Maximum Likelihood Linear Regression (MLLR) for ASR Severity Based Adaptation to Help Dysarthric Speakers

Bassam Ali Al-Qatab, Mumtaz Begum Mustafa, and Siti Salwah Salim
Faculty of Computer Science and Information Technology
University of Malaya
Kuala Lumpur, Malaysia
bassam@siswa.um.edu.my, mumtaz@um.edu.my, salwa@um.edu.my

***Abstract*— Automatic speech recognition (ASR) for dysarthric speakers is one of the most challenging research areas. The lack of corpus for dysarthric speakers makes it even more difficult. The speaker adaptation (SA) is an alternative solution to overcome the lack of dysarthric speech and enhance the performance of ASR. This paper introduces the Severity-based adaptation, using small amount of speech data, in which data from all participants in a given severity type will use for adaptation of that type. The adaptation is performed for two types of acoustic models, which are the Controlled Acoustic Model (CAM) developed using rich phonetic corpus, and Dysarthric Acoustic Model (DAM) that includes speech collected from dysarthric speakers suffering from variety level of severity. This paper compares two adaptation techniques for building ASR systems for dysarthric speakers, which are Maximum Likelihood Linear Regression (MLLR) and Constrained Maximum Likelihood Linear Regression (CMLLR). The result shows that the Word Recognition Accuracy (WRA) for the CAM outperformed DAM for both the Speaker Independent (SI) and Speaker Adaptation (SA). On the other hand, it was found that MLLR is outperformed the CMLLR for both Controlled Speaker Adaptation (CSA) and Dysarthric Speaker Adaptation (DSA).**

Keywords-Component, Automatic Speech Recognition, Dysarthric Speakers, Severity Adaptation, Maximum Likelihood Linear Regression, Constrained Maximum Likelihood Linear Regression.

I. INTRODUCTION

Dysarthria is a neuro-motor speech impairment, in which the muscles related to speech organs are weak, move slowly, or not move at all. The causes of dysarthria includes stroke, head injury, cerebral palsy, and muscles dystrophy [1]. The classification and severity of dysarthria relay on the affected area of the nervous system as well as the site damage and degree of neurological damage. Clinically, dysarthria is assessed according to the articulation and speech intelligibility based on human perceptual measures [2]. A common assessment tools include Computerized Assessment of Intelligibility of Dysarthric Speech (CAIDS) [3].

Due to the tiring and frustration of speaking for a long periods of time by dysarthric speaker, there is a lack of speech databases available that can provide sufficient speech samples to train a speaker dependent (SD) ASR system [4]. As such, two alternatives can be considered when developing ASR systems for people with dysarthria. The first one, which is a crude approach, is to use unimpaired (normal speech) speech acoustic model to recognize impaired (dysarthric) speech. However, the speech of a dysarthric speaker has very low speech intelligibility, causing typical measures of speech acoustics to have values in ranges that are very different from those for unimpaired speech [5]. It is unlikely then that the acoustic models trained in unimpaired speech will be able to adjust to this mismatch [6].

The speaker adaptation (SA) is an alternative solution to overcome the lack of dysarthric speech and enhance the performance of ASR. The speaker adaptation is a useful method to reducing the mismatch between the feature of the given speaker with the participants of the trained acoustic model ASR [4-7].

Earlier researches have evaluated different types of speech acoustic models and the performance of ASR systems in recognizing dysarthric speech [1, 6, 8, 9]. Sanders *et al*. [10] conducted a comparative study to evaluate the performance of SD and SI models with dysarthric speech. The SI systems were trained with unimpaired speech of 5,000 speaker corpuses (40 items per speaker), whereas the SD systems were trained with the speech of two dysarthric speakers for 8.5 and 12.8 minutes of speech respectively. The SI and SD speech acoustic models were tested using ten dysarthric speakers; each speaker uttered ten utterances. It was found that the performance of the ASR systems using the SD model was better than that of the systems using the SI model. The recognition error of the SI system was between 67–100%, whereas the recognition error of the SD system was 19–50%; this marks an improvement in recognition of 50–100% over the SI model. Similar findings were also reported in [5, 11]. However, building SD model for speech-impaired individuals is non-trivial because of difficulties in obtaining sufficient speech data for training from an individual afflicted with dysarthria, particularly the more severe level [1, 8, 10]. Mustafa *et al*. [12] conducted adaptation on two types of acoustic models from TIMIT and TORGO speech database. The experiment was designed as leave-one-out for testing the other speakers from the same level of severity. The result shows that CMLLR outperformed the MLLR. This because the CMLLR technique is used to extract features that are more specifically focused on the speaker-related speech properties rather than the standard spectral envelope [13].

