapping table [6, 3] using a data-driven approach. In some cases, it is not necessary to deal with phone mapping problem in isolation. For example, in cross-lingual speech recognition, the aim is to make use of a source acoustic phone models to perform word recognition on a target language. Therefore, it is possible to learn the phone to word mapping explicitly which allows phonetic contexts to be incorporated when learning the mapping [1].

This paper concentrates on explicit modelling of the phone set mapping. In this paper, a simple probabilistic framework is formulated to model the mapping function between two phone sets. This model can be easily simplified to a phone mapping table by selecting the phone pairs with the highest probability. An automatic learning methodology for the mapping model is also described. In addition, this model is also extended to incorporate phone contexts, a technique known as *context-sensitive* phone mapping. To ensure robustness, a decision tree clustering approach is used to control the model complexity. The proposed mapping method is evaluated using a cross-lingual phone recognition task.

The remaining of this paper is organised as follows. Section 2 introduces cross-lingual phone recognition. Section 3 formulates the phone set mapping problem and describes an automatic methodology for learning the mapping function. Section 4 introduces the context sensitive phone set mapping. Next, Section 5 explains the use of decision tree clustering method to improve the robustness of context sensitive phone set mapping. Experimental results are presented in Section 6. Finally, conclusions are given in Section 7.

2. CROSS-LINGUAL PHONE RECOGNITION

Phone recognition involves finding a phone sequence, $\hat{Y}$, that best match the spoken utterance, $\mathcal{O}$, *i.e.*

$$\hat{Y} = \operatorname*{argmax}_{Y} P(Y|\mathcal{O}, \mathcal{Y}, \boldsymbol{\theta}(\mathcal{Y})) \tag{1}$$

where $P(Y|\mathcal{O}, \mathcal{Y}, \boldsymbol{\theta}(\mathcal{Y}))$ is the posterior probability of the phone sequence, Y, given the observation, $\mathcal{O}$, phone set, $\mathcal{Y}$ and the acoustic model, $\boldsymbol{\theta}(\mathcal{Y})$. Usually, the phone set, $\mathcal{Y}$, is chosen to match the language of $\mathcal{O}$ and the acoustic model, $\boldsymbol{\theta}(\mathcal{Y})$ is also trained using speech data of the same language. Under certain circumstances, it may be necessary to perform a *cross-lingual* phone recognition where a phone recogniser from one language is used to decode the speech of another language. Such a scenario arises when, for example, there is insufficient speech data available to train a reliable phone recogniser and an existing well-trained phone recogniser of another language may be utilised.

For a cross-lingual phone recognition, Eqn (1) is rewritten as

$$\begin{aligned}\hat{Y} &= \operatorname*{argmax}_{Y} P(Y|\mathcal{O}, \mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}(\mathcal{X})) \\ &= \operatorname*{argmax}_{Y} \sum_{X} P(Y|X, \mathcal{M}) P(X|\mathcal{O}, \mathcal{X}, \boldsymbol{\theta}(\mathcal{X}))\end{aligned}$$

where $\boldsymbol{\theta}(\mathcal{X})$ denotes an acoustic model trained on the $\mathcal{X}$ phone set and $P(Y|X, \mathcal{M})$ is the probability of mapping phone sequence X to Y given a mapping model $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$. This can be approximated as a simple two-stage process:

$$\hat{X} = \operatorname*{argmax}_{X} P(X|\mathcal{O}, \mathcal{X}, \boldsymbol{\theta}(\mathcal{X})) \tag{2}$$

$$\hat{Y} \approx \operatorname*{argmax}_{Y} P(Y|\hat{X}, \mathcal{M}) \tag{3}$$

The first stage is simply to decode the speech utterance, $\mathcal{O}$ using the foreign phone recogniser, $\boldsymbol{\theta}(\mathcal{X})$ to yield the best phone sequence, $\hat{X}$, using the $\mathcal{X}$ phone set. The second stage is simply to map $\hat{X}$ to $\hat{Y}$ using the mapping model $\mathcal{M}$. In the next section, phone set

mapping will be introduced. The formulation and training of the mapping model, $\mathcal{M}$, will also be presented.

