International Journal of Computational Linguistics and Applications, Vol. 6, No. 1, 2015, pp. 117–129
Received 24/02/2015, Accepted 19/05/2015, Final 17/06/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Speaker Adaptation Applied to Sinhala Speech Recognition

THILINI NADUNGODAGE
RUVAN WEERASINGHE
University of Colombo School of Computing, Sri Lanka

MAHESAN NIRANJAN
University of Southampton, Highfield, UK

ABSTRACT

Sinhala, which the main spoken language of the majority of Sri Lanka, is an under-resourced language. Sinhala language is new to the speech recognition research field and faces the problem of not having suitable speech corpora available. For a language like Sinhala, it is essential to find out ways of developing good recognition models using a fewer sample of data. Speaker Adaptive methods provides the opportunity of improving speaker independent recognition systems into more speaker dependent systems by adapting the features of the user. In this paper we are presenting an experiment we carried out by adapting a pre-trained Sinhala speech recognition system (with a single voice) with several different speaker voices. Our experiment shows that although individual adaptation systems gives the best results for corresponding speakers, we can build general speaker adaptation models to get better results than building speaker independent models using the same amount of data.

1. INTRODUCTION

Sinhala, which is one of the national languages in Sri Lanka, is an under resourced language in the field of Automatic Speech Recognition (ASR) research. We have recently started looking into Sinhala speech recognition and in the need of collecting data from the scratch since there are no previously created speech corpora are available. Collecting acoustic data is not a very difficult task and can be done by placing a recorder where people are talking. However, the hard part is that getting the corresponding text transcriptions. This is a very time consuming and a very tedious task so sometimes it is not very practical to transcribe a whole set of recorded data.

For several years Speaker Adaptation has been used to match the differences between the trained model features and the input data features. Speaker Adaptation techniques are mostly used to convert speaker independent ASR systems in to more speaker dependent ones. For adaptation one does not want large number of data samples as training a ASR model. Few utterances from a user is enough to get the system to respond to that user's voice satisfactorily.

Since speaker adaptation requires a fewer sample of new data, we thought of applying speaker adaptation for Sinhala language speech recognition as it is very hard to collect a large corpus of Sinhala speech data from scratch. In this paper we present how we used speaker adaptation in Sinhala speech recognition using a small set of data available.

Rest of the paper is as follows: Section 2 gives a brief description about Sinhala language. Section 3 reviews about speaker adaptation and different techniques that are used for it. In section 4 and 5 we presents our experiments, evaluation and results. Section 6 compares the speaker adaptation model with speaker independent model. Finally we conclude our paper in section 7 and future works in section 8.

2. SINHALA LANGUAGE

Sinhala is one of the official languages of Sri Lanka and the mother tongue of the majority (about 74%) of its population. Sinhala belongs to the Indu-Aryan language family. Sinhala language words can be divided into three main categories as *Nishapanna* (words that are of local origin), *Thadbhava* (Words borrowed from other languages in their near original form) and

Thathsama (Words derived from other languages but modified to be incorporated to Sinhala – mainly from Sanskrit and Pali). There is a high impact from Sanskrit in Sinhala given the fact that they are in the same language family. Pali has also a significant impact on Sinhala vocabulary. Tamil, Portuguese, Dutch and English have also impacted the structure and vocabulary of Sinhala due to various cultural, historical factors [1].

Spoken Sinhala contains 40 segmental phonemes; 14 vowels and 26 consonants. There are two nasalized vowels occurring in two or three words in Sinhala. Spoken Sinhala also has following several Diphthongs. Sinhala characters are written left to right in horizontal lines. Words are delimited by a space in general. Vowels have corresponding full character forms when they appear in an absolute initial position of a word. In other positions, they appear as strokes and, are used with consonants to denote vowel modifiers [2].

