Eurospeech 2003, Geneva

Language-Adaptive Persian Speech Recognition

Naveen Srinivasamurthy and Shrikanth Narayanan

Department of Electrical Engineering,
Integrated Media Systems Center,
University of Southern California, Los Angeles
[snaveen,shri]@sipi.usc.edu

Abstract

Development of robust spoken language technology ideally relies on the availability of large amounts of data preferably in the target domain and language. However, more often than not, speech developers need to cope with very little or no data, typically obtained from a different target domain. This paper focuses on developing techniques towards addressing this challenge. Specifically we consider the case of developing a Persian language speech recognizer with sparse amounts of data. For language modeling, there are several potential sources of text data, e.g., available on the Internet, to help bootstrap initial models; however, acoustic data can be obtained only by tedious data collection efforts. The drawback of limited Persian acoustic data can be partially overcome by making use of acoustic data from languages that have vast resources such as English (and other languages, if available). The phoneme sets especially for diverse languages such as English and Persian differ considerably. However by incorporating knowledge-based as well as data-driven phoneme mappings, reliable Persian acoustic models can be trained using well-trained English models and small amounts of Persian re-training data. In our experiments Persian models re-trained from seed models created by data-driven phoneme mappings of English models resulted in a phoneme error rate of 19.80% as compared to a phoneme error rate of 20.35% when the Persian models were re-trained from seed models created by sparse Persian data.

1. Introduction

In the past few years automatic speech recognition (ASR) has had tremendous success in several different languages. However statistical approaches used in ASR rely on the availability of large amounts of speech and text data. The rapid transfer of ASR technologies to new languages is hampered by the non-availability of sufficient data in new languages, specifically in the target domain of the application. This is especially severe at the initial launching stages of application development. Text data required for language modeling has several potential sources e.g., newspapers from the Internet. However speech data required for acoustic modeling can typically only be obtained by tedious data collection efforts. To overcome this drawback speech data available in resource rich languages like English can be used to help build the target language acoustic models. This effectively reduces the amount of speech data that needs to be collected in the target language thus enabling savings in cost as well as reducing the time required to deploy ASR systems in new target languages.

The DARPA *Babylon* project [1] investigates the development of speech to speech (S2S) systems capable of supporting conversations between people speaking different languages without requiring human translators. The languages considered were English and Persian (Farsi), and the target task was medical domain (details of the S2S system are in [2]).

In this paper, we specifically consider the development of a Persian speech recognizer using limited amount of Persian speech data and borrowing acoustic data from English speech data. Note that we only address the problem of acoustic modeling, we do not address the problem of language modeling.

Phoneme mappings from the English phonemes to the Persian phonemes are derived to create seed models for the Persian acoustic models using English speech data. These seed models are further adapted or re-trained using the limited amount of data available in Persian. In addition to a knowledge (linguistic) based phoneme mapping we present a *novel data driven* phoneme mapping technique based on the Earth Movers Distance (EMD). The difficulty in deriving phoneme mappings between different languages is that closed form *pdf* "distance" metrics are not available for Gaussian mixture models (GMMs), which are typically used in acoustic models. Our proposed algorithm combines a bipartite network flow algorithm with distance between Gaussian *pdfs* (which have a closed form expression) to provide a low complexity phoneme (HMM) distance metric. Using only the sparse Persian speech data to build a speech recognizer resulted in 20.35% phoneme error rate. However the proposed EMD based data driven phoneme mapping technique achieved 19.80% phoneme error rate on the same task, which is a 2.70% relative reduction in phoneme error rate.

Previous work in the area of language independent acoustic modeling is presented in Section 2. Section 3 presents details of the knowledge and data driven phoneme/sub-phoneme mappings. Section 4 presents experiments and results. Conclusions are in Section 5.

2. Previous Work

The most intuitive and straightforward approach is to use knowledge (linguistic) based phonetic mappings. This has the desirable feature of not requiring any speech data in the target language. The proposed knowledge based mappings are based on IPA phoneme definitions [3, 4] or phoneme classes [5]. However it has been observed that data driven approaches usually outperform these knowledge based approaches ([3, 6] and Sec-

tion [4]).

