MULTI-STREAM TEMPORALLY VARYING WEIGHT REGRESSION FOR CROSS-LINGUAL SPEECH RECOGNITION

Shilin LIU and Khe Chai SIM

School of Computing, National University of Singapore, Singapore

ABSTRACT

Building a good Automatic Speech Recognition (ASR) system with limited resources is a very challenging task due to the existing many speech variations. Multilingual and cross-lingual speech recognition techniques are commonly used for this task. This paper investigates the recently proposed Temporally Varying Weight Regression (TVWR) method for cross-lingual speech recognition. TVWR uses posterior features to implicitly model the long-term temporal structures in acoustic patterns. By leveraging on the well-trained foreign recognizers, high quality monophone/state posteriors can be easily incorporated into TVWR to boost the ASR performance on low-resource languages. Furthermore, multi-stream TVWR is proposed, where multiple sets of posterior features are used to incorporate richer (temporal and spatial) context information. Finally, a separate state-tying for the TVWR regression parameters is used to better utilize the more reliable posterior features. Experimental results are evaluated for English and Malay speech recognition with limited resources. By using the Czech, Hungarian and Russian posterior features, TVWR was found to consistently outperform the tandem systems trained on the same features.

Index Terms— cross-lingual, decision tree clustering, context expansion

1. INTRODUCTION

Recently, multilingual and cross-lingual speech recognition has attracted many researchers due to its challenges and practical applications. This task is particularly designed to build an Automatic Speech Recognition (ASR) system with limited resources, particularly limited transcribed speech data. Although the number of tied triphone states can be reduced to provide sufficient training data for each physical state, the performance could be dramatically decreased due to the poor modelling of acoustic contexts. When the accuracy of the ASR system is low, it becomes difficult to utilize massive un-transcribed speech data. In order to improve the performance of an ASR system with limited resources, researchers began to investigate borrowing the rich resources from other languages due to the similar acoustic characteristics among human languages. For convenience, language with limited resources will be named as native (target) language, while others as foreign (source) language.

One popular approach is to train a multilingual ASR system [1, 2, 3, 4] by pooling resources from all related languages. The ASR system for a target language can be easily obtained by defining a new lexicon using the universal phone set. For a better performance, language adaptation is also applied by optimizing small number of language specific parameters. However, the complexity of the resulting multilingual system may be very high in order to model all the different contexts and language specific patterns, which can lead to inefficient decoding. Other researchers have interests in finding a probabilistic phone mapping [5, 6, 7, 8, 9] between the source language and target language. Thus, it may be applied to adapt the acoustic models before decoding [5, 9], or translate the foreign phone sequence after decoding [6, 7, 8]. The biggest challenge of phone mapping is that it is difficult to robustly map context dependent phone sets given very limited resources. Lastly, tandem features [10, 11, 12, 13] based on well-trained foreign-language neural networks phone recognizers have shown promising results for cross-lingual speech recognition. However, not all tandem features from foreign language can outperform the native acoustic features, e.g. tandem features from Spanish neural networks for Chinese recognition does not perform as good as baseline [13].

Temporally Varying Weight Regression [14] was recently proposed to improve the temporal correlation modelling for Hidden Markov Models (HMM). It extends the conventional HMM by incorporating posterior features trained on long-span acoustic features to model temporally-varying GMM weights. In this paper, TVWR is applied to cross-lingual speech recognition by leveraging on well-trained foreign monophone/state recognizers to produce high quality posterior features. In addition, multi-stream TVWR is proposed where multiple sets of posterior features are used to incorporate richer spatial and temporal context information. Finally, a separate tree-based state-tying is applied to the TVWR regression parameters to better exploit the more reliable foreign posterior features.

The remainder of the paper is organized as follows. In Section 2, an overview of the previously proposed TVWR is given. Multi-stream TVWR is formulated in Section 3.

Section 4 presents a tree-based state tying algorithm for the TVWR regression parameters. Lastly, experimental results are reported in Section 5.

