A STUDY ON MULTILINGUAL ACOUSTIC MODELING FOR LARGE VOCABULARY ASR

Hui Lin[1], Li Deng[2], Dong Yu[2], Yi-fan Gong[2], Alex Acero[2], Chin-Hui Lee[3]

Microsoft Corporation[2], University of Washington[1], Georgia Institute of Technology[3]

ABSTRACT

We study key issues related to multilingual acoustic modeling for automatic speech recognition (ASR) through a series of large-scale ASR experiments. Our study explores shared structures embedded in a large collection of speech data spanning over a number of spoken languages in order to establish a common set of universal phone models that can be used for large vocabulary ASR of all the languages seen or unseen during training. Language-universal and language-adaptive models are compared with language-specific models, and the comparison results show that in many cases it is possible to build general-purpose language-universal and language-adaptive acoustic models that outperform language-specific ones if the set of shared units, the structure of shared states, and the shared acoustic-phonetic properties among different languages can be properly utilized. Specifically, our results demonstrate that when the context coverage is poor in language-specific training, we can use one tenth of the adaptation data to achieve equivalent performance in cross-lingual speech recognition.

Index Terms— Multilingualism, acoustic modeling, language adaptation, universal phone models.

1. INTRODUCTION

Building language-specific acoustic models for automatic speech recognition (ASR) of a particular language is a reasonably mature technology when a large amount of speech data can be collected and transcribed to train the acoustic models. However when multilingual ASR for many languages is desired, data collection and labeling may become too costly that alternative solutions are often desired. One potential solution is to explore shared acoustic phonetic structures among different languages to build a large set of acoustic models (e.g., [1][2][3][5][7][8][9]) that characterize all the phone units needed in order to cover all the spoken languages being considered. This is sometimes called multilingual ASR or cross-lingual ASR when no language-specific data is available to build the acoustic models for the target language.

One central issue in multilingual speech recognition is the tradeoff between two opposing factors. On the one hand, use of multiple source languages' acoustic data creates the opportunity of greater context coverage (as well as more environmental recording conditions). On the other hand, the differences between the source and target languages create potential impurity in the training data, giving the possibility of polluting the target language's acoustic model. In addition, different languages may cause mixed acoustic dynamics and context mismatch, hurting the context-dependent models trained using diverse speech data from many language sources.

Thus, one key challenge in multilingual speech recognition is to maximize the benefit of boosting the acoustic data from multiple source languages while minimizing the negative effects of data impurity arising from language "mismatch". Many design issues arise in addressing this challenge, including the choice of language-"universal" speech units, the total size of such units, definition of context-dependent units and their size, decision-tree building strategy, optimal weighting of the individual source languages' data in training, model adaptation strategy, feature normalization strategy, etc.

Our current research explores all the above issues in a comprehensive manner so as to develop a most appropriate strategy in building multilingual ASR systems. The goal of this paper is to present our preliminary findings in this exploration. Specifically, we found that when the context coverage is weak in the language-specific (LS) acoustic model (even with over 20 hrs of LS training data), using additional new languages' training data improves the LS recognizer's performance. Further, after adapting the above language-universal (LU) acoustic model using the limited LS acoustic data, further performance improvement is observed.

2. UNIVERSAL PHONE SET

The main goal of multilingual acoustic modeling is to share the acoustic data across multiple languages to cover as much as possible the contextual variation in all languages being considered. One way to achieve such data sharing is to define a common phonetic alphabet across all languages. This common phone set can be either derived in a data-driven way, or obtained from phonetic inventories such as Worldbet [4], or International Phonetic Alphabet (IPA) [6]. In this study we use the universal phone set (UPS), which is a machine-readable phone set based on the IPA.

In general, there is a one-to-one mapping between UPS and IPA symbols, while in a few other cases UPS is a superset of IPA. For example, UPS includes some unique phone labels for commonly used sounds such as diphthongs, and nasalized vowels, while IPA treats them as compounds. Generally, UPS covers sounds in various genres, including consonants, vowels, suprasegmentals, diacritics, and tones. Table 1 illustrates the number of different types of UPS units for each of the eight languages (French, Portuguese, German, Italian, Dutch, Spanish, Swedish, and English) we used in this study.

