ACOUSTIC DATA SHARING FOR AFGHAN AND PERSIAN LANGUAGES

Arindam Mandal, Dimitra Vergyri, Murat Akbacak, Colleen Richey, Andreas Kathol

Speech Technology and Research Laboratory
SRI International, Menlo Park CA USA
{arindam,dverg,murat,colleen,kathol}@speech.sri.com

ABSTRACT

In this work, we compare several known approaches for multilingual acoustic modeling for three languages, Dari, Farsi and Pashto, which are of recent geo-political interest. We demonstrate that we can train a single multilingual acoustic model for these languages and achieve recognition accuracy close to that of monolingual (or language-dependent) models. When only a small amount of training data is available for each of these languages, the multilingual model may even outperform the monolingual ones. We also explore adapting the multilingual model to target language data, which are able to achieve improved automatic speech recognition (ASR) performance compared to the monolingual models for both large and small amounts of training data by 3% relative word error rate (WER).

Index Terms— multilingual acoustic modeling, language-independent acoustic modeling, languages of Afghanistan

1. INTRODUCTION

When linguistic knowledge (e.g., phone inventory, pronunciation dictionary) and transcribed speech data are available, continuous speech recognition technology is generalizable to new languages since similar statistical modeling assumptions hold. However when ASR systems are desired for multiple languages, large scale data collection may not be possible. Moreover it is often desirable for the recognition systems to be able to recognize multiple languages of interest, and be able to switch among them on the fly, depending on the user, in applications that run on portable devices such as smartphones. The increasing demand for speech technologies in different languages with low resources, has caused rapid deployment of these technologies to become an important research area. There have been studies on leveraging resources between languages for different speech technologies. One application is continuous speech recognition where the focus has been on using acoustic data from resource-rich languages and a target language (which is usually considered as a limited-resource language) to build an ASR system in the target language.

Large-vocabulary multilingual speech recognition has been an area of intensive work at many research centers [1, 2, 3] for resource-rich languages such as English, German, etc. for which detailed linguistic knowledge, text resources for language modeling and lexicon construction, as well as sufficient data to train acoustic models are all available. These studies employ an initial bootstrapping step to align acoustic data with the text provided in each language using the an initial source language. Acoustic models (AMs) need either knowledge-based [1, 2] or data-driven-based [4, 5] phone mapping. Depending on the amount of available aligned acoustic training data for a target language, source language acoustic models are adapted towards the mapped target language AMs, or target language AMs are trained directly.

This work focuses on Afghan and Iranian languages which have been of geo-political interest in the last few years. Our goal is to leverage acoustic data from three languages - Pashto, Dari and Farsi - and train a single multilingual AM. As with acoustic modeling for multilingual speech recognition, we make use of well-established methods for (semi-) continuous HMM training. Previous approaches used in multilingual settings include the use of multilingual seed HMMs [6], the use of language questions in phonetic tree growing [1], and polyphone decision tree specialization for better coverage of contexts from an unseen target language combined with the use of a Bayesian Information Criterion (BIC) to determine an appropriate model size [7].

In this study we evaluate the effectiveness of these well-known multilingual acoustic data sharing approaches for the three languages of interest using moderate to large amounts of acoustic training data. In addition, we investigate the effect of data sparsity on these modeling approaches by limiting the amount of acoustic training data. We also evaluate the effect that language variability has on the multilingual model by comparing against a model developed using only two of the three languages, Dari and Farsi, which are known to be more similar to each other than either is to Pashto. Lastly, we investigate the effect of giving more importance to the training data for a target language in training the multilingual acoustic model, using standard MAP adaptation techniques.

In this paper Sec. 2 presents linguistic information for the languages under consideration, Sec. 3 describes resources and ASR systems used and Sec. 4 provides results of ASR experiments.

2. LINGUISTIC BACKGROUND

Dari, Farsi and Pashto are dialects and languages within the Indo-Iranian language family. Pashto, a language spoken in Afghanistan and Pakistan, consists of two major dialects, southern and eastern. Farsi and Dari are dialects of Persian; Farsi is spoken in Iran and Dari is spoken in Afghanistan. The phone inventories of Pashto, Farsi and Dari are very similar. A basic difference is that Pashto has an additional series of retroflex consonants. In addition, [f] and [q], which are not native to Pashto, tend to be merged with [p] and [k] by Pashto speakers. A major difference between the southern and eastern dialects of Pashto is the pronunciation of the retroflex fricatives. Southern Pashto has maintained the retroflex, while eastern Pashto has merged them with the velar fricative and velar stop [x], [g]. The main difference between the phone inventories of Farsi and Dari is in the vowels. Dari maintains more vowel distinctions than Farsi. There are also a few consonant differences. Farsi has [v] where Dari has [w]. The Farsi dialect also no longer distinguishes between [q] and [g] . Figure 1 shows the set of phones for each language that we used in the acoustic models.

