UNSUPERVISED ADAPTIVE SPEECH TECHNOLOGY FOR LIMITED RESOURCE LANGUAGES: A CASE STUDY FOR TAMIL

Özgür Çetin[*] *Madelaine Plauché*[†] *Udhaykumar Nallasamy*[‡]

[*]Yahoo!, Inc.,Santa Clara, CA, USA
[†]International Computer Science Institute, Berkeley, CA, USA
[‡]Carnegie Mellon University, Pittsburgh, PA, USA
ocetin@yahoo-inc.com

ABSTRACT

This paper evaluates adaptive speech technology for creating low cost, rapidly deployable speech recognizers for new languages with very limited data. A multi-modal (speech and touch) dialog system in Tamil, which delivered agricultural information to rural villagers, is described. Based on the field recordings from this system, a number of automatic speech recognition (ASR) adaptation techniques are compared, including cross-language transfer (English to Tamil), multilingual training, bootstrapping, and model adaptation (supervised and unsupervised). For this small-vocabulary task, supervised model adaptation using a small amount of target speech data yields the best results. In the supervised mode, we find no significant performance difference between adapting models from English, and models from Tamil that used a medium-sized data set at a significant labeling cost. Unsupervised adaptation from English yields slightly inferior but comparable recognition results. In summary, we find that model adaptation from a language with existing resources, using a very small amount of target data is a viable option for rapidly building small-vocabulary speech recognizers.

Index Terms—
Speech recognition, unsupervised learning.

1. INTRODUCTION

Speech research has mainly focused on a handful of languages, for which the cost of creating large annotated corpora, is carried either by wealthy government or corporate budgets. The data are collected invariably from human subjects with considerate levels of text and even computer literacy. However, speech technologies can also be attractive channels for information dissemination in developing regions, as they require minimal infrastructure (e.g., a phone), do not require literacy, and can run on any language, even those without a written form [1]. However, there is typically very little, if any, data available for such languages. This paper looks at how existing ASR design and techniques might immediately provide more equitable access to information, especially along the criteria of literacy and local language.

We evaluate standard ASR training and adaptation methods that utilize existing language resources in one language to develop an ASR system for another language. The evaluation is performed on the speech of rural villagers of Tamil Nadu, as they used a spoken dialog system that delivered recommended agricultural techniques in a form accessible to a primarily oral population. In previous work [2], we demonstrated that the recognition of a small vocabulary of command words is sufficient to power effective speech interfaces for such a community. However, the recognition task remains challenging due to outdoor recording conditions, substantial dialectal variation across geographic and social space, and unfamiliarity of the speakers with speech interfaces.

In previous work, among other methods, bootstrapping [3, 4], multilingual training [4, 5], and model adaptation in [6, 7, 8], were found to be effective for cross-language learning. For adaptation, maximum likelihood linear regression (MLLR) was comparable to maximum a posteriori (MAP) estimation with small data sets, while the latter being better with larger data sets [8]. The previous work also found less, or no benefit at all, from these methods (as compared to direct model training), with sizable target training data [6, 7, 9].

This paper makes a number of contributions to the literature on cross-language transfer learning and adaptation. First, while the methods used are similar, they are applied to the speech from people who cannot read and write, and who are unfamiliar with speech technology. In addition, Tamil is studied as the target language, for which there has been little previous work about ASR [10], or cross-language learning [11]. Second, a small-vocabulary task is studied here, as opposed to a medium- or large-vocabulary one. Less powerful methods might be sufficient for small tasks [3]. In addition, the utility of labeled data is not obvious in such contexts. It could be that a small amount of transcribed data is enough for building a high-accuracy system. Alternatively, unsupervised methods could work very well [12], because of the generally high recognition accuracies. Third, we therefore include unsupervised model adaptation in our comparisons, using both models from another language and models from a different dialect of the same language.

2. ASR ADAPTATION METHODS

We use the term ASR adaptation to refer to not just model adaptation using MLLR or MAP estimation, but any type of transfer of knowledge from one source to another. We focus on automatic methods, which reduce the cost of human labor, instead of knowledge-based methods such as multilingual phones, or articulatory features, e.g., [7, 13, 4, 14]. These methods are presented within a cross-language setup, but they are equally applicable in other mismatched data situations, which we also explore in our experiments.