3. SPEAKER ADAPTATION

Speaker Adaptive model is an approach to obtain results which are nearly same as speaker dependent models without requiring a large amount of speaker specific data. In this process a trained model is tuned for a new speaker with relatively a few speech samples extracted from respective speaker. Speaker adaptation models have shown considerable improvement recognition over speaker independent models [3]. Speaker adaptation has become the modern interest in speech recognition because of its low cost approach. Speaker adaptation can be supervised or unsupervised, static or dynamic. In supervised adaptation, speech transcriptions are available and in unsupervised, it is not. In static adaptation, adaptation data is available prior to adaptation but in dynamic adaptation, data is incrementally available [4].

There are several statistical methods that are used for speaker adaptation. Some of them are described here.

- *Speaker normalization*
  Normalize the acoustic data to reduce mismatch with the acoustic models. One approach is to normalize the vocal tract length. Human vocal tract length varies according to their age, size, gender, etc. Frequency of human speech is inversely proportional to vocal tract length.
- *Maximum a posteriori (MAP) adaptation*
  Use the SI models as a prior probability distribution over model parameters when estimating using speaker-specific data. In this adaptation method it is required to have a well trained model to be adapted with new data. Also this requires a large amount of adaptation data since, it deals with separate phonemes.
- *Maximum likelihood linear regression (MLLR) adaptation*
  MLLR uses linear transformation of Gaussian model parameters to adapt to a given speaker. MLLR adaptation updates the mean vectors and covariance matrices using the new adaptation data.

There are several other methods such as Mixture Models which combines MAP and MLLR adaptation methods, Cluster Adaptive training, Eigenvoices, etc [5].

3.1. *Related work*

In literature there are lots of examples in applying speaker adaptation methods from better speech recognition. [6] and [3] presents MLLR speaker adaptation techniques and experiments. [7] presents a study of speaker adaptation techniques applied to hybrid HMM-ANN systems.

A acoustic-phonetic based speaker adaptation method based on decomposition of spectral variation sources is described in [8]. [9] presents a method for unsupervised instantaneous speaker adaptation by modeling the speaker variation in a continuous speech recognition system. A speaker adaptation system for limited data based on regression-trees is described in [10]. [11] and [12] presents how to use speaker adaptation methods for accent adaptation.

4. EXPERIMENT

Our experiment was carried out using a pre-built Sinhala speech recognition model which is considered as a base-line ASR model for Sinhala language. This base-line ASR model was trained using utterances from a single female speaker. The training data set was consisted with 3000 utterances (31,625 words), which were read speech of sentences extracted from the UCSC Sinhala Text Corpus [13]. The lexicon consisted with 4K unique words. The model was trained using the Hidden Markov Model Tool Kit (HTK) developed by the Cambridge University, UK [14]. The process of building this model is described in [15].

For speaker adaptation, we collected recorded speech from 5 female voices and 4 male voices. Each speaker read out 25 utterances. From these we used 10 utterances for adaptation and 15 remaining utterances for evaluation.

We have used HTK Toolkit's speaker adaptation tool and Maximum likelihood linear regression (MLLR) adaptation technique for these experiments.

We carried out the speaker adaptation experiment in 3 different ways to see how it performs with different sets of adaptation data. Following are the three methods we tried:

- Speaker Adaptation for individual speakers
- Speaker Adaptation for female voices / Speaker Adaptation for male voices
- Speaker Adaptation for both female and male voices

4.1. *Speaker adaptation for individual speakers*

In this experiment we did speaker adaptation in the general way which is to adapt the initial model with utterances from each speaker separately and built speaker dependent recognition models for each speaker.

4.2. *Speaker adaptation for female voices / speaker adaptation for male voices*

In this experiment we built two adaptation models separately for male voices and female voices. For the female adaptation model we used adaptation data from only three speakers and for the male adaptation model we used adaptation data from only two speakers. Hence, in this experiment we have been able to evaluate the built models with previously unseen voices.