	Phone	P	D	F		Phone	P	D	F
Labial	p	✓	✓	✓	Retroflex	ʈ	✓		
	f	✓	✓	✓		ʂ	✓		
	b	✓	✓	✓		ɖ	✓		
	m	✓	✓	✓		ʐ	✓		
	w	✓	✓			ɳ	✓		
	v			✓		ɻ	✓		
Dental/ Alveolar	t	✓	✓	✓	Velar/ Uvular	k	✓	✓	✓
	s	✓	✓	✓		x	✓	✓	✓
	d	✓	✓	✓		g	✓	✓	✓
	z	✓	✓	✓		ɣ	✓	✓	✓
	ʦ	✓				q	✓	✓	
	ʣ	✓			Long Vowels	iː	✓	✓	✓
	n	✓	✓	✓		uː	✓	✓	✓
	l	✓	✓	✓		ɑː	✓	✓	✓
	r	✓	✓	✓		eː	✓	✓	
Palatal	ʃ	✓	✓	✓		oː	✓	✓	
	ʒ	✓	✓	✓	Short Vowels	i	✓	✓	
	ʧ	✓	✓	✓		u	✓		
	ʤ	✓	✓	✓		ɑ	✓	✓	✓
	j	✓	✓	✓		e		✓	✓
Glottal	ʔ		✓	✓		o		✓	✓
	h	✓	✓	✓		ə	✓		

Fig. 1. Phones in Pashto(P), Dari(D) and Farsi(F)

3. SYSTEMS & RESOURCES

3.1. CORPORA

The ASR training data for the three languages was drawn from the data collected and distributed under the DARPA CAST (Farsi) and TRANSTAC (Dari, Pashto, Farsi) programs. In addition to transcriptions, the data also contain pronunciation lexicons. The transcriptions were cleaned and partially normalized, and the pronunciation dictionary was adjusted accordingly. Table 1 shows the training data available for the three different languages in terms of hours of speech, transcription word count, number of distinct word types, and number of words in the pronunciation lexicon. The evaluation sets were provided under DARPA's TRANSTAC program. In Table 1 we also show the size of these test sets in terms of number of words and minutes and percentage of out-of-vocabulary (OOV) items (words not in our training data or lexicon).

Language	Dari	Farsi	Pashto
Training data			
Hours	126	75	123
Words	1M	750K	1.4M
Word Types	15K	31K	17K
Lexicon Vocab	12K	30K	17K
LM 2-grams	275K	300K	330K
Evaluation data			
Test Words	6269	4266	22617
Minutes	56	32	113
OOV (%)	2	2	5

Table 1. Description of data resources and evaluation sets

3.2. SPEECH RECOGNITION SYSTEM

The AMs developed in this work were used for a large-vocabulary, statistical-language model (LM) recognition system that was run on low resource devices (for example the Nexus One Android smartphone) as part of a real-time speech-to-speech translation system [8]. We used a common acoustic front-end that computes 13 MFCC[1] (including energy) and their first, second and third order derivatives. HLDA[2] is used to reduce the dimensionality of the feature vector to 36. Due to the memory constraints of the application, we used smaller AMs: for each language we trained cross-word triphone models, with decision-tree state clustering that resulted in 2000 fully-tied states and each state was modeled by a 16-component Gaussian mixture model. For the experiments presented in this work, all models were trained with maximum likelihood estimation. A detailed description of the ASR system is provided in [9]. The vocabulary and size of recognition LMs used for each language are shown in Table 1.

