Transfer learning for cross-lingual automatic speech recognition

Amit Das

***Abstract*—In this study, an instance based transfer learning phoneme modeling approach is presented to mitigate the effects of limited data in a target language using data from richly resourced source languages. A maximum likelihood (ML) learning criterion is introduced to learn the model parameters of a given phoneme class using data from both the target and source languages. Each phoneme was modeled using a 3 state, 1 Gaussian mixture HMM. Turkish and English were chosen to be the target and source languages respectively. It was found that using only 20 utterances from Turkish, the monophone recognition accuracy in Turkish using transfer learned HMMs is close to the levels of accuracy achieved using standard HMMs when 100 or more utterances from the Turkish training corpus were used.**

***Index Terms*—Transfer learning, maximum likelihood, maximum mutual information**

I. Introduction

WITH the widespread use of hands-free electronic gadgets, speech applications has been gaining more importance throughout the world. The utility of speech technologies like automatic speech recognition (ASR) in these gadgets is dependent on the versatility of ASR systems across users who speak different languages depending on which part of the world they belong to. Hidden Markov Models (HMMs) have gained the widest acceptance in building ASR systems. Ideally, language dependent or monolingual HMMs can be deployed in electronic gadgets where they are expected to be used by a majority of the population speaking the most common language specific to a geographic region. Although feasible, this is not commercially attractive for two reasons. Firstly, data collection of a specific language is a time consuming and expensive process. Secondly, experienced transcribers who can mark word or phoneme boundaries with a high degree of accuracy may be available only for a limited set of more popular languages like English. Hence, the need arises for building multilingual ASR systems and/or using them for rapid adaptation to a new target (desired) language. In this section, first a brief overview of several techniques used in building multilingual systems are explored followed by a brief explanation of some of the popular language adaptation techniques.

A multilingual ASR system is sometimes known as language independent system since it is versatile across multiple languages. This implies that acoustic-phonetic similarities across languages must be exploited. In [1], multilingual phone modeling was achieved using three approaches. In the first and the most obvious approach, given a set of corpora of multiple languages, language dependent phonemes were mapped to a new mapping convention such as the WORLDBET [2] that has a wide phonetic symbol coverage across multiple languages. With this, all language dependent transcriptions can be converted to the WORLDBET convention. Therefore, this represents a sematic way of handling multilingual phoneme units. All the transcriptions and speech files from different language corpora were pooled together into one single global multilingual corpus. HMM training was performed on this global corpus to form language independent acoustic models. The main disadvantage of this approach is that sometimes subtle language dependent variations might be lost during the mapping procedure. For example, monolingual phonemes for the alveolar "r" and palato-alveolar "r" sound differently but they might be represented with the same symbol in two different languages. After mapping to WORLDBET, both the phonemes are mapped to the same symbol thereby blurring the distinct language properties.

The second approach is a data-driven approach as opposed to the sematic approach described earlier. Here, the phonemes are mapped to a multilingual set using a bottom-up clustering procedure based on log-likelihood distance measure [3] between two phoneme models. The models with least distances are merged together to form a new cluster. Because the estimation of the new phone models of the merged cluster is difficult to achieve, the distance between the two clusters is computed as the maximum of all distances found by pairing a phone model in the first cluster versus another phone model in the second cluster. This "furthest-neighbor" merging heuristic was used to encourage compact clusters and was known to work well empirically. The clustering process continues until all calculated cluster distances are higher than a pre-defined distance threshold or if a specified number of clusters were formed. The disadvantage with this data-driven approach is that the phoneme models present in a single cluster lose their original phonetic symbol and use a symbol that is the best representation for the cluster. Hence, it is possible that models for the fricatives /s/ and /f/ might be members of the same cluster whose phonetic symbol may simply be denoted by /f/. Thus, /s/, by using /f/ as its identity, would lose its original semantic representation.