2. TVWR OVERVIEW

TVWR formulation is motivated from factorizing a standard HMM system using a long span of observations as features. The goal is to model the limited temporal information into the GMM weights and keep the system complexity relatively low. Technically, the output probability of an HMM state j in TVWR framework is given as follows:

$$p(\mathbf{o}_t, \boldsymbol{\tau}_t|j) = \sum_{m=1}^{M} \underbrace{P(m|j)p(\boldsymbol{\tau}_t|\mathbf{o}_t, j, m)}_{c_{jmt}} p(\mathbf{o}_t|j, m) \quad (1)$$

where $\boldsymbol{\tau}_t$ is the limited acoustic context of the current observation $\mathbf{o}_t$, denoted as $\boldsymbol{\tau}_t = \{\mathbf{o}_{t-\delta}, \ldots, \mathbf{o}_{t-1}, \mathbf{o}_{t+1}, \ldots, \mathbf{o}_{t+\delta}\}$, δ is the context window size, and c_{jmt} is the temporally varying weight. In order to avoid estimating high dimensional probability density function, $p(\boldsymbol{\tau}_t|\mathbf{o}_t, j, m)$ for each component, the high dimensional components are shared by introducing a latent variable such as:

$$\begin{aligned} p(\boldsymbol{\tau}_t|\mathbf{o}_t, j, m) &\approx p(\boldsymbol{\tau}_t|j, m) && (2)\\ &= \sum_{i\in\mathcal{R}} p(\boldsymbol{\tau}_t|i, j, m)P(i|j, m) && (3)\\ &\approx \sum_{i\in\mathcal{R}} p(\boldsymbol{\tau}_t|i)P(i|j, m) && (4)\\ &\approx Z_t \sum_{i\in\mathcal{R}} p(i|\boldsymbol{\tau}_t)P(i|j, m) && (5) \end{aligned}$$

where $Z_t = p(\boldsymbol{\tau}_t)/P(i)$ is the component independent normalization term, which can be ignored during likelihood calculation, i is the latent discrete variable to partition the context space, $\mathcal{R}$ is a set of these latent variables and $P(i|j, m)$ is the regression parameter. Three assumptions are made: 1). since continuous dependency is hard to model and complicates the model derivation, assumption is made in Eq.2; 2). in order to reduce the system complexity, the high dimensional component, $p(\boldsymbol{\tau}_t|i)$ is shared by assumption in Eq.4; 3). uniform prior $P(i)$ is assumed for convenience in Eq.5. Typically, i is represented by the monophone/state so that its posterior, $p(i|\boldsymbol{\tau}_t)$ can be robustly predicted using a neural network.

Although TVWR has shown better performance than the conventional HMM system, it still suffers from performance degradation under limited-resource condition. Since TVWR requires a robust monophone/state predictor using a long span of observations as input, the lack of training data will lead to less accurate predictor and hence poorer performance for TVWR system. However, it is not necessary to represent the latent variables using the *native* monophone/state. For cross-lingual speech recognition, monophone/state from a different language can also be used to partition the acoustic space. As such, TVWR can leverage on other well-trained foreign recognizers to provide high quality posterior features, avoiding the need to train one with limited resources. In the next two sections, two modifications will be introduced to the TVWR system to further enhance its performance for cross-lingual speech recognition. In Section 3, multi-stream TVWR is proposed where multiple sets of posterior features are used to incorporate richer context information. In Section 4, a separate tree-based parameter tying algorithm is derived for the TVWR regression parameters to improve model complexity control under low-resource conditions.

3. MULTI-STREAM TVWR

In the previous section, monophone/state is introduced to represent the latent variable i such that the system complexity can keep relatively low. Since foreign monophone/state posterior features are used, the performance improvement might be limited by the language differences. In order to further improve the performance under limited-resource condition, multi-stream TVWR is proposed to introduce richer context information to regress the time-varying weights. Specifically, i is now a context-rich latent variable. Without losing generalization, i is defined as a structured variable:

$$\begin{aligned} i &= \{i_1, i_2, \ldots i_c \ldots i_C\} && (6)\\ \mathcal{R} &= \mathcal{R}_1 \times \mathcal{R}_2 \cdots \mathcal{R}_c \cdots \mathcal{R}_C && (7) \end{aligned}$$

where i is now composed of C context-specific "sub-variables" and $\mathcal{R}_c$ is the set of the c-th sub-variable, i_c. This can be viewed as employing multiple partition strategies such that a much higher resolution of the acoustic space can be obtained for better discrimination. However, even with C = 2 or 3, the resulting set $\mathcal{R}$ can be very large. Therefore, it is difficult to estimate the joint posterior probability and there will be a lot of regression parameters to estimate. To circumvent this problem, the context-specific latent variables are assumed to be independent such that the joint posterior probabilities, $P(i_1, \ldots, i_C|\tau_t)$, and the corresponding regression parameters, $P(i_1, \ldots, i_C|j, m)$, can be factorized as follows:

$$\begin{aligned} p(i_1 \ldots i_C|\boldsymbol{\tau}_t) &\approx p(i_1|\boldsymbol{\tau}_t) \cdots p(i_C|\boldsymbol{\tau}_t) && (8)\\ p(i_1 \ldots i_C|j, m) &\approx p(i_1|j, m) \cdots p(i_C|j, m) && (9) \end{aligned}$$

