



# An Evaluation of Cross-Language Adaptation and Native Speech Training for Rapid HMM Construction Based on Very Limited Training Data

Xufang Zhao, Douglas O'Shaughnessy

Institut natinal de la recherche scientifique, Université du Québec

zhaoxf@emt.inrs.ca, dougo@emt.inrs.ca

## Abstract

As the needs and opportunities for speech technology applications in a variety of languages have grown, methods for rapid transfer of speech technology across languages have become a practical concern. Previous works focus on the comparison of different adaptation algorithms, for example, MAP (Maximum A Posterior), Bootstrap, and MLLR (Maximum Likelihood Linear Regression) on speaker adaptation. However, a very interesting point is that, with increasing adaptation corpora, the performance of direct native speech training may already exceed the performance of cross-language adaptation. If it is true, there should be a threshold for the size of an adaptation corpus. In general, transferring acoustic knowledge is useful when there is not enough training data available. This paper presents a systematic comparison of the relative effectiveness of cross-language adaptation and native speech training, using transfer from English to Mandarin as a test case. This study found that cross-language adaptation does not produce better acoustic models than the direct native speech training approach even using limited training data.

**Index Terms:** cross-language adaptation, multilingual acoustic model

## 1. Purpose and Previous Works

Cross-language adaptation could transfer acoustic knowledge from a source language to a target language, and previous works concentrated on the comparison of different adaptation algorithms, for example, MAP (Maximum A Posterior) [1], Bootstrap [2], and MLLR (Maximum Likelihood Linear Regression) on speaker adaptation [3]. This technique should be very useful when there is only a small amount of training data available for a target language. If there is only a small amount of adaptation data available, MLLR adaptation should be more effective than MAP; [4] compared MLLR and MAP on non-native speech adaptation. As expected, MAP adaptation would be the better choice as long as there is enough adaptation data. In [5], Kohler compared MAP, Bootstrapping, and scratch training in cross-language adaptation, and his results proved that with more than 1000 utterances the scratch training shows better recognition results than MAP adaptation. There is a lack of work comparing the performance of different adaptation techniques with direct native speech training on Mandarin speech recognition. Schultz et al. [6] presented their recognition results using language-dependent models, language-independent models, and language adaptation acoustic models. They compared the performance of Portuguese recognition using different amounts of adaptation data, from 25 minutes to 90 minutes, and this experiment indicated that with an increase of adaptation data, the WER (Word Error Rate) goes down. The purpose of this paper was a comparison between cross-

language adaptation and native speech training based on a limited training set. It is true that the performance of 90 minutes adaptation is better than the performance of 15 minutes adaptation; however, if 90 minutes direct training is already better than 90 minutes adaptation, why do adaptation? So it is meaningful to find a balance point between direct training and adaptation, and the comparison work of native speech training and cross-language adaptation should be worth trying. [7] applied the MAP algorithm to adapt an English recognizer to a Chinese recognizer. The minimum size of adaptation set in [7] was  $10^5$  phonemes, and the maximum adaptation set contained  $10^{10}$  Mandarin phonemes.  $10^{10}$  phonemes is large enough even for direct native speech training. There is a lack of work on comparison of native speech training and cross-language adaptation only using a small amount of training set.

Previous works of [4] already proved that MAP only takes effect when there are enough adaptation data, and MLLR plays an important role in limited training corpora usage. Our testing case is using a MLLR adaptation algorithm adapting English acoustic models to Mandarin acoustic models.

## 2. Speech Corpora

The Mandarin training and test data are drawn from HUB-4NE [8]. All contributors to the corpus are native speakers of Mandarin. This collection consists of 30 hours of recorded broadcasts and transcripts that have been drawn from the Voice of America (VOA), People's Republic of China Television (CCTV), and Commercial radio based in Los Angeles, California. The reason for choosing the HUB-4NE database is that some research teams working in the field of adaptation acoustic models applied exactly the same database or a database that was recorded from one of the same speech resources. For example, [7] and [6] already did some initial phoneme mapping and phoneme combination work on English and Mandarin. In [7], HUB-4NE was the Mandarin training speech database, and in [6], part of the Chinese corpora was recorded from VOA news. The Mandarin language model is a bi-gram model that is derived from the HUB-4NE training set transcriptions. The English acoustic models used as seeds in cross-language adaptation are context-independent phone models trained on the TIMIT Acoustic-Phonetic corpus [9].