The third approach is a hybrid of the semantic and data driven approaches. Here, all monolingual triphone HMMs that have the same phonetic symbol for a given state (left, center, or right) are pooled together. For example, the Gaussian mixture densities of the phoneme /k/ in state 1 (left) of "cat", "cut", "kin", may be pooled together to form a pool of mixture densities modeling the phoneme /k/. Clustering is performed by taking the a weighted L1-norm of the difference of all possible pairs of mean vectors present in this pool. The motivation behind this is that performing clustering at the level of mixture densities helps retain some distinctive

language dependent properties which are otherwise lost if the clustering were to be performed at the HMM level (as in the second approach). Experiments in [1] indicate that the highest multilingual recognition of isolated words was achieved using the third approach and very little degradation was observed compared to the recognition accuracies of monolingual models.

Often there are scenarios when despite having well trained multilingual phoneme models, the target language that needs to be recognized has no data or very limited data. Recognizing a target language with zero data training data of the target language in a multilingual ASR system is known as cross-language transfer. When limited data is available from the target language, language adaptation of multilingual ASR systems can be useful. This scenario is referred to as cross-lingual recognition or cross-lingual adaptation.

One of the earlier approaches in cross-lingual recognition was to bootstrap or seed acoustic models that were not trained using the target language [4]. In the bootstrapping process, the phoneme set of the target language is mapped to the multilingual phoneme set. Using a limited amount of training data from the target language, the multilingual acoustic model was retrained with the seed model. Later, [5] showed that such a procedure outperforms models using random seeds even with very few iterations (1-3). It is quite normal to expect that larger the amount of training data of the target language better will be its recognition accuracy. The lower the phonetic dissimilarity between phonemes of the source languages and those of the target language the greater is the recognition accuracy using bootstrapped models [6].

A second approach in cross-lingual adaptation was by using polyphone decision tree specialization (PDTS) [7]. The PDTS method is especially useful for context dependent models. In the PDTS approach, the clustered multilingual polyphone decision tree is adapted to the target language by restarting the decision tree growing process according to the limited amount of training data available from the target language. For example, the non-adapted polyphone decsion tree of a multilingual model may not capture finer variations of the rhotic phoneme "r" if the target language uses several of these variations. Hence, clustering the target language phonemes using the non-adapted tree would result in poorly estimated class models. It was shown in [7] that performance gain using the PDTS method exceeds the gain achieved by using larger adaptation data. Other cross-language adaptation methods include maximum aposteriori (MAP) adaptation [6] using the multilingual acoustic models as the prior model for MAP adaptation.

Recently, in [8], a cross-dialectal Gaussian mixture model training criteria was proposed to transfer knowledge from Modern Standard Arabic to Levantine Arabic by data sharing. Furthermore, such transfer learning criteria have been successfully implemented in [9] for semi-supervised learning for phone recognition, and prosody detection. This study extends the use of such transfer learning framework for cross-lingual recognition. The rest of the paper is organized as follows. In Section II, the problem definition for training phoneme class models is stated. In Section II-A, a transfer learning algorithm using generative models is explained. This is followed by experimental results and conclusions in Sections III and IV respectively.