4.3. *Speaker adaptation for both female and male voices*

For this experiment we built only one adaptation model using both female (from 3 speakers) and male (from 2 speakers) adaptation data. As in previous experiment we were able to evaluate this model with both seen and unseen voices.

5. EVALUATION AND RESULTS

For these experiments we have not included out of vocabulary words in our adaptation data or evaluation data. Hence, all the words in the test data set were previously seen words.

Before doing the evaluation of adaptation experiments, we have evaluated the initial model (trained using single female speaker voice) using voices from different female and male speakers. Table 1 shows the word level accuracy values we obtained from this.

Table 1. *Evaluation of the Initial model with different male and female speakers*

Speaker	Word Accuracy
Female 1	36.69%
Female 2	26.43%
Female 3	3.60%
Female 4	20.14%
Female 5	0.80%
Male 1	0.00%
Male 2	0.00%
Male 3	1.52%
Male 4	0.00%

We can see that although the initial model can recognize other female voices (different from the trained voice) to some extent, it has failed in recognizing male voices.

5.1. *Speaker adaptation for individual speakers: Evaluation*
For this experiment we have used 15 utterances from each speaker as test data. Table 2 shows the recognition accuracy with respect to each voice (Each adaptation model we built were evaluated using the corresponding voice). We have compared the word level accuracy values with adaptation and without adaptation.

Table 2. *Evaluation of the individual adaptation models with corresponding male and female speakers*

Speaker	**Word Accuracy**	
	Without Adaptation	**With Adaptation**
Female 1	36.69%	75.54%
Female 2	26.43%	58.57%
Female 3	3.60%	29.50%
Female 4	20.14%	72.66%
Female 5	0.80%	72.66%
Male 1	0.00%	66.91%
Male 2	0.00%	39.20%
Male 3	1.52%	46.04%
Male 4	0.00%	33.81%

Table 2 clearly shows how the accuracy of recognizing each speaker's voice is increased although the adaptation was done using a very small sets of data as 10 utterances. We can see that most of the female voices perform with high accuracy with adaptation. Even the male voices (which are very different from the initial model’s training voice) shows a significant accuracy increasing after the adaptation.

5.2. *Speaker adaptation for female voices / speaker adaptation for male voices: Evaluation*
Here we have evaluated the two separate models (Female adaptation model and Male adaptation model) with two sets of data. One with voices from the speakers we have used to adapt the models (previously seen data) and one with voices from new speakers (previously unseen data).

Table 3. *Evaluation of the male and female adaptation models with seen and unseen male/female speakers*

Test Set	**Word Accuracy**	
	Without Adaptation	**With Adaptation**
Female voices (seen)	22.3%	40.53%
Female voices (unseen)	10.98%	23.02%
Male voices (seen)	0.00%	54.37%
Male voices (unseen)	0.81%	12.23%

In Table 3 we can see that by building separate common models for male and female voices also we can obtain an increased accuracy. Although this performance may not be good as the performance of individual adaptation models, we can see that by this method even for new (unseen) speakers we can get a good recognition accuracy compared to the initial model.

5.3. *Speaker adaptation for both female and male voices: Evaluation*

To evaluate this general adaptation model we have used the same test sets we used in the previous experiment. In Table 4 we can see that even by building a general adaptation model for both male and female voices, we can obtain an increase in the accuracy than the initial model.

Table 4. *Evaluation of the general adaptation model with seen and unseen male/female speakers*

Test Set	**Word Accuracy**	
	Without Adaptation	**With Adaptation**
Female voices (seen)	22.3%	26.14%
Female voices (unseen)	10.98%	15.11%
Male voices (seen)	0.00%	19.01%
Male voices (unseen)	0.81%	7.55%

6. Speaker Adaptation vs. Training Speaker Independent Models

In the previous sections we have described the results we got from various types of speaker adaptation models. In this section we thought of comparing these results with results we can obtain by training a speaker independent model using the adaptation data we used for speaker adaptation.

