COMBINATION OF DATA BORROWING STRATEGIES FOR LOW-RESOURCE LVCSR

*Yanmin Qian, Kai Yu**

Institute of Intelligent Human-Machine Interaction
MOE-Microsoft Key Lab. Of Intelligent Computing and Intelligent Systems
Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

Jia Liu†

Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering
Tsinghua University, Beijing, China

ABSTRACT

Large vocabulary continuous speech recognition (LVCSR) is particularly difficult for low-resource languages, where only very limited *manually transcribed* data are available. However, it is often feasible to obtain large amount of *untranscribed* data of the low-resource *target language* or sufficient transcribed data of some *non-target* languages. Borrowing data from these additional sources to help LVCSR for low-resource language becomes an important research direction. This paper presents an integrated data borrowing framework in this scenario. Three data borrowing approaches were first investigated in detail, including feature, model and data corpus. They borrow data at different levels from additional sources, and all get substantial performance improvements. As these strategies work independently, the obtained gains are likely additive. The three strategies are then combined to form an integrated data borrowing framework. Experiments showed that with the integrated data borrowing framework, significant improvement of more than 10% absolute WER reduction over a conventional baseline was obtained. In particular, the gain under the extreme limited low-resource scenario is 16%.

Index Terms— Data borrowing, Low resource speech recognition, Articulatory feature, Subspace Gaussian mixture models, Unsupervised training

1. INTRODUCTION

The performance of speech recognition systems has improved dramatically. Training acoustic models for state-of-the-art systems often require large amount of language-specific *transcribed* speech data. However, demand exists for speech recognition systems in languages which have only limited training data available [1] [2], and in these cases performance is still quite poor. The expensive cost for transcribing audio data makes the data sparseness the most pressing challenges in this scenario and is further exacerbate by the fact that today's speech technologies heavily rely on statistical models, such as hidden Markov model (HMM) and neural network (NN) models.

Obviously there are normally four types of speech data corpus in the real low-resource environment. As Fig 1 illustrated, the entire data resource space can be divided into four separated parts:

- Manually transcribed target (low-resource) language data
- Untranscribed target language data
- Manually transcribed non-target (normally rich-resource) language data
- Untranscribed non-target language data

These parts are denoted as V1∼V4, and the area of each part in Fig 1 approximates the real quantity of each data sources. Traditionally only the V1 part (transcribed target) is used to develop the speech recognition system. However since the amount of V1 is limited for the low-resource target language, the constructed system only using V1 often performs poorly. From Fig 1, there are relatively tremendous amount of data in the other three parts. Accordingly, researchers have tried to develop approaches that could make models particularly elaborate or strategies that could utilize not only the transcribed target language data but also the data in the other parts.

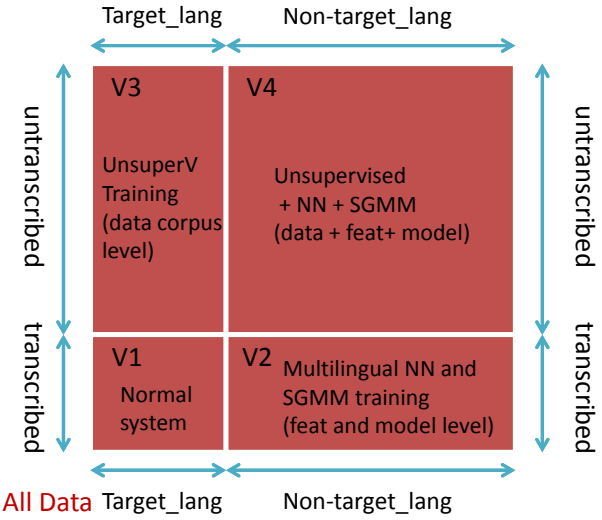

Fig. 1. Data corpus distribution in low-resource scenario.