II. Algorithm

Let $\mathcal{X}^{(l)}$ comprise of a sequence of tokens generated from a language with language identity l. Hence, $\mathcal{X}^{(l)} = \left\{\mathbf{x}_1^{(l)}, \mathbf{x}_2^{(l)}, \ldots, \mathbf{x}_{N^{(l)}}^{(l)}\right\}$ where the n^{th} token is the set of features vectors from time $t = 1$ to $t = T$ and is given by $\mathbf{x}_n^{(l)} = \left\{\mathbf{x}_{n,1}^{(l)}, \mathbf{x}_{n,2}^{(l)}, \ldots, \mathbf{x}_{n,T}^{(l)}\right\}$ such that $\mathbf{x}_{n,t}^{(l)} \in \mathcal{R}^D$. Corresponding to $\mathcal{X}^{(l)}$, there are labels in $\mathcal{Y}^{(l)} = \left\{y_n^{(l)}\right\}$ where $y_n^{(l)} \in \{1, 2, \ldots, C\}$ where C is the total number of phoneme classes in language l. Let $l \in \{1, 2\}$ where $l = 1$ is the language identity for target language and $l = 2$ is the language identity of all the other source languages. The target language is the language whose models are to be estimated. The set of source languages represent all the other languages whose data is shared with the target language in the HMM model estimation process. The set of HMM models θ is given by $\{\theta_c\}_{c=1}^{C}$. Furthermore, the parameters in each class HMM are given by $\theta_c = \{\pi_c, \mathbf{A}_c, \omega_c, \mu_c, \mathbf{\Sigma}_c\}$ corresponding to the initial state and transition probabilities, mixture weights, means, and covariance matrices associated with N states and M mixture Gaussian mixture models (GMMs). From this point onward, the subscript c has been dropped from the set of class specific parameters for ease of notation. Any reference made to θ_c will imply that the parameters in the context of the discussion are specific to class c.

A. ML Based Transfer Learning

The objective is to learn the parameters θ_c of target language 1 by using *all* available training data from the low resourced target language and selecting only *relevant* data from other richly resourced languages. This is the case of *instance based inductive transfer learning* approach. In inductive transfer learning, a few labeled data in the target domain are required as the training data inducing an objective function of the target data to be optimized. The term instance based learning comes from the fact that there are certain parts or instances of source data that can be reused together with the target data.

Usually, to learn the parameters of a HMM, the objective function to be maximized is the log-likelihood function of the training data. In this work, since the training data consists of both the target and source languages we regularize the likelihood function of the target data with a regularization term involving the likelihood of the source data. Hence, the new objective function is called as the ML-ML criterion and is given by,

$$\mathcal{J}(\theta_c) = \mathcal{L}(\mathcal{X}^{(1)}|\theta_c) + \rho\mathcal{L}(\mathcal{X}^{(2)}|\theta_c),\ c = 1, \ldots, C \qquad (1)$$

where,

$$\mathcal{L}(\mathcal{X}^{(l)}; \theta_c) = \sum_n \log\ p(\mathbf{x}_n^{(1)}; \theta_c), \quad l = 1, 2 \qquad (2)$$

and ρ is a constant such that $\rho < 1$. The optimal parameter set is given by,

$$\theta_c^\star = \arg\max_{\theta_c} \mathcal{J}(\theta_c)$$

The corresponding auxiliary function for the new objective function becomes,

$$\mathcal{J}'(\theta_c, \theta_c^0) = \frac{1}{N^{(1)}} \sum_{n,\mathbf{Q}\in\mathcal{Q}} p(\mathbf{X}_n^{(1)}, \mathbf{Q}; \theta_c^0) \log\, p(\mathbf{X}_n^{(1)}, \mathbf{Q}; \theta_c) + \frac{\rho}{N^{(2)}} \sum_{n',\mathbf{Q}\in\mathcal{Q}} p(\mathbf{X}_{n'}^{(2)}, \mathbf{Q}; \theta_c^0) \log\, p(\mathbf{X}_{n'}^{(2)}, \mathbf{Q}; \theta_c). \quad (3)$$

where the summation is taken over all possible state sequences (per token) and all tokens from both the languages. Given an initial model θ_c^0, the maximum likelihood (ML) parameters, under the constraints $\sum_{m=1}^{M} \omega_{jm} = 1$ and $\mathbf{\Sigma}_{jm} \succ 0$, are found using the Expectation-Maximization (EM) algorithm as,

