An Investigation of Heuristic, Manual and Statistical Pronunciation Derivation for Pashto

Upendra V. Chaudhari #1, Xiaodong Cui #2, Bowen Zhou #3, and Rong Zhang #4

IBM T.J. Watson Research Center
1101 Kitchawan Road, Rt. 134
Yorktown Heights, NY 10598, USA
[1] uvc@us.ibm.com
[2] cuix@us.ibm.com
[3] zhou@us.ibm.com
[4] zhangr@us.ibm.com

***Abstract*—In this paper, we study the issue of generating pronunciations for training and decoding with an ASR system for Pashto in the context of a Speech to Speech Translation system developed for TRANSTAC. As with other low resourced languages, a limited amount of acoustic training data was available with a corresponding set of manually produced vowelized pronunciations. We augment this data with other sources, but lack pronunciations for unseen words in the new audio and associated text. Four methods are investigated for generating these pronunciations, or baseforms: an heuristic grapheme to phoneme map, manual annotation, and two methods based on statistical models. The first of these uses a joint Maximum Entropy N-gram model while the other is based on a log-linear Statistical Machine Translation model. We report results on a state of the art, discriminatively trained, ASR system and show that the manual and statistical methods provide an improvement over the grapheme to phoneme map. Moreover, we demonstrate that the automatic statistical methods can perform as well or better than manual generation by native speakers, even in the case where we have a significant number of high quality, manually generated pronunciations beyond those provided by the TRANSTAC program.**

I. Introduction

We study the issue of pronunciation derivation for Automatic Speech Recognition (ASR) of Pashto, a heavily inflected, morphologically complex language with limited resources available for data collection and annotation. It is primarily spoken in Afghanistan, Pakistan and Iran, and is a member of the Indo-Iranian languages. An important aspect in building an ASR system for a given language is to properly describe the pronunciations that may be observed when the system is used. Heavily studied languages would have enough resources devoted to the task of defining these pronunciations via human annotators. These would have a significant amount of audio training data with transcripts, and all of the words in the transcripts would be covered by a human generated pronunciation dictionary. Note that words can have multiple pronunciations and a carefully annotated database would identify these differences.

Currently, state of the art MSA (Modern Standard Arabic) ASR systems are trained on over 1800 hours of data [1]. Such systems rely on a lot of annotated training data, which presumably contains multiple pronunciations that can be learned in the training procedure. Yet the resources necessary for such annotation are seldom available, and moreover the requirement for rapid deployment in new languages exacerbates the problem [2]. Low resourced languages such as Pashto [3] have limited amounts, on the order of 90 hours for the system used here, of properly annotated data. In such scenarios, it is common to try and obtain additional data. But even if new audio and associated transcripts are obtained, typically resources are not available to generate the pronunciations, also referred to as baseforms, for unseen words which are required in ASR training and decoding.

Previous work has addressed the general problem of automatic baseform generation (prediction, derivation, also referred to as grapheme to phoneme conversion) in various contexts [4] [5] [6] [7] [8] [9] [10]. In [10], automatic generation of pronunciations for Pashto based on text-to-phoneme (T2P) tools from CMU was analysed in the context of a 34 hour, ML trained ASR system. Subsets of varying size were selected from a set of manually generated (from Appen) pronunciations and used to train the T2P tools, and it was shown that for the low resource case, with only 1K words for training, the T2P tool has better performance than using a grapheme based approach [11]. Further, in [12], it was shown that ASR performance in the Iraqi-Arabic dialect is improved when multiple pronunciations of words, obtained either from pronunciation dictionaries or automatically generated, are used as opposed to a single pronunciation.