3. PROBABILISTIC PHONE SET MAPPING

In this section, phone set mapping is formally defined. Given two phone sets, $\mathcal{X}$ and $\mathcal{Y}$, the phone mapping from $\mathcal{X}$ to $\mathcal{Y}$ is defined as

$$\mathcal{X} \rightarrow \mathcal{Y} = \{(x, M(x)); M(x) \in \mathcal{Y}, \forall x \in \mathcal{X}\} \tag{4}$$

where $M(x) : \mathcal{X} \mapsto \mathcal{Y}$ denotes the mapping function. Eqn (4) defines a many-to-one mapping, *i.e.* every element in $\mathcal{X}$ must be mapped to *at most one* element in $\mathcal{Y}$, but there can be more than one elements in $\mathcal{X}$ being mapped to the same element in $\mathcal{Y}$. Thus, the inverse mapping, $M^{-1}(y) : \mathcal{Y} \mapsto \mathcal{X}$ is non-deterministic and the mapping is therefore irreversible.

In this paper, the mapping function, $M(x)$, is modelled as a probabilistic model, $\mathcal{M}$, whose parameters are estimated such that the mapping probability, $P(Y|X, \mathcal{M})$ is maximised, *i.e.*

$$\hat{\mathcal{M}} = \underset{\mathcal{M}}{\operatorname{argmax}} \sum_{(X,Y)\in\Psi} P(Y|X, \mathcal{O}, \mathcal{M}) \tag{5}$$

where Ψ denotes the training set. X and Y are the pair of phone sequences for the speech utterance, $\mathcal{O}$. In order to learn the phone mapping probabilities, it is necessary to align the two phone sequences. This alignment may be obtained through forced-alignment with an acoustic model or using a dynamic programming algorithm. In this paper, T_x and T_y, the time alignments for X and Y respectively, are assumed to be available. Let the time alignments be The phone sequence and the time alignments can be expanded as

$$X = (x_1, x_2, \ldots, x_N), \quad x_i \in \mathcal{X} \quad \forall i \tag{6}$$

$$T_x = (\tau_{x,0}, \tau_{x,1}, \tau_{x,2}, \ldots, \tau_{x,N}) \tag{7}$$

where x_i is the ith phone in the sequence which starts and ends at $\tau_{x,i-1}$ and $\tau_{x,i}$ respectively. The last elements of T_x and T_y are the same and they correspond to the length of the speech utterance, T. Given the time boundaries, the mapping probability is given by:

$$P(Y|X, \mathcal{M}) = \sum_{t=1}^{T} P_{\mathcal{M}}(y = \bar{y}_t | x = \bar{x}_t) \tag{8}$$

where $\bar{x}_t$ and $\bar{y}_t$ denote the phone at time t of X and Y respectively:

$$\bar{x}_t = x_i \quad \text{where} \quad \tau_{x,i-1} \leq t < \tau_{x,i} \tag{9}$$

$$\bar{y}_t = y_i \quad \text{where} \quad \tau_{y,i-1} \leq t < \tau_{y,i} \tag{10}$$

Therefore, the mapping probability table, $P_{\mathcal{M}}(y|x)$, can be estimated by maximising Eqn (5), which yields:

$$P_{\mathcal{M}}(y|x) = \frac{C(x, y)}{\sum_{x\in\mathcal{X}} C(x, y)} \tag{11}$$

where

$$C(x, y) = \sum_{(X,Y)\in\Psi} \sum_{t} I(x, \bar{x}_t) I(y, \bar{y}_t) \tag{12}$$

$$I(p, q) = \begin{cases} 1 & p = q \\ 0 & x \neq y \end{cases} \tag{13}$$

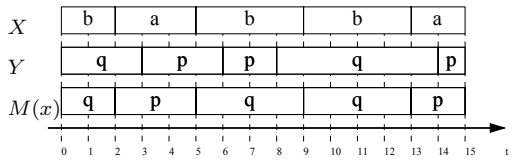
The sufficient statistics, $C(x, y)$, is the total co-occurrence frame count of x and y in the aligned training data and $I(.)$ is an indicator function. The optimum mapping function is given by

$$M(x) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}}\, P_{\mathcal{M}}(y|x) \tag{14}$$

In the following, a simple example will be presented to illustrate the training process. Given that:

$$\begin{array}{cc} \mathcal{X} = \{\text{ a, b }\} & \mathcal{Y} = \{\text{ p, q }\} \\ X = (\text{ b, a, b, b, a }) & Y = (\text{ q, p, p, p, q }) \\ T^{(x)} = (0, 2, 5, 9, 13, 15) & T^{(y)} = (0, 3, 6, 8, 14, 15) \end{array} \tag{15}$$

the time-aligned parallel transcripts can be illustrated as:



By using Eqn (12), $C(x, y)$ can be computed as:

C(x,y)	a	b
p	3	3
q	2	7

For example, `a` and `q` are aligned at the intervals [2,3] and [13,14]. Therefore, $C(a, q) = 2$. Similarly, `b` and `p` aligned at the intervals [5,6] and [6,8], giving a total aligned counts of 3. Hence, by applying Eqn (14) yields the following optimum mapping

x	a	b
M(x)	p	q

4. CONTEXT SENSITIVE MAPPING

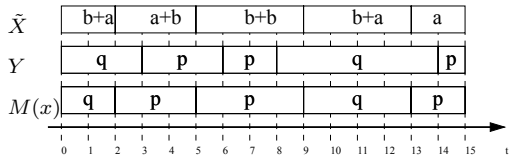
In reality, it is not sufficient to describe the relationship between two phone sets using merely a many-to-one mapping. In some cases, it is just not possible to map a phone $x \in \mathcal{X}$ to only one phone $y \in \mathcal{Y}$ all the time. It is possible to extend the aforementioned mapping method to take into account of the contexts of x. This form of mapping is known as *context sensitive* mapping. Several contexts considered in this paper include the left biphone context (`lc`), right biphone context (`rc`) and triphone context (`tri`).

To achieve context sensitive mapping, it is necessary to expand the source phone set, $\mathcal{X}$ to include all possible contexts and the phone sequences, X, to the corresponding context dependent phones. Taking the previous example, the right biphone context expanded phone set and phone sequence become:

$$\tilde{\mathcal{X}} = \{\text{ a, b, a+a, b+b, a+b, b+a }\}$$

$$\tilde{X} = (\text{ b+a, a+b, b+b, b+a, a })$$

and the modified time-aligned phone sequences are shown below:



The corresponding sufficient statistics is now a 6×2 matrix:

C(x,y)	a	b	a+a	b+b	a+b	b+a
p	1	0	0	3	2	0
q	0	0	0	1	1	6

By treating each context dependent phone as a distinct element in $\mathcal{X}$, the sufficient statistics may be computed in the same way as given in Eqn (12) and the optimal mapping becomes (using Eqn (14)):

x	a	b	a+a	b+b	a+b	b+a
M(x)	p	$\emptyset$	$\emptyset$	p	p	q

From the above example, there are *unseen* contexts (b and a+a) where no mapping was learned (denoted by a $\emptyset$). A simple solution to circumventing this problem is to *back-off* to the mapping without contextual information, in which case, b and a+a will be mapped to q and p respectively. Instead of using a simple back-off strategy, the issue of unseen context-dependent phones can also be resolved by employing a decision tree clustering algorithm similar to that used for building context-dependent acoustic models in speech recognition [7]. This will be discussed further in the next section.

5. DECISION TREE CLUSTERING

Decision tree clustering provides an alternative solution to handling unseen contexts when wider contexts are used. It also provides an elegant scheme to control the complexity of the mapping model to ensure robustness. This is inspired by the success of using decision tree clustering to maintain a balance between model complexity and available training data [7].

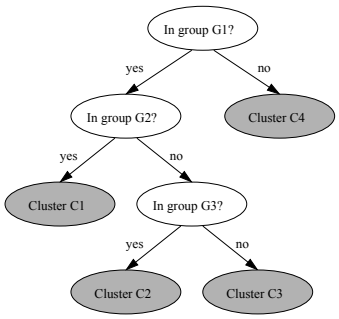


Fig. 1. A binary decision tree for phone clustering

A typical clustering decision tree is depicted in Figure 1. One decision tree is built for each group of context-dependent phones that share the same centre phone. Starting with all these phones at the root node, a question is asked about the contexts at each non-terminal node and the phones in the current node is split into two groups. This process is repeated recursively until a termination criterion is met. The leave nodes of the final tree represents the phone clusters. Usually, a series of overlapping groups are defined so that clustering is performed by asking if the phone belongs to a particular group. Grouping can be derived based on expert knowledge, such as nasals, fricatives, liquids and so on. Alternatively, automatic grouping can also be done based on acoustic similarities. In the example in Figure 1, a phone which is in group `G1` and `G3` but not in `G2` will be clustered into `C2`.