$$\pi_i = \frac{\frac{1}{N^{(1)}} \sum_n \alpha_{n,1}^{(1)}(i)\beta_{n',1}^{(1)}(i) + \frac{\rho}{N^{(2)}} \sum_{n'} \alpha_{n',1}^{(2)}(i)\beta_{n',1}^{(2)}(i)}{\frac{1}{N^{(1)}} + \rho\frac{1}{N^{(2)}}}, \quad (4)$$

$$a_{ij} = \frac{\frac{1}{N^{(1)}} \sum_{n,t} \xi_{n,t}^{(1)}(i,j) + \frac{\rho}{N^{(2)}} \sum_{n',t'} \xi_{n',t'}^{(2)}(i,j)}{\frac{1}{N^{(1)}} \sum_{n,t} \gamma_{n,t}^{(1)}(i) + \frac{\rho}{N^{(2)}} \sum_{n',t'} \gamma_{n',t'}^{(2)}(i)}, \quad (5)$$

$$\omega_{j,m} = \frac{\frac{1}{N^{(1)}} n_{jm}^{(1)}(1) + \frac{\rho}{N^{(2)}} n_{jm}^{(2)}(1)}{\frac{1}{N^{(1)}} \sum_m n_{jm}^{(1)}(1) + \frac{\rho}{N^{(2)}} \sum_m n_{jm}^{(2)}(1)}, \quad (6)$$

$$\mu_{jm} = \frac{\frac{1}{N^{(1)}} n_{jm}^{(1)}(\mathbf{X}) + \frac{\rho}{N^{(2)}} n_{jm}^{(2)}(\mathbf{X})}{\frac{1}{N^{(1)}} n_{jm}^{(1)}(1) + \frac{\rho}{N^{(2)}} n_{jm}^{(2)}(1)}, \quad (7)$$

$$\mathbf{\Sigma}_{jm} = \frac{\frac{1}{N^{(1)}} n_{jm}^{(1)}(\mathbf{X}^2) + \frac{\rho}{N^{(2)}} n_{jm}^{(2)}(\mathbf{X}^2)}{\frac{1}{N^{(1)}} n_{jm}^{(1)}(1) + \frac{\rho}{N^{(2)}} n_{jm}^{(2)}(1)}, \quad (8)$$

where,

$$n_{jm}^{(l)}(1) = \sum_{n,t} \gamma_{n,t}^{(l)}(j,m), \quad (9)$$

$$n_{jm}^{(l)}(\mathbf{X}) = \sum_{n,t} \gamma_{n,t}^{(l)}(j,m)\mathbf{x}_{n,t}^{(l)}, \quad (10)$$

$$n_{jm}^{(l)}(\mathbf{X}^2) = \sum_{n,t} \gamma_{n,t}^{(l)}(j,m)\Delta_{n,t}^{(l)}(j,m){\Delta^{(l)}}^T_{n,t}(j,m), \quad (11)$$

$$\Delta_{n,t}^{(l)}(j,m) = \mathbf{x}_{n,t}^{(l)} - \mu_{jm}. \quad (12)$$

The probabilities $\alpha_{n,t}^{(l)}(i)\beta_{n,t}^{(l)}(i), \gamma_{n,t}^{(l)}(j,m), \xi_{n,t}^{(l)}(i,j)$ are as given in [10, eq.(18, 23, 26, 36)].

Ignoring the superscript in parenthesis for the language identity momentarily, we represent the conditional distribution $p(\mathbf{x}_i^{(1)}|y_i = j; \theta_j)$ as $p(\mathbf{x}_i|y_i; \theta)$. There are three inherent problems with the estimation of the conditional distribution $p(\mathbf{x}_i|y_i; \theta)$. Firstly, the choice of the distribution for real world problems is mostly governed by how well it is mathematical tractable rather than how well it fits the real world data. Even though a GMM can model arbitrary distributions, ambiguities still remain in its prototype design. For example, there exists no well defined procedure to determine the optimal choice of the number of mixtures or the type of covariance matrix (diagonal, full) to be used. Secondly, the estimation method may not produce consistent estimated parameters. Finally, if the amount of training data is limited the quality of the estimated parameters cannot be guaranteed to be reliable. The third point is the most relevant for future work.