One might assume that the best performance would be obtained if human annotators could generate these baseforms, however inter annotator differences might affect the results, especially for a language such as Pashto. And, in the likely case where human annotators are not available, automatic techniques, e.g. those that use heuristics or statistical models for generating the pronunciations, must be used. We focus on two methods for this task that have proven to be successful in dealing with the types of complexity inherent in the baseform generation task for Pashto. These are the joint Maximum

Entropy (ME) n-gram model described in [7] and a method that uses sequence translation with a phrase based log-liner translation model [9]. We compare the performance of these methods and relate them to human performance as well as to an heuristic approach that maps graphemes to phonemes with a context independent map. Our experiments are based on a total of 105 hours of training data and we report results on a state of the art, discriminatively trained, evaluation ASR system. We show that the statistical methods can perform as well or better than human annotators and represent a viable approach to baseform generation in an evaluation context for languages with limited annotation resources.

The paper is organized as follows. Section II describes the components of the ASR system and section III describes the experimental setup. The methods for baseform generation are presented in section IV. Results are given in section V and conclusions in section VI.

II. ASR Overview

The experiments presented are performed on an ASR module that is one in a chain which comprises the IBM English-Pashto translation system developed for applications such as TRANSTAC: Spoken Language Communication and Translation System for Tactical Use [13], [14]. Accuracy in this module is critical so as not to propagate errors. We focus specifically on Pashto ASR.

A. Acoustic Modeling

For the Speaker Independent Maximum likelihood (ML) model, the base feature space is constructed from 24-dimensional PLPs including energy. Cepstral mean normalization (CMN) is followed by linear discriminative analysis (LDA) which projects 9 concatenated PLP frames down to 40 dimensions. Maximum likelihood linear transformation (MLLT) is also applied at this stage. A diagonal, quinphone acoustic model (AM) with states tied by a decision tree is used. Based on the ML model, feature space discriminative training using the boosted MMI criterion (FMMI)[15] is applied. This is followed by discriminative training in the model space with BMMI[15].

B. Language Modeling

Language modeling utilizes both bilingual and monolingual training data. Subsets are created from the bilingual data corresponding for example to whether the text came from transcription or translation. A separate language model is built (with Modified Kneser-Ney smoothing [16]) for each subset as well as for the monolingual data. These are interpolated to from the final language model.

C. Decoding Structure

The ASR module has a Viterbi decoder running on a static graph, which leads to a very fast decoding process at run-time compared to traditional Viterbi dynamic decoders but comes at the expense of large memory consumption.

III. Experimental Setup

Our experiments are conducted on a 105.5 hour Pashto ASR system (3K states and 80K Gaussians) developed for the TRANSTAC evaluation (see section II). The data breaks down as follows: Appen data taken from the Pashto side of conversational data accounts for 90.6 hours (E2F, F2E, Monolingual) TTS data accounts for 5 hours. Scripts were generated and native speakers spoke the utterances. Data from SRI constitutes 8.9 hours and there is 1 hour of data from the Web. We report results at the ML level and with discriminatively trained models with BMMI [15].

A lexicon of about 17K words with human generated vowelized pronunciations from Appen was used in the experiments (and referred to as the provided lexicon). Note that this is the size of the lexicon at the time the experiments were conducted. The Pashto phoneset used for ASR contained 44 phones. Approximately 50% of training vocabulary from TTS, Web, and SRI data was not covered by the provided lexicon. Overall, the dictionary used for ASR had 27K words. Of those, 10396 words were not covered by the provided lexicon and had pronunciations that were generated by the methods described in the paper or by new native transcribers.

The test data consisted of 4986 utterances with 79236 words from 68 speakers. Heldout Appen data was used as the test set, for which 90% of the words were covered by the provided lexicon described above. The overall OOV rate was 1.7%.

In the experiments, the training and decoding vocabularies are identical across all experiments. Thus we focus on how pronunciation derivation affects the quality of the acoustic model training and ASR performance.

IV. Graphemem to Phoneme Conversion for Pashto

In order to build robust ASR systems, a vocabulary with common pronunciations, or baseforms, is required. Low resource languages will not typically have a large annotated vocabulary. Moreover, highly inflected languages and languages for which mixing regional dialtects is common, such as Pashto, will magnify the issue because of the variety of word forms and pronunciations that occur. As compared to languages lacking such variation, relatively more annotation per concept may be required. This also implies that manual annotation is not straightforward for Pashto. Given this, the question is how successful can automatic methods be at generating baseforms for ASR and what is the impact on word error rate as compared to manual baseform generation.