This assumption significantly reduces the system complexity and makes the TVWR formulation tractable. This leads to a multi-stream TVWR where Eq.5 can be rewritten as:

$$p(\boldsymbol{\tau}_t|\mathbf{o}_t, j, m) \approx Z_t \prod_{c=1}^{C} \sum_{i_c\in\mathcal{R}_c} p(i_c|\boldsymbol{\tau}_t)p(i_c|j, m) \quad (10)$$

Note that multi-stream TVWR is different from multi-stream HMM system. In multi-stream HMM system, each state has

multiple stream acoustic features and each stream is represented by a GMM, while every state in multi-stream TVWR has only one stream acoustic feature represented by a single GMM. In the following subsections, multi-stream TVWR will be used to incorporate both temporal and spatial contexts.

3.1. Temporal Context Expansion

In continuous speech, the sound of a phone can be easily influenced by its preceding and succeeding phones, a phenomenon called *co-articulation*, this correlation can be modelled by introducing a cross word triphone/state. However, this important information is lost when using monophone/state to represent the latent variable i in TVWR. Therefore, a temporal context dependent latent variable i is introduced by setting $C = 3$ such that i_1, i_2, i_3 can be used to indicate left, middle and right monophone/state of current frame, respectively. Since these "sub-variables" are from the same monophone/state set but with different position information, X_1, X_2, X_3 are literally the same without considering context position.

Instead of using three separate recognizers to produce three sets of posteriors, only one recognizer is used to predict the middle monophone/state posterior probabilities. The corresponding left and right posterior probabilities are derived from the sequence of middle monophone/state posterior probabilities as follows: starting from the current frame, search left and right until the identity of the monophone/state with the largest posterior probability changes and use the corresponding posterior probabilities as the left and right posteriors. As a results, the left posterior feature, $p(i_1|\boldsymbol{\tau}_t)$ is given by $p(i_2|\boldsymbol{\tau}_{t-\phi})$ where:

$$\arg\max_{i_2} p(i_2|\boldsymbol{\tau}_{t-\phi}) \neq \arg\max_{i_2} p(i_2|\boldsymbol{\tau}_t) \tag{11}$$

$$\arg\max_{i_2} p(i_2|\boldsymbol{\tau}_{t-k}) = \arg\max_{i_2} p(i_2|\boldsymbol{\tau}_t) \quad k \in [1, \phi) \tag{12}$$

At the same time, the right posterior feature, $p(x_3|\boldsymbol{\tau}_t)$ can also be obtained in a similar way. Since silence does not need context, its left and right posteriors are assumed to be the same as the middle posteriors.

3.2. Spatial Context Expansion

Alternatively, multiple monophone/state predictors from different foreign languages can be applied to build a spatial context. In general, multiple foreign languages with more differences can lead to a better discrimination, since they are more likely to be complementary for each other. Spatial context can be more useful when each individual foreign language does not provide a good prediction of monophone/state posterior features. Therefore, C in Eq.10 will represent the total number of foreign languages to be applied, while $\mathcal{R}_c$ is the monophone/state set for c-th language.

3.3. Parameter Estimation

After ignoring independent terms, the auxiliary function w.r.t. regression parameters can be written as:

$$Q(\Lambda, \hat{\Lambda}) = \sum_{t,j,m} \gamma_{jm}(t) \log \Big(\prod_{c=1}^{C} \sum_{i_c \in \mathcal{R}_c} p(\boldsymbol{\tau}_t|i_c) P(i_c|j,m) \Big)$$

$$\geq \sum_{t,j,m,c,i_c} \gamma_{jmi_c}(t) \Big(\log P(i_c|j,m) + \log p(\boldsymbol{\tau}_t|i_c) \Big) \tag{13}$$

where the component occupancy is now given as:

$$\gamma_{jm}(t) = \gamma_j(t) \frac{\hat{c}_{jmt} p(\mathbf{o}_t|j,m)}{\sum_{m=1}^{M} \hat{c}_{jmt} p(\mathbf{o}_t|j,m)} \tag{14}$$