Several strategies have been previously proposed to address low-resource problems: one main branch can take advantage of transcribed data from other languages to build multilingual acoustic models, including the universal phone based multilingual ASR method [3] [4], multilingual trained Subspace Gaussian Mixture (SGMM) modelling [5] [6] and multi-layer perceptron (MLP) based data-driven approaches with training process in a multilingual or cross-lingual mode [7] [8]. In addition unsupervised [9] or lightly-supervised [10] training is another type of popular strategy which could enlarge the size of target language data quickly and cheaply [11] [12]. The multilingual or cross-lingual approaches borrow

*Yanmin Qian and Kai Yu were supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning and the China NSFC project No. 61222208.

†Jia Liu was supported by the Project 2009BAH41B00 and 2009BAH41B01 of National Science and Technology Pillar Program of China, the Project 90920302 and 61273268 of NSFC.

data from the V2 (transcribed non-target language data), and the unsupervised training usually develop technologies to borrow data from the untranscribed target language data, i.e. V3 part of Fig 1. However, these methods are mostly isolated developed. To our knowledge, there has not been any efforts to investigate data borrowing strategies systematically in the low-resource scenarios, from all the parts of the data resources.

In this paper, data is borrowed from all types of resources as illustrated in Fig 1. First, several data borrowing strategies are revisited and investigated, which focus on different modules of speech recognition system respectively. Three different approaches are proposed to borrow data from target or non-target, transcribed or untranscribed data individually (V2 and V3 parts), including feature, model and data corpus level, to obtain more discriminative feature or robust models for the low-resource applications.

As these approaches work at different levels, they are relatively independent of each other and the gains may be additive. With this assumption, a fully integrated data borrowing strategy is proposed to combine these approaches ideas tightly into one united framework. Moreover, with the integrated framework, it is possible to extend existing approaches to borrow data from untranscribed non-target languages data corpus more easily and flexibly (V4 part of Fig 1). This reveals new possibilities of data borrowing. As far as we know, this work may be the first comprehensive attempt to investigate data borrowing strategy using all data resources for the low-resource speech recognition.

The remainder of this paper is organized as follows. Section 2 describes individual data borrowing approaches, including feature, model and data corpus level refinement. Then, the integrated data borrowing framework is proposed in Section 3. In Section 4, experimental setup, results as well as detailed analysis are given. Section 5 concludes the paper and discusses future research directions.

2. INDEPENDENT DATA BORROW STRATEGIES AT THREE LEVELS

In this section three effective different data borrowing strategies, focusing on feature, model and data corpus level are reviewed in detail.

2.1. Feature Level: Multilingual Articulatory MLP

The universal phone set is a popular multilingual method [4], however the phone mapping or clustering across languages induces confusion among models, and it also needs tremendous large quantity of training languages data to get complete coverage of a universal phone set. Articulatory features (AF) are alternative modeling units. Articulatory Features [13] have been demonstrated as more fundamental units shared across languages than phones, since they are independent of the underlying language. Table 1 shows an example of the mapping between some phones and AFs in English. All phones can be modeled by a set of articulatory attributes and each attribute is possessed by a set of phones.

Table 1. Mappings between articulatory features and phones

AFs	Phones	Phones	AFs
Fricative	jh ch s sh z f th v dh	f	Fricative Labial
Nasal	m n ng	th	Fricative Dental
Labial	b f m p v w	aa	Vowel Low Back
.....	.....	.....	.....

In the main setup of this work, English is selected as the target language, and Spanish and German as the non-target languages. 28 AFs for English, 29 for German and 27 for Spanish are used. The articulatory features across three languages, classified according whether it is universal or language-specific, are shown in Table 2. There are in total 40 articulatory features (AFs) for the three languages, of which 24 AFs are common. The number of unique AFs for each language is small, only 6 for English, 5 for German and 5 for Spanish. It is found that more than 80% of the articulatory features of English are shared with German and/or Spanish. Accordingly AF mapping across languages is more robust than the traditional phone mapping and borrowing data from non-target languages on AFs becomes more reliable than phones.