The basic task of grapheme to phoneme (G2P) conversion is to map the spelling of a word, in terms of graphemes (letters), into a baseform, or pronunciation, which is a sequence of phones (we use the terms phones and phonemes interchangeably in the text). The general problem is as follows. We have a vocabulary $\mathrm{V} = \cup\mathbf{g}$ where $\mathbf{g}$ is the grapheme (letter) sequence corresponding to a word. We also have a dictionary

$$\mathrm{D} = \cup_{\mathbf{g}\in\mathrm{V}} \cup_i (\mathbf{g}, \mathbf{p}_i^{\mathbf{g}}) : \mathbf{p}_i^{\mathbf{g}} \text{ is a baseform for } \mathbf{g}.$$

This definition of the dictionary allows for multiple pronunciations of a vocabulary item. Here, $\mathbf{g}$ is a sequence of letters

l_i from the Pashto alphabet and $\mathbf{p}$ is a sequence of phones p_i from the Pashto phoneset.

We can refine the definition as follows. The vocabulary is further split into one subset that has elements which are covered by the provided vowelized lexicon and one containing those that are not, e.g. $\mathrm{V} = \mathrm{V_c} \cup \mathrm{V_u}$. As mentioned before, $\mathrm{V_u}$ contains 10396 words. The paradigm for our experiments is that the covered subset $\mathrm{V_c}$ is always the same with baseforms given by the provided lexicon, whereas the baseforms for the uncovered subset $\mathrm{V_u}$ come from the various methods we study. Final acoustic models are trained all the way through starting from ML to FMMI/BMMI (see section II) for each method that is studied using the resulting dictionary D.

We investigate four approaches for G2P conversion. An heuristic grapheme to phoneme map, manual generation of pronunciations, and two statistical approaches based on the joint Maximum Entropy (ME) n-gram model[7] (Joint-ME) and sequence translation with a phrase based log-liner translation model[9] (SMT).

A. Heuristic Grapheme to Phoneme Map

Grapheme based pronunciations (Grapheme-Map) map the spelling of a word into a sequence of symbols essentially in one to one correspondence to the sequence of letters in the spelling, perhaps modified by some general heuristics. These are simple to generate and have yielded acceptable results, especially when the spelling of a word matches very closely to the pronunciation. Here, this approach is defined as a mapping of graphemes to phones that uses no contextual information and which was generated by hand via heuristic rules. In particular, since there were more letters than phones, a many to one mapping of some letters to phones was necessary in a few cases. Note that this approach is not the same as using the graphemes as the phone set. This procedure results in a set of unvowelized pronunciations for the elements of $\mathrm{V_u}$.

B. Manual Generation

On the other hand, the most accurate baseforms can be expected to be those that are manually derived (Manual). Here, these were generated for data that was not covered by the provided lexicon (primarily from TTS, but also from the Web, and SRI datasets) by native speakers with significant experience with annotation. Further, we stress that the manually generated pronunciations were of a very high quality and the TTS system generated using them was the state of the art. We have 8K such pronunciations and define this set as $\mathrm{V_{manual}}$. Given that this annotation focused on TTS, the overlap with $\mathrm{V_u}$ is not complete, but approximately 4K words. We denote this subset as $\mathrm{V_{u,manual}}$. Grapheme-Map or another method must be used to generate pronunciations for the remaining words in $\mathrm{V_u}$.

C. Pronunciations via Statistical Models

There are many issues that complicate the grapheme to phoneme (G2P) mapping. Contextual local and non local effects, deletion and insertion, many to many mappings and reordering are among the phenomena that can be observed. In light of these issues, we choose and compare two different statistical methods for grapheme to phoneme conversion (baseform prediction) that have proven successful in dealing with them. One of them uses the joint Maximum Entropy (ME) n-gram model [7] and the other uses a phrase based log-linear translation model with alignments based on sequences of context dependent units [9]. The provided lexicon with 17K human generated vowelized pronunciations is taken as data to train models of the pronunciation process. These are subsequently used to predict pronunciations for new unseen words.