Language mixed/tagged approaches [7] have been proposed for acoustic model combination. These approaches initially map phonemes based on a knowledge (IPA) based approach. Then for phonemes mapped to the same IPA symbol in the language mixed approach, all parameters of the Gaussian mixture are kept the same and in the language tagged approach, only the means and variances are kept the same and language specific mixture weights are trained for each of the languages. While this makes parameter estimation easy it might be restrictive in modeling the differences in phonemes from different languages.

A data driven confusion matrix based phoneme mapping technique [3] has been previously proposed. The confusion matrix is created by running a free phoneme recognizer using the acoustic models of the source phonemes. The phoneme alignment achieved by this phoneme recognizer is compared to the target language phoneme alignment of the utterance. The target phoneme alignment is then created either by human transcription or by force alignment. The entries in the confusion matrix indicate the co-occurrences between the source and target phonemes. The target phoneme is created by choosing the source phoneme which has the highest co-occurrences with the target phoneme. The problem with this approach is that, in general, phoneme recognition rates are not very reliable even when applied to the same language in which the phonemes were designed, it can be expected that when source phonemes are used for phoneme recognition of target language utterances the phoneme recognition performance will be poor. This might result in bad estimates for the source to target phoneme mappings.

A language adaptive clustering [3] previously proposed, imposes for computational reasons, the use of affine transformation while transforming feature vectors between languages. Since the differences between the same phonemes in different languages is definitely influenced by contexts, the restriction of using only affine transformation maybe be restrictive.

Different phoneme distance matrices have been evaluated for deriving phoneme mappings [6]. It is shown that the Jeffreys-Matusia distance measure achieves better performance than knowledge based phoneme mappings. However since most of the distances they have considered do not have closed form expressions for *pdfs* (other than Gaussians), their solution in using these distance metrics for mixture Gaussians is to calculate the distance per mixture per state and add the distances to get the distance between phonemes. We will show in Section [3] that there exists a more formal method of combining the distance between Gaussian to calculate the distance between mixture Gaussians and find the phoneme mappings.

Unlike previous approaches we do not impose either transformation or parameter constraints. The phoneme mappings are determined completely by the acoustic models in the source and target languages.

3. English to Persian Phoneme Mappings

Similar to previous work [3, 7] we also mapped English to Persian phonemes using linguistic knowledge. The detailed phoneme mappings are available at *<http://sail.usc.edu/transonics/documents.html>*. While it can be expected that the knowledge driven approach will result in meaningful phoneme mappings, its failing are two-fold (i) speaking styles and phoneme contexts especially between diverse languages like Persian and English can result in differences between "similar" phonemes, thus the mapping may not be optimal at the acoustic feature level; and (ii) unseen phonemes in Persian (the Velar fricative and the Uvular stop) do not have equivalent representations in English. These reasons result in significant degradations in recognition performance when pure knowledge driven phoneme mappings are adopted.

To overcome these drawbacks a data driven approach is required. Here the phoneme mappings are automatically derived from Persian and English speech data. The advantages of this approaches are (i) acoustic models in the source and target languages determine the "optimal" mappings, this enables better ASR performance since we operate directly on the models being used for speech recognition and (ii) sub-phonetic mappings can be derived which enables us to create better target acoustic models from source acoustic model components, furthermore the sub-phonetic mapping enables the creation of acoustic models for unseen Persian phonemes, which was not possible by the knowledge driven approach.

3.1. Phoneme mapping using the Earth Movers Distance

One of the main difficulties in using a data driven approach is the lack of a suitable distance metric between acoustic models which can be used to categorize phoneme similarities. We adopt the Earth Movers Distance (EMD) to find similarities between the phonemes. EMD was originally introduced for navigation in image databases [8]. It has subsequently been extended to 3-D vector fields [9] and also has been used for content based music similarity [10]. EMD is a method to find distances between "signatures". For our problem the acoustic models of the English and Persian phonemes are regarded as signatures, and EMD is used to find the distance between them. These phoneme distances are calculated for all phoneme pairs. Each Persian phoneme is then mapped to the English phoneme which has the smallest distance from it. Note that with this approach unseen Persian phonemes will be assigned to the closest English phoneme. Since this might result in bad seed models for the unseen phonemes, we can operate at the HMM state level. Now the seed HMMs are constructed by borrowing states from different HMMs to construct a Persian HMM.