The objective of this paper is to examine the use of adaptation with small amount of data and the effectiveness of using the Severity-based adaptation for obtaining high recognition accuracy. Severity-based adaptation collects the

adaptation data from all participants related to certain severity model.

This paper is organized as follows: section II describes the adaptation techniques for ASR, the experimental setup in which include the experiment design, baseline acoustic model, speech data, speech data coding, adaptation technique, as well as other preparation for the experiment is described in section III. The result of the experiment is described in section IV. Section V discusses the major findings and section VI concludes the experiment.

II. ADAPTATION TECHNIQUES IN ASR

Due to the speech variability factor that effect the speech recognition performance. There are many approaches used to enhance the performance of the ASR. The adaptation technique is one of the techniques used to enhance the performance of the ASR. The factor of speaker variability that affects the ASR performance has been addressed using the adaptation techniques [13]. The state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems use speaker-adapted model. The use of the less speaker data in speaker adaptation model comparing to that full speaker data used to train speaker dependent model is essential goal for using the speech adaptation technique [14]. In particular, in the speaker independent ASR, the use of utterances from many speakers to train the speaker independent ASR allows these models to represent both the phonetic features and speakers features. This caused the speaker independent to perform less than speaker dependent in which the target user for speaker dependent has sufficient amount of utterances to estimate the HMM parameters. In addition, speaker differences is not the only factor that decrease the accuracy of the ASR, the environmental noise and channels also contribute to decrease the recognition accuracy. As a result, the adaptation techniques are important to ASR [15].

The acoustic model adaptation is applied to overcome the problem of data sparseness which results in low accuracy of ASR. The data sparseness appeared when there are a large number of parameters in a triphone acoustic model ASR leads to less accurate in estimation the model parameter in Maximum-Likelihood (ML) estimation using Expectation Maximization (EM). Let the θ_i be the parameter set of the initial model i given the beforehand, and $\hat{\theta}$ be that of the target model to be determined. Acoustic model adaptation is defined as the process to find a mapping function f from the space of parameters of the initial model to the space of the target model using the adaptation data [15, 16].

$$\hat{\theta} = f(\theta_1, \dots\dots, \theta_n) \tag{1}$$

where n is the number of initial models provided. This mapping function f called the adaptation model. Figure 1 illustrates acoustic model adaptation.

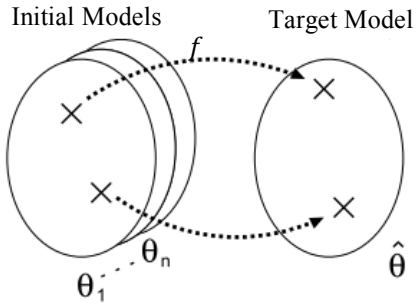

Figure 1. Acoustic model adaptation. f is a mapping function (an adaptation model), $\theta_1, \dots\dots, \theta_n$ are the parameter sets of initial models, and $\hat{\theta}$ is the parameter set of the target model.

Shinoda [15] listed the criteria need to accomplish by an adaptation model. First, is regarding to the recognition accuracy in which the adaptation model should improve it even with small amount of adaptation data. Second, the recognition accuracy of the adaptation model is approximately similar to the accuracy of initial (matched) model, as the amount of adaptation data increases.

There are three widely families used in the speaker adaptation system. The maximum a posterior (MAP) adaptation family [17]; parameter transformation based adaptation using maximum likelihood linear regression (MLLR) family [18]; the third family related to the speaker clustering methods or speaker-space method [19-21].

A. Maximum Likelihood Linear Regression

This paper is concerned in the second family of adaptation techniques which is maximum likelihood linear regression (MLLR). The MLLR is one of the linear mapping strategies between the acoustic feature spaces of many speakers as the adaptation model. It is one of the most popular adaptation techniques used in ASR [22, 23].