The simplest form of transfer learning between languages is *cross-language transfer*, where an acoustic model trained in one language is directly applied to another language, without any modification. In *multilingual training*, a common acoustic model for many languages is built using data from those languages. Both cross-language transfer and multilingual training require some form of a common encoding method for phones of different languages. Linguistically similar languages offer the best results [4, 7]. In *bootstrapping*, the target acoustic model training is initialized using seed models from another language. The models are then iteratively rebuilt, using the native data. In *model adaptation*, the initial acoustic models are transformed using a small set of adjustable parameters to reduce mismatch with the target domain. We use MLLR [15] for this purpose, which in previous work was found to be comparable to the adaptation methods based on MAP estimations, in cases with limited adaptation data [8]. The adaptation could be done in a supervised, or an unsupervised manner, depending on whether the transcripts, or the ASR hypotheses are used for estimating MLLR transforms. Unsupervised adaptation minimizes the human participation, but its effectiveness depends on the accuracy of the initial models used for generating the ASR hypotheses [12].

3. A SPOKEN DIALOG SYSTEM IN TAMIL

In this section, we describe a spoken dialog system, providing Banana crop information, which we used to collect speech data from speakers of Tamil who have no formal schooling, and might therefore benefit from a speech interface as a means to access local, relevant information. We describe our more direct means of data collection in this section, as well.

3.1. System Design

The Banana Crop application is a simple, template spoken dialog system, built by converting a portion of M. S. Swaminathan Research Foundation's (MSSRF's, a nonprofit network of village centers in India) resource website into a command menu. Its 27 command words correspond to the original website subheadings (e.g., *banana varieties* and *soil preparation*), and were one to three words in length. The menu system was only three levels deep, and presented no more than eight options at a time. The dialog output consisted of pre-recorded Tamil speech. The system did not assume literacy, or previous computer experience. With no prior training, Tamil speakers could quickly learn to operate the system using their voice, or a hand-made touch screen. See [2] for more details.

Data set	Vocab. (words)	Size (words)	Description
TAMIL-05 (Tamil)	43	10245	Digits and verbs read or guessed out loud, indoors and out in three districts
OFFICE (Tamil)	20	174	Agricultural words read out loud by NGO staff in a fairly quiet office
FIELD (Tamil)	27	377	More agricultural words spoken by villagers indoors and out in one district (Dindigul)
TIMIT (English)	6226	39834	Phonetically balanced sentences read out loud in a laboratory setting

Table 1. Tamil and English annotated corpora.

3.2. Data Collection

Once the content and command words for Banana Crop were selected, two sets of recordings were performed. First, five MSSRF staff members read these words out loud, three times each in a quiet office. These recordings, referred to as OFFICE, are used for system development in our experiments. Second, the dialog system was evaluated by rural villagers across six different sites in Dindigul district of Tamil Nadu, where approximately 50 people (roughly equal women and men) actively navigated the system using either touch, or speech input. (The literacy rate of the district is 40%. We estimate the literacy rate of our subject pool to be slightly higher.) The system employed a recognizer, trained using TAMIL-05 (cf. Section 3.3) [2]. The participant's audio commands to the system were recorded during use. Sessions were generally short, and involved very little training. We did not attempt a formal user study of the two modalities for input. The second set of recordings, referred to as FIELD, constitutes the target recognition condition, and is used for evaluation.

3.3. Other Corpora

In addition to the Banana Crop recordings, another database, referred to as TAMIL-05, is available to us from a 2005 field study [2]. This data set contains speech from 80 speakers with a range of literacy levels, collected across three districts. Native Tamil speakers in offices, schools, and fields were shown flash cards, or a number of fingers to elicit digits and verbs. This database differs from FIELD in speaking style (read vs. spoken) and in dialect (different districts). The literacy levels of its participants are comparable. TAMIL-05 provides a better match to FIELD than English TIMIT database, which we utilize for cross-language transfer learning. The data sets used in our experiments are summarized in Table 1.

	Training Data	Accuracy (OFFICE)	Accuracy (FIELD)
1	TIMIT	66.1	30.2
2	TAMIL-05	97.1	68.7
3	TAMIL-05 (Bootstrapped from TIMIT)	97.1	63.1
4	TAMIL-05 & TIMIT (Multilingual training)	85.1	54.9

Table 2. Accuracies (%) of various model training methods on OFFICE and FIELD.

	Training Data	Adaptation (on OFFICE)	Accuracy (on FIELD)
1	TIMIT	None	30.2
		Unsupervised	75.6
		Supervised	80.4
2	TAMIL-05	None	68.7
		Unsupervised	80.1
		Supervised	82.2

Table 3. Accuracies (%) of supervised and unsupervised adaptation methods. The results without adaptation are repeated from Table 2, for the ease of comparison. All adaptation results are after one iteration of MLLR, except the unsupervised adaptation from TIMIT, which used five.