4. EXPERIMENTS

4.1. BASELINE ACOUSTIC MODELS

In Table 2 we present the performance of each of the monolingual (or language-dependent) baseline acoustic models. For each language we present: (1) the results with the acoustic

[1]Mel Frequency Cepstral Coefficients
[2]Heteroscedastic Linear Discriminant Analysis

Experiment	Model size	Dari	Farsi	Pashto
All data	2000x16	33.6	36.7	40.0
10 hours	2000x16	37.1	39.5	42.9

Table 2. WER (%) for language-dependent acoustic models

Experiment	Model size	Dari	Farsi	Pashto
Baseline: Lang dep	2000x16	33.6	36.7	40.0
Trilingual (D,F,P)	2000x16	36.8	44.6	45.6
	4000x16	34.7	42.3	44.5
	6000x16	34.9	42.5	44.0
+ Language flag	2000x16	35.2	41.7	43.0
	4000x16	33.3	40.1	41.2
	6000x16	**32.9**	**38.9**	**40.9**
Bilingual (D,F) + Language flag	2000x16	33.5	39.4	N/A

Table 3. WER (%) for multilingual acoustic models trained on all available data, (D is Dari, F is Farsi, P is Pashto)

model trained on all available data and (2) using only a random sample of 10 hours of acoustic training data for each language, thus simulating the scenario of more extreme acoustic data sparseness. The same 2-gram LM, trained on transcripts of all available acoustic data, was used for all ASR experiments.

4.2. MULTILINGUAL ACOUSTIC MODELS

As a first effort for multilingual acoustic models, we merged all the training data and lexicons, using the common phonetic inventory. We built a common decision tree and examined the performance of models with 16 Gaussian distributions per state and 2000, 4000 and 6000 fully-tied triphone states as shown in Table 3. The model size was increased to 6000 in order to compare more fairly with the original monolingual models of the same size.[3] We can see that the ASR performance of this multilingual model (referred to as Trilingual in Table 3) degrades significantly compared to the monolingual models.

Next, we added the language identity as a flag for each phone, and allowed the decision tree to ask questions about the value of that flag. This approach is largely based on the work described in [1]. When the decision tree can use questions about the phone language flag it allows the creation of language-specific triphone states based on the observed distribution similarity. This way, even though the lexicons use a common phone set, language specific models can be used if the observations for each phone in each language set are dissimilar enough. The lexicon was also augmented with a language flag, so that only language specific pronunciations are used for each word. This was not the case in the previous approach which, did not use the language flag and merged all pronunciations found across languages in the training lexicon under the same word entry. In order to verify that the final model is still sharing data across all languages (and that the language flag was not the primary question for splitting in the decision tree) we looked at the distribution of the number of triphone states from each of the three languages that were clustered at each genone[4] state. Figure 2 shows the histogram of the entropy of these distributions and it can be seen that most of the genones have an entropy close to that of uniform distribution[5]. It can thus be interpreted that the vast majority of the states in the 6000-state model still share data, almost evenly, from all three languages, and only very few are actually c

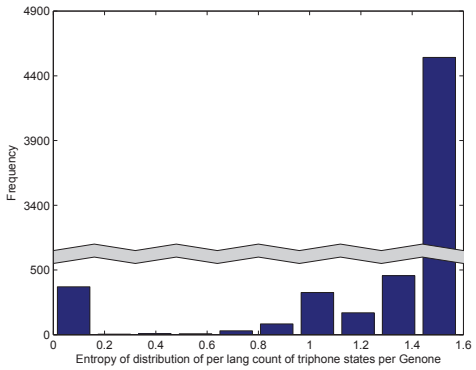

Fig. 2. Entropy of distribution that counts number of triphones states for each language in a genone state.

In Table 3 we can see that this approach of multilingual model training achieves improved ASR performance over the previous one. For the 6000x16-model size the results are close to the original monolingual baseline. Using this approach we can train a single larger acoustic model rather than three separate ones and use a recognition system that allows switching between languages. To examine the effect of the amount of data on the performance of the multilingual model (compared to the monolingual one), we also trained multilingual models using the 10 hours of randomly sampled data.

[3]Three separate 2000x16 Gaussian monolingual models were used for decoding each language, implying that effectively 6000x16 Gaussians were used to recognize the three languages.

Experiment	Model size	Dari	Farsi	Pashto
Baseline:Lang dep	2000x16	37.1	39.5	42.9
Trilingual (D,F,P) + Language flag	2000x16	**37.0**	41.7	44.9
	4000x16	**36.3**	**40.7**	44.4
	6000x16	37.0	41.0	44.9
Bilingual (D,F) + Language flag	2000x16	**36.4**	**40.2**	N/A

Table 4. WER (%) for multilingual acoustic models trained on 10 hours per language (D is Dari, F is Farsi, P is Pashto)

[4]Genones refer to the leaves of the decision tree that clusters triphone HMM states

[5]For a uniform distribution the entropy is $-log_2 X$, where X is the number of possible values of a discrete distribution

This is similar to the amount used for the multilingual acoustic modeling work described in [1]. The results using these models are shown in Table 4. We can see similar trends as with the larger amount of data. The multilingual, 4000-state AM in the case of Dari is better than the baseline monolingual model by 0.8% absolute.