The mean vectors of the Gaussian distribution of the HMMs in MLLR is represented as

$$\mu = (\mu_1, \dots\dots, \mu_n)' \tag{2}$$

where n is the dimension of a feature vector, to update the mean vector in the equation (2) the following transformation is used:

$$\hat{\mu} = A\boldsymbol{\mu} + \boldsymbol{b}, \tag{3}$$

where, A is an $n \times n$ matrix, and $\boldsymbol{b}$ is a n-dimentional vector. Equation (3) can be written into the linear mapping as the following:

$$\hat{\mu} = W\boldsymbol{\xi}, \tag{4}$$

where $\boldsymbol{\xi} = (1, \mu_1, \dots\dots, \mu_n)'$. W is an $n \times (n+1)$ matrix whose first column is identical to $\boldsymbol{b}$.

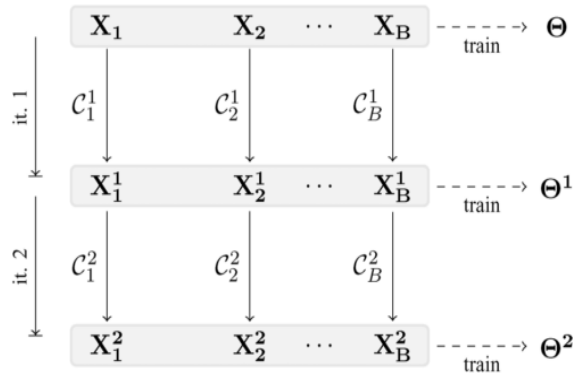

Figure 2. Block diagram of two iteration of SAT.

To obtain the ML estimation, W is calculated by using the EM algorithm. Let X is the sequence of the feature vector as the following:

$$X = \{x_1, \ldots\ldots, x_T\} \tag{5}$$

The auxiliary function can be rewrite as the following:

$$Q(W, \bar{W}) = K - \frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_m(t)[K_m + \log|\sum_m| + (x_t - W\xi)'\sum_m^{-1}(x_t - W\xi)] \tag{6}$$

where $\gamma_m(t)$ is the posterior probability of being in mixture component m at time t, K is a term independent from the output probability, and K_m is a normalization factor for mixture component m. From the equation (6) using the ML estimation to estimate the $\bar{W}$ of W as the following

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1}x_t\xi_m' = \sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1}\bar{W}\xi_m\xi_m' \tag{7}$$

When the covariance matrix for each mixture component is diagonal the equation (7) can be solved. By compensating the left-hand side of equation (7) with Z:

$$Z = \sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1}x_t\xi_m' \tag{8}$$

Furthermore, defining the matrix $G^{(i)}$ whose (j,q)-th the element g_{iq}, is

$$g_{iq} = \sum_{m=1}^{m} v_{ij}^{(m)} d_{qj}^{(m)} \tag{9}$$

where v_{ij} is the (i,j)-th element of matrix V, d_{ij} is the (i,j)-th element of matrix D, V and D are as follows:

$$V^m = \sum_{m=1}^{M}\gamma_m(t)\sum_m^{-1} \tag{10}$$

$$D^m = \xi_m\xi_m' \tag{11}$$

Using these equations, $\widetilde{W}$ is obtained as follows.

$$\widetilde{w}_i' = G^{(i)-1}z_i' \tag{12}$$

where $\widetilde{w}_i$is the i-th column vector of $\widetilde{w}_m$, and z_i is the i-th column vector Z.

B. Speaker Adaptive Training (SAT)

Speaker adaptive training (SAT) rely on a good initial model for model adaptation [24]. One way to simplify the regression model is by using diagonal or block-diagonal covariance matrices which is reducing the number of parameters in linear regression model or to share the mean and variance transforms [22]. Constrained MLLR (CMLLR) [25] is estimate parameters from the mean vector and covariance matrix to provide the estimate feature of the adaptation model. In CMLLR the parameter of the Gaussian component in the regression class transformed as:

$$\hat{\mu} = \boldsymbol{A\mu} + \boldsymbol{b} \tag{13}$$

$$\widehat{\sum} = \boldsymbol{A}\sum\boldsymbol{A^T} \tag{14}$$

where the linear transform $\boldsymbol{A}$ is used for adaptation of both $\boldsymbol{\mu}$ and $\sum$. The equation (13) and equation (14) for CMLLR is similar to that estimation equation of MLLR with the different is that the former adapt both the mean vector and the covariance matrices.