## 3. MLLR Adaptation Algorithm

MLLR computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data. More specifically MLLR is a model adaptation technique that estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. The effect of these transformations is to shift the component means and alter the variances in the initial system

so that each state in the HMM system is more likely to generate the adaptation data. We utilized HTK to simulate using MLLR to adapt to a new language. Equation (1) gives the transformation matrix used to give a new estimate of the adapted mean.

$$\mu = W\xi \quad (1)$$

where  $W$  is the  $n * (n + 1)$  transformation matrix (where  $n$  is the dimensionality of the data) and  $\xi$  is the extended mean vector,

$$\xi = [w\mu_1\mu_2\mu_3\dots\mu_n]^T \quad (2)$$

where  $w$  represents a bias offset whose value is fixed at 1. Hence  $W$  can be decomposed into

$$W = [bA] \quad (3)$$

where  $A$  represents an  $n * n$  transformation matrix and  $b$  represents a bias vector. The transformation matrix  $W$  is obtained by solving a maximization problem using the Expectation-Maximization (EM) technique, which is also used to compute the deviation transformation matrix using EM results in the maximization of a standard auxiliary function.

#### 4. Experimental Results

We used HTK [10] to simulate our ASR system. The acoustic model of the Mandarin syllable recognition system was mono-phone modeling; each mono-phone model was HMM (Hidden Markov Model) with 5 states, including start and end states; each state had 15 mixture Gaussian distributions. The dimension of the acoustic model was 12. The characteristics of the English acoustic model were the same as the Mandarin one's. The Mandarin language model was a bi-gram model that was build using HTK HBuild tool and transcriptions of the HUB-4NE training corpora.

The test dataset contained another 100 natural Mandarin utterances, 856 Mandarin syllables, which were spoken by 4 men and 4 women. Speech signals were monaural sound and sampled at a rate of 16 kHz. There are seven different measure points as shown in Table 1; MP1 (Measure point 1) to MP7 (Measure Point 7) contain the Mandarin utterances from 40 to 5861, which applied the seven different sizes of training sets to check the performance of cross-language adaptation and native speech training. By this way, we could find the point at which the native training begins to outperform the cross-language adaptation. In Table 1, "Utt#" stands for "Utterance number"; "Syl#" stands for "Syllable number"; and "Phn#" stands for "Phoneme number".

	MP1	MP2	MP3	MP4	MP5	MP6	MP7
Utt#	40	97	168	249	466	2937	5861
Syl#	556	1207	2188	3161	5926	34789	67677
Phn#	1014	2220	4005	5769	10845	63348	123392

Table 1: Seven Measure Points

Table 2 showed the syllable accuracy applied to different approaches. In Table 2, NST stands for Native Speech Training. It was from the Measure Point MP4 that the native speech training started to outperform the cross-language adaptation approach. MP4 contains 249 utterances, 3161 syllables. We calculate the length of the training set and the number of total syllables in training corpora, and the speech rate was estimated to be 226 msec. for each syllable; therefore, the length of the MP4 is around 12 minutes. Based on

experimental results of Table 2, it was shown that if very limited training utterances are available, let's say, shorter than 12 minutes, 249 utterances, MLLR cross-language adaptation is better than direct native speech training; otherwise, utilize native speech data to train acoustic models directly and skip the adaptation step. It is a big work to collect a large amount of training corpora with accurate transcriptions, but it is relatively simple to record a small training set of around 15 minutes and label it. From this point of view, cross-language adaptation does not look that much useful. The possible best way to create a recognizer for a new language with limited training data is to train initial acoustic models only using native speech, and then apply MLLR and MAP for online unsupervised adaptation.

Figure 1 showed the recognition results of comparison between direct native speech training and cross-language adaptation. When only trained by 40 Mandarin utterances, the recognition accuracy of native speech training is only 14.5%. With the Mandarin training set increased in size, the recognition accuracy goes up to 82.7%. Next, let's check the performance of Mandarin speech adaptation using MLLR. First of all, when the adapted Mandarin speech set contained 40 utterances, the Mandarin recognition accuracy was 47.8%, which is better than direct native speech training, but at the point of around 200 utterances, the recognition result of direct native language training began to be better than the recognition result of adaptation. Therefore, if we could collect a native speech corpus of around 200 utterances, 5000 phonemes, we would use it to train the acoustic model directly without using an adaptation approach.