1) Joint ME N-Gram Model: Here, a Maximum Entropy (ME) N-gram model [7] (Joint-ME) is used to estimate the component conditional distributions in

$$Pr(\mathbf{g}, \mathbf{p}) = \sum_{C \to \{\mathbf{g},\mathbf{p}\}} \prod_{i=1}^{m} Pr(c_i | c_1, \ldots, c_{i-1}), \qquad (1)$$

where

$$C = \{\ldots, c_i, \ldots\}$$

is a sequence of m "graphonemes"

$$c_i = \begin{cases} (l, p) \\ (\epsilon, p) \\ (l, \epsilon) \end{cases}$$

with l an element of the Pashto alphabet, p an element of the Pashto phone set, and ϵ representing the empty symbol. This sequence C describes the alignment, h, of letters to phones, which is hidden and often not unique. The notation $C \to \{\mathbf{g}, \mathbf{p}\}$ indicates all valid alignments h for the grapheme sequence $\mathbf{g}$ and phone sequence $\mathbf{p}$. Decoding to produce the baseform $\mathbf{p}^*$ involves the following optimization over phone sequences.

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} \sum_{C \to \{\mathbf{g},\mathbf{p}\}} \prod_{i=1}^{m} Pr(c_i | c_1, \ldots, c_{i-1}), \qquad (2)$$

and for N-grams

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} \sum_{C \to \{\mathbf{g},\mathbf{p}\}} \prod_{i=1}^{m} Pr(c_i | c_{i-N+1}, \ldots, c_{i-1}). \qquad (3)$$

2) Statistical Machine Translation: Here, a log-linear model [9] (SMT)

$$Pr(\mathbf{p}, h | \mathbf{g}) = \frac{\exp[\Sigma_{m=1}^{M} \lambda_m f_m(\mathbf{p}, \mathbf{g}, h)]}{\Sigma_{\mathbf{p}', h'} \exp[\Sigma_{m=1}^{M} \lambda_m f_m(\mathbf{p}', \mathbf{g}, h')]} \qquad (4)$$

is used as the means of combining scores from a phrase translation model, a lexicon translation model, and a phone n-gram language model, each of which is a feature function $f_m(\mathbf{p}, \mathbf{g}, h)$ whose arguments are a phone sequence $\mathbf{p}$, a grapheme sequence $\mathbf{g}$, and an alignment between them h. Thus M = 3. Each function has an associated weight λ_m determined in training. Decoding involves optimization over both the phone sequences and the alignments.

TABLE I
PERFORMANCE OF GRAPHEME-MAP WITH ML MODELS.

Model Type	*WER%*
ML	40.20
DT	34.47

$$\mathbf{p}^* \leftarrow \text{argmax}_{\mathbf{p},h} \left\{ \exp[\Sigma_{m=1}^{M} \lambda_m f_m(\mathbf{p}, \mathbf{g}, h)] \right\} \quad (5)$$

In [9], it was observed that the use of context dependent units when generating alignments h, as followed here, led to an approximately 20% relative reduction in phoneme error rate over context independent units.

V. EXPERIMENTAL RESULTS

Results for the simplest approach, Grapheme-Map, can be seen in table I for both Maximum Likelihood (ML) and discriminatively trained (DT) models. In this case, all of the pronunciations for words in $\mathrm{V_u}$ are derived via the map. Recall that those for $\mathrm{V_c}$ are given by the provided lexicon for all experiments. We expect that this performance is sub-optimal since the new pronunciations are unvowelized and produced without taking context into account.

Recall that $\mathrm{V_{u,manual}}$ is the set of uncovered words (with respect to the provided lexicon) that have new manual pronunciations (see section IV-B). Table II compares the performance with ML models when using pronunciations derived manually, by using Joint-ME, or by using SMT for the words in $\mathrm{V_{u,manual}}$, with the remaining words handled using Grapheme-Map. In every case, there is an improvement over the ML performance when using Grapheme-Map for the words in $\mathrm{V_{u,manual}}$ (table I).