6. EXPERIMENTAL RESULTS

This section presents the experimental results of phone recognition on the Czech `SpeechDat-E` database [8]. The portion of the data containing phonetically rich sentences was chosen for this experiment. The data is split into 10.20 and 2.66 hours of training and testing sets respectively.

Firstly, the conventional GMM/HMM phone recognisers were trained on the 10-hour training set. 39-dimensional Mel Frequency Cepstral Coefficients (MFCC) features (12 MFCC + Energy + Δ + $\Delta\Delta$) were used to train context-independent phone HMM models. Each HMM model adopts a 3-state left-to-right topology with the observation probability distribution represented by a Gaussian Mixture Model (GMM). Model parameters were estimated using both maximum Likelihood (ML) and Maximum Mutual Information (MMI)criteria. A simple phone loop is used during decoding. The phone accuracy performance of the 32-component GMM/HMM

Phone recogniser	Training method	Phone Accuracy (%)
GMM/HMM	ML	39.14
	MMI	41.51
NN/HMM	backprop	67.47

Table 1. Phone Accuracies of 32-component GMM/HMM and NN/HMM phone recognisers on the Czech database

systems are reported in Table 1. The MMI trained phone recogniser gave the best performance of 41.51%. In addition, the performance of a NN/HMM hybrid Czech (CZ) phone recogniser [4] [1] based on long temporal context was found to be 67.47%, significantly better than the best performing GMM/HMM system.

Let us consider the case where we only have the Hungarian (HU) and Russian (RU) NN/HMM hybrid phone recognisers. These recognisers are also trained on the `SpeechDat-E` databases [8] and their phone accuracies are 66.68% and 60.73% on their respective languages. Since the performance of these foreign recognisers are better than the GMM/HMM Czech phone recogniser, we would like to perform cross-lingual phone recognition using the two-stage methodology outlined in Section 2. A subset of the training data ($\sim$ 3.15 hours) was selected to learn the mapping function. No further adaptation of the HU and RU phone recognisers were performed to concentrate on the effect of phone set mapping alone. Firstly, the HU and RU recognisers were used to decode the subset data. The output phone sequences, along with the reference transcriptions (in Czech phone set) were used to train the mapping function, as described in detail in Section 3. The cross-lingual phone recognition

Evaluation Set	Phone Recogniser	Phone Accuracy (%)			
		mono	lc	rc	tri
Train	HU	45.09	46.76	47.36	50.13
	RU	37.75	40.56	40.56	44.73
Test	HU	44.53	47.01	47.05	45.26
	RU	38.27	41.75	41.20	40.52

Table 2. Phone accuracies of the HU and RU recognisers on the CZ test set using various context-sensitive phone set mappings.

performance using the HU and RU recognisers are shown in Table 2. `mono`, `lc`, `rc` and `tri` denote the monophone, left-context biphone, right-context biphone and triphone respectively. The results show that using context-insensitive mapping achieved phone accuracies of 45.09% and 37.75% for the HU and RU recognisers respectively on the training set; 44.53% and 38.27% on the test set. Since the HU phone recogniser has a higher phone accuracy, the resulting cross-lingual recognition performance using this phone recogniser is

[1] Available for download at `http://www.fit.vutbr.cz/research/groups/speech/index_e.php?id=phnrec`

also better. There is a consistent improvement in phone accuracy on the training set as more context information was considered. While similar improvements were observed on the test set for the `lc` or `rc` contexts, there is a clear performance degradation for the triphone context on the test set. This problem is largely due to the many unseen triphone contexts in the test and can be circumvented using a simple back-off strategy. We repeated the experiment in Table 2

Phone Recogniser	Use back-off	Phone Accuracy (%)			
		mono	lc	rc	tri
HU	No	44.53	47.01	47.05	45.26
	Yes	—	47.08	47.10	47.63
RU	No	38.27	41.75	41.20	40.52
	Yes	—	41.82	41.26	42.68

Table 3. Phone accuracies of the HU and RU recognisers on the CZ test set using various context-sensitive phone set mappings, with and without back-off.

using back-off for the unseen contexts. The results, as shown in Table 3, indicate that the back-off strategy improves the triphone context mapping and the phone accuracies are now consistently better than that using biphone context mapping. Furthermore, the triphone back-off phone mapping model trained on only a small subset of the training data is already better than the baseline GMM/HMM system using the complete training data.