There are a total of 80 distinct UPS symbols, including 30 vowels, 47 consonants, 1 suprasegmental, and 2 diacritics for the eight languages we used in this study. The language sharing factor is 3.85, meaning that on average one symbol is shared by 3 to 4 languages. Adding four other symbols used for silence and noise, we finally have 84 units for the eight languages to train language-universal (LU) acoustic models. The UPS-based pronunciation dictionariy for each language is converted from its original lexicon using the mapping of old phone set to UPS. The mapping is defined by linguistic experts, and is usually one-to-many.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

Table 1. *Number of vowel (vow), consonant(con), suprasegmentals (sup), and diacritics (dia) units for the 8 languages used in this study*

	FRA	PTB	DEU	ITA	NLD	ESP	SVE	ENG
vow	19	13	19	11	15	6	19	18
con.	23	22	25	24	22	21	23	24
sup	-	-	1	1	-	-	1	-
Dia	1	1	-	-	-	-	-	-
Total	42	36	45	36	37	27	43	42

3. EXPLORING ISSUES IN MULTILINGUAL ASR

The LU acoustic training data used in this study are from telephone recordings in all eight languages. They are selected from several corpora. Most of them contain phonetic-rich sentences and application oriented utterances such as application words, number strings, dates embedded application word, spelled names/words, natural number, money amounts, etc. There are around 20-30 hours' data from each language, resulting in a total of 180 hours data for language-universal training. We used these data sets to explore several key issues in multilingual ASR, as reported in this section.

Table 2. *Description of training data*

Language	Corpus	# Speakers	Hours
FRA	ELRA-S0016	1000	30.8
PTB	MOBIL4PB	922	23.4
DEU	ELRA-S0162	3474	16.7
ITA	ELRA-S0052	989	23.5
NLD	SSP35097	703	18.6
ESP	ELRA-S0065	992	33.9
SVE	PHIL196	493	23.2
ENG	ELRA-S0074	1000	17.5

Training of both language-specific and language-universal models follow the same procedure described below. 13 MFCCs were extracted along with their first and second time derivatives, giving a feature vector of 39 dimensions. Cepstral mean normalization was used for feature normalization. All the models mentioned in this paper are cross-word triphone models. Phonetic decision tree tying was utilized to cluster triphones. A set of linguistically motivated questions were derived from the phonetic features defined in the UPS set. The number of tied states, namely *senones*, can be specified at the decision tree building stage to control the size of the model. The top-down tree building procedure is repeated until the increase in the log-likelihood falls below a preset threshold. The number of mixtures per senone is increased to 4 along with several EM iterations. This leads to an initialized cross-word triphone model. The transcriptions are then re-labeled using the initialized cross-word triphone models, which were used to run the training procedure once again – to reduce number of mixture components to 1, untie states, re-cluster states and increase number of mixture components. The final cross-word triphone is modeled with 12 Gaussian components per senone.

As for testing, we are interested in telephony ASR under various environments, including home, office, and public places. We chose Italian as our target language, which is *seen* during the language-universal training. Other languages can also be used to conduct the same series of tests to be discussed in detail in the following. Several test sets were used as shown in Table 3. In all of our experiments, The Microsoft speech recognition engine was used for acoustic modeling and the grammar-based decoding.

Table 3. *Test set descriptions including number of utterances (#utt) and number of speakers (#sp.)*

ID	Corpus	#utt	# sp.	Environments
Test I	ELRA-S0052	2140	99	Office/home
Test II	ELRA-S0116	2199	400	Office/home/street/public place/vehicle
Test III	PHIL18	4827	197	Quite environment
Test IV	PHIL42	3916	129	Office/home/Quite environment

3.1. Language-Specific & Language-Universal Training

In our language-specific (LS) exploration, we used only the Italian data from corpus ELRA-S0052 as the training set, which contains over 20 hours of speech. In contrast, our language-universal (LU) training made use of all training data from eight languages, including the Italian data, for building the multilingual acoustic model. For the LS model, we had 3000 senones based on the amount of data for LS training, while for the LU model, we had 5000 senones since more training data were used in the training. For fair comparisons, we also increased the number of senones for the LS model, which was stopped at around 4600 when the problem of data insufficiency was detected.