Our initial training set was consisted with 3000 utterances from one female speaker. To this initial training set we added the data we used for speaker adaptation (90 utterances from 5 females and 4 males: 10 utterances each) to create a new training data set. Using this new train data we trained a new acoustic model which is no longer dependent on one speaker. We evaluated this model using the same test sets we have used for speaker adaptation evaluation.

Table 5. *Comparison of the performance of general adaptation model and speaker independent model*

Test Set	**Word Accuracy**	
	Speaker Independent Model	**General SA Model**
Female voices (seen)	16.07%	26.14%
Female voices (unseen)	7.91%	15.11%
Male voices (seen)	0.00%	19.01%
Male voices (unseen)	0.00%	7.55%

Table 5 shows a comparison of evaluation accuracy for general speaker adaptation model and speaker independent model. I clearly shows that speaker adaptation is far more better than building speaker independent systems where only a small sample of data is available.

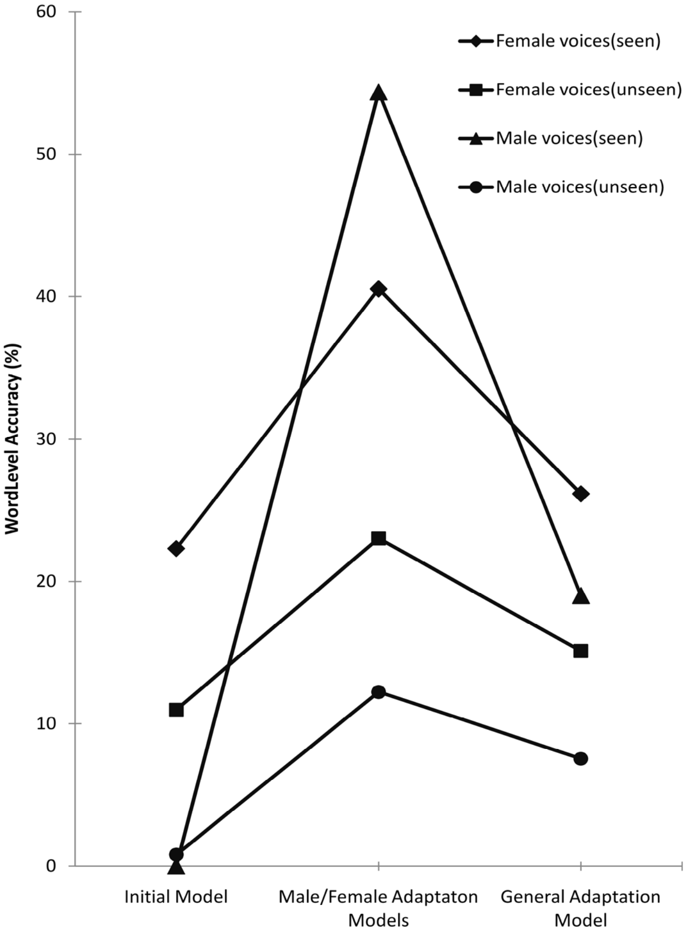

Figure 1. *Word Level performances of the Initial model, Male/Female Adaptations models and the General Adaptation model with different evaluation data sets.*

7. CONCLUSIONS

From the above experiments we can say that any level of adaptation can lead to increasing of recognition accuracy than the initial model. Figure 1 shows the summary of what we have gathered in these experiments. We can say that although individual male / female adaptation systems are best to obtain a good recognition accuracy, we can built a general adaptation system where we are able to use for new speakers without prior adaptation. Also we have shown that by building a general speaker adaptation system we can achieve better recognition accuracy than building a speaker independent recognition model where there are fewer samples of data available. We can use speaker adaptation is suitable for an under-resourced language like Sinhala, where full corpora of transcribed speech is hard to come by.

8. FUTURE WORKS

The experiments and results presented in this paper were based on a very small data set with a very few vocabulary where we did not consider out of vocabulary words. As future work we intend to improve this work by collecting more data from more speakers and by increasing the vocabulary size and also considering out of vocabulary words in the evaluation sets.