Table 2. The universal and the unique language-specific articulatory features (AF) for English, Spanish and German

Language	Articulatory features
Universal	Alveolar, Approximant, Back, Bilabial, Central, Close, Consonant, Fricative, Front, Glottal, Labiodental, Middel, Nasal, Open, Palatal, Plosive, Postalveolar, Rounded, Silences, Spause, Unrounded, Velar, Voiced, Vowel
English	Liquid, Obstruent, Semi-vowel, Sonorant, Stop, Unvoiced
German	Diphthongs, Long, Open-mid, Short, Uvular
Spanish	Labial-velar, Lateral, Tap, Trill, Voiceless

Based on the articulatory feature units, AF based MLP system is constructed as Fig 2, similarly to the front-end of the system in [14] [1]. It mainly consists two main blocks:

1. **Articulatory NNs**, which consist of a bank of speech event detectors and produce the posterior probability of the articulatory attribute.
2. **Phone MLPs**, which take as input the outputs of the articulatory-feature detectors, and are trained to classify phones.

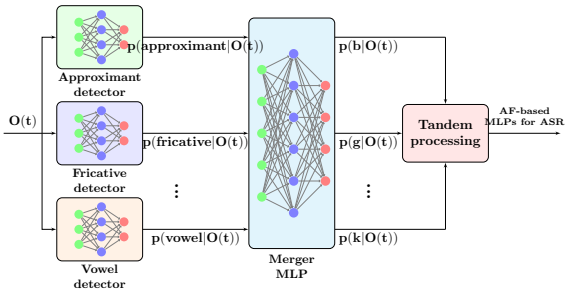

Fig. 2. Articulatory feature based MLPs framework.

Taking advantage of the multilingual articulatory features, multilingual networks are trained to obtain the universal AF detectors. During this training, data from both the target and non-target languages are pooled. Since the AF detectors are trained with much more data than is available in the target language, they are much more robust and discriminative for AF classification. Using these robust estimated detectors, the phone NN classifier also produces more accurate performance. The normal tandem processing [15] is then applied on the phone merger outputs to generate the MLP features. This AF framework can be further extended to generate more

[1] Note that in contrast to [14], we are using the articulatory detectors in the MLP feature extraction phase rather than as probabilities in the model.

elaborated multilingual MLP features [16]. In this work, the multilingual AFs and NN models are used to borrow data from the transcribed non-target languages to improve the low resource system at *feature level*.

2.2. Model Level: Multilingual SGMM

The Subspace Gaussian mixture models have a more compact representation than GMMs. It is a special case of the *canonical state* model framework [17], which combines phone-specific and phone-independent information in an adaptive training fashion. General form of the SGMM can be expressed as:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} \omega_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i) \quad (1)$$

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j \quad (2)$$

$$\omega_{ji} = \frac{\exp(\mathbf{w}_i^T \mathbf{v}_j)}{\sum_{k=1}^{I} \exp(\mathbf{w}_k^T \mathbf{v}_j)} \quad (3)$$

where $p(\mathbf{x}|j)$ is the state output distribution of feature vector $\mathbf{x}$ at HMM state j. The model is a mixture of Gaussians, but unlike the conventional GMM, the number of mixture components I is the same for all states and is typically quite large, e.g. several hundred. The covariance $\boldsymbol{\Sigma}_i$ for each Gaussian in the mixture is globally shared across states (full covariance matrices are used here). The most important difference is that the mean $\boldsymbol{\mu}_{ji}$ and mixture weights ω_{ji} are not direct parameters of the model, and instead they are expanded from a state-specific vector $\mathbf{v}_j$, via globally shared parameters $\mathbf{M}_i$ and $\mathbf{w}_i$, as illustrated in Equations (2) and (3).