4. EXPERIMENTS

In this section, we report results using various ASR adaptation methods to build a recognizer for the Tamil field data. After a brief description of the recognition system, we will first present results for the methods that train models from scratch, and then for the methods that adapt existing models.

4.1. Recognition System

All recognition systems are trained and tested using HTK [16]. 13 mel-frequency cepstral coefficients, and their first- and second-order differences are used as acoustic features. The cepstral mean subtraction is applied per utterance. The acoustic models are word-internal triphones, with a mixture of 16 diagonal-covariance Gaussian output distributions. (Monophone models were significantly inferior, for example, 73% vs. 97.1% accuracy when trained on TAMIL-05, and tested on OFFICE.) The decoding is first-pass using a simple grammar, which outputs one of the menu options of the dialog system, without loop. No language model is used. The lexicon was hand-constructed by mapping Tamil phones to English phonemes as closely as possible. After mapping, the Tamil dictionary included a 29-phone subset of 46 TIMIT phones. MLLR-based model adaptation is done at two stages, first by applying a global transformation, and then class-specific transformations. Both the MLLR transformation classes and triphone state-tying clusters are derived using data-dependent clustering methods, automatically adjusting for the different amounts of training data available in our experiments [16].

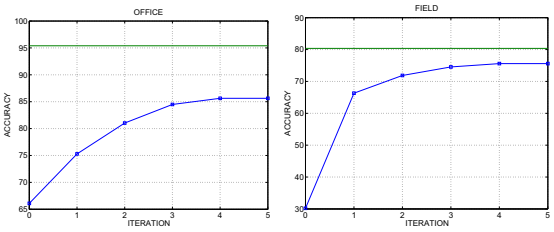
4.2. Model Training Methods

The first set of experiments compared the utility of using a significant amount of transcribed English data in various model training paradigms, cross-language transfer, multilingual training, and bootstrapping, when a significant amount of transcribed data from the target language, TAMIL-05, is available. First, separate acoustic models are trained using only TIMIT and TAMIL-05, and tested without any modification. Second, TIMIT monophone models are used for initializing the training on TAMIL-05 (bootstrapping). Third, a joint model is trained using both TIMIT and TAMIL-05 (multilingual training). The accuracies of these systems are given in Table 2 (the target task is FIELD; results on OFFICE are given for analysis purposes). All pairs of results on FIELD are statistically significant ($p < 0.011$), using McNemar's test [17].

Table 2 contains a number of interesting results. First, the accuracies on OFFICE are about 30% (absolute) higher than those on FIELD, for all systems. Because the vocabularies for these two tasks are almost identical, the degradation on FIELD is due to the speaking style (read vs. spoken), and environmental noise (office vs. outdoors). Second, the best results on both data sets are obtained with training using only the native speech data. This finding is consistent with the previous work on other pairs of languages, where a sizable amount of target language data was available, e.g., [7, 9]. The approximate phonetic mapping between English and Tamil limits the utility of the English data. Third, while the models bootstrapped from English and the models trained using only Tamil, perform identically on OFFICE, the bootstrapped models are significantly inferior on FIELD. Therefore, while model adaptation on OFFICE cannot help for variations in speaking style, dialect, and environmental noise observed in FIELD, it can ultimately improve recognition accuracy for FIELD, by emphasizing the triphones of the Banana Crop vocabulary. (OFFICE is not large enough to train a model from scratch—our attempts were unsuccessful.)

4.3. Model Adaptation Methods

The second set of experiments tested the utility of MLLR-based adaptation in supervised and unsupervised settings, starting from the English- and Tamil-trained models. Except for the unsupervised adaptation of TIMIT, the multiple iterations of MLLR did not help beyond a single iteration. The results of those experiments are given in Table 3. We make the following observations. First, the adaptation improves the accuracy of both English- and Tamil-trained models, but more for the English models. After supervised adaptation, there is no statistically significant difference between them. Therefore, there is very little gain to be had by collect-



(a) (b)

Fig. 1. (a) The accuracies (%) of TIMIT models on OFFICE, after each iteration of unsupervised adaptation on the same data set, (b) the accuracies of the same adapted models on FIELD. The straight lines correspond to the model with one iteration of supervised adaptation.