$\gamma_j(t)$ is the state occupancy at time t given the current model $\hat{\Lambda}$, $\hat{c}_{jm}(t)$ is the current time-varying weight of multi-stream TVWR and

$$\gamma_{jmi_c}(t) = \gamma_{jm}(t) \frac{P(i_c|j,m,\hat{\Lambda}) p(\boldsymbol{\tau}_t|i_c)}{\sum_{i_c \in \mathcal{R}_c} P(i_c|j,m,\hat{\Lambda}) p(\boldsymbol{\tau}_t|i_c)} \tag{15}$$

$$\approx \gamma_{jm}(t) \frac{P(i_c|j,m,\hat{\Lambda}) p(i_c|\boldsymbol{\tau}_t)}{\sum_{i_c \in \mathcal{R}_c} P(i_c|j,m,\hat{\Lambda}) p(i_c|\boldsymbol{\tau}_t)} \tag{16}$$

The optimal estimation can be then obtained by using Lagrange multiplier such that:

$$P(i_c|j,m) = \frac{\sum_t \gamma_{jmi_c}(t)}{\sum_{i_c \in \mathcal{R}_c} \sum_t \gamma_{jmi_c}(t)} \; \forall c \in C, i_c \in \mathcal{R}_c \tag{17}$$

Note that this update formula is similar to applying the standard TVWR estimation (by setting $C = 1$) to each stream independently, except that the component occupancy is calculated using the multi-stream TVWR system when performing forward-backward calculations in the E-step of the Baum-Welch training.

4. TREE-BASED STATE CLUSTERING

Typically, the complexity of a triphone system is controlled using the decision tree state tying technique [15]. When training a system with limited resources, the number of distinct triphone state clusters is kept small to ensure robust estimation of all the parameters associated with the tied states. However, this may limit the potential of the TVWR system where the regression parameters cannot take full advantage of the high quality posterior features. To alleviate this problem, a separate tree-based tying algorithm is applied to the TVWR regression parameters so that the model complexity with respect to the regression parameters can be controlled independent of the regular GMM parameters.

Due to the limited training data, state tying algorithm will be performed on the system with only one mixture per state. The essence of the tree-based state tying algorithm is the

derivation of the likelihood increase as a result of splitting a state cluster into two. This allows the appropriate questions to be chosen for each node when constructing the decision tree. The following state-tying derivation for the TVWR regression parameters is largely based on the framework given in [16]. The auxiliary function to be maximized with respect to the TVWR regression parameters can be written as:

$$Q(\Lambda, \hat{\Lambda}) = \sum_t \sum_{j \in \mathcal{J}} \gamma_j(t) \log \Big(\sum_{i \in \mathcal{R}} p(\boldsymbol{\tau}_t|i) P(i|j) \Big) \quad (18)$$

$$\geq \sum_t \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{R}} \gamma_{ji}(t) \Big(\log P(i|j) + \log p(\boldsymbol{\tau}_t|i) \Big) \quad (19)$$

$$= \sum_t \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{R}} \gamma_{ji}(t) \log P(i|j) + K(\mathcal{J}) \quad (20)$$

where $\gamma_j(t)$ is the state occupancy at time t given the current model $\hat{\Lambda}$ and transcription, $\mathcal{J}$ is a state cluster including all triphone states, $\gamma_{ji}(t)$ can be similarly obtained by setting $m = 1$ and $C = 1$ in Eq.16, and

$$K(\mathcal{J}) = \sum_t \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{R}} \gamma_{ji}(t) \log p(\boldsymbol{\tau}_t|i) \quad (21)$$

the optimal solution of $P(i|j)$ can be similarly found by setting $m = 1$ in Eq.17. Assuming that the alignment, $\gamma_{ji}(t)$ is unchanged during state tying, the auxiliary likelihood function for a state cluster $\mathcal{S}$ can be obtained as:

$$Q(\mathcal{S}) = \sum_{i \in \mathcal{R}} \sum_{j \in S} \beta_j P(i|j) \log P(i|\mathcal{S}) + K(\mathcal{S}) \quad (22)$$