As we expected, LU modeling is supposed to utilize more data and more context information borrowed from other languages. To examine this, we analyzed the context coverage defined as the percentage of the number of test set triphones that appear in the training set more than *N* times over the total number of triphones in the test set. As shown in Table 4, for *N*=1, LU training usually has larger value than the LS training has, indicating that some triphones seen in the LU training is unseen in the LS training. Also, more triphone units are seen over 100 times in LU training than in LS training.

Table 4 *Context coverage on four test sets from the LU and LS training, respectively*

	Test I		Test II		Test III		Test IV	
	LS	LU	LS	LU	LS	LU	LS	LU
N=1	99%	99%	95%	99%	96%	100%	91%	97%
N=100	70%	82%	59%	76%	85%	93%	73%	83%

However, the resource-rich advantage of LU acoustic modeling does not seem to directly reflect the advantage on recognition performance in terms of word error rate (WER). Row two and row four in Table 5 show the WER of the LS and LU models over the four test sets. LU model only works better than LS model on test set III. Note that test set III is the only test set that LU training covers *all* the triphones in the test and still have over 90% context coverage for *N*=100. Error analysis reveals that the improvement might come from the fact that LU training has better context coverage. One example is triphone "*sil-e+m*". It is covered by LU training but unseen in LS training. In the test, LS model made 6 deletion errors of word "*emergenza*", while there was only one deletion in the LU case.

On other test sets, LS model outperforms LU. The observation that LU training usually decreases the model precision for recognition of seen languages is consistent with other studies

[2][5][8][9], which we believe are caused by the same reason as our results above for the test sets other than set III.

Although LU modeling enjoys richer data resources, it also has a larger set of units to model and these units shared across different languages may not be "pure''. In our experiments, LU training may not be considered to have contained more data because the number of cross-word triphones for the LU model is about 9 times larger than the LS model. Further, senones in the LU model are supposed to cover all languages and thus are less "dedicated'' to a specific language compared with senones in the LS model. This situation is similar to that of speaker-independent models vs. speaker-dependent models. Finally, identical UPS symbols across languages may not correspond to acoustic similarity. In the following sections, several methods are investigated to address these issues in order to improve LU acoustic modeling.

3.2. Improving LU acoustic modeling

3.2.1. UPS Unit Reduction

The size of the UPS derived from phonetic knowledge related to IPA definition is usually quite large. There are over 200 units in order to cover all world languages. In our case, we have over 80 monophones for the eight languages. One issue we intend to address in this study is whether we need so many "monophone" units? For some units, although they might have subtle differences from the phonetic consideration, it is difficult to distinguish them acoustically. Moreover, for cross-word triphones, a compact speech unit representation is usually favored in acoustic modeling since the explosion of the number of triphone units can be avoided.

In this study, we propose an acoustic-phonetic clustering algorithm to reduce the number of UPS units derived from IPA. Two conditions should be satisfied for merging two monophones. First, the two monophones should be acoustically similar. We used Kullback-Leibler distances to measure the acoustic similarity between two phones, which were approximated from the HMMs we have in the LU model. Second, two units can be merged if representing them with the same symbol does not increase confusability in the pronunciation dictionaries of all eight languages. The confusability here is defined as the number of homo-lexicons in the dictionary. After unit merging, the number of units in our universal phone set was reduced from 84 to 62, which are then used to train a smaller LU model. The WER results are shown in Table 5 (rows with LU model and monophone number 62). Although the improvement is not significant here in the exploratory experiments, after language adaption the advantage of the new parsimonious representation becomes more obvious, as we will see in the next section.

3.2.2. Language Adaptation (LA)

Adapting the multilingual model to a new target language was shown to be very useful in some earlier studies [1][3]. In this study, we adapt the LU model to language-adapted (LA) model by applying both maximum a posteriori (MAP) and maximum log likelihood linear regression (MLLR). Language-specific data were used for adaptation. Although these data have already been seen during LU training, their functionality during adaptation is to make the LU model better fit the characteristics of the target language.