TABLE I
TURKISH AND ENGLISH PHONE SET

Language	Vowels		Consonants		Total
	Monopthongs	Dipthongs	Non-Syllabics	Syllabics	
Turkish	10	0	28	0	38
English	13	5	27	3	48
Common	4	0	20	0	24

III. EXPERIMENTS AND RESULTS

The transfer learned HMM parameters using the ML-ML criterion in (1) was applied to build a Turkish recognition system using English utterances. Modern standart Turkish almost has a one-to-one mapping between written text and its pronunciation [11]. The Turkish corpus in [11] was used. Its training set consists of a total of 3976 utterances spoken across 100 speakers. Its test set consists of 752 utterances spoken acorss 19 speakers. For English, the TIMIT training and test set consists of 3696 and 192 utterances respectively. The Turkish corpus follows a set of METUbet phonetic alphabets [11]. Since the phonetic alphabet systems are different for Turkish and TIMIT, it is important for both the alphabet systems to be mapped to a single alphabet system prior to building speech recognizers. In this study, the WORLDBET [2] system was used since its alphabets cover a wdie range of multilingual phones and it is represented in the ASCII format. In Table I, the phone sets specific to Turkish and English are given in addition to the phones that are common across both languages. It was observed that by using English as the source language for a Turkish recognizer, about 24/38 (63%) monophone coverage in Turkish is achieved. In scenarios where there is very limited or no data in Turkish, the common phones can be sourced from the English corpus.

A 3 state, 1 mixture left to right HMM was modeled for each Turkish monophone using all of the available utterances in TIMIT and a few (20/50/100/200/500) utterances from the Turkish train set. For each training set, several ρ values (0.01, 0.02, 0.2, 0.4, 0.6, 0.8) were used. The trained HMM models were tested using the Turkish test set of 752 utterances which consists of about 33750 monophones. In Fig. 1, the monophone recognition accuracies are plotted for different training sets and for each training set different values of ρ were used. The case where $\rho = 0$ is the scenario where no English data was used to train the Turkish recognizer. Starting with this case, it is normal to expect that the least accuracy (31.30%) is observed when the least number of Turkish utterances (20 utterances) were used. However, the accuracies do not improve considerably beyond 100 utterances. It was also noted that if the entire (3976 utterances) Turkish train set was used, the improvement in recognition accuracy is only 0.2%. This indicates that modeling the emissions using 1

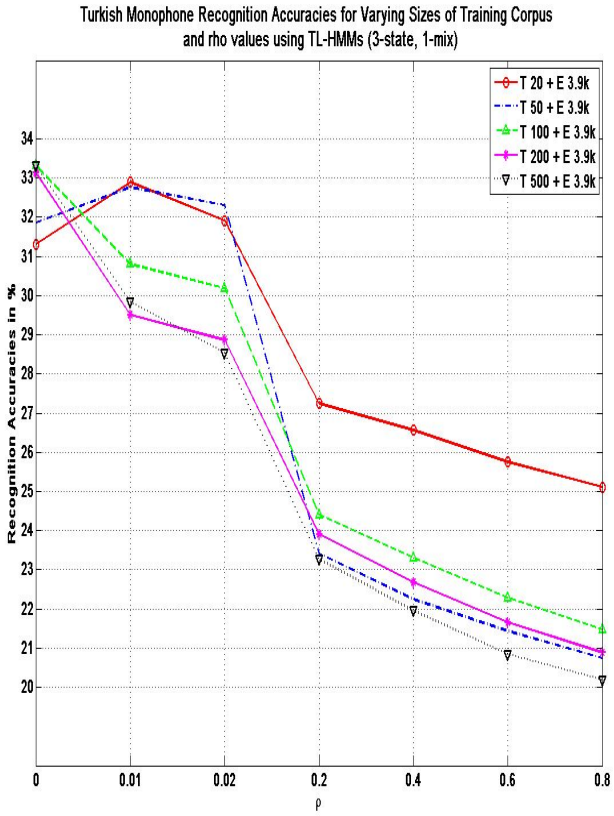

Fig. 1. Turkish monophone recognition accuracies using Transfer Learned (TL) HMM for varying sizes of Turkish training set (20/50/100/200/500) trained with a fixed size of English training set (3.9k) for different values of $0 \leq \rho \leq 0.8$