A well-tuned SGMM typically has fewer parameters than a well-tuned GMM system [6]. Moreover, the majority of the parameter count in a SGMM system consists of shared parameters $\mathbf{M}_i$, $\boldsymbol{\Sigma}_i$ and $\mathbf{w}_i$, which can be 8$\sim$10 times larger than the state-specific parameters $\mathbf{v}_j$. This leads to a natural method of training SGMMs in a multilingual way: the state-specific parameters are trained as separate language-specific states, and the common SGMM parameters are, however, shared across languages. In addition to sharing global parameters, it has also been shown that it is more advantageous to also share the state-specific parameters across languages, using an accurate distance calculation metric with state occupation consideration [18].

With the multilingual intrinsic ability, in this work, SGMM is used to borrow data from the transcribed non-target languages data to improve systems at *model level*.

2.3. Data Corpus Level: Unsupervised Training

Compared to manual transcriptions, raw audio is more easily collected and relatively much cheaper and less time consuming, so a large amount of speech data is presented as the V3 part of Fig 1. In this scenario, unsupervised training will be an effective approach, which gains more and more popularity [11] [12]. Typical procedure of unsupervised training includes using a seed model, trained on a small amount of manual transcribed corpus, to recognize a large quantity of unlabelled speech data. The recognized hypothesis should be filtered firstly and then pool the chosen hypothesis transcriptions with manual transcriptions to retrain the acoustic model.

Efficient and effective data selection method is crucial in unsupervised training, due to the fact that there are many recognition error in the hypothesis transcriptions. Confidence scores are normally used for data selection. In this paper, in addition to confidence scores, phone frequencies are also used for data selection [19]. When checking the accuracy of the initial recognized transcriptions, we will be more inclined to select the data with lower phone frequency in the limited manually labelled corpus.

Most unsupervised training approaches focus on modifying HMM acoustic models [11] [12]. However, it is also possible to extend the idea to the feature-level. A NN-based unsupervised training method for robust MLP feature extraction has been proposed [19]. In this approach, an enlarged transcribed speech corpus is used to train a neural network for feature extraction, rather than retrain the HMM as usual.

In this paper, the NN-based and HMM-based unsupervised training approaches are used together [19]. First, data are borrowed from the unlabelled target languages corpus to unsupervised train NN and get robust MLP features, then the acoustic model are re-constructed using MLP feature and hypotheses are re-generated. Finally, the data from the unlabelled data are borrowed for the second time to retrain the HMM model and get more improved performance. In summary, in this paper, refined unsupervised training approach is used to borrow data from the untranscribed target language data, i.e. V3 part of Fig 1.

3. COMBINATION OF DATA BORROWING STRATEGIES

Several state-of-the-art data borrowing strategies have been reviewed in the previous section. They focus on distinct independent modules of speech recognition system, and borrow data from transcribed non-target languages data or untranscribed target-language data respectively, V2 and V3 parts of Fig 1. Due to different working level, these approaches are relatively independent of each other. It is then interesting to investigate the combination effect of these strategies. In this section, the different data borrowing strategies on three levels are combined into one unified framework to achieve more efficient and effective data borrowing. The final architecture makes data borrowing more easily and flexibly, and particularly useful for the low-resource speech recognition. Moreover the integrated framework reveals new possibility to borrow data from *all* parts of data resources space. For example, the untranscribed non-target languages data corpus may be exploited (V4 part of Fig 1) in future work.

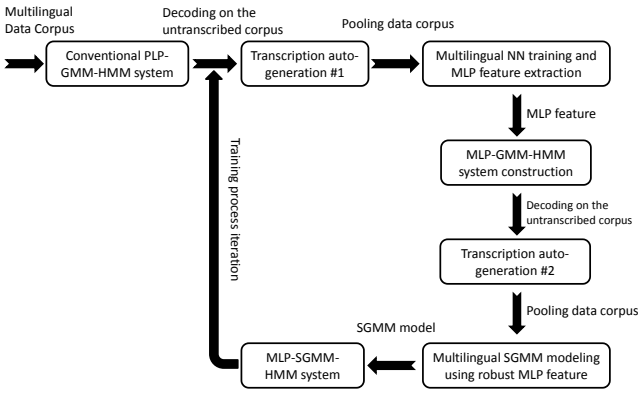

Fig. 3. Fully combined data borrowing framework.