	MP1 (%)	MP2 (%)	MP3 (%)	MP4 (%)	MP5 (%)	MP6 (%)	MP7 (%)
MLLR	47.8	52.8	55.6	56.1	55.5	55.4	55.3
NST	14.5	34.4	51.9	62.9	69.4	78.1	82.7

Table 2: Syllable accuracy of different approaches to build Mandarin acoustic models

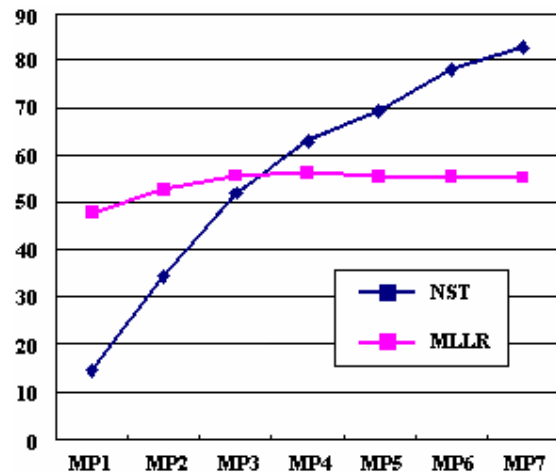


Figure 1: Percentage of syllable recognition accuracy for adaptation compared with native speech training

#### 4.1 Error-bars Using Different Training Sets

		MP1	MP2	MP3	MP4	MP5
Training Set 1	Utt#	40	100	168	250	500
	Syl#	453	1111	1880	2822	5682
	Phn#	829	2045	3444	5171	10395
Training Set 2	Utt#	40	100	168	250	500
	Syl#	421	1127	1894	2943	6511
	Phn#	773	2073	3492	5381	11888
Training Set 3	Utt#	40	100	168	250	500
	Syl#	553	1355	2222	3458	6309
	Phn#	1014	2486	4071	6308	11398
Training Set 4	Utt#	40	100	168	250	500
	Syl#	588	1277	1976	2685	5544
	Phn#	1061	2310	3577	4899	10061
Training Set 5	Utt#	40	100	168	250	500
	Syl#	393	995	1877	2709	5220
	Phn#	721	1822	3394	4936	9499
Average	Utt#	40	100	168	250	500
	Syl#	482	1173	1970	2923	5853
	Phn#	880	2147	3596	5339	10648

Table 3: Training sets for error-bars experiments

		MP1 Acc. (%)	MP2 Acc. (%)	MP3 Acc. (%)	MP4 Acc. (%)	MP5 Acc. (%)
Training Set 1	MLLR	40.8	49.6	51.9	51.2	52.2
	NST	15.6	40	51.7	54.2	64.9
Training Set 2	MLLR	45.3	54.5	54.5	53.1	55.4
	NST	17.1	41.8	48	52.4	65.9
Training Set 3	MLLR	37.3	52	55.5	53.5	55.8
	NST	12.4	45.9	55	63.5	68
Training Set 4	MLLR	41.4	45.1	44.8	57.3	54.5
	NST	10.8	32.9	37.3	59.1	66.7
Training Set 5	MLLR	39.3	45	47.4	52.8	54.3
	NST	16.9	35.8	39.5	51.3	61.7
Mean	MLLR	42.6	49.2	50.8	53.6	54.4
	NST	14.6	39.3	46.3	56.1	65.4

Table 4: Mandarin recognition accuracy using five different training sets

The above experiment discovered that the point where the native training outperforms the adaptation training is around 200 utterances, 2500 syllables. To get the accuracy variation, another five experiments were repeated using five separated training sets; these training sets were separated on both speech and speakers. In experiments of training set error-bars, the test set was fixed in Table 3; MP1 to MP5 are five measure points around the cross point that was gotten from Figure 1. Because the cross point was between 168 utterances and 250 utterances in the first experiment, measure points were distributed around this cross point in experiments of the train set error-bars. Table 3 showed the distribution of MP1 to MP5. For example, in the err-bars experiment 1, MP1 has 500

utterances, 5682 syllables, 10395 phonemes, which means these 500 utterances will be used to adapt the English acoustic model to attain a Mandarin adaptation model.