where $\beta_j = \sum_{t,i} \gamma_{ji}(t)$ is the state occupancy, and the regression parameter for cluster $\mathcal{S}$ is given as

$$P(i|\mathcal{S}) = \frac{\sum_{j \in \mathcal{S}} \beta_j P(i|j)}{\sum_{i \in \mathcal{R}} \sum_{j \in \mathcal{S}} \beta_j P(i|j)} \quad (23)$$

the question is selected to maximize the following function:

$$\Delta Q_q = Q(\mathcal{S}_y(q)) + Q(\mathcal{S}_n(q)) - Q(\mathcal{S}) \quad (24)$$

where $\mathcal{S}$ is the initial state cluster, $\mathcal{S}_y(q), \mathcal{S}_n(q)$ are the split state clusters for "yes" and "no" answers, respectively. Given the fact that

$$K(\mathcal{S}_y(q)) + K(\mathcal{S}_n(q)) - K(\mathcal{S}) = 0 \quad (25)$$

the objective function, ΔQ_q will only depend on the regression parameters of each cluster, $P(i|\mathcal{S})$, $P(i|\mathcal{S}_y(q))$ and $P(i|\mathcal{S}_n(q))$. Although the above state tying algorithm is described for a TVWR system without context expansion as in Section.3, it can be easily extended to support multi-stream TVWR by replacing the objective function Eq.18 with Eq.13 and setting $m = 1$.

5. EXPERIMENTS

The experiments are conducted for two native (target) language recognition tasks: 1) 5k close vocabulary English speech recognition, 2) 22k open vocabulary Malay speech recognition. The full English dataset (WSJ0) contains 7k+ utterances (15 hours) with 84 speakers for training, and 330 utterances with 8 speakers for testing, while the full Malay dataset contains 35k+ utterances (74.5 hours) with 28 speakers for training, and 600 utterances with 6 speakers for testing. Both English and Malay corpora are reading speech recorded in clean environments. 39-dimension MFCC features are used for both corpora, including 13 static parameters and first two derivatives as dynamic parameters. A 3-state left-to-right HMM is applied as the acoustic model for each triphone, and tree based state tying is applied to cluster triphone states. To perform the recognition, both use full bigram decoding and trigram lattice rescoring, while four-gram lattice rescoring is additionally employed for Malay recognition.

For cross-lingual experiments, an 1.2 hours of English subset are extracted, including 500 utterances with 5 speakers. 6 hours of Malay subset are extracted, including 3k utterances with 6 speakers. Three foreign (source) language resources are employed, including Czech (CZ), Hungarian (HU) and Russian (RU). Specifically, three foreign phone recognizers [17] well trained by respective telephone speech data are employed. For clarification, English, Malay, CZ, HU, RU have 40, 33, 45, 61, 52 monophones, respectively. In order to use these three foreign phone recognizers, all speech waveform files were down-sampled to 8kHz, which are also used to extract acoustic features for consistency. In our experiments, phone recognizers were used to generate respective monophone/state posterior features instead of monophone/state sequence.

5.1. Baseline Mono-lingual Recognition

English HMM fullset baseline system is obtained with 3151 tied states and 16 mixtures per state; Malay HMM fullset baseline is estimated with 5043 tied states and 16 mixture per state. Due to limited data, English HMM subset baseline contains only 445 tied states and 8 mixtures per state, while Malay HMM subset baseline contains 1178 tied states and 4 mixtures per state, which will be the default number of components for all subsequent experiments if not mentioned explicitly. Performance degradation was observed by further increasing the number of mixtures using Maximum Likelihood criterion. In order to build a TVWR subset baseline, two neural networks are estimated to predict English and Malay monophone posterior features. Both neural networks are obtained by training a 3-layer neural network using the subset and quicknet [1] with $\delta = 4$ and 1000 hidden units.

[1] ICSI quicknet software package, `http://www.icsi.berkeley.edu/speech/qn.htm`

TVWR subset baseline is estimated by starting from respective HMM subset baseline and using respective monophone posterior features. As shown in Table.1, dramatical performance degradation was observed on both HMM and TVWR subset baseline systems. Although TVWR obtained consistent improvements over respective HMM subset baseline, its performance is still far from the HMM fullset baseline. These results clearly show that the performance of both HMM and TVWR is sensitive to the amount of available training data.