The flow chart of the integrated data borrowing framework is shown in Fig 3. And the detailed algorithm is described in algorithm 1. To our knowledge, this is the first attempt to form a unified framework of borrowing data at all working levels.

Algorithm 1 DATA BORROWING STRATEGY COMBINATION

1. Initial GMM-HMM system construction:
Build an initial GMM-HMM systems for each target and non-target language separately using small amount of manually transcribed data. PLP feature is used for all initial systems here.

2. Automatic Transcription Generation #1:
Use the language-specific initial GMM-HMM systems to recognize the unlabelled target or non-target languages speech utterances respectively. The results are denoted as *initial hypotheses #1*. Then use the data selection approach in section 2.3 to filter the hypotheses, and pool the selected hypotheses with the manually transcribed utterances to form an enlarged target and non-target languages corpus #1.

3. Multilingual NN training and MLP feature extraction:
Use the enlarged target and non-target languages corpus #1 to train NN as described in section 2.1, and then extract MLP features.

4. MLP-GMM-HMM system construction:
Use the robust MLP features obtained from procedure 3 to train an initial MLP-GMM-HMM models.

5. Transcription auto-generation #2:
Use the MLP-GMM-HMM systems in procedure 4 to recognize the unlabelled target or non-target languages speech utterances for the second time, and obtain new hypotheses #2. Again use the data selection method in section 2.3 to filter the new hypothesis, and pool the new selected hypothesis utterances with the manually transcribed utterances to form the enlarged target and non-target languages corpus #2.

6. Multilingual SGMM modeling using robust MLP feature:
Use the enlarged target and non-target languages corpus #2 to train the SGMMs as described in section 2.2. With the MLP feature extracted from procedure 3, the multilingual trained MLP-SGMM-HMM system can be built.

7. Training process iteration:
Go back to step 2 and repeat the whole process to refine the system. This step is optional.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Data and Baseline System

In the experiments, Callhome multilingual databases and Switchboard I English corpus are used. The conversational nature of speech in these databases along with high out-of-vocabulary rates, the use of foreign words and the telephone channel distortions make the task of speech recognition on this database challenging.

In the first part of the experiments, English was selected as the target language. Two groups of systems with different configurations were constructed to mimic the realistic low-resource environments, whose data details are described in Table 3.

Configuration #1 : Extremely low-resource situation. Only 1 hour of randomly chosen Callhome English were selected as transcribed data and the retained 14 hours as the unlabelled data.

Configuration #2 : Normal low-resource situation. All of the 15 hours of Callhome English were selected as labelled data and 100 hours of randomly chosen Switchboard English as the untranscribed data.

In both configurations, the entire 15 hours of German and 16 hours of Spanish training data were used as the non-target language data. The test set consists of 20 conversations of the Callhome English evaluation set, roughly containing 2 hours of speech.

From the description, training data resources comprises three parts:

- Limited transcribed target language data
- Plenty of transcribed non-target language data
- Plenty of untranscribed target language data

These correspond to the V1$\sim$V3 parts of Fig 1. Data borrowing from the untranscribed non-target language data, V4 part, is not evaluated here due to lack of data in the current setup. It will be investigated in the future.

Table 3. Data resources for two experimental configurations

	Lang	Trans	Corpus	Amount (hr)
Configuration #1				
Target	English	manual	CHE	1
		none	CHE	14
Configuration #2				
Target	English	manual	CHE	15
		none	SWB I	100
Common parts				
Non-target #1	German	manual	CHG	15
Non-target #2	Spanish	manual	CHS	16
Test set	English	manual	CHE	2

The baseline GMM-HMM systems were built using 39 dimensional PLP features with energy, first and second derivatives, plus per-speaker mean and variance normalization. After state-clustering, there are 550 tied states with 4 Gaussians per state in config #1; 1930 states with 16 Gaussians per state in config #2. The SRILM tools [20] were used to build a trigram language model with a wordlist of 62K words. The trigram model is an interpolated model, where the individual components were respectively trained on English Callhome corpus, the Switchboard corpus and the Gigaword corpus. HDecode and Kaldi decoders were used to decode the GMM or SGMM model respectively.