Before presenting the phonetic and sub-phonetic mappings, the EMD algorithm is briefly described. Assume we want to find the distance between two Gaussian mixture models (GMMs); $G_t \sim \sum_{i=1}^{N_t} c_i^t N(\mu_i^t, \sigma_i^t)$ and $G_s \sim \sum_{j=1}^{N_s} c_j^s N(\mu_j^s, \sigma_j^s)$ where μ_k^x, σ_k^x and c_k^x are the mean, standard deviation and mixture weight, respectively, of the k^{th} Gaussian in the GMM G_x (note that the number of mixtures, N_t and N_s, in the two GMMs can be different). In the spirit of EMD the distance between G_t and G_s can be formulated as the "amount of work" needed to convert the *pdf* defined by G_s to the *pdf* defined by G_t[1]. Let d_{ij} be the work needed to transform a unit probability mass from $N(\mu_j^s, \sigma_j^s)$ $j \in N_s$ to $N(\mu_i^t, \sigma_i^t)$ $i \in N_t$. Then we have a bipartite network flow prob-

[1]Note that the distance between G_s and G_t can be defined as $D(G_t||G_s)$, i.e., the KL distance, however there is no closed form expression for KL distance between GMMs and the problem is further complicated when the GMMs model high dimensional vectors, as is the case in acoustic models of practical speech recognizers.

lem of finding the flows, f_{ij}, that minimize the cost

$$\sum_{i \in N_t} \sum_{j \in N_s} d_{ij} f_{ij} \tag{1}$$

subject to the constraints

$$f_{ij} \geq 0, \ \sum_{j \in N_s} f_{ij} = c_i^t \text{ and } \sum_{i \in N_t} f_{ij} \leq c_j^s \ i \in N_t, \ j \in N_s \tag{2}$$

The distance d_{ij} should measure the difference between *pdfs* $N(\mu_i^t, \sigma_i^t)$ and $N(\mu_j^s, \sigma_j^s)$. A suitable metric to measure this dissimilarity is the symmetric KL distance which for Gaussians is

$$d_{ij} = \frac{1}{2}\left[\frac{(\sigma_i^t)^2}{(\sigma_j^s)^2} + \frac{(\sigma_j^s)^2}{(\sigma_i^t)^2} + (\mu_i^t - \mu_j^s)^2 \cdot \left(\frac{1}{(\sigma_i^t)^2} + \frac{1}{(\sigma_j^s)^2}\right)\right] - \frac{1}{2} \tag{3}$$

This distance has the desirable property of being zero only when the two *pdfs* are same and being small when the *pdfs* are similar and large when they are not. If f_{ij}^* is the flow minimizing the cost in Equation (1) then the distance between G_s and G_t is

$$D_{GMM}(G_s, G_t) = \frac{\sum_{i \in N_t} \sum_{j \in N_s} d_{ij} f_{ij}^*}{\sum_{i \in N_t} \sum_{j \in N_s} f_{ij}^*} \tag{4}$$

Since acoustic features used in ASRs are vectors which are usually modeled by diagonal co-variance matrices, the distance between states can be calculated as

$$D_{STATE}(S_s, S_t) = \sum_{m=1}^{M} D_{GMM}(G_{S_s}^m, G_{S_t}^m) \tag{5}$$

where G_x^m is the GMM of the m^{th} acoustic feature in the x^{th} state. Equation (5) can be used to find the distance between HMMs. Let P_f, P_e be the set of all Persian and English phonemes respectively. Let each of the phonemes $p_f \in P_f$ and $p_e \in P_e$ be modeled by a Persian HMM H_{p_f} and an English HMM H_{p_e} respectively. Let S_x^s denote state s of an HMM H_x. Then the distance between the HMMs (phonemes) is

$$D_{HMM}(p_e, p_f) = \sum_{s=1}^{S} D_{STATE}(S_{p_e}^s, S_{p_f}^s) \tag{6}$$

We assume that both the Persian and English phonemes have the same number of states, which is a reasonable assumption.