REFERENCES

1. Weerasinghe, R., Wasala, A., Gamage, K. 2005. A rule based syllabification algorithm for sinhala. In *Natural Language Processing - IJCNLP 2005* (pp. 438-449), Springer.
2. Wasala, A., Weerasinghe, R., Gamage, K. 2006. Sinhala grapheme-to-phoneme conversion and rules for schwa epenthesis. In proceedings of the *COLING/ACL on Main Conference Poster Sessions* (pp. 890-897), Association for Computational Linguistics.
3. Ganitkevitch, J. 2005. Speaker adaptation using maximum likelihood linear regression. In Rheinish-Westesche Technische Hochschule Aachen, the course of Automatic Speech Recognition, www-i6. informatik. rwthaachen. <de/web/Teaching/Seminars/ SS05/ASR/Juri Ganitkevitch Ausarbeitung.pdf>.
4. Renals, S. 2013. Speaker adaptation (automatic speech recognition asr lecture 10 &11.
5. Shinoda, K. 2011. Speaker adaptation techniques for automatic speech recognition. *Proc. APSIPA ASC 2011 Xi'an*.

6. Leggetter, C. J. & Woodland, P .C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9, 171-185.
7. Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S. & Robinson, T. 1995. Speaker-adaptation for hybrid hmm-ann continuous speech recognition system. In *Fourth European Conference on Speech Communication and Technology, EUROSPEECH* (pp. 18-21), International Speech Communication Association.
8. Zhao, Y. 1994. An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition. *Speech and Audio Processing*, 2, 380-394.
9. Strom, N. 1996. Speaker adaptation by modeling the speaker variation in a continuous speech recognition system. In proceedings of *ICSLP 96* (pp. 989-992).
10. Wang, S., Cui, X. & Alwan, A. 2007. Speaker adaptation with limited data using regression-tree-based spectral peak alignment. *Audio, Speech, and Language Processing, IEEE Transactions on 15* (pp. 2454-2464).
11. Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R. & Yoon, S. Y. 2005. Accent detection and speech recognition for shanghai-accented mandarin. In *Interspeech Citeseer* (pp. 217-220).
12. Clarke, C. M .& Garrett, M .F. 2004. Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, 116, 3647-3658.
13. Weerasinghe, R., Herath, D., Welgama, V., Medagoda, N., Wasala, A., Jayalatharachchi, E. 2007. Ucsc sinhala corpus - pan localization project-phase i.
14. Young, S. 1993. The htk hidden markov model toolkit: Design and philosophy. Technical report, Department of Engineering, Cambridge University, UK.
15. Nadungodage, T. & Weerasinghe, R. 2011. Continuous sinhala speech recognizer. In *Conference on Human Language Technology for Development* (pp. 141-147), Alexandria, Egypt.

ACKNOWLEDGMENTS

Authors of this paper would like to acknowledge the National Research Council (NRC) of Sri Lanka for funding this research. They are also indebted to the research team members of the

Language Technology Research Laboratory of the University of Colombo School of Computing, Sri Lanka, for assisting in numerous ways.

THILINI NADUNGODAGE
LANGUAGE TECHNOLOGY RESEARCH LABORATORY,
UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING,
SRI LANKA.
E-MAIL: <FHND@UCSC.LK>

RUVAN WEERASINGHE
LANGUAGE TECHNOLOGY RESEARCH LABORATORY,
UNIVERSITY OF COLOMBO SCHOOL OF COMPUTING, SRI LANKA
E-MAIL: <ARWG@UCSC.LK>

MAHESAN NIRANJAN
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE,
UNIVERSITY OF SOUTHAMPTON, HIGHFIELD,
SOUTHAMPTON SO17 1BJ, UK.
E-MAIL: <M.NIRANJAN@SOUTHAMPTON.AC.UK>