Tgt	HMM_full	HMM_sub	TVWR_sub
English	**6.9**	24.3	22.1
Malay	**13.1**	24.4	23.1

Table 1. *WER(%) performance of HMM and TVWR fullset/subset baseline systems for English and Malay speech recognition.*

5.2. Tandem Cross-lingual Recognition

To obtain tandem features, three foreign phone recognizers were used to generate respective monophone-state posterior features for the English and Malay subset. Log posterior features are then obtained for a more Gaussian-like distribution [18]. Principle Component Analysis (PCA) was applied to obtain 13-dimension features, which was concatenated to the original 39-dimension MFCC features. Tandem systems using 52-dimension features were estimated using two-model re-estimation and 4 iteration ML estimation. The best performance on English subset was found on the tandem systems with 8 mixtures per state, while Malay subset was with 12 mixtures per state. As shown in Table.2, different tandem system performs slightly differently but generally obtained significantly improvements over the HMM baseline. For English speech recognition, tandem systems using single foreign phone recognizer achieved 7-9% absolute improvements, while Russian language with best performance probably has more commons to the target English language. However, for Malay speech recognition, the absolute improvements is only 4% by using single foreign recognizer, while Hungarian seems more similar to the target Malay language. After combining three tandem systems, further 2-3% absolute improvements are observed for both English and Malay languages. These results show that tandem features using foreign language phone recognizers can help improve the performance of these two target languages with limited resources.

Tgt	CZ	HU	RU	CZ⊗HU⊗RU
English	16.7	16.3	15.4	**14.1**
Malay	20.4	19.9	20.1	**17.2**

Table 2. *WER(%) performance of various tandem systems with limited resources for target English and Malay speech recognition.*

5.3. TVWR Cross-lingual Recognition

As shown in Table.3, all TVWR systems using foreign posteriors outperformed both HMM and TVWR subset baselines in Table.1. When no context expansion is performed, 6-7% absolute improvements for English speech recognition over HMM subset baseline are observed using single stream of posterior features, while 3-4% are observed for Malay speech recognition. This tells that using a well trained foreign phone recognizer can provide a better partition for the acoustic space for TVWR. However, when compared to tandem systems in Table.2, TVWR without context expansion is consistently inferior to tandem systems. This may be because TVWR depends more on unreliable GMM by MFCC features.

Tgt: English	w/o temporal	w/ temporal
CZ	17.7	13.0
HU	17.8	11.9
RU	17.9	13.4
spatial context	12.1	**9.8**

Tgt: Malay	w/o temporal	w/ temporal
CZ	21.9	18.1
HU	20.8	17.7
RU	18.0	16.7
spatial context	16.2	**14.5**

Table 3. *WER(%) performance of TVWR systems with or without context expansion for target English and Malay speech recognition.*

After applying temporal/spatial context expansion, multi-stream TVWR consistently outperformed both conventional TVWR systems and tandem systems. When compared to TVWR without context expansion, 4-6% absolute improvements over respective TVWR using single foreign posteriors are observed for English language, while 3-4% absolute improvements are observed for Malay language. This shows that temporal context expansion can significantly improve TVWR system performance without suffering over-fitting issue despite introducing many regression parameters. When compared to the individual tandem systems in Table.2, 2-4% absolute improvements are shown for English, while 2-3% absolute improvements for Malay. Particularly, TVWR using single HU for English and single RU for Malay already shows better than multiple stream tandem systems. These results show that multi-stream TVWR with temporal context expansion can learn more information from single stream of posterior features. TVWR with spatial context expansion by three languages performs similar to the best temporal context expansion based TVWR, i.e. HU for English and RU for Malay, which shows spatial context expansion may have a more robust acoustic partition, while temporal context expansion is more sensitive to the difference between source and target languages. After combining both temporal and spatial

context, another 1-2% absolute improvements are observed, which tells that temporal and spatial context are different and complementary.

5.4. State Tying for TVWR Parameters

Last, multi-stream TVWR with a second state tying method is evaluated. In order to obtain strong Gaussian bases for TVWR system, the number of tied states for GMM parameters is reduced to about 330 for English (8 mix per state) and 900 for Malay (4 mix per state), while the number of tied states for TVWR parameters increased to about 1.2-1.3k for English and 3.0-3.4k for Malay. When combining both temporal and spatial context, 1.9k and 4.8k tied states for TVWR parameters are used for English (8 mix per state) and Malay (8 mix per state), respectively. Recognition results for various TVWR systems are reported in Table.4. First, after introducing a second state tying method, consistent improvements are found for all TVWR systems. However, the amount of improvements varies as foreign language. Generally speaking, foreign posteriors with better performance in Table.3 can gain more by introducing more tied states. Since the rational of introducing more tied states for TVWR parameters is that posteriors are more reliable than acoustic features, this method may not work well if foreign posteriors are not reliable enough. Finally, combination of temporal and spatial context with more tied states achieved very close performance to the HMM fullset baseline, i.e. 1-2% difference. However, it is important to note that discriminative training is applied to obtain posterior features for TVWR, which can definitely lead to a better HMM baseline.