The WER results are shown in Table 5 (see the rows labeled with LA-MLLR or LA-MAP). It can be seen that LA consistently improves over LU models for all 4 test sets. Note the amount of adaptation data is reasonably large, and MAP was shown to be more effective than MLLR. Note also that the performance gap between the LU and LS models becomes smaller after adaptation. Comparing WER results of LA-MAP to LS, the LS model no longer has significant advantage for test sets I and II. Meanwhile, LA-MAP outperforms LS on test sets III and IV.

Table 5. *Word Error Rate (WER) results*

Models	#.Mono	# Senone	Test I	Test II	Test III	Test IV
LS	40	3000	3.62	5.57	6.34	17.90
LS	40	4600	3.82	5.68	7.67	19.61
LU	84	5000	4.93	6.75	5.84	19.66
LU	62	5000	5.05	6.67	5.55	19.08
LA-MLLR	84	5000	4.74	6.49	5.45	18.59
LA-MLLR	62	5000	4.81	6.48	5.51	18.34
LA-MAP	84	5000	4.30	6.24	5.20	17.11
LA-MAP	62	5000	4.17	5.89	**4.66**	16.98
LW-5	84	5000	4.23	5.92	5.22	17.03
LW-10	84	5000	4.08	5.60	5.06	**16.97**

3.2.3. Language Data Weighting (LW)

A third way to improve multilingual acoustic modeling is to explore weighting the language data in multi-lingual training according to their similarity to the target language. In this study, we performed the weighting after the LU training was finished. Specifically, four more EM iterations were performed on the LU model while weighting the data from different languages according to the following Gaussian mean update:

$$\boldsymbol{\mu}_{im} = \frac{\sum_{l}\sum_{\tau \in D(l)} w_l\, \gamma_{im}(\tau)\, \boldsymbol{o}(\tau)}{\sum_{l}\sum_{\tau \in D(l)} w_l\, \gamma_{im}(\tau)}$$

where w_l is the weight for language *l*, *D(l)* represents all the observation vectors from language l, $\boldsymbol{o}(\tau)$ is the τ-th vector and $\gamma_{im}(\tau)$, the probability that $\boldsymbol{o}(\tau)$ belongs to the *m*-th Gaussian component in the *i*-th senone which was computed in the initial LU model training. The corresponding variances and mixture weights were updated in a similar way.

The WER results are shown in Table 5. Here, the target language, Italian, was seen during LU training. LW-5 and LW-10 in Table 5 correspond to setting the language weight for Italian to 10 and 5, respectively, while the weight for all other languages remained to be unity. It is clear that increasing the weights for Italian improves the performance. On test sets I and II, no significant differences between the performances of LW and LS were seen. Indeed, when the weight for Italian is large enough, the senones in LU corresponding to triphones appear in LS training tend to be very similar to those in LS models. On the other hand, senones corresponding to other triphones unseen in LS training remains to be complementary. This is best illustrated by the results on test set

III (as well as IV), where LW outperforms LS significantly.

4. EXPERIMENTS ON CROSS-LINGUAL MODEL ADAPTATION

With three ways of improving the multilingual acoustic model after exploring several key issues as described in the preceding section, we report our experimental results on cross-lingual acoustic model adaptation. Building a system from scratch for a new target language can be too time consuming and expensive, since the process involves a huge amount of work in speech data collection (as well as lexicon building, etc.). If we can adapt the existing multilingual acoustic models to the target language, the time-to-market cycle for the new language can be significantly reduced.

In our experiments, we examine the case when a new target language is *unseen* during LU training. Canadian French was used in the experiments, with the data sets described in Table 6. The FRC-TRAIN data contain 10.5 hours audio, and were used as the training data for the LS model, and as the adaptation data for the LA models. Two sets of test data were used, with well matched and poorly matched context coverage, respectively.

Table 6. *Data description*

ID	Corpus	#utt	# Sp.	Environments
FRC-Train	PHIL101	44357	n/a	Quite environment
FRC-Test I	PHIL101	4921	4443	Quite environment
FRC-Test II	PHIL23	1293	1293	Quite environment

In our experiments, we exploited UPS size reduction and language data weighting, where Continental French received higher weights during training since we know that the target Canadian French is more similar to Continental French than other languages in the training set. For adaptation, we used a varying size of the adaptation data in order to examine the relationship between the amount of adaptation data and the adaptation performance. The performance results are shown in Figures 1 and 2.