Since we can use Joint-ME and SMT to predict all of the baseforms in $\mathrm{V_u}$, results for this case are shown in table III. Also included in this table are the results when Manual is used for the words in $\mathrm{V_{u,manual}}$ and SMT for the remaining words in $\mathrm{V_u}$ (since there are no manual pronunciations for these words and SMT had the better performance of the two statistical methods in table II). This is denoted by Manual+SMT. Note that SMT has the best performance, 0.8% absolute better than the results with Grapheme-Map. Particularly interesting is that using SMT for all of the words is better than Manual+SMT. Thus we train discriminative models for Manual+SMT, Joint-ME, and SMT.

See table IV for the results after discriminative training (all of the words in $\mathrm{V_u}$ have derived baseforms as for table III). Here, the Joint-ME approach gives the best performance, 1% absolute better than with Grapheme-Map and better than using Manual+SMT pronunciations. While the use of manually derived baseforms is clearly better than using Grapheme-Map, the statistical methods ultimately give better performance. These results are important, because discriminatively trained models are generally very good at accounting for sub-optimal ML models [15]. Thus, a difference at this level can be assumed to be a real difference.

TABLE II
PERFORMANCE WITH ML MODELS AND DIFFERENT METHODS FOR $\mathrm{V_{U,MANUAL}}$.

Baseform Generation	*WER%*
Manual	39.63
Joint-ME	39.65
SMT	39.54

TABLE III
PERFORMANCE WITH ML MODELS AND DIFFERENT METHODS FOR $\mathrm{V_U}$.

Baseform Generation	*WER%*
Manual+SMT	39.58
Joint-ME	39.53
SMT	39.38

A. Analysis

Given our experimental setup, with 90% of the test data covered by the provided lexicon, we hypothesize that the differences in observed performance are essentially due to the quality of the acoustic models that are trained. With regard to the Joint-ME and SMT models, since the same training data is used, it is reasonable to expect some overlap in the outputs. In fact, examining the subset of the 10396 uncovered words which were subsequently covered by the manual baseforms ($\mathrm{V_{u,manual}}$), the outputs of the two models on these words overlap by about 50%. By overlap, we mean that identical baseforms were generated. On the other hand, the overlap of each automatic method to manual generation is approximately 30%.

The fundamental question is how do differences in baseform generation affect the quality of the acoustic models. Since we are using context dependent models, differences in baseforms (pronunciations) can have an effect through the phone N-grams that are observed in training. Here, we are using quinphone acoustic models, so we look at the variety of 5-gram contexts that are generated for the training data when the various dictionaries (as used for table IV) are used to expand the training text into phone sequences. Interestingly, we find that when we compare the sets of quinphones where at least one element represents a short vowel, the Joint-ME and SMT approaches respectively produce about 2.8% and 2.3% more unique contexts beyond those produced with the Manual+SMT approach, which in turn produces 4.5% more unique contexts than Grapheme-Map. This trend correlates with the performance differences in table IV, and we hypothesize that the additional contexts lead to richer decision trees and better

TABLE IV
PERFORMANCE WITH DT MODELS AND DIFFERENT METHODS FOR $\mathrm{V_U}$.

Baseform Generation	*WER%*
Grapheme-Map	34.47
Manual+SMT	33.71
Joint-ME	33.54
SMT	33.62

TABLE V
PERFORMANCE WITH DT MODELS AND JOINT-ME N-BEST LIST USED FOR ALTERNATE PRONUNCIATIONS.

Baseform Generation	*WER%*
Joint-ME	33.54
Joint-ME (N=2)	33.55
Joint-ME (N=3)	34.56

TABLE VI
JOINT-ME WITH SMT AS ALTERNATE PRONUNCIATIONS.

Model Type	*WER%*
ML	40.00
DT	34.75

acoustic models.

B. Experiments with Multiple Predicted Pronunciations

It has been observed that using multiple pronunciations leads to improved performance [12] (for Iraqi-Arabic). We also experiment with using multiple pronunciations, taken from the best performing systems in IV. For the first experiment, we use the N-best list obtained from decoding with the Joint-ME models to get alternate pronunciations. Table V gives the results for 1 and 2 alternate pronunciations (N = 2 or 3 total pronunciations). The results are nearly identical to the single pronunciation case.