Given the State and HMM distance definitions (Equations (5) and (6)), we are ready to propose our data driven phonetic/sub-phonetic mapping techniques.

Algorithm 1 *(Phonetic mapping: English to Persian)*
Step 1: *Design HMMs for all English phoneme models and HMMs for all Persian phonemes (with limited available Persian speech data).*
Step 2: *for each* $p \in P_f$ $M(p) = argmin_{q \in P_e} D_{HMM}(q, p)$
Step 3: *Set* $H_{M(p)}$ *as the seed model for Persian phoneme* p

Algorithm 2 *(Sub-phonetic mapping: English to Persian)*
Step 1: *Design HMMs for all English phoneme models and HMMs for all Persian phonemes (with limited available Persian speech data).*
Step 2: *for each* $p \in P_f$
Step 3: *for each* $s \in S$ $M_s(p) = argmin_{q \in P_e} D_{STATE}(S_q^s, S_p^s)$
Step 4: *For Persian phoneme* p*, use* $M_s(p)$ *as state* s *of the seed HMM* H_p

The seed HMM models constructed by either knowledge or data driven approaches can be used for either adaptation or re-training using the limited amount of target domain Persian data available.

4. Experiments and Results

HTK 3.1 was used to design the recognizers. All speech data used in the experiments were downsampled to 8 kHz. The feature vectors used were 12 MFFCs and the zeroth cepstral coefficient and their Δ and $\Delta\Delta$ derivatives, using hamming windows of 25 ms with a feature vector calculated every 10 ms. Both English and Persian phoneme models had 3 states with 16 GMMs per state. The English models were created from the train subset of TIMIT database. For Persian data we used FARSDAT database, which consists of 22.5 kHz recordings by 300 Persian speakers from 10 different dialect regions of Iran. For our experiments we transcribed data from 80 speakers which gave us approximately 2800 sec of speech. A new transcription scheme USCPers+/USCPron, for Persian was developed [11]. This transcription scheme uses only ASCII characters for representation and also adds the vowels which are usually not present in written Persian (which uses the Arabic script).

Additionally we collected and transcribed speech from 18 native Persian speakers (9 females and 7 males), each of whom read approximately 250 short phrases in the target medical domain. This collected data were used for adaptation/re-training and testing, but not for creation of baseline Persian acoustic models. Since the data available was less, we adopted a hold-one-out strategy (or cross validation) in our experiments, wherein the models were adapted/re-trained using data from 17 speakers and tested on the 18^{th}. This was repeated for all 18 speakers and results reported are the average WERs.

We compare Persian adapted/re-trained ASRs using seed models from (i) sparse Persian speech data (FARSDAT), (ii) knowledge based English phonemes, (iii) data driven phonetic models and (iv) data driven sub-phonetic models, in a medical domain task to evaluate our proposed techniques. Our proposed techniques differ from previous techniques in that we use the phoneme mappings (both knowledge driven and data driven) to create seed models which are further adapted/re-trained using the sparse speech data available in the target Persian language, while most previous techniques propose the development of acoustic models in the target language using data from all available speech data.

To evaluate our proposed techniques we performed a free phoneme recognition with phonotactic bigram constraints and a finite state grammar (FSG) based phrase recognition. Table 1 and 2 show the phoneme recognition error rates and the WER obtained in our experiments. Observe that for both the phoneme and phrase recognition experiment while knowledge based phoneme mappings achieved good results, better results are achieved with data driven sub-phonetic mapping. Specifically after re-training, knowledge based ASR achieved 20.00% phoneme error rate, sub-phonetic mapping based ASR achieved 19.80% phoneme error rate while sparse Persian data ASR achieved a phoneme error rate of 20.35% i.e., we achieved a 2.70% relative reduction in phoneme error rate using sub-phonetic mapping over the sparse data ASR. For the FSG based phrase recognition, sub-phonetic mapping achieved a WER of

20.89%. This is better than the 21.94% WER achieved by knowledge based mapping. Also note that sub-phonetic mapping achieves better results than phonetic mapping in both the experiments. This is due to the fact there is considerable difference between the English and Persian phonemes. Sub-phonetic mapping assembles a Persian HMM by combining states from different English HMMs. This enables construction of "better" Persian HMMs than that was possible by the phonetic mapping techniques.