Gaussian mixture per state is not complex enough to discover the finer structural variations in the underlying distribution. Now, keeping the training size fixed to 20 Turkish utterances and varying ρ, it was observed that when $\rho = 0.01$ the transfer learned HMMs are able to achieve recognition accuracies (32.76-32.89%) close to those (33.1 - 33.32%) achieved using 100-500 utterances. Therefore, the transfer learned HMM are able to boost the recognition accuracies using only one-fifth (20 utterances) of the minimum training size (100 utterances) required to achieve close to maximum levels recognition accuracy (about 33.3%). On further increasing ρ, the recognition accuracies decreased indicating that the recognizer was learning the English language rather than Turkish. This trend was observed across all training sizes of Turkish training set. It is quite counterintuitive that at higher values of ρ, i.e. $\rho \geq 0.2$, the degradation in recognition accuracy tends to be higher if more number of Turkish utterances are used. For example, at $\rho = 0.4$, the recognition accuracy for the training set with most number (500) Turkish utterances is least whereas for the training set with least number of utterances (20) the recognition accuracy is highest.

IV. Conlcusions and Future Work

Since modeling monopones using a single mixture Gaussian is too simplistic, the next goal is to extend the modeling to multiple mixture Gaussian mixture models. Furthermore, modeling the phones in the triphone context is usually known to boost recognition accuracies considerably. By increasing modeling complexity or extending the models from context independent to context dependent ones, it is natural to expect that more Turkish data would be required. Under the assumption that limited Turkish data is available, the role of transfer learning becomes even more important. Also, instead of monophone recognition accuracies it would be more meaningful to evaluate the word recognition accuracies. Another question that needs to be addressed is the effect of recognition when there is no Turkish training data. To conclude, the most important result from this study is that one could achieve the desired monophone recognition accuracy (3 state, 1 mixture HMM) by using only one-fifths of the minimum training size from the Turkish training set.

References

[1] J. Kohler, "Multilingual phone models for vocabulary-independent speech recognition tasks," *Speech Communication.*, vol. 35, no. 1-2, pp. 21–30, Aug. 2001.

[2] J. L. Hieronymus, "Ascii phonetic symbols for the world's languages: Worldbet," Bell Labs Technical Memorandum, Tech. Rep.

[3] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden markov models," AT&T Technical Journal, Tech. Rep. 2.

[4] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid hmm development in a new language," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*

[5] T. Schultz and A. Waibel, "Fast bootstrapping of lvcsr systems with multilingual phoneme sets," in *Eurospeech.*, 1997.

[6] J. Kohler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, vol. 1, pp. 417–420.

[7] T. Schultz and A. Waibel, "Polyphone decision tree specialization for language adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 3, pp. 1707–1710.

[8] P. Huang and M. Hasegawa-Johnson, "Cross-dialectal data transferring for gaussian mixture model training in arabic speech recognition," *4th International Conference on Arabic Language Processing*, pp. 119–123, 2012.

[9] J.-T. Huang, "Semi-supervised learning for acoustic and prosodic modeling in speech applications," Ph.D. dissertation, University of Illinois at Urbana-Champaign.

[10] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] Özgül Salor and M. Demirekler, "On developing new text and audio corpora and speech recognition tools for the turkish language," in *International Conf. Spoken Language Processing*, 2002, pp. 349–352.