An adaptation of CMLLR can be performed in the model space if one regression class is used. One of the common used of the feature space CMLLR is the SAT [24, 25]. SAT can estimate the transformed feature space of the speaker independent that including feature distribution of a large number of speakers with the transformed feature space per speaker of CMLLR [13]. Let $\boldsymbol{B}$ is a set of speakers and $\boldsymbol{X_i}$ is a set of adaptation cepstra regarding to the given speaker in which $1 \leq i \leq B$, SAT optimizes the maximum likelihood criterion on a per speaker basis as

$$arg\max_{\boldsymbol{\theta}, C_i}\prod_{i=1}^{B} p(C_i(X_i)|\boldsymbol{\theta}) \tag{15}$$

where C_i is the individual speaker-dependent transforms, $\boldsymbol{\theta}$ is the model parameters where $\boldsymbol{\theta} = (\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_N}, \boldsymbol{\sum_1}, \ldots, \boldsymbol{\sum_N})$.

Both C_i and $\boldsymbol{\theta}$ are jointly estimated if two steps. First, the parameters of the feature space for the initial model of each training speaker C_i using CMLLR and the parameter of the speaker-independent model are estimated. Second, the transformation of the speech data of each training speaker using the feature estimation on the first step is retrain the speaker-independent model $\boldsymbol{\theta}$. EM can be used to iterated this process several times as illustrated in Figure 2 [13, 15, 16].

III. EXPERIMENTAL SETUP

The experimental design in this research is to show the effectiveness of the adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) and Constrained Maximum Likelihood Linear Regression (CMLLR) to the performance of ASR system for dysathria especially when the amount of data is small.

The experiment in this research is twofold. As the focus is on dysarthric speakers, the first part of the experiment is the training data from the dysarthric speakers (TORGO dysarthric speech corpus). The second one, which is controlled experiment will be developed using the TIMIT speech corpus.

Figure 3 shows the overall design of the experiment and components in accomplishing the experiments. Two main components are the acoustic model of impaired speech (Dysarthric Speakers) (DAM) and acoustic model of unimpaired speech (Control Speakers) (CAM) which is the baseline acoustic model for control purposes. The Dysarthric speech corpus of NEMOURS which consists of all several types of severity is used to adapt the baseline acoustic models (DAM and CAM). The MLLR and CMLLR techniques were used for speaker adaptation (SA) of DMA and CAM. The four SA experiments are Dysarthric Speaker Adaptation DSA using MLLR and CMLLR, named DSA-MLLR and DSA-CMLLR respectively, and Controlled Speaker Adaptation CSA using MLLR and CMLLR named CSA-MLLR and CSA-CMLLR respectively. The test data is taken from dysarthric severity type speech corpus. The percentage of the testing and adaptation speech samples are shown in Figure 3.

A. Speech Data

The experiment includes three types of speech data, which are used to build and test the acoustic models. These types are used according to the speaker's participant in the speech corpus. The three types are:

- **TIMIT speech corpus**: the TIMIT Acoustic-Phonetic Continuous speech corpus was developed at Texas instruments and MIT, and distributed by the US National institute of standards and technology. It represents eight major dialect division of American English with a total of 630 speakers divided into 438 male speakers and 192 female speakers [26]. We refer to this data as an unimpaired speech data or controlled speech data and it used to build a controlled acoustic model (CAM).
- **TORGO speech corpus:** consist of 15 subjects classified into dysarthric and control speakers comprise of eight and seven subjects respectively. The dysarthric speakers are our concern in this paper in which divided into five male and three female. The corpus is covering a wide range of intelligibility. The speakers were in the age between 16 and 50 years old. The participant included in the speech corpus suffering from cerebral palsy (SP) and amyotrophic lateral sclerosis (ALS). The spastic, athetoid, or ataxic are the examples of SP impairment for the participant in this corpus [7]. The TORGO speech corpus has been diagnosed by speech-language pathologist based on the Frenchay Dysarthria Assessment [27]. Based on this test, the four subjects are severely dysartheric, one subject is moderate-to-severely dysarthric, one subject is moderately dysarthric, and two subjects are very mild dysarthric[28]. We refer to this data as dysartheric speech data and used to build dysarthric acoustic model (DAM).
- **NEMOURS speech corpus**: is a collection of 814 short nonsense sentences. These sentences have been spoken by 11 male speakers. Each speaker prompted to speech 74 sentences. The sentence has the form of "The X is Ying the Z" where $X \neq Z$ [29]. The target words X, Y, and Z had the constraints to provide closed-set phonetic contrasts (e.g. place, manner and voicing contrasts) similar to those in [30]. The NEMOURS speech database dysarthric intelligibility was the main focus of adaptation and testing for this experiment. Nine speakers were chosen for adaptation and testing for this experiment. Two speakers were excluded from this experiment because some data from one speaker was missing and the other speaker was left out to obtain the balancing of the total participants for each severity type which states to three participants per severity type.

B. Speech Data Coding (Feature Extraction)

An important step in (ASR) is feature extraction. Thus, we extracted 12 Mel-frequency cepstral coefficients (MFCCs) including C0 as energy components for every 10ms analysis frame using 25-ms Hamming window, and their first and second derivatives computed to obtain a 39 dimensional feature vector. During training and testing, the cepstral mean normalization and energy normalization has been applied to feature vectors. All configuration parameters for feature vector extraction were similar for both controlled corpus and dysarthric corpus as well as for testing and adaptation corpus.

C. Building Baseline Acoustic Model

There are two types of baseline acoustic model built in this experiment, which are the acoustic model built from controlled speech data (CAM) and the acoustic model built from the dysarthric speech data (DAM).

The 3-state left to right, context-dependent, and triphone based were used to build the HMM topology and train the acoustic model using the HTK tool version 3.4.1 toolkit [31]. All the triphones models were constructed from 41 monophones in which contained a silence and a short pause models. The context-dependent triphones state of acoustic model was tied by applying the decision tree clustering to enhance the acoustic performance and share the common feature among the states. Additionally, the 16 mixture Gaussians per state was performed as an additional advantage for gain an extra acoustic performance. As a result, the number of triphones, states and utterance used to build a trained acoustic model for both controlled and impaired acoustic model are shown in Table 1.

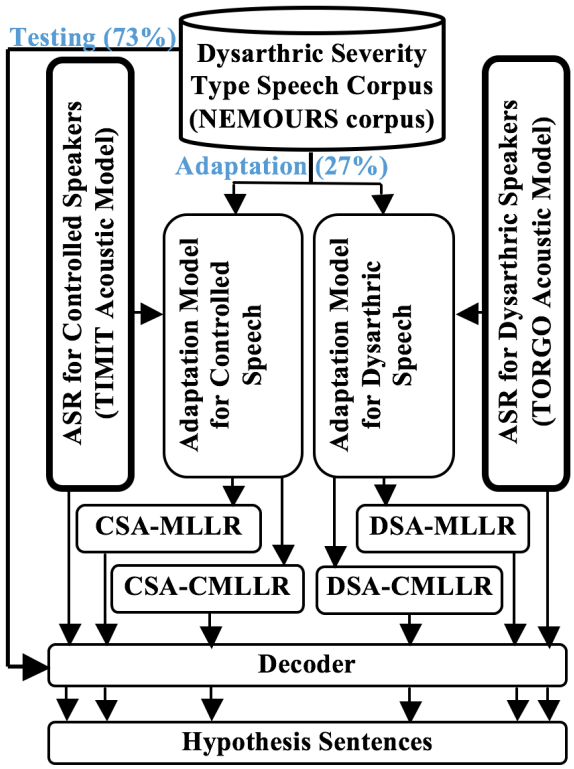

Figure 3. Severity based adaptation experiment design.

Table 1. The number of triphones, states and utterance used to train CAM and DAM.