In this study, only Maximum Likelihood (ML) parameter estimation is considered because the main focus is the evaluation of data borrowing strategies. The proposed methods could also be extended to discriminative training.

The first line of Table 4 show the performance of baseline systems which only use the labelled target language data, and it is comparable to other people's work [5] [7]. It is clear that the ASR systems built with low resource perform poorly. The conventional approach relies on the amounts of labeled target language data heavily. The proposed strategies aim to relieve the demanded data quantity, and improve system performance for low resource scenarios.

4.2. Performance of Individual Data Borrowing Strategies

Systems using the distinct data borrowing strategy as described in Section 2 were first built, including the multilingual AF based MLP method, multilingual SGMM based method and unsupervised training approach. As the Table 4 illustrates, all the three data borrowing strategies obtained large improvement on each level of speech recognition system (more than 10% relative improvement). The AF-based and SGMM-based methods borrow data from the cross-lingual non-target languages, and the unsupervised approach generates new data from the untranscribed target language data.

4.3. Performance of Data Borrowing Strategy Combination

Finally, the data borrowing strategies on different levels were combined as described in section 3. Only one iteration of combination was used in this experiment for simplicity. The last line in Table 4 shows that the integrated data borrowing framework obtained more

Table 4. Performance comparison (WER) of individual data borrowing strategies

System	Data Borrowing	Conf#1	Conf#2
Baseline	–	72.6%	55.2%
S1	Feature	64.1%	49.7%
S2	Model	60.0%	45.9%
S3	Untranscribed Tgt Data	64.3%	48.6%
Combination	S1⊕S2⊕3	**56.5%**	**43.1%**

than 10% absolute WER reduction. Particularly in the extreme limited low-resource environment (Conf#1), the gain is even larger than absolute 16%. This is much better than the recently reported results using the similar configuration in [5] and [8]. Based on GMM-HMM systems, the proposed combination approach even achieved the same performance improvement of using the DNN technology [21].

Fig 4 shows a performance comparison of all the data borrowing methods investigated in this paper. Compared to traditional systems, the proposed approaches show substantial improvements. Moreover the gains from the individual strategy are additive, and the fully combined data borrowing approach achieves the best overall performance, with significant improvements over individual strategy.

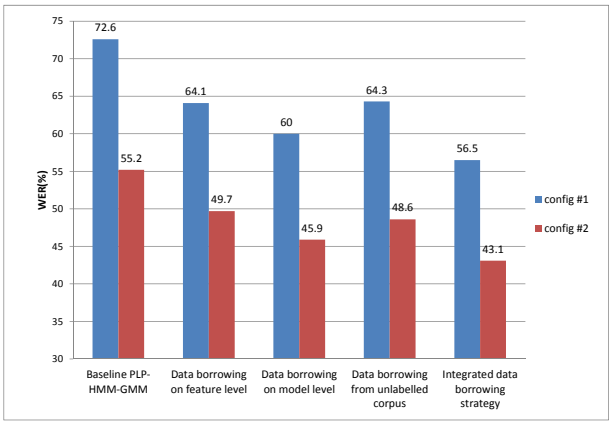

Fig. 4. Performance comparison of the proposed data borrowing strategies investigated in this paper.

4.4. Resource Saving Evaluation

To easily evaluate that how much resource could be saved when using the paper proposed methods, conventional PLP-GMM-HMM systems with different amounts of transcribed target language training data were built. The data are randomly chosen from Callhome and Switchboard I English corpus. The performance of the baseline systems and the proposed methods in Configuration #2 are illustrated in Fig 5.