4.3. Effect of language similarity for multilingual AMs

Farsi and Dari are more similar to each other than either is to Pashto. We wanted to evaluate the effect that using a language dissimilar to the rest has on the multilingual model training so we decided to remove Pashto from the pool of training data and train a model only from the two other languages. We see the results in the rows named "Bilingual" in Tables 3 & 4. The new Bilingual model trained on Farsi and Dari outperforms the monolingual model for Dari, and is very close to the Farsi one. We only trained this model with the limited training data (20 hours total) so increasing the model size to more than 2000 states did not help.

4.4. Weighted adaptation to target language

In multilingual acoustic modeling it is often desirable to give higher weight to the acoustic data of the target language when training an acoustic model for that language. MAP adaptation [10] is an efficient way to accomplish this, since it adapts an existing multilingual model using acoustic data of the target language and a pre-determined weight. This approach has also been used previously by researchers to compensate for sparseness of training data in the target language by leveraging data from related languages. This approach can produce adapted models that have better performance than the monolingual models of the target language. We performed MAP adaptation of the multilingual models using the acoustic data of the target language, in this case Dari[6]. In MAP adaptation we chose a weight of 20 for the Dari adaptation data based on recognition accuracy on a held-out set. In Table we can see MAP adaptation results, 5 in the rows labeled "+MAP adapt", for both Trilingual and Bilingual models for both cases of using all acoustic training data or 10 hours of data for the 2000x16 models for the Dari evaluation set. It can be seen that MAP-adapted acoustic models are an absolute 1-2% better in WER compared to multilingual and monolingual models trained using either all data or 10 hours of data.

Experiment	All data	10 hrs
Baseline: Lang dep	33.6	37.1
Trilingual (D,F,P) + Language flag	35.2	37.0
+MAP adapt	**33.2**	**35.0**
Bilingual (D,F) + Language flag	33.5	36.4
+MAP adapt	**32.6**	**36.0**

Table 5. WER (%) for Dari using 2000x16-Gaussians multilingual acoustic models before and after MAP-adaptation on Dari-only data

5. CONCLUSIONS

We have analyzed in detail multilingual acoustic modeling approaches for three closely related languages, Dari, Farsi and Pashto by leveraging prior linguistic knowledge. Under simulated conditions when the amount of available training data is low, the standard multilingual modeling approaches reported previously are able to achieve speech recognition performance very close to the monolingual model. When the amount of available data for each language increases, then the performance of the multilingual model is significantly lower compared to the monolingual one, for the same number of total model parameters. We also explored MAP adaptation of the multilingual models for the target language, which is able to improve performance compared to the monolingual baseline models by about 3% relative WER.

Acknowledgments: This work is supported by the Defense Advanced Research Projects Agency (DARPA) under contract number N10PC20000. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. We also thank our colleague Jing Zheng at SRI International for valuable discussions.

6. REFERENCES

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

[2] J. Kohler, "Multilingual phone models for vocabulary-independent speech recognition," *Speech Communication*, vol. 35, 2001.

[3] U. Uebler, "Multilingual speech recognition in seven languages," *Speech Communication*, vol. 35, 2001.

[4] W. Byrne et al., "Towards language independent acoustic modeling," in *Proceedings of ICASSP*, 2000, vol. 2.

[5] J. J. Sooful and E. C. Botha, "An acoustic distance measure for automatic cross-language phoneme mapping," in *Proceedings of PRASA*, 2001.

[6] T. Schultz, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Eurospeech '97*, 1997.

[7] T. Schultz and A. Waibel, "Polyphone decision tree specialization for language adaptation," in *Proceedings of ICASSP*, 2000, vol. 3.

[8] J. Zheng, A. Mandal, et al., "Implementing SRIs Pashto speech-to-speech translation system on a smart phone," in *Proc. of IEEE SLT Workshop*, 2010.

[9] Akbacak et al., "Recent advances in SRIs Iraqcomm Iraqi-Arabic-English speech-to-speech translation system," in *Proc. ICASSP*, 2009.

[10] J.L. Gauvian and C.H. Lee, "Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, 1994.

[6]The acoustic data of the target language was also used for training the multilingual model