Type of Acoustic Model	No. of Triphones	No. of States	No. of Utterance
CAM	19196	1952	6300
DAM	3352	364	3121

The tool used for performing training and testing of the experiments is HTK (version 3.4.1). The word-bigram language model was estimates from the 74 sentences provided by each dysarthric speaker. The dictionary was extracted from the total of 74 sentences used for testing with average of six words per sentence. All configuration parameters for feature vector extraction were similar for both controlled corpus and dysarthric corpus.

D. Severity-based Adaptation

In several study, speaker dependent (SD) performs better than speaker independent (SI). Building SD systems or even just region dependent systems is still impractical in most services for general public due to the requirement of amassing sufficient amount of data from each speaker or region. Speaker adaptation techniques (SA), which eliminate training and testing mismatches by adapting the speaker independent (SI) model's parameters to the new individual user's voice with a small amount of data from that user, could bridge up the gap between the SI and SD performances. The idea of the adaptation is to adjust the parameters of the acoustic model used for decoding the speech to the relevant acoustic characteristics of the target speaker [16, 32].

The MLLR [23] adaptation uses linear mapping between the acoustic feature spaces of different speakers. It is one of the most commonly used adaptation method due to its simplicity and performance. CMLLR [25] performs transformation of mean vectors and covariance matrix whereas the MLLR performs transformation only of mean vector adaptation. CMLLR is commonly used in the speaker adaptive training (SAT) which estimates a set of CMLLR transforms for each speaker and a speaker-independent model in the transformed feature space [25].

The procedure of the adaptation includes the data from all speakers of Nemours according to the severity type. The first step of adaptation in this experiment, the 20 sentences out of 74 sentences uttered by the speaker (almost 27% of sentences uttered) were selected for adaptation. Second, the total of 60 sentences per severity type has been designed for the final data set in which include 20 samples extracted from each member of the severity type for that is called the Severity-based adaptation. For example: if the mild severity model includes three participants A, B, and C. Then the adaptation set for mild severity model will include some utterance from participants A, B, and C. As a result, the severity-based adaptation set for mild severity model in our experiment will include 60 utterances (20 utterances from A, 20 utterances from B, and 20 utterances from C). In general severity-based adaptation for the severity type will include utterances from all members of specific severity type.

Two passes were involved to develop the severity-based adaptation model. The global adaptation was performed in the first pass. In the second pass, the model set has been transformed, using the global transformation as an input transform, for more estimation of specific transformation using a regression class tree with 32 terminal nodes or base classes.

E. Testing

The testing of the proposed method includes almost 73% (53 sentences out of 74 sentences) of sentences recorded for each speaker of NEMOURS. The testing set classified based on the severity type (mild, moderate, and severe). The testing was performed for the both baseline ASR and the adaptive ASR. The word recognition accuracy (WRA) is the measurement used in this experiment. Equation (16) shows the formula of WRA applied in this experiment.

$$\text{Word Recognition Accuracy (WRA)} = \frac{N-D-S-I}{N} \times 100\% \quad (16)$$

where N is the total number of labels in the reference transcriptions with correct recognition, D is the number of deletion errors, S is number of substitution errors and I is the number of insertion errors.

F. Experimental procedure

The experimental procedure comprises nine sub-experiments which related to the number of participants in testing speech corpus. Each sub experiment has the following form (**SeverityType-ParticipantID**), where the

SeverityType states the type of severity for given dysarthric speaker. The four characters code **SeverityType** describes severity type that will use for adaptation of the acoustic model. For example, the **MILD**, **MODR** and **SEVR** are the code for the mild, moderate and severe severity type for dysarthric speaker respectively. The **ParticipantID** is two character codes that represent the speaker used for testing of the acoustic model. The experiment identified by the name **MODR-RK** means the adaptation data set is the data from participants of moderate severity type which is **MORD** and the testing for this experiment is the participant **RK**.

IV. RESULTS

The results of this experiment were divided in to two parts: the results from the controlled acoustic ASR and dysarthric acoustic ASR. For both ASR, the results include the performance of the ASR for Speaker Independent (SI), speaker adaptation (SA) with MLLR, and speaker adaptation with CMLLR.