We also observe that our proposed techniques while having better performance when the re-training is used does not perform as well when only adaptation is used. The possible reason for this is that the adaptation scheme used, MLLR, is restricted to only linear transformations which as mentioned before may not be sufficient to model differences in phonemes between different languages, where phoneme contexts play an important role.

However these results are very encouraging, illustrating that it is possible to make use of acoustic data even between diverse languages like English and Persian to improve the performance of ASRs in languages constrained by sparse data.

Seed Models	Phoneme Error Rate	
	Re-training	Adaptation
FARSDAT	20.35%	38.95%
Knowledge based	20.00%	39.87%
Phonetic mapping	20.13%	57.03%
Sub-phonetic mapping	19.80%	51.48%

Table 1: Phoneme error rates obtained for different approaches. Observe that sub-phonetic mapping ASR achieved the best recognition performance when re-training was used.

Seed Models	Word Error Rate	
	Re-training	Adaptation
FARSDAT	20.52%	40.97%
Knowledge based	21.94%	58.54%
Phonetic mapping	24.30%	78.52%
Sub-phonetic mapping	20.89%	69.62%

Table 2: WERs obtained for different approaches. Observe that the result achieved by the data driven sub-phonetic mapping ASR, out-performs the knowledge based ASR when re-training was used.

5. Conclusions

We proposed a data driven phoneme mapping technique which can be used to use data from "resource" rich languages to create seed models in resource poor target languages to enable design of good acoustic models in the target language. The proposed technique was extended to enable sub-phonetic mappings which enables modeling of unseen target language phonemes. We have addressed only one area of language independent ASR, namely acoustic modeling. However this is arguably the most important issue which currently prevents rapid transfer of ASR technologies to new languages.

As future work we are interested in using these techniques for Arabic. Furthermore our data driven mappings ignored temporal information in deriving the phonetic mappings, we are currently exploring techniques to use temporal information which could enable better data driven phonetic mappings. The current work derived mappings only for context independent models, we are interested in extending the data driven techniques to derive mappings for context-dependent models.

6. Acknowledgments

The authors want to thank Shadi Ganjavi for helping with the knowledge based mappings from English to Persian and Shahryar Karimi-Asthiani for transcribing the speech data used from FARSDAT. This work was supported by the DARPA Babylon program, contract N66001-02-C-6023.

7. References

[1] http://darpa-babylon.mitre.com.

[2] S. Narayanan et. al., "Transonics: A speech to speech system for English-Persian interaction," Submitted to Eurospeech 2003.

[3] Bryne et. al., "Toward language-independent acoustic modeling," tech. rep., John Hopkins University, Language Engineering Workshop, 1999. http://www.clsp.jhu.edu/ws99/projects/asr/.

[4] J. Kohler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *EEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 417 –420, May 1998.

[5] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, "A study of multilingual speech recognition," in *Eurospeech*, (Rhodes, Greece), pp. 359–362, 1997.

[6] J. J. Sooful and E. C. Botha, "An acoustic distance measure for automatic cross-language phoneme mapping," in *Twelfth Annual Symposium of the South African Pattern Recognition Association*, 2001.

[7] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, August 2001. Issues 1-2.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *IEEE International Conference on Computer Vision*, pp. 59–66, January 1998.

[9] R. K. Batra and L. Hesselink, "Feature comparisons of 3-D vector fields using earth movers distance," in *10th IEEE Visualization 1999 Conference (VIS '99)*, (San Francisco, CA), October 1999.

[10] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *ICME 2001*, August 2001.

[11] S. Ganjavi, P. Georgiou, and S. Narayanan, "ASCII based transcription schemes with the Arabic script: The case of Persian language," Submitted to Eurospeech 2003.