Although the simple back-off strategy circumvented the unseen triphone problem, many of the seen triphones in the training set are rare, which may lead to unreliable estimation of the mapping model. The decision tree clustering method (see Section 5) provides a way of controlling the number of distinct triphones to be used to achieve a balance between model complexity and availability of training data. Figure 2 illustrate the change in phone accuracies with respect to the

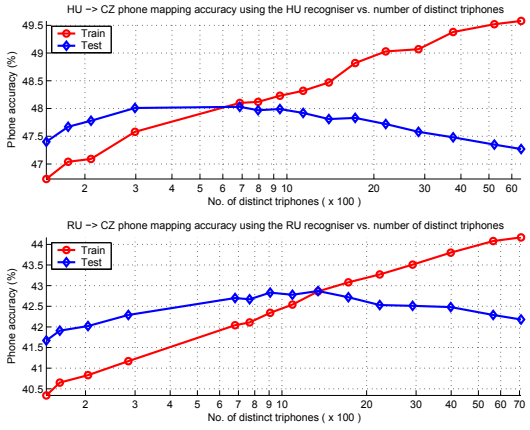


Fig. 2. Phone accuracies variation w.r.t. the number of distinct triphones used in context sensitive mapping, using the HU (top) and RU (bottom) phone recognisers

number of distinct triphones using in context sensitive mapping. On the training data, the phone accuracy increases consistently as the number of distinct triphones increases. However, there exists an optimum point at which the best phone accuracy is achieved. There is obviously a trade-off between the model complexity and training data size. If the number of distinct triphone is too low, the model becomes too simple to take advantage of the contextual information and if the number is too high, reliable statistics cannot be obtained to ensure robust estimation of the mapping model. The best performance of 48.03% and 42.87% were obtained for the HU and RU respectively. The optimum number of distinct triphones were 687 and 1344 respectively. These numbers were also lower than the total number biphones seen in the training data (∼1800).

7. CONCLUSIONS

This paper has proposed a context-sensitive phone set mapping method to enhance the mapping accuracy. It was shown in this paper that using wider phonetic contexts improves the mapping quality provided sufficient training data is available. In order to overcome the data sparseness problem, decision tree clustering approach was adopted to control the model complexity and to ensure robust parameter estimation. The proposed phone set mapping technique was applied to cross-lingual phone recognition, where a foreign language phone recogniser was used to decode the phone sequence and later mapped to the desired phone set. Using triphone mapping was found to consistently outperform context-insensitive mapping by 5.0 – 7.0% relative. The use of decision tree also helped to reduce model complexity, improve robustness and achieve a further marginal gain. We hope to extend the work further in two directions: 1) perform phone alignment using language independent attributes so that a target acoustic model is not required; 2) perform soft phone mapping by incorporating the mapping probabilities in cross-lingual word recognition.

8. REFERENCES

[1] R. Bayeh, S. Lin, G. Chollet, and C. Mokbel, "Towards multilingual speech recognition using data driven source/target acoustical units association," in *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, May 2004, pp. 521–524.

[2] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, August 2001.

[3] T. Martin and T. Svendsen, "Cross-lingual pronunciation modelling for indonesian speech recognition," in *Proceedings of EUROSPEECH*, 2003, pp. 3125–3128.

[4] Jan Cernocky Pavel Matejka, Petr Schwarz and Pavel Chytil, "Phonotactic language identification using high quality phoneme recognition," in *Proceedings Eurospeech*, September 2005.

[5] Leonardo Badino, Claudia Barolo, and Silvia Quazza, "Language independent phoneme mapping for foreign TTS," in *5th ISCA Speech Synthesis Workshop, Pittsburgh*, June 2004.

[6] Le Viet-Bac and L. Besacier, "First steps in fast acoustic modeling for a new target language: Application to vietnamese," in *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, March 2005, pp. 821–824.

[7] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings ARPA Workshop on Human Language Technology*, 1994, pp. 307–312.

[8] P. Pollak *et al.*, "SpeechDat(E) - eastern european telephone speech databases," in *Proceedings LREC'2000 Satellite workshop XLDB - Very large Telephone Speech Databases*, May 2000, pp. 20–25.