Tgt	w/ temporal			spatial	temporal +spatial
	CZ	HU	RU		
English	11.7	11.3	12.6	10.6	**8.7**
Malay	18.0	16.9	15.6	14.9	**13.9**

Table 4. *WER(%) performance of various multi-stream TVWR systems with a second state tying and limited resources for target English and Malay speech recognition.*

6. CONCLUSIONS

In this paper, the recently proposed TVWR is investigated for cross-lingual speech recognition under limited resources. First, various foreign monophone/state posterior features are employed to replace the native unreliable features so that a better acoustic partition can be obtained. Second, multi-stream TVWR is proposed by incorporating much richer temporal and spatial context information for a better representation of the context variable. Third, a separate state tying algorithm for the TVWR regression parameters is proposed to introduce more distinct triphone state with reliable regression parameters. Various TVWR systems were evaluated for English and Malay speech recognition with limited resources. TVWR systems using foreign monophone/state posterior features have shown significant improvements over both HMM and tandem systems. Introducing multi-stream TVWR and more tied states can obtain further improvement, which results in less than 2% inferior to respective English and Malay HMM fullset baselines.

7. REFERENCES

[1] W. Byrne, P. Beyerlein, J.M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and T. Wang, "Towards language independent acoustic modeling," in *Proceedings of ICASSP*, 2000.

[2] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, et al., "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Proceedings of ICASSP*, 2010, pp. 4334–4337.

[3] H. Lin, L. Deng, D. Yu, Y-F Gong, A. Acero, and C-H Lee, "A study on multilingual acoustic modeling for large vocabulary asr," in *Proceedings of ICASSP*, 2009, pp. 4333–4336.

[4] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace gaussian mixture models for cross-lingual speech recognition," in *Proceedings of ASRU*, 2011, pp. 365–370.

[5] V.B. Le and L. Besacier, "First steps in fast acoustic modeling for a new target language: application to vietnamese," in *Proceedings of ICASSP*, 2005, pp. 821–824.

[6] Dong Yu, Li Deng, Peng Liu, Jian Wu, Yifan Gong, and Alex Acero, "Cross-lingual speech recognition under runtime resource constraints," in *Proceedings of ICASSP*, 2009, pp. 4193–4196.

[7] K.C. Sim and H. Li, "Context sensitive probabilistic phone mapping model for cross-lingual speech recognition," in *Proceedings of Interspeech*, 2008, pp. 2715–2718.

[8] K.C. Sim and H. Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *Proceedings of ICASSP*, 2008, pp. 4309–4312.

[9] K.C. Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *Proceedings of ASRU*, 2009, pp. 546–551.

[10] S. Dupont, C. Ris, O. Deroo, and S. Poitoux, "Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents," in *Proceedings of ASRU*, 2005, pp. 29–34.

[11] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multistream posterior features for low resource lvcsr systems," in *Proceedings of Interspeech*, 2010.

[12] A. Stolcke, F. Grézl, M-Y Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proceedings of ICASSP*, 2006.

[13] P. Lal, *Cross-lingual Automatic Speech Recognition using Tandem Features*, Ph.D. thesis, University of Edinburgh, 2011.

[14] S. Liu and K.C. Sim, "Implicit Trajectory Modelling Using Temporally Varying Weight Regression for Automatic Speech Recognition," in *Proceedings of ICASSP*, 2012, pp. 4761–4764.

[15] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of HLT*, 1994, pp. 307–312.

[16] G. Wang and K.C. Sim, "An investigation of tied-mixture gmm based triphone state clustering," in *Proceedings of ICASSP*, 2012, pp. 4717–4720.

[17] "Phoneme recognizer based on long temporal context," in *Brno University of Technology, http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-author-context*.

[18] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proceedings of ICASSP*, 2000, pp. 1635–1638.