Interestingly, when using the SMT pronunciations as alternates for the Joint-ME system, the performance degrades, as seen in table VI, and indicates that in this case the added pronunciations are confusable.

VI. CONCLUSIONS

We have conducted an investigation into methods of deriving pronunciations for Pashto, an example of a language with limited resources available in terms of human annotation. Four methods were studied: an heuristic grapheme to phoneme map, manual generation of pronunciations, and two automatic methods based on statistical models, Joint-ME and SMT. We have shown that it is possible to achieve state of the art performance, comparable to or better than with high quality human annotation, using the automatic methods in the context of a discriminatively trained, evaluation Pashto ASR system. Analysis revealed that for our particular data, the statistical methods generated a richer set of quinphone contexts. In future experiments, we will explore whether using all the manual pronunciations we have, from Appen and internal sources, to train the statistical models can further improve performance. Also, we plan to investigate whether N-gram statistics can play a role in model training for grapheme to phoneme conversion.

REFERENCES

[1] B. Kingsbury, H. Soltau, G. Saon, S. Chu, H.-K. Kuo, L. Mangu, S. Ravuri, N. Morgan, and A. Janin, "The IBM 2009 gale arabic speech transcription system," in *Proc. ICASSP*, 2011.

[2] T. Schultz and A. W. Black, "Challenges with rapid adaptation of speech translation systems to new language pairs," in *In the Proceedings of ICASSP*, 2006.

[3] A. Kathol, K. Precoda, D. Vergyri, W. Wang, and S. Riehemann, "Speech translation for low-resource languages: The case of pashto," in *In the Proceedings of Interspeech*, 2005, lisbon, Portugal.

[4] A. W. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *Proc. of ESCA*, 1998, pp. 77–80.

[5] W. Byrne, M. Finke, S. Khunanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation modelling using a hand-labelled corpus for conversational speech recognition," in *Proc. ICASSP*, 1998.

[6] M. Bisani and H. Ney, "Investigations on joint-multigram models for grapheme-to-phoneme conversion," in *in Proc. Int. Conf. on Spoken Language Processing*, 2002, pp. 105–108.

[7] S. F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proceedings of Eurospeech*, 2003.

[8] T. Rama, A. K. Singh, and S. Kolachina, "Modeling letter-to-phoneme conversion as a phrase based statistical machine translation problem with minimum error rate training," in *Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium*, 2009, pp. 90–95, boulder, Colorado.

[9] R. Zhang and B. Zhou, "Applying log linear model based context dependent machine translation techniques to grapheme-to-phoneme conversion," in *Proc. ICASSP*, 2010.

[10] R. Prasad, S. Tsakalidis, I. Bulyko, C.-L. Kao, and P. Natarajan, "Pashto speech recognition with limited pronunciation lexicon." in *ICASSP'10*, 2010, pp. 5086–5089.

[11] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, and F. Kubala, "Audio indexing of arabic broadcast news."

[12] H. Al-Haj, R. Hsiao, I. Lane, A. Black, and A. Waibel, "Pronunciation modeling for dialectal arabic speech recognition," in *ASRU*, 2009, merano, Italy.

[13] DARPA, "Spoken language communication and translation system for tactic use," http://www.darpa.mil/IPTO/programs/transtac/transtac.asp.

[14] Y. Gao, B. Zhou, L. Gu, R. Sarikaya, H. kwang Kuo, A.-V. Rosti, M. Afify., and W. Zhu, "IBM MASTOR: Multilingual automatic speech-to-speech translator," in *ICASSP*, 2006, pp. 1205–1208.

[15] D. Povey, D. Kanevsky, and B. Kinsbury, "Boosted MMI for model and feature-space discriminative training," *Proc. of ICASSP*, pp. 4075–4060, 2008.

[16] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proc. of ACL*, 1996, pp. 310–318.