ing and annotating a corpus in a new language, like TAMIL-05, which took an estimated 100 hours of expert time, when adapting from an existing language system yields similar results, using a fraction of labeled data. Second, there is not any significant difference between supervised and unsupervised adaptation of the TAMIL-05 models, which is expected given the near-perfect recognition accuracy of those models on OFFICE. Third, while the accuracy difference between the supervised and unsupervised adaptation of the TIMIT models, is statistically significant ($p = 0.01$), the unsupervised adaptation yields comparable results to the best systems, without any transcription costs. In the unsupervised mode, the multiple iterations of adaptation are necessary, due to the very low recognition accuracy of the English models on the adaptation data (see Figures 1(a) and (b)). The accuracy significantly improves until about the fourth iteration.

5. CONCLUSIONS

In this paper, we evaluated a number of ASR adaptation methods, including cross-language transfer, multilingual training, bootstrapping, and supervised and unsupervised model adaptation, for creating low cost, rapidly deployable speech recognizer for new languages. We described Banana Crop, a small-vocabulary, agricultural spoken dialog system in Tamil. Evaluations on the field recordings from this system indicated that supervised adaptation of English- or Tamil-trained models yields the best results. Unsupervised adaptation of the English models, however, achieves comparable results, and saves the cost of hand annotation. In summary, model adaptation from a language with existing resources using a very small amount of unlabeled data is a viable option for rapidly building speech recognizers in new languages. However, methods that are robust against environmental noise and dialectal variations, are ultimately needed for further improving accuracy in spoken dialog systems, like Banana Crop, that aim to serve primarily oral communities. Better cross-language phone mappings could also improve results.

Acknowledgments The authors acknowledge ICSI for technical support, ICSI speech group, TIER group, MSSRF VRC Sempatti, Amrita University, Richard Carlson, and Srinivasan Ramaswamy. This work in part was performed while the first author was at ICSI. This material is based upon work supported by NSF (0326582) and by Elsevier. Any opinions, findings and conclusions are those of the authors and do not necessarily reflect the views of the sponsors.

6. REFERENCES

[1] E. Brewer *et al.*, "Challenges for technology transfer for developing regions," *IEEE Pervasive Computing*, vol. 5, pp. 15–23, 2006.

[2] M. Plauché, Ö. Çetin, and U. Nallasamy, "How to build a spoken dialog system with limited (or no) language resources," in *Proc. IJCAI Workshop on AI in ICT for Development*, 2006.

[3] B. Wheatley *et al.*, "An evaluation of cross-language adaptation for rapid HMM development in a new language," in *Proc. ICASSP*, 1994, pp. 237–240.

[4] T. Schultz and A. Waibel, "Language portability in acoustic modeling," in *Proc. Workshop on Multilingual Speech Communication*, 2000, pp. 59–64.

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

[6] J. Kohler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Proc. ICASSP*, 1998, pp. 417–420.

[7] W. Byrne *et al.*, "Towards laguage-independent acoustic modeling," in *Proc. ICASSP*, 2000, pp. 1029–1032.

[8] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. ICASSP*, 2003, pp. 540–543.

[9] Z. Zhao and D. O'Shaughnessy, "An evaluation of cross-language adaptation and native speech training for rapid HMM construction based on very limited training data," in *Proc. INTERSPEECH*, 2007, pp. 1433–1436.

[10] A. Lakshmi and H.A. Murphy, "A syllable based continuous speech recognizer for Tamil," in *Proc. INTERSPEECH*, 2006, pp. 1055–1058.

[11] U. Nallasamy, R. Swaminathan, and S.K. Ramakrihnan, "Multilingual speech recognition for information retrieval in Indian context," in *Proc. Student Research Workshop, HLT/NAACL*, 2004, pp. 1–6.

[12] G. Zavaliagkos *et al.*, "Using untranscribed training data to improve performance," in *Proc. ICSLP*, 1998, pp. 2551–2554.

[13] A.-K. Kienappel, D. Geller, and R. Bippus, "Cross-language transfer of multilingual phoneme models," in *Proc. ASR: Challenges for the new Millenium*, 2000, pp. 155–159.

[14] S. Stüker *et al.*, "Multilingual articulatory features," in *Proc. ICASSP*, 2003, pp. 144–147.

[15] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *CSL*, vol. 9, pp. 171–185, 1995.

[16] S. Young et al., *The HTK Book (for HTK Version 3.2)*, Cambridge University, 2002.

[17] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, 1989, pp. 532–535.