Table 2 and Table 3 show the recognition rate for CAM and DAM, respectively. For each table, the results of speaker independent (SI) and speaker adaptation (SA) systems, applied on three types of dysarthric severity namely; mild, moderate and severe. In addition, the results of each severity shows the results of SI and SA for each speaker. Two adaptation techniques, MLLR (SA-MLLR) and CMLLR (SA-CMLLR), are used for speaker adaptation systems. In addition, the systems are evaluated in terms of Word Recognition Accuracy (WRA). The ID and adaptation set in both tables represent the experiment name and adaptation data respectively.

V. DISCUSSION

When comparing the results of two baseline acoustic model shown in Table 2 and Table 3, the performance of the CAM is performing better than the DAM. The CAM trained using TIMIT speech corpus which is phonetically rich in compared to DAM.

The SA shows significant improvement in the WRA for both CAM and DAM. For example, in CAM the WRA for the severe speech types have improved from 42.76 to 51.11 when using SI and SA-MLLR respectively. Fig. 4 shows the comparative WRA for the two acoustic models CAM and DAM which applied the MLLR and CMLLR adaptation techniques.

The results from both dysarthric ASR and control ASR show that low word recognition accuracy obtained for more severe dysarthric speech is the highest. This is because the more severe dysarthric speech consists of involuntary breathing, irregular articulatory breakdowns, prosodic disruptions, stuttering, and accidental pauses which results in low word recognition accuracy [28].

In comparing this results with the work in [12], the small amount of adaptation data used in this experiment result in better performance of MLLR then CMLLR which can make use of the standard spectral envelop to detect the impaired speech with high accuracy.

Table 2. The Word Recognition Accuracy (WRA) for Controlled Acoustic Model (CAM).

ID	SI	Adaptation Set	SA-MLLR	SA-CMLLR
MILD-BB	71.85	**MILD (BB-MH-FB)**	74.55	75.76
MILD-MH	78.60		79.39	80.61
MILD-FB	72.97		76.97	75.15
MILD-All	**74.47**		**76.97**	**77.17**
MODR-RK	59.23	**MODR (RK-RL-JF)**	62.42	54.85
MODR-RL	46.17		56.36	61.82
MODR-JF	57.88		59.70	58.79
MODR -All	**54.43**		**59.49**	**58.49**
SEVR-SC	52.7	**SEVR (SC-BK-BV)**	60.30	59.70
SEVR-BK	15.99		31.82	36.06
SEVR-BV	59.68		61.21	57.27
SEVR-All	**42.79**		**51.11**	**51.01**

Table 3. The Word Recognition Accuracy (WRA) for Dysarthric Acoustic Model (DAM).

ID	SI	Adaptation Set	SA-MLLR	SA-CMLLR
MILD-BB	57.43	**MILD (BB-MH-FB)**	66.67	64.24
MILD-MH	62.16		66.97	65.76
MILD-FB	59.68		73.94	72.12
MILD-All	**59.76**		**69.19**	**67.37**
MODR-RK	38.74	**MODR (RK-RL-JF)**	48.79	40.3
MODR-RL	47.30		60.30	56.97
MODR-JF	58.56		63.94	61.21
MODR -All	**48.20**		**57.68**	**52.83**
SEVR-SC	57.66	**SEVR (SC-BK-BV)**	59.39	58.18
SEVR-BK	44.59		42.12	44.85
SEVR-BV	42.57		49.09	47.88
SEVR-All	**48.27**		**50.20**	**50.30**

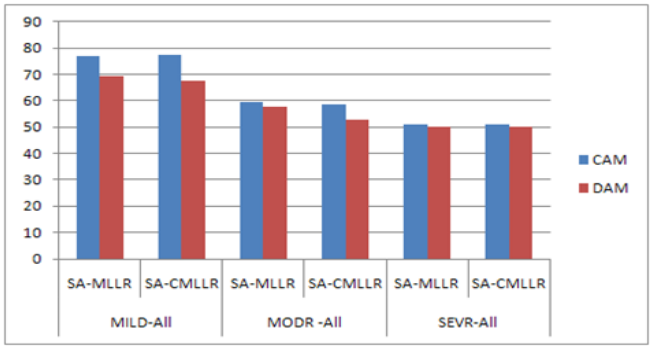

Figure 4. The comparative WRA for the two acoustic models CAM and DAM.