Table 4 is the syllable recognition accuracy report using different training sets. In Table 4, Acc. stands for "Accuracy", so "MP1 Acc." means the accuracy on measure point 1. Figure 2 showed ranges of accuracy variation for English-Mandarin cross-language adaptation. From lines of mean MLLR and mean NST, we could find that with the increasing of the adaptation data, the MLLR adaptation did boost the recognition accuracy, but such improvement converges around a 55% accuracy. Cross-language adaptation got a 42.6% recognition accuracy even when there was only 40 Mandarin utterances available. However, at the point of 4400 phonemes, native training began to outperform the MLLR adaptation. In Mandarin, 4400 phonemes are about 2200 syllables, which is about 8 minutes of speech at the rate of 226 msec. per syllable. So, on the condition of little Mandarin training data, if the training set is less than 8 minutes, the MLLR cross-language adaptation is useful; otherwise, we could use Mandarin native speech directly to train a Mandarin acoustic model.

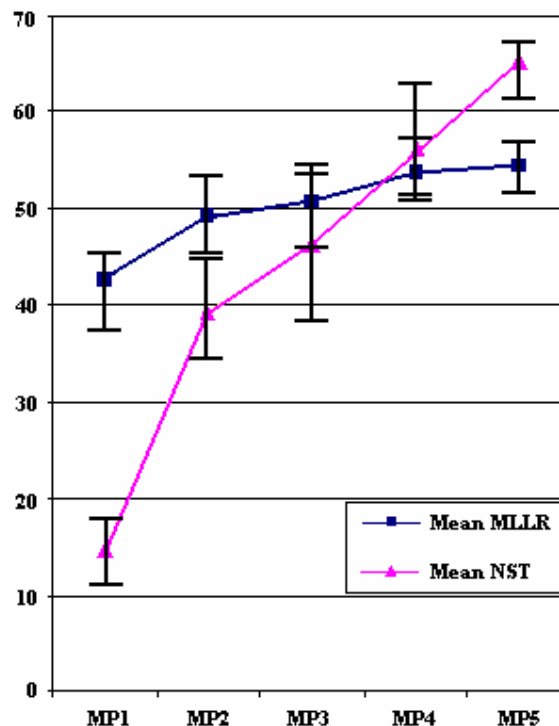


Figure 2: Error-bars using different training sets

#### 4.2 Error-bars Using Different Testing Sets

	Testing Set 1	Testing Set 2	Testing Set 3	Testing Set 4	Testing Set 5
Utt#	100	100	100	100	100
Syl#	1022	1069	1105	1079	1071
Phn#	1859	1949	2010	1968	1975

Table 5: Testing sets for syllable accuracy variation experiments

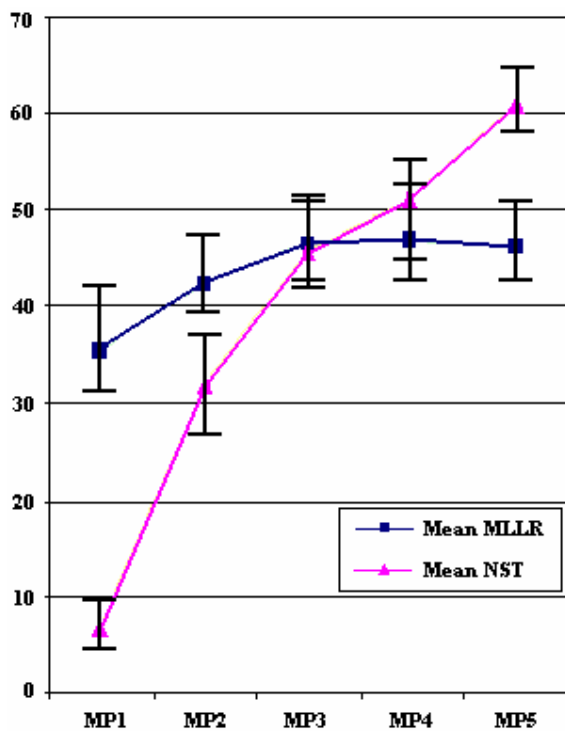


Figure 3: Error-bars using different testing sets

To get performance ranges on different testing sets, another 5 experiments were repeated using five different testing sets. Table 5 showed the size of five testing sets.