In test set I, LS training has very good context coverage since the training and test conditions are closely matched. As seen in Figure 1, LS performs best in this case. In test II, however, LS training has a poorer context converge (67% for *N*=1) compared with LU training (82% for *N*=1). We observe from Figure 2 that with the use of fewer than one-hour data for adaptation, LA models already achieved a similar performance to that of LS training with 10-hour data. With more than two hours of data for adaptation, significant improvement was observed. In both test sets, language weighting is shown to be beneficial. Also, reducing the UPS unit size is shown to be successful in adaptation.

5. CONCLUSTIONS

In this study, we explored issues of multilingual ASR and found that when the context coverage is weak in the LS acoustic model, LU model can outperform language-specific model. LA model further improves language-universal model, and provides a good alternative way to deal with resource-limited languages as shown in Fig. 2. We conclude in this study that it is possible to build general-purpose LU and LA acoustic models that outperform LS ones when we carefully design and properly utilize the set of shared units, the structure of shared states, and the shared acoustic-phonetic properties among different languages.

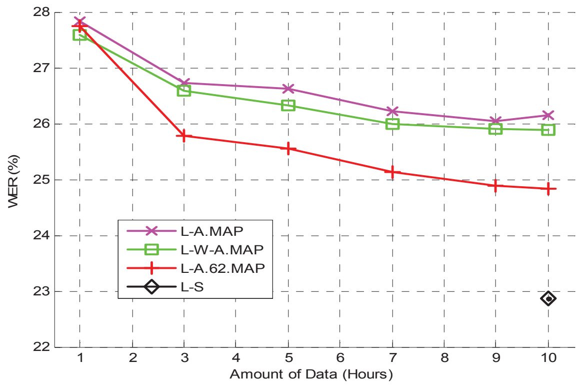

Figure 1. WER curves on FRC-Test I. L-A.MAP represents the LA training using MAP adaptation. L-W-A.MAP means MAP adapaton was performed on the model after language weighting,, and L-A-62.MAP is the LA model trained on the reduced phone set.

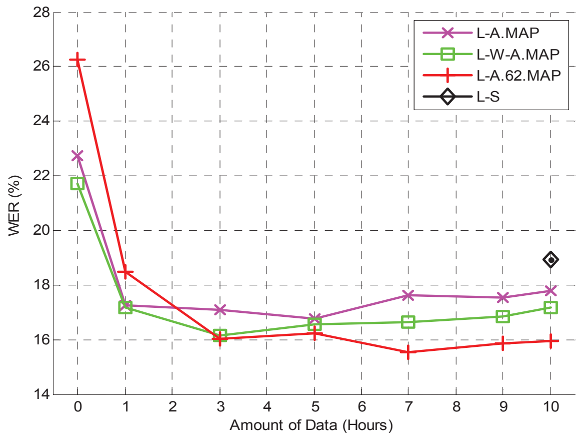

Figure 2. WER curves on FRC-Test II

6. REFERENCES

[1]. W. Byrne, et al., "Towards language independent acoustic modeling", *Proc. ICASSP*, 2000

[2]. P. Cohen, et al., "Towards a universal speech recognizer for multiple languages", in *Proc. ASRU 1997.*

[3] L. Deng, "Integrated-multilingual speech recognition using universal phonological features in a functional speech production model," *Proc. ICASSP,* Munich, Germany, 1997, Vol. 2, pp. 1007-1010.

[4] L. Hieronymus. "ASCII Phonetic Symbols for the World's Languages: Worldbet," *Journal of the International Phonetic Association* Vol. 23.

[5] E. Garcia, E. Mengusoglu and E. Janke, "Multilingual acoustic models for speech recognition in low-resource devices", *Proc. ICASSP 2007*.

[6] International Phonetic Association, "Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet," Cambridge University Press, 1999, pp. 1-204.

[7]. T. Niesler, "Language-dependent state clustering for multilingual acoustic modeling", *Speech Communication*, vol. 49, 2007

[8] T. Schultz, and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition", *Speech Communication*, vol. 35, 2001.

[9]. T. Schultz and A. Waibel, "Language Independent and language adaptive large vocabulary speech recognition", in *Proc. ICSLP 1998*