It can be seen that the three level approaches achieved the same performance of baselines with about 50 hours, 100 hours, and 50 hours training data respectively (borrowing data on feature, model or data corpus), and our fully combined data borrowing strategy, which utilizes only 15 hours transcribed target language data, achieved better performance than the conventional system using 150 hours, and even approached the performance of the 200 hours baseline. The resource saving percentage is larger than 90%.

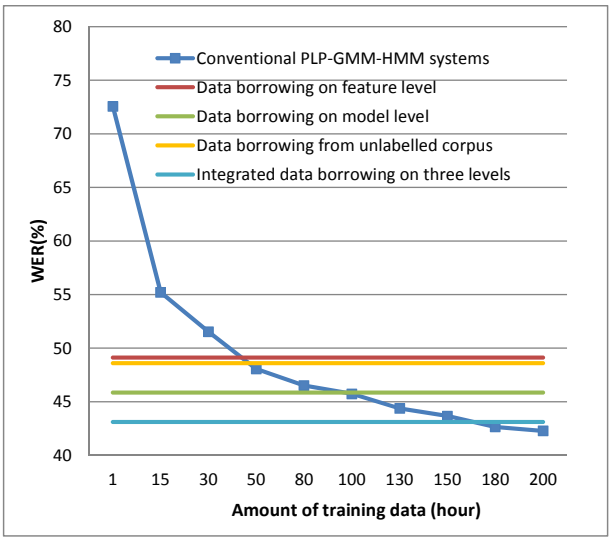

Fig. 5. Performance comparison of conventional systems using different amounts of training data and the data borrowing strategies proposed in this paper

4.5. Cross-Languages Validation

To validate that the methods proposed here are language-independent, the target language is changed from English to other languages. In this paper, Spanish, German and Uigur were chosen as the target languages. The entire Callhome English, Spanish, German corpus and the authors' lab collected Uigur data were used as the transcribed data. Each language has roughly 15 hours training data and 2 hours test set.

As in English, language-specific conventional baseline PLP-GMM-HMM systems were built for each language using the limited transcribed target language data. The model size was tuned to obtain the best performance for each language. Maximum Likelihood criterion was also adopted as before.

Due to lack of word-level transcriptions and language models, phone recognition was performed for the four languages. Since the work mainly focus on acoustic modelling, not the language modelling, we believe phone recognition is sensible for our purpose. Phone reference transcripts were obtained using forced alignment performed with the language-specific baseline systems. The language-specific phone bigram language models were trained on forced aligned phone transcripts of acoustic training data.

In the multilingual validation, each language was selected as the target language in turn, and the other three as the non-target languages. In the experiments, only MLP feature level and SGMM model level were performed because there were no untranscribed data in Spanish, German and Uigur. The procedures in section 3 was slightly modified to remove the unsupervised training procedures and firstly train to obtain multilingual MLP features and then use these features to build the multilingual trained SGMM model. Data from the transcribed cross-lingual non-target languages data were borrowed.

Table 5 lists the phone error rate (PER) of different systems on the four languages. It can be seen that data borrowing strategies achieved consistent improvements on the four languages in low-resource scenarios. Compared to the single level refinement, combination of data borrowing strategies obtained larger improvement, which is the same as the English case.

It can also be observed that the improvement on Uigur is rela-

tively smaller than the other three. This is because all the other three languages belong to the Indo-european family, and Uigur belongs to the Altaic family. Different language families will influence the effectiveness of data borrowing. Even with this inferiority of Uigur, the improvement is still clear compared to the baseline system. The four languages validation experiments showed that the proposed approaches are language-independent. They can be extended to other languages easily.