VI. CONCLUSIONS

As a result, the MLLR adaptation technique is preferred when only small data is available. The rich phonetic acoustic model performs better with high recognition accuracy. Hence, we can make use of the available phonetically rich speech corpus of normal speaker to adapt with limited dysarthric speech.

ACKNOWLEDGMENT

This research is supported by UM High Impact Research Grant UM-MOHE UM.C/HIR/MOHE/FCSIT/05 from the Ministry of Higher Education Malaysia.

REFERENCES

[1] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. S. Hawley, and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers," in *INTERSPEECH*, 2003.

[2] P. Kayasith, T. Theeramunkong, and N. Thubthong, "Recognition Rate Prediction for Dysarthric Speech Disorder Via Speech Consistency Score," in *PRICAI 2006: Trends in Artificial Intelligence*. vol. 4099, Q. Yang and G. Webb, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 885-889.

[3] Y. KM, B. DR, and T. CD, "Computerized assessment of intelligibility of dysarthric speech . Tigard, OR:C.C . Publications," 1984.

[4] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Computer Speech & Language*.

[5] H.-M. Liu, F.-M. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *The Journal of the Acoustical Society of America,* vol. 117, pp. 3879-3889, 2005.

[6] S. O. C. Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP Journal on Advances in Signal Processing,* vol. 2009, p. 2, 2009.

[7] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation,* pp. 1-19, 2011.

[8] S.-A. Selouani, M. S. Yakoub, and D. O'Shaughnessy, "Alternative speech communication system for persons with severe speech disorders," *EURASIP Journal on Advances in Signal Processing,* vol. 2009, p. 6, 2009.

[9] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech."

[10] E. Sanders, M. B. Ruiter, L. Beijer, and H. Strik, "Automatic recognition of dutch dysarthric speech: a pilot study," in *INTERSPEECH*, 2002.

[11] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication,* vol. 16, pp. 48-60, 2000.

[12] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, "Severity-Based Adaptation with Limited Data for ASR to Aid Dysarthric Speakers," *PLoS ONE,* vol. 9, p. e86285, 2014.

[13] M. Ferras, L. Cheung-Chi, C. Barras, and J. Gauvain, "Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, pp. 1366-1378, 2010.

[14] G. Jayaram and K. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks," *Journal of rehabilitation research and development,* vol. 32, pp. 162-162, 1995.

[15] K. Shinoda, "Acoustic model adaptation for speech recognition," *IEICE transactions on information and systems,* vol. 93, pp. 2348-2362, 2010.

[16] K. Shinoda, "Speaker Adaptation Techniques for Automatic Speech Recognition," *Proc. APSIPA ASC 2011 Xi'an,* 2011.

[17] J. Gauvain and L. Chin-Hui, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *Speech and Audio Processing, IEEE Transactions on,* vol. 2, pp. 291-298, 1994.

[18] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language,* vol. 9, p. 171, 1995.

[19] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *Speech and Audio Processing, IEEE Transactions on,* vol. 8, pp. 417-428, 2000.

[20] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *Speech and Audio Processing, IEEE Transactions on,* vol. 8, pp. 695-707, 2000.

[21] Y. Kai and M. J. F. Gales, "Discriminative cluster adaptive training," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 1694-1703, 2006.

[22] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language,* vol. 10, pp. 249-264, 10// 1996.

[23] C. J. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language,* vol. 9, pp. 171-185, 1995.

[24] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, pp. 1137-1140 vol.2.

[25] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *Speech and Audio Processing, IEEE Transactions on,* vol. 3, pp. 357-366, 1995.

[26] J. S. Garofolo and L. D. Consortium, *TIMIT: acoustic-phonetic continuous speech corpus*: Linguistic Data Consortium, 1993.

[27] P. Enderby, "Frenchay dysarthria assessment," *International journal of language & communication disorders,* vol. 15, pp. 165-173, 1980.

[28] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 4924-4927.

[29] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, 1996, pp. 1962-1965 vol.3.

[30] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders,* vol. 54, pp. 482-499, 1989.

[31] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu*, et al.*, *The HTK Book (For HTK Version 3.4)*: Microsoft Corporation and Cambridge University Engineering Department, 2009.

[32] H. Xuedong and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *Speech and Audio Processing, IEEE Transactions on,* vol. 1, pp. 150-157, 1993.