Figure 3 showed syllable accuracy variation using different testing set. When the training set was bigger than about 4000 phonemes, the native training outperform the cross-language adaptation. In Mandarin, 4000 phonemes are around 2000 syllables considering that the majority of Mandarin syllables are bi-phone syllables. The length of 2000 syllables is about 7.5 minutes at the speech rate of 226 msec. per syllable. We derived the consistence result from both training set and testing sets syllable accuracy variation.

		MP1	MP2	MP3	MP4	MP5
		Acc.	Acc.	Acc.	Acc.	Acc.
		(%)	(%)	(%)	(%)	(%)
Testing Set 1	MLLR	33.6	39.6	44.2	45.1	42.9
	NST	9.3	33.9	47.9	49.3	59.1
Testing Set 2	MLLR	37.5	42.1	44.6	46.4	45.5
	NST	5.4	31.6	43.1	51.6	59.7
Testing Set 3	MLLR	42.9	47.9	50.8	52.1	51.7
	NST	4.7	28.9	42.8	52.2	61
Testing Set 4	MLLR	32.7	42.5	50.1	48.1	49.5
	NST	9.6	37.6	51.7	55.6	65.7
Testing Set 5	MLLR	30.5	38.9	42.7	42	41.8
	NST	4.3	26.4	43	47.4	58.3
Mean	MLLR	35.4	42.2	46.5	46.7	46.3
	NST	6.7	31.7	45.7	51.2	60.8

Table 6: Mandarin syllable recognition accuracy using five different training sets

Table 6 is the syllable accuracy report using different testing sets. In Table 6, Acc. stands for "Accuracy". For example, in the first row, MP1 means "Measure Point 1", which use training set MP1 of table 1 to adapt the English acoustic model or directly train a Mandarin acoustic model. These models will be tested using different testing sets that were showed in Table 5, and syllable recognition accuracies were listed in Table 6, from column 3 to column 7.

## 5. Conclusion

This study based on limited training corpora indicates that when only a small amount of training set is available, cross-language adaptation is not a more effective technique than direct native speech training for HMM development. When the training data is very tiny, for example, less than 5 minutes, the cross-language method produced the better models, while if the training set is more than about 8 minutes, the native speech training outperform the cross-language adaptation. Further research will explore the extensibility of this result and attempt to confirm that it is not specific to this English-Mandarin pair or this type of application.

## 6. References

- [1] L.Bahl, J.Jelinek, J.Raviv, and F.Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate", *IEEE Transactions on Information Theory*, Vol. IT-20, pp. 284-287, March 1974.
- [2] Osterholtz, L., Augustine, C., McNair, A., Rogina, I., Saito, H., Sloboda, T., Tebelskis, J., Waibel, A., and Woszczyna, M., "Testing Generality in JANUS: A Multi-lingual Speech Translation System," *Proceedings of ICASSP '92*, pp. 209-212, 1992.
- [3] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, pp. 171-185, 1995.
- [4] Zhirong Wang, Schultz, T., Waibel, A., "Comparison of Acoustic Model Adaptation Techniques on Non-native Speech," *Proceedings of ICASSP*, Vol. 1, pp.540-543, 2003.
- [5] Kohler, J., "Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks," *Proceedings of ICASSP '98*, pp. 417-420, 1998.
- [6] Tanja Schultz and Alex Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, Vol. 35, pp. 31-51, 2001.
- [7] Pascale Fung, MA Chi Yuen, and LIU Wai Kat, "MAP-based Cross-Language Adaptation Augmented by Linguistic Knowledge: from English to Chinese," *Proceedings of EUROSPEECH '99*, pp. 871-874, 1999.
- [8] Linguistic Data Consortium, "1997 Mandarin Broadcast News Speech (HUB4-NE)," ISBN: 1-58563-125-6, 1998.
- [9] Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM." Linguistic Data Consortium, 1986.
- [10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book," Entropic, Inc., 1999.