Table 5. Phone error rate (PER) comparison of systems for four target languages

Data Borrowing	English	Spanish	German	Uigur
—	56.4%	47.3%	57.6%	73.5%
MLP	49.9%	44.8%	54.2%	71.8%
SGMM	50.4%	44.2%	53.3%	71.4%
MLP⊕SGMM	**48.4%**	**42.7%**	**52.1%**	**70.7%**

5. CONCLUSION AND FUTURE WORK

In this paper, an integrated data borrowing strategy framework is proposed for low-resource speech recognition. Three types of data borrowing approaches at different levels are reviewed, including feature, model and data corpus. Then these data borrowing strategies are combined into an unified framework to make the data borrowing more efficiently and effectively. The integrated strategy enables researcher to borrow data from all the parts of data space more easily and flexible. Experiments showed that the combination obtained more than 10% absolute WER reduction for the low-resource speech recognition.

Moreover the consistent performance improvement on the other four languages also illustrate that the proposed methods are language-independent. They can be extended to other languages easily. The results on the Uigur experiment also indicates that it is important and effective to select linguistically close languages to borrow the data for the low-resource scenario.

When applying the proposed fully integrated data borrowing strategy, it is possible to use only 15 hours transcribed target-language data to exceed the traditional system using 150 hours manually transcribed speech corpus. The result is even approaching the baseline using 200 hours of data, in which case the resource saving percentage is larger than 90%. It demonstrated that using the novel combination of data borrowing strategies can significantly reduce the manual effort of transcribing speech data, save costs and accelerate the development for low-resource speech recognition.

In the future we hope to combine these ideas with the recent popularized deep neural network technique [22], and use several hundreds of hours of multilingual data and more untranscribed data to get a better system.

6. REFERENCES

[1] P. Fung and T. Schultz, "Multilingual spoken language processing," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 89–97, 2008.

[2] X. Cui, J. Xue, P. L. Dognin, U. V. Chaudhari, and B. Zhou, "Acoustic modeling with bootstrap and restructuring for low-resourced languages," in *INTERSPEECH*, 2010.

[3] B. Walker, B. Lackey, J. Muller, and P. Schone, "Language-reconfigurable universal phone recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[4] 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 *ICASSP*. IEEE, 2009, pp. 4333–4336.

[5] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *ICASSP*. IEEE, 2010, pp. 4334–4337.

[6] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow *et al.*, "The subspace gaussian mixture modelła structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

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

[8] ——, "Multilingual mlp features for low-resource lvcsr systems," in *ICASSP*. IEEE, 2012, pp. 4269–4272.

[9] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *ICASSP*, vol. 3. IEEE, 2006, pp. III–III.

[10] M. Paulik and P. Panchapagesan, "Leveraging large amounts of loosely transcribed corporate videos for acoustic model training," in *ASRU*. IEEE, 2011, pp. 95–100.

[11] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for lvcsr," *Speech Communication*, vol. 52, no. 7, pp. 652–663, 2010.

[12] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *ICASSP*. IEEE, 2011, pp. 4656–4659.

[13] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *ICASSP*, vol. 1. IEEE, 2003, pp. I–144.

[14] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," in *ASRU*. IEEE, 2007, pp. 566–569.

[15] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, vol. 3. IEEE, 2000, pp. 1635–1638.

[16] Y. Qian and J. Liu, "Articulatory feature based multilingual mlps for low-resource speech recognition." in *INTERSPEECH*, 2012.

[17] M. J. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Proc. Interspeech*, 2010, pp. 58–61.

[18] Y. Qian, D. Povey, and J. Liu, "State-level data borrowing for low-resource speech recognition based on subspace gmms," in *INTERSPEECH*, 2011.

[19] Y. Qian and J. Liu, "Mlp-hmm two-stage unsupervised training for low-resource languages on conversational telephone speech recognition," in *INTERSPEECH*, 2013, to appear.

[20] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit," in *Proceedings of the international conference on spoken language processing*, vol. 2, 2002, pp. 901–904.

[21] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *ICASSP*. IEEE, 2013, pp. 6704–6708.

[22] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.